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

February 2019

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 is titled "The Coupling Time for the Ising Heat-Bath Dynamics." In it, we consider the Ising heat-bath Glauber dynamics on both the 1 dimensional cycle in Chapter 3, and on certain transitive graphs in Chapter 4. These dynamics describe a continuous time Markov chain, whose states are assignments of spins (either $+1$ or -1) to each vertex in a given graph. We construct a coupling of two such Markov chains, one starting with all spins $+1$ and one starting with all spins -1. The time it takes for these coupled chains to have the same spin configuration is the coupling time. Our main results show that the coupling time for these dynamics, when appropriately scaled, converges in distribution to a Gumbel distribution as the size of our graph goes to infinity. On the cycle, our results hold at all temperatures, and on certain transitive graphs, our results hold for sufficiently high temperatures. The two main tools we use are compound Poisson approximation, and the relatively new framework of information percolation, introduced by Lubetzky and Sly in [1]. We use Chapter 2 to summarise these techniques, as well as to introduce various preliminaries.

Part II is titled "Efficient Optimization for Statistical Inference." In this part, we look at two optimization problems that arise in statistical inference. In Chapter 6 we consider maximum likelihood mixtures where the components are parameterized by a single location parameter, and in Chapter 7 we look at a new method for deconvolution introduced by Delaigle and Hall in [2]. In each of these, we have to solve some optimization problem to find a discrete probability distribution. A key similarity between these two problems is that the optimal probability distribution is typically supported on a very small number of points. For the case of maximum likelihood mixtures, there are some results in the literature bounding the number of points in these optimal distributions; however we consider cases in which these bounds can be considerable overestimates. We provide various new results which provide tighter bounds in some cases, or extend known bounds to different classes of component densities. We compare these to empirical solutions. In the deconvolution case, the optimization problem is more complex than in the maximum likelihood setting, and obtaining and proving results concerning the number of points of support in the optimal distribution is difficult. However, we are still able to numerically explore the behaviour of these optimizing distributions and consider how we might take advantage of their properties.

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 [3] under the supervision of Wilhelm Lenz, who introduced the model in [4]. 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 [6] 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 [7]. 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 [5, 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 [5]. 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 $\beta > \beta_c(d)$ [5]. 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 [8]. 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 [9]. 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 [10, 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 [9, 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

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

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

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

The results of [9, 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

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

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 $d(t) \leq \bar{d}(t) \leq 2d(t)$ [11, Lemma 4.10].

On the complete graph, a complete characterisation of the mixing time as a function of temperature is obtained in [12]. On other graphs, the mixing time is treated in [13] for sufficiently high temperature. More recently, a series of papers by Lubetzky and Sly ([14], [15], [1], and [16]) 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

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

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

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" [11]. 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

$(\mathscr{B}_t, \mathscr{T}_t) = (\mathscr{B}^{\mathrm{DIS}}_{N_t}, \mathscr{T}^{\mathrm{DIS}}_{N_t})$ 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

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

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

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

(see [17, 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 [17, 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,

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

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

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

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

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

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 [17, 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 $c_n > 0$,

$$\begin{aligned}\mathbb{P}\left(\frac{N_n}{\lambda a_n} > \frac{4b_n}{a_n}\right) &\leq \mathbb{P}\left(\frac{T(N_n)}{a_n} > \frac{2b_n}{a_n}\right) + \mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \frac{1}{2}\right) &(2.51)\\ &= \mathbb{P}\left(Y_n > \frac{b_n}{a_n}\right) + \mathbb{P}\left(\left|\frac{T(N_n)\lambda}{N_n} - 1\right| > \frac{1}{2}\right). &(2.52)\end{aligned}$$

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 Proposition 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 [9] 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} [9], 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 [11], [18], and [19] 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 [1]. 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 [20] 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 [1], the best result known for general graphs was that cutoff occurs with a $\mathcal{O}(1)$ window in the simple case when $\beta = 0$ [21]. However, no results were known for $\beta > 0$, despite a conjecture by Peres in 2009 [11, 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 [14].

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 [16], 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 [22], 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 $Y_0 = (1, 1, \ldots, 1)$ and $Y_t = \mathscr{B}_t$ if $Y_0 = (-1, -1, \ldots, -1)$. 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 $Y_{t_1}, Y'_{t_1} \in \{-1, +1\}^V$

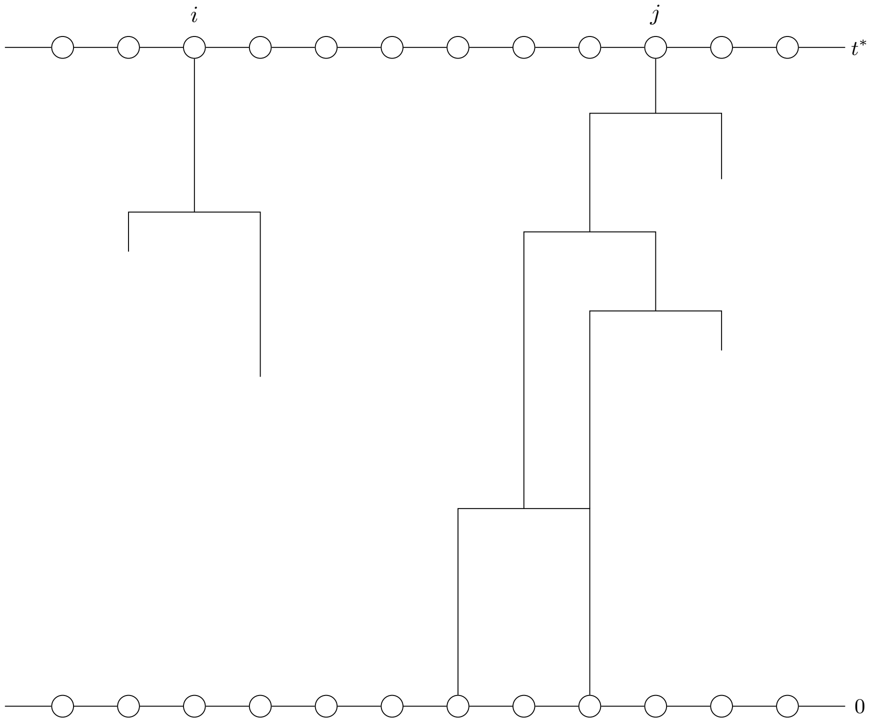


Figure 2.1 – *A section of the space-time slab $V \times [0, t_*]$ along with a typical appearance of the update histories for two vertices on the cycle. Time runs vertically from bottom to top, and the vertices are represented by circles, laid out horizontally. If there is a path in the update history 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 $V \times [0, t_*]$ 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*, $\mathscr{F}_{\text{UPD}}(A, t_1, t_2)$, to be the set of all vertices that A can 'reach' through the update function. That is, $i \in \mathscr{F}_{\text{UPD}}(A, t_1, t_2)$ if and only if there exists a subsequence of the updates, $(\mathcal{V}_k, U_k, S_k)$, such that $t_1 < S_1 < S_2, \dots, \leq S_m$ and $i, \mathcal{V}_1, \mathcal{V}_2, \dots, \mathcal{V}_m$ 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 $\mathscr{F}_{\text{UPD}}(i, t, t_*)$ backwards through time to create the analogous *update trace*, $\mathcal{G}_i$. It is clear that $\mathscr{F}(A, t_1, t_2) \subseteq \mathscr{F}_{\text{UPD}}(A, t_1, t_2)$ 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

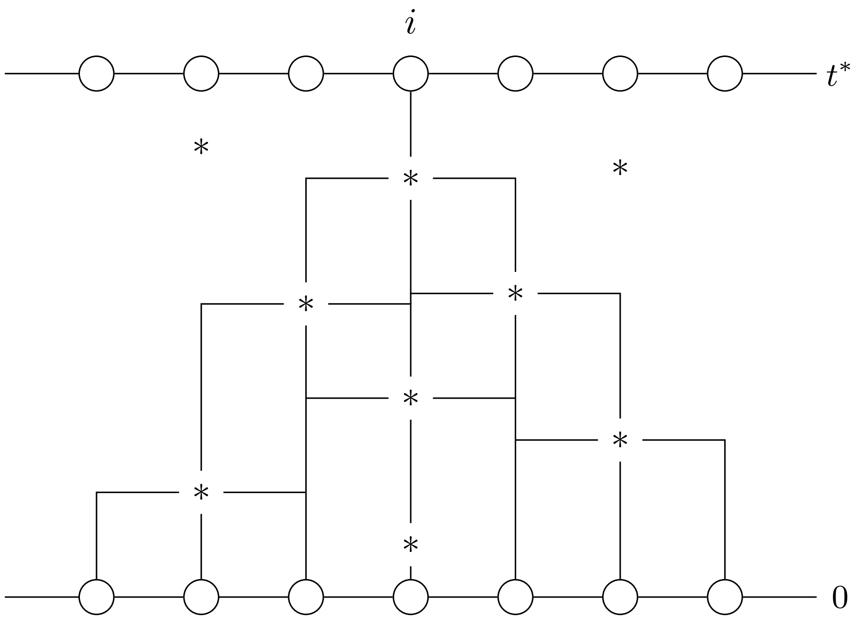


Figure 2.2 – *The update trace of i. Each update $(\mathcal{V}, U, t)$ in the update sequence is represented by a $*$ at $(\mathcal{V}, t)$.*

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 [1] 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

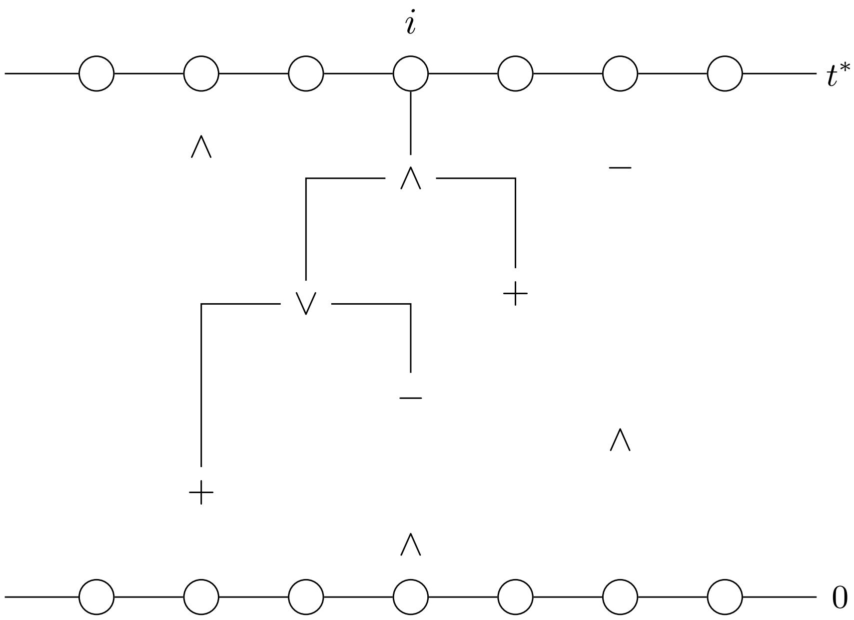


Figure 2.3 – *The update sequence for a section of the cycle and the corresponding update history from vertex i. For this particular update sequence, i takes a final spin of $+1$ regardless of the initial configuration.*

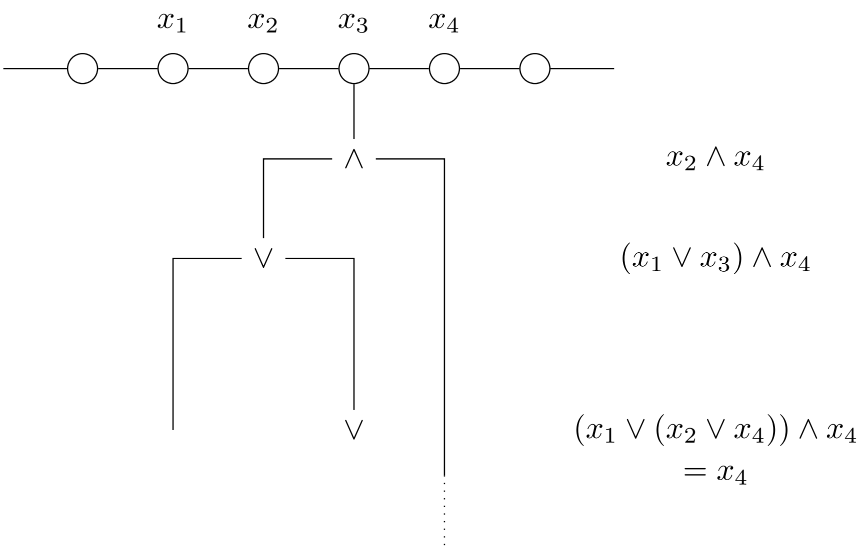


Figure 2.4 – *[Example taken from [1]]. A non-oblivious update that shrinks the size of the update history. On the right is written the final spin of x_3 as a function of the configuration at that time. The update $x_3 \mapsto x_2 \vee x_4$ causes the entire function to collapse to x_4, and so removes 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.

2.4 Compound Poisson Approximation

In addition to the framework of information percolation, we will also make use of 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].

A compound Poisson distribution, $\mathrm{CP}(\lambda, \boldsymbol{\mu})$, is the distribution of the sum of a rate λ Poisson number of independent random variables, each with distribution $\boldsymbol{\mu}$. That is, a compound Poisson random variable can be defined by

$$\mathrm{CP}(\lambda, \boldsymbol{\mu}) = \mathscr{L}\left(\sum_{j=1}^{M} Y_j\right) \tag{2.72}$$

for any $\lambda > 0$, and any probability distribution μ on $\mathbb{N}$, where the Y_j are independent and identically distributed with distribution $\boldsymbol{\mu}$ and are also independent of $M \sim \mathrm{Po}(\lambda)$.

The goal of compound Poisson approximation is to bound the distance between the distribution of a random variable W, and some canonically chosen compound Poisson distribution. A common scenario, and the one which we address here, is that

$$W = \sum_{i \in V} X_i \tag{2.73}$$

for some countable index set V, where each X_i is a nonnegative integer valued random variable. The following approach to compound Poisson approximation is taken from [23, Section 2.2].

For each $i \in V$, partition V into subsets $\{i\}$, B_i, C_i, and D_i and set

$$U_i = \sum_{j \in B_i} X_j, \qquad Z_i = \sum_{j \in C_i} X_j, \qquad W_i = \sum_{j \in D_i} X_j. \tag{2.74}$$

so that for each $i \in V$ we have the decomposition

$$W = X_i + U_i + Z_i + W_i. \tag{2.75}$$

For our approximation to be good, we want B_i to contain those X_j which strongly influence X_i, D_i to contain those X_j which have a neglibible effect on X_i and C_i to contain the remainder.

Using $I[\cdot]$ to denote the indicator function, we now define the canonical parameters λ and $\boldsymbol{\mu}$ by

$$\lambda = \sum_{i \in V} \mathbb{E}\left[\frac{X_i}{X_i + U_i} I[X_i + U_i \geq 1]\right], \tag{2.76}$$

$$\mu_l = \frac{1}{l\lambda} \sum_{i \in V} \mathbb{E}\left[X_i \, I[X_i + U_i = l]\right], \qquad l \geq 1. \tag{2.77}$$

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

$$\delta_1 = \sum_{i \in V} \sum_{k \geq 0} \mathbb{P}[X_i = 1, U_i = k] \mathbb{E}\left|\frac{\mathbb{P}[X_i = 1, U_i = k | W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1\right|, \tag{2.78}$$

$$\delta_4 = \sum_{i \in V} \left(\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]\right), \tag{2.79}$$

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

We can bound the total variation distance between W and the canonically chosen compound Poisson using the following theorem, reworked from [23, CPA 1A] with bounds for the constants provided there substituted from [23, Equation 2.17].

Theorem 2.4 ([23]). *Let W, λ, $\boldsymbol{\mu}$, δ_1 and δ_4 be as defined above. Then*

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{CP}(\lambda, \boldsymbol{\mu})) \leq (\delta_1 + \delta_4)e^{\lambda}. \tag{2.80}$$

Since the canonical parameter $\boldsymbol{\mu}$ is constructed to only place mass on the positive integers, we have that

$$\mathbb{P}\left(\mathrm{CP}(\lambda, \boldsymbol{\mu}) = 0\right) = \mathbb{P}\left(\mathrm{P}(\lambda) = 0\right) \tag{2.81}$$

$$= e^{-\lambda}. \tag{2.82}$$

This gives us the following corollary of Theorem 2.4.

Corollary 2.5. *Let W, λ, δ_1 and δ_4 be as defined above. Then*

$$\left|\mathbb{P}\left(W = 0\right) - e^{-\lambda}\right| \leq (\delta_1 + \delta_4)e^{\lambda}. \tag{2.83}$$

2.4.1 Application to our problem

We will now give an overview of how we will apply compound Poisson approximation to the problem of finding the limiting distribution of the coupling time of the Ising heat-bath process. The exact implementation will depend on the class of graphs on which the heat-bath process is running, and so we leave some of the specifics for the following chapters.

For a given graph, $G_L = (V, E)$, in a given sequence of graphs, $(G_L)_{L\geq 1}$, we fix $z \in \mathbb{R}$ and a time of interest

$$t_* = a_L z + d_L. \tag{2.84}$$

The specific choices for a_L and d_L depend on the sequence $(G_L)_{L\geq 1}$. We choose these such that for large enough L, $t_* > 0$, and then only consider these L. For each $i \in V$, define the indicator

$$X_i = I\left[\mathscr{B}_{t_*}[i] \neq \mathscr{T}_{t_*}[i]\right] \tag{2.85}$$

and set

$$W = \sum_{i\in V} X_i. \tag{2.86}$$

For each $i \in V$, we partition V into the sets $\{i\}, B_i, C_i$, and D_i defined by

$$B_i = \{j \neq i : d(i,j) \leq b_L\}, \tag{2.87}$$

$$C_i = \{j \notin B_i \cup \{i\} : d(i,j) \leq c_L\}, \tag{2.88}$$

$$D_i = V \setminus (B_i \cup C_i \cup \{i\}), \tag{2.89}$$

where we use $d(i,j)$ to denote the graph distance between vertices i and j. The specific choices for b_L and c_L again depend on the graph G_L.

We can now define U_i, Z_i, and W_i as above, along with $\lambda, \boldsymbol{\mu}, \delta_1$, and δ_4. We can then directly apply Theorem 2.4 to get that

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{CP}(\lambda, \boldsymbol{\mu})) \leq (\delta_1 + \delta_4)e^{\lambda}. \tag{2.90}$$

To relate W to the coupling time, observe that W is zero precisely when $T \leq t_*$ and so the events $\{W = 0\}$ and $\{T \leq t_*\}$ coincide. So using Corollary 2.5,

$$\left|\mathbb{P}\left(T \leq t_*\right) - e^{-\lambda}\right| \leq (\delta_1 + \delta_4)e^{\lambda}. \tag{2.91}$$

The bulk of the work is therefore in calculating λ, and in showing that δ_1 and δ_4 go to zero as $L \to \infty$. (Note that $W, T, t_*, \lambda, \boldsymbol{\mu}, \delta_1$, and δ_4 all depend on L but we have omitted this in the notation for convenience).

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 $\beta < \beta_c$ [10, 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 [5], 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 [25].

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,*

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

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

$$\frac{1}{2\sqrt{\frac{4}{\theta}-3}-1} \leq C_\theta \leq 1. \tag{3.2}$$

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

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

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 $\mathscr{B}_t[i] \le \mathscr{T}_t[i], \forall t$. 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.

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 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),

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

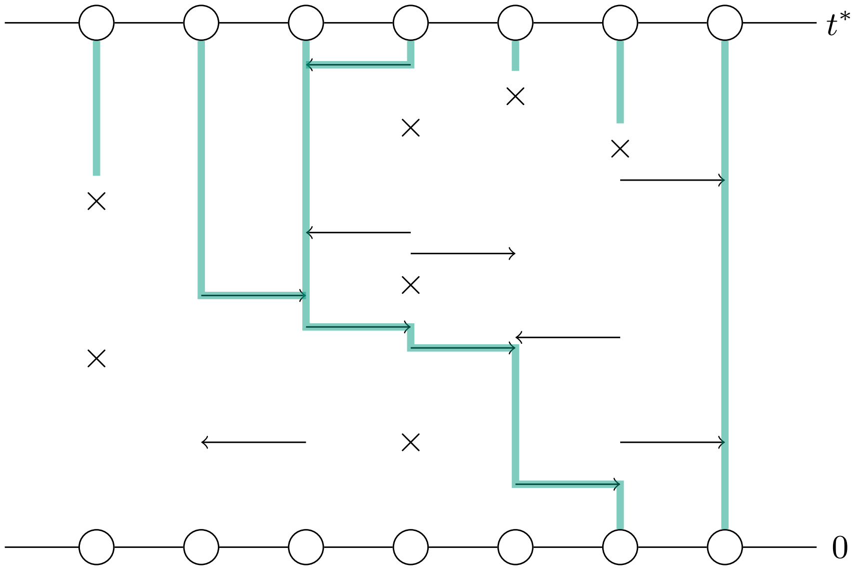


Figure 3.1 – *The update sequence for a section of the cycle represented by the thinned Poisson processes K_t^i ($\times$), L_t^i (left arrows), and R_t^i (right arrows) for each vertex. The corresponding update histories are overlaid in blue.*

Remark 3.2. The graphical representation described here follows the standard graphical representation for interacting particle systems. See [26, Chapter 3, Section 6] for an overview of these representations. In particular, our graphical representation for the Ising heat-bath dynamics on the cycle (and therefore the dynamics themselves) is exactly equivalent to that of the noisy voter model, studied in [27]. This equivalence does not hold for the Ising heat-bath dynamics on higher dimensional tori.

3.2 Problem set-up

In order to prove Theorem 3.1, we will use the method sketched out in Section 2.4.1. Here we provide the specific choices for the various quantities that were left unspecified there. The graph of interest here is of course $G_n = (\mathbb{Z}/n\mathbb{Z})$.

Fix $z \in \mathbb{R}$ and set

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

For any fixed $z \in \mathbb{R}$, $t_* > 0$ for all sufficiently large n and we only consider such n in what follows. We define X_i and W as in Section 2.4.1 using this choice of t_* and note that from (3.7) we get

$$\mathbb{P}[X_i = 1] = e^{-\theta t_*} = \frac{e^{-z}}{n}. \tag{3.9}$$

The vertex sets B_i, C_i, and D_i are chosen by

$$B_i = \{j \neq i : d(i,j) \leq b_n\}, \tag{3.10}$$

$$C_i = \{j \notin B_i \cup \{i\} : d(i,j) \leq c_n\}, \tag{3.11}$$

$$D_i = V \setminus (B_i \cup C_i \cup \{i\}), \tag{3.12}$$

with $b_n = \ln(n)$ and $c_n = \ln^2(n)$. We then define $U_i, Z_i, W_i, \lambda, \boldsymbol{\mu}, \delta_1$, and δ_4 as in Section 2.4.1 using these particular choices for B_i, C_i, and D_i. From (2.91), we get the following corollary of Theorem 2.4.

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*

$$\left| \mathbb{P}\left[T_n \leq \frac{z + \ln(n)}{\theta}\right] - e^{-\lambda} \right| \leq (\delta_1 + \delta_4)e^{\lambda}, \tag{3.13}$$

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, 3.7, and 3.8. 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.

Lemma 3.4. *Using the above set-up*

$$\limsup_{n\to\infty} \lambda \leq e^{-z} \tag{3.14}$$

and

$$\liminf_{n\to\infty} \lambda \geq C_\theta e^{-z} \tag{3.15}$$

for some

$$C_\theta \in \left[\frac{1}{2\sqrt{\frac{4}{\theta} - 3} - 1}, 1\right]. \tag{3.16}$$

Proof. Beginning with the definition of λ, we have

$$\lambda = \sum_{i \in V} \mathbb{E}\left[\frac{X_i}{X_i+U_i} I[X_i+U_i \geq 1]\right] \tag{3.17}$$

$$= \sum_{i=1}^{n} \mathbb{P}(X_i=1)\mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \tag{3.18}$$

$$= \sum_{i=1}^{n} \frac{e^{-z}}{n}\mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \tag{3.19}$$

$$= e^{-z}\mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \tag{3.20}$$

where we have used that X_i is either zero or one, (3.9), and the transitivity of the graph in each step respectively. Clearly

$$\mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \leq 1, \tag{3.21}$$

and we therefore obtain the upper bound.

