Linear and Parallel Learning of Markov Random Fields

by

Yariv Dror Mizrahi

A THESIS SUBMITTED IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF

Doctor of Philosophy

in

THE FACULTY OF GRADUATE AND POSTDOCTORAL STUDIES
(Mathematics)

The University of British Columbia
(Vancouver)

November 2014

Abstract

In this thesis, we introduce a new class of embarrassingly parallel parameter learning algorithms for Markov random fields (MRFs) with untied parameters, which are efficient for a large class of practical models. The algorithms parallelize naturally over cliques and, for graphs of bounded degree, have complexity that is linear in the number of cliques. We refer to these algorithms with the acronym LAP, which stands for Linear And Parallel. Unlike their competitors, the marginal versions of the proposed algorithms are fully parallel and for log-linear models they are also data efficient, requiring only the local sufficient statistics of the data to estimate parameters. LAP algorithms are ideal for parameter learning in big graphs and big data applications.

The correctness of the newly proposed algorithms relies heavily on the existence and uniqueness of the normalized potential representation of an MRF. We capitalize on this theoretical result to develop a new theory of correctness and consistency of LAP estimators corresponding to different local graph neighbourhoods. This theory also establishes a general condition on composite likelihood decompositions of MRFs that guarantees the global consistency of distributed estimators, provided the local estimators are consistent.

We introduce a conditional variant of LAP that enables us to attack parameter estimation of fully-observed models of arbitrary connectivity, including fully connected Boltzmann distributions. Once again, we show consistency for this distributed estimator, and relate it to distributed pseudo-likelihood estimators.

Finally, for linear and non-linear inverse problems with a sparse forward operator, we present a new algorithm, named *i*LAP, which decomposes the inverse problem into a set of smaller dimensional inverse problems that can be solved independently. This parallel estimation strategy is also memory efficient.

Preface

This dissertation is original, independent work by the author. A version of Chapter 3 was accepted to the International Conference on Machine Learning, 2014 (with Misha Denil and Nando de Freitas as co authors). A version of Chapter 4 was submitted to the Advances in Neural Information Processing Systems conference, 2014 (with Misha Denil and Nando de Freitas as co authors). Chapter 6 describes joint work with Fred Roosta and Nando de Freitas.

Table of Contents

List of Figures

Acknowledgments

I wish to thank Prof. Nando de Freitas, for his supervision, patience and support. I also thank Prof. Eldad Haber for his early stage supervision, and Prof. Joel Friedman and Prof. Luis Tenorio for commenting on an early version of this work. Special thanks to Fred Roosta for all his help and many useful discussions.

I thank my family for their love and support.

Chapter 1

Introduction

1.1 Motivation

Markov random fields (MRFs), also known as undirected probabilistic graphical models, are ubiquitous structured statistical models that have impacted a significantly large number of fields, including computer vision [Li, 2001, Szeliski et al., 2008], computational photography and graphics [Agarwala et al., 2004, Boykov and Veksler, 2006, Chen et al., 2008], computational neuroscience [Ackley et al., 1985, Hopfield, 1984], bio-informatics [Yanover et al., 2007], natural language processing [Lafferty et al., 2001, Galley, 2006, Sutton and McCallum, 2012] and statistical physics [Marinari et al., 1997, Braunstein et al., 2005]. As pointed out in MacKay [2003] and Wainwright and Jordan [2008] there are also many applications in classical statistics, constraint satisfaction and combinatorial optimization, error-correcting codes and epidemiology. Not surprisingly, many comprehensive treatments of this important topic have appeared in the last four decades; see for example [Kindermann and Snell, 1980, Lauritzen, 1996, Li, 2001, Bremaud, 2001, Koller and Friedman, 2009, Murphy, 2012].

Despite the huge success and impact of these models, fitting them to data via maximum likelihood (ML) is prohibitively expensive in most practical situations. Although the likelihood is typically convex in the parameters, each optimization step requires solving inference problems that in the worst case are ♯P-hard [Murphy, 2012]. As stated, in bold, in the authoritative book of Koller and Friedman [2009]: "*a full inference step is required at every iteration of the gradient ascent procedure. Because inference is almost always costly in time and space, the computational cost of parameter estimation in Markov networks is usually high, sometimes prohibitively so.*"

Ideally, we would like to be able to compute the maximum likelihood estimates as these are consistent and maximally asymptotic efficient [Fisher, 1922]. We remind the reader that an estimator is asymptotically consistent if it converges to the true parameters as the sample size goes to infinity. An asymptotically consistent estimator is maximally efficient if the variance in the estimated parameters attains the minimum possible value among all consistent estimators as the sample size goes to infinity. Of course, in many applications, we are interested in penalized maximum likelihood estimates. That is, the goal is to find the

maximum a posteriori (MAP) estimates after the addition of smoothness or sparsity priors. For presentation simplicity, we will focus the discussion ML estimates, as our results will follow straightforwardly for typical MAP estimates.

In many cases, maximum likelihood in these models is *data efficient* in the sense that the data term in the gradient can be easily precomputed, making its evaluation trivial during optimization. The main difficulty with maximum likelihood is that it is not *model efficient* since evaluating the gradient involves computing expectations over the model distribution. This requires evaluating a sum with exponentially many terms, which is intractable for even moderately sized models, as pointed out above.

If the MRF under study has low tree-width, then the junction-tree algorithm can be adopted as the inference engine [Lauritzen, 1996, Murphy, 2012]. However, for many MRFs of interest, such as square lattices, the complexity of inference with the junction-tree algorithm grows exponentially with the size of the grid. This is also a severe problem when deploying skip-chain conditional random fields (CRFs) in natural language processing applications, such as named entity recognition and co-reference resolution, and computer vision tasks, such as dense stereo reconstruction [Galley, 2006, Sutton and McCallum, 2012, Murphy, 2012, Bradley, 2013].

The computational difficulties associated with exact inference have motivated researchers to adopt alternative estimators even if these are not as statistically efficient as maximum likelihood. Examples of these estimators include ratio matching, score matching, stochastic maximum likelihood, contrastive divergence, composite likelihood and pseudolikelihood [Besag, 1975, Younes, 1989, Hinton, 2000, Hyvärinen, 2005, 2007, Marlin et al., 2010, Varin et al., 2011, Marlin and de Freitas, 2011, Swersky et al., 2011].

An important class of approximate methods for this problem are stochastic approximation techniques, which approximate the model term by drawing samples from the model distribution, typically via Markov chain Monte Carlo (MCMC). This simulation is costly and often many samples are required for accurate estimation. Moreover, in settings where the parameters or data must be distributed across many machines such simulation poses additional difficulties.

Another approach is to approximate the maximum likelihood objective with a factored alternative. The leading method in this area is pseudo-likelihood. In this approach the joint distribution over all variables in the MRF is replaced by a product of conditional distributions for each variable. Applying pseudo likelihood in a distributed setting is may difficult, because the conditional distributions share parameters. Several researchers have addressed this issue by proposing to approximate pseudo-likelihood by disjointly optimizing each conditional and combining the parameters using some form of averaging [Ravikumar et al., 2010, Wiesel and Hero III, 2012, Liu and Ihler, 2012]. Yet, as pointed out by its creator, Julian Besag, "*My own view is that the technique is really a creature of the 1970s and 1980s and I am surprised to see it recommended in the computer age*" [Besag, 2001].

In this thesis, we introduce a new approach to parameter estimation in MRFs with untied parameters, which avoids the model inefficiency of maximum likelihood for an important class of models while preserving its data efficiency. Moreover, the proposed algorithms are naturally parallel and can be implemented in a distributed setting without modification. The algorithms replace the joint maximum likelihood problem with

a collection of much smaller auxiliary maximum likelihood problems which can be solved independently.

We prove that if the auxiliary problems satisfy certain conditions, the relevant parameters in the auxiliary problems converge to the values of the true parameters in the joint model. The experiments show that good performance is achieved in this case and that good performance is still achieved when these conditions are not satisfied. Violating the conditions for convergence sacrifices theoretical guarantees in exchange for even further computational savings while maintaining good empirical performance.

Under a strong assumption, we prove that a proposed Linear And Parallel (LAP) algorithm is exactly equal to maximum likelihood on the full joint distribution. While not directly applicable, this result provides additional insight into why our approach is effective.

