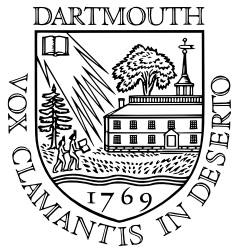
Dartmouth College
College of Arts and Sciences
Department of Mathematics

Bachelor's Thesis

STATISTICAL ESTIMATION OF ISING GRAPHICAL MODELS

FROM THEORY TO PRACTICE

by James Brofos



Advisor: Prof. Eugene Demidenko
June 2015

This thesis is submitted in partial fulfillment of the requirements for the Bachelor of Arts in Mathematics.

COMMITTEE

Eugene Demidenko, Professor at Dartmouth College (Advisor)

ABSTRACT

Graphical model estimation is the reconstruction of the edge weights of a network based on data. In this work, we focus on analyzing Ising graphical models, a kind of graphical model where variables are strictly dichotomous. We construct and analyze an upper bound on the error probability of a maximum likelihood estimator for estimating Ising graphical models. We additionally derive some results using information theory that allows us to give a necessary condition on the sample size so that estimation has any hope of being reliable. Finally, we demonstrate an embarrassingly parallel algorithm that uses variational inference to estimate the parameters of an Ising graphical model. Numerical experiments accompany this algorithm to demonstrate that it is competitive with state-of-the-art serial estimation methods.

ACKNOWLEDGMENTS

To begin with, I would like to thank Professor Eugene Demidenko for agreeing to supervise this research. I would also like to thank Rui Shu '15 for assisting me with the technical correctness of my claims.

This work was supported by the Kaminski Family Fund for senior theses.

CONTENTS

1 INTRODUCTION

1.1 THESIS OUTLINE

This thesis is generally related to the theory and practice of graph estimation on Markov random fields. Our main contribution is the development of a new algorithm for estimating edge weights using distributed computing. Our approach leverages sufficient statistics that can be estimated efficiently from data and then solves a variational optimization problem to recover the edge weights.

We begin in Chapter 2 by introducing some important components of statistical learning on graphs, particularly the idea of the maximum likelihood estimator, which is ubiquitous throughout statistics. We discuss the capabilities of the maximum likelihood estimator and provide an upper bound on the probability of graph estimation error under a zero-one loss function. We also discuss necessary conditions on the sample size such that recovery of the edge weights is a realistic goal.

Chapter 3 introduces an important graph estimator that leverages the sufficient statistics of the Markov random field. We show that the ℓ_2 error of these sufficient statistics may be made arbitrarily small as a function of the sample size with high probability.

Beginning in Chapter 4 we discuss the underlying strategy of estimating Markov random fields in practice. We present our major contribution: a distributed algorithm for graph estimation, which combines important existence and uniqueness, bijection, sufficient statistics, and consistency theorems to estimate graphical structure in parallel.

Finally, the manuscript concludes in Chapter 5 with some numerical simulations that demonstrate the performance of the distributed algorithm on synthetic data. This section illustrates that our algorithms offer a viable substitute to standard graph estimation procedures, especially when distributed resources are available.

The remainder of this chapter is devoted to providing an introduction to the fundamentals of Markov random fields and graph estimation.

1.2 INTRODUCTION TO THE THEORY OF MARKOV RANDOM FIELDS

A Markov random field gives a graphical representation of the joint probability distribution of random variables. The graph associated to a Markov random field contains information completely describing the conditional independence properties over that joint distribution. These models have applications in statistical physics, social networks, and image analysis.

An undirected graph G is an ordered pair (V, E) with $E \in V \times V$ for $V = \{1, \ldots, p\}$. In this case, we say that G has p vertices and an edge set E that connects those vertices. In the case of Markov random fields, we do not allow self-loops to present themselves in E. That is, for all $v \in V$, $(v, v) \notin E$. The neighborhood of a vertex v is defined as,

$$\mathrm{ne}_v = \{u \in V : (u, v) \in E\} \tag{1.1}$$

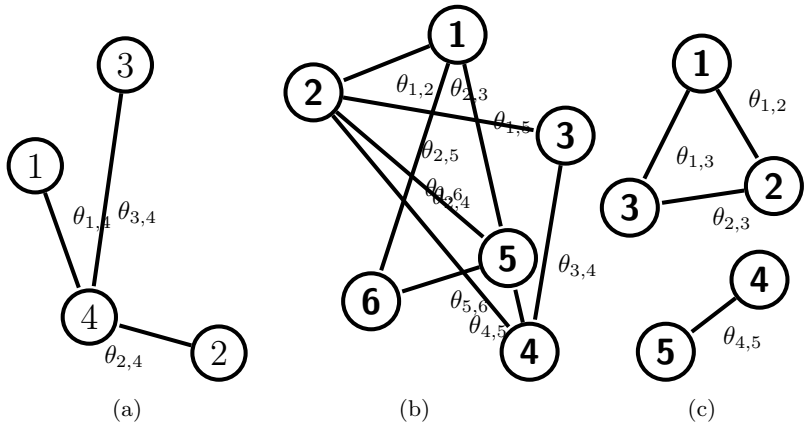


Figure 1.1: We show here some examples of the underlying graphical structure of a Markov random fields. In each case, the vertices in the graph are associated with a random variable assuming values in an specified set. Notice that the notation $\boldsymbol{\theta}_{ij}$ refers to the edge weight connecting vertices i and j. It is convenient to say $\boldsymbol{\theta}_{ij} = 0$ if and only if there is no edge connecting the vertices i and j in the graph.

Given the undirected graph G a Markov random field is straightforward to construct. In essence, each vertex v is associated with a random variable X_v and a joint probability distribution $\mathbb{P}$ is specified over the random vector $\mathbf{X} = \{X_1, \ldots, X_p\}$. In general we let $X_v \in J$ for all v, where J is a abstract but finite set of values that X_v may take on with nonzero probability. For standard Ising models, we simply let $J = \{-1, +1\}$. In this work, we study a probability mass function $\mathbb{P}$ which assumes the form,

$$\mathbb{P}\left[X_1 = x_1, \ldots, X_p = x_p; \boldsymbol{\theta}\right] = \frac{1}{Z\left(\boldsymbol{\theta}\right)} \exp\left\{ \sum_{(u,v)\in E} \boldsymbol{\theta}_{uv} x_u x_v \right\} \tag{1.2}$$

$$= \frac{1}{Z\left(\boldsymbol{\theta}\right)} \exp\left\{ \boldsymbol{x}' \boldsymbol{\Theta} \boldsymbol{x} \right\} \tag{1.3}$$

Here, $\boldsymbol{x}$ is a column vector corresponding to $\{x_i\}_{i=1}^p$ and $\boldsymbol{\theta}$ is a vector in $\mathbb{R}^{\binom{p}{2}}$ such that $\boldsymbol{\theta}_{uv} = 0$ if and only if $(u, v) \notin E$. Additionally, $\boldsymbol{\Theta}$ is an upper triangular matrix such that $\boldsymbol{\Theta}_{uv} = \boldsymbol{\theta}_{uv}$ if $u < v$ and $(u, v) \in E$.

The function $Z\left(\boldsymbol{\theta}\right)$ is often called the partition function and is used to ensure that $\mathbb{P}$ constitutes a valid probability mass function. Indeed, $Z\left(\boldsymbol{\theta}\right)$ can (and should be) thought of as a normalizing coefficient. In this work, however, we use the language partition function to refer to this quantity. It can be readily verified that

$$Z\left(\boldsymbol{\theta}\right) = \sum_{\mathbf{x}\in J^p} \exp\left\{ \sum_{(u,v)\in E} \boldsymbol{\theta}_{uv} x_u x_v \right\} \tag{1.4}$$

Here J^p denotes the Cartesian product space of the set J.

Exponential Families

It is worth mentioning that the equation in 1.2 belongs to the exponential family of probability distributions. These kinds of distributions arise naturally in a

variety of disciplines including social network analysis, econometrics, and computer vision. More generally, an exponential distribution may be expressed,

$$\mathbb{P}\left[Y = y; \boldsymbol{\theta}\right] = \frac{\exp\left\{\boldsymbol{\theta}^T \boldsymbol{\phi}\left(y\right)\right\}}{c\left(\boldsymbol{\theta}\right)} \tag{1.5}$$

Here, the vector $\boldsymbol{\phi}\left(y\right)$ is known as a vector of *sufficient statistics* and $c\left(\boldsymbol{\theta}\right)$ assumes the role of the normalizing constant.

It is common to leverage equation 1.2 to create a *conditional random field* model. In this case, the problem becomes the prediction of a particular X_j after one has conditioned on the remaining variables appearing in the joint distribution. The problem of structure estimation for this kind of graphical model, then, can be viewed equivalently as the problem of estimating the nonzero edge weights throughout the network that are connect the j^{th} node to the remaining vertices in the graph.

1.3 STRUCTURE ESTIMATION OF GRAPHICAL MODELS

The edge weight $\boldsymbol{\theta}_{st}$ captures the conditional dependence between the random variables X_s and X_t. Given that the set of random variables $\left\{X_j : j \notin \{s,t\}\right\}$ assume fixed values, a straightforward calculation yields,

$$\mathbb{P}\left[X_s, X_t \middle| \left\{X_j = x_j : j \notin (s,t)\right\}; \boldsymbol{\theta}\right] \tag{1.6}$$

$$\propto \exp\left\{\boldsymbol{\theta}_{st} x_s x_t + \sum_{i \in \text{ne}_s \setminus t} \boldsymbol{\theta}_{is} x_i x_s + \sum_{j \in \text{ne}_t \setminus s} \boldsymbol{\theta}_{jt} x_j x_t\right\} \tag{1.7}$$

ISING MODEL

Consider the setting where each covariate in the Markov random field may assume either of -1 or $+1$. Under these circumstances, we have the case of the Ising model, which arises out of statistical physics. The Ising model was developed by Ernst Ising who used it to describe the interactions between magnetic dipoles in a system. In this case, the -1 state corresponded to the negative spin state of an atom, whereas $+1$ indicated a positive spin. In this dichotomous case, a positive value of $\boldsymbol{\theta}_{st}$ implies that, conditional on the other atoms's states, atoms X_s and X_t are more likely than not to take on the same value. Conversely, a negative edge weight suggests that the states are more likely to differ (and in particular $X_s = -X_t$).

Assuming that section 1.2 holds in this fully dichotomous setting, it can be further shown that the full conditional distribution must satisfy [Foygel and Drton, 2014],

$$\log\left(\frac{\mathbb{P}\left[X_s = 1 \middle| X_t = x_t, t \neq s\right]}{1 - \mathbb{P}\left[X_s = 1 \middle| X_t = x_t, t \neq s\right]}\right) = \sum_{t \in \text{ne}_s} 2\theta_{st} x_t \tag{1.8}$$

This demonstrates that the Ising model has conditional distributions which resemble a logistic regression relation where X_s is the target random variable, and the remaining $\{X_t : t \neq s\}$ variables represent the covariates. Therefore, graph estimation in this case can be achieved by estimating the p logistic regression models given in eq. (1.8).

2 INFERENCE ON MARKOV RANDOM FIELDS

Outline
In this section we introduce and discuss the theory of maximum likelihood estimation on Ising graphical models. Due to the general intractability of the maximum likelihood estimator for large graphical models, what we present here amounts particularly to a theoretical analysis. In particular, we are able to present a non-trivial upper bound on the error probability for such an estimator by leveraging the Chernoff and union bounds.

2.1 MAXIMUM LIKELIHOOD ESTIMATION

Let $\boldsymbol{\theta} = \boldsymbol{\theta}(G)$ be the network edge weight parameter associated to the undirected graph $G \in \mathcal{G}_p$. We assume that $\boldsymbol{\theta} \in \Theta$ is a finite subset of $\mathbb{R}^d$ so that, $\mathcal{G}_p$ is also finite (and of the same cardinality). We therefore consider a reduction from estimation to testing so that instead of estimating $\boldsymbol{\theta}$ from data, we are instead attempting to choose that $\boldsymbol{\theta} \in \Theta$ which best explains the data according to some criterion.

Notice that this defines a probability distribution $\mathbb{P}_{\boldsymbol{\theta}}$ for the Markov random field whose structure is determined by G. Suppose then that we observe n i.i.d. random vectors from $\mathbb{P}_{\boldsymbol{\theta}}$ and denote them, $\mathbf{X}^n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$, where $\mathbf{X}_i \in J^p \; \forall \; i$. (This yields the consequence that $\mathbf{X}^n \in (J^p)^n$.) We are interested in the problem of graph structure estimation. Consider functions of the form $\phi : (J^p)^n \to \Theta$. We call this function a *graph estimator*. We analyze the zero-one loss function for ϕ,

$$\mathbf{1}\left\{\phi\left(\mathbf{X}^n\right) \neq \boldsymbol{\theta}\right\} \tag{2.1}$$

Notice that because $|\Theta| < \infty$, it is meaningful to consider the zero-one loss. The operator $\mathbf{1}$ denotes the indicator function, which assumes the value one when the argument is true, and otherwise is false. Correspondingly, this yields a natural risk definition,

$$\mathbb{P}\left[\phi\left(\mathbf{X}^n\right) \neq \boldsymbol{\theta}\right] = \mathbb{E}\left[\mathbf{1}\left\{\phi\left(\mathbf{X}^n\right) \neq \boldsymbol{\theta}\right\}\right] \tag{2.2}$$

Explanation of the Testing Framework
We note that the problem we consider is a simplification from usual maximum likelihood estimation. We restrict the class of possible graphical model parametrizations so that ϕ may be analyzed from an information-theoretic perspective. Under this class-constrained framework, maximum likelihood estimators are inappropriate since, in general, the θ' which maximizes the likelihood will not be an element of Θ. Hence, we require testing functions that are appropriate to a finite set of parametrizations.