By Jensen's inequality

$$\mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \geq \frac{1}{1+\mathbb{E}[U_i|X_i=1]}. \tag{3.22}$$

Now

$$\mathbb{E}[U_i|X_i=1] = \sum_{j \in B_i} \mathbb{P}[X_j=1|X_i=1] \tag{3.23}$$

$$= \sum_{k=1}^{\lfloor b_n \rfloor} \sum_{j \in \{i-k, i+k\}} \mathbb{P}[X_j=1|X_i=1] \tag{3.24}$$

$$= 2\sum_{k=1}^{\lfloor b_n \rfloor} \mathbb{P}[X_{i+k}=1|X_i=1] \tag{3.25}$$

where we have used the symmetry of X_{i+k} and X_{i-k} in the last step. From Lemma 3.13,

$$\mathbb{E}[U_i|X_i=1] \leq 2\sum_{k=1}^{\lfloor b_n \rfloor}\left(\frac{e^{-z}}{n} + 2\left(\frac{2-\theta-\sqrt{\theta(4-3\theta)}}{2-2\theta}\right)^k\right) \tag{3.26}$$

$$< 2\sum_{k=1}^{\lfloor b_n \rfloor}\frac{e^{-z}}{n} + 4\sum_{k=1}^{\infty}\left(\frac{2-\theta-\sqrt{\theta(4-3\theta)}}{2-2\theta}\right)^k \tag{3.27}$$

$$= \frac{2\lfloor b_n \rfloor}{n}e^{-z} + 2\left(\sqrt{\frac{4}{\theta}-3}-1\right). \tag{3.28}$$

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

$$\liminf_{n\to\infty} \mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \geq \frac{1}{2\sqrt{\frac{4}{\theta}-3}-1}. \tag{3.29}$$

□

Having bounded λ, we now turn to proving that it converges. Up until this point, we have chosen in our notation of several variables to omit references to their dependence on n. However, it will prove useful here, and in Lemma 3.10 in the next section, to use a superscript on some variables to make this dependence explicit.

Lemma 3.5. *The limit of λ^n as $n \to \infty$ exists.*

Proof. From Lemma 3.10 and 3.6, along with the fact that $0 \leq \mathbb{P}(U_i^n \leq k | X_i^n = 1) \leq 1$, we have that $U_i^n | X_i^n = 1$ converges in distribution. Via the Portmanteau Theorem [17, Theorem 25.8] we immediately obtain that

$$\lambda^n = e^{-z}\mathbb{E}\left[\frac{1}{1+U_i^n}\middle| X_i^n = 1\right] \tag{3.30}$$

converges. □

Lemma 3.6. *Let $(c_n)_n$ be a sequence of real numbers, with $c_n \geq 0$ for all n and $\lim_{n\to\infty} c_n = 0$. Let $(a_n)_n$ be a sequence of real numbers such that for all $m \in \mathbb{N}$, we have $a_n - a_{n+m} \geq -c_n$.*

If $|a_n|$ is uniformly bounded, then $\lim_{n\to\infty} a_n$ exists and is finite.

Proof. Since $|a_n|$ is uniformly bounded, $\liminf_n a_n$ and $\limsup_n a_n$ both exist and are finite. Let $(n_k)_k$ and $(l_k)_k$ be two increasing sequences of natural numbers such that $\lim_{k\to\infty} = \liminf_n a_n$, $\lim_{k\to\infty} a_{l_k} = \limsup_n a_n$, and $n_k < l_k$ for all k. We have that

$$a_{n_k} - a_{n_k+(l_k-n_k)} \geq -c_{n_k} \tag{3.31}$$

for all k. Hence

$$\liminf_n a_n - \limsup_n a_n \geq 0 \tag{3.32}$$

and so the limit exists. □

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 (3.13) to go to zero as $n \to \infty$. In the following proofs we use the notation $B(i, l)$ to mean the set of vertices at distance at most l from vertex i.

Lemma 3.7. *Let δ_1 be as defined above in (2.78). Then*

$$\lim_{n\to\infty} \delta_1 = 0. \tag{3.33}$$

Proof. Starting with the definition of δ_1, we have

$$\delta_1 = \sum_{i=1}^{n} \sum_{k=0}^{2\lfloor b_n \rfloor} \mathbb{P}[X_i = 1, U_i = k] \mathbb{E} \left| \frac{\mathbb{P}[X_i = 1, U_i = k | W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1 \right|, \tag{3.34}$$

$$= n \sum_{k=0}^{2\lfloor b_n \rfloor} \mathbb{E} \left| \mathbb{P}[X_i = 1, U_i = k | W_i] - \mathbb{P}[X_i = 1, U_i = k] \right| \tag{3.35}$$

by the transitivity of the cycle. Denote using

$$B(i, l) = \{j \in V : d(i, j) \leq l\} \tag{3.36}$$

the set of points within distance l of i and define the events

$$A_1 = \{\exists j \in B_i \cup \{i\}, \exists t \in [0, t_*] : \mathcal{H}_j(t) \nsubseteq B(i, (c_n + b_n)/2)\} \tag{3.37}$$

and

$$A_2 = \{\exists j \in D_i, \exists t \in [0, t_*] : \mathcal{H}_j(t) \cap B(i, (c_n + b_n)/2) \neq \emptyset\} \tag{3.38}$$

as well as their intersection

$$A = A_1 \cap A_2. \tag{3.39}$$

From Lemma 3.11,

$$\mathbb{P}[X_i = 1, U_i = j | A^c, W_i] = \mathbb{P}[X_i = 1, U_i = j | A^c]. \tag{3.40}$$

Continuing on from (3.35), we split the probabilities into

$$\delta_1 = n \sum_{k=0}^{2\lfloor b_n \rfloor} \mathbb{E}\,|\mathbb{P}[X_i = 1, U_i = k | W_i, A]\mathbb{P}[A|W_i] - \mathbb{P}[X_i = 1, U_i = k|A]\mathbb{P}[A] + \tag{3.41}$$

$$\mathbb{P}(X_i = 1, U_i = k|A^c)(\mathbb{P}[A^c|W_i] - \mathbb{P}[A^c])|$$

$$\leq n(2\lfloor b_n \rfloor + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + |\mathbb{P}[A^c|W_i] - \mathbb{P}[A^c]|\right] \tag{3.42}$$

$$= n(2\lfloor b_n \rfloor + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + |1 - \mathbb{P}[A|W_i] - (1 - \mathbb{P}[A])|\right] \tag{3.43}$$

$$\leq n(2\lfloor b_n \rfloor + 1)\mathbb{E}\left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \mathbb{P}[A|W_i] + \mathbb{P}[A])\right] \tag{3.44}$$

$$= 2n(2\lfloor b_n \rfloor + 1)\left(\mathbb{E}[\mathbb{P}[A|W_i]] + \mathbb{P}[A]\right) \tag{3.45}$$

$$= 4n(2\lfloor b_n \rfloor + 1)\mathbb{P}[A]. \tag{3.46}$$

For A_1 to hold, we must have a history $\mathcal{H}_j$ that extends from a distance of no more than b_n from i to a distance of at least $(b_n + c_n)/2$ from i. That is, it must extend a distance of at least $(c_n - b_n)/2$. For A_2 to hold, we must have a history $\mathcal{H}'_j$ that extends from a distance of at least c_n from i to a distance of no more than $(b_n + c_n)/2$ from i. That is, it must extend a distance of at least $(c_n - b_n)/2$. So for either A_1 or A_2 to hold, there must exist a history that spreads at least distance $(c_n - b_n)/2$ away from its starting vertex. By a union bound

$$\mathbb{P}[A] \leq \sum_{i=1}^{n} \mathbb{P}[\mathcal{H}_i \not\subseteq B(i, (c_n - b_n)/2) \times [0, t_*]] \tag{3.47}$$

$$= n\mathbb{P}\left[\bigcup_{u \in [0,t_*]} \mathcal{H}_i(t_* - u) \not\subseteq B(i, (c_n - b_n)/2)\right] \tag{3.48}$$

Combining this with Lemma 3.12, and recalling our choices of $b_n = \ln(n)$ and $c_n = \ln(n)^2$ we get that

$$\delta_1 \leq 8n^2(2\lfloor b_n \rfloor + 1)\exp\left((z + \ln n)/\theta - \ln 2(c_n - b_n)/2\right) \tag{3.49}$$

$$\leq 16\exp(z/\theta)n^{3+1/\theta+\ln 2/2-\ln n \ln 2/2} \tag{3.50}$$

where we have used that $2n > 2\ln n + 1$ for all $n \geq 1$. It is now clear that δ_1 goes to 0 as $n \to \infty$. □

Lemma 3.8. *Let δ_4 be as defined above in (2.79). Then*

$$\lim_{n\to\infty} \delta_4 = 0. \tag{3.51}$$

Proof. Starting with the definition of δ_4, we have

$$\delta_4 = \sum_{i=1}^{n} \left(\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]\right), \tag{3.52}$$

$$= n\mathbb{P}[X_i = 1]\mathbb{E}[Z_i | X_i = 1] + e^{-z} \sum_{j \in \{i\} \cup B_i \cup C_i} \mathbb{E}[X_j], \tag{3.53}$$

$$= e^{-z}\mathbb{E}[Z_i | X_i = 1] + \frac{(2\lfloor c_n \rfloor + 1)e^{-2z}}{n}. \tag{3.54}$$

Now

$$\mathbb{E}[Z_i | X_i = 1] = \sum_{j \in C_i} \mathbb{P}[X_j = 1 | X_i = 1], \tag{3.55}$$

$$= 2 \sum_{k=\lfloor b_n \rfloor + 1}^{\lfloor c_n \rfloor} \mathbb{P}[X_{i+k} = 1 | X_i = 1]. \tag{3.56}$$

From Lemma 3.13,

$$\mathbb{E}[Z_i | X_i = 1] \leq 2 \sum_{k=\lfloor b_n \rfloor + 1}^{\lfloor c_n \rfloor} \left(\frac{e^{-z}}{n} + 2 \left(\frac{2 - \theta - \sqrt{\theta(4 - 3\theta)}}{2 - 2\theta} \right)^k \right), \tag{3.57}$$

$$\leq \frac{2(c_n - b_n + 1)e^{-z}}{n} + 4(c_n - b_n + 1) \left(\frac{2 - \theta - \sqrt{\theta(4 - 3\theta)}}{2 - 2\theta} \right)^{b_n + 1}. \tag{3.58}$$

Altogether,

$$\delta_4 \leq \frac{2(c_n - b_n + 1)e^{-z}}{n} + 4(c_n - b_n + 1) \left(\frac{2 - \theta - \sqrt{\theta(4 - 3\theta)}}{2 - 2\theta} \right)^{b_n + 1} + \frac{(2c_n + 1)e^{-2z}}{n} \tag{3.59}$$

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

3.4 Additional lemmas

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

Lemma 3.9. *Let E be an event that depends only on the update sequence along $[t_1, t_*]$. Then for any vertex in the cycle, i, and any $t_0 \leq t_1$,*

$$\mathbb{P}\left(E|\mathcal{H}_i(t_1) \neq \emptyset\right) = \mathbb{P}\left(E|\mathcal{H}_i(t_0) \neq \emptyset\right) \tag{3.60}$$

Proof. Let $\bar{\mathcal{H}}_i$ be the update history that results from ignoring all oblivious updates (that is, the history does not die). Let $K_t^{\mathcal{H}_i}$ be the point process that places a point every time $\bar{\mathcal{H}}_i$ encounters an oblivious update. It is obvious that this is a Poisson process with waiting times that are exponential with rate θ. We have that the event $\{\mathcal{H}_i(t_1) \neq \emptyset\}$ coincides with the event $\{K_t^{\mathcal{H}_i}$ encounters no oblivious updates along $[t_1, t_*]\}$. Since the waiting times of $K_t^{\mathcal{H}_i}$ are exponential, conditioning on the behaviour of $K_t^{\mathcal{H}_i}$ along $[t_0, t_1)$ does not affect the behaviour of $K_t^{\mathcal{H}_i}$ along $[t_1, t_*]$. Since E only depends only on the update sequence along $[t_1, t_*]$, conditioning on the behaviour of $K_t^{\mathcal{H}_i}$ along $[t_0, t_1)$ does not affect its probability. Hence

$$\begin{aligned}
\mathbb{P}\left(E|\mathcal{H}_i(t_1) \neq \emptyset\right) &= \mathbb{P}\left(E|K_t^{\mathcal{H}_i} \text{ encounters no oblivious updates along } [t_1, t_*]\}\right) && (3.61)\\
&= \mathbb{P}\left(E|K_t^{\mathcal{H}_i} \text{ encounters no oblivious updates along } [t_0, t_*]\}\right) && (3.62)\\
&= \mathbb{P}\left(E|\mathcal{H}_i(t_0) \neq \emptyset\right). && (3.63)
\end{aligned}$$

□

Lemma 3.10. *For all $k \in \mathbb{R}$, any positive integer n, and any integer $m > n$,*

$$\begin{aligned}
&\mathbb{P}(U_i^n \leq k|X_i^n = 1) - \mathbb{P}(U_i^{n+m} \leq k|X_i^{n+m} = 1)\\
&\quad \geq -2e^{z(1+1/\theta)} n^{1/\theta + \ln 2 + z}(2\ln n + 1)\left(\frac{1}{2}\right)^{n/2} - 2e^{-z}\ln n/n. && (3.64)
\end{aligned}$$

Proof. On the length n cycle, let A_j be the event

$$A_j = \{\mathcal{H}_j^n \not\subseteq B(i, n/2 - 1) \times [0, t_*^n]\} \tag{3.65}$$

and define A to be

$$A = \bigcup_{j \in B_i^n \cup \{i\}} A_j. \tag{3.66}$$

Similarly, on the length $n + m$ cycle, define the time $t_0 = t_*^{n+m} - t_*^n$ and let α_j be the event

$$\alpha_j = \{\mathcal{H}_j^{n+m} \cap V^{n+m} \times [t_0, t_*^{n+m}] \not\subseteq B(i, n/2 - 1) \times [t_0, t_*^{n+m}]\} \tag{3.67}$$

and define α to be

$$\alpha = \bigcup_{j \in B_i^n \cup \{i\}} \alpha_j. \tag{3.68}$$

We have

$$\begin{aligned}
&\mathbb{P}(U_i^n \leq k | X_i^n = 1) - \mathbb{P}(U_i^{n+m} \leq k | X_i^{n+m} = 1) \\
&\quad = \mathbb{P}(U_i^n \leq k, A | X_i^n = 1) - \mathbb{P}(U_i^{n+m} \leq k, \alpha | X_i^{n+m} = 1) \\
&\qquad + \mathbb{P}(U_i^n \leq k, A^{\mathrm{c}} | X_i^n = 1) - \mathbb{P}(U_i^{n+m} \leq k, \alpha^{\mathrm{c}} | X_i^{n+m} = 1) \qquad (3.69)\\
&\quad \geq -\mathbb{P}(\alpha | X_i^{n+m} = 1) \\
&\qquad + \mathbb{P}(U_i^n \leq k, A^{\mathrm{c}} | X_i^n = 1) - \mathbb{P}(U_i^{n+m} \leq k, \alpha^{\mathrm{c}} | X_i^{n+m} = 1). \qquad (3.70)
\end{aligned}$$

For a given vertex i, we may couple the histories on the length n cycle and the length $n+m$ cycle in the following way. Recall that the update sequence on the cycle may be represented using the Poisson processes K_t^i, L_t^i, and R_t^i, as introduced in Section 3.1.1. Let K_t^j, L_t^j, and R_t^j, $j \in \{1, \ldots, n+m\}$, be the Poisson processes which represent the update sequence on the length $n+m$ cycle. Let $\hat{K}_t^j$ be the Poisson processes resulting from taking the updates of K_t^j in the interval $[t_0, t_*^{n+m}]$ and temporally shifting them onto the interval $[0, t_*^n]$. Likewise, define $\hat{L}_t^j$ and $\hat{R}_t^j$ in the same way. On the size n cycle, we use the processes $\hat{K}_t^j, \hat{L}_t^j$, and $\hat{R}_t^j$ for $j \in [i - \lceil n/2 - 1 \rceil, i + \lfloor n/2 \rfloor]$.

The picture here is that we construct the update sequence on the length n cycle from the top part of the update sequence on the length $n+m$ cycle. Of course, we must discard the update sequences of some vertices from the $n+m$ cycle and we choose to discard the update sequences of those vertices that are furthest away from i.

Under this coupling, if α^{c} holds, then we note that the histories of vertices in $B_i \cup \{i\}$ on the length n cycle are equal to the histories along the interval $[t_0, t_*^{n+m}]$ of the same vertices on the length $n+m$ cycle. Hence, if α^{c} holds, for $j \in B_i \cup \{i\}$, $X_j^n = 1$ if and only if $\mathcal{H}_j^{n+m}(t_0) \neq 0$. We also have that α^{c} holds if and only if A^{c} holds. Therefore,

$$\mathbb{P}(U_i^n \leq k, A^{\mathrm{c}} | X_i^n = 1) = \mathbb{P}\left(\sum_{j \in B_i^n} I(\mathcal{H}_j^{n+m}(t_0) \neq \emptyset) \leq k, \alpha^{\mathrm{c}} \,\middle|\, \mathcal{H}_i^{n+m}(t_0) \neq \emptyset \right) \tag{3.71}$$

$$= \mathbb{P}\left(\sum_{j \in B_i^n} I(\mathcal{H}_j^{n+m}(t_0) \neq \emptyset) \leq k, \alpha^{\mathrm{c}} \,\middle|\, \mathcal{H}_i^{n+m}(0) \neq \emptyset \right) \tag{3.72}$$

$$= \mathbb{P}\left(\sum_{j \in B_i^n} I(\mathcal{H}_j^{n+m}(t_0) \neq \emptyset) \leq k, \alpha^{\mathrm{c}} \,\middle|\, X_i^{n+m} = 1 \right). \tag{3.73}$$

where we have used Lemma 3.9 to go from (3.71) to (3.72).

Define

$$V_i = \sum_{j \in B_i^n} I[\mathcal{H}_i^{n+m} \cap \mathcal{H}_j^{n+m} \cap \left(V \times [t_0, t_*^{n+m}]\right) \neq \emptyset], \tag{3.74}$$

the number of $j \in B_i^n$ whose histories intersect with $\mathcal{H}_i$ in $[t_0, t_*]$, and

$$Y_i = \sum_{j \in B_i^n} I[\mathcal{H}_j^{n+m}(t_0) \neq \emptyset] I[\mathcal{H}_i^{n+m} \cap \mathcal{H}_j^{n+m} \cap \left(V \times [t_0, t_*^{n+m}]\right) = \emptyset], \tag{3.75}$$

the number of $j \in B_i^n$ whose histories survive to time t_0 without intersecting with $\mathcal{H}_i$, as well as

$$\bar{V}_i = \sum_{j \in B_i^n} I[\mathcal{H}_i^{n+m} \cap \mathcal{H}_j^{n+m} \neq \emptyset] \tag{3.76}$$

and

$$\bar{Y}_i = \sum_{j \in B_i^n} X_j^{n+m} I[\mathcal{H}_i^{n+m} \cap \mathcal{H}_j^{n+m} = \emptyset] \tag{3.77}$$

the analogous quantities defined with t_0 replaced by $t = 0$.

Clearly, if $X_i^{n+m} = 1$, we have

$$\sum_{j \in B_i^n} I(\mathcal{H}_j^{n+m}(t_0) \neq \emptyset) = V_i + Y_i \tag{3.78}$$

and

$$U_i^{n+m} = \bar{V}_i + \bar{Y}_i. \tag{3.79}$$

Also note that under the coupling given above, if α^{c} holds, then $V_i \leq \bar{V}_i$. Now we can bound the last two terms of (3.70). We have

$$\begin{aligned}
&\mathbb{P}(U_i^n \leq k, A^{\mathrm{c}} | X_i^n = 1) - \mathbb{P}(U_i^{n+m} \leq k, \alpha^{\mathrm{c}} | X_i^{n+m} = 1) \\
&= \mathbb{P}\left(V_i + Y_i \leq k, \alpha^{\mathrm{c}} \middle| X_i^{n+m} = 1\right) - \mathbb{P}(\bar{V}_i + \bar{Y}_i \leq k, \alpha^{\mathrm{c}} | X_i^{n+m} = 1) &(3.80)\\
&\geq \mathbb{P}\left(V_i + Y_i \leq \bar{V}_i + \bar{Y}_i, \bar{V}_i + \bar{Y}_i \leq k, \alpha^{\mathrm{c}} \middle| X_i^{n+m} = 1\right) \\
&\quad - \mathbb{P}(\bar{V}_i + \bar{Y}_i \leq k, \alpha^{\mathrm{c}} | X_i^{n+m} = 1) &(3.81)\\
&\geq \mathbb{P}\left(\bar{V}_i + \bar{Y}_i \leq k, \alpha^{\mathrm{c}} \middle| X_i^{n+m} = 1\right) - \mathbb{P}\left(V_i + Y_i > \bar{V}_i + \bar{Y}_i, \alpha^{\mathrm{c}} \middle| X_i^{n+m} = 1\right) \\
&\quad - \mathbb{P}(\bar{V}_i + \bar{Y}_i \leq k, \alpha^{\mathrm{c}} | X_i^{n+m} = 1) &(3.82)\\
&\geq -\mathbb{P}\left(V_i + Y_i > \bar{V}_i + \bar{Y}_i, \alpha^{\mathrm{c}} \middle| X_i^{n+m} = 1\right) &(3.83)\\
&\geq -\mathbb{P}\left(Y_i > \bar{Y}_i | X_i^{n+m} = 1\right) &(3.84)\\
&\geq -\mathbb{P}\left(Y_i \neq 0 | X_i^{n+m} = 1\right) &(3.85)
\end{aligned}$$

since conditioned on $X_i^{n+m} = 1$, if α^c holds then $V_i \leq \bar{V}_i$. By Lemma 3.9, since Y_i only depends on the update sequence along $[t_0, t_*^{n+m}]$, we may replace the conditioning above by

$$-\mathbb{P}\left(Y_i \neq 0 | X_i^{n+m} = 1\right) = -\mathbb{P}\left(Y_i \neq 0 | \mathcal{H}_i^{n+m}(t_0) \neq \emptyset\right). \tag{3.86}$$

Now,

$$\begin{aligned}
\mathbb{P}\left(Y_i \neq 0 | \mathcal{H}_i^{n+m}(t_0) \neq \emptyset\right) \leq \sum_{j \in B_i^n} \mathbb{P}\Big(&\mathcal{H}_j^{n+m}(t_0) \neq \emptyset, \ldots \\
&\ldots \mathcal{H}_i^{n+m} \cap \mathcal{H}_j^{n+m} \cap (V \times [t_0, t_*^{n+m}]) = \emptyset \,\big|\mathcal{H}_i^{n+m}(t_0) \neq \emptyset\Big) \qquad (3.87)\\
= \mathbb{P}(\mathcal{H}_i^{n+m}(t_0) \neq \emptyset)^{-1} \sum_{j \in B_i^n} \mathbb{P}\Big(&\mathcal{H}_j^{n+m}(t_0) \neq \emptyset, \ldots \\
&\ldots \mathcal{H}_i^{n+m} \cap \mathcal{H}_j^{n+m} \cap (V \times [t_0, t_*^{n+m}]) = \emptyset, \mathcal{H}_i^{n+m}(t_0) \neq \emptyset\Big) \qquad (3.88)
\end{aligned}$$

and by Lemma 3.14 we get

$$\mathbb{P}(Y_i \neq 0 | \mathcal{H}_i^{n+m}(t_0) \neq \emptyset) \leq |B_i^n| \mathbb{P}(\mathcal{H}_i^{n+m}(t_0) \neq \emptyset) \tag{3.89}$$

We also observe that the event $\{\mathcal{H}_i^{n+m}(t_0) \neq \emptyset\}$ is simply the event that $\mathcal{H}_i$ does not encounter an obvlivious update over an interval of length t_*^n. So

$$\begin{aligned}
\mathbb{P}(Y_i \neq 0 | \mathcal{H}_i^{n+m}(t_0) \neq \emptyset) &\leq |B_i^n| e^{-\theta t_*^n} &(3.90)\\
&= e^{-z} |B_i^n| / n. &(3.91)
\end{aligned}$$

Turning to the other term of interest in (3.70) we have

$$\begin{aligned}
\mathbb{P}(\alpha | X_i^{n+m} = 1) &\leq \sum_{j \in B_i^n \cup \{i\}} \mathbb{P}(\alpha_j | X_i^{n+m} = 1) &(3.92)\\
&= \sum_{j \in B_i^n \cup \{i\}} \mathbb{P}(\alpha_j | \mathcal{H}_i^{n+m}(t_0) \neq \emptyset) &(3.93)\\
&\leq \mathbb{P}(\mathcal{H}_i^{n+m}(t_0) \neq \emptyset)^{-1} \sum_{j \in B_i^n \cup \{i\}} \mathbb{P}(\alpha_j). &(3.94)
\end{aligned}$$

For α_j to occur, $j \in B_i^n \cup \{i\}$, the history of vertex j must reach a vertex at distance at least $n/2 - b_n$ from j in time t_*^n. By Lemma 3.12 we have

$$\mathbb{P}(\alpha_j) \leq 2 \exp(t_*^n - \ln 2(n/2 - b_n)). \tag{3.95}$$

and so,

$$\mathbb{P}(\alpha|X_i^{n+m}=1) \leq ne^z(|B_i|+1)2\exp(t_*^n - \ln 2(n/2-b_n)). \tag{3.96}$$

So overall we have

$$\begin{aligned}&\mathbb{P}(U_i^n \leq k|X_i^n=1) - \mathbb{P}(U_i^{n+m} \leq k|X_i^{n+m}=1)\\ &\quad \geq -ne^z(|B_i|+1)2\exp(t_*^n - \ln 2(n/2-b_n)) - e^{-z}|B_i^n|/n,\end{aligned} \tag{3.97}$$

which, recalling our choices of $b_n = \ln n$ and $t_*^n = (z+\ln n)/\theta$, becomes

$$\begin{aligned}&\mathbb{P}(U_i^n \leq k|X_i^n=1) - \mathbb{P}(U_i^{n+m} \leq k|X_i^{n+m}=1)\\ &\quad \geq -2e^{z(1+1/\theta)}n^{1/\theta+\ln 2+z}(2\ln n+1)\left(\frac{1}{2}\right)^{n/2} - 2e^{-z}\ln n/n.\end{aligned} \tag{3.98}$$

□

Lemma 3.11. *Let i be a vertex in a graph and define the events*

$$A_1 = \{\exists j \in B_i \cup \{i\}, \exists t \in [0,t_*] : \mathcal{H}_j(t) \not\subseteq B(i,(b_n+c_n)/2)\} \tag{3.99}$$

and

$$A_2 = \{\exists j \in D_i, \exists t \in [0,t_*] : \mathcal{H}_j(t) \cap B(i,(b_n+c_n)/2) \neq \emptyset\} \tag{3.100}$$

as well as their union

$$A = A_1 \cup A_2. \tag{3.101}$$

Then

$$\mathbb{P}[X_i=1, U_i=j|A^c, W_i] = \mathbb{P}[X_i=1, U_i=j|A^c]. \tag{3.102}$$

Proof. If A_1^c holds, then the events $\{X_i=1\}$ and $\{U_i=j\}$ depend only on the values of the update sequence inside $B(i,(b_n+c_n)/2)$. If A_2^c holds then the events $\{W_i=k\}$, $k \geq 0$, depend only on the values of the update sequence outside of $B(i,(b_n+c_n)/2)$. Since the update sequences of each vertex are independent of each other vertex, if A_2^c holds, conditioning on W_i does not affect the update sequences inside $B(i,(b_n+c_n)/2)$ and so

$$\mathbb{P}[X_i=1, U_i=j|A^c, W_i] = \mathbb{P}[X_i=1, U_i=j|A^c]. \tag{3.103}$$

□

The following Lemma bounds how fast updates can percolate through the cycle. The proof is a slight modification of a similar result in [1, Lemma 2.1].

