The University of Melbourne

Doctoral Thesis

Efficient Optimization for Statistical Inference

Author:
Timothy Hyndman

Supervisors:
Prof. Peter Taylor
Prof. Aurore Delaigle

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

Operations Research
School of Mathematics and Statistics

July 2018

THE UNIVERSITY OF MELBOURNE

Abstract

Faculty of Science
School of Mathematics and Statistics

Doctor of Philosophy

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.

We note that while a typical thesis is a cohesive whole, this is certainly not a requirement to obtaining a doctorate. Indeed, to obtain the degree of Doctor of Philosphy, the candidate is required to produce a substantial piece of original research. We believe that the sum of the research contained in both parts of this thesis is sufficiently substantial, and thus adequate for submission.

Part I

The Distribution of the Coupling Time for the Ising Heat-Bath Dynamics on Lattices

Chapter 2

Introduction

2.1 The Ising Model

The Ising model was first studied by Ernst Ising in his 1924 thesis [1] under the supervision of Wilhelm Lenz who invented the model. It was originally motivated by the phenomenon of ferromagnetism but it has since been used in various other applications CITE. [Gas thingo, Potts model - image denoising, why is it popular etc]

The Ising model has enjoyed a prominent position in the statistical physics literature. This is largely due to the existence of a phase transition [CITE], a sharp transition in the large scale behaviour of the model as a parameter moves past a critical value. Additionally, the Ising model is both relatively simple, and also mathematically tractable in some non-trivial cases [CITE]. 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 - an assignment 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 $i \in V$ with $\sigma[i]$. Each configuration has an associated energy, given by

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

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,

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

2.2 Coupling from the Past

One of the primary concerns regarding the Ising model is how to efficiently sample from the Gibbs measure. A direct approach is computationally intractable as the number of configurations grows exponentially in the size of the graph 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, initially, one does not know how long to run the chain for. In principal, bounds on this time can be achieved, but in practise, proving these bounds can be very challenging.

An alternative to MCMC was introduced by Propp and Wilson called Coupling from the Past (CFTP) [2]. 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 in Chapter 3, and on any vertex transitive graph in Chapter 4. Our main result in each chapter will be proving that, when appropriately scaled, the coupling time essentially converges to a Gumbel distribution as the size of the graph increases.

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 i.i.d. rate-one Poisson clock. Define the probability

$$p_i(\sigma) = \frac{e^{\beta S_i(\sigma)}}{e^{\beta S_i(\sigma)} + e^{-\beta S_i(\sigma)}} \tag{2.5}$$

where

$$S_i(\sigma) = \sum_{j \sim i} \sigma[j] \tag{2.6}$$

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.

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. In order to do this, it will prove convenient to use a random mapping representation for the jump process. This will also help us outline an implementation of CFTP for our coupling.

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

$$\sigma'[i] = \begin{cases} 1, & u \leq p_i(\sigma), \\ -1, & u > p_i(\sigma). \end{cases} \tag{2.7}$$

Let $\mathscr{V}$ and U be independent, with $\mathscr{V}$ uniform on V and U uniform on $[0,1]$. Then, updating our chain at rate $n = |V|$, and performing updates from σ to σ' according to $\sigma' = f(\sigma, \mathscr{V}, U)$, we recover the dynamics described in Section 2.2.1.

We also note that f is monotonic. 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 chains, $(\mathscr{T}_t)_{t \geq 0}$ and $(\mathscr{B}_t)_{t \geq 0}$, with initial states

$$\mathscr{T}_0 = (1, 1, \ldots, 1) \tag{2.8}$$

$$\mathscr{B}_0 = (-1, -1, \ldots, -1) \tag{2.9}$$

that update together at rate n. On the kth update at time t_k, update $\mathscr{T}_{t_k}$ to $f(\mathscr{T}_{t_k}, \mathscr{V}_k, U_k)$ and update $\mathscr{B}_{t_k}$ to $f(\mathscr{B}_{t_k}, \mathscr{V}_k, U_k)$.

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

A more descriptive explanation of the coupling is that 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 (2.7).

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

$$T = \inf \left\{t : \mathscr{T}_t = \mathscr{B}_t\right\}. \tag{2.10}$$

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 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.

[EXTEND?]

2.3 Information percolation

A cornerstone to the proofs contained in Chapters 3 and 4 is the framework of information percolation, first introduced by Lubetzky and Sly in 2016 [4]. This paper proved the existence of cutoff with a constant order window for the Glaubery dynamics for the Ising model. Cutoff is a BLAHBLAHBLAH

In this section we write a summary of the basic framework and definitions that we will use.

2.3.1 The update sequence

We can encode each update with the tuple (i, u, t), where i is the vertex that is updated, t is the time of the update, and u is the value of the uniform random variable that tells us whether i is a plus or minus according to (2.7). The *update sequence* along an interval $(t_0, t_1]$ is the set of these tuples with $t_0 < t \leq t_1$. Given the state of our Markov Chain 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 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 $Y_{t_1}, Y'_{t_1} \in \{-1, +1\}^V$ 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 any two coupled chains Y and Y', $Y_t[i] = Y'_t[i]$. 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

$$\mathbb{P}[\mathscr{T}_t[i] \neq \mathscr{B}_t[i]] = \mathbb{P}[\mathscr{F}(i, 0, t) \neq \emptyset]. \tag{2.11}$$

For ease of notation, we will often use the shorthand

$$\mathcal{H}_i(t) := \mathscr{F}(i, t, t^*). \tag{2.12}$$

where t^* is some target time that should be clear from context. We call this the *update history of vertex i at time t*. Tracing $\mathcal{H}_i(t)$ backwards in time from t^* produces a subgraph of $\Omega \times [0, t^*]$ which we write as $\mathcal{H}_i$ and which we simply 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 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 any sufficiently small $\epsilon > 0$.

2.3.3 Oblivious updates

Important to the idea of the update support function is that of an oblivious update. In general, an oblivious update is one that does not depend on its neighbours. Write Δ_i for the degree of a vertex i. Recalling (2.5),

$$\frac{e^{-\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}} \leq p_i(\sigma) \leq \frac{e^{\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}}. \tag{2.13}$$

For a particular update (i,u,t), if $u \leq \frac{e^{-\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}}$ then i is updated to a plus regardless of the state of the neighbours of i and similarly, if $u > \frac{e^{\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}}$ then i is updated to a minus regardless of the state of the neighbours of i. If one of these two things happens then the update is an *oblivious update*. The rate of these updates at vertex i is

$$\theta_i = 1 - \left(\frac{e^{\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}} - \frac{e^{-\beta\Delta_i}}{e^{\beta\Delta_i}+e^{-\beta\Delta_i}}\right) \tag{2.14}$$

$$= 1 - \tanh(\beta\Delta_i). \tag{2.15}$$

If G is a Δ-regular graph then we can drop the subscript and write $\theta = 1 - \tanh(\beta\Delta)$ for the rate of oblivious updates at each vertex.

Of particular note is the effect of an oblivious update on the update history of a vertex. 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. These are not necessarily the only updates that can shrink the size of the update history of i, but they are essential in doing so.

Chapter 3

One Dimensional Case

In this chapter we prove the following theorem.

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

$$e^{-\sqrt{\theta/(4-3\theta)}e^{-z}} \leq \lim_{n\to\infty} \mathbb{P}[T_n < (z+\ln n)/\theta] \leq e^{-e^{-z}} \tag{3.1}$$

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

The simplification of only looking at the Ising model on the cycle means that BLAHBLAHBLAH

3.1 Information percolation on the cycle

On the cycle, we will use a different coupling of $\mathscr{T}_t$ and $\mathscr{B}_t$ via a new set of update rules that will replace those from (2.7). The new update rules will ensure that each of the update histories never contain more than one vertex at any one time. However, since the coupling time is a property of the specific coupling we choose, we must also verify that these new rules do not affect T.

The new update rules state that when vertex i updates, a spin σ_i' is chosen via

$$\sigma_i' = \begin{cases} +1 & U < \theta/2, \\ \sigma_{i-1} & \theta/2 \leq U < 1/2, \\ \sigma_{i+1} & 1/2 \leq U < 1-\theta/2, \\ -1 & U \geq 1-\theta/2. \end{cases} \tag{3.2}$$

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

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 (2.7). To show that the coupling time is unchanged, it is sufficient to verify that the joint jump probabilities of $(\mathscr{T}_t[i], \mathscr{B}_t[i])$ are unchanged for each possible configuration of spins of vertices $i-1$ and $i+1$. There are only nine possible configurations for the two neighbours of i in the top and bottom chain since $\mathscr{B}_t[i] \leq \mathscr{T}_t[i], \forall t$. Likewise, there are only three possible configurations for the updated spins $(\mathscr{T}_t[i]', \mathscr{B}_t[i]')$. Hence, given vertex i updates at time t, we can easily calculate all the required jump probabilities as shown in Table 3.1. These are unchanged whether using (2.7) or (3.2) and so the new rules do not change the coupled dynamics.

3.1.1 Update histories on the cycle

Under the update rules in (3.2), 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. So as t decreases from t^*, $\mathcal{H}_i(t)$ is a continuous-time random walk that dies at rate θ, moves left at rate $(1-\theta)/2$, and moves right at rate $(1-\theta)/2$. The probability that $\mathcal{H}_i(0) \neq \emptyset$ is simply the probability that the continuous-time random walk survives until time $t = 0$.