We consider a graph estimator that exploits the normalized log-likelihood score. Given a collection of n i.i.d. vectors drawn from the probability distribution $\mathbb{P}_{\boldsymbol{\theta}}$, the normalized log-likelihood is,

$$\log \mathcal{L}\left(\mathbf{X}^n\right)_{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^{n} \log \mathbb{P}\left[\mathbf{X}_i; \boldsymbol{\theta}\right] \tag{2.3}$$

The maximum likelihood graph estimator is defined by,

$$\phi^* \left(\mathbf{X}^n\right) = \arg\max_{\boldsymbol{\theta}} \ \log \mathcal{L}\left(\boldsymbol{\theta}; \mathbf{X}^n\right) \tag{2.4}$$

We assume that the maximum likelihood graph estimator satisfies the identifiability property of maximum likelihood such that $\boldsymbol{\theta}\left(G\right) \neq \boldsymbol{\theta}\left(G'\right) \iff \phi^*_{\boldsymbol{\theta}}\left(\mathbf{X}^n\right) \neq \phi^*_{\boldsymbol{\theta}'}\left(\mathbf{X}^n\right)$. We have that ϕ^* fails to predict the true model if and only if there exists a $\boldsymbol{\theta}' \neq \boldsymbol{\theta}$ such that $\log \mathcal{L}\left(\boldsymbol{\theta}'; \mathbf{X}^n\right) > \log \mathcal{L}\left(\boldsymbol{\theta}; \mathbf{X}^n\right)$.

We will exhibit a bound on the error probability of the maximum likelihood graph estimator. That is, we will construct an upper bound for $\mathbb{P}\left[\phi^*\left(\mathbf{X}^n\right) \neq \boldsymbol{\theta}\right]$. The proof this result requires the development of first a few technical lemmas.

Lemma 2.1 (Chernoff Bound)**.**

$$\mathbb{P}\left[X \geq t\right] \leq \inf_{r>0} \left\{\exp\left\{-t \cdot r\right\} \cdot \mathbb{E}\left[\exp\left\{r \cdot X\right\}\right]\right\} \tag{2.5}$$

Proof. Fix $r > 0$. Then we obtain,

$$\begin{aligned} \mathbb{P}\left[X \geq t\right] &= \mathbb{P}\left[r \cdot X \geq r \cdot t\right] = \mathbb{P}\left[\exp\left\{r \cdot X\right\} \geq \exp\left\{r \cdot t\right\}\right] && (2.6)\\ &\leq \exp\left\{-r \cdot t\right\} \mathbb{E}\left[\exp\left\{r \cdot X\right\}\right] && (2.7) \end{aligned}$$

The inequality in section 2.1 follows from a direct application of Markov's Inequality. The term $\mathbb{P}\left[X \geq t\right]$ is r-independent, so it is valid to take the infimum over r. □

Convexity of Chernoff Bound

It is easy to show that, as a function of r, the right-hand side in Lemma 2.1 is convex. To show this, we first write,

$$f(r) = e^{rt}\mathbb{E}\left[e^{rX}\right]. \tag{2.8}$$

By direct calculation, the second derivative with respect to r can be shown to be,

$$\begin{aligned} \frac{\partial^2 f}{\partial r^2} &= -2te^{-rt}\mathbb{E}\left[Xe^{rX}\right] + t^2e^{-rt}\mathbb{E}\left[e^{rX}\right] + e^{-rt}\mathbb{E}\left[X^2e^{rX}\right] && (2.9)\\ &= e^{-rt}\mathbb{E}\left[e^{rX}\left(-2tX + t^2 + X^2\right)\right] && (2.10)\\ &= e^{-rt}\mathbb{E}\left[e^{rX}\left(t - X\right)^2\right]. && (2.11) \end{aligned}$$

Since each of the terms in the above equation are non-negative we have trivially that their product is non-negative. Hence $f(r)$ is convex. (Note that this remark suppresses some important regularity conditions with respect to the integrand under the expectation).

Tightness of the Chernoff Bound

The tightness of the Chernoff bound as presented essentially comes down to the tightness of Markov's inequality. Unfortunately, Markov's inequality is, in general, not tight. This is a natural consequence of the fact that Markov's inequality is formulated in full generality: the inequality is true for any non-negative random variable.

There are situations in which Markov's inequality is tight, however. Consider for instance a fixed $k \in \mathbb{N}$. Let $X = k$ with probability $1/k$ and $X = 0$

with probability $(k-1)/k$. Thus, it is apparent that $\mathbb{E}[X] = 1$. We therefore have that Markov's inequality is tight in the sense that,

$$\mathbb{P}[X \geq k] = \mathbb{E}[X]/k = 1/k \tag{2.12}$$

However our eventual upper bound on the error probability will be a decreasing function of the sample size such that the upper bound can be made arbitrarily close to zero for the testing case.

Lemma 2.2.

$$\mathbb{P}[\phi^*(\mathbf{X}^n) \neq \boldsymbol{\theta}; \boldsymbol{\theta}] \leq \sum_{\boldsymbol{\theta}' \neq \boldsymbol{\theta}} \mathbb{P}\left[\log \mathcal{L}(\boldsymbol{\theta}'; \mathbf{X}^n) - \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}^n) > 0; \boldsymbol{\theta}\right] \tag{2.13}$$

Proof. This follows immediately from a union bound. □

Theorem 2.3 (Maximum Error Probability)**.** *The error probability of the maximum likelihood graph estimator is bounded and that bound is,*

$$\mathbb{P}[\phi^*(\mathbf{X}^n) \neq \boldsymbol{\theta}] \leq \tag{2.14}$$

$$\sum_{\boldsymbol{\theta}' \neq \boldsymbol{\theta}} \exp\left\{ n \cdot \inf_{r>0} \log\left[\sum_{x \in J^p} \left(\mathbb{P}[x; \boldsymbol{\theta}']\right)^r \left(\mathbb{P}[x; \boldsymbol{\theta}]\right)^{1-r} \right] \right\} \tag{2.15}$$

Proof of Theorem 2.3. Define the random variable V as follows,

$$\begin{aligned} V &= \log \mathcal{L}(\boldsymbol{\theta}'; \mathbf{X}^n) - \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}^n) && (2.16) \\ &= \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}'] - \frac{1}{n}\sum_{i=1}^n \log \mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}] && (2.17) \end{aligned}$$

Notice first of all that each of the $\mathbf{X}_i$ is a p-vector whose elements are objects in J. It is assumed that $\mathbf{X}^n = \{\mathbf{X}_1, \ldots, \mathbf{X}_n\}$ is a set of observations drawn in an i.i.d. fashion from the probability mass function $\mathbb{P}[x]_{\boldsymbol{\theta}}$. The function $\mathbb{P}[x; \boldsymbol{\theta}']$ is another mass function which is not the distribution that yielded $\mathbf{X}^n$. We have by Lemma 2.1,

$$\frac{1}{n}\log \mathbb{P}[V > 0] \leq \frac{1}{n}\inf_{s>0} \log \mathbb{E}[\exp\{s \cdot V\}] \tag{2.18}$$

Then using the definition of V we obtain,

$$\begin{aligned} e^{sV} &= e^{\frac{s}{n}\sum_{i=1}^n \log \mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}']} e^{\frac{-s}{n}\sum_{i=1}^n \log \mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}]} && (2.19) \\ &= \prod_{i=1}^n \left(\mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}']\right)^{s/n} \prod_{i=1}^n \left(\mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}]\right)^{-s/n} && (2.20) \\ &= \prod_{i=1}^n \left(\mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}']\right)^{s/n} \left(\mathbb{P}[\mathbf{X}_i; \boldsymbol{\theta}]\right)^{-s/n} && (2.21) \end{aligned}$$

Then by leveraging expectations and the assumed independence of the $\mathbf{X}_i$ we have,

$$\begin{aligned} \mathbb{E}\left[e^{sV}\right] &= \prod_{i=1}^n \mathbb{E}\left[\left(\mathbb{P}[\mathbf{X}_i]_{\boldsymbol{\theta}'}\right)^{s/n} \left(\mathbb{P}[\mathbf{X}_i]_{\boldsymbol{\theta}}\right)^{-s/n}\right] && (2.22) \\ &= \prod_{i=1}^n \sum_{\mathbf{x} \in J^p} \left(\mathbb{P}[\mathbf{x}]_{\boldsymbol{\theta}}\right) \left(\mathbb{P}[\mathbf{x}]_{\boldsymbol{\theta}'}\right)^{s/n} \left(\mathbb{P}[\mathbf{x}]_{\boldsymbol{\theta}}\right)^{-s/n} && (2.23) \end{aligned}$$

$$= \prod_{i=1}^{n} \sum_{\mathbf{x} \in J^p} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}'}\right)^{s/n} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-s/n} \tag{2.24}$$

$$= \left(\sum_{\mathbf{x} \in J^p} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}'}\right)^{s/n} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-s/n} \right)^{n} \tag{2.25}$$

From this, a straightforward calculation exhibits,

$$\frac{1}{n} \inf_{s>0} \log \mathbb{E}\left[e^{sV}\right] = \inf_{s>0} \log \left(\sum_{\mathbf{x} \in J^p} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}'}\right)^{s/n} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-s/n} \right) \tag{2.26}$$

$$= \inf_{r>0} \log \left(\sum_{\mathbf{x} \in J^p} \left(\mathbb{P}\left[\mathbf{X}_i; \boldsymbol{\theta}'\right]\right)^{r} \left(\mathbb{P}\left[\mathbf{X}_i; \boldsymbol{\theta}\right]\right)^{1-r} \right) \tag{2.27}$$

The last step substitutes $r = \frac{s}{n}$ where it is understood that the infimum over $\frac{s}{n}$ is equivalent to the infimum over s itself. Then some simple algebra yields,

$$\frac{1}{n} \log \mathbb{P}\left[V > 0\right] \leq \inf_{r>0} \log \left(\sum_{\mathbf{x} \in J^p} \left(\mathbb{P}\left[\mathbf{X}_i\right]_{\boldsymbol{\theta}'}\right)^{r} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-r} \right) \tag{2.28}$$

$$\mathbb{P}\left[V > 0\right] \leq \exp\left\{ n \cdot \inf_{r>0} \log \left(\sum_{\mathbf{x} \in J^p} \left(\mathbb{P}\left[\mathbf{X}_i\right]_{\boldsymbol{\theta}'}\right)^{r} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-r} \right) \right\} \tag{2.29}$$

Substituting into the result from Lemma 2.2, we obtain that,

$$\mathbb{P}\left[\phi^*\left(\mathbf{X}^n\right) \neq \boldsymbol{\theta}\right] \leq \sum_{\boldsymbol{\theta}' \neq \boldsymbol{\theta}} \mathbb{P}\left[\log \mathcal{L}\left(\boldsymbol{\theta}'; \mathbf{X}^n\right) - \log \mathcal{L}\left(\boldsymbol{\theta}; \mathbf{X}^n\right) > 0\right] \tag{2.30}$$

$$\leq \sum_{\boldsymbol{\theta}' \neq \boldsymbol{\theta}} \exp\left\{ n \cdot \inf_{r>0} \log \sum_{\mathbf{x} \in J^p} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}'}\right)^{r} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-r} \right\} \tag{2.31}$$

□

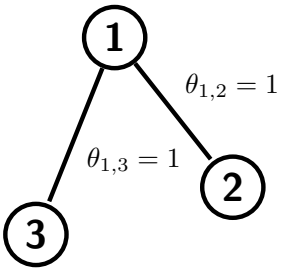
Improvements on What Comes Before

The bound demonstrated in eq. (2.27) represents a significant improvement over the bounds claimed in existing literature. This bound is, for example, precisely the *logarithm* of that provided in [Santhanam and Wainwright, 2012].

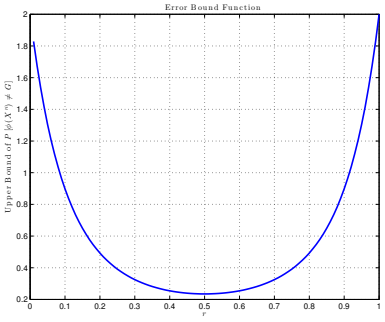
It is important to emphasize that ϕ^* represents an oracle estimator that is largely unhelpful in applications. This is because ϕ^* assumes knowledge of the underlying joint distribution coefficients (the $\boldsymbol{\theta}_{u,v}$ in section 1.2). Since it is unlikely that one would have knowledge of this vector in true application, ϕ^* becomes something of a theoretical object that can serve as a benchmark for the performance of other graph estimators.

Error Bounds for Simple Ising Graphs

We consider the case where G is a graph of three nodes and two edges, where each node is associated with a dichotomous random variable as in the Ising model. For simplicity, we assume that the two edges have equal and non-zero weight parameters. In this case, let $\boldsymbol{\theta} = \{\boldsymbol{\theta}_{1,2}, \boldsymbol{\theta}_{1,3}, \boldsymbol{\theta}_{2,3}\} = \{+1, +1, 0\}$. It is easy to check that this setting of parameters induces a family of three possible



(a) The underlying graphical structure of the example Markov random field. The family of three Markov random fields parameterized by θ are simply the graph above, and the additional two graphs with a missing edge.