A method similar to the naive LAP algorithm was recently, and independently, introduced in the context of *Gaussian graphical models* by Meng et al. [2013]. In that paper, the authors consider local neighbourhoods of nodes, whereas we consider neighbourhoods of cliques, and they rely on a convex relaxation via the *Schur complement* to derive their algorithm for inverse covariance estimation. At the time of writing this thesis, the same authors have shown that the convergence rate to the true parameters with their method is comparable to centralized maximum likelihood estimation [Meng et al., 2014].

Although our work and that of Meng et al. arrive at distributed learning via different paths, and while theirs is restricted to (pair-wise) Gaussian graphical models, both works show that it is possible to capitalize on graph structures beyond low tree-width to design algorithms that are both data and model efficient and exhibit good empirical performance.

In this thesis, we also introduce the *Strong LAP Condition*, which characterises a large class of composite likelihood factorisations for which it is possible to obtain global consistency, provided the local estimators are consistent. This much stronger sufficiency condition enables us to construct linear and globally consistent distributed estimators for a much wider class of models than Mizrahi *et al.*, including fully-connected Boltzmann machines.

Using this framework, we also show how the asymptotic theory of Liu and Ihler [2012] applies more generally to distributed composite likelihood estimators. In particular, the Strong LAP Condition provides a sufficient condition to guarantee the validity of a core assumption made in the theory of Liu and Ihler, namely that each local estimate for the parameter of a clique is a consistent estimator of the corresponding clique parameter in the joint distribution. By applying the Strong LAP Condition to verify the assumption of Liu and Ihler, we are able to import their M-estimation results into the LAP framework directly, bridging the gap between LAP and consensus estimators. In particular we illustrate how LAP can be applied to sparse Gaussian graphical models and discrete tables.

Finally, this thesis also offer an alternative approach for solving inverse problems by introducing an efficient parallel algorithm, named *i*LAP, which appropriately divides the large problem into smaller subproblems of much lower dimension. This process of localization offers substantial advantages in terms of computational efficiency and memory allocation.

1.2 Contribution list

The main novel result of this thesis is a mathematical observation that leads to a parameter estimation technique, LAP, which can be used, with linear complexity, for many important graphs and parametric families. The algorithm uses a set of local, independent, and low-dimensional estimators.

LAP is a naturally parallel algorithm, with significantly reduced complexity, that uses data statistics for efficient online and distributed memory allocation. We prove the consistency of LAP and prove that (under some conditions) LAP is identical to ML.

We further develop the algorithm by proving a strong LAP theorem that significantly reduces its complexity. The strong LAP theorem establishes the relationship between the graph topology and the consistency of local estimators.

We introduce the conditional LAP (CLAP), which is applicable to a larger class of parametric distribution families and prove the consistency of CLAP. We link the LAP and CLAP results to the design of algorithms for distributed parameter estimation in MRFs by showing how the work of Liu and Ihler [2012] and LAP can both be seen as special cases of *distributed composite likelihood*. Casting these two works in a common framework allows us to transfer results between them, strengthening the results of both works.

We study sparse Gaussian graphical models and discrete tables, and demonstrate the proper usage of LAP for these models. We investigate the complexity versus accuracy trade-off for these models.

We present an efficient algorithm, *i*LAP, for solving large scale inverse problems with sparse forward operators. Using *i*LAP we reduce the large inverse problem into a set of local inverse problems of smaller dimension. This approach is naturally parallel and significantly reduces the memory requirements of each solver. We experiment with *i*LAP by applying it to an image de-blurring problem and compare the results to baseline estimators.

1.3 Thesis organization

Chapters 1 and 2 provide motivation and background material on the well established theory of MRFs. Chapter 3 introduces the LAP methodology. We define the first neighbourhood of a clique, introduce the first LAP algorithms and prove their consistency and relation to ML estimation.

Chapter 4 refines the result of Chapter 3. In particular it includes the strong LAP theorem, which relates the graph topology to our ability to obtain consistent global estimators from local estimators. In the same chapter, we introduce a conditional version of LAP (named CLAP) and relate the LAP algorithm to other estimators. In particular, we develop distributed composite likelihood estimators with emphasis on obtaining sufficient conditions for their consistency.

In Chapter 5 we explain in detail how to apply LAP to sparse Gaussian graphical models and discrete tables. We define higher order neighbourhoods as domains and compare the resulting trade-off of complexity versus accuracy.

In Chapter 6 we introduce *i*LAP, which is a LAP based method for solving inverse problems. Chapter 7 suggests directions for future research.

Chapter 2

Background

2.1 Definitions and notation

2.1.1 Random fields

A real-valued random field defined on a set S is a collection of random variables indexed by the elements in S: $\{x_t\}_{t\in S}$. Since we will only consider finite sets S, a random field is simply a $k \times 1$ random vector $\mathbf{x}$, where k is the number of elements in S. The correlation structure of $\mathbf{x}$ will be defined by the relationships among the elements in S. Thus, from now on we will assume that the random field is defined on $S = \{1, ..., K\}$.

If $A = \{i_1, ..., i_n\} \subset S$, we define $\mathbf{x}_A = (\mathbf{x}_{i_1}, .., \mathbf{x}_{i_n})$, but for simplicity we shall write $\mathbf{x}$ instead of $\mathbf{x}_S$. The random field is completely defined by the joint distribution function of $\mathbf{x}$. We will assume that this distribution has the probability density function (PDF) p. The marginal PDF corresponding to $\mathbf{x}_A$ can be obtained from p by integration:

$$p(\mathbf{x}_A) = \int p(\mathbf{x}) d\mathbf{x}_\mathbf{c}, \quad \mathbf{c} = \mathbf{S} \setminus \mathbf{A}. \tag{2.1}$$

2.1.2 Graphs, neighbourhoods and cliques

Let S index a discrete set of K sites, where a site may represent a point location in a Euclidian space, a pixel location in an image, etc. Assume that for each site index $i \in S$, there exists a corresponding subset of S, denoted as $\mathcal{N}_i$. We denote the collection of the corresponding subsets $\{\mathcal{N}_i\}$ as $\mathcal{N}$.

Definition 1. *$\mathcal{N} = \{\mathcal{N}_i\}$ is called a neighbourhood system if and only if*

1. $i \in \mathcal{N}_j \iff j \in \mathcal{N}_i \quad \forall i, j \in S$ *(mutual relationship)*
2. *A site is not a neighbour of itself:* $i \notin \mathcal{N}_i$

The finite set S and the neighbourhood system $\mathscr{N}$ can be described as an undirected graph, with the sites as the nodes and the neighbourhood relationships as the edges, $G = \{S, E\}$, where S is the discrete set of K nodes and E is a list of edges satisfying $\{i, j\} \in E \iff i \in \mathscr{N}_j$.

Definition 2. *A subset $c \subseteq S$ is a clique for neighbourhood system $\mathscr{N}$, if any two different i, j in c are neighbours. That is:*

$$\forall i, j \in c, \;\; i \neq j, \;\; i \in \mathscr{N}_j.$$

Clearly any subset of a clique, $c_s \subseteq c$, is itself a clique, and any singleton $\{i\}$ is a clique. In the graphical model representation, any fully connected subset is a clique.