So recalling (2.11)

$$\mathbb{P}\left[\mathscr{B}_{t^*}[i] \neq \mathscr{T}_{t^*}[i]\right] = \mathbb{P}\left[\mathcal{H}_i(0) \neq \emptyset\right] = e^{-\theta t^*}. \tag{3.3}$$

3.2 Compound Poisson Approximation

Fix z and a time of interest, $t_* = (z + \ln n)/\theta$. For each vertex $i \in G$, we define indicators

$$X_i = \begin{cases} 1 & \mathscr{B}_{t^*}[i] \neq \mathscr{T}_{t^*}[i], \\ 0 & \mathscr{B}_{t^*}[i] = \mathscr{T}_{t^*}[i] \end{cases} \tag{3.4}$$

and set $W = \sum_{i \in V} X_i$. Note that from (3.3) we get

$$\mathbb{P}[X_i = 1] = e^{-\theta t_*} = \frac{e^{-z}}{n}. \tag{3.5}$$

The random variable W is closely related to the distribution of the coupling time T in that the events $\{W = 0\}$ and $\{T \leq t^*\}$ are the same. We will show that the limiting distribution of W as n increases is compound Poisson using Theorem 3.2 which we have taken from [5] which in turn is based on Stein's method for the compound Poisson distribution, introduced in [6]. Before stating the theorem as it applies to our problem, there are a few more quantities we need to define.

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

$$U_i = \sum_{j \in B_i} X_j, \qquad Z_i = \sum_{j \in C_i} X_j, \qquad W_i = \sum_{j \in D_i} X_j. \tag{3.6}$$

and B_i, C_i, and D_i are the vertex sets

$$B_i = \{j \neq i : |j - i| \leq b_n\}, \tag{3.7}$$

$$C_i = \{j \notin B_i \cup \{i\} : |j - i| \leq c_n\}, \tag{3.8}$$

$$D_i = V \setminus (B_i \cup C_i \cup \{i\}). \tag{3.9}$$

We have some freedom in choosing b_n and c_n, but they must be chosen such that various quantities go to zero as $n \to \infty$. One choice that will work in our circumstances is $b_n = \ln(n)$ and $c_n = \ln(n)^2$.

Define the following which are the parameters of the approximating compound Poisson distribution to W.

$$\lambda = \sum_{i\in V} \mathbb{E}\left[\frac{X_i}{X_i+U_i} I[X_i+U_i \geq 1]\right], \tag{3.10}$$

$$\mu_l = \frac{1}{l\lambda}\sum_{i\in V} \mathbb{E}\left[X_i I[X_i+U_i = l]\right], \qquad l \geq 1. \tag{3.11}$$

Also define quantities

$$\delta_1 = \sum_{i\in V}\sum_{k\geq 0} \mathbb{P}[X_i = 1, U_i = k]\mathbb{E}\left|\frac{\mathbb{P}[X_i = 1, U_i = k|W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1\right|, \tag{3.12}$$

$$\delta_4 = \sum_{i\in V} \left(\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i+U_i+Z_i]\right). \tag{3.13}$$

which we require to vanish as $n \to \infty$.

The following theorem (reworked from [5]) bounds the distance between the distributions of W and the approximating compound Poisson.

Theorem 3.2 ([5])**.** *Let W, λ, μ, δ_1 and δ_4 be as defined above. Then there exist constants $C_1 = C_1(\lambda, \boldsymbol{\mu})$ and $C_2 = C_2(\lambda, \boldsymbol{\mu})$ such that*

$$d_{\mathrm{TV}}(\mathscr{L}(W), \mathrm{CP}(\lambda, \boldsymbol{\mu})) \leq C_1\delta_1 + C_2\delta_4. \tag{3.14}$$

As an immediate corollary, and from the equivalence of the events $\{W = 0\}$ and $\{T \leq t^*\}$, we get that

$$\left|\mathbb{P}\left[T \leq \frac{z+\ln(n)}{\theta}\right] - e^{-\lambda}\right| \leq C_1\delta_1 + C_2\delta_4. \tag{3.15}$$

3.3 Proof of Theorem 3.1

The proof of Theorem 3.1 comes as a result of equation (3.15) along with Lemmas 3.3, 3.4, and 3.5 which bound the various quantities required.

Lemma 3.3. *Using the above setup*

$$\sqrt{\frac{\theta}{4-3\theta}}e^{-z} \leq \lim_{n\to\infty} \lambda \leq e^{-z}. \tag{3.16}$$

Proof.

$$\begin{aligned}
\lambda &= \sum_{i \in V} \mathbb{E}\left[\frac{X_i}{X_i + U_i} I[X_i + U_i \geq 1]\right] && (3.17)\\
&= \sum_{i=1}^{n} \mathbb{P}(X_i = 1)\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] && (3.18)\\
&= \sum_{i=1}^{n} \frac{e^{-z}}{n}\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] && (3.19)\\
&= e^{-z}\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] && (3.20)
\end{aligned}$$

where we have used that X_i is zero-one, (3.5), and the transitivity of the graph. Clearly

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

and so $\lambda \leq e^{-z}$.

By Jensen's inequality

$$\begin{aligned}
\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] &\geq \frac{1}{\mathbb{E}[1+U_i|X_i = 1]} && (3.22)\\
&= \frac{1}{1+\mathbb{E}[U_i|X_i = 1]}. && (3.23)
\end{aligned}$$

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

$$\begin{aligned}
\mathbb{E}[U_i|X_i = 1] &= \sum_{j \in B_i} \mathbb{P}[X_j = 1|X_i = 1] && (3.24)\\
&= \sum_{k=1}^{b_n} \sum_{|j-i|=k} \mathbb{P}[X_j = 1|X_i = 1] && (3.25)\\
&= 2\sum_{k=1}^{b_n} \mathbb{P}[X_{i+k} = 1|X_i = 1] && (3.26)
\end{aligned}$$

where we have used the symmetry of X_{i+k} and X_{i-k} in the last step. From Lemma 3.8,

$$\begin{aligned}
\mathbb{E}[U_i|X_i = 1] &\leq 2\sum_{k=1}^{b_n}\left(\frac{e^{-z}}{n} + \left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)}\right)^k\right) && (3.27)\\
&< 2\sum_{k=1}^{b_n}\frac{e^{-z}}{n} + 2\sum_{k=1}^{\infty}\left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)}\right)^k && (3.28)\\
&= \frac{2b_n}{n}e^{-z} + \sqrt{\frac{4}{\theta}-3} - 1. && (3.29)
\end{aligned}$$

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

$$\lim_{n\to\infty} \lambda \geq \sqrt{\frac{\theta}{4-3\theta}} e^{-z}. \tag{3.30}$$

□

Lemma 3.4.

$$\lim_{n\to\infty} \delta_1 = 0 \tag{3.31}$$

Proof.

$$\delta_1 = \sum_{i=1}^{n} \sum_{k=0}^{2b_n} \mathbb{P}[X_i = 1, U_i = k] \mathbb{E} \left| \frac{\mathbb{P}[X_i = 1, U_i = k | W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1 \right| \tag{3.32}$$

$$= n \sum_{k=0}^{2b_n} \mathbb{E} \left| \mathbb{P}[X_i = 1, U_i = k | W_i] - \mathbb{P}[X_i = 1, U_i = k] \right| \tag{3.33}$$

Let A be the event that the history of a vertex in D_i merges with the history of a vertex in $B_i \cup \{i\}$. More formally,

$$A = \{\exists j \in B_i \cup \{i\}, l \in D_i : \mathcal{H}_j \cap \mathcal{H}_l \neq \emptyset\}. \tag{3.34}$$

We note that if A does not happen, then W_i cannot affect X_i or U_i. That is,

$$\mathbb{P}[X_i = 1, U_i = j | A^{\complement}, W_i] = \mathbb{P}[X_i = 1, U_i = j | A^{\complement}]. \tag{3.35}$$

Continuing on from (3.33), we split the probabilities into

$$\begin{aligned}
\delta_1 &= n \sum_{k=0}^{2b_n} \mathbb{E} \Big| \mathbb{P}[X_i = 1, U_i = k | W_i, A] \mathbb{P}[A|W_i] - \mathbb{P}[X_i = 1, U_i = k | A] \mathbb{P}[A] + & (3.36)\\
&\quad \mathbb{P}(X_i = 1, U_i = k | A^{\complement})(\mathbb{P}[A^{\complement} | W_i] - \mathbb{P}[A^{\complement}]) \Big| \\
&\leq n(2b_n + 1) \mathbb{E} \left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \left| \mathbb{P}[A^{\complement}|W_i] - \mathbb{P}[A^{\complement}] \right| \right] & (3.37)\\
&= n(2b_n + 1) \mathbb{E} \left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + |1 - \mathbb{P}[A|W_i] - (1 - \mathbb{P}[A])| \right] & (3.38)\\
&\leq n(2b_n + 1) \mathbb{E} \left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \mathbb{P}[A|W_i] + \mathbb{P}[A]) \right] & (3.39)\\
&= 2n(2b_n + 1) \left(\mathbb{E}[\mathbb{P}[A|W_i]] + \mathbb{P}[A] \right) & (3.40)\\
&= 4n(2b_n + 1) \mathbb{P}[A] & (3.41)
\end{aligned}$$