Lemma 3.12. *Let i be a vertex on the length n cycle. The probability that the update history of vertex i escapes $B(i,l)$ in time s satisfies*

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i,l)\right] \leq 2\exp\left(s - l\ln 2\right). \tag{3.104}$$

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_+. For M_- to occur, first i must update, then $i-1$ must update, and so on. Each vertex updates at rate 1, and there are l updates required within time s. So the probability of this event is the probability that a rate 1 Poisson clock goes off at least l times in time s, or equivalently

$$\mathbb{P}[M_-] = \mathbb{P}[M_+] = \mathbb{P}[Y \geq l] \tag{3.105}$$

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

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i,l)\right] \leq 2\mathbb{P}[\mathrm{Po}(s) \geq l]. \tag{3.106}$$

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

$$M(t) = \exp\left(s\left(e^t - 1\right)\right). \tag{3.107}$$

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

$$\mathbb{P}[\mathrm{Po}(s) \geq l] \leq \exp\left(s\left(e^t - 1\right) - tl\right). \tag{3.108}$$

Overall we have

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i,l)\right] \leq 2\exp\left(s\left(e^t - 1\right) - tl\right). \tag{3.109}$$

Choosing $t = \ln 2$ then yields

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i,l)\right] \leq 2\exp\left(s - l\ln 2\right). \tag{3.110}$$

□

Lemma 3.13. *Let i and j be two vertices on the cycle $(\mathbb{Z}/n\mathbb{Z})$ separated by distance k. Then*

$$\mathbb{P}[X_j = 1 | X_i = 1] \leq \frac{e^{-z}}{n} + 2\left(\frac{2-\theta-\sqrt{\theta(4-3\theta)}}{2-2\theta}\right)^k. \tag{3.111}$$

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

$$\begin{aligned} \mathbb{P}[X_j = 1 | X_i = 1] &= \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] \\ &\quad + \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] &(3.112)\\ &\leq \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] + \mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1]. &(3.113) \end{aligned}$$

The result follows from Lemmas 3.15 and 3.16. □

Lemma 3.14. *Let i and j be two vertices in a transitive graph, $G = (V, E)$. Let $0 \leq t_0 < t_*$. Then*

$$\mathbb{P}\left(\mathcal{H}_j(t_0) \neq \emptyset, \mathcal{H}_i(t_0) \neq \emptyset, \mathcal{H}_j \cap \mathcal{H}_i \cap (V \times [t_0, t_*]) \neq \emptyset\right) \leq \mathbb{P}[\mathcal{H}_i(t_0) \neq \emptyset]^2. \tag{3.114}$$

Proof. Define S to be the time closest to t_* that the histories of i and j intersect, or define $S = t_0$ if the histories do not intersect along $[t_0, t_*]$. We construct a new history, $\mathcal{H}'_j$, in the following way. For all $t \in [S, t_*]$, $\mathcal{H}'_j(t) = \mathcal{H}_j(t)$. However, along the interval $[t_0, S)$, we replace the update sequence with an another i.i.d. copy of the update sequence along the interval $[t_0, S)$ and construct $\mathcal{H}'_j$ using this new update sequence.

Clearly, $\mathcal{H}'_j$ and $\mathcal{H}_j$ have the same distribution. Moreover, we now argue that $\mathcal{H}'_j$ is independent of $\mathcal{H}_i$. By construction, $\cup_{t=0}^{S} \mathcal{H}'_j(t)$ is independent of $\cup_{t=0}^{S} \mathcal{H}_k(t)$ for all $k \in V$, and in particular, for $k = i$. For $t \geq S$ we note that no updates are a part of both histories. Since every update comes from independent Poisson processes, the updates in $\mathcal{H}'_j \cap (V \times [S, t^*])$ are independent of those in $\mathcal{H}_i \cap (V \times [S, t^*])$. Hence $\mathcal{H}'_j$ and $\mathcal{H}_i$ are independent.

Note that in the event $\{\mathcal{H}_i \cap \mathcal{H}_j \cap (V \times [t_0, t_*]) = \emptyset\}$, we have $\mathcal{H}_j = \mathcal{H}'_j$ along the interval $[t_0, t_*]$ and so

$$\begin{aligned}
&\mathbb{P}[\mathcal{H}_i(t_0) \neq \emptyset, \mathcal{H}_j(t_0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j \cap (V \times [t_0, t_*]) = \emptyset] \\
&\quad = \mathbb{P}[\mathcal{H}_i(t_0) \neq \emptyset, \mathcal{H}'_j(t_0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j \cap (V \times [t_0, t_*]) = \emptyset] && (3.115) \\
&\quad \leq \mathbb{P}[\mathcal{H}_i(t_0) \neq \emptyset, \mathcal{H}'_j(t_0) \neq \emptyset] && (3.116) \\
&\quad = \mathbb{P}[\mathcal{H}_i(t_0) \neq \emptyset]\mathbb{P}[\mathcal{H}'_j(t_0) \neq \emptyset] && (3.117) \\
&\quad = \mathbb{P}[\mathcal{H}_i(t_0) \neq \emptyset]^2 && (3.118)
\end{aligned}$$

since $\mathcal{H}_i$ and $\mathcal{H}'_j$ are independent and $\mathbb{P}[\mathcal{H}_i(t_0) \neq \emptyset] = \mathbb{P}[\mathcal{H}'_j(t_0) \neq \emptyset]$ by the transitivity of G. □

Lemma 3.15. *Let i and j be the indices of two vertices on the cycle $(\mathbb{Z}/n\mathbb{Z})$. Then*

$$\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] \leq e^{-z}/n. \tag{3.119}$$

Proof. To begin

$$\begin{aligned}
\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] &= \mathbb{P}[X_i = 1]^{-1}\mathbb{P}[X_i = 1, X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] && (3.120) \\
&= \frac{\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset, \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset]}{\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]}. && (3.121)
\end{aligned}$$

By Lemma 3.14,

$$\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset, \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] \leq \mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]^2. \tag{3.122}$$

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

Lemma 3.16. *Let i and j be two vertices on the length n cycle that are separated by distance k. Then*

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_j = 1] \leq 2\left(\frac{2 - \theta - \sqrt{\theta(4 - 3\theta)}}{2 - 2\theta}\right)^k. \tag{3.123}$$

Proof. We must first deal with the effect that conditioning on $X_j = 1$ has on the probability that the two update histories merge. Recall from 3.1.1 that we could interpret the update history as a collection of independent Poisson processes for each vertex. We may construct a different collection of filtered Poisson processes for each vertex by following the updates that occur along a single vertex's history. Let $\bar{\mathcal{H}}_i$ be the update history that results from ignoring all oblivious updates (that is, the history does not die). Then $L_t^{\mathcal{H}_i}$ is the Poisson process that places a point every time $\bar{\mathcal{H}}_i$ encounters a left update, $R_t^{\mathcal{H}_i}$

is the Poisson process that places a point every time $\bar{\mathcal{H}}_i$ encounters a right update, and $K_t^{\mathcal{H}_i}$ is the Poisson process that places a point every time $\bar{\mathcal{H}}_i$ encounters an oblivious update. It is easy to see that these are Poisson processes with waiting times that are exponential with rate 1, and that $L_t^{\mathcal{H}_i}$, $R_t^{\mathcal{H}_i}$, and $K_t^{\mathcal{H}_i}$ are all independent of each other. Furthermore, for two vertices i and j, while $\mathcal{H}_i$ and $\mathcal{H}_j$ do not share any updates, $L_t^{\mathcal{H}_i}$, $R_t^{\mathcal{H}_i}$, and $K_t^{\mathcal{H}_i}$ are all independent of each of $L_t^{\mathcal{H}_j}$, $R_t^{\mathcal{H}_j}$, and $K_t^{\mathcal{H}_j}$.

We now see that conditioning on $X_j = 1$ is equivalent to conditioning that $K_t^{\mathcal{H}_j}$ has no updates along $[0, t_*]$. Since $L_t^{\mathcal{H}_j}$ and $R_t^{\mathcal{H}_j}$ are independent of $K_t^{\mathcal{H}_j}$, conditioning on $X_j = 1$ does not affect the left and right updates and so we have that

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_j = 1] = \mathbb{P}[\mathcal{H}_i \cap \bar{\mathcal{H}}_j \neq \emptyset]. \tag{3.124}$$

Consider a process which measure the distance between the two modified histories $\bar{\mathcal{H}}_i$ and $\bar{\mathcal{H}}_j$,

$$P(t) = d(\bar{\mathcal{H}}_i(t_* - t), \bar{\mathcal{H}}_j(t_* - t)), \qquad 0 \leq t \leq t_*. \tag{3.125}$$

Define the time S_0 via $S_0 = \inf\{t : P(t) = 0\}$ or $S_0 = \infty$ if $P(t) > 0$ for all $t \in [0, t_*]$. Define the time S_d via $S_d = \inf\{t : \mathcal{H}_i(t_* - t) = \emptyset\}$ or $S_d = t_*$ if $\mathcal{H}_i(t) \neq \emptyset$ for all $t \in [0, t_*]$. Then

$$\mathbb{P}[\mathcal{H}_i \cap \bar{\mathcal{H}}_j \neq \emptyset] = \mathbb{P}[S_0 \leq S_d]. \tag{3.126}$$

For most configurations of $\bar{\mathcal{H}}_i(t)$ and $\bar{\mathcal{H}}_j(t)$, while $P(t) > 0$, $P(t)$ reduces by 1 at rate $1 - \theta$ and increases by 1 at rate $1 - \theta$. The only exception being when $\bar{\mathcal{H}}_i(t_* - t)$ and $\bar{\mathcal{H}}_j(t_* - t)$ are maximally separated on the cycle, in which case the distance between them cannot increase.

To account for this, we construct two related birth and death processes. We assume that i is to the left of j (if it is to the right, we may simply swap 'left' and 'right' in what follows). The first, $P_1(t)$, starts at $k = d(i, j)$ and increases by one every time $\bar{\mathcal{H}}_i$ moves left or $\bar{\mathcal{H}}_j$ moves right and decreases by one every time $\bar{\mathcal{H}}_i$ moves right or $\bar{\mathcal{H}}_j$ moves left. The second, $P_2(t)$, starts at $n - k$ and decreases by one every time $P_1(t)$ increases by one, and increases by one every time $P_1(t)$ decreases by one. We may also construct times $S_0^{P_1}$ and $S_0^{P_2}$ as above to be the first times that P_1 and P_2 reach 0 (or ∞ if they do not reach 0). The events $\{P(t) = 0\}$ and $\{P_1(t) = 0\} \cup \{P_2(t) = 0\}$ coincide and so

$$\begin{aligned}
\mathbb{P}[\mathcal{H}_i \cap \bar{\mathcal{H}}_j \neq \emptyset] &= \mathbb{P}[\min(S_0^{P_1}, S_0^{P_2}) \leq S_d] && (3.127)\\
&\leq \mathbb{P}[S_0^{P_1} \leq S_d] + \mathbb{P}[S_0^{P_2} \leq S_d] && (3.128)\\
&\leq 2\mathbb{P}[S_0^{P_1} \leq S_d] && (3.129)
\end{aligned}$$

since $P_1(t)$ starts closer to 0 than $P_2(t)$ does. Let S_θ be exponentially distributed with rate θ, and independent of $S_0^{P_1}$. We have that $S_d \stackrel{\mathrm{d}}{=} \min(S_\theta, t_*)$, and since both S_θ and S_d are independent of $S_0^{P_1}$, we have

$$\mathbb{P}[\mathcal{H}_i \cap \bar{\mathcal{H}}_j \neq \emptyset] \leq 2\mathbb{P}[S_0^{P_1} \leq \min(S_\theta, t_*)] \tag{3.130}$$

$$\leq 2\mathbb{P}[S_0^{P_1} \leq S_\theta]. \tag{3.131}$$

In the case that $P_1(t) > 0$ for all $t \in [0, t_*]$, instead of defining $S_0^{P_1}$ to be infinite, we may instead choose to extend $P_1(t)$ into the interval (t_*, ∞) by continuing the birth and death transitions with rates $\lambda = \mu = 1 - \theta$, independent of S_θ. Call this extended process $P_1'(t)$. We can then define $S_0^{P_1'} = \inf\{t : P_1'(t) = 0\}$ which is constructed such that $S_0^{P_1'} \leq S_0^{P_1}$. We therefore have

$$\mathbb{P}[\mathcal{H}_i \cap \bar{\mathcal{H}}_j \neq \emptyset] \leq 2\mathbb{P}[S_0^{P_1'} \leq S_\theta]. \tag{3.132}$$

While $t < S_\theta$ and $P_1'(t) > 0$, there are three possibilities for what can happen to P_1' next. Either the next event is a birth with probability $(1-\theta)/(2-\theta)$, the next event is a death with the same probability or we reach time S_θ with probability $\theta/(2-\theta)$. Writing $\zeta_k = \mathbb{P}_k(S_0^{P_1'} < S_\theta)$, where $\mathbb{P}_k$ denotes that $P_1(t)$ starts at k, this gives us the recurrence relation

$$\zeta_k = \frac{1-\theta}{2-\theta}\zeta_{k-1} + \frac{1-\theta}{2-\theta}\zeta_{k+1} \tag{3.133}$$

which is subject to the conditions

$$\zeta_0 = 1 \tag{3.134}$$

$$\zeta_k \leq 1, \forall k \in \mathbb{N}. \tag{3.135}$$

This recurrence has characteristic equation

$$x^2 - \frac{2-\theta}{1-\theta}x + 1 = 0 \tag{3.136}$$

which has roots

$$r_1 = \frac{2-\theta+\sqrt{\theta(4-3\theta)}}{2-2\theta} \tag{3.137}$$

$$r_2 = \frac{2-\theta-\sqrt{\theta(4-3\theta)}}{2-2\theta} \tag{3.138}$$

and so

$$\zeta_k = ar_1^k + br_2^k \tag{3.139}$$

where a and b are constants to be determined from (3.134) and (3.135). We note that $r_1 \geq 1, \forall \theta \in [0,1]$ and so from (3.135) we have that $a = 0$. Finally from (3.134), $b = 1$ and so

$$\zeta_k = \left(\frac{2 - \theta - \sqrt{\theta(4 - 3\theta)}}{2 - 2\theta} \right)^k. \tag{3.140}$$

Overall,

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_j = 1] \leq 2 \left(\frac{2 - \theta - \sqrt{\theta(4 - 3\theta)}}{2 - 2\theta} \right)^k. \tag{3.141}$$

$\square$

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 [10] 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 [14] 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 4.14. In particular, this Lemma concerns a quantity (see the comments immediately preceding Lemma 4.14) that is very similar to a quantity used in [15] 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

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

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

$$\sum_{i=1}^{d} x_i \leq k \tag{4.2}$$

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

$$\star \mid \star\star \mid\mid \star\star \tag{4.3}$$

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,

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

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 *expected 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 expected magnetization at any vertex at time t.

This quantity is not to be confused with the total magnetization of a stationary Ising configuration, as defined in Section 2.1.1. Recall from (2.5) that given a fixed configuration, the magnetization on a volume Λ is the sum of the spins on Λ. We could consider an analogous random variable, M_t, that is the sum of the spins of the top process, $\mathcal{T}_t$, at any fixed time t. Then we have

$$\frac{\mathbb{E}[M_t]}{n} = m_t. \tag{4.7}$$

As we do not need to refer to the total magnetization of a stationary Ising configuration in this chapter, we will refer to the expected magnetization as simply the magnetization.

We can now define the time

$$t_c(n) = \inf\left\{t > 0 : m_t \leq \frac{1}{n}\right\}. \tag{4.8}$$

which is around the time it takes for the top and bottom chains to couple. One way of interpreting this is that at time $t_c(n)$, $\mathbb{E}[M_{t_c(n)}] = \mathcal{O}(1)$. By way of comparison, at the inital state, $\mathbb{E}[M_0] = n$, and at stationarity, $\mathbb{E}[M] = 0$.

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*

$$\lim_{m\to\infty} \mathbb{P}[T_m < z + t_c(m)] = e^{-C_1 e^{-C_2 z}} \tag{4.9}$$

for some

$$C_1 = C_1(\beta, \mathscr{G}) \in (0, 1] \tag{4.10}$$

and

$$C_2 = C_2(\beta, \mathscr{G}) \in [1 - \beta\Delta, 1]\,. \tag{4.11}$$

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 $V \times [0, t_*]$ 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 $(t, t_*]$.

It is not immediately obvious what these histories look like. To gain some intuition, and to simplify some upcoming proofs, we will construct a different subgraph, $\hat{\mathcal{H}}_A$ which contains $\mathcal{H}_A$. In doing so we also define an analogous update support $\hat{\mathcal{H}}_A(t)$ for every $t \in [0, t_*]$.

Construct $\hat{\mathcal{H}}_A$ 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. Since oblivious updates do not depend on any other vertices, and since a non-oblivious update to i depends on at most the neighbours of i, we have that $\mathcal{H}_A(t) \subseteq \hat{\mathcal{H}}_A(t)$, and that $\mathcal{H}_A$ is a subgraph of $\hat{\mathcal{H}}_A$.

Note that it is possible for vertices to be removed from $\mathcal{H}_A(t)$ by updates that are not oblivious (see Figure 2.4 for an example on the cycle). Since our method for constructing $\hat{\mathcal{H}}_A$ does not take this into account, in general we expect that $\mathcal{H}_A$ and $\hat{\mathcal{H}}_A$ are not equal.

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

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

Now if event A_t^{c} 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 [16] and is valid on any graph, not just transitive ones.

Lemma 4.3 ([16], 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*

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

[FOLLOWING NEEDS LOWER BOUNDS FOR LAMBDA!!!!!]

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

For $z \geq 0$*,*

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

For $z \leq 0$*,*

$$m_{t_*} \leq \frac{e^{-z}}{n}. \tag{4.21}$$

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, to prove Theorem 4.2 we will use the method sketched out in Section 2.4. The various quantities left unspecified there are defined almost identically to Chapter 3;

the main difference in construction being that here we are interested in graph sequences $(G_n) \in \mathscr{G}$ and we define t_* as in (4.17).

Recalling the definition of $t_c(n)$ in (4.8), fix $z \in \mathbb{R}$ and set

$$t_* = t_c(n) + z. \tag{4.23}$$

By Corollary 4.6, for any fixed $z \in \mathbb{R}$, $t_* > 0$ for all sufficiently large n and we only consider such n in what follows. As in Chapter 3, we define the vertex sets

$$B_i = \{j \neq i : d(i,j) \leq b_n\}, \tag{4.24}$$

$$C_i = \{j \notin B_i \cup \{i\} : d(i,j) \leq c_n\}, \tag{4.25}$$

$$D_i = V \setminus (B_i \cup C_i \cup \{i\}), \tag{4.26}$$

where $b_n = \ln(n)$ and $c_n = \ln^2(n)$. From here we define X_i, U_i, W_i, δ_1, δ_4, λ, and μ exactly as in Section 2.4 using our new definition for t_*. From (2.91), we get the following Corollary of Theorem 2.4.

Corollary 4.7. *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*

$$\left|\mathbb{P}\left[T_n \leq z + t_c(n)\right] - e^{-\lambda}\right| \leq (\delta_1 + \delta_4)e^{\lambda}. \tag{4.27}$$

4.3 Proof of Theorem 4.2

In this section we use Corollary 4.7 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 4.8, 4.9, and 4.10. 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 4.8. *Using the above set-up,*

$$\limsup_{n\to\infty} \lambda \leq \max(e^{-z}, e^{-(1-\beta\Delta)z}) \tag{4.28}$$

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

$$\liminf_{n\to\infty} \lambda \geq C \min(e^{-z}, e^{-(1-\beta\Delta)z}). \tag{4.29}$$

Proof. Starting with the definition of λ, we have

$$\lambda = \sum_{i \in V} \mathbb{E}\left[\frac{X_i}{X_i + U_i} I[X_i + U_i \geq 1]\right] \tag{4.30}$$

$$= \sum_{i=1}^{n} \mathbb{P}(X_i = 1)\mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \tag{4.31}$$

$$= n\, m_{t_*}\, \mathbb{E}\left[\frac{1}{1+U_i}\middle| X_i = 1\right] \tag{4.32}$$

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

$$\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \leq 1 \tag{4.33}$$

and so by Corollary 4.5, $\lambda \leq nm_{t_*} \leq \max(e^{-z}, e^{-(1-\beta\Delta)z})$.