Definition 3. *A clique c is called maximal if $c \cup \{i\}$ is not a clique for any $i \notin c$.*

2.1.3 Markov random fields

The random vector $\mathbf{x}$ is said to have the Markov property with respect to a neighbourhood system $\mathscr{N}$ if its marginal and conditional distributions are such that:

$$p(\mathbf{x}_i | \mathbf{x}_{S \setminus i}) = p(\mathbf{x}_i | \mathbf{x}_{\mathscr{N}_i}).$$

Definition 4. *A family of random variables $\mathbf{x}_1, ..., \mathbf{x}_K$ is called a Markov Random Field on S with respect to a neighbourhood system $\mathscr{N}$ if it satisfies the Markov property.*

2.1.4 MRFs and Gibbs distributions

A Gibbs random field (GRF) on S with respect to neighbourhood system $\mathscr{N}$, is a set of random variables with the joint density function of the form

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(-U(\mathbf{x})\right). \tag{2.2}$$

U is called the *energy function* and Z is a constant called *the partition function* which plays the role of the normalization factor. The constant Z can be calculated by integration over the high dimensional space:

$$Z = \int ... \int \exp\left(-U(\mathbf{x})\right) d\mathbf{x}. \tag{2.3}$$

The Hammersley-Clifford theorem [Hammersley and Clifford, 1971] establishes the equivalence between MRFs and GRFs. The theorem states that a probability distribution that has a positive mass or density satisfies the Markov property under neighbourhood system $\mathscr{N}$ if and only if it is a Gibbs random field, where the energy function U is a summation of functions called *potentials* over the set of cliques:

$$U(\mathbf{x}) = \sum_{c \in \mathscr{C}} E_c(\mathbf{x}_c). \tag{2.4}$$

$E_c(\mathbf{x}_c)$ is the *energy* or *clique potential* associated with the variables in clique c, and it depends only on the values at the site inside the clique c. $\mathscr{C}$ is the collection of all cliques.

Combining Equations (2.2) and (2.4) yields

$$p(\mathbf{x}) = \frac{1}{Z} \exp \sum_{c \in \mathscr{C}} E_c(\mathbf{x}_c). \tag{2.5}$$

From the above expression it is clear that p is determined by local characteristics. *The representation of p using the potentials in Equation* (2.5) *is not unique*, but there is a way to impose such uniqueness using the concept of *normalized potentials*.

Definition 5. *An energy function, $E_c(\mathbf{x}_c)$ (or potential), is said to be normalized with respect to the zero vector if $E_c(\mathbf{x}_c) = 0$ whenever there exists $t \in c$ such that $x_t = 0$.*

Theorem 1. (Uniqueness of normalized potential) *There exists one and only one normalized potential representation with respect to zero corresponding to a Gibbs distribution.*

Proof. See Bremaud [2001], pages 262-265. □

The normalized potential is also known as the *canonical potential* [Griffeath, 1976, Kindermann and Snell, 1980]. One can choose normalized potentials with respect to reference values other than the zero vector. However, in order to ease the notation we consider only the normalization with respect to the zero vector.

The uniqueness of the normalized potential representation enables us to talk about the uniqueness of cliques. In particular, each clique has one and only one normalized potential. Henceforth, this thesis will focus on clique potentials in normalized potential representation and, specifically, normalized with respect to the zero vector.

2.2 Model specification and objectives

We are interested in estimating the parameter vector θ of the positive distribution $p(\mathbf{x} \mid \theta) > 0$ that satisfies the Markov properties of an undirected graph G. That is, a distribution that can be represented as a Gibbs distribution:

$$p(\mathbf{x} \mid \theta) = \frac{1}{Z(\theta)} \exp(-\sum_c E(\mathbf{x}_c \mid \theta_c)), \tag{2.6}$$

where $Z(\theta)$ is the partition function:

$$Z(\theta) = \int \exp(-\sum_c E(\mathbf{x}_c \mid \theta_c)) dx. \tag{2.7}$$

We will assume that the energy terms $E(\mathbf{x}_c \,|\, \theta_c) \in \mathbb{R}$ are chosen so that the parameters are identifiable. That is,

$$\theta_1 \neq \theta_2 \iff E(\mathbf{x}_c \,|\, \theta_1) \neq E(\mathbf{x}_c \,|\, \theta_2). \tag{2.8}$$

When the energy is a linear function of the parameters,

$$E(\mathbf{x}_c \,|\, \theta_c) = -\theta_c^T \phi_c(\mathbf{x}_c),$$

where $\phi_c(\mathbf{x}_c)$ is a feature vector derived from the values of the variables $\mathbf{x}_c$, we will refer to the model as a *maximum entropy* representation or *log-linear* model (Wasserman [2004], Buchman et al. [2012], Murphy [2012]). The features in these models are also referred to as local sufficient statistics.

At this stage, we will make an additional remark regarding notation. We will use $\mathbf{x}$ to refer to the vector of all variables (nodes). When needed, we increase the precision in our notation by using S to denote the set of all variables and use $\mathbf{x}_S$ for the vector of all variables in the MRF. We restrict the symbols n and c so that $\mathbf{x}_n$ refers to the n-th observation of all the variables in the MRF, and $\mathbf{x}_c$ refers to the subset of variables associated with clique c. Finally x_{mn} refers to the n-th observation of node m.

2.2.1 Maximum Likelihood estimation

Maximum Likelihood (ML) is maybe the most natural principle for estimating θ. Denoted by $\widehat{\theta}_{ml}$, the maximum likelihood estimator is defined as the θ that maximizes the *likelihood function* $L(\cdot)$. Specifically, let $\mathbf{x}_1, .., \mathbf{x}_N$ be independent realizations of $p(\mathbf{x}|\theta)$, then the *likelihood function* L is given by

$$\mathscr{L}(\theta; \mathbf{x}_1, .., \mathbf{x}_N) = \prod_{n=1}^{N} p(\mathbf{x}_n|\theta). \tag{2.9}$$

There is (in general) no closed form solution for the maximum likelihood estimate of the parameters of an MRF, so gradient-based optimizers are needed.

Consider the fully-observed maximum entropy model

$$p(\mathbf{x} \,|\, \theta) = \frac{1}{Z(\theta)} \exp(\sum_c \theta_c^T \phi_c(\mathbf{x})). \tag{2.10}$$

The scaled log-likelihood is given by

$$\ell(\theta) = \frac{1}{N} \sum_{n=1}^{N} \log p(\mathbf{x}_n \,|\, \theta). \tag{2.11}$$

By substituing Equation 2.10 we get

$$\ell(\theta) = \frac{1}{N} \sum_{n=1}^{N} \left[\sum_c \theta_c^T \phi_c(\mathbf{x}_n) - \log Z(\theta) \right], \tag{2.12}$$

which is a convex function of θ. The derivative with respect to the parameters of a particular clique, q, is given by

$$\frac{\partial \ell}{\partial \theta_q} = \frac{1}{N}\sum_{n=1}^{N}\left[\phi_q(\mathbf{x}_n) - \frac{\partial \log Z(\theta)}{\partial \theta_q}\right] \quad , \tag{2.13}$$

where

$$\frac{\partial \log Z(\theta)}{\partial \theta_q} = \mathbb{E}\left[\phi_q(\mathbf{x}) \,|\, \theta\right] = \int \phi_q(\mathbf{x}) p(\mathbf{x} \,|\, \theta) dx. \tag{2.14}$$

Equation (2.14) is the expectation of the feature $\phi_q(\mathbf{x})$ over the model distribution.

The full derivative of the log-likelihood contrasts the model expectation against the expected value of the feature over the data,

$$\frac{\partial \ell}{\partial \theta_q} = \frac{1}{N}\sum_{n=1}^{N}\phi_q(\mathbf{x}_n) - \mathbb{E}\left[\phi_q(\mathbf{x}) \,|\, \theta\right] \quad . \tag{2.15}$$

At the optimum these two terms will be equal and the empirical distribution of the features will match the model predictions. Asymptotically, ML is the optimal estimator since it reaches the Cramer-Rao lower bound. Specifically, the central limit theorem tell us that the ML estimate $\hat{\theta}_{ML}$ converges to the true parameter vector θ_{true} at the following rate (which is the fastest possible asymptotic rate):

$$\lim_{N\to\infty}(\theta_{true} - \hat{\theta}_{ML}) = \mathcal{N}(0, \mathbf{I}^{-1}), \tag{2.16}$$

where N is the number of samples and $\mathbf{I}$ is the Fisher Information matrix defined by

$$I_{ql} = -\mathbb{E}\left[\left(\frac{\partial^2 \log p(\mathbf{x}|\theta)}{\partial \theta_q \partial \theta_l}\right)\right]. \tag{2.17}$$

This statistical optimality is not the only important aspect of learning. For many models of interest $\hat{\theta}_{ML}$ is intractable. We must therefore also consider the computational costs associated with learning. In this thesis, we will seek to design classes of algorithms that are both computationally and statistically efficient. This will not always be possible and in some cases we will have to present trade-offs between efficient computation and asymptotic statistical efficiency.