Let $A_{j,k}$ denote the event that the histories of vertices j and k merge. That is,

$$A_{j,k} = \{\mathcal{H}_j \cap \mathcal{H}_k \neq \emptyset\}. \tag{3.42}$$

By a union bound,

$$\delta_1 \leq 4n(2b_n+1) \sum_{j \in B_i \cup \{i\}} \sum_{k \in D_i} \mathbb{P}[A_{j,k}] \tag{3.43}$$

$$\leq 8n^2(2b_n+1)^2 \left(\frac{1-\sqrt{\theta(2-\theta)}}{1-\theta} \right)^{c_n - b_n} \tag{3.44}$$

by Lemma 3.6. This goes to 0 as $n \to \infty$. □

Lemma 3.5.

$$\lim_{n\to\infty} \delta_4 = 0 \tag{3.45}$$

Proof.

$$\delta_4 = \sum_{i=1}^{n} (\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]) \tag{3.46}$$

$$= \sum_{i=1}^{n} \mathbb{E}[X_i Z_i] + e^{-z} \sum_{j \in \{i\} \cup B_i \cup C_i} \mathbb{E}[X_j] \tag{3.47}$$

$$= n\mathbb{E}[X_i Z_i] + \frac{2e^{-2z} c_n}{n} \tag{3.48}$$

$$= n\mathbb{P}[X_i = 1]\mathbb{E}[Z_i | X_i = 1] + \frac{2e^{-2z} c_n}{n} \tag{3.49}$$

$$= e^{-z}\mathbb{E}[Z_i | X_i = 1] + \frac{2e^{-2z} c_n}{n} \tag{3.50}$$

Now

$$\mathbb{E}[Z_i | X_i = 1] = \sum_{j \in C_i} \mathbb{P}[X_j = 1 | X_i = 1] \tag{3.51}$$

$$= 2 \sum_{k=b_n+1}^{c_n} \mathbb{P}[X_{i+k} = 1 | X_i = 1] \tag{3.52}$$

From Lemma 3.8,

$$\mathbb{E}[Z_i | X_i = 1] \leq 2 \sum_{k=b_n+1}^{c_n} \left(\frac{e^{-z}}{n} + 2 \left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)} \right)^k \right) \tag{3.53}$$

$$\leq \frac{2(c_n-b_n)e^{-z}}{n} + 4(c_n-b_n) \left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)} \right)^{b_n+1}. \tag{3.54}$$

Altogether

$$\delta_4 \leq \frac{2(c_n-b_n)e^{-z}}{n} + 4(c_n-b_n) \left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)} \right)^{b_n+1} + \frac{2e^{-2z} c_n}{n} \tag{3.55}$$

which goes to 0 as $n \to \infty$. □

3.3.1 Additional Lemmas

Lemma 3.6. *Let $A_{i,j}$ be the event that the update history of vertex i merges with the update history of vertex j. Then*

$$\mathbb{P}[A_{i,j}] \leq 2\left(\frac{1-\sqrt{\theta(2-\theta)}}{1-\theta}\right)^k \tag{3.56}$$

where $k = |i-j|$.

Proof. We note that, while the update histories survive, the distance between the update histories of i and j is a birth and death process that starts at $k = |i-j|$ and has birth and death rates, $\lambda = \mu = 1-\theta$. Let $P(t)$ be such a process and define $s_0 = \inf\{t : P(t) = 0\}$ to be the first time the process reaches zero (this corresponds to the update histories merging). Let s_d be exponentially distributed with rate 2θ (this corresponds to the first time that one of the update histories dies). Then

$$\mathbb{P}[A_{i,j}] \leq 2\mathbb{P}_k(s_0 < s_d) \tag{3.57}$$

where $\mathbb{P}_k$ indicates that $P(0) = k$. The factor of two comes from the fact that the update histories may meet by going the other direction around the cycle.

At any time before s_d there are three possibilities for what can happen to P next. Either the next event is a birth with probability $(1-\theta)/2$, the next event is a death with the same probability or we reach time s_d with probability θ. Writing $\zeta_k = \mathbb{P}_k(s_0 < s_d)$ this gives us the recurrence relation

$$\zeta_k = \frac{1-\theta}{2}\zeta_{k-1} + \frac{1-\theta}{2}\zeta_{k+1} \tag{3.58}$$

which is subject to the conditions

$$\zeta_0 = 1 \tag{3.59}$$

$$\zeta_k \leq 1, \forall k \in \mathbb{N}. \tag{3.60}$$

This recurrence has characteristic equation

$$x^2 - \frac{2}{1-\theta}x + 1 = 0 \tag{3.61}$$

which has roots

$$r_1 = \frac{1+\sqrt{\theta(2-\theta)}}{1-\theta} \tag{3.62}$$

$$r_2 = \frac{1-\sqrt{\theta(2-\theta)}}{1-\theta} \tag{3.63}$$

and so

$$\zeta_k = ar_1^k + br_2^k \tag{3.64}$$

where a and b are constants to be determined from (3.59) and (3.60). We note that $r_1 \geq 1, \forall \theta \in [0,1]$ and so from (3.60) we have that $a = 0$. Finally from (3.59), $b = 1$ and so

$$\zeta_k = \left(\frac{1-\sqrt{\theta(2-\theta)}}{1-\theta} \right)^k . \tag{3.65}$$

□

Lemma 3.7. *Let $A_{i,j}$ be the event that the update history of vertex i merges with the update history of vertex j. Then*

$$\mathbb{P}[A_{i,j}|X_j = 1] \leq 2 \left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)} \right)^k \tag{3.66}$$

where $k = |i-j|$.

Proof. We note that, while the update histories survive, the distance between the update histories of i and j is a birth and death process that starts at $k = |i-j|$ and has birth and death rates, $\lambda = \mu = 1-\theta$. Let $P(t)$ be such a process and define $s_0 = \inf\{t : P(t) = 0\}$ to be the first time the process reaches zero (this corresponds to the update histories merging). Let s_d be exponentially distributed with rate θ (this corresponds to the update history of vertex i dying). Then

$$\mathbb{P}[A_{i,j}|X_j = 1] \leq 2\mathbb{P}_k(s_0 < s_d) \tag{3.67}$$

where $\mathbb{P}_k$ indicates that $P(0) = k$. The factor of two comes from the fact that the update histories may meet by going the other direction around the cycle.

At any time before s_d there are three possibilities for what can happen to P next. Either the next event is a birth with probability $(1-\theta)/(2-\theta)$, the next event is a death with the same probability or we reach time s_d with probability $\theta/(2-\theta)$. Writing

$\zeta_k = \mathbb{P}_k(s_0 < s_d)$ this gives us the recurrence relation

$$\zeta_k = \frac{1-\theta}{2-\theta}\zeta_{k-1} + \frac{1-\theta}{2-\theta}\zeta_{k+1} \tag{3.68}$$

which is subject to the conditions

$$\zeta_0 = 1 \tag{3.69}$$

$$\zeta_k \leq 1, \, \forall k \in \mathbb{N}. \tag{3.70}$$

This recurrence has characteristic equation

$$x^2 - \frac{2-\theta}{1-\theta}x + 1 = 0 \tag{3.71}$$

which has roots

$$r_1 = \frac{2-\theta+\sqrt{4\theta-3\theta^2}}{2(1-\theta)} \tag{3.72}$$

$$r_2 = \frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)} \tag{3.73}$$

and so

$$\zeta_k = ar_1^k + br_2^k \tag{3.74}$$

where a and b are constants to be determined from (3.69) and (3.70). We note that $r_1 \geq 1, \forall\theta \in [0,1]$ and so from (3.70) we have that $a = 0$. Finally from (3.69), $b = 1$ and so

$$\zeta_k = \left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)}\right)^k. \tag{3.75}$$

□

Lemma 3.8.

$$\mathbb{P}[X_{i+k} = 1 | X_i = 1] \leq \frac{e^{-z}}{n} + 2\left(\frac{2-\theta-\sqrt{4\theta-3\theta^2}}{2(1-\theta)}\right)^k. \tag{3.76}$$

Proof. There are two ways in which the update history of vertex $i+k$ 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 $i+k$ can survive long enough to merge with the update history of vertex i (whose surival we are conditioning on). (Note that these events are not mutually exclusive as the update history of vertex $i+k$ could merge with the update history of vertex i after it reaches time 0.) So we have

$$\mathbb{P}[X_{i+k} = 1 | X_i = 1] \leq \mathbb{P}[X_{i+k} = 1] + \mathbb{P}[A_{i,i+k} | X_i = 1]. \tag{3.77}$$

The result follows from (3.5) and Lemma 3.7. □

Chapter 4

General Results

In this chapter we prove the following theorem.

Conjecture 4.1. *Let T_L be the coupling time for the continuous-time Ising heat-bath dynamics for the zero-field ferromagnetic Ising model on the torus $(\mathbb{Z}/L\mathbb{Z})^d$. Then for any small enough inverse-temperature β,*

$$\lim_{L\to\infty} \mathbb{P}[T_L < a_L z + b_L] = e^{-e^{-z}} \tag{4.1}$$

where a_L and b_L have yet to be determined.