By Jensen's inequality

$$\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \geq \frac{1}{\mathbb{E}[1+U_i|X_i = 1]} \tag{4.34}$$

$$= \frac{1}{1+\mathbb{E}[U_i|X_i = 1]}. \tag{4.35}$$

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 4.12, there exists a $C_1 > 0$ such that for small enough β,

$$\mathbb{E}[U_i|X_i = 1] = \sum_{j \in B_i} \mathbb{P}[X_j = 1|X_i = 1] \tag{4.36}$$

$$\leq |B_i|m_{t_*} + C_1 \sum_{k=1}^{\lfloor b_n \rfloor} \sum_{j:d(i,j)=k} e^{-k}. \tag{4.37}$$

From (4.1),

$$\mathbb{E}[U_i|X_i = 1] \leq |B_i|m_{t_*} + C_1 \sum_{k=1}^{\lfloor b_n \rfloor} P_n(k)e^{-k} \tag{4.38}$$

$$\leq |B_i|m_{t_*} + C_1 \sum_{k=1}^{\infty} P_n(k)e^{-k} \tag{4.39}$$

$$\leq C_z|B_i|/n + C_2 \tag{4.40}$$

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

$$\liminf_{n\to\infty} \lambda \geq \frac{1}{1+C_2} n m_{t_*} \tag{4.41}$$

$$\geq C \min(e^{-z}, e^{-(1-\beta\Delta)z}) \tag{4.42}$$

for some $C \in (0,1)$. $\square$

Lemma 4.9. *Let δ_1 be as defined in (2.78). Then at any inverserse temperature $\beta \geq 0$,*

$$\lim_{n\to\infty} \delta_1 = 0. \tag{4.43}$$

Proof. Starting with the definition of δ_1, we have

$$\delta_1 = \sum_{i=1}^{n} \sum_{k=0}^{|B_i|} \mathbb{P}[X_i = 1, U_i = k] \mathbb{E} \left| \frac{\mathbb{P}[X_i = 1, U_i = k | W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1 \right| \tag{4.44}$$

$$= n \sum_{k=0}^{|B_i|} \mathbb{E} \left| \mathbb{P}[X_i = 1, U_i = k | W_i] - \mathbb{P}[X_i = 1, U_i = k] \right| \tag{4.45}$$

by the transitivity of the graph. Denote using

$$B(i,l) = \{j \in V : d(i,j) \leq l\} \tag{4.46}$$

the set of points within distance l of i and define the events

$$A_1 = \{\exists j \in B_i \cup \{i\}, \exists t \in [0, t_*] : \mathcal{H}_j(t) \nsubseteq B(i, (c_n + b_n)/2)\} \tag{4.47}$$

and

$$A_2 = \{\exists j \in D_i, \exists t \in [0, t_*] : \mathcal{H}_j(t) \cap B(i, (c_n + b_n)/2) \neq \emptyset\} \tag{4.48}$$

as well as their intersection

$$A = A_1 \cap A_2. \tag{4.49}$$

From Lemma 3.11,

$$\mathbb{P}[X_i = 1, U_i = j | A^c, W_i] = \mathbb{P}[X_i = 1, U_i = j | A^c]. \tag{4.50}$$

Continuing on from (4.45), we split the probabilities into

$$\begin{aligned}
\delta_1 &= n\sum_{k=0}^{|B_i|}\mathbb{E}\,|\mathbb{P}[X_i=1,U_i=k|W_i,A]\mathbb{P}[A|W_i]-\mathbb{P}[X_i=1,U_i=k|A]\mathbb{P}[A]+ && (4.51)\\
&\quad \mathbb{P}(X_i=1,U_i=k|A^c)(\mathbb{P}[A^c|W_i]-\mathbb{P}[A^c])| \\
&\le n(|B_i|+1)\mathbb{E}\left[\mathbb{P}[A|W_i]+\mathbb{P}[A]+|\mathbb{P}[A^c|W_i]-\mathbb{P}[A^c]|\right] && (4.52)\\
&= n(|B_i|+1)\mathbb{E}\left[\mathbb{P}[A|W_i]+\mathbb{P}[A]+|1-\mathbb{P}[A|W_i]-(1-\mathbb{P}[A])|\right] && (4.53)\\
&\le n(|B_i|+1)\mathbb{E}\left[\mathbb{P}[A|W_i]+\mathbb{P}[A]+\mathbb{P}[A|W_i]+\mathbb{P}[A])\right] && (4.54)\\
&= 2n(|B_i|+1)\left(\mathbb{E}[\mathbb{P}[A|W_i]]+\mathbb{P}[A]\right) && (4.55)\\
&= 4n(|B_i|+1)\mathbb{P}[A] && (4.56)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{P}[A] &\le \mathbb{P}[A_1\cup A_2] && (4.57)\\
&\le \mathbb{P}\left[\bigcup_{j\in V}\{\mathcal{H}_j\not\subseteq B(j,(c_n-b_n)/2)\times[0,t_*]\}\right] && (4.58)\\
&\le \sum_{j\in V}\mathbb{P}[\mathcal{H}_j\not\subseteq B(j,(c_n-b_n)/2)\times[0,t_*]] && (4.59)\\
&= n\mathbb{P}\left[\bigcup_{u\in[0,t_*]}\mathcal{H}_j(t_*-u)\not\subseteq B(j,(c_n-b_n)/2)\right] && (4.60)
\end{aligned}$$

Combining this with Lemma 4.11 we get that

$$\begin{aligned}
\delta_1 &\le 4n^2(|B_i|+1)\exp(t_*\Delta^2-\ln(\Delta)(c_n-b_n)/2) && (4.61)\\
&\le 4\exp(\Delta^2 z)n^{2+\Delta^2/(1-\beta\Delta)}(|B_i|+1)\exp(-\ln(\Delta)(c_n-b_n)/2) && (4.62)
\end{aligned}$$

using Corollary 4.6. Recalling our choices of $b_n=\ln(n)$ and $c_n=\ln(n)^2$, we have

$$\begin{aligned}
\delta_1 &\le 4e^{\Delta^2 z}n^{2+\Delta^2/(1-\beta\Delta)}Q_n(b_n)e^{-\frac{\ln(\Delta)}{2}c_n}e^{\frac{\ln(\Delta)}{2}b_n} && (4.63)\\
&= 4e^{\Delta^2 z}n^{2+\Delta^2/(1-\beta\Delta)+\log(\Delta)/2}Q_n(b_n)e^{-\frac{\ln(\Delta)}{2}\ln^2(n)} && (4.64)
\end{aligned}$$

which by (4.1) goes to 0 as $n\to\infty$.

□

Lemma 4.10. *Let δ_4 be as defined in (2.79). Then for small enough β,*

$$\lim_{n\to\infty}\delta_4=0. \qquad (4.65)$$

Proof. Starting with the definition of δ_4, we have

$$\delta_4 = \sum_{i=1}^{n} (\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]) \tag{4.66}$$

$$= n\mathbb{E}[X_i Z_i] + nm_{t_*}^2 (1 + |B_i| + |C_i|) \tag{4.67}$$

$$= nm_{t_*}\mathbb{E}[Z_i|X_i = 1] + nm_{t_*}^2 (1 + |B_i| + |C_i|) \tag{4.68}$$

$$\leq C_z\mathbb{E}[Z_i|X_i = 1] + \frac{C_z^2}{n} (1 + |B_i| + |C_i|) \tag{4.69}$$

where using Corollary 4.5

$$C_z = \max(e^{-z}, e^{-(1-\beta\Delta)z}). \tag{4.70}$$

Now from (4.1), the second term above vanishes as $n \to \infty$. So we turn our attention to the first term. From Lemma 4.12, there exists a $C > 0$ such that for small enough β,

$$\mathbb{E}[Z_i|X_i = 1] = \sum_{j\in C_i} \mathbb{P}[X_j = 1|X_i = 1] \tag{4.71}$$

$$\leq \sum_{j\in C_i} \left[m_{t_*} + Ce^{-d(i,j)}\right] \tag{4.72}$$

$$\leq \sum_{j\in C_i} \left[\frac{C_z}{n} + Ce^{-b_n}\right] \tag{4.73}$$

$$= |C_i| \left(\frac{C_z}{n} + \frac{C}{n}\right) \tag{4.74}$$

where we have again used Corollary 4.5. Altogether, we see that δ_4 goes to zero as $n \to \infty$. □

4.4 Additional lemmas

Our first Lemma is based on [1, Lemma 2.1] and bounds the speed at which the histories can spread through the graph.

Lemma 4.11. *The probability that the history of vertex i escapes a ball of radius l in time s is bounded by*

$$\mathbb{P}\left[\bigcup_{u\in[0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i,l)\right] \leq \exp\left(s\Delta^2 - l\ln\Delta\right). \tag{4.75}$$

Proof. Let $\mathcal{W} = \{\boldsymbol{w} = (w_1, w_2, \ldots, w_l) : w_1 = i, w_{k-1} \sim w_k\}$ be the set of length l sequences of adjacent vertices starting at vertex i. That is, $\mathcal{W}$ is the set of all walks of length $l - 1$ on G_n that start at 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_k was updated in order along the interval $[t_* - s, t_*]$. For any particular sequence w, let M_w be the event that each w_k was updated at some time t_k such that $t_* > t_k > t_* - s$ and $t_{k-1} > t_k$.

For M_w to occur, we require that the l independent rate 1 Poisson clocks associated with each w_k rings in a particular order along $[t_* - s, t_*]$. The probability of this is equal to the probability that a single rate 1 Poisson clock rings at least l times in $[0, s]$, or equivalently, a rate s Poisson clock rings at least l times in $[0, 1]$. So we have

$$\mathbb{P}[M_w] = \mathbb{P}[\mathrm{Po}(s) \geq l] \tag{4.76}$$

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

$$\mathbb{P}\left[\bigcup_{u \in [0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i, l)\right] \leq \Delta^{l-1} \mathbb{P}[\mathrm{Po}(s) \geq l]. \tag{4.77}$$

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

$$M(t) = \exp\left(s\left(e^t - 1\right)\right). \tag{4.78}$$

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

$$\mathbb{P}[\mathrm{Po}(s) \geq l] \leq \exp\left(s\left(e^t - 1\right) - tl\right). \tag{4.79}$$

Overall we have

$$\begin{aligned}\mathbb{P}\left[\bigcup_{u \in [0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i, l)\right] &\leq \Delta^{l-1} \exp\left(s\left(e^t - 1\right) - tl\right) &(4.80)\\ &\leq \exp\left(s\left(e^t - 1\right) + l(\ln \Delta - t)\right). &(4.81)\end{aligned}$$

Choosing $t = 2 \ln \Delta$,

$$\begin{aligned}\mathbb{P}\left[\bigcup_{u \in [0,s]} \mathcal{H}_i(t_* - u) \not\subseteq B(i, l)\right] &\leq \exp\left(s\left(\Delta^2 - 1\right) - l \ln \Delta\right) &(4.82)\\ &\leq \exp\left(s\Delta^2 - l \ln \Delta\right). &(4.83)\end{aligned}$$

□

Lemma 4.12. *There exists a constant $C > 0$ such that for sufficiently small β,*

$$\mathbb{P}[X_j = 1 | X_i = 1] \leq m_{t_*} + Ce^{-k}. \tag{4.84}$$

where $k = d(i,j)$ 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

$$\mathbb{P}[X_j = 1 | X_i = 1] = \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] + \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] \tag{4.85}$$

$$\leq \mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] + \mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1]. \tag{4.86}$$

The result follows from Lemmas 4.13 and 4.14. □

Lemma 4.13. *Let i and j be two vertices on a vertex transitive graph. Then*

$$\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] \leq m_{t_*} \tag{4.87}$$

Proof. To begin

$$\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] = \mathbb{P}[X_i = 1]^{-1} \mathbb{P}[X_i = 1, X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] \tag{4.88}$$

$$= \frac{\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset, \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset]}{\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]}. \tag{4.89}$$

By Lemma 3.14,

$$\mathbb{P}[\mathcal{H}_i(0) \neq \emptyset, \mathcal{H}_j(0) \neq \emptyset, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset] \leq \mathbb{P}[\mathcal{H}_i(0) \neq \emptyset]^2. \tag{4.90}$$

and so

$$\mathbb{P}[X_j = 1, \mathcal{H}_i \cap \mathcal{H}_j = \emptyset | X_i = 1] \leq \mathbb{P}[\mathcal{H}_i(0) \neq \emptyset] \tag{4.91}$$

$$= m_{t_*}. \tag{4.92}$$

□

The final two Lemmas use 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

$$\chi(\mathcal{H}_i) = \# \{((u,t),(v,t)) \in \mathcal{H}_i\} \tag{4.93}$$

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

$$\mathcal{L}(\mathcal{H}_i) = \sum_{i \in V} \int_0^{t_*} I_{(i,t)\in\mathcal{H}_i} \,\mathrm{d}t \tag{4.94}$$

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

The following Lemma contains some similarities to the proof of [15, Lemma 2.1]. Indeed, the quantity

$$\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\}] \tag{4.95}$$

which appears in (4.101) below, is equivalent to the expression

$$\mathbb{P}[A \in \mathrm{RED}_A^*] \tag{4.96}$$

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

$$\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] \leq Ce^{-k}. \tag{4.97}$$

Proof. Firstly,

$$\begin{aligned}
\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] &= \frac{\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{X_i = 1\}]}{\mathbb{P}[X_i = 1]} && (4.98)\\
&\leq m_{t_*}^{-1} \mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{X_i = 1\}] && (4.99)\\
&\leq m_{t_*}^{-1} \mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap (\{X_i = 1\} \cup \{X_j = 1\})] && (4.100)\\
&= m_{t_*}^{-1} \mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\}] && (4.101)
\end{aligned}$$

since

$$\{X_i = 1\} = \{\mathcal{H}_i(0) \neq \emptyset\}\,. \tag{4.102}$$

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. (We cannot say that S is the random time at which $\mathcal{H}_i(t) \cup \mathcal{H}_j(t)$ first reduces to a single vertex since a single update may remove more than one vertex. See Figure 2.4 for an example.) Note that

$$\{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\} \subseteq \{\mathscr{F}(v, 0, S) \neq \emptyset\} \tag{4.103}$$

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

$$\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \subseteq \{\chi\left((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S, t_*]\right) \geq k\} \tag{4.104}$$

since there must be at least k spatial edges for the histories to meet. The event on the right hand side of (4.104) depends only on the update sequence in $[S, t_*]$. The event on the right hand side of (4.103) depends only on the update sequence in $[0, S)$. Therefore, given S, these events are independent and

$$\begin{aligned}
&\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\} \, | S = t_s] \\
&\quad \leq \mathbb{P}[\mathscr{F}(v, 0, S) \neq \emptyset | S = t_s] \mathbb{P}[\chi\left((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S, t_*]\right) \geq k | S = t_s] &(4.105)\\
&\quad = m_{t_s} \mathbb{P}[\chi\left((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S, t_*]\right) \geq k | S = t_s] &(4.106)
\end{aligned}$$

and so

$$\begin{aligned}
\mathbb{P}[\{\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset\} \cap \{\mathcal{H}_i(0) \cup \mathcal{H}_j(0) \neq \emptyset\}] &\leq \mathbb{E}[I_{\chi((\mathcal{H}_i \cup \mathcal{H}_j) \cap V \times [S, t_*]) \geq k} m_S] &(4.107)\\
&\leq \mathbb{E}[I_{\chi(\mathcal{H}_i \cup \mathcal{H}_j) \geq k} m_S]. &(4.108)
\end{aligned}$$

From Corollary 4.4,

$$m_S \leq e^{t_* - S} m_{t_*} \tag{4.109}$$

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

$$t_* - S \leq \mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2. \tag{4.110}$$

So

$$\begin{aligned}
\mathbb{P}[\mathcal{H}_i \cap \mathcal{H}_j \neq \emptyset | X_i = 1] &\leq m_{t_*}^{-1} m_{t_*} \mathbb{E}[I_{\chi(\mathcal{H}_i \cup \mathcal{H}_j) \geq k} e^{\mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2}] &(4.111)\\
&\leq \mathbb{E}[e^{\chi(\mathcal{H}_i \cup \mathcal{H}_j) - k} e^{\mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2}] &(4.112)\\
&= e^{-k} \mathbb{E}[e^{\chi(\mathcal{H}_i \cup \mathcal{H}_j) + \mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2}] &(4.113)
\end{aligned}$$

[OPTIMIZE ALPHA?]

From Lemma 4.15, for any $\alpha > \ln(2)$ if

$$\tanh(\beta\Delta) \leq \frac{1 - 2e^{-\alpha}}{2(e^{\Delta(\alpha+1)} - e^{-\alpha})} \tag{4.114}$$

then

$$\mathbb{E}[e^{\chi(\mathcal{H}_i \cup \mathcal{H}_j)+\mathcal{L}(\mathcal{H}_i \cup \mathcal{H}_j)/2}] \leq e^{2\alpha} \tag{4.115}$$

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

The last of our additional lemmas comes from [15, 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.

Lemma 4.15 ([15]). *For any $0 \leq \eta < 1$, $\lambda \in \mathbb{R}$, $\alpha > -\ln(1-\eta)$, if*

$$\tanh(\beta\Delta) \leq \frac{1-\eta-e^{-\alpha}}{e^{(\alpha+\lambda)\Delta}-e^{-\alpha}}, \tag{4.116}$$

then for any $A \subseteq V$,

$$\mathbb{E}[\exp(\lambda\chi(\mathcal{H}_A)+\eta\mathcal{L}(\mathcal{H}_A))] \leq \exp(\alpha|A|). \tag{4.117}$$

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

$$\chi(\mathcal{H}_A) \leq \chi(\hat{\mathcal{H}}_A), \qquad \mathcal{L}(\mathcal{H}_A) \leq \mathcal{L}(\hat{\mathcal{H}}_A). \tag{4.118}$$

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

Initially, $W_0 = |A|$, $Y_0 = 0$, 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

$$\frac{d}{ds}\mathbb{E}\left[Q_s|Q_{s_0}\right]\Big|_{s=s_0} = \Big(\eta + \theta(e^{-\alpha}-1) + (1-\theta)(e^{(\alpha+\lambda)\Delta}-1)\Big)\,\bar{W}_{s_0}Q_{s_0} \tag{4.119}$$

which is non-positive when

$$\theta \geq \frac{\eta + e^{(\alpha+\lambda)\Delta} - 1}{e^{(\alpha+\lambda)\Delta} - e^{-\alpha}} \tag{4.120}$$

or in terms of the inverse temperature, β,

$$\tanh(\beta\Delta) \leq \frac{1 - \eta - e^{-\alpha}}{e^{(\alpha+\lambda)\Delta} - e^{-\alpha}}. \tag{4.121}$$

Hence Q_s is a supermartingale when (4.121) holds. Define the stopping time

$$\tau = \inf\{s : \bar{W}_s = 0\}. \tag{4.122}$$

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 $\bar{Z}_\tau \geq \mathcal{L}(\mathcal{H}_A)$, and so $\mathbb{E}\exp(\lambda\chi(\mathcal{H}_A) + \eta\mathcal{L}(\mathcal{H}_A)) \leq \mathbb{E}Q_\tau$. From optional stopping,

$$\mathbb{E}[Q_\tau] \leq \mathbb{E}[Q_0] \tag{4.123}$$

$$= \exp(\alpha|A|). \tag{4.124}$$

□

Chapter 5

Conclusion to Part I

In the first part of this thesis, we have examined the dynamics of the continuous-time Ising heat-bath Glauber process. We started by describing the discrete time dynamics, as well as the two coupled chains from which we define the coupling time. We then described how to 'continuize' the discrete time dynamics. It is the distribution of the coupling time of these continuous time dynamics that was the central object of our analysis. We proved that the asymptotic distributions of the discrete coupling time and the continuous coupling time are equivalent via Proposition 2.1 which is applicable in much more generality. This was important to show since our analysis focuses on the continuous-time dynamics, whereas our motivation for the study of the coupling time was that the discrete coupling time has the same distribution as the 'Coupling From the Past' algorithm. We ended the introduction by giving a description of information percolation, as used by Lubetzky and Sly in [1], and of compound Poisson approximation, as described in [23].

In Chapter 3, we looked at the coupling-time on the cycle. Our main result was that at any inverse-temperature β, the coupling time converges to a Gumbel distribution. This confirmed, for $d = 1$, a conjecture in [10, Conjecture 7.1] 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 for all $\beta < \beta_c$. We used the two techniques of information percolation, and compound Poisson approximation to create the proof.

In Chapter 4, we extended the class of graphs we considered from the cycle, to a family of vertex-transitive graphs with fixed degree. We showed that as well as the d-dimensional lattices, this family also included other classes of graphs which are of practical interest with respect to the Ising model. Our main result was that for sufficiently small β, the coupling time converges to a Gumbel distribution. This partially confirms, for $d > 1$, the afore mentioned conjecture in [10]. While our result does not hold up to the critical temperature as conjectured, it holds for a larger family of graphs and we commented that not holding up to the critical temperature on lattices is an expected consequence of the increased generality of graphs to which our result is applicable.

Of particular note is that the main part of the proof in which we require β to be small concerns an object which is almost identical to a quantity used by Lubetzky and Sly in [15]. In this paper, Lubetzky and Sly proved the existence of cutoff on a wide class of transitive graphs for β small enough. They later refined this proof in [1] to apply up to the critical temperature on the d-dimensional lattices. In [16] they were able to specify how small β must be for cutoff to occur on any sequence of graphs with maximum degree Δ.

This encourages future efforts to both refine our proof on the d-dimensional lattices so that it holds up all the way up to the critical β, and to specify a sufficiently small value of β for our result to hold in general. We would also like to investigate other problems for which information percolation can provide solutions. The success of information percolation in establishing cutoff (a previously notoriously difficult problem), and our success in applying it to our problem, suggests that it is a powerful and flexible tool for analysing spin dynamics. Given that it is a relatively recent technique, we think that there exists a variety of new results which can be obtained using information percolation.

[ANYTHING MORE SPECIFIC?]

Part II

Efficient Optimization for Statistical Inference

Chapter 6

Maximum Likelihood Location Mixtures

6.1 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" [28, Section 1.1.1]. Mixture models have been in use for over 100 years. In 1894, 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 [29]. 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 [30] 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 would contain 10000 terms.

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

$$f_Q(\boldsymbol{x}) = \int_\Omega f(\boldsymbol{x}; \boldsymbol{\theta}) \, \mathrm{d}Q(\boldsymbol{\theta}) \tag{6.1}$$

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 (6.1) is a *finite mixture* and can be written

$$f_Q(\boldsymbol{x}) = f_{\boldsymbol{\theta},\boldsymbol{p}}(\boldsymbol{x}) = \sum_{j=1}^{m} p_j f(\boldsymbol{x}; \boldsymbol{\theta}_j). \tag{6.2}$$

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

$$f_{\boldsymbol{p},\boldsymbol{\theta}} = \sum_{j=1}^{m} p_j f(x - \theta_j) \tag{6.3}$$

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] [28].

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

$$L(Q; \boldsymbol{x}, f) = \prod_{i=1}^{n} f_Q(x_i). \tag{6.4}$$

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

$$l(Q; \boldsymbol{x}, f) = \log L(Q; \boldsymbol{x}, f) = \sum_{i=1}^{n} \log f_Q(x_i), \tag{6.5}$$

since $\log(x)$ is a strictly increasing function. In [31], 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 (6.5) 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

$$\hat{Q}_{\boldsymbol{x},f} = \{Q \in \mathcal{Q}_n : \forall Q' \in \mathcal{Q}_n, l(Q; \boldsymbol{x}) \geq l(Q'; \boldsymbol{x})\} \tag{6.6}$$

be the set of all global maximizers of (6.5). Then define

$$K_{\boldsymbol{x}} = K(\boldsymbol{x}; f) = \min\{m \leq n : \mathcal{Q}_m \cap \hat{Q}_{\boldsymbol{x},f} \neq \emptyset\} \tag{6.7}$$

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 (6.5) 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 when 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 [31] and in [32] for continuous univariate component densities in the exponential family. A generalization of these results to discrete exponential families be found in [33].

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 [34]).

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 [28, 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) [35] or the Bayesian information criterion (BIC) [36], does not underestimate this quantity asymptotically [37]. 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 $g(x)$.

6.2 Summary of Lindsay

To obtain results concerning $K_{\boldsymbol{x}}$ we will make use of the geometrical approach employed by Lindsay in [31] and [32] and summarized later in [38, 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.

6.2.1 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*

$$\boldsymbol{\gamma}(Q;\boldsymbol{x},f) = (f_Q(x_1),\ldots,f_Q(x_n)) \tag{6.8}$$

and the objective function

$$\mathcal{L}(\boldsymbol{\gamma}) = \sum_{i=1}^{n} \ln(\gamma_i). \tag{6.9}$$

Let

$$\mathcal{M} = \{\boldsymbol{\gamma}(Q;\boldsymbol{x},f) : Q \text{ is a probability distribution}\} \tag{6.10}$$

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

$$\boldsymbol{\gamma}(Q;\boldsymbol{x},f) = \hat{\boldsymbol{\gamma}} \tag{6.11}$$

for Q.

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

$$\boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f) = (f_{\boldsymbol{\theta}}(x_1), \ldots, f_{\boldsymbol{\theta}}(x_n)) \tag{6.12}$$

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

$$f_{\boldsymbol{\theta}}(x) = f(x; \boldsymbol{\theta}). \tag{6.13}$$

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

$$\boldsymbol{\gamma}(Q; \boldsymbol{x}, f) = \int \boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f)\, \mathrm{d}Q(\boldsymbol{\theta}) \tag{6.14}$$

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

$$\boldsymbol{\gamma}(Q; \boldsymbol{x}, f) = \sum_{j=1}^{m} p_j \boldsymbol{\gamma}(\boldsymbol{\theta}_j; \boldsymbol{x}, f). \tag{6.15}$$

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

$$\Gamma_{\boldsymbol{x},f} = \{\boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f) : \boldsymbol{\theta} \in \Omega\}. \tag{6.16}$$

Then

$$\mathcal{M} = \mathrm{conv}(\Gamma_{\boldsymbol{x},f}), \tag{6.17}$$

where we use $\mathrm{conv}(A)$ to denote the convex hull of A. We now state the following Theorem taken directly from [38, Theorem 18] which is a consequence of the convexity of $\mathcal{M}$, and the concavity of $\mathcal{L}(\boldsymbol{\gamma})$.