2.2.2 Maximum Pseudo-Likelihood estimation

To surmount the intractable problem of computing expectations over the model distribution, the popular pseudo-likelihood estimator considers a simpler factorized objective function,

$$\ell^{PL}(\theta) = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M}\log p(x_{mn} \,|\, \mathbf{x}_{-mn}, \theta) \tag{2.18}$$

where $\mathbf{x}_{-mn}$ denotes all the components of the n-th data vector, except for component m. (For models with sparse connectivity, we only need to condition on the neighbours of node m.) In the binary, log-linear case, the gradient of this objective can be expressed in contrastive form,

$$\frac{\partial \ell^{PL}}{\partial \theta_q} = \frac{1}{N} \sum_{n,m} p(\bar{x}_{mn}^m \,|\, \mathbf{x}_{-mn}, \theta) \left[\phi_q(\mathbf{x}_n) - \phi_q(\bar{\mathbf{x}}_n^m) \right] \quad ,$$

where $\bar{\mathbf{x}}_n^m$ is the data vector $\bar{\mathbf{x}}_n$ with the m-th bit flipped. That is, $\bar{x}_{mn}^i = 1 - x_{mn}$ if $i = m$ and x_{mn} otherwise (Marlin et al. [2010]). Pseudo-likelihood will appear in most chapters of this thesis, first as a baseline and later as an application of LAP when considering the distributed pseudo-likelihood setting.

Chapter 3

The Marginal LAP

In this chapter, we describe a parameter estimation algorithm named LAP. We prove that LAP is both model efficient (*i.e.*, estimating the parameters requires low computational complexity) and data efficient (*i.e.*, it is sufficient to have data statistics distributed in a network). In other words, LAP avoids the model inefficiency of maximum likelihood for an important class of models while preserving its data efficiency.

LAP is naturally parallel and can be implemented in a distributed setting without modification. LAP replaces the joint maximum likelihood problem with a (linear) collection of much smaller auxiliary maximum likelihood problems that can be solved independently.

We prove that if the auxiliary problems satisfy certain conditions, the relevant parameters in the auxiliary problems converge to the values of the true parameters in the joint model. Under a strong assumption, we also prove that LAP is exactly equal to maximum likelihood on the full joint distribution. While hard to verify in practice, this result provides additional insight into why the LAP approach is effective.

3.1 Model and data efficiency

There are two terms in the gradient of Equation 2.15. The first term is an empirical expectation,

$$\frac{1}{N}\sum_{n=1}^{N}\boldsymbol{\phi}_q(\mathbf{x}_n)$$

and depends only on the data. The value of this term for each clique can be pre-computed before parameter optimization begins, making this term of the gradient extremely cheap to evaluate during optimization.
The data term in the ML gradient is contrasted with an expectation over the model distribution,

$$\mathbb{E}\left[\boldsymbol{\phi}_q(\mathbf{x}) \,|\, \boldsymbol{\theta}\right],$$

which is a sum over exponentially many configurations. For large models this term is intractable.

We describe this situation by saying that ML estimation is *data efficient*, since the terms involving only the data can be computed efficiently. However, ML is not *model efficient*, since the model term in the

Algorithm 1 LAP

Input: MRF with maximal cliques $\mathscr{C}$
for $q \in \mathscr{C}$ **do**
  Construct auxiliary MRF over the variables in A_q.
  Estimate parameters $\hat{\alpha}^{ML}$ of auxiliary MRF.
  Set $\hat{\theta}_q \leftarrow \hat{\alpha}^{ML}_q$.
end for

gradient is intractable, and the difficulty in evaluating it is the primary motivation for the development of alternative objectives like pseudo-likelihood.

Pseudo-likelihood addresses the model inefficiency of ML by eliminating the model term from the gradient, which makes pseudo-likelihood model efficient. However, pseudo-likelihood is not data efficient, since computing the gradient requires access to the full conditional distributions $p(\bar{x}^m_{mn} \,|\, \mathbf{x}_{-mn}, \theta)$. Because of this the outer sum over data examples must be computed for each gradient evaluation. (Note that for binary models the full conditionals correspond to logistic regressions, so any advances in scaling logistic regression to massive models and datasets would be of use here.)

In the following section we introduce a Linear And Parallel (LAP) algorithm, which uses a particular decomposition of the graph to avoid the exponential cost in ML, but unlike pseudo-likelihood LAP is fully parallel and maintains the data efficiency of ML estimation. LAP is therefore both model and data efficient.

3.2 Algorithm description

The LAP algorithm operates by splitting the joint parameter estimation problem into several independent sub-problems which can be solved in parallel. Once the sub-problems have been solved, it combines the solutions to each sub-problem together into a solution to the full problem.

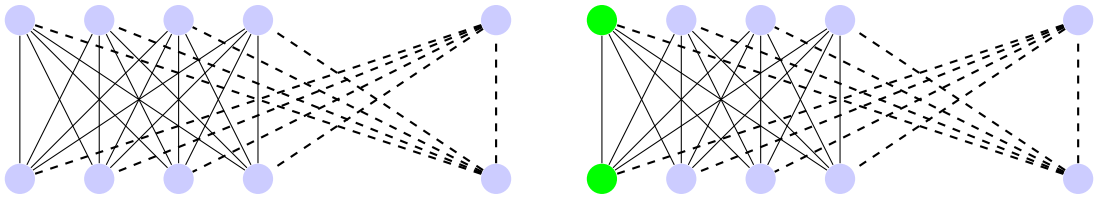
Definition 6. *For a fixed clique q we define its* 1-neighbourhood A_q *as the union of all cliques with non empty intersection with q:*

$$A_q = \bigcup_{c \cap q \neq \emptyset} c. \tag{3.1}$$

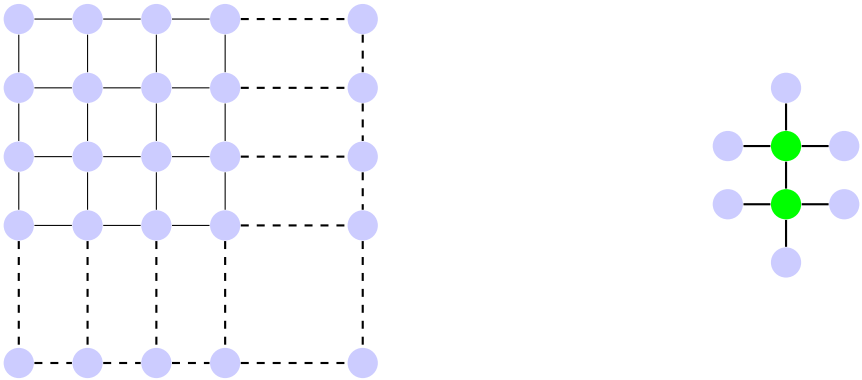
Alternatively, we say that A_q contains all of the variables of q itself as well as the variables with at least one neighbour in q (See proof in Appendix A.1). LAP creates one sub-problem for each maximal clique in the original problem by defining an *auxiliary MRF* over the variables in A_q. Details on how to construct the auxiliary MRF will be discussed later, for now we assume we have an auxiliary MRF on A_q and that it contains a clique over the variables in q that is parametrized the same way as q in the original problem.

LAP derives the parameter vector θ_q for the full problem by estimating parameters in the auxiliary MRF on A_q using maximum likelihood and reading off the parameters for the clique q directly. The steps of the algorithm are summarized in Algorithm 1.