(b) Above is a plot of the infimum of the upper bound construction being achieved in the example. In this case, $r = \frac{1}{2}$ yields the minimum. In the case of achieving this minimum, the upper bound on the error probability is nearly $\frac{1}{4}$.

Figure 2.1: We show here some graphical representations of the information presented in the example that leverages Theorem 2.3. We show the structure of the Markov random field for a graph with three vertices and two edges, where the non-zero edge weights are equal to one. We also demonstrate how taking the infimum over r as in Theorem 2.3 can yield a meaningful upper bound on the error probability of the maximum likelihood graph estimator.

Markov random fields. We assume that we have drawn $n = 50$ i.i.d. samples from a system obeying this model construction. We will now seek to provide an upper bound on the probability that the maximum likelihood graph estimator will yield the incorrect model.

By the result in Theorem 2.3, we have that an upper bound on the error probability takes the form,

$$\begin{aligned}
&\mathbb{P}\left[\phi\left(\mathbf{X}^n\right) \neq \boldsymbol{\theta}\right] \leq \\
&\exp\left\{n \cdot \inf_{r>0} \log \sum_{\mathbf{x}\in\{-1,+1\}^3} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}'=\{+1,0,+1\}}\right)^r \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-r}\right\} \\
&+ \exp\left\{n \cdot \inf_{r>0} \log \sum_{\mathbf{x}\in\{-1,+1\}^3} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}'=\{0,+1,+1\}}\right)^r \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-r}\right\} \\
&= 2\exp\left\{n \cdot \inf_{r>0} \log \sum_{\mathbf{x}\in\{-1,+1\}^3} \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}'=\{+1,0,+1\}}\right)^r \left(\mathbb{P}\left[\mathbf{x}\right]_{\boldsymbol{\theta}}\right)^{1-r}\right\} \quad (2.32)
\end{aligned}$$

The last inequality follows from the symmetry of the setting in which we have constructed the problem. This reduces evaluating the upper bound merely to the case of treating the case $\boldsymbol{\theta}' = \{+1, 0, +1\}$. Using MATLAB to calculate the infimum, we arrive at the upper bound,

$$\mathbb{P}\left[\phi\left(\mathbf{X}^n\right) \neq \boldsymbol{\theta}\right] \leq 0.234624 < \frac{1}{4} \quad (2.33)$$

From this, we can conclude that the maximum likelihood estimator will incorrectly construct the underlying graphical structure with probability not

exceeding one-quarter when fifty samples have been drawn from the distribution. Notice that this bound can be crushed to an arbitrarily small value via increased sampling. Indeed, by 100 samples the upper bound has been reduced to 0.027524.

3 AN ALTERNATIVE GRAPH ESTIMATOR

Outline

Motivated by the intractability of the general maximum likelihood estimator, we illustrate in this section an alternative graph estimator that relies on the "sufficient statistics" of the Markov random field, from which a precise reconstruction of the original edge weight parameters may be obtained. Importantly, we are able to prove a result that upper bounds the Euclidean error of the estimator from the true value with high probability, increasing in the number of observations. We also introduce an information-theoretic treatment of graph estimation that provides sample size lower bounds for desirable properties in particular graph estimation problems.

Here we demonstrate a method of constructing an estimator of the mean-value parameters associated with the canonical network parameters in the Markov random field. Let $\mathbf{X}_i \in \{-1, +1\}^p$ for $i = 1, \ldots, n$. Notice that for simplicity we restrict ourselves to the binary ("Ising field" case). We can define a mean-value parameter $\boldsymbol{\mu}$,

$$\mu_{st} = \mathbb{E}\left[X_s X_t\right] \tag{3.1}$$

We then construct an estimator of μ_{st}, denoted $\hat{\mu}_{st}$ as follows:

$$\hat{\mu}_{st} = \frac{1}{n} \sum_{i=1}^{n} X_{is} X_{it} \tag{3.2}$$

It is easy to confirm that $\hat{\boldsymbol{\mu}}$ is an unbiased estimator of the true $\boldsymbol{\mu}$ parameter. This can be seen from the straightforward calculation as follows,

$$\mathbb{E}\left[\hat{\mu}_{st}\right] = \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\left[X_{is} X_{it}\right] = \mathbb{E}\left[X_s X_t\right] = \mu_{st} \tag{3.3}$$

3.1 MOTIVATION

At first glance, it may seem that the the mean parameter specified as $\boldsymbol{\mu}$ will not prove particularly useful for the purposes of graph structure estimation. However, as we intimated in the title of this section, it can be shown that $\boldsymbol{\mu}$ is inherently connected to the graph edge parameter $\boldsymbol{\theta}$. We take this section to exhibit that connection and to show why $\boldsymbol{\mu}$ is useful in our problem.

Proposition 3.1 (Cumulant Function)**.** *Consider the function* $A\left(\boldsymbol{\theta}\right) = \log Z\left(\boldsymbol{\theta}\right)$. *Let* $\boldsymbol{\eta}\left(\mathbf{X}\right) = \left\{\eta_1\left(\mathbf{X}\right), \ldots, \eta_d\left(\mathbf{X}\right)\right\}^T$ *be a set of* d *sufficient statistics. Commonly, each sufficient statistic takes the form of a pairwise multiplication of random variables:* $\eta_\alpha\left(\mathbf{X}\right) = X_i X_j$, *where* α *indexes elements of* $\{1, \ldots, p\} \times \{1, \ldots, p\}$ *and where* p *is the number of vertices of the graph. It can be shown that in the general case of Markov random fields,*

$$A\left(\boldsymbol{\theta}\right) = \log \int_{J^p} \exp\left\{\left\langle \boldsymbol{\theta}, \boldsymbol{\eta}\left(x\right)\right\rangle\right\} \nu\left(dx\right) \tag{3.4}$$

Here, we denote by ν the counting measure for the probability mass function $\mathbb{P}$. Then A is referred to as the cumulant function of the associated exponential distribution. It can be shown that A has the following properties [Wainwright and Jordan, 2008],

(a) The first two derivatives of A yield the cumulants of the random vector $\boldsymbol{\eta}(\mathbf{X})$, where $\mathbf{X}$ is a random vector from the Markov random field parametrized by $\boldsymbol{\theta}$. In particular,

$$\frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\alpha} = \mathbb{E}\left[\boldsymbol{\eta}_\alpha(\mathbf{X})\right]_{\boldsymbol{\theta}} \tag{3.5}$$

$$\frac{\partial A^2(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\alpha \partial \boldsymbol{\theta}_\beta} = \mathbb{E}\left[\boldsymbol{\eta}_\alpha(\mathbf{X})\,\boldsymbol{\eta}_\beta(\mathbf{X})\right]_{\boldsymbol{\theta}} \tag{3.6}$$

$$-\mathbb{E}\left[\boldsymbol{\eta}_\alpha(\mathbf{X})\right]_{\boldsymbol{\theta}} \mathbb{E}\left[\boldsymbol{\eta}_\beta(\mathbf{X})\right]_{\boldsymbol{\theta}} \tag{3.7}$$

(b) It can be demonstrated that A is a convex function over its domain Θ. Moreover, this convexity can be strengthened to strictness if the representation of the exponential distribution is minimal.

Proof. We can yield the first claim through direct computation,

$$\frac{\partial A(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_\alpha} = \frac{\partial}{\partial \boldsymbol{\theta}_\alpha}\left\{\log \int_{J^p} \exp\left\{\langle \boldsymbol{\theta}, \boldsymbol{\eta}(x)\rangle\right\} \nu(dx)\right\} \tag{3.8}$$

$$= \frac{\int_{J^p} \frac{\partial}{\partial \boldsymbol{\theta}_\alpha} \exp\left\{\langle \boldsymbol{\theta}, \boldsymbol{\eta}(x)\rangle\right\} \nu(dx)}{\int_{J^p} \exp\left\{\langle \boldsymbol{\theta}, \boldsymbol{\eta}(u)\rangle\right\} \nu(du)} \tag{3.9}$$

$$= \int_{J^p} \boldsymbol{\eta}_\alpha(x) \frac{\exp\left\{\langle \boldsymbol{\theta}, \boldsymbol{\eta}(x)\rangle\right\} \nu(dx)}{\int_{J^p} \exp\left\{\langle \boldsymbol{\theta}, \boldsymbol{\eta}(u)\rangle\right\} \nu(du)} \tag{3.10}$$

$$= \mathbb{E}\left[\boldsymbol{\eta}_\alpha(\mathbf{X})\right] \tag{3.11}$$

The proof for the higher order derivatives is no different and can be exhibited in a similarly straightforward manner. The convexity result is less trivial and is presently withheld from this document. However, a proof may be references in the text in [Wainwright and Jordan, 2008]. □

Proposition 3.2. *Consider the gradient mapping defined by,*

$$\nabla A = \left\{\frac{\partial}{\partial \boldsymbol{\theta}_\alpha} \log \int_{J^p} \exp\left\{\langle \boldsymbol{\theta}, \boldsymbol{\eta}(x)\rangle\right\} \nu(dx)\right\}_{\alpha=1}^{\binom{p}{2}} \tag{3.12}$$

$$= \left\{\mathbb{E}\left[\boldsymbol{\eta}_\alpha(\mathbf{X})\right]\right\}_{\alpha=1}^{\binom{p}{2}} \tag{3.13}$$

Then ∇A is a bijective mapping from $\Theta \to \mathcal{M}$ if and only if the exponential representation is minimal, where $\mathcal{M}$ is the space of mean-parameters for each pair of random variables in the Markov random field. Furthermore, to recover the $\boldsymbol{\theta}$ parameter from $\boldsymbol{\mu}$ one needs only to solve the variational optimization problem,

$$A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Theta} \boldsymbol{\theta}^T \boldsymbol{\mu} - A(\boldsymbol{\theta}) \tag{3.14}$$

Proof. This proof is reproduced from [Wainwright and Jordan, 2008] ; refer to that text for a more thorough treatment and additional details. Suppose that the representation is not minimal. Then there exists a vector, denoted $\gamma \in \mathbb{R}^d$ that is not the zero vector such that the inner product $\langle \gamma, \eta(x)\rangle = C$. Denote by $\boldsymbol{\theta}$ a parameter is Θ and define further an alternative parameter $\boldsymbol{\theta}' = \boldsymbol{\theta} + t \cdot \gamma$ for $t \in \mathbb{R}$. It can be demonstrated easily that in an open

parameter space Θ, choosing t appropriately small gives $\boldsymbol{\theta}' \in \Theta$ as well. Notice that γ, by definition, will generate two probability distributions $\mathbb{P}_{\boldsymbol{\theta}}$ and $\mathbb{P}_{\boldsymbol{\theta}'}$ that differ only by a normalization constant. Then unfortunately we are faced with the consequence that $\nabla A(\boldsymbol{\theta}) = \nabla A(\boldsymbol{\theta}')$ and we have the implication that the mapping induced by ∇A is not bijective.

For the opposite direction, we this time assume that the representation is minimal. We have by Theorem 3.1 that A is strictly convex. Indeed, for any convex and differentiable function, we have that the following inequality holds,

$$\langle \nabla A(\boldsymbol{\theta}) - \nabla A(\boldsymbol{\theta}'), \boldsymbol{\theta} - \boldsymbol{\theta}' \rangle > 0 \tag{3.15}$$

This is true for all distinct $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$. This shows that ∇A is a bijection.

As is proved in [Wainwright and Jordan, 2008], we obtain an alternative representation for the function A,

$$A(\boldsymbol{\theta}) = \sup_{\boldsymbol{\mu} \in \mathcal{M}} \{\langle \boldsymbol{\theta}, \boldsymbol{\mu} \rangle - A^*(\boldsymbol{\theta})\} \tag{3.16}$$

The *conjugate dual function* of A, denoted A^*, is written as in eq. (3.14). It is clear that maximizing the dual function as specified yields the unique (by bijectivity) $\boldsymbol{\theta}$ parameter corresponding to $\boldsymbol{\mu}$. □

This demonstrates a useful and surprising fact about the mean parameter $\boldsymbol{\mu}$: That it is equally descriptive as, and in fact equivalent to, the parameterization induced on the Markov random field by $\boldsymbol{\theta}$. Therefore, by estimating $\hat{\boldsymbol{\mu}}$, we recover as much information relating to graphical structure as we could through the direct estimation of $\boldsymbol{\theta}$ itself.