Proof. □

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

In practise, we may choose to construct this history as follows: For each $i \in A$, create a temporal edge between (i, t^*) and (i, t_i) where t_i is the time of the latest update to

i (or 0 if i is never updated). Then for each update (i, u, t_i), we either terminate the edge if u is such that the update is oblivious, or we add spatial branches to each of the neighbours of i. We repeat this process recursively for the neighbours of i until every branch has been terminated due to an oblivious update or has reached time 0.

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

$$\mathcal{H}_A(t) \subseteq \hat{\mathcal{H}}_A(t) \tag{4.2}$$

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

4.1.1 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^*.

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

$$m_t(i) = \mathbb{E}[\mathcal{T}_t[i]] \tag{4.3}$$

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

$$\begin{aligned} m_t(i) &= \mathbb{E}[Y_t^+[i]] && (4.4)\\ &= \mathbb{P}[A_t]\left(\mathbb{P}\left[Y_t^+[i] = 1|A_t\right] - \mathbb{P}\left[Y_t^+[i] = -1|A_t\right]\right) && (4.5)\\ &\quad + \mathbb{P}\left[A_t^{\complement}\right]\left(\mathbb{P}\left[Y_t^+[i] = 1|A_t^{\complement}\right] - \mathbb{P}\left[Y_t^+[i] = -1|A_t^{\complement}\right]\right). \end{aligned}$$

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

$$m_t(i) = \mathbb{P}[A_t]. \tag{4.6}$$

Finally, given a target time t^*, X_i is defined such that $\{X_i = 1\} = A_{t^*}$. So $\mathbb{P}[X_i = 1] = m_{t^*}(i)$. This motivates the following restatement of part of Lemma 2.1 from [4].

Lemma 4.2 ([4], Lemma 2.1). *There exist some constant $c_{\beta,d} > 0$ such that for any $t > 0$,*

$$m_t \leq 2e^{-c_{\beta,d}t} \tag{4.7}$$

Corollary 4.3.

$$\mathbb{P}[X_i = 1] \leq 2e^{-c_{\beta,d}t^*} \tag{4.8}$$

4.2 Setup

Define the time

$$t_c(n) = \inf\left\{t > 0 : m_t = \frac{1}{n}\right\}. \tag{4.9}$$

Fix z and a time of interest $t^* = t_c(n) + z$.

Lemma 4.4 ([7], Claim 3.3). *On any graph with maximum degree Δ, for any $t, s > 0$ we have*

$$e^{-2s} \leq \frac{\sum_i m_{t+s}[i]^2}{\sum_i m_t[i]^2} \leq e^{-2(1-\beta\Delta)s}. \tag{4.10}$$

The following corollary is then straightfoward.

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

For $z \geq 0$,

$$\frac{e^{-z}}{n} \leq m_{t^*} \leq \frac{e^{-(1-\beta\Delta)z}}{n}. \tag{4.11}$$

For $z \leq 0$,

$$\frac{e^{-(1-\beta\Delta)z}}{n} \leq m_{t^*} \leq \frac{e^{-z}}{n}. \tag{4.12}$$

4.3 Proof of Theorem 4.1

Lemma 4.6. *Using the above setup*

$$\leq \lim_{n\to\infty} \lambda \leq nm_{t^*} \tag{4.13}$$

Proof.

$$\lambda = \sum_{i \in V} \mathbb{E}\left[\frac{X_i}{X_i + U_i} I[X_i + U_i \geq 1]\right] \tag{4.14}$$

$$= \sum_{i=1}^{n} \mathbb{P}(X_i = 1)\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \tag{4.15}$$

$$= nm_{t^*}\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \tag{4.16}$$

where we have used that X_i is zero-one, (3.5), and the transitivity of the graph. Clearly

$$\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \leq 1 \tag{4.17}$$

and so $\lambda \leq nm_{t^*}$.

By Jensen's inequality

$$\mathbb{E}\left[\frac{1}{1+U_i}|X_i = 1\right] \geq \frac{1}{\mathbb{E}[1+U_i|X_i = 1]} \tag{4.18}$$

$$= \frac{1}{1+\mathbb{E}[U_i|X_i = 1]}. \tag{4.19}$$

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

$$\mathbb{E}[U_i|X_i = 1] = \sum_{j \in B_i} \mathbb{P}[X_j = 1|X_i = 1] \tag{4.20}$$

$$= \sum_{k=1}^{b_n} \sum_{|j-i|=k} \mathbb{P}[X_j = 1|X_i = 1] \tag{4.21}$$

$$\leq \sum_{k=1}^{b_n} \sum_{|j-i|=k} (m_{t^*} + \mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_i \neq \emptyset|X_i = 1]) \tag{4.22}$$

□

Lemma 4.7.

$$\lim_{n\to\infty} \delta_1 = 0 \tag{4.23}$$

Proof.

$$\delta_1 = \sum_{i=1}^{n} \sum_{k=0}^{|B_i|} \mathbb{P}[X_i = 1, U_i = k]\mathbb{E}\left|\frac{\mathbb{P}[X_i = 1, U_i = k|W_i]}{\mathbb{P}[X_i = 1, U_i = k]} - 1\right| \tag{4.24}$$

$$= n\sum_{k=0}^{|B_i|} \mathbb{E}\left|\mathbb{P}[X_i = 1, U_i = k|W_i] - \mathbb{P}[X_i = 1, U_i = k]\right| \tag{4.25}$$

Let A be the event that the history of a vertex in D_i merges with the history of a vertex in $B_i \cup \{i\}$. More formally,

$$A = \{\exists j \in B_i \cup \{i\}, l \in D_i : \mathcal{H}_j \cap \mathcal{H}_l \neq \emptyset\}. \tag{4.26}$$

We note that if A does not happen, then W_i cannot affect X_i or U_i. That is,

$$\mathbb{P}[X_i = 1, U_i = j | A^{\complement}, W_i] = \mathbb{P}[X_i = 1, U_i = j | A^{\complement}]. \tag{4.27}$$

Continuing on from (4.25), we split the probabilities into

$$\begin{aligned}
\delta_1 &= n \sum_{k=0}^{|B_i|} \mathbb{E} \Big| \mathbb{P}[X_i = 1, U_i = k | W_i, A] \mathbb{P}[A|W_i] - \mathbb{P}[X_i = 1, U_i = k | A] \mathbb{P}[A] + \qquad (4.28)\\
&\quad \mathbb{P}(X_i = 1, U_i = k | A^{\complement})(\mathbb{P}[A^{\complement}|W_i] - \mathbb{P}[A^{\complement}]) \Big| \\
&\leq n(|B_i| + 1) \mathbb{E} \left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \left| \mathbb{P}[A^{\complement}|W_i] - \mathbb{P}[A^{\complement}] \right| \right] \qquad (4.29)\\
&= n(|B_i| + 1) \mathbb{E} \left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + |1 - \mathbb{P}[A|W_i] - (1 - \mathbb{P}[A])| \right] \qquad (4.30)\\
&\leq n(|B_i| + 1) \mathbb{E} \left[\mathbb{P}[A|W_i] + \mathbb{P}[A] + \mathbb{P}[A|W_i] + \mathbb{P}[A]) \right] \qquad (4.31)\\
&= 2n(|B_i| + 1) \left(\mathbb{E}[\mathbb{P}[A|W_i]] + \mathbb{P}[A] \right) \qquad (4.32)\\
&= 4n(|B_i| + 1) \mathbb{P}[A] \qquad (4.33)
\end{aligned}$$

By a union bound,

$$\delta_1 \leq 4n(|B_i| + 1) \sum_{j \in B_i \cup \{i\}} \sum_{k \in D_i} \mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_k \neq \emptyset]. \tag{4.34}$$

From Lemma 4.9,

$$\delta_1 \leq 4n^2 (|B_i| + 1)^2 \exp(2\alpha) \exp(-\lambda(c_n - b_n)). \tag{4.35}$$

On the torus $(\mathbb{Z}/L\mathbb{Z})^d$, $n = L^d$, $c_n = \log(L)^2$, $b_n = \log(L)$, and $|B_i| \leq C \log(L)^d$. So

$$\delta_1 \leq C L^{2d} \log(L)^{2d} \exp(-\lambda(\log(L)^2 - \log(L)) \tag{4.36}$$

which goes to 0 as $L \to \infty$. □

Lemma 4.8.

$$\lim_{n \to \infty} \delta_4 = 0 \tag{4.37}$$

Proof.

$$\begin{aligned}
\delta_4 &= \sum_{i=1}^{n} \left(\mathbb{E}[X_i Z_i] + \mathbb{E}[X_i]\mathbb{E}[X_i + U_i + Z_i]\right) && (4.38)\\
&= n\mathbb{E}[X_i Z_i] + nm_{t^*}^2 \left(1 + |B_i| + |C_i|\right) && (4.39)\\
&= nm_{t^*}\mathbb{E}[Z_i | X_i = 1] + nm_{t^*}^2 \left(1 + |B_i| + |C_i|\right) && (4.40)\\
&= nm_{t^*} \left(\sum_{j \in C_i} \mathbb{P}[X_j = 1 | X_i = 1] + m_{t^*} \left(1 + |B_i| + |C_i|\right) \right) && (4.41)\\
&\leq nm_{t^*} \left(\sum_{j \in C_i} (m_{t^*} + \mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_i \neq \emptyset | X_i = 1) + m_{t^*} \left(1 + |B_i| + |C_i|\right) \right) && (4.42)
\end{aligned}$$

NEED TO PROVE LEMMA 4.10 □

4.3.1 Additional Lemmas

Lemma 4.9.

$$\mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_k \neq \emptyset] \leq \exp(2\alpha)\exp(-\lambda|j - k|). \tag{4.43}$$

Proof. The following is based on the proof used in Section 3.2 of [8].