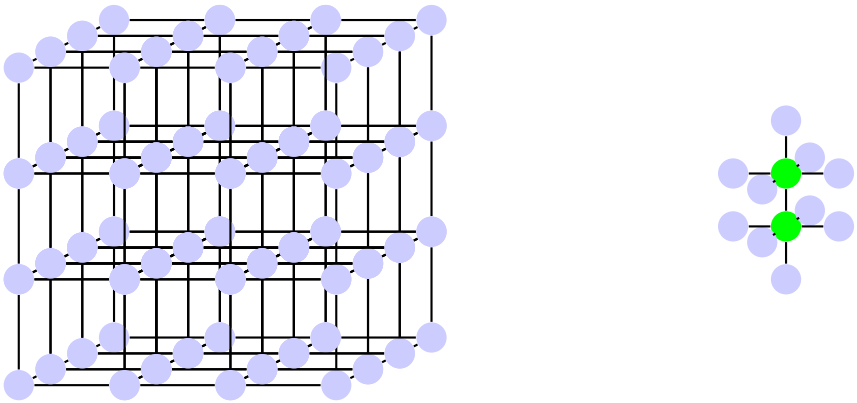
In a log-linear model, when estimating the vector of parameters α of the auxiliary MRF by maximum



(a)



(b)



(c)

Figure 3.1: The left column shows several popular MRFs: (a) a restricted Boltzmann machine (RBM), (b) a 2-D Ising model, and (c) a 3-D Ising model. The right hand side shows the corresponding 1-neighbourhoods. Models (b) and (c) have small 1-neighbourhoods compared to the full graph.

likelihood, the relevant derivative is

$$\frac{\partial \ell^{\mathscr{M}_q}}{\partial \alpha_q} = \frac{1}{N}\sum_{n=1}^{N} \phi_q(\mathbf{x}_{A_q n}) - \mathbb{E}\left[\phi_q(\mathbf{x}_{A_q})|\alpha\right] \quad . \tag{3.2}$$

This approach is data efficient, since the sufficient statistics $\frac{1}{N}\sum_{n=1}^{N}\phi_q(\mathbf{x}_{A_q n})$ can be easily pre-computed. Moreover, the data vector $\mathbf{x}_n$ can be stored in a distributed fashion, with the node estimating the auxiliary MRF only needing access to the sub-vector $\mathbf{x}_{A_q n}$. In addition, LAP is model efficient since the expectation $\mathbb{E}\left[\phi_q(\mathbf{x}_{A_q})|\alpha\right]$ can be easily computed when the number of variables in A_q is small. To illustrate this point, consider the models shown in Figure 3.1. For dense graphs, such as the restricted Boltzmann machine, the exponential cost of enumerating over all the variables in A_q is prohibitive. However, for other practical MRFs of interest, including lattices and Chimeras [Denil and de Freitas, 2011], this cost is acceptable.

3.2.1 Construction of the auxiliary MRF

The effectiveness of LAP comes from proper construction of the auxiliary MRF. As already mentioned, the auxiliary MRF must contain the clique q, which must be parametrized in the same way as in the joint model. This requirement is clear from the previous section, otherwise the final step in Algorithm 1 would be invalid.

We will see in the analysis section that it is desirable for the auxiliary MRF to be as close to the marginal distribution on $\mathbf{x}_{A_q}$ as possible. This means we must include all cliques from the original MRF which are subsets of A_q. Additionally, marginalization may introduce additional cliques not present in the original joint distribution. It is clear that these cliques can only involve variables in $A_q \setminus q$, but determining their exact structure in general can be difficult.

We consider three strategies for constructing auxiliary MRFs, which are distinguished by how they induce clique structures on $A_q \setminus q$. The three strategies are as follows.

Exact: Here we compute the exact structure of the marginal distribution over A_q from the original problem. We have chosen our test models to be ones where the marginal structure is readily computed.

Dense: For many classes of model the marginal over A_q involves a fully parametrized clique over $A_q \setminus q$ for nearly every choice of q (for example, this is the case in lattice models). The dense variant assumes that the marginal always has this structure. Making this choice will sometimes over-parametrize the marginal, but avoids the requirement of explicitly computing its structure.

Pairwise: Both the exact and dense strategies create high order terms in the auxiliary MRF. While high order terms do exist in the marginals of discrete MRFs, it is computationally inconvenient to include them, since the add many parameters to each sub-problem. In the pairwise variant we use the same graph structure as in dense, but here we introduce only unary and binary potentials over $A_q \setminus q$. This results in a significant computational savings for each sub-problem in LAP, but fails to capture the true marginal distribution in many cases (including all of the example problems we consider).

3.3 Experiments

In this section we describe some experiments designed to show that the LAP estimator has good empirical performance. We focus on small models where exact maximum likelihood is tractable in order to allow

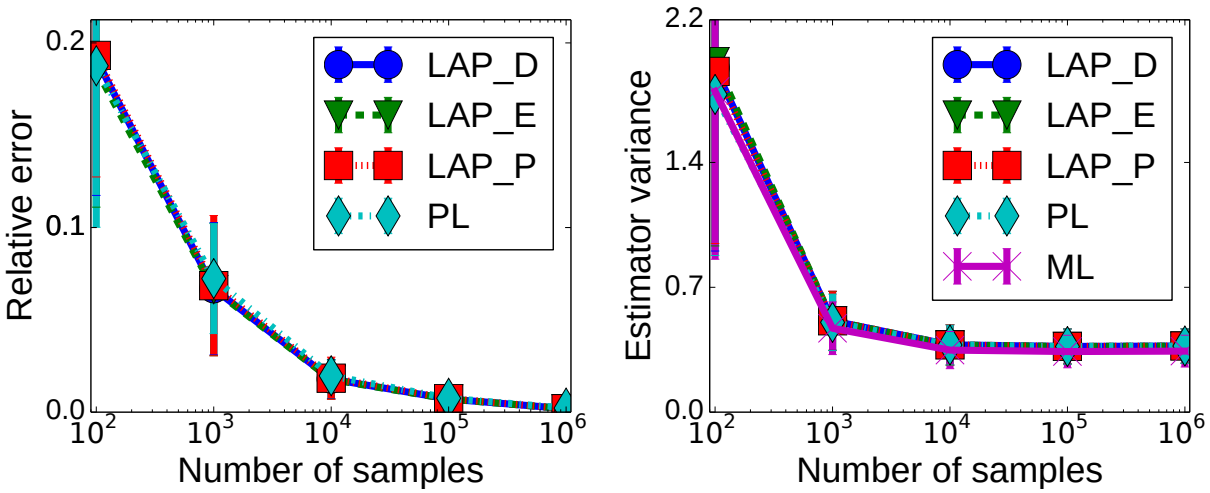


Figure 3.2: **Left:** Relative error of parameter estimates compared to maximum likelihood for LAP and pseudo-likelihood on a 4×4 Ising grid. Error bars show the standard deviation over several runs. **Right:** Variance of the parameter estimates for each algorithm.

performance to be measured. We chose to focus our experiments on demonstrating accuracy rather than scalability since the scaling and data efficiency properties of LAP are obvious.

The purpose of the experiments in this section is to show two things:

1. The accuracy of LAP estimates is not worse than its main competitor, pseudo-likelihood; and

2. LAP achieves good performance even when the exact marginal structure is not used.

In all of our experiments we compare pseudo-likelihood estimation against LAP using the three different strategies for constructing the auxiliary MRF discussed in the previous section. In each plot, lines labeled `PL` correspond to pseudo-likelihood and `ML` corresponds to maximum likelihood. `LAP_E`, `LAP_D` and `LAP_P` refer respectively to LAP with the exact, dense and pairwise strategies for constructing the auxiliary MRF.

We compare LAP and pseudo-likelihood to maximum likelihood estimation on three different model classes. The first is a 4×4 Ising grids with 4-neighbourhoods, and the results are shown in Figure 3.2. The second is a $4 \times 4 \times 4$ Ising lattice with 6-neighbourhoods, which is shown in Figure 3.3. Finally, we also consider a Chimera $3 \times 3 \times 3$ model, with results shown in Figure 3.4.