Theorem 6.1 ([38]). *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 $\hat{\gamma} \in \partial\mathcal{M}$, the boundary of $\mathcal{M}$, such that $\hat{\gamma}$ maximizes $\mathcal{L}(\gamma)$ over $\mathcal{M}$.*

In addition to this, in [38, Theorem 21] is stated the following.

Theorem 6.2 ([38]). *The solution $\hat{\gamma}$ 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,

$$K_{\boldsymbol{x}} \leq n. \tag{6.18}$$

6.2.2 All points separated by α

Lindsay states that for normal mixtures with fixed variance, σ^2, "one can construct sets of data for which the bound $[K_{\boldsymbol{x}} = n]$ is attained simply by spreading the observations widely apart," [38, Section 5.2]. Here we make this intuition concrete by saying how far apart we should spread our observations to ensure that we use all n components.

Theorem 6.3. *Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ be the sample for which we are finding a maximum likelihood location mixture using $f(x)$ as our component density. Let $f(x)$ be a unimodal density on $\mathbb{R}$, symmetric about $x = 0$, and such that the conditions of Theorem 6.2 hold. Let $\alpha > 0$ be such that*

$$\frac{f(\alpha/2)}{f(0)} < \frac{1}{n}\left(\frac{n-1}{n}\right)^{n-1}. \tag{6.19}$$

Then if for all $i \neq j$, $|x_i - x_j| > \alpha$ we must have $K_{\boldsymbol{x}} = n$.

Proof. First consider constructing the maximum likelihood mixture density using no more than $n-1$ components. Let $\hat{f}_{n-1}$ be this 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

$$\hat{f}_{n-1}(x_{i^*}) < f(\alpha/2), \tag{6.20}$$

and so

$$L_{n-1} < f(\alpha/2) \prod_{i \neq i^*} \hat{f}_{n-1}(x_i). \tag{6.21}$$

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 $(\theta, p) = (x_{i^*}, \frac{1}{n})$. Call this function f_n^* and the corresponding likelihood L_n. At each $x_i \neq x_{i^*}$ the likelihood has decreased by $(n-1)/n$ and at x_{i^*} the likelihood is at least $f(0)/n$. So

$$L_n > \frac{f(0)}{n}\left(\frac{n-1}{n}\right)^{n-1} \prod_{i \neq i^*} \hat{f}_{n-1}(x_i). \tag{6.22}$$

By (6.19), we get that $L_n > L_{n-1}$. Since the likelihood strictly increases when we allow an additional component, the maximum likelihood solution must have at least n components. By Theorem 6.2, the maximum likelihood solution cannot have any more than n components and so $K_{\boldsymbol{x}} = n$. □

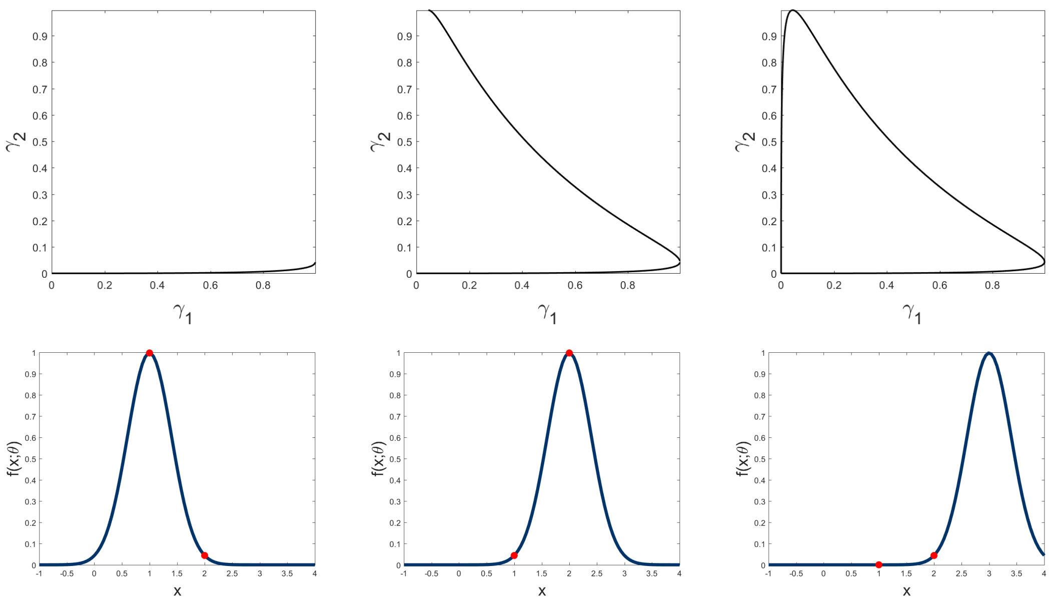
6.2.3 An example likelihood curve

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 $\sigma_2 = 0.4^2$ and parametrized by θ, that is,

$$f(x; \theta) = \frac{1}{0.4\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2 \cdot 0.4^2}\right). \tag{6.23}$$

Then the corresponding likelihood curve can be traced out as we increase θ from $-\infty$ to ∞ as shown in Figure 6.1.



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

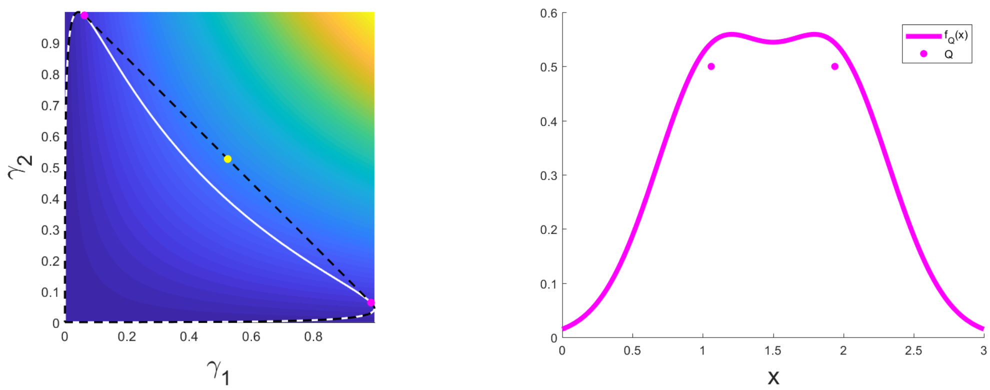
Figure 6.1 – *A simple example of a likelihood curve. The component density, $f(x;\theta)$, as defined in (6.23), is shown in blue for $\theta = 1, 2,$ and 3. Each value of θ contributes a point to $\Gamma_{\boldsymbol{x},f}$ whose coordinates are given by $(f(1;\theta), f(2;\theta))$ (represented by the heights of the red circles). As we increase θ from $-\infty$ to ∞ we trace out more of $\Gamma_{\boldsymbol{x},f}$.*

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 6.2a 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 6.1. This point may be written as the convex combination of two points in $\Gamma_{\boldsymbol{x},f}$,

$$\hat{\gamma} = \sum_{j=1}^{2} p_j \boldsymbol{\gamma}(\theta_j; \boldsymbol{x}, f) \tag{6.24}$$

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 $\hat{\gamma} = \boldsymbol{\gamma}(\hat{Q}; \boldsymbol{x}, f)$ is the one that places masses p_1 and p_2 at locations θ_1 and θ_2 and it is plotted in Figure 6.2b 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 6.2 – *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* $\hat{\gamma}$ *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).*

Remark 6.4. 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 6.1 and 6.2. 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 [38, Section 5.2.2.].

6.2.4 Gradient characterization

Another important construction due to Lindsay is the gradient function

$$D_Q(\boldsymbol{\theta}; \boldsymbol{x}, f) = -n + \sum_{i=1}^{n} \frac{f(x_i; \boldsymbol{\theta})}{f_Q(x_i)}. \tag{6.25}$$

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

$$(1-p)Q + p\Delta_{\boldsymbol{\theta}} \tag{6.26}$$

evaluated at $p = 0$ and where $\Delta_{\boldsymbol{\theta}}$ is a degenerate distribution that places all of its mass at $\boldsymbol{\theta}$. In [31, Theorem 4.1], Lindsay showed you could characterise the maximizing mixture by three equivalent conditions. The statement of the Theorem here is taken from [38, Theorem 19].

Theorem 6.5 ([38]). *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 [31, Theorem 4.1] was the following result concerning the locations of the support points of $\hat{Q}$. The statement here is taken from [38, Theorem 20].

Theorem 6.6 ([38]). *The support of any maximum likelihood estimator $\hat{Q}$ lies in the set*

$$\left\{\boldsymbol{\theta} : D_{\hat{Q}}(\boldsymbol{\theta};\boldsymbol{x},f) = 0\right\}. \tag{6.27}$$

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

$$D_{\hat{Q}}(\theta^*;\boldsymbol{x},f) = 0, \tag{6.28}$$

$$D'_{\hat{Q}}(\theta^*;\boldsymbol{x},f) = 0, \tag{6.29}$$

$$D''_{\hat{Q}}(\theta^*;\boldsymbol{x},f) \leq 0. \tag{6.30}$$

6.2.4.1 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

$$\mathcal{H}_Q = \{\boldsymbol{z} : \langle\boldsymbol{\gamma}^{-1}(Q;\boldsymbol{x},f),\boldsymbol{z}\rangle = n\}, \tag{6.31}$$

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

$$D(Q;\boldsymbol{x},f) = \langle\boldsymbol{\gamma}^{-1}(Q;\boldsymbol{x},f),\boldsymbol{\gamma}(\boldsymbol{\theta};\boldsymbol{x},f)\rangle - n. \tag{6.32}$$

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

$$\langle \boldsymbol{\gamma}^{-1}(\hat{Q}; \boldsymbol{x}, f), \boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f)\rangle \leq n \tag{6.33}$$

for all $\boldsymbol{\theta} \in \Omega$. This means that $\mathcal{M} = \text{conv}(\Gamma_{\boldsymbol{x},f})$ lies entirely on one side of $\mathcal{H}$ and Theorem 6.6 tell us that if $\boldsymbol{\theta}$ is in the support of $\hat{Q}$, then

$$\langle \boldsymbol{\gamma}^{-1}(\hat{Q}; \boldsymbol{x}, f), \boldsymbol{\gamma}(\boldsymbol{\theta}; \boldsymbol{x}, f)\rangle = n \tag{6.34}$$

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?]

6.2.5 KKT Conditions

[SHOW THAT KKT CONDITIONS ARE RELATED TO DIRECTIONAL DERIVATIVE CONSTRAINTS]

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

Theorem 6.2 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 $\boldsymbol{x}$.

In [32], Lindsay looked at the special case of component densities in the exponential family. He related the behaviour of certain polynomials to the location of the support points of the maximizing mixture. A consequence of this was a sufficient condition for the maximizing mixture to have exactly one point of support. We restate part of this theorem here.

Theorem 6.7 (Theorem 4.1, [32]). *Let f_θ belong to the exponential class of densities and be parameterized by its mean value θ. Let $\boldsymbol{x} = (x_1, \ldots, x_n)$, $x_1 \leq \cdots \leq x_n$, be the sample to which we are finding the maximum likelihood mixture of f_θ. Define the function*

$$M(\theta) = (x_1 - \theta)(x_2 - \theta) + \text{Var}_\theta(X). \tag{6.35}$$

If $M(\theta)$ is strictly positive on $[x_1, x_n]$, then the maximimum likelihood mixing distribution, $\hat{Q}$, has exactly one point of support located at $\theta = \bar{x}$.

The above is a sufficient condition for $K_{\boldsymbol{x}} = 1$. However, Linsday also showed that in the case of two observations ($n = 2$), one could easily visualise the likelihood curve for a location mixture of normals and observe that the above sufficient condition also seems to be necessary for a normal density. However, he also stated that "these results were very difficult to obtain in higher dimensions, and hard to generalize outside the exponential family" [38]. Some generalizations to the results in [32] were presented in [33]. This paper gave improved bounds on $K_{\boldsymbol{x}}$ for discrete component densities.

In Section 6.4 we will present a necessary and sufficient condition for $K_{\boldsymbol{x}} = 1$ in the case of two observations for a certain class of densities that include the normal density. We will also present a necessary condition for $K_{\boldsymbol{x}} = m$, $m = 1, \ldots, n$, for normal densities in the case of n observations.

6.3 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 (6.3)). 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	$\frac{1}{\pi\gamma+\pi x^2/\gamma}$

6.3.1 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 6.5. 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 [39] 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

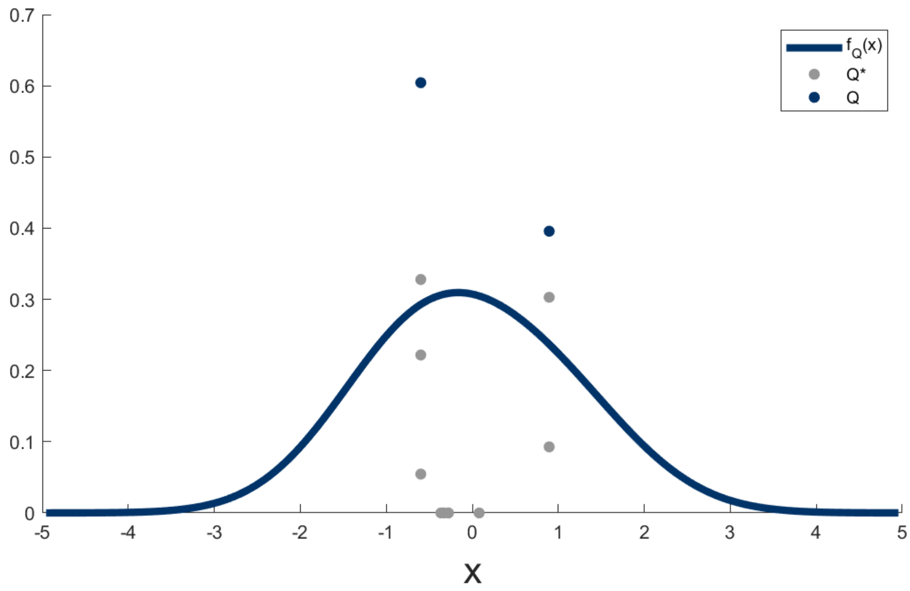


Figure 6.3 – *The unsimplified mixing distribution Q^* that results from using more points of support than required when finding our maximizing mixture, as well as the equivalent simplified distribution, Q, and the resulting mixture density, $f_Q(x)$. Each probability mass with support θ and weight p is represented by a point at (θ, p).*

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 6.3). 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 6.6, 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 $K_{\boldsymbol{x}}$ 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?]

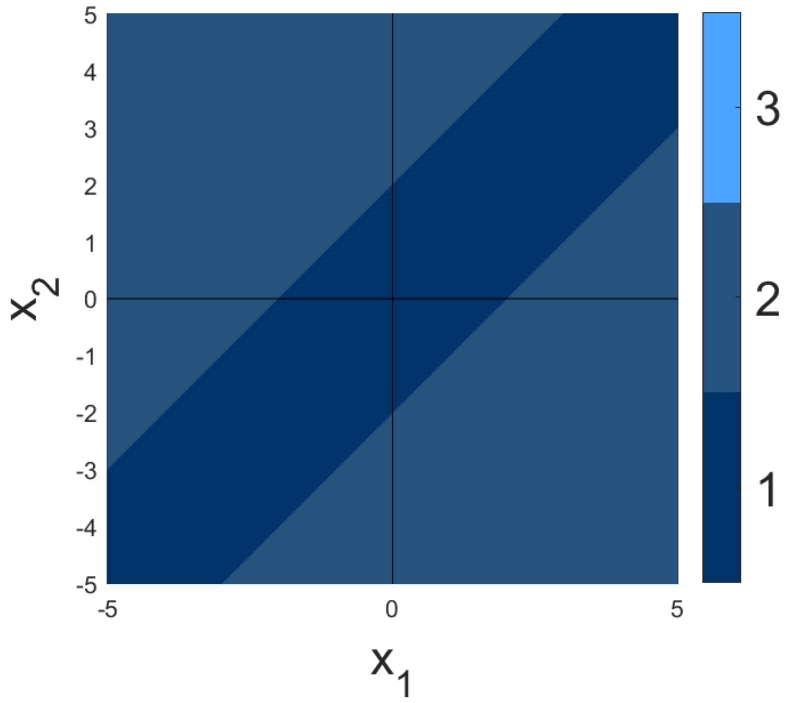


Figure 6.4 – *$K_{\boldsymbol{x}}$ as a function of $\boldsymbol{x} = (x_1, x_2)$, for a normal component density with fixed variance $\sigma^2 = 1$.*

[EXTEND AS REQUIRED]

6.3.2 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 6.4 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 $\boldsymbol{x}+(c, \ldots, c)$ 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 grid. We do this for both a normal component density with variance $\sigma^2 = 1$ (Figure 6.5) and a Cauchy density with scale $\gamma = \sqrt{3}$ (Figure 6.6). We have chosen the scale of the

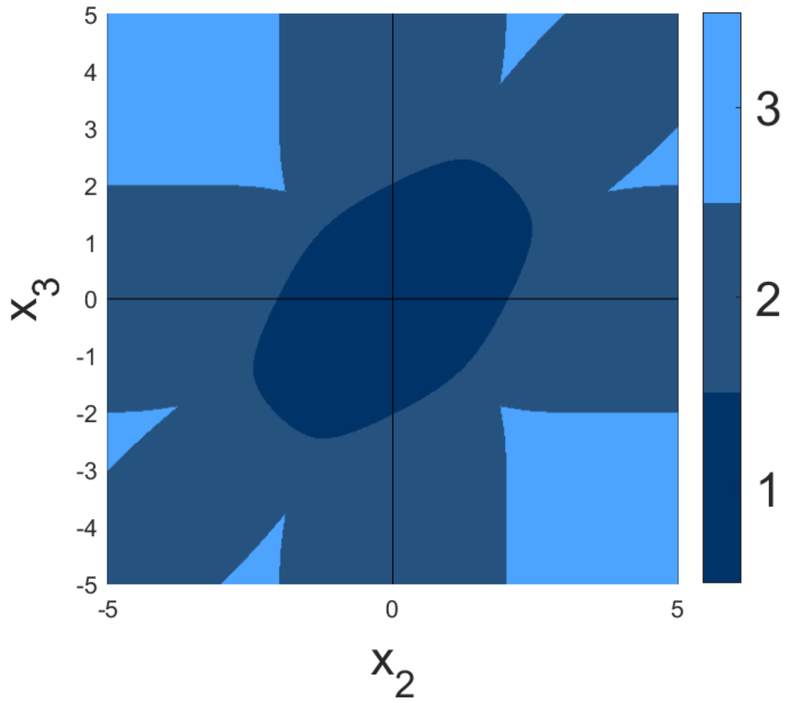


Figure 6.5 – $K_{\boldsymbol{x}}$ *as a function of* $\boldsymbol{x} = (0, x_2, x_3)$ *for a normal component density with fixed variance* $\sigma^2 = 1$.

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 6.8 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

$$C_k = \{\boldsymbol{x} \in \mathbb{R}^n | K_{\boldsymbol{x}} = k\}, \qquad k = 1, \ldots, n. \tag{6.36}$$

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 6.4, we present various results which bound the regions C_k in various settings.

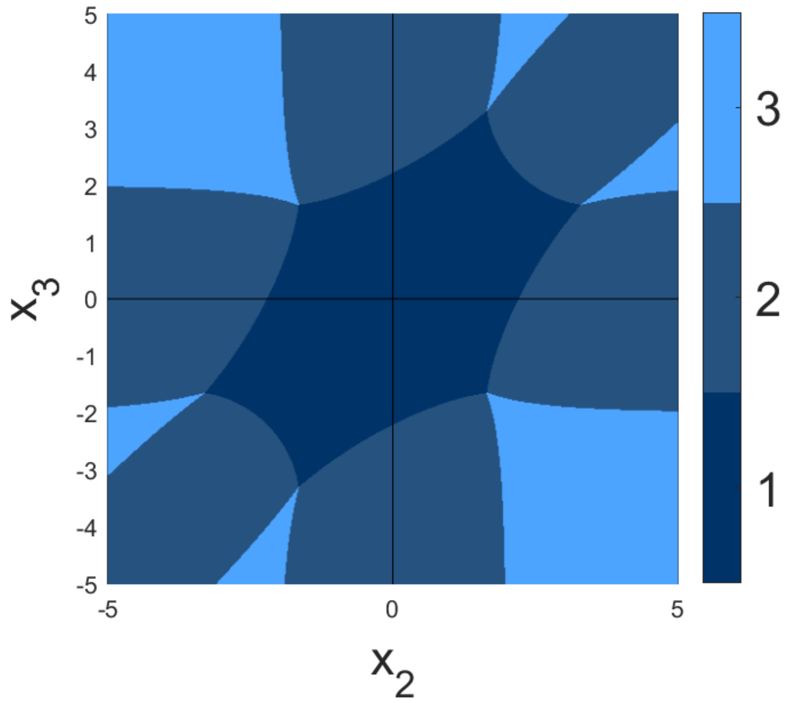


Figure 6.6 – *$K_{\boldsymbol{x}}$ as a function of $\boldsymbol{x} = (0, x_2, x_3)$ for a Cauchy component density with fixed scale $\gamma = \sqrt{3}$.*

6.3.3 Other interesting things

We end our empirical results section with

[PUT MORE RESULTS IN HERE]

6.4 Results

The figures obtained in Section 6.3 suggest that the bounds on $K_{\boldsymbol{x}}$ from Section 6.2 could be significantly tightened for location mixtures. The main bounds that we have discussed so far are Theorem 6.2 which states that under mild conditions, $K_{\boldsymbol{x}} \leq n$; and Theorem 6.7 which provides a sufficient condition for $K_{\boldsymbol{x}} = 1$ for component densities in the exponential class of densities. In this section we will present new bounds on $K_{\boldsymbol{x}}$ that either tighten these bounds, or extend them to different classes of component densities.

6.4.1 Results for $n = 2$

The first bound we present is Theorem 6.8 which extends Theorem 6.7 to a different class of component densities in the case that $n = 2$, and which tightens the result to be both sufficient and necessary on this class. The theorem concerns component densities, $f(x)$, which satisfy the following assumptions.

A1 (Continuity). The density $f(x)$ is continuous and is supported on the whole real line.

A2 (Differentiability). The first and second derivatives of $f(x)$ exist and are continuous.

A3 (Unimodality). The density $f(x)$ has a single mode at $x = 0$. That is, $f'(x) > 0$ 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. The density $f(x)$ has only two points of inflection, located at $x = \pm i$.

A6. The density $f(x)$ satisfies $f'(x) < -f'(x - 2i)$ for $x \in (i, \infty)$.

Theorem 6.8. *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 location mixture using f as the component density. Then*

$$K_{\boldsymbol{x}} = K(\boldsymbol{x}; f) = \begin{cases} 1 & |x_2 - x_1| \leq 2i, \\ 2 & \textit{otherwise.} \end{cases} \tag{6.37}$$

The proof of Theorem 6.8 will be given in Section 6.4.3 after some discussion concerning how the likelihood curve behaves when $n = 2$.

It is worth verifying that Assumptions A1 through to A6 are satisfied by some common densities. We remark here on the following common unimodal densities:

$$\text{Normal with variance } \sigma^2 \qquad f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/(2\sigma^2)} \tag{6.38}$$

$$\text{Cauchy with scale } \gamma \qquad g(x) = \frac{1}{\pi(\gamma + x^2/\gamma)} \tag{6.39}$$

Clearly, Assumptions A1 through to A4 are satisfied for both $f(x)$ and $g(x)$. The inflection points of $f(x)$ are located at $x = \pm\sigma$ and the inflection points of $g(x)$ are located at $x = \pm\gamma/\sqrt{3}$, satisfying Assumption A5. A plot of $f'(x)$ against $-f'(x - 2\sigma)$ and of $g'(x)$ against $-g'(x - 2\gamma/\sqrt{3})$ in Figure 6.7 makes it clear that Assumption A6 is satisfied too. Of course, one could make this rigourous if desired.