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

$$\mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_k \neq \emptyset] \leq \mathbb{P}[\hat{\mathcal{H}}_j \cap \hat{\mathcal{H}}_k \neq \emptyset]. \tag{4.44}$$

Let $W_s = |\hat{\mathcal{H}}_{\{j,k\}}(t^* - s)|$ and let $Y_s = \# \left\{ ((u,t),(v,t)) \in \hat{\mathcal{H}}_{\{j,k\}} : t \in [t^* - s, t^*) \right\}$ count the total number of spatial edges observed in the history by time $t^* - s$. For $\hat{\mathcal{H}}_j$ and $\hat{\mathcal{H}}_k$ to intersect at time $t^* - s$, there must be enough spatial edges for the histories to reach each other. That is, we require that

$$Y_s \geq |i - j|. \tag{4.45}$$

Initially, $W_0 = 2$ and $Y_0 = 0$. 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. Therefore we can create a coupled process $(\bar{W}_s, \bar{Y}_s)$ such that $\bar{W}_s \geq W_s$ and $\bar{Y}_s \geq Y_s$ in the following way. We start with $(\bar{W}_s, \bar{Y}_s) = (2, 0)$

and at rate $\theta\bar{W}_s$, $\bar{W}_s$ decreases by 1, and at rate $(1-\theta)\bar{W}_s$, both $\bar{W}_s$ and $\bar{Y}_s$ increase by Δ.

Let $Q_s = \exp\left(\alpha\bar{W}_s + \lambda\bar{Y}_s\right)$ where α and λ are some fixed constants yet to be determined. We will show that Q_s is a supermartingale. Let h be some small time-step. Then

$$\begin{aligned} \mathbb{E}\left[Q_{s_0+h} - Q_{s_0} | Q_{s_0}\right] &= h\theta\bar{W}_{s_0}\left(\exp(\alpha(\bar{W}_{s_0}-1)+\lambda\bar{Y}_{s_0}) - \exp(\alpha\bar{W}_{s_0}+\lambda\bar{Y}_{s_0})\right) + \quad (4.46) \\ &\quad h(1-\theta)\bar{W}_{s_0}\left(\exp(\alpha(\bar{W}_{s_0}+\Delta)+\lambda(\bar{Y}_{s_0}+\Delta)) - \exp(\alpha\bar{W}_{s_0}+\lambda\bar{Y}_{s_0})\right) + \\ &\quad \mathcal{O}(h^2) \\ &= \Big(\theta(e^{-\alpha}-1) + (1-\theta)(e^{(\alpha+\lambda)\Delta}-1)\Big)\, h\bar{W}_{s_0}Q_{s_0} + \mathcal{O}(h^2). \quad (4.47) \end{aligned}$$

Dividing through by h and taking h to 0, we have

$$\frac{d}{ds}\mathbb{E}\left[Q_s|Q_{s_0}\right]\Bigg|_{s=s_0} = \Big(\theta(e^{-\alpha}-1) + (1-\theta)(e^{(\alpha+\lambda)\Delta}-1)\Big)\,\bar{W}_{s_0}Q_{s_0} \quad (4.48)$$

which is negative for θ sufficiently close to 1. Hence Q_s is a supermartingale for sufficiently large θ.

Define the stopping time

$$\tau = \inf\{s : \bar{W}_s = 0\}. \quad (4.49)$$

Since $(\bar{W}_s, \bar{Y}_s) \geq (W_s, Y_s)$, we have that $W_\tau = 0$ which corresponds to the event that both histories, $\mathcal{H}_j$ and $\mathcal{H}_k$, have terminated by time τ. Hence

$$\mathbb{P}[\hat{\mathcal{H}}_j \cap \hat{\mathcal{H}}_k \neq \emptyset] \leq \mathbb{P}[Y_\tau \geq |j-k|]. \quad (4.50)$$

Now

$$\mathbb{E}[\exp(\lambda Y_\tau)] \leq \mathbb{E}[\exp(\lambda\bar{Y}_\tau)]. \quad (4.51)$$

and from optional stopping [FIGURE THIS OUT PROPERLY],

$$\begin{aligned} \mathbb{E}[\exp(\lambda\bar{Y}_\tau)] &\leq \mathbb{E}[Q_0] \quad (4.52) \\ &= \exp(2\alpha). \quad (4.53) \end{aligned}$$

Hence

$$\mathbb{P}[Y_\tau \geq |j-k|] \leq \exp(2\alpha)\exp(-\lambda|j-k|) \quad (4.54)$$

and overall we have that

$$\mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_k \neq \emptyset] \leq \exp(2\alpha)\exp(-\lambda|j-k|). \quad (4.55)$$

□

Lemma 4.10.

$$\mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_k \neq \emptyset | X_j = 1] \leq \tag{4.56}$$

Proof. The proof here is similar to that of Lemma 4.9. We first relax to our alternative histories

$$\mathbb{P}[\mathcal{H}_j \cap \mathcal{H}_k \neq \emptyset | X_j = 1] \leq \mathbb{P}[\hat{\mathcal{H}}_j \cap \hat{\mathcal{H}}_k \neq \emptyset | X_j = 1]. \tag{4.57}$$

By symmetry,

$$\mathbb{P}[\hat{\mathcal{H}}_j \cap \hat{\mathcal{H}}_k \neq \emptyset | X_j = 1] = \mathbb{P}[\hat{\mathcal{H}}_j \cap \hat{\mathcal{H}}_k \neq \emptyset | X_k = 1] \tag{4.58}$$

and also

$$\mathbb{P}[X_j = 1] = \mathbb{P}[X_k = 1]. \tag{4.59}$$

So

$$\mathbb{P}[\hat{\mathcal{H}}_j \cap \hat{\mathcal{H}}_k \neq \emptyset | X_j = 1] \leq 2\mathbb{P}[\hat{\mathcal{H}}_j \cap \hat{\mathcal{H}}_k \neq \emptyset | \{X_j = 1\} \cup \{X_k = 1\}]. \tag{4.60}$$

We now define W_s and Y_s as before, except now conditioned on the event $\{X_j = 1\} \cup \{X_k = 1\}$. The effect of this conditioning is to forbid updates that reduce W_s to 0.

HANDWAVING: If we stop when $W_s = 1$ then we can ignore the conditioning???

Define the stopping time

$$\tau = \inf\{s : \bar{W}_s = 1\}. \tag{4.61}$$

□

Chapter 5

Conclusion

Part II

Efficient Optimization for Statistical Inference

Chapter 6

Maximum Likelihood Location Mixtures

6.1 Introduction

6.1.1 Definitions

A location mixture on $\mathbb{R}$ with mixing distribution Q and component density $f(x)$ can be written as

$$f_Q(x) = \int_{-\infty}^{\infty} f(x-\theta)\,\mathrm{d}Q(\theta). \tag{6.1}$$

Given a sample $\boldsymbol{x} = (x_1, \ldots, x_n)$, we wish to find a distribution that maximizes the log likelihood

$$L(Q; \boldsymbol{x}) = \sum_{i=1}^{n} \log(f_Q(x_i)). \tag{6.2}$$

Lindsay showed that under quite general conditions, such a maximizing distribution exists and has no more than n points of support [9]. It is therefore common to use

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

instead of (6.1) as our definition of a location mixture. We note that (6.1) is equivalent to (6.3) in the case where Q is a discrete distribution which places probability masses of weight p_j at locations θ_j, for $j = 1, \ldots, m$. The order of the x_i does not matter and so we will assume without loss of generality that $x_1 \leq x_2 \leq \cdots \leq x_n$ throughout.

In this paper we are primarily interested in the number of probability masses that are required in the maximizing mixture. We will call this quantity $K_{\boldsymbol{x}}$. It should be noted that there is not always a unique mixing distribution that maximizes (6.2). However we can choose $K_{\boldsymbol{x}}$ to be the smallest number of probability masses that any of the maximizing distributions have.

6.1.2 Flag graphs

One point which we wish to emphasize is that, given a component density f, K is a function of $\boldsymbol{x}$ only. We can visualise this for small values of n. In Figure 6.1, we empirically found the MLE for $\boldsymbol{x} = (0, x_2, x_3)$ and recorded the number of probability masses in the maximizing mixture. We took values for x_2 and x_3 from an evenly spaced grid with $-6 \leq x_2, x_3 \leq 6$. We fixed $x_1 = 0$ since the shape of the maximum likelihood mixture (and therefore $K_{\boldsymbol{x}}$) depends only on the relative location of the x_i, not their absolute location.