Recovery of Ising Parameters by Variational Mean Estimation

We now illustrate a direct calculation of $\boldsymbol{\mu}$ for a simple case of a Ising model with three random variables. The possible states that can be assumed in this Markov random field are,

$$\begin{aligned} J^3 = \{-1,+1\}^3 = \{&(-1,-1,-1), (-1,-1,+1), (-1,+1,-1), \\ &(-1,+1,+1), (+1,-1,-1), \\ &(+1,-1,+1), (+1,+1,-1), \\ &(+1,+1,+1)\} \end{aligned}$$

We suppose that $\boldsymbol{\theta} = \{\theta_{12}, \theta_{23}, \theta_{13}\} = \{0, -2, +1\}$. Aided by a computer, it is easy to calculate the partition function for this example situation. We find that,

$$Z(\boldsymbol{\theta}) = \sum_{\mathbf{x} \in J^3} \exp\{0 \cdot \mathbf{x}_1 \cdot \mathbf{x}_2 - 2 \cdot \mathbf{x}_2 \mathbf{x}_3 + 1 \cdot \mathbf{x}_1 \mathbf{x}_3\} \approx 46.443 \tag{3.17}$$

An application of the usual expected value formula quickly yields,

$$\boldsymbol{\mu} = \{\mu_{12}, \mu_{13}, \mu_{23}\} \approx \{-0.7342, -0.9640, 0.7616\} \tag{3.18}$$

Then in order to recover the $\boldsymbol{\theta}$ parameter from $\boldsymbol{\mu}$ it is necessary to solve the following variational problem from eq. (3.14):

$$A^*(\boldsymbol{\mu}) = \sup_{\boldsymbol{\theta} \in \Theta} \Big(\theta_{12}\mu_{12} + \theta_{13}\mu_{13} + \theta_{23}\mu_{23} -$$

$$\log\left(\sum_{x\in J^3}\exp\left\{\theta_{12}x_1x_2+\theta_{13}x_1x_3+\theta_{23}x_2x_3\right\}\right)\Bigg)$$

Using a programming language such as MATLAB or Mathematica we solve this optimization problem numerically from $\boldsymbol{\theta}_0 = \{0,0,0\}$ thus reflecting the initial belief that there is no conditional relationship between the random variables. Optimizing with respect to $\boldsymbol{\theta}$ yields a maximum at $\{0,-2,1\}$, which successfully recovers the original problem parameters.

It is apparent then that a major consideration involved in using the $\boldsymbol{\mu}$ parameter as a apparatus by which to reconstruct $\boldsymbol{\theta}$ is the extent to which $\boldsymbol{\mu}$ is reflected in $\hat{\boldsymbol{\mu}}$. In the next section we construct a probabilistic bound on the norm of the difference vector between the true $\boldsymbol{\mu}$ and its estimator.

3.2 BOUNDING THE ERROR OF $\hat{\boldsymbol{\mu}}$

We seek to bound the probability that the L_2-norm of the difference of the estimator and the true parameter is greater than a specified amount. The following is a naive approach that leverages an absolute bound on the differences of individual entries in either parameter vector. We proceed as follows,

$$\mathbb{P}\left[||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}||_2^2 \geq t\right] = \mathbb{P}\left[\sum_{(s,t)}\left(\hat{\mu}_{st}-\mu_{st}\right)^2 \geq t\right] \tag{3.19}$$

It is easy to verify that since each of the $X_s \in \{-1,+1\}$ that the following inequality must necessarily hold,

$$\left(\hat{\mu}_{st}-\mu_{st}\right)^2 \leq 4 \tag{3.20}$$

From here, it is possible to leverage the Chernoff bound to give an upper probabilistic limit on the L_2-norm deviation of the estimator $\hat{\boldsymbol{\mu}}$ from the true $\boldsymbol{\mu}$. The results are as follows,

$$\mathbb{P}\left[||\hat{\boldsymbol{\mu}}-\boldsymbol{\mu}||_2^2 \geq t\right] = \mathbb{P}\left[\sum_{(s,t)}\left(\hat{\mu}_{st}-\mu_{st}\right)^2 \geq t\right] \tag{3.21}$$

$$\leq \inf_{\gamma>0}\exp\left\{-\gamma\cdot t\right\}\mathbb{E}\left[\exp\left\{\gamma\sum_{(s,t)}\left(\hat{\mu}_{st}-\mu_{st}\right)^2\right\}\right] \tag{3.22}$$

$$\leq \inf_{\gamma>0}\exp\left\{-\gamma\cdot t\right\}\mathbb{E}\left[\exp\left\{\gamma\sum_{(s,t)}4\right\}\right] \tag{3.23}$$

$$= \inf_{\gamma>0}\exp\left\{-\gamma\cdot t\right\}\mathbb{E}\left[\exp\left\{\gamma\cdot\binom{p}{2}\cdot 4\right\}\right] \tag{3.24}$$

$$= \exp\left\{-\gamma\cdot t\right\}\exp\left\{\gamma\cdot\binom{p}{2}\cdot 4\right\} \tag{3.25}$$

For $0<\delta\leq 1$, we have that the following value for t bounds the desired probability by δ.

$$t > 2\cdot p\cdot(p-1) - \frac{\log\delta}{\gamma} \leq 2\cdot p\cdot(p-1) - \log\delta \tag{3.26}$$

3.3 IMPROVING THE ERROR BOUND

The limiting factor of the above proof is that we are not able to provide a better bound on the value $(\hat{\mu}_{st} - \mu_{st})^2$ that is better than four. In this section we are able to bound this quantity with a certain probability that depends on the number of observations.

Proposition 3.3. *For any $t^\star > 0$, suppose that $\max_{(s,t)} |\hat{\mu}_{st} - \mu_{st}| < t^\star$. It can be verified that this occurs with probability not less than,*

$$1 - 2\exp\left\{\frac{-n\,(t^\star)^2}{2} + \log\binom{p}{2}\right\} \tag{3.27}$$

Then with probability not less than the value in Equation (3.27), the deviation of $\boldsymbol{\mu}$ from its estimate is bounded in probability as,

$$\mathbb{P}\left[||\hat{\boldsymbol{\mu}} - \boldsymbol{\mu}||_2^2 \geq t\right] \leq \inf_{\gamma > 0} \exp\{-\gamma \cdot t\} \exp\left\{\gamma \cdot \binom{p}{2} \cdot (t^\star)^2\right\} \tag{3.28}$$

Proof. Notice that the random variables $\{\mathbf{X}_i\}_{i=1}^n$ are i.i.d. (by assumption) and lie in the set $\{-1, +1\}^p$. A straightforward application of Hoeffding's inequality yields,

$$\mathbb{P}\left[|\hat{\mu}_{st} - \mu_{st}| \geq t^\star\right] \leq 2\exp\left\{\frac{-n\,(t^\star)^2}{2}\right\} \tag{3.29}$$

By a union bound over all of the $\binom{p}{2}$ edges, we obtain,

$$\begin{aligned}
\mathbb{P}\left[\max_{(s,t)} |\hat{\mu}_{st} - \mu_{st}| \geq t^\star\right] &\leq \sum_{(s,t)} 2\exp\left\{\frac{-n\,(t^\star)^2}{2}\right\} && (3.30)\\
&= 2\binom{p}{2}\exp\left\{\frac{-n\,(t^\star)^2}{2}\right\} && (3.31)\\
&= 2\exp\left\{\frac{-n\,(t^\star)^2}{2} + \log\binom{p}{2}\right\} && (3.32)
\end{aligned}$$

From here we proceed as follows,

$$\begin{aligned}
\mathbb{P}\left[\max_{(s,t)} |\hat{\mu}_{st} - \mu_{st}| < t^\star\right] &= 1 - \mathbb{P}\left[\max_{(s,t)} |\hat{\mu}_{st} - \mu_{st}| \geq t^\star\right] && (3.33)\\
&\geq 1 - 2\exp\left\{\frac{-n\,(t^\star)^2}{2} + \log\binom{p}{2}\right\} && (3.34)
\end{aligned}$$

At this point we have bounded $\mathbb{P}\left[\max_{(s,t)} |\hat{\mu}_{st} - \mu_{st}| < t^\star\right]$ by an amount decreasing in the number of observations n. In particular, we now claim that for all (s,t) we have $(\hat{\mu}_{st} - \mu_{st})^2 \leq (t^\star)^2$ and that this occurs with probability no less than,

$$1 - 2\exp\left\{\frac{-n\,(t^\star)^2}{2} + \log\binom{p}{2}\right\} \tag{3.35}$$

Thus, with the claimed probability in eq. (3.35), we have that Theorem 3.3 follows immediately. □

Corollary 3.4. *We wish to construct an upper bound as follows,*

$$\inf_{\gamma>0} \exp\left\{-\gamma \cdot t\right\} \exp\left\{\gamma \cdot \binom{p}{2} \cdot \left(t^{\star}\right)^{2}\right\} < \delta \tag{3.36}$$

We require that t satisfy the inequality,

$$t > \frac{(p-1) \cdot p \cdot \left(t^{\star}\right)^{2}}{2} - \frac{\log \delta}{\gamma} \tag{3.37}$$

Furthermore, we will wish to bound the required conditional probability,

$$1 - 2\exp\left\{\frac{-n \cdot \left(t^{\star}\right)^{2}}{2} + \log\binom{p}{2}\right\} > 1 - \delta' \tag{3.38}$$

This in turn requires that we sample from the distribution $\mathbb{P}$ a number of times n that obeys the inequality,

$$n > \left(\log\binom{p}{2} - \log\frac{\delta'}{2}\right) \cdot \frac{2}{\left(t^{\star}\right)^{2}} \tag{3.39}$$

3.4 CHARACTERIZING ESTIMATION IN TERMS OF $\boldsymbol{\mu}$

After establishing an upper bound on the probability of deviation from the true mean, we proceed now to characterize *any* graph estimator as it depends on the estimate of $\boldsymbol{\mu}$. In particular, we seek to establish a theory for describing for an arbitrary graph estimator,

$$\mathbb{P}\left[\phi\left(\mathbf{X}^{n}\right) \neq G\right] \tag{3.40}$$

Theorem 3.5 (Fano's Inequality)**.** *We consider a set of models, each of which is parameterized by one of $\{\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_m\}$, each assumed to be equally likely. Then suppose we let $K \in \{1, \ldots, M\}$ at random and draw n i.i.d. samples from a true probability distribution $\mathbb{P}_{\boldsymbol{\theta}_K}$, denoted $\mathbf{X}^n$. Furthermore, suppose we have an estimator $\phi : \mathbf{X}^n \to \{1, \ldots, M\}$. Then Fano's lemma yields a lower bound on the error probability specified in terms of entropy from information theory. In particular, this bound is,*

$$\max_{k \in \{1,\ldots,M\}} \mathbb{P}\left[\phi\left(\mathbf{X}^{n}\right) \neq G; \boldsymbol{\theta}_k\right] \geq 1 - \frac{\mathcal{I}\left(\mathbf{X}^{n}; \boldsymbol{\theta}_k\right) + \log 2}{\log\left(M-1\right)} \tag{3.41}$$

We denote by $\mathcal{I}\left(\mathbf{X}^{n}; \boldsymbol{\theta}\right)$ the mutual information between $\mathbf{X}^n$ and $\boldsymbol{\theta}$.

A major consideration is applying Fano's Inequality to estimation problems is the difficulty that accompanies calculating the mutual information directly. However, for our purposes it will suffice to bound the mutual information from above. We accomplish this by the following lemma.

Lemma 3.6.

$$\mathcal{I}\left(\boldsymbol{\theta}_k, \mathbf{X}^{n}\right) \leq \frac{n}{M^{2}} \sum_{i=1}^{M} \sum_{j=1}^{M} S\left(p(x; \theta_i) \,||\, p(x; \theta_j)\right) \tag{3.42}$$

Here we have denoted by $S\left(\cdot \,||\, \cdot\right)$ the symmetrized Kullback-Leibler divergence, which assumes the form,

$$S\left(\mathbb{P} \,||\, \mathbb{P}'\right) = \sum_{(s,t) \in E} \left(\theta_{st} - \theta'_{st}\right)\left(\mu_{st} - \mu'_{st}\right) \tag{3.43}$$

Proof. Let γ be a random variable on $\{1, \ldots, M\}$, each outcome occurring with equal probability. For notational simplicity we take $\mathbb{P}[x; \theta_\gamma] = \mathbb{P}[x; \gamma]$. We also use the notation $\mathbb{P}_{\theta_\gamma} = \mathbb{P}[\cdot; \gamma]$ when refering to the distribution itself. Then for $\gamma = m$ and a matrix $\mathbf{X}^n$ of i.i.d. observations drawn from $p(\cdot; \gamma)$, we obtain,

$$\begin{aligned} I(\gamma; \mathbf{X}^n) &= H(\mathbf{X}^n) - H(\mathbf{X}^n|\gamma) && (3.44)\\ &= H(\mathbf{X}_1; \ldots; \mathbf{X}_n) - H(\mathbf{X}_1; \ldots; \mathbf{X}_n|\gamma) && (3.45)\\ &\leq H(\mathbf{X}_1) + \ldots + H(\mathbf{X}_n) - H(\mathbf{X}_1; \ldots; \mathbf{X}_n|\gamma) && (3.46) \end{aligned}$$

$$\begin{aligned} = \; & H(\mathbf{X}_1) + \ldots + H(\mathbf{X}_n) - [H(\mathbf{X}_1|\gamma) + \\ & H(\mathbf{X}_2|\gamma) + \ldots + H(\mathbf{X}_n|\gamma)] \end{aligned} \quad (3.47)$$

$$\begin{aligned} &= \sum_{i=1}^{n} I(\gamma; \mathbf{X}_i) && (3.48)\\ &= nI(\gamma; \mathbf{X}_1) && (3.49) \end{aligned}$$

Here, the first inequality is a standard bound on the entropy of a collection of random variables, and the equality below it (that is, the third-to-last equality above) follows because the $\{\mathbf{X}_i\}_{i=1}^n$ are conditionally independent given γ.