6.4.2 The likelihood curve when $n = 2$

The shape of $\Gamma_{\boldsymbol{x}}$ can provide us with some insight into the behaviour of $K_{\boldsymbol{x}}$. In Figure 6.8, we give some examples of $\Gamma_{\boldsymbol{x}}$ for $n = 2$ using a normal component density with variance $\sigma^2 = 1$. [FIX AND MAKE REPRODUCIBLE (Gamma not U on labels)] In particular, we note that the distance between x_1 and x_2 has a strong effect on the shape

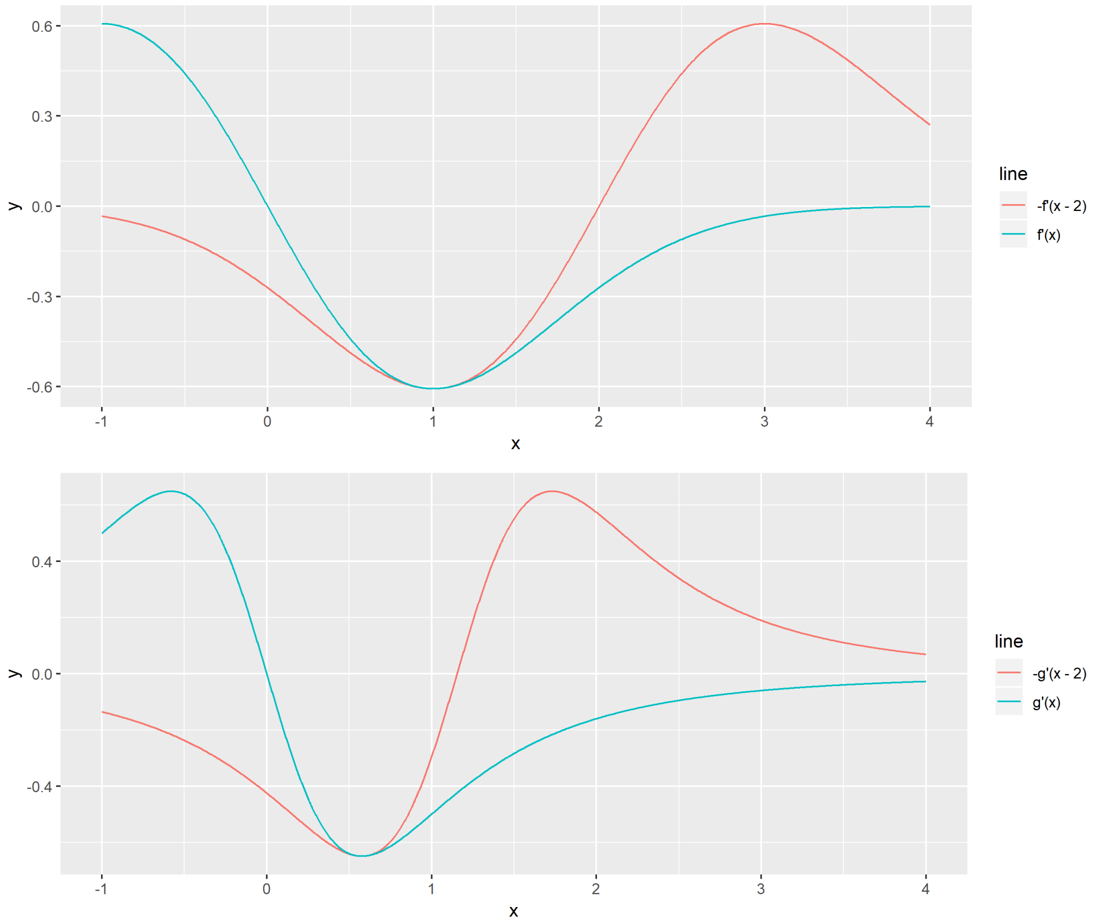


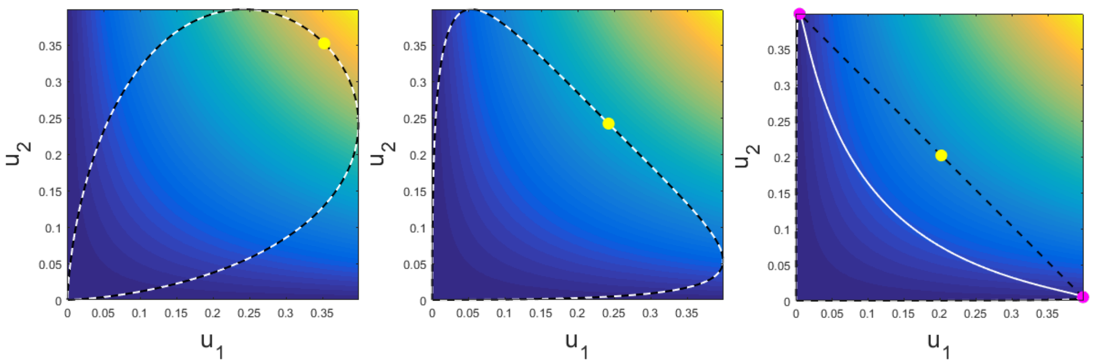
Figure 6.7 – *Plots of $f'(x)$ against $-f'(x-2\sigma)$ and $g'(x)$ against $-g'(x-2\gamma/\sqrt{3})$ for $\sigma = 1$ and $\gamma = 1$.*

of $\Gamma_{\boldsymbol{x}}$. In Figure 6.8a, the points are distance 1 apart and $\Gamma_{\boldsymbol{x}}$ is the boundary of $\mathrm{conv}(\Gamma_{\boldsymbol{x}})$. In this case, it is clear that $K_{\boldsymbol{x}} = 1$. In Figure 6.8c, 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 6.8c, is shown in Figure 6.8b.

Obtaining results about where these boundaries lie is very difficult in higher dimensions. In [32], 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.

[WE USE UNIMODALITY TO BE ABLE TO JUST LOOK AT CURVATURE ALONG INTERVAL]

[WHAT ABOUT IF $N > 2$ AND ALL POINTS WITHIN DISTANCE 2I?]



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

Figure 6.8 – *The curve* $\Gamma_{\boldsymbol{x}}$ *for three different* $\boldsymbol{x}$ *along with the boundary of* $\mathrm{conv}(\Gamma_{\boldsymbol{x}})$. *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.*

6.4.3 Proof of Theorem 6.8

The proof of Theorem 6.8 is based on the discussion contained in [32, Section 4], in which Lindsay discussed how the curvature of Γ relates to the support points of $\hat{Q}$. The important realisation we use here is that points of support must correspond to regions of non-negative curvature of Γ.

Proof. If $x_1 = x_2$ then clearly $K_{\boldsymbol{x}} = 1$. Otherwise, without loss of generality, assume $x_1 < x_2$. By the unimodality of f (A3), the points of support of any maximizing mixing distribution, $\hat{Q}$, must lie between x_1 and x_2 [38, Proposition 25]. Hence $\hat{\boldsymbol{\gamma}} = \boldsymbol{\gamma}(\hat{Q}, \boldsymbol{x}, f)$ must lie in the convex hull of $\Gamma^*_{\boldsymbol{x},f} = \{\boldsymbol{\gamma}(\theta; \boldsymbol{x}, f) : \theta \in [x_1, x_2]\}$.

Consider the behaviour of $\boldsymbol{\gamma}(\theta; \boldsymbol{x}, f) = (f(x_1 - \theta), f(x_2 - \theta))$ as we increase θ from x_1 to x_2. Since $f(x)$ has a single mode at $x = 0$, $f(x_1 - \theta)$ is non-increasing and $f(x_2 - \theta)$ is non-decreasing along this interval. So $\boldsymbol{\gamma}(\theta; \boldsymbol{x}, f)$ crosses the line $\gamma_1 = \gamma_2$ once only and by the symmetry of $f(x)$, this occurs at $\theta = (x_1 + x_2)/2$. We also have that by Lemma 6.9, $\hat{\boldsymbol{\gamma}}$ must lie on the line $\gamma_1 = \gamma_2$.

Since f is continuous and twice differentiable, $\boldsymbol{\gamma}$ traces out a continuous, smooth curve. The signed curvature

$$k(\theta) = \frac{-f'(x_1 - \theta)f''(x_2 - \theta) + f'(x_2 - \theta)f''(x_1 - \theta)}{(f'(x_1 - \theta)^2 + f'(x_2 - \theta)^2)^{\frac{3}{2}}} \tag{6.40}$$

is defined everywhere since $f'(x)$ is zero only at $x = 0$ and so for $x_1 \neq x_2$, the denominator is non-zero everywhere. The sign of $k(\theta)$ matches the sign of

$$S(\theta) = \begin{vmatrix} \gamma_1'(\theta; \boldsymbol{x}) & \gamma_1''(\theta; \boldsymbol{x}) \\ \gamma_2'(\theta; \boldsymbol{x}) & \gamma_2''(\theta; \boldsymbol{x}) \end{vmatrix} = \begin{vmatrix} -f'(x_1 - \theta) & f''(x_1 - \theta) \\ -f'(x_2 - \theta) & f''(x_2 - \theta) \end{vmatrix}. \tag{6.41}$$

Recall from Section 6.2.4.1, that the points of support of a maximizing mixture $\hat{Q}$, correspond with the contact points of $\Gamma_{\boldsymbol{x},f}$ with the support hyperplane of $\mathrm{conv}(\Gamma_{\boldsymbol{x},f})$ that contains $\hat{\boldsymbol{\gamma}}$. Note that any point $\boldsymbol{\gamma}(\theta_j;\boldsymbol{x},f)$ that is in contact with a support hyperplane of $\mathrm{conv}(\Gamma_{\boldsymbol{x},f})$ must have non-negative curvature.

First, let us 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 $\theta=(x_1+x_2)/2$. At this point, the curvature of $\boldsymbol{\gamma}$ has the same sign as

$$S\left(\frac{x_1+x_2}{2}\right)=\begin{vmatrix}-f'(\frac{x_1-x_2}{2}) & f''(\frac{x_1-x_2}{2})\\ -f'(\frac{x_2-x_1}{2}) & f''(\frac{x_2-x_1}{2})\end{vmatrix}. \tag{6.42}$$

Since $x_2-x_1>2i$, by A5, $f''((x_2-x_1)/2)>0$. Similarly, $f''((x_1-x_2)/2)>0$. 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. Since $\boldsymbol{\gamma}$ has negative curvature when it crosses the line $\gamma_1=\gamma_2$, and since $\hat{\gamma}$ lies on that line, the maximizing point $\hat{\gamma}$ must be the convex combination of two separate points in $\Gamma_{\boldsymbol{x},f}$ and so $K_{\boldsymbol{x}}=2$.

Now assume that $x_2-x_1\leq 2i$. By Lemma 6.10, there is only one point at which $\boldsymbol{\gamma}$ is pointing along the direction $(1,-1)$. Now, by the symmetry of f, $\Gamma_{\boldsymbol{x},f}$ must have a line of symmetry along $\gamma_1=\gamma_2$. Since $\hat{\boldsymbol{\gamma}}$ lies on the line $\gamma_1=\gamma_2$ by Lemma 6.9, the support line of $\Gamma_{\boldsymbol{x},f}$ that contains $\hat{\boldsymbol{\gamma}}$ must point along the direction $(1,-1)$. Each contact point of $\Gamma_{\boldsymbol{x},f}$ with one of its support lines must be tangent to that support line. Hence, there is only one possible point of contact with the support line that contains $\hat{\gamma}$ and so $K_{\boldsymbol{x}}=1$.

□

Lemma 6.9. *Let $f(x)$ be a component density symmetric around $x=0$, and $\boldsymbol{x}=(x_1,x_2)\in\mathbb{R}^2$ such that the conditions of Theorem 6.1 are met. Let $\hat{Q}$ be a maximum likelihood location mixture of $f(x)$ to $\boldsymbol{x}$. Then $f_{\hat{Q}}(x_1)=f_{\hat{Q}}(x_2)$.*

Proof. Let (u,v) denote the coordinates of $\hat{\boldsymbol{\gamma}}=(f_{\hat{Q}}(x_1),f_{\hat{Q}}(x_2))$. By Theorem 6.2 we have

$$(u,v)=p\boldsymbol{\gamma}(\theta_1;\boldsymbol{x},f)+(1-p)\boldsymbol{\gamma}(\theta_2;\boldsymbol{x},f) \tag{6.43}$$

$$=p(f(x_1-\theta_1),f(x_2-\theta_1))+(1-p)(f(x_1-\theta_2),f(x_2-\theta_2)) \tag{6.44}$$

for some choice of $p \in (0,1]$, and $\theta_1, \theta_2 \in \mathbb{R}$. By the symmetry of $f(x)$, $f(x) = f(-x)$ and so

$$
\begin{aligned}
u &= pf(x_1 - \theta_1) + (1-p)f(x_1 - \theta_2) && (6.45)\\
&= pf(\theta_1 - x_1) + (1-p)f(\theta_2 - x_1) && (6.46)\\
&= pf(x_2 - [x_1 + x_2 - \theta_1]) + (1-p)(f(x_2 - [x_1 + x_2 - \theta_2]) && (6.47)
\end{aligned}
$$

and likewise

$$v = pf(x_1 - [x_1 + x_2 - \theta_1]) + (1-p)f(x_1 - [x_1 + x_2 - \theta_2]). \tag{6.48}$$

Hence the point (v, u) can be written as

$$(v, u) = p\boldsymbol{\gamma}(x_1 + x_2 - \theta_1; \boldsymbol{x}, f) + (1-p)\boldsymbol{\gamma}(x_1 + x_2 - \theta_2; \boldsymbol{x}, f), \tag{6.49}$$

and so $(v, u) \in \text{conv}(\Gamma_{\boldsymbol{x},f})$. Since both $(u, v), (v, u) \in \text{conv}(\Gamma_{\boldsymbol{x},f})$, we must have that $\frac{1}{2}(u + v, u + v) \in \text{conv}(\Gamma_{\boldsymbol{x},f})$. Now

$$
\begin{aligned}
\mathcal{L}\left(\frac{1}{2}(u+v, u+v)\right) &= 2\ln((u+v)/2) && (6.50)\\
&= \ln\left(\left(\frac{u+v}{2}\right)^2\right) && (6.51)\\
&\geq \ln(uv) && (6.52)\\
&= \mathcal{L}(u, v). && (6.53)
\end{aligned}
$$

However, $(u, v) = \hat{\boldsymbol{\gamma}}$, which by Theorem 6.1 is the unique maximizing point of $\mathcal{L}$ in $\text{conv}(\Gamma_{\boldsymbol{x},f})$. Hence $(u, v) = \frac{1}{2}(u + v, u + v)$ and so $u = v$. □

Lemma 6.10. *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*

$$-f'(x_1 - \theta) = f'(x_2 - \theta) \tag{6.54}$$

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 $f'(x)$:

$$f'(x) \text{ is } \begin{cases} \text{positive and increasing,} & x \in (-\infty, -i) \\ \text{positive and decreasing,} & x \in (-i, 0) \\ \text{negative and decreasing,} & x \in (0, i) \\ \text{negative and increasing,} & x \in (i, \infty). \end{cases} \tag{6.55}$$

We also note, from A4, that $f'(x)$ is an odd function. Using this and rearranging (6.54) we obtain the equivalent equation

$$g(\theta) = h(\theta) \tag{6.56}$$

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 (6.56) 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 (6.56) on this interval.

For $\theta \in (0, -i + x_2 - x_1)$ we can again use assumption A6 by observing that since

$$f'(\theta) < -f'(\theta - 2i) \tag{6.57}$$

for $\theta \in (i, \infty)$, we also have that for $\theta \in (-\infty, -i)$

$$f'(-\theta) < -f'(-\theta - 2i) \tag{6.58}$$

and so for $\theta \in (-\infty, i)$,

$$f'(-\theta + 2i) < -f'(-\theta) \tag{6.59}$$

which, since f' is odd, is equivalent to

$$-f'(\theta - 2i) < f'(\theta). \tag{6.60}$$

So 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 (6.56) 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, (6.54) must have only one solution. □

6.4.4 Results for general n

To obtain results for $n \geq 2$ we now restrict our component density to be normal with fixed variance σ^2. In this case, we can make use of the gradient function defined in Section 6.2.4 and the equations, (6.28) to (6.30), which it must satisfy at a maximizing mixture.

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

$$f_\sigma(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \tag{6.61}$$

the gradient function defined in (6.25), evaluated at a mixture Q which places masses $\boldsymbol{p}$ at locations $\boldsymbol{\theta}$, becomes

$$D_Q(\theta; \boldsymbol{x}, f_\sigma) = -n + \sum_{i=1}^{n} \frac{\exp(-(x_i - \theta)^2/2\sigma^2)}{\sum_{j=1}^{m} p_j \exp(-(x_i - \theta_j)^2/2\sigma^2)}, \tag{6.62}$$

and equations (6.28) to (6.30) become

$$\frac{1}{n}\sum_{i=1}^{n} \Psi_k(x_i; \boldsymbol{\theta}, \boldsymbol{p}) = 1, \qquad k = 1, \ldots, m, \tag{6.63}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \Psi_k(x_i; \boldsymbol{\theta}, \boldsymbol{p}) = \theta_k, \qquad k = 1, \ldots, m, \tag{6.64}$$

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \theta_k)^2 \Psi_k(x_i; \boldsymbol{\theta}, \boldsymbol{p}) \leq \sigma^2, \qquad k = 1, \ldots, m, \tag{6.65}$$

where we have written

$$\Psi_k(x; \boldsymbol{\theta}, \boldsymbol{p}) = \Psi_k(x; \boldsymbol{\theta}, \boldsymbol{p}, f_\sigma) = \frac{f_\sigma(x - \theta_k)}{\sum_{j=1}^{m} p_j f_\sigma(x - \theta_j)}. \tag{6.66}$$

for ease of notation.

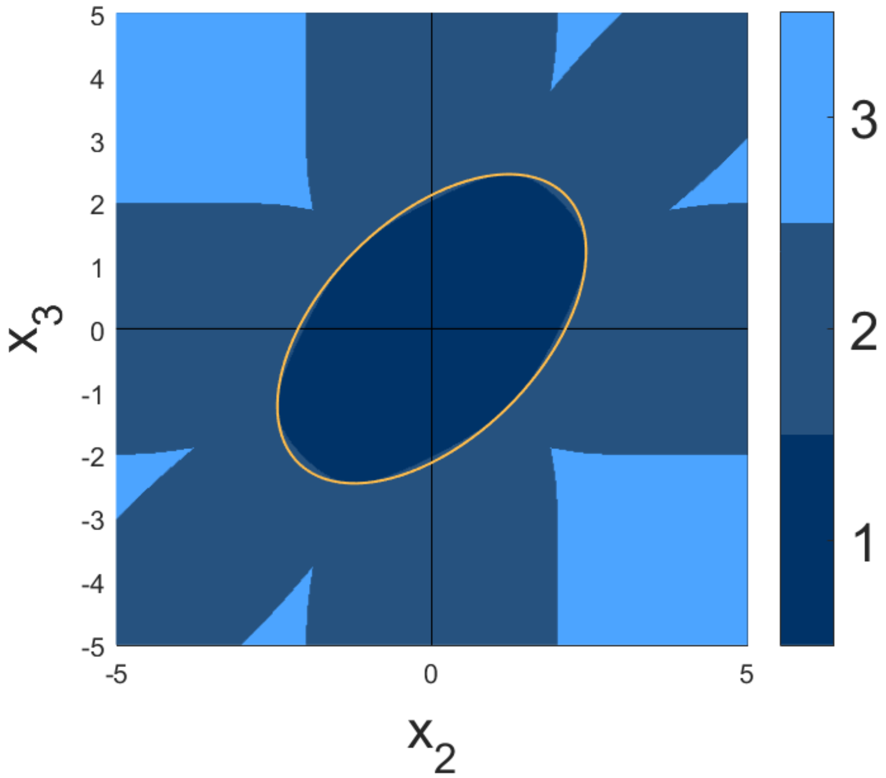


Figure 6.9 – *The bound obtanied in Theorem 6.11 tells us that C_1 must lie within the orange ellipse. The true shape of C_1 is given by the dark blue region.*

We will show that these three equations constrain the regions $C_1, \ldots, C_n$ as defined in (6.36). This will be done in Theorem 6.12. However, as a gentle introduction, we will start with the much simpler problem of just bounding C_1.

Theorem 6.11. *If $\boldsymbol{x} \in C_1$ then*

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{\boldsymbol{x}})^2 \leq \sigma^2, \tag{6.67}$$

where

$$\bar{\boldsymbol{x}} = \frac{1}{n}\sum_{i=1}^{n} x_i. \tag{6.68}$$

Proof. If $\boldsymbol{x} \in C_1$ then the maximizing mixture has one component and so $\Psi_1(x; \boldsymbol{\theta}, \boldsymbol{p}) = 1$. Then (6.64) gives us that $\theta_1 = \bar{\boldsymbol{x}}$ and combining this with (6.65) completes the proof. $\square$

In Figure 6.9 we compare the bound above to the 'flag graphs' we produced in section 6.3.2.

We now state a generalisation of Theorem 6.11 that bounds the regions $C_1, \ldots, C_n$.

Theorem 6.12. *If $\boldsymbol{x} \in C_m$ for some $m \leq n$, then there exists a subset of the x_i, A, such that A contains at least n/m elements and*

$$\frac{1}{n}\sum_{i \in A}(x_i - \bar{x}_A)^2 \leq \sigma^2, \tag{6.69}$$

where

$$\bar{x}_A = \frac{1}{|A|}\sum_{i \in A} x_i. \tag{6.70}$$

Proof. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{p}}$ represent the m point maximizing mixture of $\boldsymbol{x} \in C_m$. By Lemma 6.13, there exists a k^* such that $\Psi_{k^*}(x_i; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}) \geq 1$ for at least n/m different x_i. Call the set of these x_i, A_{k^*}.

Let $A_{\theta_{k^*}}$ be the set of the $|A_{k^*}|$ closest x_i to θ_{k^*}. Then

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \theta_{k^*})^2 \Psi_{k^*}(x_i; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}) \geq \frac{1}{n}\sum_{i \in A_{\theta_{k^*}}}(x_i - \theta_{k^*})^2 \tag{6.71}$$

$$\geq \frac{1}{n}\sum_{i \in A_{\theta_{k^*}}}\left(x_i - \bar{x}_{A_{\theta_{k^*}}}\right)^2. \tag{6.72}$$

From (6.65),

$$\frac{1}{n}\sum_{i \in A_{\theta_{k^*}}}\left(x_i - \bar{x}_{A_{\theta_{k^*}}}\right)^2 \leq \sigma^2. \tag{6.73}$$

□

The bound for the C_2 region given by Theorem 6.12 is compared to the 'flag graphs' from Section 6.3.2 in Figure 6.10. We observe that the bound does not appear as tight as the one bounding C_1.