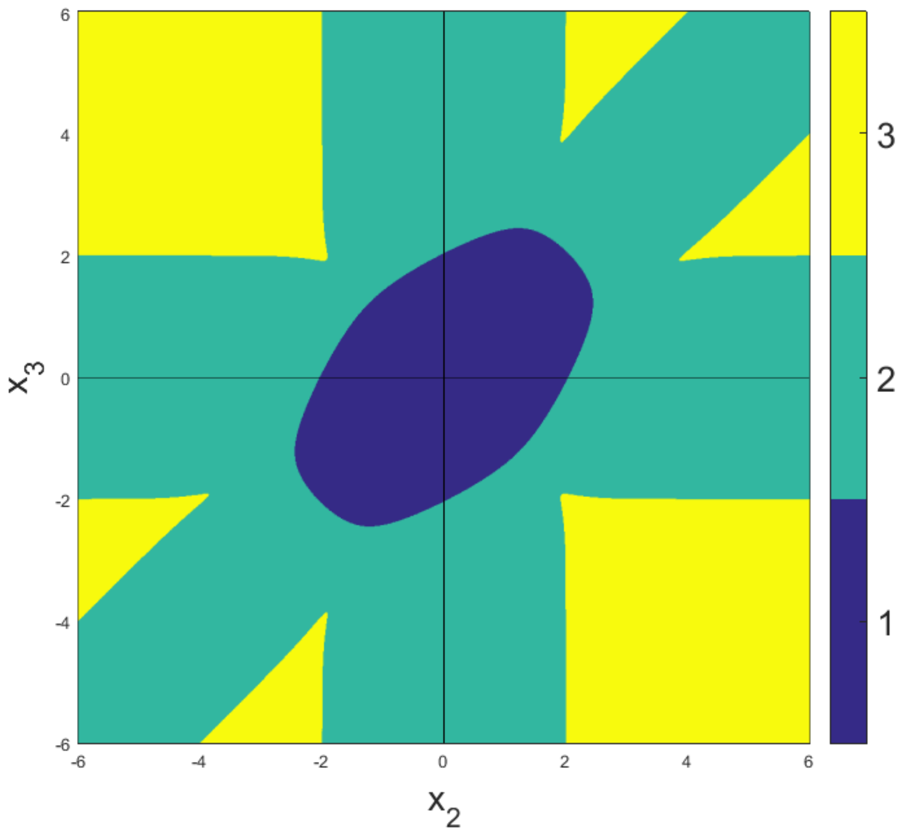


FIGURE 6.1: K as a function of x_2 and x_3 (with $x_1 = 0$) for a normal component density with unit variance.

For a particular choice of component density, we can partition $\mathbb{R}^n$ into sets

$$C_k = \{\boldsymbol{x} \in \mathbb{R}^n | K_{\boldsymbol{x}} = k\}, \qquad k = 1, \ldots, n. \tag{6.4}$$

The problem of determining $K_{\boldsymbol{x}}$ is then the same as determining in which of the C_k $\boldsymbol{x}$ lies.

6.1.3 Motivating Example

6.2 Results for $n = 2$

In this section, we present Theorem 6.4 which expands upon the results found in [9] and [10].

6.2.1 Things that are referenced

REWORK CONTENTS OF THIS SUBSECTION INTO MAIN TEXT.

$$M(\boldsymbol{u}) = \sum_{i=1}^{n} \log(u_i) \tag{6.5}$$

Theorem 6.1. *If* Γ *is compact then there exists a unique point on the boundary of* $\mathrm{Conv}(\Gamma)$ *which maximizes the likelihood. This point corresponds to a distribution* Q *which maximizes the likelihood and that has no more than* n *point masses.*

6.2.2 The likelihood curve

In [9], the problem of mixture likelihoods was looked at from a geometrical perspective. One key construction introduced by Lindsay was the *likelihood curve*,

$$\boldsymbol{\gamma}(\theta; \boldsymbol{x}) = (f(x_1 - \theta), \dots, f(x_n - \theta)) \tag{6.6}$$

and it's trace,

$$\Gamma_{\boldsymbol{x}} = \{\boldsymbol{\gamma}(\theta; \boldsymbol{x}) | \theta \in \mathbb{R}\}. \tag{6.7}$$

A useful property of the likelihood curve is that any convex combination of elements from $\Gamma_{\boldsymbol{x}}$ can be written as

$$\boldsymbol{u}(\boldsymbol{p}, \boldsymbol{\theta}; \boldsymbol{x}) = (f_{\boldsymbol{p},\boldsymbol{\theta}}(x_1), \dots, f_{\boldsymbol{p},\boldsymbol{\theta}}(x_n)) = \sum_{j=1}^{m} p_j \boldsymbol{\gamma}(\theta_j; \boldsymbol{x}), \qquad \sum_{j=1}^{m} p_j = 1 \tag{6.8}$$

where $f_{\boldsymbol{p},\boldsymbol{\theta}}(x)$ is as defined in (6.3). The log likelihood of the corresponding distribution is simply the sum of the log of the components of $\boldsymbol{u}(\boldsymbol{p}, \boldsymbol{\theta}; \boldsymbol{x})$.

One of Lindsay's main results, which follows from this observation, was that if

$$\hat{\boldsymbol{u}} = \underset{\boldsymbol{u} \in \mathrm{Conv}(\Gamma_{\boldsymbol{x}})}{\arg\max} \sum_{i=1}^{n} \log(u_i) \tag{6.9}$$

then we can write

$$\hat{\boldsymbol{u}} = \boldsymbol{u}(\boldsymbol{p}, \boldsymbol{\theta}; \boldsymbol{x}) \tag{6.10}$$

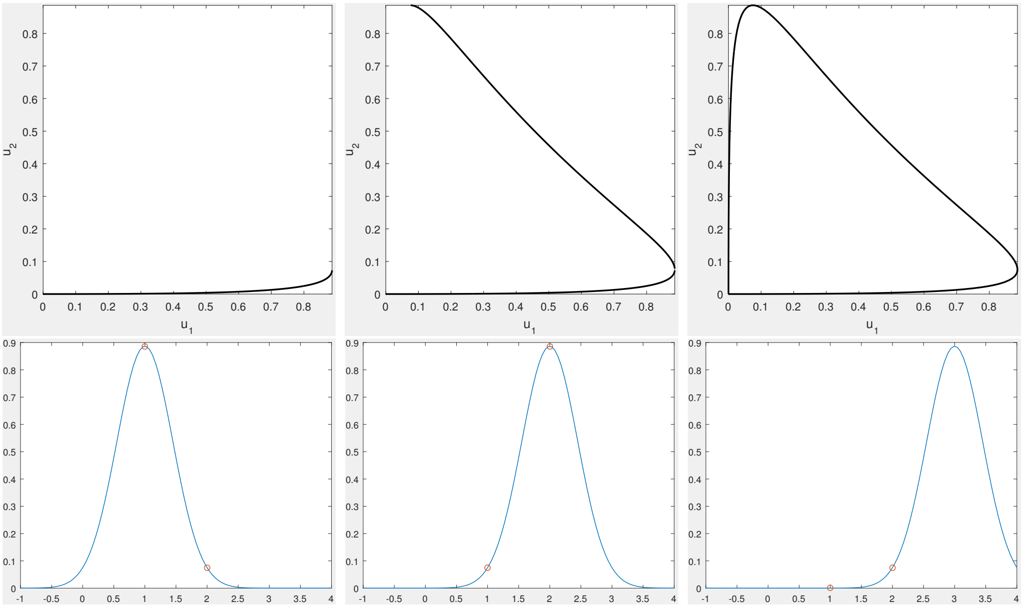
for some $\boldsymbol{p}$ and $\boldsymbol{\theta}$ whose dimension is no more than n. Furthermore, the corresponding distribution that places masses p_j at locations θ_j maximizes (6.2). There are some minor conditions on this result, but they will not cause any problems for our purposes and so will not be discussed (see [9] for details).

6.2.2.1 An example

The following example illustrates the geometrical approach given above. We consider the case where $n = 2$. Our sample is made up of two points, $X_1 = 1$ and $X_2 = 2$. We define

$$f_\theta(x) = \frac{1}{0.45\sqrt{2\pi}} \exp\left(-\frac{(x-\theta)^2}{2 \cdot 0.45^2}\right) \tag{6.11}$$

(i.e. a normal density with mean θ and variance $\sigma^2 = 0.45^2$) to be our component density. We trace out the set Γ in Figure 6.2.