The procedure for all models is the same: we choose the generating parameters uniformly at random from the interval $[-1, 1]$ and draw samples approximately from the model. We then fit exact maximum likelihood parameters based on these samples, and compare the parameters obtained by pseudo-likelihood and LAP to the maximum likelihood estimates. The left plot in each figure shows the mean relative error of the parameter estimates using the maximum likelihood estimates as ground truth. Specifically, we measure

$$\mathrm{err}(\theta) = \|\theta^{ML}\|^{-1} \cdot \|\theta - \theta^{ML}\|$$

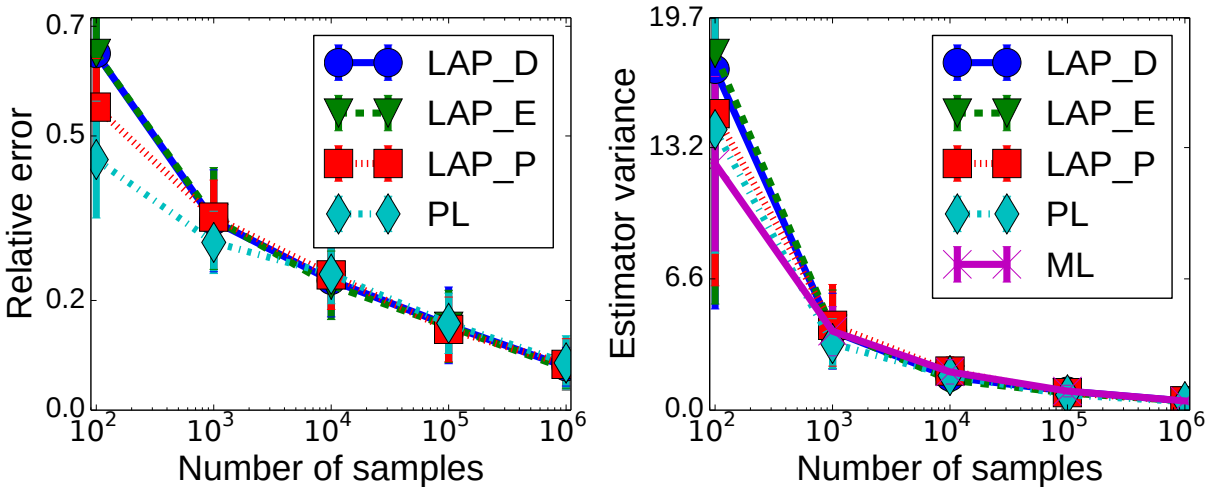


Figure 3.3: **Left:** Relative error of parameter estimates compared to maximum likelihood for LAP and pseudo-likelihood on a $4 \times 4 \times 4$ Ising lattice. Error bars show the standard deviation over several runs. **Right:** Variance of the parameter estimates for each algorithm.

for each estimate on each set of samples and average over several runs. We also measure the variance of the estimates produced by each algorithm over several runs. In this case we measure the variance of the estimates of each parameter separately and average these variances over all parameters in the model. These measurements are shown in the right plot in each figure. For reference we also show the variance of the maximum likelihood estimates in these plots.

In all of the experiments we see that the performance of all of the LAP variants is basically indistinguishable from pseudo-likelihood, except for small numbers of samples. Interestingly, `LAP_P` does not perform noticeably worse than the other LAP variants on any of the problems we considered here. This is interesting because `LAP_P` approximates the marginal with a pairwise MRF, which is not sufficient to capture the true marginal structure in any of our examples. `LAP_P` is also the most efficient LAP variant we tested, since the auxiliary MRFs it uses have the fewest number of parameters.

3.4 Theory

In this section show that matching parameters in the joint and the marginal distributions is valid, provided the parametrizations are chosen correctly. We then prove consistency of the LAP algorithm and illustrate its connection to ML.

Undirected probabilistic graphical models can be specified, locally, in terms of Markov properties and conditional independence and, globally, in terms of an energy function $\sum_c E(\mathbf{x}_c|\theta_c)$. As shown in Equations (2.2), (2.3), and (2.4). This is the direct outcome of the Hammersley-Clifford theorem [Hammersley and Clifford, 1971] which establishes the equivalence of these two representations.

One important fact that is often omitted is that the energy function and the partition function are not

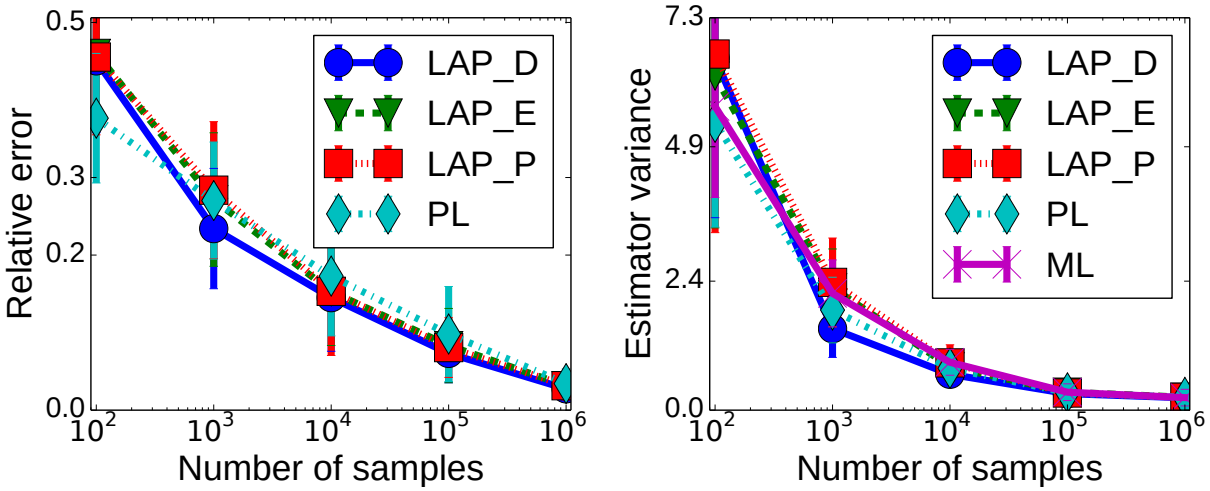


Figure 3.4: **Left:** Relative error of parameter estimates compared to ML for LAP and pseudo-likelihood on a Chimera $3 \times 3 \times 3$ model. Error bars show the standard deviation over several runs. **Right:** Variance of the parameter estimates for each algorithm.

unique. It is however possible to obtain uniqueness, for both of these functions, by imposing normalization with respect to a setting of the random variables of the potential. This gives rise to the concept of *normalized potential* [Bremaud, 2001]:

Definition 7. *A Gibbs potential* $\{E(\mathbf{x}_c|\theta_c)\}_{c\in\mathscr{C}}$ *is said to be normalized with respect to zero if* $E(\mathbf{x}_c|\theta_c) = 0$ *whenever there exists* $t \in c$ *such that* $\mathbf{x}_t = 0$.

(In this section, we use the term *Gibbs potential*, or simply *potential*, to refer to the energy so as to match the nomenclature of [Bremaud, 2001].) The following theorem plays a central role in understanding the LAP algorithm. The proof can be found in [Griffeath, 1976, Bremaud, 2001]:

Theorem 8. [Existence and Uniqueness of the normalized potential] *There exists one and only one (Gibbs) potential normalized with respect to zero corresponding to a Gibbs distribution.*

3.4.1 The LAP argument

Suppose we have a Gibbs distribution $p(\mathbf{x}_S \,|\, \theta)$ that factors according to the clique system $\mathscr{C}$, and let $q \in \mathscr{C}$ be a clique of interest. Let the auxiliary MRF

$$p(\mathbf{x}_{A_q} \,|\, \alpha) = \frac{1}{Z(\alpha)} \exp(-\sum_{c\in\mathscr{C}_q} E(\mathbf{x}_c \,|\, \alpha_c)) \tag{3.3}$$

have the same form as the marginal distribution on A_q (with clique system $\mathscr{C}_q$) parametrized so that the potentials are normalized with respect to zero. We can obtain the marginal from the joint in the following

way:

$$p(\mathbf{x}_{A_q} \mid \theta) = \int p(\mathbf{x}_S \mid \theta) d\mathbf{x}_{S \setminus A_q}. \tag{3.4}$$

Substituting Equation 3.3 into Equation 3.4, yields

$$p(\mathbf{x}_{A_q} \mid \theta) = \frac{1}{Z(\theta)} \int \exp(-\sum_{c \in \mathscr{C}} E(\mathbf{x}_c \mid \theta_c)) d\mathbf{x}_{S \setminus A_q} \tag{3.5}$$

Pulling out from the integral all the clique potentials with empty intersection with A_q yields:

$$p(\mathbf{x}_{A_q} \mid \theta) = \frac{1}{Z(\theta)} \exp(-\sum_{c \subseteq A_q} E(\mathbf{x}_c \mid \theta_c)) \int \exp(-\sum_{c \subsetneq A_q} E(\mathbf{x}_c \mid \theta_c)) d\mathbf{x}_{S \setminus A_q} \tag{3.6}$$

The domain of the function

$$\int \exp(-\sum_{c \subsetneq A_q} E(\mathbf{x}_c \mid \theta_c)) d\mathbf{x}_{S \setminus A_q}$$

has empty intersection with the clique of interest q (since A_q is defined as the union of all the cliques with non empty intersection).