Now it suffices to prove a bound on $I(\gamma; \mathbf{X}_1)$. We have $I(\gamma; \mathbf{X}_1) = H(\mathbf{X}_1) - H(\mathbf{X}_1|\gamma)$ and furthermore,

$$\begin{aligned} H(\mathbf{X}_1) &= -\sum_x p(x) \log(p(x)) && (3.50)\\ &= -\sum_{\gamma=1}^{M} \sum_x \frac{1}{M} \mathbb{P}[x; \gamma] \log \left(\frac{1}{M} \sum_{\gamma'=1}^{M} \mathbb{P}[x; \gamma'] \right) && (3.51)\\ &\leq -\sum_{\gamma=1}^{M} \sum_x \frac{1}{M} \mathbb{P}[x; \gamma] \left[\frac{1}{M} \sum_{\gamma'=1}^{M} \log(\mathbb{P}[x; \gamma']) \right] && (3.52)\\ &= \frac{1}{M^2} \sum_{\gamma'=1}^{M} \sum_{\gamma=1}^{M} \left[\sum_x \mathbb{P}[x; \gamma] \log \left(\frac{1}{\mathbb{P}[x; \gamma']} \right) \right] && (3.53) \end{aligned}$$

Here, the inequality uses Jensen's inequality on the concave function log. Finally,

$$\begin{aligned} H(\mathbf{X}_1|\gamma) &= -\sum_{\gamma=1}^{M} \sum_x \frac{\mathbb{P}[x; \gamma]}{M} \log(\mathbb{P}[x; \gamma]) && (3.54)\\ &= -\frac{1}{M^2} \sum_{\gamma'=1}^{M} \sum_{\gamma=1}^{M} \sum_x \mathbb{P}[x; \gamma] \log(\mathbb{P}[x; \gamma]) && (3.55) \end{aligned}$$

This then yields,

$$\begin{aligned} I(\gamma; \mathbf{X}_1) &= H(\mathbf{X}_1) - H(\mathbf{X}_1|\gamma) && (3.56)\\ &\leq \frac{1}{M^2} \sum_{\gamma'=1}^{M} \sum_{\gamma=1}^{M} \sum_x \mathbb{P}[x; \gamma] \log \left(\frac{\mathbb{P}[x; \gamma]}{\mathbb{P}[x; \gamma']} \right) && (3.57)\\ &= \frac{1}{M^2} \sum_{\gamma'=1}^{M} \sum_{\gamma=1}^{M} D_{KL}\left(\mathbb{P}[x; \gamma] \,||\, \mathbb{P}[x; \gamma'] \right) && (3.58) \end{aligned}$$

The result follows then from the realization that,

$$\frac{1}{M^2}\sum_{\gamma'=1}^{M}\sum_{\gamma=1}^{M} D_{KL}\left(\mathbb{P}\left[x;\gamma\right]||\mathbb{P}\left[x;\gamma'\right]\right) \tag{3.59}$$

$$= \frac{1}{M^2}\sum_{\gamma'=1}^{M}\sum_{\gamma=1}^{\gamma'-1}\left[D_{KL}\left(\mathbb{P}\left[x;\gamma\right]||\mathbb{P}\left[x;\gamma'\right]\right)\right. \tag{3.60}$$

$$\left. + D_{KL}\left(\mathbb{P}\left[x;\gamma'\right]||\mathbb{P}\left[x;\gamma\right]\right)\right] \tag{3.61}$$

This completes the proof. □

Definition 3.1 (δ-inconsistent)**.** *We say that a classifier is δ-inconsistent over the parameterization family $\{\boldsymbol{\theta}_1,\ldots,\boldsymbol{\theta}_M\}$ if,*

$$\max_{k\in\{1,\ldots,M\}} \mathbb{P}\left[\phi\left(\mathbf{X}^n\right)\neq G;\boldsymbol{\theta}_k\right] > \delta \tag{3.62}$$

Accordingly, Fano's Lemma provides an upper bound over the number of observations n that will *guarantee* a δ-inconsistent estimator:

$$n < \frac{\left[(1-\delta)\cdot\log\left(M-1\right)-\log 2\right]\cdot M^2}{\sum_{i=1}^{M}\sum_{j=1}^{M} S\left(\mathbb{P}_{\boldsymbol{\theta}_i}||\,\mathbb{P}_{\boldsymbol{\theta}_j}\right)} \tag{3.63}$$

Any n that obeys this inequality implies that an estimator is δ-inconsistent. Here, $\mathbb{P}$ and $\mathbb{P}'$ have edge weights $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$ and mean parameters $\boldsymbol{\mu}$ and $\boldsymbol{\mu}'$, respectively. We prove this characterization of the symmetric Kullback-Leibler divergence in the following lemma.

Lemma 3.7. *Let $\mathbb{P}$ and $\mathbb{P}'$ be probability distributions as in section 1.2. The (symmetrized) Kullback-Leibler divergence measures the the closeness of two distributions. More precisely, this divergence measures the amount of information lost when $\mathbb{P}'$ is used as an estimate of the distribution $\mathbb{P}$. A convenient representation of the symmetrized Kullback-Leibler divergence is presented as follows,*

$$S\left(\mathbb{P}\,||\,\mathbb{P}'\right) = \sum_{(s,t)\in E}\left(\theta_{st}-\theta'_{st}\right)\left(\mu_{st}-\mu'_{st}\right) \tag{3.64}$$

Proof. This a straightforward, but extensive, calculation. We present the more crucial steps as follows,

$$S\left(\mathbb{P}\,||\,\mathbb{P}'\right) = \sum_{x\in J^p}\mathbb{P}\left[x\right]\log\frac{\mathbb{P}\left[x\right]}{\mathbb{P}\left[x\right]'} + \sum_{x\in J^p}\mathbb{P}\left[x\right]'\log\frac{\mathbb{P}\left[x\right]'}{\mathbb{P}\left[x\right]} \tag{3.65}$$

$$= \sum_{x\in J^p}\mathbb{P}\left[x\right]\log\mathbb{P}\left[x\right] - \mathbb{P}\left[x\right]\log\mathbb{P}\left[x\right]' + \mathbb{P}\left[x\right]'\log\mathbb{P}\left[x\right]' - \mathbb{P}\left[x\right]'\log\mathbb{P}\left[x\right] \tag{3.66}$$

$$= \sum_{(s,t)\in E}\theta_{st}\mathbb{E}\left[\mathbf{X}_s\mathbf{X}_t\right]_{\mathbb{P}} - \boldsymbol{\theta}'_{st}\mathbb{E}\left[\mathbf{X}_s\mathbf{X}_t\right]_{\mathbb{P}} + \boldsymbol{\theta}'_{st}\mathbb{E}\left[\mathbf{X}_s\mathbf{X}_t\right]_{\mathbb{P}} - \boldsymbol{\theta}_{st}\mathbb{E}\left[\mathbf{X}_s\mathbf{X}_t\right]_{\mathbb{P}'} \tag{3.67}$$

$$= \sum_{(s,t)\in E}\left(\theta_{st}-\theta'_{st}\right)\left(\mu_{st}-\mu'_{st}\right) \tag{3.68}$$

This demonstrates the desired result. □

In this section we argue essentially as did Santhanam et al. [Santhanam and Wainwright, 2012] to demonstrate an upper bound on n that guarantees that any graph estimator must be δ-inconsistent. We consider a family of graphical models on p vertices and at most k nonzero edges. Then we demonstrate an upper bound on the number of observations that guarantees (by Theorem 3.5) that any graph estimator will be δ-inconsistent over that family of graphs by explicitly calculating the symmetrized Kullback-Leibler divergence for a particular sub-family of graphs. In turn, this allows us to create a lower bound (although *not* a greatest lower bound) on the number of observations required to prevent any graph estimator from being failing for this family.

Consider the sub-family of graphs $\mathcal{G}_{p,1}$. These are the graphs which contain p vertices and only a single edge. Let use denote by $G_{(u,v)} \in \mathcal{G}_{p,1}$ that graph whose single edge connects vertices u and v. A direct result of this yields that $E = \{(u, v)\}$. We focus here on analyzing the mean parameter $\mathbb{E}[X_u X_v]$ as opposed to the edge weight parameter θ_{uv}.

Lemma 3.8. *Consider the mean parameter $\mu_{uv} = \lambda \neq 0$ with a corresponding graphical model $G_{(u,v)}$. Then the probability distribution $\mathbb{P}$ that is parameterized by $\boldsymbol{\theta}$ (and consequently, by bijection, $\boldsymbol{\mu}$), obeys the symmetrized Kullback-Leibler divergence with respect to another distribution $\mathbb{Q}$ with underlying graph $G_{(s,t)}$ such that $(s, t) \neq (u, v)$ as follows:*

$$S\left(\mathbb{P}_{G_{(s,t)}} \,||\, \mathbb{Q}_{G_{(u,v)}}\right) = 2\lambda \tanh^{-1}(\lambda) \tag{3.69}$$

Proof. This is a straightforward application of the definition of $S(\cdot \,||\, \cdot)$. Let $\theta_{ij}(G_{st})$ refer to weight of the edge between vertex i and vertex j in the graph G_{st}, the graph whose only non-zero edge connects vertices s and t. We proceed as follows, leveraging the representation of the symmetrized Kullback-Leibler divergence in Theorem 3.7,

$$S\left(\mathbb{P}_{G_{(s,t)}} \,||\, \mathbb{Q}_{G_{(u,v)}}\right) = \tag{3.70}$$

$$\lambda\left\{\left(\theta_{uv}(G_{uv}) - \theta_{st}(G_{uv})\right) - \left(\theta_{uv}(G_{st}) - \theta_{st}(G_{st})\right)\right\} \tag{3.71}$$

$$= 2\lambda\theta_{st} \tag{3.72}$$

Here, the equality comes from the following additional fact relating to the μ_{st} parameter. In particular,

$$\mu_{st} = \frac{\exp\{\theta_{st}\} - \exp\{\theta_{st}\}}{\exp\{\theta_{st}\} + \exp\{\theta_{st}\}} = \tanh(\theta_{st}) \tag{3.73}$$

Thus, we obtain the result that $\theta_{st} = \tanh^{-1}(\mu_{st})$. From this, we have the implication that $\mu_{st} = 0 \iff \theta_{st} = 0$ in this particular families of graphs. The claim follows immediately from an elementary combination of these results. □

Proposition 3.9. *By leveraging Theorem 3.5, we can apply the result directly to yield an upper bound on n that guarantees that any graph estimator is δ-inconsistent. We have the result, that for the family of graphs $\mathcal{G}_{p,k}$, a graph estimator will be δ-inconsistent if the sample size has an upper limit that obeys the inequality,*

$$n < \frac{\left[(1-\delta)\cdot\log\left(\binom{p}{2} - 1\right) - \log 2\right]}{2\lambda\tanh^{-1}(\lambda)} \tag{3.74}$$

Proof. Upon realizing that eq. (3.63) can be obtained quite immediately by simply rearranging terms (and that the upper bound on the mutual information can be substituted into eq. (3.41)), this result simply becomes a matter of substitution. Since in this case, $S\left(\mathbb{P}_{\boldsymbol{\theta}_i} \,||\, \mathbb{P}_{\boldsymbol{\theta}_j}\right)$ is in fact independent of i and j, there is significant cancellation. The result follows quickly. □

4

PARALLEL ESTIMATION ON MARKOV RANDOM FIELDS

Outline

Having at this point discussed and derived in detail some important characteristics of estimators of Markov random fields, particularly in the case of the Ising model, we turn our attention now to modeling such objects from data. This work is based off of the model proposed originally in [Y. Mizrahi and de Freitas, 2014], but investigates the parallel estimation of Markov random fields using pseudo-likelihood as well. In order to facilitate our discussion, however, we require a more detailed discussion of the structure of the probability distribution over the underlying graphical structure.

4.1 MODEL SPECIFICATION

For a graph G with vertex set $V = \{1, 2, ..., k\}$, consider the set of maximal cliques of G, which we denote by $\mathcal{C}$. For each c in $\mathcal{C}$ (with c a subset of V) let $\mathbf{x}_c = \{x_i\}_{i \in c}$. An equivalent way to think of $\mathbf{x}_c$ is a subvector of $\mathbf{x}$ with components from c. We consider probability distributions of the form,

$$p(\mathbf{x}; \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c; \boldsymbol{\theta}_c) \tag{4.1}$$

The term ψ_c is often referred to as the potential function for the clique c and is non-negative. Naturally the partition function under this framework, again denoted $Z(\boldsymbol{\theta})$ is just the sum over every $\mathbf{x}$ over the product $\prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c; \boldsymbol{\theta}_c)$, thereby guaranteeing that the cumulative distribution function is properly normalized.

Arising from statistical physics is a new notion of *energy*, given by the notation $Q[\mathbf{x}_c; \boldsymbol{\theta}_c]$. Typically the potential functions are constructed to be of the form $\psi_c(\mathbf{x}_c; \boldsymbol{\theta}_c) = \exp\{-Q[\mathbf{x}_c; \boldsymbol{\theta}_c]\}$, which yields a Gibbs distribution. We will in particular be interested in energy functions that are linear in the sense that $Q[\mathbf{x}_c; \boldsymbol{\theta}_c] = \boldsymbol{\theta}_c^T \phi(\mathbf{x}_c)$, where ϕ is a function of the data, commonly a vector of terms that correspond with the sufficient statistics. Usually ϕ will involve pairwise products as appeared earlier in this work.