Lemma 6.13. *Let f be a density supported on all of $\mathbb{R}$. For all $x \in \mathbb{R}$, and all discrete probability distributions with masses $\boldsymbol{p}$ at locations $\boldsymbol{\theta}$,*

$$\max_k \left[\Psi_k(x; \boldsymbol{p}, \boldsymbol{\theta}, f)\right] \geq 1. \tag{6.74}$$

Proof. Let $x \in \mathbb{R}$, and $\boldsymbol{\theta}, \boldsymbol{p}$ be the location and weights of an m point discrete probability distribution. Choose θ_{k^*} to be the smallest θ_j that satisfies

$$f(x - \theta_{k^*}) \geq f(x - \theta_j) \qquad j = 1, \ldots, m. \tag{6.75}$$

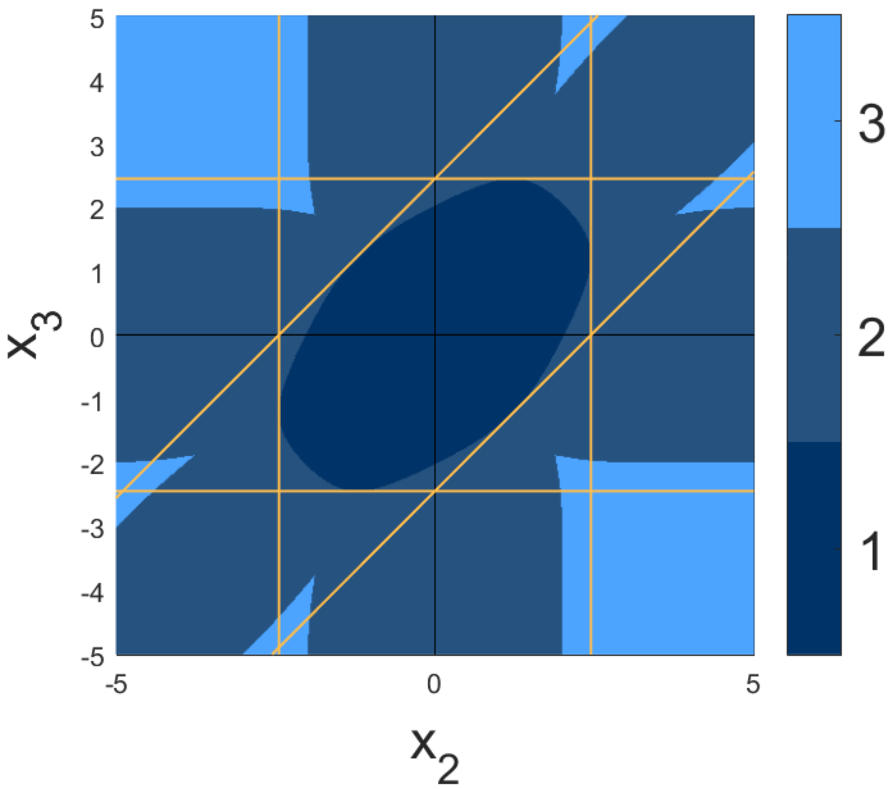


Figure 6.10 – *The bound obtanied in Theorem 6.12 tells us that C_2 must lie between the pairs of parallel orange lines. The true shape of C_2 is given by the middle blue shaded region.*

(We choose the smallest in case there is more than one θ_j which satisfies the above). Then

$$\Psi_{k^*}(x; \boldsymbol{p}, \boldsymbol{\theta}, f) = \frac{f(x - \theta_{k^*})}{\sum_{j=1}^m p_j f(x - \theta_j)} \geq \frac{f(x - \theta_{k^*})}{\sum_{j=1}^m p_j f(x - \theta_{k^*})} = 1. \tag{6.76}$$

$\square$

6.4.5 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 consider a random sample

$$\boldsymbol{X} = (X_1, \ldots, X_n) \tag{6.77}$$

where the X_i are i.i.d. with normal distribution

$$X_i \sim N(\mu, \sigma_1^2). \tag{6.78}$$

Given a component density, we may consider the probabilities

$$\mathbb{P}[\boldsymbol{X} \in C_m], \qquad m = 1, \ldots, n. \tag{6.79}$$

Theorem 6.14. *Let the component density with which we are finding a maximum likelihood location mixture be normal with variance σ_2^2. Then*

$$\mathbb{P}[\boldsymbol{X} \in C_1] \leq \mathbb{P}\left(\chi^2_{n-1} \leq \frac{n\sigma_2^2}{\sigma_1^2}\right) \tag{6.80}$$

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

Proof. From Theorem 6.11,

$$\mathbb{P}(\boldsymbol{X} \in C_1) \leq \mathbb{P}\left(\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})^2 \leq n\sigma_2^2\right) \tag{6.81}$$

$$= \mathbb{P}\left(\frac{1}{\sigma_1^2}\sum_{i=1}^{n}(X_i - \bar{\boldsymbol{X}})^2 \leq \frac{n\sigma_2^2}{\sigma_1^2}\right) \tag{6.82}$$

$$= \mathbb{P}\left(\chi^2_{n-1} \leq \frac{n\sigma_2^2}{\sigma_1^2}\right). \tag{6.83}$$

□

Theorem 6.14 is of particular interest when the $\sigma_1 = \sigma_2$. In this case, our mixture can select the 'true' number of components, which will happen if $\boldsymbol{X} \in C_1$. The probability of this occuring satisfies

$$\mathbb{P}(\boldsymbol{X} \in C_1) \leq \mathbb{P}\left(\chi^2_{n-1} \leq n\right) \tag{6.84}$$

which converges to $1/2$ as $n \to \infty$. This is a simple proof that the number of components chosen by a normal location mixture is not a consistent estimator for the true number of components.

Similar results to this have already been discovered. In [40], Hartigan considered the mixture

$$(1-p)N(0,1) + pN(\theta, 1) \tag{6.85}$$

and showed that the likelihood ratio between $\theta = 0$ and $\theta \neq 0$ converges to ∞ as $n \to \infty$. Also related is the result [37] that certain penalised likelihoods do not underestimate the true number of components (and so it is reasonable to expect that the unpenalised likelihood might overestimate the true number of components).

[SOMETHING RELATED TO RESULT FOR N = 2?]

[UP TO HERE]

6.4.6 Final Observation

[EXTRA STUFF]

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

Theorem 6.15. *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 $\boldsymbol{x} \notin \partial E_m$. □

6.5 Conclusion

[WRITE THIS]

Chapter 7

Deconvolution

7.1 Introduction

In this final chapter, we look at another statistical problem in which we perform an optimization over discrete probability distributions and find that the number of points of support of the optimal distribution is much smaller than expected. In contrast to Chapter 6, we will take a descriptive and practically minded point of view, rather than the more theory driven previous chapter. This is partly due to the additional complexity of the problem, which prevents many of the tools used on maximum likelihood mixtures from working here. However, this does not stop us from pointing out similarities between the two problems, and an empirical exploration allows us to benefit from the similar phenomenon that occurs in a variety of scenarios.

The problem is one of *deconvolution*, the recovery of the distribution, F_X, or density, f_X, of some random variable X, from measurements of

$$W = X + U \tag{7.1}$$

where the measurement error U is independent of X.

In the case that we know the distribution of U, Carroll and Hall [41] and Carroll and Stefanski [42] proposed a deconvolution kernel density estimator of f_X. If the distribution of U is unknown, then we may hope to estimate it from replicate measurements, $W_{jk} = X_j + U_{jk}$, for $1 \leq k \leq N_j$ and $1 \leq j \leq n$. Delaigle, Hall, and Meister provide an estimator for f_X in this scenario. However, up until Delaigle's and Hall's 2016 paper, "Methodology for nonparametric deconvolution when the error distribution is unknown," [2], there had been no nonparametric method to estimate the distribution of X when we had no data on the distribution of U (see the discussion in Section 1 of [2]).

In this chapter, we focus on the methods introduced in this paper. We will start in Section 7.1.1 by summarizing deconvolution methods for when the error distribution is known, or estimated from replicates. This will serve as a basis for a discussion on Delaigle's and Hall's "Methodology for nonparametric deconvolution when the error distribution is unknown," in Section 7.2. In Section 7.3, we will empirically demonstrate that these methods produce a similar phenomenon to the one encountered when using maximum likelihood location mixtures, and we will explore how this phenomenon manifests under various conditions. We will also point out how we can take advantage of this phenomenon. Finally, in Section 7.4, we will make a general observation about the deconvolution problem, and suggest that its behaviour might not be totally unexpected.

7.1.1 Prior deconvolution methods

We start with the estimator of Stefanski and Carroll in [42]. In this scenario, we let X and U be independent random variables with probability density functions f_X and f_U respectively. We observe a set of n independent observations, $\{W_j\}_{j=1}^{n}$, where each W_j is an observation of

$$W = X + U. \tag{7.2}$$

Furthermore, we assume that f_U is known. Given this information, the goal is to estimate f_X.

The estimator makes use of characteristic functions. A random variable X with density f_X has characteristic function

$$\phi_X(t) = \int e^{itx} f_X(x)\,\mathrm{d}x \tag{7.3}$$

which can be inverted via

$$f_X(x) = \frac{1}{2\pi}\int e^{-itx}\phi_X(t)\,\mathrm{d}t \tag{7.4}$$

if ϕ_X is integrable. One property of characteristic functions that is convenient for this deconvolution problem is that since X and U are independent, we know that the characteristic function of $W = X + U$ is

$$\phi_W = \phi_X \phi_U, \tag{7.5}$$

or equivalently,

$$\phi_X = \frac{\phi_W}{\phi_U} \tag{7.6}$$

if $|\phi_U(t)| > 0$ for all real t. We assume this in the following estimator.

Let f_K be a bounded, even, probability density function whose characteristic function, ϕ_K, satisfies

$$\sup_t |\phi_K(t)/\phi_U(t/h)| < \infty, \qquad \int |\phi_K(t)/\phi_U(t/h)| \, \mathrm{d}t < \infty, \tag{7.7}$$

for any fixed $h > 0$ (we impose these conditions make various functions integrable). We may estimate the density of W via the usual kernel density estimator

$$\hat{f}_W(t) = \frac{1}{nh} \sum_{j=1}^{n} K((W_j - x)/h), \tag{7.8}$$

which has characteristic function

$$\hat{\phi}_W(t) = \phi_{W_{\text{EMP}}}(t)\phi_K(ht) \tag{7.9}$$

where $\phi_{W_{\text{EMP}}}(t)$ denotes the empirical characteristic function of $\{W_j\}_{j=1}^n$,

$$\phi_{W_{\text{EMP}}}(t) = \frac{1}{n} \sum_{j=1}^{n} e^{itW_j}. \tag{7.10}$$

Motivated by (7.6), we may estimate the characteristic function of X by

$$\hat{\phi}_X = \hat{\phi}_W / \phi_U \tag{7.11}$$

and the density of X by inverting $\hat{\phi}_X$. Overall, the estimator for f_X is given by

$$\hat{f}_X(x) = \frac{1}{2\pi} \int e^{-itx} \phi_K(ht)\phi_{W_{\text{EMP}}}(t)/\phi_U(t) \, \mathrm{d}t. \tag{7.12}$$

If the distribution of U is unknown, we may still be able to estimate it if replicate measurements are present. Methods for deconvolution with the presence of replicates are presented in [43], [44], [45], and also [46] in the case of heteroscedastic errors. Here we look at the estimator of Delaigle, Hall, and Meister [45] as it has the most similarities with the estimator of interest in this chapter.

In this setting, we measure

$$W_{jk} = X_j + U_j \qquad j = 1, \dots, n \text{ and } k = 1, \dots, N_j \tag{7.13}$$

where the X_j and U_{jk} are independent and are identically distributed as X and U respectively. A consistent estimator of ϕ_U is

$$\hat{\phi}_U(t) = \left| \frac{1}{N} \sum_{j=1}^{n} \sum_{(k1,k2)\in \mathscr{S}_j} \cos\left(t(W_{jk_1} - W_{jk_2})\right) \right|^{1/2}. \tag{7.14}$$

In light of (7.12), this suggests an estimator of f_X given by

$$\hat{f}_X(x) = \frac{1}{2\pi} \int e^{-itx} \phi_K(ht) \phi_{W_{\mathrm{EMP}}}(t) / (\hat{\phi}_U(t) + \rho) \, \mathrm{d}t \tag{7.15}$$

with

$$\phi_{W_{\mathrm{EMP}}}(t) = \frac{1}{M} \sum_{j=1}^{n} \sum_{k=1}^{N_j} e^{itW_{jk}} \tag{7.16}$$

where $h > 0$ is a bandwidth, $\rho \geq 0$ is a ridge parameter, and K is a symmetric kernel density with compact support, whose characteristic function, ϕ_K, has compact support.

[WHY COMPACT SUPPORT?]

The presence of the ridge parameter is to account for potential fluctuations in $\hat{\phi}_U$ that could cause the denominator in (7.15) to be zero, or too close to zero. We may alternatively use another method to prevent $\hat{\phi}_U$ from getting too close to zero such as replacing $\hat{\phi}_U(t)$ with some other function for $|t|$ larger than some threshold t^*.

A similar approach could be used in scenarios where we are able to sample directly from the distribution of U. See also [47] and [48] for further discussion of cases where samples of the errors are available.

In the case where we do not know the distribution of U, nor have access to replicate measurements or direct measurements from U, methods for deconvolution generally require F_U to have a parameteric or semi-parametric form. Such methods can be found in [49], [50], [51], and [52].

7.2 Method for deconvolution when the error is unknown

Delaigle's and Hall's 2016 paper, "Methodology for nonparametric deconvolution when the error distribution is unknown," [2], presents a non-parametric deconvolution estimator for when the error distribution is unknown, and we do not have access to replicates or samples from the error distribution. This is unusual in deconvolution problems. As in the methods described above, deconvolution methods usually require assumptions about the shape or scale of the error distribution, or rely on additional samples to estimate it.

In this section we summarise the estimator of [2], and discuss some practicalities that arise in the implementation.

7.2.1 Problem Setup

Suppose that we have n observations of

$$W_j = X_j + U_j \qquad\qquad j = 1, \ldots, n \tag{7.17}$$

where X_j and U_j are independent and identically distributed as X and U respectively, and X and U are independent. We use ϕ_X and ϕ_U to denote the characteristics functions of X and U, and F_X and F_U for the distributions. We use f_X and f_U for the densities if they exist.

We still make some assumptions about U and its characteristic function, ϕ_U. These are

A7. ϕ_U is real-valued.

A8. For U discrete, ϕ_U is non-negative and is zero at at most a countable number of points, and for U continuous, ϕ_U is strictly positive on the whole real line.

Assumption A7 is equivalent to assuming that F_U is symmetric. Assumption A8 is a standard assumption in deconvolution problems.

[WHAT DOES $\phi_U \geq 0$ MEAN FOR f_U]

We also make assumptions on the distribution of X.

A9. F_X is not symmetric.

A10. It is not possible to decompose X as

$$X = Y + Z \tag{7.18}$$

for nondegenerate and independent random variables Y and Z with F_Z symmetric.

We require these assumptions because all we know about F_U is that it is symmetric. So if F_X was also symmetric, we could not distinguish it from F_U, and if X is itself made up of a symmetric part Z, we could not distinguish U from $Z + U$.

To form the estimator for F_X, we will make use of the *phase function*, which for a random variable V, is defined by

$$\rho_V = \frac{\phi_V}{|\phi_V|} \tag{7.19}$$

on all points where $\phi_V \neq 0$. Note that from Assumption A7, we have that

$$\phi_U = |\phi_U| \tag{7.20}$$

and so $\rho_U = 1$. Since $W = X + U$, with X and U independent, we have that

$$\phi_W = \phi_X \phi_U \tag{7.21}$$

and so from (7.20), on all points where $\phi_U \neq 0$,

$$\rho_W = \rho_X. \tag{7.22}$$

In fact, all random variables of the form $V = X + Z$, where Z is symmetric and independent of X, will have phase function ρ_W, and the variance of these will satisfy $\mathrm{Var}(V) \geq \mathrm{Var}(X)$. This motivates the final assumption we make for X.

A11. F_X has the uniquely smallest variance out of all distributions with phase function ρ_X.

7.2.2 Estimator

From the discussion above, we might think to take our estimator for F_X to be the distribution that has smallest variance out of all distributions with phase function equal to some estimator of ρ_W constructed from $W_1, \ldots, W_n$.

We can estimate ρ_W by

$$\hat{\rho}_W = \frac{\hat{\phi}_W}{\left|\hat{\psi}_W\right|^{1/2}} \tag{7.23}$$

where

$$\hat{\phi}_W(t) = \frac{1}{n}\sum_{j=1}^{n} \exp(itW_j) \tag{7.24}$$

and

$$\hat{\psi}_W(t) = \frac{1}{n(n-1)}\sum_{k=1}^{n}\sum_{j\neq k} \exp(it(W_j - W_k)) \tag{7.25}$$

are consistent estimators of ϕ_W and $\psi_W = |\phi_W|^2$ respectively.

Ideally, we would like to choose our estimate for F_X to have phase function $\hat{\rho}_X = \hat{\rho}_W$, or equivalently, $\hat{\phi}_W(t)|\hat{\phi}_X(t)| - |\hat{\phi}_W(t)|\hat{\phi}_X(t) = 0$ for all t. However, $\hat{\rho}_W(t)$ grows less reliable for large values of $|t|$, and we should preference matching $\hat{\rho}_X$ and $\hat{\rho}_W$ around $t = 0$ over large values of $|t|$. Taking this into consideration, our goal is to instead choose an F that makes small

$$T(F) = \int_{-\infty}^{\infty} \left| \hat{\phi}_W(t) |\phi_F(t)| - \left| \hat{\psi}_W(t) \right|^{1/2} \phi_F(t) \right|^2 w_1(t)\, \mathrm{d}t, \tag{7.26}$$

where $w_1(t)$ is some non-negative weight function that assigns greater weight when $|t|$ is small. We also expect that $\hat{\phi}_U = \hat{\phi}_W/\phi_F$ should be symmetric, and have magnitude less than or equal to 1. We try to satisfy this by constructing penalties which we desire to be small,

$$P1(F) = \int \Im \left(\hat{\phi}_W \bar{\phi}_F \right) w_2(t)\, \mathrm{d}t \tag{7.27}$$

and

$$P2(F) = \int \left(\hat{\phi}_U(t) - 1 \right) I_{\hat{\phi}_U(t)>1} w_2(t)\, \mathrm{d}t \tag{7.28}$$

where $\bar{a}$ represents the complex conjugate of a and $w_2(t)$ is some non-negative weight function with bounded support.

Overall, the method is as follows. Let $\mathscr{F}$ be the set of distributions over which we search for our estimator (for example, discrete distributions with no more than m points of support). Find

$$F_0 = \arg\min_{F\in\mathscr{F}} \left[T(F) + \lambda_1 P1(F) + \lambda_2 P2(F) \right], \tag{7.29}$$

for some choice of scaling factors λ_1 and λ_2, and set

$$T_{\min} = T(F_0) \tag{7.30}$$

$$P1_{\min} = P1(F_0) \tag{7.31}$$

$$P2_{\min} = P2(F_0). \tag{7.32}$$

Then our estimator for F_X is

$$\hat{F}_X = \arg\min_{F\in\mathscr{F}} \mathrm{Var} F \tag{7.33}$$

subject to the constraints $T(F) \leq T_{\min}$, $P1(F) \leq P1_{\min}$, and $P2(F) \leq P2_{\min}$.

7.2.3 Numerical Implementation

Delaigle and Hall suggest performing this optimization problem by approximating F_X with a discrete distribution on m points. They suggest choosing the location of the probability masses, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)$, to be fixed and distributed uniformly but randomly along $[\min(W_i), \max(W_i)]$ and letting the probability weights $\boldsymbol{p} = (p_1, \ldots p_m)$ be the variables over which we find the solution. We will denote by $F_{\boldsymbol{\theta},\boldsymbol{p}}$ the distribution that places mass p_j at θ_j for $j = 1, \ldots, m$. Its characteristic function is given by

$$\phi_{\boldsymbol{\theta},\boldsymbol{p}} = \sum_{j=1}^{m} p_j \exp(it\theta_j) \tag{7.34}$$

and its phase function by

$$\rho_{\boldsymbol{\theta},\boldsymbol{p}} = \frac{\sum_{j=1}^{m} p_j \exp(it\theta_j)}{\left|\sum_{j=1}^{m} p_j \exp(it\theta_j)\right|}. \tag{7.35}$$

It has variance

$$\mathrm{Var}(F_{\boldsymbol{\theta},\boldsymbol{p}}) = \sum_{j=1}^{m} p_j\theta_j^2 - \left(\sum_{j=1}^{m} p_j\theta_j\right)^2. \tag{7.36}$$

For the weight function, $w_1(t)$, Delaigle and Hall use the Epanechnikov kernel, rescaled to an interval $[-t^*, t^*]$ where t^* is the smallest $t > 0$ such that

$$\left|\hat{\phi}_W(t)\right| \leq n^{-1/4}. \tag{7.37}$$

For the weight function in the penalty terms, they use $w_2(t) = I_{|t|\leq t^*}(\delta)^{-1}$, where δ is the distance between consecutive points in their discretization of the integrals in (7.27) and (7.28). They use $\lambda_1 = \lambda_2 = 500$ for the scaling factors in (7.29). There is some freedom of choice in the number of points in the approximating discrete distribution but the authors suggest that in their experience, $m = 5\sqrt{n}$ is a reasonable choice.

7.2.4 Converting to continuous distribution

If X is continuous, then we will often want to use $\hat{F}_X = F_{\boldsymbol{\theta},\boldsymbol{p}}$ to create an estimate of the density, f_X. Perhaps the most obvious solution is to use

$$\hat{f}_X(x) = \sum_{j=1}^{m} p_j K_h(x - \theta_j) \tag{7.38}$$

where $K_h(x)$ is a kernel with some bandwidth $h > 0$. This is exactly equivalent to

$$\hat{f}_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\phi_{\boldsymbol{\theta},\boldsymbol{p}}\phi_K(ht)\,\mathrm{d}t \tag{7.39}$$

where ϕ_K is the Fourier transform of $K_1(x)$. However, since we only used $t \in [-t^*, t^*]$ when constructing $\phi_{\boldsymbol{\theta},\boldsymbol{p}}$, it is reasonable to assume that $\phi_{\boldsymbol{\theta},\boldsymbol{p}}$ is less reliable for t outside this range. This motivates replacing $\phi_{\boldsymbol{\theta},\boldsymbol{p}}$ with

$$\tilde{\phi} = \begin{cases} \phi_{\boldsymbol{\theta},\boldsymbol{p}}, & t \in [-t^*, t^*], \\ r(t), & \text{otherwise}, \end{cases} \tag{7.40}$$

where $r(t)$ is some ridge function. Delaigle and Hall suggest following [45] and using

$$r = \hat{\phi}_W / \hat{\phi}_{U,P} \tag{7.41}$$

where $\hat{\phi}_{U,P}$ is the characteristic function of a Laplace distribution with variance equal to an estimator of the variance of U. They also suggest choosing the bandwidth, h, using the two-stage plug-in bandwidth of Delaigle and Gijbels [53] [54].

We wish to remark here that the conversion from the discrete $F_{\boldsymbol{\theta},\boldsymbol{p}}$ to the continuous $\hat{f}_X(x)$ has no effect on the phenomenon of interest in this chapter. We are interested only in the discrete distribution $F_{\boldsymbol{\theta},\boldsymbol{p}}$. However, we have included this section for completeness.

7.3 Empirical Results

Following the methods outlined above, the authors noted that $F_{\boldsymbol{\theta},\boldsymbol{p}}$ was usually supported on a small number of points. That is, only a few of the p_j were non-zero. A typical example of this is given in Figure 7.1.

A theoretical justification for this behaviour is hard to obtain. In Section 7.4 we make an observation about the problem which could be thought of as being in favour of this phenomenon, but for the most part, we have no explanation for why it occurs.

In this section, we instead explore the phenomenon empirically, and suggest that we might benefit by allowing both the θ_j and p_j to be the variables of our optimization, rather than fixing the θ_j as suggested in [2].

We start by demonstrating the phenomenon using a variety of parameters and distributions. The base example (Figure 7.1) has the following setup:

- The true distribution, F_X, is chi-squared with 3 degrees of freedom, rescaled to have variance $\sigma_X = 1$.
- The error distribution, F_U, is normal.
- The noise to signal ratio (NSR) is 1/5. That is, $\sigma_U^2 = \sigma_X^2/5$.

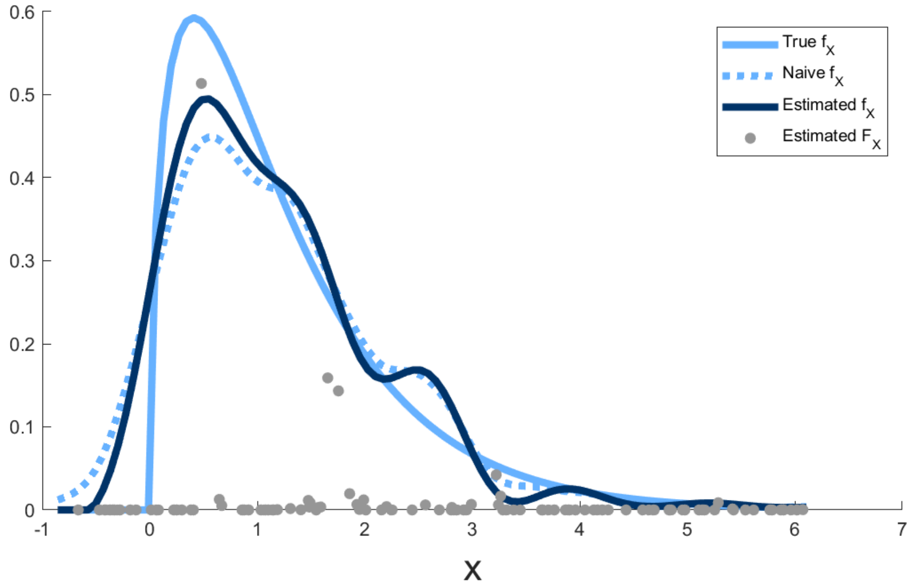
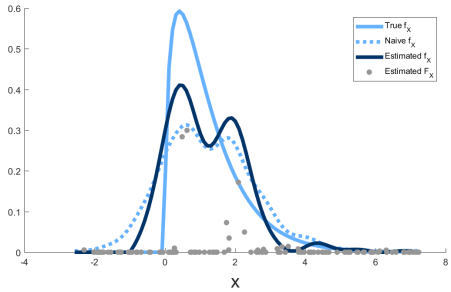


Figure 7.1 – *A typical example of $F_{\boldsymbol{\theta},\mathbf{p}}$ being supported on only a few points. The probability masses (θ_j, p_j) are represented by grey points, and the resulting estimating density is in dark blue. The true density is a chi-squared density with 3 degrees of freedom, and the errors were normal with variance $\sigma_U^2 = \sigma_X^2/5$. There were $n = 500$ samples taken from W. The naive density is a kernel density estimator of the W_j.*

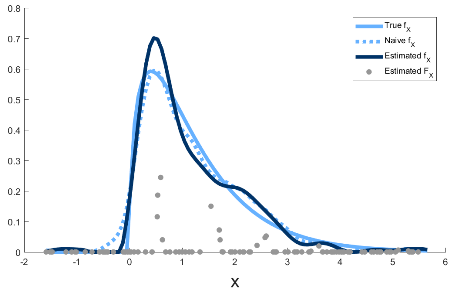
- We sample $n = 500$ points of $W = X + U$.