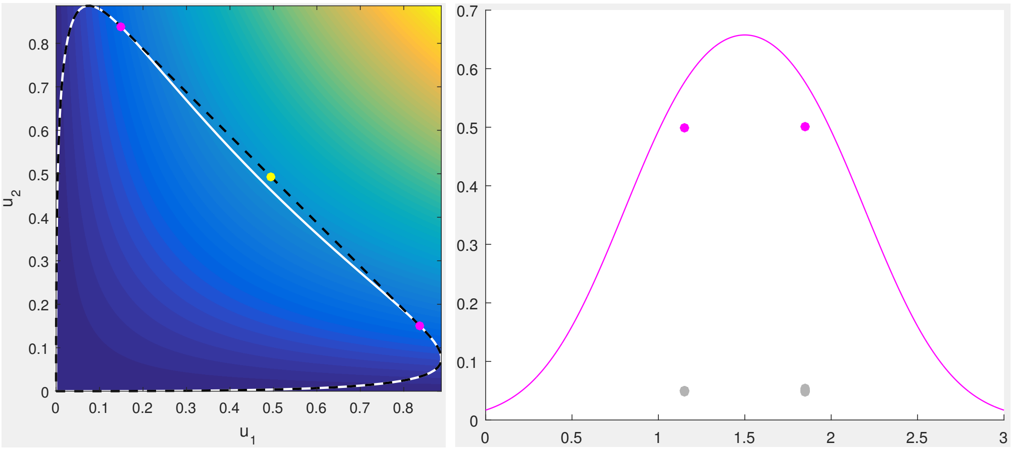


(A) $\theta = 1$ (B) $\theta = 2$ (C) $\theta = 3$

FIGURE 6.2: The blue density is f_θ for $\theta = 1, 2, 3$. Each value of θ contributes a point to Γ whose coordinates are given by $(f_\theta(X_1), f_\theta(X_2))$ (represented by the red circles). As we increase θ from $-\infty$ to ∞ we trace out more of Γ (shown above).

Note that while Γ is bounded, it is not closed (it does not contain the limit point $(0, 0)$), and so Γ is not compact (as required by Theorem 6.1). 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 Theorem 6.1 if $\Gamma \cup \{\mathbf{0}\}$ is compact.

We trace the boundary of $\mathrm{Conv}(\Gamma)$ in Figure 6.3 along with a heat map of the objective function (6.5). The optimal point is on the boundary of $\mathrm{Conv}(\Gamma)$ as expected and it can be written as the $p_1\boldsymbol{f}(\theta_1) + p_2\boldsymbol{f}(\theta_2)$ (where $p_1 + p_2 = 1$). These two points correspond to the two probability masses in the maximizing mixture distribution shown in Figure 6.3b. These masses are located at θ_1 and θ_2 with weights p_1 and p_2.

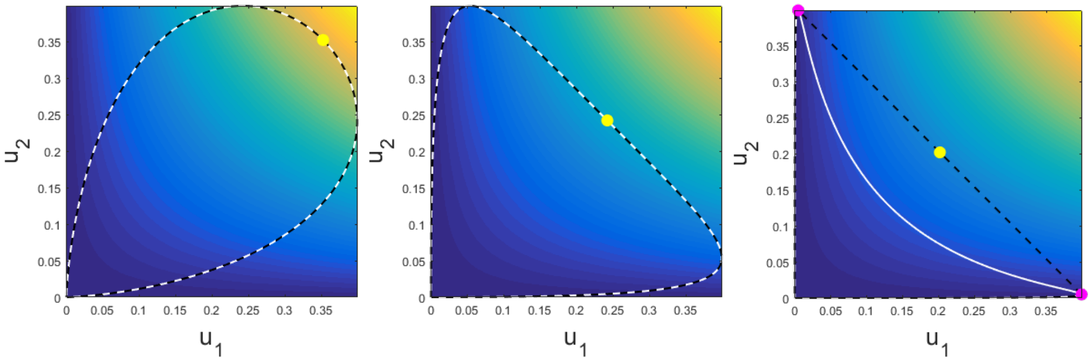


(A) (B)

FIGURE 6.3: In (a), the boundary of $\mathrm{Conv}(\Gamma)$ is shown as a dashed black line, Γ is the white curve, the heat map shows the objective function (likelihood increases from blue to yellow) and the yellow point is the maximizing point which can be written as the convex combination of the two magenta points. These two magenta points correspond to the two probability masses in the maximizing mixing distribution (b).

6.2.3 The likelihood curve for $n = 2$

The shape of $\Gamma_{\boldsymbol{x}}$ can provide us with some insight into the behaviour of $K_{\boldsymbol{x}}$. In Figure 6.4, we give some examples of $\Gamma_{\boldsymbol{x}}$ for $n = 2$ using a normal component density with variance $\sigma^2 = 1$. In particular, we note that the distance between x_1 and x_2 has a strong effect on the shape of $\Gamma_{\boldsymbol{x}}$. In Figure 6.4a, 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 6.4c, 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 6.4c, is shown in Figure 6.4b.



(A) $\boldsymbol{x} = (0, 1)$ (B) $\boldsymbol{x} = (0, 2)$ (C) $\boldsymbol{x} = (0, 3)$

FIGURE 6.4: The curve $\Gamma_{\boldsymbol{x}}$ for three different $\boldsymbol{x}$ along with the boundary of Conv$(\Gamma_{\boldsymbol{x}})$. The objective function from (6.9) is represented as a heat map. The optimal point $\hat{\boldsymbol{u}}$ is shown in yellow, and where applicable, the points $\boldsymbol{\gamma}(\theta_j)$ that make up $\hat{\boldsymbol{u}}$ (as in (6.8)) are shown in magenta.

Obtaining results about where these boundaries lie is very difficult in higher dimensions. In [10], Lindsay used the sign of the curvature of $\boldsymbol{\gamma}(\theta; \boldsymbol{x})$ to obtain results for $n = 2$ when the component density is in the exponential family. Here we will present an extension to Lindsay's results by considering densities that satisfy the following assumptions.

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

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

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

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

A5. f has only two points of inflection

Definition 6.2. If f satisfies assumptions A1 through to A3, then define $[i^-, i^+]$ to be the largest interval that contains 0 and on which $f''(x) \leq 0$.

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

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

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

Lemma 6.3. *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*

$$-f'(x_1 - \theta) = f'(x_2 - \theta) \tag{6.12}$$

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 $f'(x)$:

$$f'(x) \text{ is } \begin{cases} \text{positive and increasing,} & x \in (-\infty, -i) \\ \text{positive and decreasing,} & x \in (-i, 0) \\ \text{negative and decreasing,} & x \in (0, i) \\ \text{negative and increasing,} & x \in (i, \infty). \end{cases} \tag{6.13}$$

We also note, from A4, that $f'(x)$ is an odd function. Using this and rearranging (6.12) we obtain the equivalent equation

$$g(\theta) = h(\theta) \tag{6.14}$$

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 (6.14) 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 (6.14) on this interval.

Similarly by symmetry of f, for $\theta \in (0, -i + x_2 - x_1)$, $f'(\theta) > -f'(\theta - 2i) > -f'(\theta - x_2 - x_1)$ and there are no solutions to (6.14) 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, (6.12) must have only one solution. □

Theorem 6.4. *Let $f(x)$ satisfy assumptions A1 through to A6. Let $\boldsymbol{x} = (x_1, x_2)$ be the sample for which we a finding a maximum likelihood mixture using f as the component density. Then $K_{\boldsymbol{x}} = 1$ if and only if*

$$x_2 - x_1 \leq 2i \tag{6.15}$$

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

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

$$S(\theta) = \begin{vmatrix} \gamma_1'(\theta; \boldsymbol{x}) & \gamma_1''(\theta; \boldsymbol{x}) \\ \gamma_2'(\theta; \boldsymbol{x}) & \gamma_2''(\theta; \boldsymbol{x}) \end{vmatrix} = \begin{vmatrix} -f'(x_1 - \theta) & f''(x_1 - \theta) \\ -f'(x_2 - \theta) & f''(x_2 - \theta) \end{vmatrix}. \tag{6.16}$$

and so

$$S\left(\frac{x_1 + x_2}{2}\right) = \begin{vmatrix} -f'(\frac{x_1 - x_2}{2}) & f''(\frac{x_1 - x_2}{2}) \\ -f'(\frac{x_2 - x_1}{2}) & f''(\frac{x_2 - x_1}{2}) \end{vmatrix}. \tag{6.17}$$

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

Now we complete the if direction. Assume that $x_2 - x_1 \leq 2i$. By Lemma 6.3, there is only one point at which the curve is pointing perpendicular to the line $u_1 = u_2$ SAY THIS BETTER. By the symmetry of f this occurs when $\gamma(\theta; \boldsymbol{x})$ is crossing the line $u_1 = u_2$. Since f is continuous, the direction that $\gamma(\theta; \boldsymbol{x})$ is moving is also continuous. At $\theta = x_1$, $\gamma(\theta; \boldsymbol{x})$ is pointing straight up and so we have that for $\theta \in [x_1, (x_1 + x_2)/2]$, $\gamma(\theta; \boldsymbol{x})$ is travelling in a direction pointing above the line perpendicular to $u_1 = u_2$. For $\theta \in [(x_1 + x_2)/2, x_2]$, $\gamma(\theta; \boldsymbol{x})$ points below the line. It is now obvious that $\gamma((x_1 + x_2)/2; \boldsymbol{x})$ is the furthest point from the origin that lies on $u_1 = u_2$ and is in the convex hull of $\Gamma_{\boldsymbol{x}}$. Since the likelihood increases as we move away from the origin along the line $u_1 = u_2$ in the positive quadrant, we must have

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

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

[1] This is obvious but may need a lemma

□

6.3 Results for general n

6.3.1 Directional Derivative

One of the tools introduced in [9] was the function

$$D(\theta; \boldsymbol{p}, \boldsymbol{\theta}, \boldsymbol{x}) = -n + \sum_{i=1}^{n} \frac{f(x_i - \theta)}{\sum_{j=1}^{m} p_j f(x_i - \theta_j)}. \tag{6.18}$$

Lindsay showed that if $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{p}}$ form a maximum likelihood location mixture of f for $\boldsymbol{x}$ then (under appropriate differentiability assumuptions) the function D satisfies the following:

$$D(\theta_k; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}, \boldsymbol{x}) = 0, \qquad k = 1, \ldots, m, \tag{6.19}$$

$$D'(\theta_k; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}, \boldsymbol{x}) = 0, \qquad k = 1, \ldots, m, \tag{6.20}$$

$$D''(\theta_k; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}, \boldsymbol{x}) \leq 0, \qquad k = 1, \ldots, m. \tag{6.21}$$

These three constraints restrict what a potential maximum likelihood solution can look like. HOW DID LINDSAY USE THEM VS US?