We now use the notation $q(X|Y)$ to refer to the conditional probability distribution of X given a particular realization of Y. Furthermore, if $\mathbf{x}$ is a vector in $\mathbb{R}^d$, we use the notation $\mathbf{x}_{-i}$ to refer to all components of $\mathbf{x}$ *except* for the i^{th}. More generally, the notation $\mathbf{x}_{-I}$ where I is an index set refers to all the components of $\mathbf{x}$ *except* for those components indexed in I.

Conditional Distribution of Ising Models

We now use the notation $q(X_j|\mathbf{X}_{i,-j}; \boldsymbol{\theta})$ to refer to the conditional probability density function of the j^{th} random variable given all of the other random variables in the Markov random field. Fortunately, it is easy to compute q

because it can be shown that for Ising models, the full conditional distribution satisfies a logistic regression relationship of the form,

$$\log\left(\frac{q\left(X_j=1\middle|\mathbf{X}_{i,-j};\boldsymbol{\theta}\right)}{1-q\left(X_j=1\middle|\mathbf{X}_{i,-j};\boldsymbol{\theta}\right)}\right)=\sum_{i\in \mathrm{ne}_j} 2\theta_{ij}X_i \tag{4.2}$$

Here, ne_j refers to the set of neighbors of the j^{th} vertex. This should remind the reader of Equation (1.8).

Most difficulty in maximum likelihood approaches to estimation on Markov random fields comes from the complexity of the gradient of the log-likelihood function. In particular, the complexity is exponential in the number of vertices in the underlying graph. To overcome this unfortunate reality, the method of pseudo-likelihood was introduced in 1975 by Julian Besag [Besag, 1975]. For a Markov random field on a graph G with vertex set of cardinality $k=|V|$. The log pseudo-likelihood is written for n observations,

$$\mathcal{L}^{\text{pseudo-likelihood}}\left(\boldsymbol{\theta}\right)=\frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{k}\log q\left(x_{ij}|\mathbf{x}_{i,-j};\boldsymbol{\theta}\right) \tag{4.3}$$

Using the pseudo-likelihood is generally easier than direct computation of maximum likelihood estimator because it avoids evaluating the partition function. Indeed, as will be shown in the following example, it is true that unlike the complete likelihood (where a summation is taken over all the edges) pseudo-likelihood only considers neighboring edges in its formulation.

Selection of Step-Size in Gradient Ascent

In the style of traditional gradient ascent algorithms, maximization occurs by "stepping" in a particular direction that is believed to lead to an increase in the pseudo-likelihood. The size of this step is determined in our algorithm by using he strategy of reducing the step size by one-half. In particular, after choosing a direction of maximization, we begin with an initial step size of 1 and we thereafter reduce the step size by half after each iteration.

Pseudo-Likelihood for Simple Ising Model

Consider a graph G with vertex set $V=\{1,2,3\}$ and a a set of edge weights, $\boldsymbol{\theta}=\{\theta_{12},\theta_{23}\}\subset\mathbb{R}^2$. For a visualization of this network, refer to Figure 2.1. Let $\mathbf{X}=\{X_1,X_2,X_3\}$ be $\{-1,+1\}$-valued random variable with probability mass function given by,

$$p\left(\mathbf{X};\boldsymbol{\theta}\right) \quad = \quad \frac{1}{Z\left(\boldsymbol{\theta}\right)}\exp\left\{\theta_{12}X_1X_2\right\}\exp\left\{\theta_{23}X_2X_3\right\} \tag{4.4}$$

Here the partition function is given by,

$$Z\left(\boldsymbol{\theta}\right)=\sum_{\mathbf{X}\in\{-1,+1\}^3}\exp\left\{\theta_{12}X_1X_2\right\}\exp\left\{\theta_{23}X_2X_3\right\} \tag{4.5}$$

In this example, computing $Z\left(\boldsymbol{\theta}\right)$ only involves taking a summation over eight possible configurations of $\mathbf{X}$, but generally the partition function will be intractable for a large number of $\{-1,+1\}$-valued random variables.

Now say that n samples are drawn i.i.d. such that $\mathbf{x}_i \sim p$ for $i = 1, 2, \ldots, n$. Recalling the logistic relation that characterizes the conditional probability distribution of Ising models, we obtain the marginal probability that $X_1 = +1$ given $\{X_2, X_3\}$,

$$p_1(\boldsymbol{\theta}) = \mathbb{E}\left[\mathbf{1}\{X_1 = 1\} \,|\, \{X_2, X_3\}; \boldsymbol{\theta}\right] = \frac{1}{1 + \exp\{-2\theta_{12} X_2\}} \tag{4.6}$$

$$p_2(\boldsymbol{\theta}) = \mathbb{E}\left[\mathbf{1}\{X_2 = 1\} \,|\, \{X_1, X_3\}; \boldsymbol{\theta}\right] = \frac{1}{1 + \exp\{-2\theta_{12} X_1 - 2\theta_{23} X_3\}} \tag{4.7}$$

$$p_3(\boldsymbol{\theta}) = \mathbb{E}\left[\mathbf{1}\{X_3 = 1\} \,|\, \{X_1, X_2\}; \boldsymbol{\theta}\right] = \frac{1}{1 + \exp\{-2\theta_{23} X_2\}} \tag{4.8}$$

Hence, $q(X_1|\{X_2, X_3\}) = p_1^{\mathbf{1}\{X_1=1\}} (1 - p_1)^{1-\mathbf{1}\{X_1=1\}}$ and similarly for $q(X_2|\{X_1, X_3\})$ and $q(X_3|\{X_1, X_2\})$. The conditional probability distribution of the remaining random variables X_2 and X_3 assume very similar forms. It can now be checked that the pseudo-likelihood function assumes the form,

$$\begin{aligned}\mathcal{L}^{\text{pseudo-likelihood}}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^{n} &\log\left(p_1(\boldsymbol{\theta})^{\mathbf{1}\{x_{i1}=1\}} (1 - p_1(\boldsymbol{\theta}))^{1-\mathbf{1}\{x_{i1}=1\}}\right) + \\ &\log\left(p_2(\boldsymbol{\theta})^{\mathbf{1}\{x_{i2}=1\}} (1 - p_2(\boldsymbol{\theta}))^{1-\mathbf{1}\{x_{i2}=1\}}\right) + \\ &\log\left(p_3(\boldsymbol{\theta})^{\mathbf{1}\{x_{i3}=1\}} (1 - p_3(\boldsymbol{\theta}))^{1-\mathbf{1}\{x_{i3}=1\}}\right).\end{aligned} \tag{4.9}$$

It is the practice to maximize this pseudo-likelihood via a gradient ascent algorithm. As stated in [Y. Mizrahi and de Freitas, 2014], in the case where we are dealing exclusively with binary Markov random fields (for instance, the Ising model), the gradient of the pseudo-likelihood can be expressed in a contrastive form. Consider a particular clique c of the underlying graphical structure. Then the gradient of the log pseudo-likelihood simplifies,

$$\frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_c} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} q\left(\bar{x}_{ij}^{(j)} | \mathbf{x}_{i,-j}; \boldsymbol{\theta}\right) \left(\phi_c(\mathbf{x}_i) - \phi_c\left(\bar{\mathbf{x}}_i^{(j)}\right)\right) \tag{4.10}$$

We denote by $\bar{\mathbf{x}}_i^{(j)}$ the i^{th} observation such that the j^{th} component is $-1 \cdot x_{ij}$.

Theorem 4.1 (Consistency of Pseudo-Likelihood)**.** *As the number of observations n of data generated by a Markov random field model goes to infinity, we have that with probability approaching one, $\boldsymbol{\theta}^\star$ optimizes both the likelihood and pseudo-likelihood functions.*

Proof. For a proof of this theorem, refer to the work of [Koller and Friedman, 2009].

□

4.1.1 *Advantages of Pseudo-Likelihood*

We present several arguments for the motivation behind using the pseudo-likelihood function as opposed to the complete likelihood.

(a) The likelihood (and thus the pseudo-likelihood) involves only complete summations over the observations $\{\mathbf{x}_i\}_{i=1}^n$. It is clear that this implies that the pseudo-likelihood is quite tractable.

(b) The pseudo-likelihood function is concave in the edge-weight parameter $\boldsymbol{\theta}$. Therefore, under the standard assumption of identifiability of the $\boldsymbol{\theta}$ parameter, there is a unique optimizer of the pseudo-likelihood and no danger of local minima.

(c) As was stated earlier in Theorem 4.1, the pseudo-likelihood is consistent in the sense that as the number of observations goes to infinity, the pseudo-likelihood and the full likelihood will have the same optimizer.

This being said, it is important to note that there are some disadvantages of pseudo-likelihood estimation. Namely, the pseudo-likelihood estimator is less efficient that the full maximum likelihood. This means that the pseudo-likelihood generally requires more samples to achieve the same estimation performance of the maximum likelihood estimator.

4.2 PARALLEL ESTIMATION WITH MAXIMUM PSEUDO-LIKELIHOOD

We now introduce a parallel numerical algorithm for estimating the edge-weight structure of the graph underlying a Markov random field using the pseudo-likelihood. The algorithm is presented below.

PARALLEL PSEUDO-LIKELIHOOD ESTIMATION

Algorithm 4.1. *Prediction of the weights of a Markov random field in a parallel manner.*

function PREDICTMAXIMUMLIKELIHOOD(A graph G)
  for Each clique c of G **do**
    Construct an auxilliary Markov random field $\mathcal{A}_c$ (defined below) from G.
    Solve the maximum pseudo-likelihood problem on $\mathcal{A}_c$.
    Set $\hat{\boldsymbol{\theta}}_c = \hat{\boldsymbol{\theta}}^{\text{pseudo-likelihood}}_{\mathcal{A}_c}$.
  end for
end function

We devote the remainder of this section to developing some theory that shows the consistency of creating auxiliary Markov random fields and solving smaller optimization problems. It is easy to show that the estimator computed via the parallel algorithm is consistent because its components are pseudo-likelihood estimates, which are themselves consistent.

Definition 4.1 (Normalized Gibbs Potential)**.** *The notion of a normalized potential arises originally from [Bremaud, 2001]. We say that a Gibbs potential $Q[\mathbf{x}_c; \boldsymbol{\theta}_c]$ is normalized (sometimes normalized with respect to zero) if $Q[\mathbf{x}_c|\boldsymbol{\theta}_c] = 0$ whenever it is the case that $\mathbf{x}_i = 0$ for any $i \in c$.*

NORMALIZED POTENTIALS IN COMMON MARKOV RANDOM FIELDS

Notice that in the case of the Gaussian Markov random field, any connected pairwise vertices clearly yield a potential that is normalized with respect to zero. This is easy to see by considering the pair of vertices $\mathbf{x}_i$ and $\mathbf{x}_j$ with connection weight $\theta_{ij} \neq 0$. Then the product $\theta_{ij}\mathbf{x}_i\mathbf{x}_j = 0 \iff$ either $\mathbf{x}_i = 0$ or $\mathbf{x}_j = 0$ (or both). In Ising models, the statement is vacuously true.

As pointed out in Mizrahi *et al.* [Y. Mizrahi and de Freitas, 2014], a perhaps under-appreciated result in the theory of Markov random fields is the existence and uniqueness result that derives from the normalized potential property. The statement and proof of this theorem is due again to [Bremaud, 2001].

Theorem 4.2 (Existence and Uniqueness for Normalized Potentials)**.** *There exists one and only one potential normalized with respect to zero corresponding to a Gibbs distribution.*

Definition 4.2 (1-Neighborhood)**.** *Consider a clique c of a graph $G = (V, E)$. We define the 1-neighborhood of c to be,*

$$\left(\bigcup_{v \in c} ne_v \right) \cup v \tag{4.11}$$

That is, the clique c itself and all those additional vertices which have at least one edge connecting it to a member of the clique.

Proposition 4.3. *Let $\boldsymbol{\theta}$ be the true edge-weight parameters corresponding to a Markov random field on k vertices. Consider a clique c of the underlying graph and suppose that we construct a new Markov random field by taking the subgraph of the 1-neighborhood of c. Denote the resulting graph $\mathcal{A}_c$. Let the set of cliques of $\mathcal{A}_c$ be denoted by $\mathcal{C}_c$.*

Suppose that the full distribution over the graph is parameterized by $\boldsymbol{\theta}$; that is, $p(\mathbf{x}; \boldsymbol{\theta})$ is the full distribution. Then denote by $p(\mathbf{x}_{\mathcal{A}_c}; \boldsymbol{\theta}')$ the marginal distribution of $\mathbf{x}$ restricted to $\mathcal{A}_c$, which is an auxiliary Markov random field parameterized by $\boldsymbol{\theta}'$. Then if both potentials are normalized with respect to zero (and assuming that parameters are identifiable) then $\boldsymbol{\theta}_c = \boldsymbol{\theta}'_c$.