Let us define

$$g(\mathbf{x}_{A_q \setminus q}) = \log(\int \exp(-\sum_{c \subsetneq A_q} E(\mathbf{x}_c \mid \theta_c)) d\mathbf{x}_{S \setminus A_q}).$$

With this definition, the marginal distribution over A_q is

$$p(\mathbf{x}_{A_q} \mid \theta) = \frac{1}{Z(\theta)} \exp(-\sum_{c \subseteq A_q} E(\mathbf{x}_c \mid \theta_c) + g(\mathbf{x}_{A_q \setminus q})). \tag{3.7}$$

The energy function in the Gibbs distribution of Equation 3.7 satsifies the sufficency condition in the Hammersley-Clifford theorem, that is $p(\mathbf{x}_{A_q} \mid \theta)$ is the PDF of a MRF on its own.

Proposition 9. *If the parametrisations of $p(\mathbf{x}_S \mid \theta)$ and $p(\mathbf{x}_{A_q} \mid \alpha)$ are chosen to be normalized with respect to zero, and if the parameters are identifiable with respect to the potentials, then $\theta_q = \alpha_q$.*

Proof. The terms $E(\mathbf{x}_q \mid \theta_q)$ and $E(\mathbf{x}_q \mid \alpha_q)$ appear as separate factors in $p(\mathbf{x}_{A_q} \mid \theta)$ in Equation 3.7 and in $p(\mathbf{x}_{A_q} \mid \alpha)$ in Equation 3.3 respectively. By existence and uniqueness of the normalized potentials (Theorem 8), we have

$$E(\mathbf{x}_q \mid \alpha_q) = E(\mathbf{x}_q \mid \theta_q) \tag{3.8}$$

which implies that $\theta_q = \alpha_q$ if the parameters are identifiable. □

3.4.2 Consistency of LAP

Let $\theta^\star$ be the true vector of parameters taken from the unknown generating distribution $p(\mathbf{x}_S \mid \theta^\star)$ parametrized such that the potentials are normalized with respect to zero. Suppose we have N samples drawn *iid* from this

distribution. Let $\hat{\theta}^{ML}$ be the ML estimate of θ given the data and let $\hat{\alpha}^{ML}$ the corresponding ML estimate for the auxiliary MRF with true parameters $\alpha^\star$.

Proposition 10. *If the true marginal distributions are contained in the class of auxiliary MRFs, we have* $\hat{\alpha}^{ML} \to \theta^\star$ *as* $N \to \infty$.

Proof. Let $q \in \mathscr{C}$ be an arbitrary clique of interest. It is sufficient to show that $\hat{\alpha}_q^{ML} \to \theta_q^\star$. By marginalization, we have

$$p(\mathbf{x}_{A_q} \,|\, \theta^\star) = \sum_{S \setminus A_q} p(\mathbf{x}_S \,|\, \theta^\star).$$

By the lap argument (Proposition 3), we know that $\alpha_q^\star = \theta_q^\star$. Since ML in consistent under smoothness and identifiability assumptions (for example, see [Fienberg and Rinaldo, 2012]), we also have

$$\hat{\alpha}^{ML} \to \alpha^\star,$$

so,

$$\hat{\alpha}_q^{ML} \to \theta_q^\star.$$

□

Note that in the above proposition, the class of auxiliary MRFs can be more general than the class of marginal MRFs, but must contain the latter. Asymptotically, superfluous terms in the auxiliary MRF vanish to zero.

3.4.3 Relationship to ML

Here we prove that, under certain (strong) assumptions in the discrete case, LAP is exactly equal to ML. The main result here will be that under the required assumptions, estimation by ML and marginalization commute. Suppose we have a discrete MRF on $\mathbf{x}_S$ which factorizes according to the cliques $\mathscr{C}$, and let $q \in \mathscr{C}$ be a particular clique of interest.

We will make use of the following characterization of ML estimates, which is proved in [Jordan, 2002].

Lemma 11. *If a distribution* $\hat{p}(\mathbf{x}_S)$ *satisfies that for each* $c \in \mathscr{C}$

$$\hat{p}(\mathbf{x}_c) = \tilde{p}(\mathbf{x}_c)$$

then $\hat{p}(\mathbf{x}_S)$ *is an ML estimate for the empirical distribution* $\tilde{p}(\mathbf{x}_S)$.

This characterization allows us to derive an explicit expression for an ML estimate of $\hat{p}(\mathbf{x}_S)$.

Proposition 12. *The distribution*

$$\hat{p}(\mathbf{x}_S) = \frac{\tilde{p}(\mathbf{x}_{A_q})\tilde{p}(\mathbf{x}_{S \setminus q})}{\tilde{p}(\mathbf{x}_{A_q \setminus q})}$$

is an ML estimate for $\tilde{p}(\mathbf{x}_S)$.

Proof. To see this we compute

$$\sum_q \hat{p}(\mathbf{x}_S) = \sum_q \frac{\tilde{p}(\mathbf{x}_{A_q})\tilde{p}(\mathbf{x}_{S\setminus q})}{\tilde{p}(\mathbf{x}_{A_q\setminus q})} = \tilde{p}(\mathbf{x}_{S\setminus q})$$

and

$$\sum_{S\setminus A_q} \hat{p}(\mathbf{x}_S) = \sum_{S\setminus A_q} \frac{\tilde{p}(\mathbf{x}_{A_q})\tilde{p}(\mathbf{x}_{S\setminus q})}{\tilde{p}(\mathbf{x}_{A_q\setminus q})} = \tilde{p}(\mathbf{x}_{A_q})$$

For an arbitrary clique $c \in \mathscr{C}$, either $c \subset S\setminus q$ or $c \subset A_q$, and we see that $\hat{f}(x_c) = \tilde{f}(x_c)$ by further marginalizing one of the above expressions. This shows that our expression for $\hat{p}(\mathbf{x}_S)$ satisfies the criteria of Lemma 11, and is therefore an ML estimate for $\tilde{p}(\mathbf{x}_S)$. □

Suppose we have a family of distributions $\mathscr{F}$ on $\mathbf{x}_S$ which satisfy the Markov properties of the MRF, and suppose that $\hat{p}(\mathbf{x}_S) \in \mathscr{F}$ where $\hat{p}(\mathbf{x}_S)$ is defined as in Proposition 12. Define the auxiliary family $\mathscr{F}_q$ associated with the clique q as follows.

$$\mathscr{F}_q = \{\sum_{S\setminus A_q} p(\mathbf{x}_S) \,|\, p(\mathbf{x}_S) \in \mathscr{F}\}$$

That is, $\mathscr{F}_q$ is the family of distributions obtained by marginalizing the family $\mathscr{F}$ over $S\setminus A_q$.

Proposition 13. *The auxiliary family* $\mathscr{F}_q$ *contains the marginal empirical distribution* $\tilde{p}(\mathbf{x}_{A_q})$. *Moreover* $\hat{p}(\mathbf{x}_{A_q}) = \tilde{p}(\mathbf{x}_{A_q})$ *is an ML estimate for* $\tilde{p}(\mathbf{x}_{A_q})$ *in* $\mathscr{F}_q$.

Proof. Recall that $\hat{p}(\mathbf{x}_S)$ from Proposition 12 is in $\mathscr{F}$ by assumption. Thus,

$$\sum_{S\setminus A_q} \hat{p}(\mathbf{x}_S) = \tilde{p}(\mathbf{x}_{A_q})$$

is in $\mathscr{F}_q$ by definition. That $\hat{p}(\mathbf{x}_{A_q}) \in \mathscr{F}_q$ is an ML estimate follows since the log likelihood gradient in Equation 2.15 is zero when the model and empirical distributions are equal. □

Suppose we can represent the family $\mathscr{F}$ as a Gibbs family, i.e.

$$\mathscr{F} = \mathscr{F}(\Theta) = \{p(\mathbf{x}_S \,|\, \theta) \,|\, \theta \in \Theta\}$$

for some domain of parameters Θ, where

$$p(\mathbf{x}_S|\theta) = \frac{1}{Z(\theta)} \exp(-\sum_{c\in\mathscr{C}} E(\mathbf{x}_c \,|\, \theta_c)) \ .$$

Moreover, suppose we have chosen this parametrisation so that the potential functions are normalized with respect to zero.