6.3.2 Normal Constraints

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

$$f(x; \sigma) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2}, \tag{6.22}$$

equations (6.19) to (6.21) become

$$\frac{1}{n}\sum_{i=1}^{n} \Gamma_k(x_i; \boldsymbol{p}, \boldsymbol{\theta}) = 1 \tag{6.23}$$

$$\frac{1}{n}\sum_{i=1}^{n} x_i \Gamma_k(x_i; \boldsymbol{p}, \boldsymbol{\theta}) = \theta_k \tag{6.24}$$

$$\frac{1}{n}\sum_{i=1}^{n} (x_i - \theta_k)^2 \Gamma_k(x_i; \boldsymbol{p}, \boldsymbol{\theta}) \leq \sigma^2 \tag{6.25}$$

where we have written

$$\Gamma_k(x; \boldsymbol{p}, \boldsymbol{\theta}) = \frac{f(x-\theta_k;\sigma)}{\sum_{j=1}^m p_j f(x-\theta_j;\sigma)}. \tag{6.26}$$

for ease of notation. Using these constraints, we will bound the regions $C_1, \ldots, C_n$ in Theorem 6.10. However, as a gentle introduction we will start with the much simpler problem of just bounding C_1.

Theorem 6.5. *Write $\bar{\boldsymbol{x}}$ for the mean of $\boldsymbol{x}$. If $\boldsymbol{x} \in C_1$ then*

$$\frac{1}{n}\sum_{i=1}^n (x_i - \bar{\boldsymbol{x}})^2 \leq \sigma^2. \tag{6.27}$$

Proof. If $\boldsymbol{x} \in C_1$ then the maximizing mixture has one component and so $\Gamma_1(x; \boldsymbol{\theta}, \boldsymbol{p}) = 1$. Then (6.24) gives us that $\theta_1 = \bar{\boldsymbol{x}}$ and combining this with (6.25) completes the proof. □

6.3.2.1 Treating $\boldsymbol{x}$ as random

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

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

for $i = 1, \ldots, n$. Our component density, f_θ, is normal with variance σ_2^2. From Theorem 6.5,

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

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

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

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

Remark 6.6. Of particular interest is the case where $\sigma_1 = \sigma_2$. In this case, $p_u \to 1/2$ as $n \to \infty$. While not a new result [CITE SOMETHING HERE], this tells us that the maximum likelihood estimator is not a consistent estimator for the number of components.

6.3.3 Properties of Γ

In order to bound regions where $m \geq 2$, we will need to get a handle on $\Gamma_k(x; \boldsymbol{\theta}, \boldsymbol{p})$. In this section, we list and prove some properties that will be required in Section 6.3.4.

Lemma 6.7.

$$\max_k \left(\Gamma_k(x; \boldsymbol{p}, \boldsymbol{\theta})\right) \geq 1 \tag{6.28}$$

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

$$\Gamma_{k_0}(x; \boldsymbol{p}, \boldsymbol{\theta}) = \frac{f(x-\theta_{k_0}; \sigma)}{\sum_{j=1}^m p_j f(x-\theta_j; \sigma)} \geq \frac{f(x-k_0; \sigma)}{\sum_{j=1}^m p_j f(x-\theta_{k_0}; \sigma)} = 1. \tag{6.29}$$

□

Lemma 6.8.

$$\Gamma_k(x; \boldsymbol{p}, \boldsymbol{\theta}) \leq \frac{1}{p_k} \tag{6.30}$$

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

$$\Gamma_k(x; \boldsymbol{p}, \boldsymbol{\theta}) = \frac{f(x-\theta_k; \sigma)}{\sum_{j=1}^m p_j f(x-\theta_j; \sigma)} \leq \frac{f(x-\theta_k; \sigma)}{p_k f(x-\theta_k; \sigma)} = \frac{1}{p_k}. \tag{6.31}$$

□

Lemma 6.9. *Let $\gamma(x)$ be a non-negative function that satisfies*

$$\frac{1}{n}\sum_{i=1}^n \gamma(x_i) = 1. \tag{6.32}$$

Then the θ that minimizes

$$\frac{1}{n}\sum_{i=1}^n (x_i-\theta)^2 \gamma(x_i) \tag{6.33}$$

is

$$\theta = \frac{1}{n}\sum_{i=1}^n x_i \gamma(x_i). \tag{6.34}$$

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

6.3.4 Bounding C_m

Theorem 6.10. *THM AND PROOF IS BELOW*

Let $\boldsymbol{x} = (x_1, \ldots, x_n)$ and assume that the maximum likelihood mixture for $\boldsymbol{x}$ has no more than m components. Let $\hat{\boldsymbol{\theta}}$ and $\hat{\boldsymbol{p}}$ denote this maximizing mixture. Then by Lemma 6.7, there exists a k^* such that $\Gamma_{k^*}(x_i; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}) \geq 1$ for at least $\lceil \frac{n}{m} \rceil$ different x_i. Let A_{k^*} be the set of all these x_i. Let $A_{\theta_{k^*}}$ be the set of the $|A_{k^*}|$ closest x_i to θ_{k^*}. Then

$$\frac{1}{n}\sum_{i=1}^{n}(x_i - \theta_{k^*})^2 \Gamma_{k^*}(x_i; \hat{\boldsymbol{p}}, \hat{\boldsymbol{\theta}}) \geq \frac{1}{n}\sum_{i \in A_{\theta_{k^*}}}(x_i - \theta_{k^*})^2 \tag{6.35}$$

$$\geq \frac{1}{n}\left\lceil \frac{n}{m} \right\rceil \mathrm{Var}\left(A_{\theta_{k^*}}\right) \qquad \text{(Biased Variance)} \tag{6.36}$$

From (6.25),

$$\mathrm{Var}\left(A_{\theta_{k^*}}\right) \leq \frac{n\sigma^2}{\lceil \frac{n}{m} \rceil}. \tag{6.37}$$

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

6.3.5 A particular class of optimization problem

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

Theorem 6.11. *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 $\boldsymbol{x} \notin \partial E_m$. □

6.3.6 Derive Constraints again

WE SHOULD BE ABLE TO DERIVE (6.23) THROUGH (6.25) AGAIN USING THEOREM 6.11.

6.4 General Results

6.4.1 All points separated by α

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

Theorem 6.12. *If our component density is unimodal and symmetric about zero, and*

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

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

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

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

and so

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

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

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

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

6.4.2 Discussion about what we hope to acheive

The few original results above (Theorems 6.4 and 6.10) seem to be special cases of what looks to be a much more general rule. Theorem 6.10 seems to be too large by a factor of m when you compare to numerics, and the distance between inflection points in Theorem 6.4 seems to come up again when when you look at images like Figure 6.1 (eg the thickness of the 'bands' is this distance). It is therefore our hope that we can either generalize or add significantly to the Theorems stated so far.

Chapter 7

Deconvolution

7.1 Introduction

From here to the end of Section 7.1.2 is a summary of [11]. We want to find the distribution of a random variable X but only measure

$$W = X + U$$

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

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

for the phase function of X. Then

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

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

$$\rho_Y = \rho_W.$$

7.1.1 Optimization problem

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

$$T(Y) = \int_{-\infty}^{\infty} \left| \hat{\phi}_W(t) - |\hat{\phi}_W(t)|\rho_Y(t) \right|^2 w(t)\,\mathrm{d}t \tag{7.1}$$

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

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

7.1.2 Kernel Smoothing

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

$$\hat{f}_Y(x) = \frac{1}{2\pi} \int e^{-itx} \phi_Y(t) \phi_K(ht)\,\mathrm{d}t \tag{7.2}$$

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

$$\hat{f}_Y(x) = \sum_{j=1}^{m} p_j K_h(x - x_j). \tag{7.3}$$

However, we can get a better result by using (7.2) and replacing $\phi_Y(t)$ with an appropriate ridge function for $t \geq t^*$.

7.2 Examples and Relation to Mixture Phenomenon

7.3 R Package

Bibliography

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

[2] James Gary Propp and David Bruce Wilson. Exact sampling with coupled markov chains and applications to statistical mechanics. 1996.

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

[4] Eyal Lubetzky and Allan Sly. Information percolation and cutoff for the stochastic ising model. *J. Amer. Math. Soc.*, 2016.

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

[6] 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.

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

[8] Eyal Lubetzky and Allan Sly. An exposition to information percolation for the ising model. December 2014.

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

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

[11] 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.