As recommended in the paper, we use $m = 5\sqrt{n}$ point masses in our approximating distribution $\hat{F}_X$. In each of Figures 7.2a, 7.2b, 7.2c, and 7.2d, we change one of these properties and plot the result. Of particular interest in each plot is $\hat{F}_X$ which we represent by placing a grey point at each probability mass. In each of these figures, as well as in the base example in Figure 7.1, we observe that the majority of the probability masses of $\hat{F}_X$ take values very close to, or equal to zero.

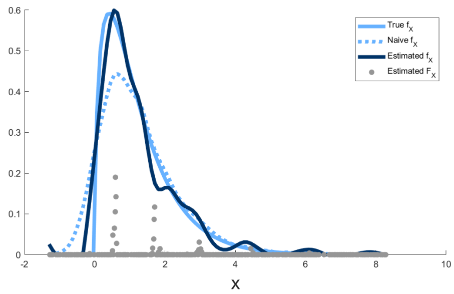
Given that most of the probability masses of $\hat{F}_X$ do not contribute to the final density, we might hope to reduce the complexity of our optimization problem by using a more appropriate number of masses. Of course, we do not know a priori where these masses should be located along the x axis, and so should make their locations variables in our optimization. This essentially doubles the dimension of our optimization and so to achieve any computational speed up we should aim to use fewer than half the original number of points. In the figures above, we observe that roughly 10 to 20 points out of $m = 112$ or $m = 500$ masses have weights that are visibly greater than 0. Furthermore, it is feasible that two points located close to each other might coalesce into a single point in such a way as to improve the objective if their locations are allowed to vary. This encourages us to proceed using roughly 10 to 20 probability masses with variable weights and locations for $\hat{F}_X$.



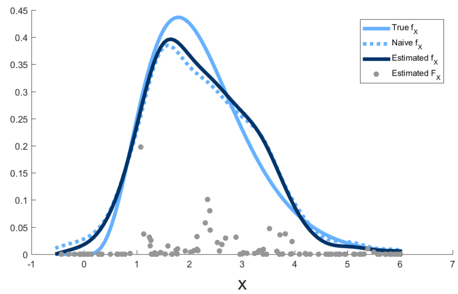
(a) *Using a NSR of 1*



(b) *Using a cauchy error distribution*



(c) *Using* $n = 10000$ *points*

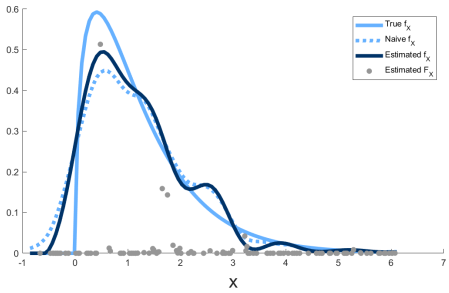


(d) *Using a gamma(5,2) true distribution*

Figure 7.2 – *Four variations on the base example in Figure 7.1.*

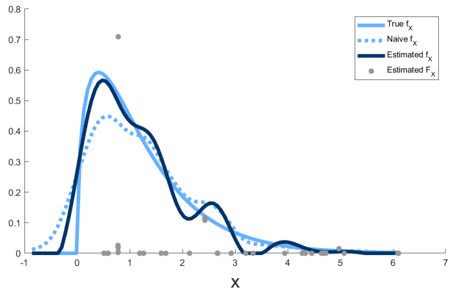
We start with Figure 7.3 in which we compare the result we obtain using the fixed masses method of Delaigle and Hall, and the results we obtain when we allow the location of the probability masses in $F_{\boldsymbol{\theta},\boldsymbol{p}}$ to vary. As well as plotting $\hat{F}_X$ and $\hat{f}_X$, we also provide a table with the various objective values obtained in the final result, as well as the time taken to run the code on a i5-4670K CPU running at 4.3 GHz, running MATLAB R2018a. For comparison, we also give the various objective values obtained by using the empirical distribution of W as $\hat{F}_X$. The values T, $P1$, $P2$ are as defined in Section 7.2.2, we use $OBJ1$ to denote the objective of our first minimization, $T(F) + \lambda_1 P1(F) + \lambda_2 P2(F)$, and Var to denote the variance of $\hat{F}_X$ (our second objective).

For this particular example, we note that we need only 10 masses to obtain smaller values for both of our objectives when compared to the original fixed masses, and that this results in a significant speed up, as expected. We get even further improvement going from 10 to 20 masses. However, when we use 40 masses, we note that OBJ_1 is significantly larger than in any of the cases where we use fewer masses, and that $\mathrm{Var}\hat{F}_X$ is smaller than in any other case. One potential explanation here is that the parameter space becomes too complex for our optimization routine to find good solutions if we allow too many moving masses and so OBJ_1 is much larger than the global minimum. This means that when we come to our second objective of minimizing $\mathrm{Var}\hat{F}_X$, the constraints $T(F) \leq T_{\min}$, $P1(F) \leq P1_{\min}$, and $P2(F) \leq P2_{\min}$ are lax, and so we have more room to search for feasible distributions with small variance.



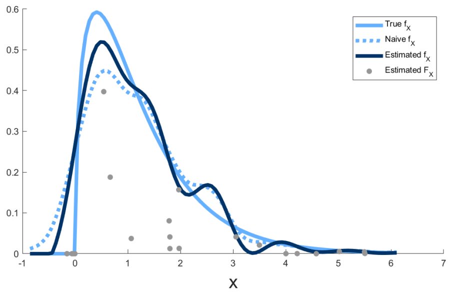
OBJ_1	0.2284
$\text{Var}\hat{F}_X$	0.8613
$T(\hat{F}_X)$	3.8697e-07
$P1$	4.5679e-04
$P2$	0
t	195s

(a) *112 fixed masses*



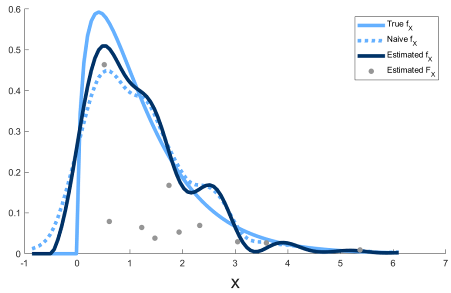
OBJ_1	54.3622
$\text{Var}\hat{F}_X$	0.6554
$T(\hat{F}_X)$	1.7882e-06
$P1$	0.1087
$P2$	0
t	76s

(b) *40 moving masses*



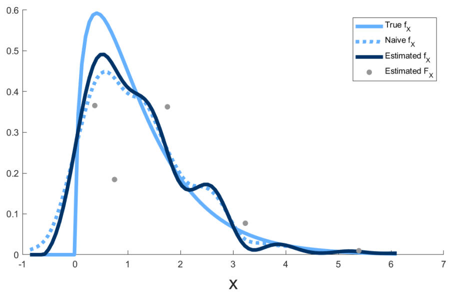
OBJ_1	0.0396
$\text{Var}\hat{F}_X$	0.8042
$T(\hat{F}_X)$	4.4981e-07
$P1$	7.9209e-05
$P2$	0
t	37s

(c) *20 moving masses*



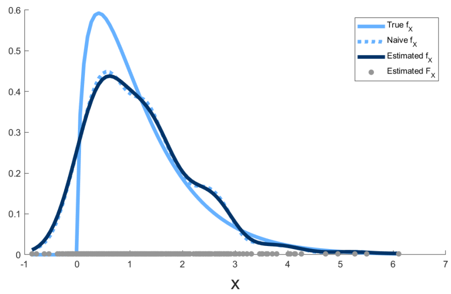
OBJ_1	0.1374
$\text{Var}\hat{F}_X$	0.8269
$T(\hat{F}_X)$	4.2464e-07
$P1$	2.7478e-04
$P2$	0
t	15s

(d) *10 moving masses*



OBJ_1	0.3905
$\text{Var}\hat{F}_X$	0.8888
$T(\hat{F}_X)$	3.6213e-07
$P1$	7.8093e-04
$P2$	0
t	4s

(e) *5 moving masses*



OBJ_1	2.7811e-07
$\text{Var}\hat{F}_X$	1.0350
$T(\hat{F}_X)$	2.7810e-07
$P1$	1.0079e-14
$P2$	1.7097e-14
t	NA

(f) *The empirical distribution of W*

Figure 7.3 – *Comparison of results between fixed and variable probability mass locations.*

It also appears as if the 40 mass case is a closer fit for the true curve than any of the other estimates, despite acheiving worse results in the first objective. Although this is purely conjecture, we suggest that this could be because the constraints $T(F) \leq T_{\text{min}}$, $P1(F) \leq P1_{\text{min}}$, and $P2(F) \leq P2_{\text{min}}$ are often too strict to acheive the variance we desire in the second optimization. The first optimization found a solution which just happened to be far enough away from the global minimum to allow for enough freedom in minimizing the variance that we obtained a result that acheived closer to the true variance. One could try using as constraints, $T(F) \leq (1+\delta)T_{\text{min}}$, $P1(F) \leq (1+\delta)P1_{\text{min}}$, and $P2(F) \leq (1+\delta)P2_{\text{min}}$, to allow for this behaviour when the first minimization attains a result closer to the global optimum, but we cannot see any obvious way to determine good values for δ.

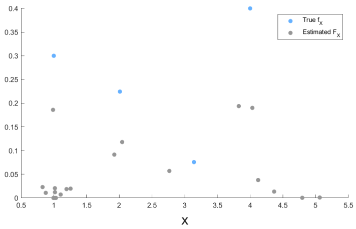
The results in Figure 7.3 are encouraging, and in our experience are consistent in a wide variety of scenarios. We present some more examples in Figure 7.4. Of course, it is possible that there is a wide class of deconvolution problems in which a large number of point masses are required to acheive a good solution, and that we have just happened to avoid examples of these. However, we do not think that this is a large concern. One can simply repeat the deconvolution with an increasing number of point masses in $\hat{F}_X$ until it does not result in an improvement in our objectives.

7.3.1 R Package

Given all the discussion above, we see no reason not to use masses with variable locations. We have used this new method in the R [55] Package 'deconvolve' [56]. This package is a collection of tools for performing non-parametric deconvolution on measurement error problems. It contains functions for finding bandwidths, deconvolved densities, and non-parametric regression estimates. The methods discussed in this Chapter for when the error distribution is unknown form just one part of the package. We give a brief overview of the performance of our implementation of it in R.

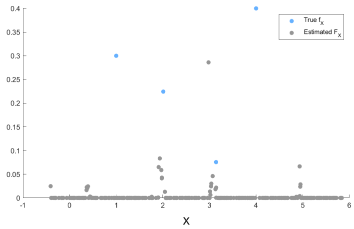
In Figure 7.5, we show the output of our package when the contaminated data is identical to that used in producing Figure 7.1. As in Figure 7.3, we provide the values obtained of each of our objectives, as well as the time taken to run.

To perform the two non-linear optimizations we use the package Nlcoptim [57]. In the example in Figure 7.5, we used 20 moving masses in our estimate $\hat{F}_X$. This make it directly comparable to 7.3c, since we are using identical data and parameters. While the objective values obtained are worse than those in the equivalent MATLAB implementation, the resulting density is still reasonable and we still benefit from a significant



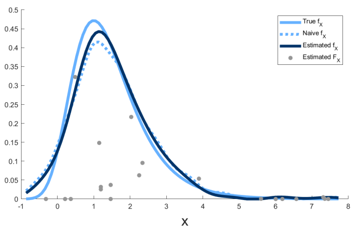
OBJ_1	0.8786
$\mathrm{Var}\hat{F}_X$	1.6542
$T(\hat{F}_X)$	1.1094e-08
$P1$	0.0018
$P2$	0
t	62s

(a) *True distribution discrete (represented by blue points) with normal errors, $n = 5000$, moving masses.*



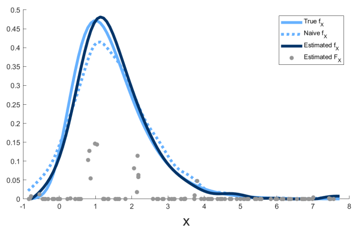
OBJ_1	106.6533
$\mathrm{Var}\hat{F}_X$	1.5731
$T(\hat{F}_X)$	0.0192
$P1$	0.2133
$P2$	0
t	1354s

(b) *True distribution discrete (represented by blue points) with normal errors, $n = 5000$, fixed masses.*



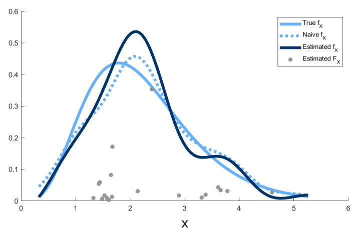
OBJ_1	0.3742
$\mathrm{Var}\hat{F}_X$	1.0063
$T(\hat{F}_X)$	3.5329e-07
$P1$	7.4839e-04
$P2$	0
t	64s

(c) *True distribution gumbel with laplace errors, $n = 500$, moving masses.*



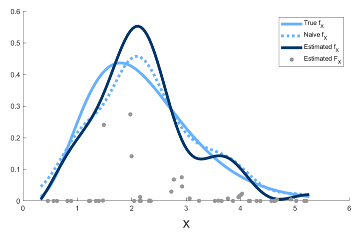
OBJ_1	0.2603
$\mathrm{Var}\hat{F}_X$	0.9413
$T(\hat{F}_X)$	4.9281e-07
$P1$	5.2056e-04
$P2$	0
t	195s

(d) *True distribution gumbel with laplace errors, $n = 500$, fixed masses.*



OBJ_1	0.1506
$\mathrm{Var}\hat{F}_X$	0.6176
$T(\hat{F}_X)$	1.2286e-05
$P1$	3.0108e-04
$P2$	0
t	47s

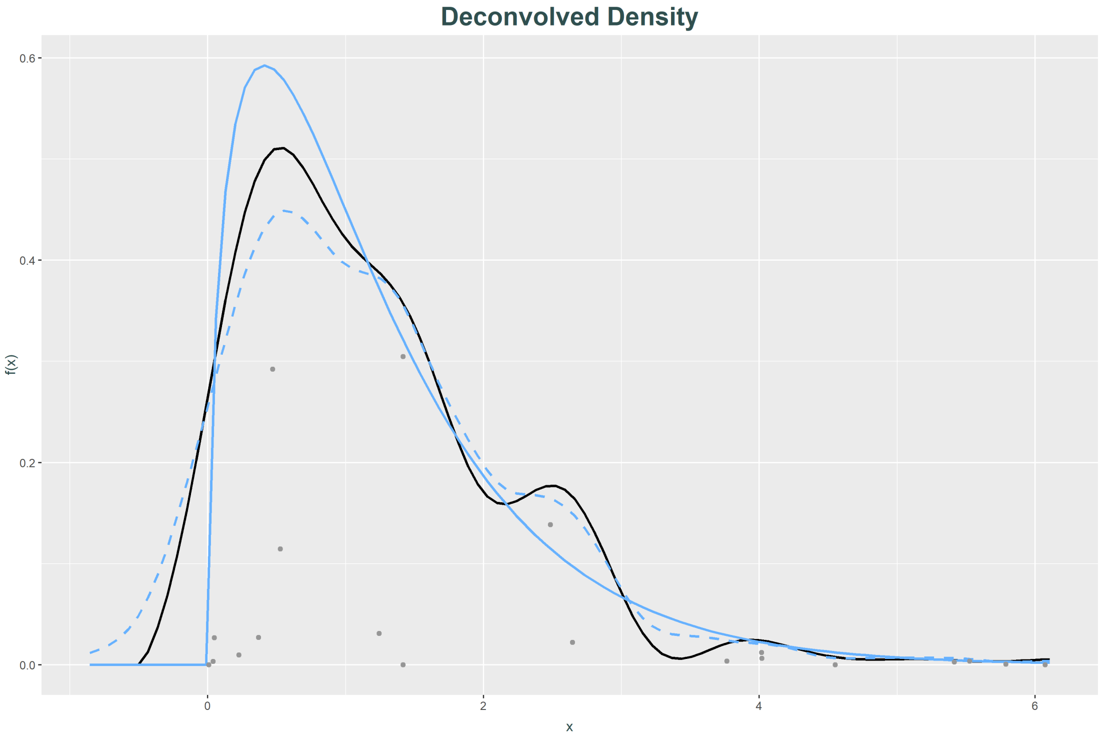
(e) *True distribution Gamma(5,2), with errors discrete (-1,0,1) with probabilities (0.2,0.6,0.2), $n = 100$, moving masses.*



OBJ_1	0.1681
$\mathrm{Var}\hat{F}_X$	0.5953
$T(\hat{F}_X)$	1.2921e-05
$P1$	3.3615e-04
$P2$	0
t	62s

(f) *True distribution Gamma(5,2), with errors discrete (-1,0,1) with probabilities (0.2,0.6,0.2), $n = 100$, fixed masses.*

Figure 7.4 – *Three more comparisons between moving masses and fixed masses.*



OBJ_1	2.277
$\mathrm{Var}\hat{F}_X$	0.8684
$T(\hat{F}_X)$	1.1672e-06
$P1$	4.554e-03
$P2$	0
t	50s

Figure 7.5 – *The output of the 'deconvolve' package with data as in Figure 7.1 and using* $m = 20$ *moving masses.*

speed up over the original fixed mass implementation. In practise, we found this to be fairly typical of our implementation in 'deconvolve'.

7.4 General Observations and Results

We end this chapter with a simple observations about our deconvolution problem. We do not claim that it is new, but we think that it is still worth stating.

Theorem 7.1. *Let* F_X *and* F_Y *be two distributions, and let*

$$F_Z = (1-\lambda)F_X + \lambda F_Y \qquad \lambda \in [0,1] \tag{7.42}$$

be a convex combination of these distributions. If $\rho_X = \rho_Y$*, then*

$$\rho_Z = \rho_X = \rho_Y. \tag{7.43}$$

Proof. Write ϕ_X and ϕ_Y for the characteristic functions of X and Y. These are complex valued functions which we can write in the form

$$\phi_X = r_X(t)e^{i\theta_X(t)}, \tag{7.44}$$

$$\phi_Y = r_Y(t)e^{i\theta_Y(t)}, \tag{7.45}$$

where $r_X(t), r_Y(t)$ take values on $[0, 1]$ and where

$$\rho_X = e^{i\theta_X(t)}, \tag{7.46}$$

$$\rho_Y = e^{i\theta_Y(t)}. \tag{7.47}$$

The characteristic function of F_Z is

$$\phi_Z = \int e^{itx} \,\mathrm{d}\left[(1-\lambda)F_X(x) + \lambda F_Y(x)\right] \tag{7.48}$$

$$= (1-\lambda)\phi_X + \lambda\phi_Y \tag{7.49}$$

$$= (1-\lambda)r_X(t)\rho_X + \lambda r_Y(t)\rho_Y. \tag{7.50}$$

If $\rho_X = \rho_Y$ then

$$\phi_Z = \left((1-\lambda)r_X(t) + \lambda r_Y(t)\right)\rho_X \tag{7.51}$$

and so

$$\rho_Z = \rho_X = \rho_Y. \tag{7.52}$$

□

Recall that ideally, we would like to search the set of all distributions with phase function equal to that of W to find the distribution with smallest variance. Theorem 7.1 tells us that this set is convex. Furthermore, given any two distributions F_X and F_Y with variances σ_X^2 and σ_Y^2 respectively, a convex combination of them, $F_Z = (1-\lambda)F_X + \lambda F_Y$, will have variance $\sigma_Z^2 \geq (1-\lambda)\sigma_X^2 + \lambda\sigma_Y^2$. Hence, the minimum variance distribution with phase function exactly equal to ρ_W will be on the boundary of this set, and will not be a non-trivial convex combination of two distributions with phase functions equal to ρ_W. Any convex combination of two discrete distributions, F_X and F_Y, has at least as many points of support as either F_X or F_Y, and so it is perhaps not totally surprising that minimizing the variance in our problem results in distributions with few points of support.

Of course, in reality we do not search over distributions with phase function exactly equal to W, but rather over distributions with phase function in some sense 'close' to an empirical estimate of the phase function of W. However the intuition gained above could potentially still be relevant.

7.5 Conclusion

[WRITE THIS]

Chapter 8

Conclusion to Part II

Bibliography

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

[2] 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.

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

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

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

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

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

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

[9] 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.

[10] 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.

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

[12] 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.

[13] 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.

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

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

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

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

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

[19] 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.

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

[21] 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.

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

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

[26] Thomas M Liggett. *Interacting Particle Systems:*. Grundlehren der mathematischen Wissenschaften. Springer New York, 1985.

[27] Boris L Granovsky and Neal Madras. The noisy voter model. *Stochastic Process. Appl.*, 55(1):23–43, January 1995.

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

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

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

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

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

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

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

[35] H Akaike. A new look at the statistical model identification. *IEEE Trans. Automat. Contr.*, 19(6):716–723, December 1974.

[36] Gideon Schwarz. Estimating the dimension of a model. *Ann. Stat.*, 6(2):461–464, March 1978.

[37] Brian G Leroux. Consistent estimation of a mixing distribution. *Ann. Stat.*, 20(3): 1350–1360, September 1992.

[38] Bruce G Lindsay. *Mixture Models: Theory, Geometry, and Applications.* IMS, 1995.

[39] Dankmar Böhning. A review of reliable maximum likelihood algorithms for semiparametric mixture models. *J. Stat. Plan. Inference*, 47(1):5–28, October 1995.

[40] J A Hartigan. A failure of likelihood asymptotics for normal mixtures. In *Proceedings of the Berkeley conference in honor of Jerzy Neyman and Jack Kiefer*, volume 2, pages 807–810, 1985.

[41] Raymond J Carroll and Peter Hall. Optimal rates of convergence for deconvolving a density. *J. Am. Stat. Assoc.*, 83(404):1184–1186, December 1988.

[42] Leonard A Stefanski and Raymond J Carroll. Deconvoluting kernel density estimators. *Statistics*, 21(2):169–184, January 1990.

[43] Tong Li and Quang Vuong. Nonparametric estimation of the measurement error model using multiple indicators. *J. Multivar. Anal.*, 65(2):139–165, 1998.

[44] Xihong Lin and Raymond J Carroll. Semiparametric estimation in general repeated measures problems. *J. R. Stat. Soc. Series B Stat. Methodol.*, 68(1):69–88, 2006.

[45] Aurore Delaigle, Peter Hall, and Alexander Meister. On deconvolution with repeated measurements. *Ann. Stat.*, 36(2):665–685, April 2008.

[46] Julie McIntyre and Leonard A Stefanski. Density estimation with replicate heteroscedastic measurements. *Ann. Inst. Stat. Math.*, 63(1):81–99, February 2011.

[47] Peter J Diggle and Peter Hall. A fourier approach to nonparametric deconvolution of a density estimate. *J. R. Stat. Soc. Series B Stat. Methodol.*, 55(2):523–531, 1993.

[48] Michael H Neumann and O Hössjer. On the effect of estimating the error density in nonparametric deconvolution. *J. Nonparametr. Stat.*, 7(4):307–330, January 1997.

[49] Cristina Butucea and Catherine Matias. Minimax estimation of the noise level and of the deconvolution density in a semiparametric convolution model. *Bernoulli*, 11 (2):309–340, April 2005.

[50] Alexander Meister. Density estimation with normal measurement error with unknown variance. *Stat. Sin.*, 16(1):195–211, 2006.

[51] Cristina Butucea, Catherine Matias, and Christophe Pouet. Adaptivity in convolution models with partially known noise distribution. *Electron. J. Stat.*, 2:897–915, 2008.

[52] Alois Kneip, Léopold Simar, Ingrid Van Keilegom, and Others. Boundary estimation in the presence of measurement error with unknown variance. *J. Econom.*, 2012.

[53] A Delaigle and I Gijbels. Estimation of integrated squared density derivatives from a contaminated sample. *J. R. Stat. Soc. Series B Stat. Methodol.*, 64(4):869–886, October 2002.

[54] A Delaigle and I Gijbels. Practical bandwidth selection in deconvolution kernel density estimation. *Comput. Stat. Data Anal.*, 45(2):249–267, March 2004.

[55] R Core Team. R: A language and environment for statistical computing, 2018.

[56] Aurore Delaigle, Timothy Hyndman, and Tianying Wang. deconvolve: Deconvolution tools for measurement error problems, 2019.

[57] X Chen and X Yin. Nlcoptim: Solve nonlinear optimization with nonlinear constraints, 2017.