Proof. Assume that all potentials presented in both the full and the auxiliary Markov random fields are normalized with respect to zero. Let $c \in \mathcal{C}$ be the clique under consideration. We seek to to estimate $\boldsymbol{\theta}_c$, the edge-weight parameters within the clique. It is easy to see that the marginal distribution on $\mathcal{A}_c$ with parameterization $\boldsymbol{\theta}_c{}'$ assumes the form,

$$p\left(\mathbf{x}_{\mathcal{A}_c}; \boldsymbol{\theta}'\right) = \frac{1}{Z\left(\boldsymbol{\theta}'\right)} \exp\left\{ - \sum_{v \in \mathcal{C}_c} Q\left[\mathbf{x}_v; \boldsymbol{\theta}'_v\right] \right\} \tag{4.12}$$

Recall that we use the notation $\mathcal{C}_c$ to refer to the cliques on the auxiliary Markov random field $\mathcal{A}_c$. The marginal given $\boldsymbol{\theta}$ (the parameters across the full Markov random field) is,

$$p\left(\mathbf{x}_{\mathcal{A}_c}; \boldsymbol{\theta}\right) = \sum_{\mathbf{x}_{-\mathcal{A}_c}} p\left(\mathbf{x}; \boldsymbol{\theta}\right) \tag{4.13}$$

$$= \frac{1}{Z\left(\boldsymbol{\theta}\right)} \exp\left\{ -Q\left[\mathbf{x}_c; \boldsymbol{\theta}_c\right] - \sum_{v \in \mathcal{C}_c - c} Q\left[\mathbf{x}_v; \boldsymbol{\theta}_{-c}\right] \right\} \tag{4.14}$$

By Theorem 4.2 and its implications for existence and uniqueness for normalized potentials, we know that the partition functions must be equal so that in particular the energy functions $Q\left(\mathbf{x}_c; \boldsymbol{\theta}_c\right)$ and $Q\left(\mathbf{x}_c; \boldsymbol{\theta}'_c\right)$ are equal. Under the assumption of identifiability, then, it therefore follows that $\boldsymbol{\theta}' = \boldsymbol{\theta}$. □

Much unlike the algorithms relying on the maximum likelihood to reconstruct graphs, leveraging the pseudo-likelihood gives more flexibility in dealing with densely connected 1-neighborhoods of cliques. In the next chapter we evaluate the efficacy of this approach through computational experiments and compare it with other state-of-the-art techniques.

4.3 PARALLEL ESTIMATION WITH VARIATIONAL MEAN-PARAMETERS

Here we prove that under a certain (very strong) assumption, there exists an embarrassingly parallel algorithm that will recover the original parameters of the Markov random field precisely.

Assumption 4.1. *Assume that the empirical mean-value on the marginal equals precisely the analytical $\boldsymbol{\mu}$ parameters. This is unlikely to hold in practice in the case of finite observations from the Markov random field, and is especially unlikely in cases of small observations. However, it is asymptotically true for infinite data.*

Lemma 4.4. *Consider an arbitrary clique c of a graph G. Let $\boldsymbol{\theta}$ be the edge-weight parameter for the complete graph G, while $\boldsymbol{\theta}'$ parameterizes the marginal over the auxiliary Markov random field given by $\mathcal{A}_c$. By the argument in Theorem 4.3 we saw that, in fact, $\boldsymbol{\theta}'_c = \boldsymbol{\theta}_c$ under certain conditions.*

Then the following also holds:

$$\boldsymbol{\theta}_c^{variational} = A^\star\left(\boldsymbol{\mu}_{\mathcal{A}_c}\right) = \sup_{\boldsymbol{\theta}\in\Omega} \boldsymbol{\theta}^T \boldsymbol{\mu}_{\mathcal{A}_c} - A\left(\boldsymbol{\theta}\right) = \boldsymbol{\theta}_c \tag{4.15}$$

Here, we denote by $A\left(\boldsymbol{\theta}\right)$ the log partition function; that is, $A\left(\boldsymbol{\theta}\right) = \log Z\left(\boldsymbol{\theta}\right)$. Furthermore it is clear that if $\hat{\boldsymbol{\mu}}_{\mathcal{A}_c} \to \boldsymbol{\mu}_{\mathcal{A}_c}$ then $\hat{\boldsymbol{\theta}}_c^{variational} = A^\star\left(\hat{\boldsymbol{\mu}}_{\mathcal{A}_c}\right) \to \boldsymbol{\theta}_c^{variational}$.

Proof. It is important to emphasize that $\boldsymbol{\theta}'_c = \boldsymbol{\theta}_c$. Indeed, we have that $\boldsymbol{\theta}'_c$ represents a valid parameterization of the marginal on $\mathcal{A}_c$. By the representation contained in Theorem 3.2, we see there is an injective map from the space of mean parameters to the space of edge weights. Therefore, we see the relation,

$$A^\star\left(\boldsymbol{\mu}_{\mathcal{A}_c}^{\boldsymbol{\theta}'}\right) = A^\star\left(\boldsymbol{\mu}_{\mathcal{A}_c}^{\boldsymbol{\theta}}\right) \iff \boldsymbol{\mu}_{\mathcal{A}_c}^{\boldsymbol{\theta}'} = \boldsymbol{\mu}_{\mathcal{A}_c}^{\boldsymbol{\theta}} \tag{4.16}$$

Since the edge weight parameters for the full Markov random field and the auxiliary one are equal, the implication holds that the mean-value parameters are equal as well. Therefore, if Assumption 4.1 holds, then solving the variational optimization problem recovers the original parameters exactly.

The subsequent result regarding the estimate of the mean parameters of $\mathcal{A}_c$ is easy to see as well. If the parameters converge directly to the true mean value, then clearly the analytical and empirical variational problems are equivalent, and so the optimal solutions are identical as well. □

PARALLEL VARIATIONAL ESTIMATION

Algorithm 4.2. *Prediction of the weights of a Markov random field in a parallel manner via variational mean-value estimation.*

1: **function** PREDICTVARIATIONALMEAN(A graph G)
2: **for** Each clique c of G **do**
3: Construct $\mathcal{A}_c$ densely from G.
4: Estimate the mean-value parameters $\boldsymbol{\mu}_{\mathcal{A}_c}$ for $\mathcal{A}_c$.
5: Solve the variational optimization problem,

$$\boldsymbol{\theta}^{\mathcal{A}_c} = \sup_{\boldsymbol{\theta}\in\Omega} \boldsymbol{\theta}^T \boldsymbol{\mu}_{\mathcal{A}_c} - A\left(\boldsymbol{\theta}\right) \tag{4.17}$$

6: Set $\boldsymbol{\theta}_c = \boldsymbol{\theta}_c^{\mathcal{A}_c}$.
7: **end for**
8: **end function**

Theorem 4.5. *The solution of the variational optimization problem when $\hat{\boldsymbol{\mu}}$, the empirical pairwise means, is used as an estimate of $\boldsymbol{\mu}$ is equivalent to the maximum likelihood estimator.*

Proof. Observe that variational optimization estimator solves the maximization problem,

$$\hat{\boldsymbol{\theta}} = \sup_{\boldsymbol{\theta} \in \Omega} \boldsymbol{\theta}^T \hat{\boldsymbol{\mu}} - \log Z(\boldsymbol{\theta}). \tag{4.18}$$

Generally for a Markov random field the likelihood function is given by (assuming i.i.d. observations),

$$\begin{aligned} \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}) &= \prod_{i=1}^{n} p(\mathbf{X}_i; \boldsymbol{\theta}) && (4.19) \\ &= \prod_{i=1}^{n} \frac{1}{Z(\boldsymbol{\theta})} \exp\left\{ \sum_{(j,k) \in E} \theta_{jk} X_{ij} X_{ik} \right\} && (4.20) \end{aligned}$$

Thus, the log-likelihood function is,

$$\begin{aligned} \log \mathcal{L}(\boldsymbol{\theta}; \mathbf{X}) &= \sum_{i=1}^{n} \left(\sum_{(j,k) \in E} \theta_{jk} X_{ij} X_{ik} - \log Z(\boldsymbol{\theta}) \right) && (4.21) \\ &= n \left(\sum_{(j,k) \in E} \theta_{jk} \hat{\mu}_{jk} \right) - n \log Z(\boldsymbol{\theta}) && (4.22) \end{aligned}$$

When maximizing the log-likelihood, constants may be discarded such that the objective function becomes,

$$\left(\sum_{(j,k) \in E} \theta_{jk} \hat{\mu}_{jk} \right) - \log Z(\boldsymbol{\theta}) = \boldsymbol{\theta}^T \hat{\boldsymbol{\mu}} - \log Z(\boldsymbol{\theta}) \tag{4.23}$$

Hence the variational objective function is equivalent to maximizing the log-likelihood. Thus, it is the case that the two solutions of the optimization problems must be equal. □

Corollary 4.6. *The variational estimator is asymptotically fully efficient. In particular, it achieves the Cramér-Rao bound in samples of arbitrarily large size.*

We will compare these embarrassingly parallel approaches to estimating the parameters of Markov random fields. We emphasize that unlike many of the competing algorithms used in graph estimation, these approaches enable distributed computing to occur. Importantly, when the auxilliary Markov random field is small, the variational optimization problem becomes tractable, so that it becomes useful in practice.

5 NUMERICAL EXPERIMENTS

Outline

In this chapter we describe and give results for some numerical experiments we performed to evaluate the "real-world" efficacy of the algorithms we presented in Chapter 4. We proceed with a discussion on algorithm performance and an brief introduction to our chosen error metric, the relative error. Finally, we conclude the chapter with a brief summary of the experiments and certain conclusions and implications yielded by this research.

To begin with, we measure error using the squared Euclidean distance metric. This is an alternative to the absolute error. To make this concept concrete, we give a mathematical definition.

Definition 5.1 (Mean Squared Error)**.** *Assume that the Ising graphical model has p vertices so that there are $\binom{p}{2}$ total edge weights in the graph. The error of a true parameter vector $\boldsymbol{\theta}$ from its estimate $\hat{\boldsymbol{\theta}}$ is given by the equation,*

$$MSE\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\right) = \mathbb{E}\left[\sum_{i=1}^{\binom{p}{2}} \left(\hat{\theta}_i - \theta_i\right)^2\right] \tag{5.1}$$

$$= \mathbb{E}\left[\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right] \tag{5.2}$$

Since the error is indeed a random variable, we do not compute it directly, but rather estimate it through simulations.

The error is useful for defining an loss function that penalizes large deviations from the truth.

5.1 DESCRIPTIONS OF EXPERIMENTS

For consistency, we emulate the experiments proposed in [Y. Mizrahi and de Freitas, 2014], which proposes to analyze algorithm performance on a variety of Markov random fields. For each experiment, we generate $\boldsymbol{\theta}$ at random from a standard normal distribution. In particular, our experiments use (a) a chain graph with twenty vertices, (b) a 2-dimensional Ising graph with sixteen vertices, and (c) a chimera 3×3 lattice graph, and (d) a $4 \times 4 \times 4$ Ising lattice. These graphs are large enough so that, for at least two of them, direct maximum likelihood estimation on the full graph is prohibitively expensive and for the remainder takes a long time. In each of our experiments, we initialize the starting point of the objective function ascent algorithm *at the true parameter values*. We motivate this choice by noting that the objective function in each case is convex so that, in an ideal world, the starting point is essentially irrelevant anyway.

The chain graph with twenty vertices has nineteen edge weight parameters to be estimated. The Ising grid with sixteen vertices has twenty-four edges to be estimated. The Ising lattice with sixty-four vertices has 144 edges to be estimated. Finally, the Ising chimera graph has 192 edge weights to be estimated.

Assumption 5.1. *When constructing the auxiliary Markov random field over the clique marginal, we assume that the structure is known exactly. That is, for all nodes in the 1-neighborhood (excluding those in the clique itself), we assume that these nodes and their connectivity to know another are known. This assumption is also made in the work of Mizrahi et al.*

In turn, this amounts to assuming that the marginal over the clique is parameterized in precisely the same way as in the complete Markov random field, without which the result contained in Theorem 4.2 could not have been applied. The precise representation presents the obstacle of potentially being too strong, but presents an advantage in the sense that the theory is well-supported.

As a point of comparison, we consider the performance of the two algorithms proposed in conjunction with the original variational optimization problem given in Proposition 3.2 (using estimates of $\boldsymbol{\mu}$) and the standard pseudo-likelihood approach proposed in [Besag, 1975]. The procedure for graph estimation among each of the models is essentially the same:

(a) We randomly generate edge weights by drawing from a normal distribution with mean zero and variance one. We then draw some number of samples approximately from the underlying probability mass function using Gibbs sampling.

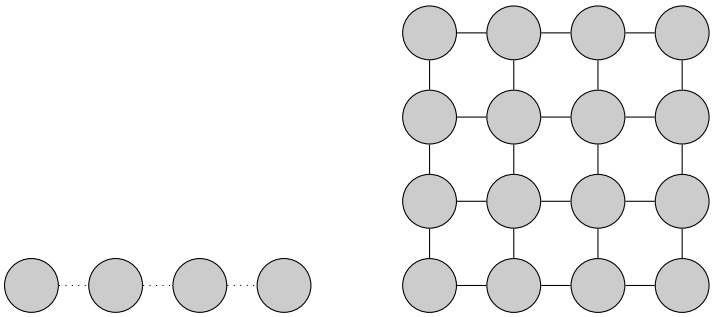
(b) We then construct approximations to $\boldsymbol{\theta}$ according to the model and compare to the remaining models using the relative error metric.

(c) We repeat these experiments one thousand times while varying the number of samples drawn approximately from the distribution. We report the average relative error as well as that quantity's variance.

5.2 RESULTS

The experiments we performed indicate that both the pseudo-likelihood and variational optimization over the marginal approaches are viable for parameter estimation on Markov random fields. Indeed, the results demonstrate that, with the exception of those experiments with very few samples, the differences in relative error are virtually undetectable between the pseudo-likelihood and parallel estimation strategies. In this case, it is important to emphasize the embarrassingly parallel algorithm, which enables the problem to be solved in a fraction of time. These results are consistent with the findings in [Y. Mizrahi and de Freitas, 2014].

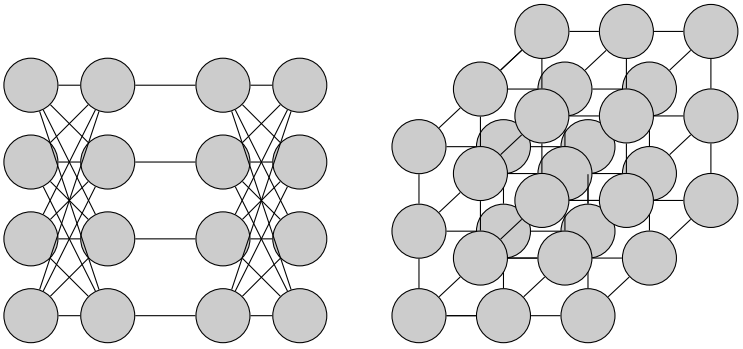
Interestingly, the variational optimization approach, when applied to the joint Markov random field, demonstrated exceptionally poor performance. Indeed, running a function optimization algorithm on MATLAB version 2014b failed to converge at all and consumed exceedingly large amounts of computation time for the two larger underlying graphical structures (the chimera and the three dimensional Ising graph). This proved to hold true even when the algorithm was lended a helping hand in the form of domain constraint bounds. Thus, this result could not be reported. This can almost doubtlessly be attributed to the complexity of the objective function, which is forced to grow exponentially in the number of vertices in the graph. This demonstrates that the variational problem, while theoretically exact, is simultaneously intractable for these larger computational problems.



(a) Shown above is a chain graph consisting of some number of nodes lined up one after the other. Notice that some elements of the chain are omitted for conservation of space, but the complete graph is analogous to this visualization.

(b) Here is presented a graph that is commonly know as an Ising grid. The name arises out of statistical physics, where each molecules are placed in a grid to demonstrate, though proximity, the electromagnetic relationships between them.

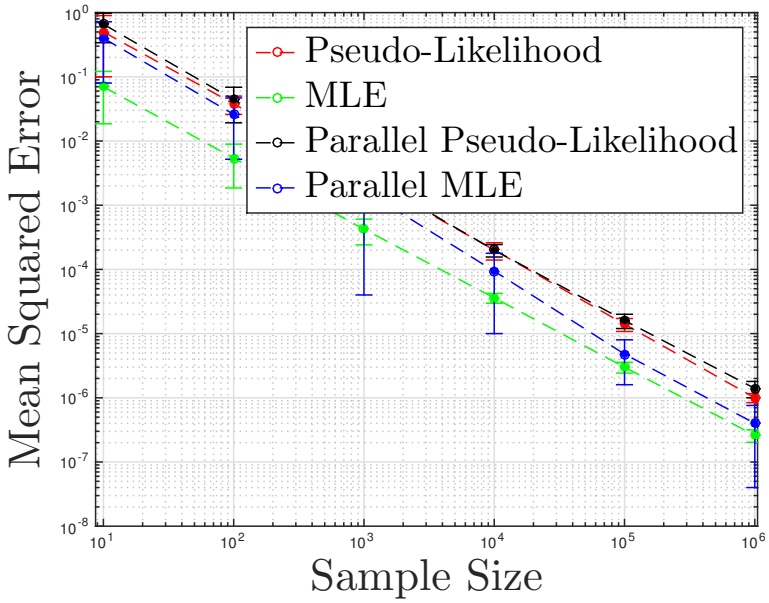
Figure 5.1: In order to aid in the proper understanding of the experiments proposed earlier, we show here some visual representations of the graphical models used in the experiments. To clarify, the purpose here is to associate with each edge appearing in the graph a real-valued weight, and then to estimate those weights from data drawn in an i.i.d. fashion from the Markov random field. To assign the edge weights in our experiments, we draw from a random variable distributed as $\mathcal{N}(0, 5)$. We justify this selection on the grounds that it provides a decent range of positive and negative edge weights, which simultaneously have the possibility of being near-zero or substantial in magnitude. In this figure we show an example of a chain graph with twenty nodes, and additionally a 2-dimensional Ising graph with sixteen nodes.



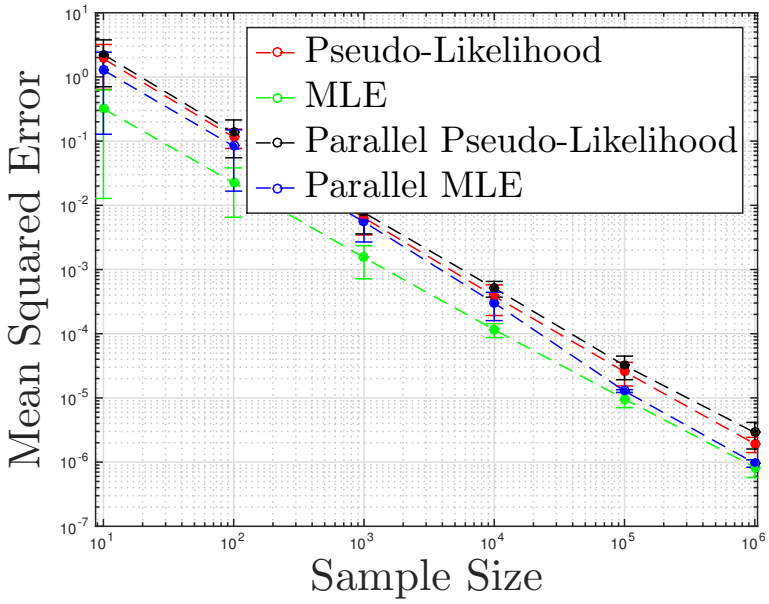
(a) A 1×2 chimera graph, which constitutes another popular underlying Markov random field. These graphs have applications in quantum mechanics.

(b) Shown above is an Ising graph in three dimensions. Interestingly, these models correspond to Fermionic lattice strings as shown by Alexander Polyakov.

Figure 5.2: As above, we present visual representations of the remaining two graphical models we selected for experimentation. In particular, we show a chimera $3 \times 3 \times 3$ lattice graph, and, lastly, a $4 \times 4 \times 4$ Ising lattice.

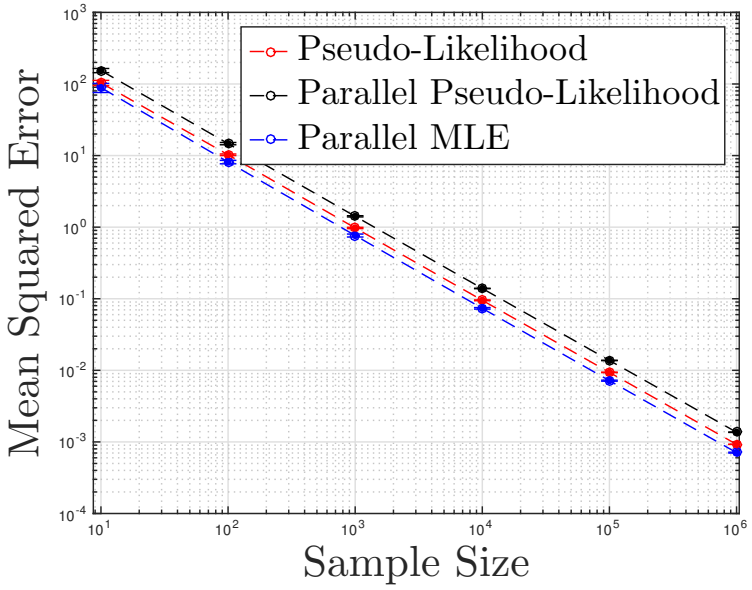


(a) 20-Chain Graph Error.

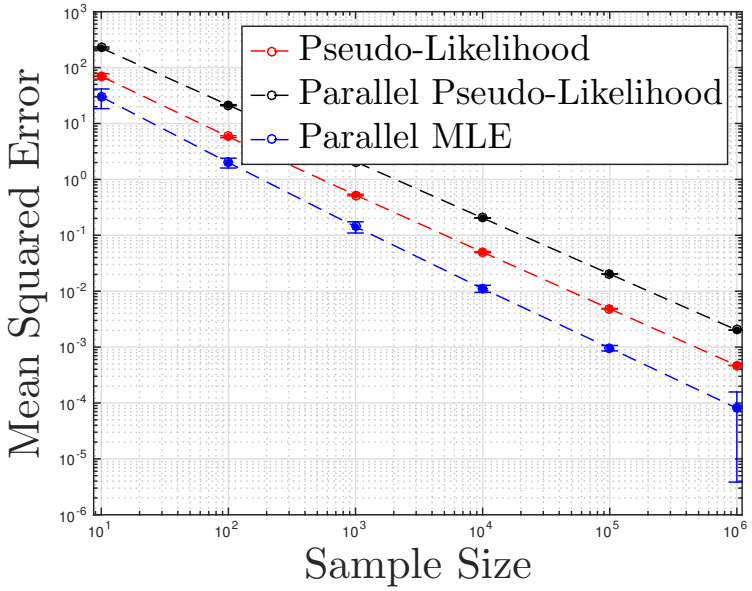


(b) Ising Grid Error.

Figure 5.3: Depicted here are the relative error rates of the varying approaches to learning Ising random fields from data. This graph shows performance on the chain graph and the lattice. Notice that, in general, the algorithms exhibit a trend of improving accuracy as a function of the number of observations drawn approximately from the underlying distribution. We depict using error bars the standard deviation of the estimators so as to give an idea of their consistency.

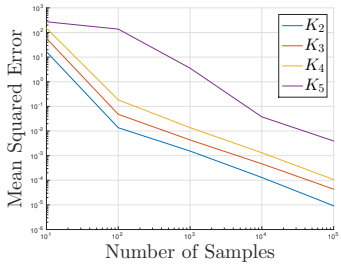


(a) Chimera Graph Error.



(b) Ising Lattice Graph Error.

Figure 5.4: This figure shows performance of the estimators on larger chimera and lattice graphs. The performance of direct maximum likelihood routine on larger graphs was poor and in fact did not converge, which we attribute below to the complexity of the objective function. Indeed, the partition functions for these graphs is intractable.



We show the mean squared error of the maximum likelihood estimator for varying sample sizes. For each sample size, we generate edge weights from a standard normal distribution for a complete graph K_m for $m \in \{2, 3, 4, 5\}$. We repeat this experiment $1,000$ times for each sample size and show the empirical mean squared error for every complete graph.

Figure 5.5: As a point of reference, we compute the empirical efficiency (as measured by the mean squared error) for the variational optimization estimator. Mean squared error for multivariate estimators is defined to be the quantity $\text{MSE}\left(\boldsymbol{\theta}, \hat{\boldsymbol{\theta}}\right) = \mathbb{E}\left[\left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)^T \left(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\right)\right]$. Since, as was shown in the previous chapter, variational inference is equivalent to maximum likelihood estimation, this estimator is asymptotically fully efficient.

5.3 CONCLUSIONS AND RECOMMENDATIONS FOR FUTURE RESEARCH

We have presented a distributed algorithm for estimating the parameters of a Markov random field by leveraging the uniqueness property of the marginal about a particular clique emerging from the graph. In particular, our variational optimization approach, because it is equivalent to maximum likelihood, enjoys some theoretical bounds on the error of the $\boldsymbol{\mu}$ parameter, a guarantee that is absent in other methods. Furthermore, we are able to construct some necessary conditions on maximum likelihood graph estimation using information theory. The methods we propose exhibit behavior that is consistent with both the current state-of-the-art approaches to parameter estimation in graphs.

A direction for future work is to extend the error guarantees on $\boldsymbol{\mu}$ through the variational problem to extend to $\boldsymbol{\theta}$ as well. This is a non-trivial problem due to the complexity of the log partition function. This would almost certainly require a very careful analysis of the behavior of partition functions, though it would be of great import to be able to bound the error of $\boldsymbol{\theta}$ according to the error of $\boldsymbol{\mu}$.

Additionally, though it is not investigated in this work, it would be desirable to evaluate the proposed parallel algorithms in the context of a real application. This would, we believe, generally increase interest in the study of Markov random fields and related fields. This represents an avenue for immediate future research.

BIBLIOGRAPHY

J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3): 179–195, 1975. (Cited on pages 22 and 30.)

P. Bremaud. Markov chains: Gibbs fields, monte carlo simulation, and queues. 2001. (Cited on pages 24 and 25.)

R. Foygel and M. Drton. High-dimensional ising model selection with bayesian information criteria. 2014. (Cited on page 3.)

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT Press, 2009. (Cited on page 23.)

N. Santhanam and M. Wainwright. Information-theoretic limits of selecting binary graphical models in high dimensions. *IEEE Transactions on Information Theory*, 58(7):4117–4134, 2012. (Cited on pages 8 and 19.)

M. J. Wainwright and M. I. Jordan. *Graphical models, exponential families and variational inference*. Found. Trends Mach. Learn., 2008. (Cited on pages 12 and 13.)

M. D. Y. Mizrahi and N. de Freitas. Linear and parallel learning of markov random fields. *International Conference on Machine Learning*, 2014. (Cited on pages 21, 23, 25, 29, and 30.)