Since $\mathscr{F}$ is representable as a Gibbs family then the auxiliary family $\mathscr{F}_q$ is also representable as a Gibbs family with

$$\mathscr{F}_q = \mathscr{F}_q(\Psi) = \{p(\mathbf{x}_{A_q} \,|\, \alpha) \,|\, \alpha \in \Psi\}$$

for some domain of parameters Ψ. We will again suppose that this parametrization is chosen so that the potential functions are normalized with respect to zero.

We have already shown that ML estimates for $\tilde{p}(\mathbf{x}_S)$ and $\tilde{p}(\mathbf{x}_{A_q})$ exist in the families $\mathscr{F}$ and $\mathscr{F}_q$, respectively. Since we have chosen the parametrizations of these families to be normalized we also have unique ML parameters $\hat{\theta} \in \Theta$ and $\hat{\alpha} \in \Psi$ such that $p(\mathbf{x}_S \,|\, \hat{\theta}) \in \mathscr{F}(\Theta)$ is an ML estimate for $\tilde{p}(\mathbf{x}_S)$ and $p(\mathbf{x}_{A_q} \,|\, \hat{\alpha}) \in \mathscr{F}(\Psi)$ is an ML estimate for $\tilde{p}(\mathbf{x}_{A_q})$.

We can now prove the main result of this chapter.

Theorem 14. *Under the assumptions used in this section, estimating the joint parameters by ML and integrating the resulting ML distribution gives the same result as integrating the joint family of distributions and performing ML estimation in the marginal family. Concisely,*

$$\sum_{S \setminus A_q} p(\mathbf{x}_S \,|\, \hat{\theta}) = p(\mathbf{x}_{A_q} \,|\, \hat{\alpha})$$

Proof. We have the following sequence of equalities:

$$\begin{aligned} p(\mathbf{x}_S \,|\, \hat{\theta}) &\overset{(1)}{=} \hat{p}(\mathbf{x}_S) \\ &\overset{(2)}{=} \frac{\tilde{p}(\mathbf{x}_{A_q})\tilde{p}(\mathbf{x}_{S \setminus q})}{\tilde{p}(\mathbf{x}_{A_q \setminus q})} \\ &\overset{(3)}{=} \frac{\hat{p}(\mathbf{x}_{A_q})\tilde{p}(\mathbf{x}_{S \setminus q})}{\tilde{p}(\mathbf{x}_{A_q \setminus q})} \\ &\overset{(4)}{=} \frac{p(\mathbf{x}_{A_q} \,|\, \hat{\alpha})\tilde{p}(\mathbf{x}_{S \setminus q})}{\tilde{p}(\mathbf{x}_{A_q \setminus q})} \end{aligned}$$

The first equality follows from the parametrisation of $\mathscr{F}$, the second follows from Proposition 12, the third from Proposition 13 and the fourth follows from the parametrisation of $\mathscr{F}_q$. The theorem is proved by summing both sides of the equality over $S \setminus A_q$. □

Applying the LAP argument to Theorem 14 we see that $\hat{\theta}_q = \hat{\alpha}_q$.

Remark 15. *The assumption that $\hat{p}(\mathbf{x}_S) \in \mathscr{F}$ amounts to assuming that the* empirical *distribution of the data factors according to the MRF. This is very unlikely to hold in practice for finite data. However, if the true model structure is known then this property does hold in the limit of infinite data.*

Chapter 4

Strong LAP theorem, conditional LAP and distributed parameter estimation

In this chapter, we advanced the theoretical understanding of LAP in two directions and great practical consequence. First, we proved that it is possible to reduce the size of the clique 1-neighbourhoods used in the construction of auxiliary marginal MRFs in Chapter 3. Second, we extended the LAP argument to the conditional case, thereby enabling the methodology to become applicable to densely connected graphs.

Finally, we link it to the design of algorithms for distributed parameter estimation in MRFs by showing how the work of Liu and Ihler [2012] and Chapter 3 can both be seen as special cases of *distributed composite likelihood*. Casting these two works in a common framework allows us transfer results between them, strengthening the results of both works. It also appears that this thesis is the first work to address distributed composite likelihood estimators.

In Chapter 3 we introduced a theoretical result to show that it is possible to learn MRFs with untied parameters in a fully-parallel but globally consistent manner. That result lead to the construction of a globally consistent estimator, whose cost is linear in the number of cliques as opposed to exponential as in centralized maximum likelihood estimators. That result applies only to a specific factorization, with the cost of learning being exponential in the size of the factors. While these factors are small for lattice-MRFs and other models of low degree, they can be as large as the original graph for other models, such as fully-observed Boltzmann machines [Ackley et al., 1985]. In this chapter, we introduce the *Strong LAP Condition*, which characterizes a large class of composite likelihood factorizations for which it is possible to obtain global consistency, provided the local estimators are consistent. This much stronger sufficiency condition enables us to construct linear and globally consistent distributed estimators for a much wider class of models, including fully-connected Boltzmann machines.

Using this framework we also show how the asymptotic theory of Liu and Ihler applies more generally to distributed composite likelihood estimators. In particular, the Strong LAP Condition provides a sufficient condition to guarantee the validity of a core assumption made in the theory of Liu and Ihler, namely that each local estimate for the parameter of a clique is a consistent estimator of the corresponding clique parameter

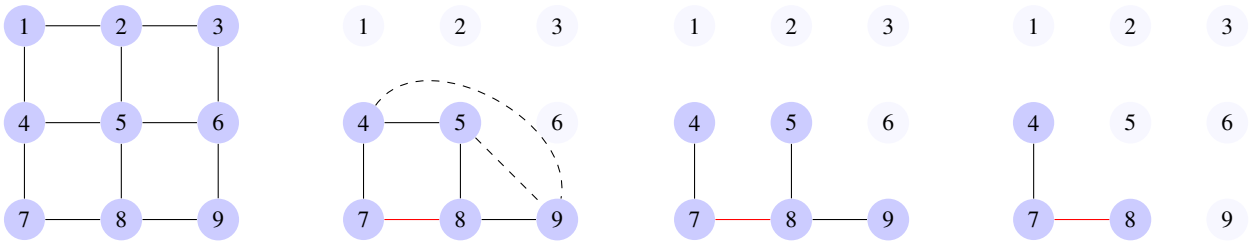


Figure 4.1: **Left:** A simple 2d-lattice MRF to illustrate our notation. For node $j = 7$ we have $\mathcal{N}(x_j) = \{x_4, x_8\}$. **Centre left:** The 1-neighbourhood of the clique $q = \{x_7, x_8\}$ including additional edges (dashed lines) present in the marginal over the 1-neighbourhood. Factors of this form are used by the LAP algorithm of Chapter 3 **Centre right:** The MRF used by our conditional estimator of Section 4.3 when using the same domain as Chapter 3 **Right:** A smaller neighbourhood which we show is also sufficient to estimate the clique parameter of q.

in the joint distribution. By applying the Strong LAP Condition to verify the assumption of Liu and Ihler, we are able to import their M-estimation results into the LAP framework directly, bridging the gap between LAP and consensus estimators.

4.1 Centralised estimation

Recall that our goal is to estimate the D-dimensional parameter vector θ of an MRF with the following *Gibbs density* or *mass function*:

$$p(\mathbf{x} \mid \theta) = \frac{1}{Z(\theta)} \exp(-\sum_{c} E(\mathbf{x}_c \mid \theta_c)). \tag{4.1}$$

As in previous chapters, $c \in \mathscr{C}$ is an index over the cliques of an undirected graph $G = (S, E)$, $E(\mathbf{x}_c \mid \theta_c)$ is the *energy* or *Gibbs potential*, and $Z(\theta)$ is a normalizing term known as the *partition function*.

The standard approach to parameter estimation in statistics is through maximum likelihood, which chooses parameters θ by maximizing

$$\mathscr{L}^{ML}(\theta) = \prod_{n=1}^{N} p(\mathbf{x}_n \mid \theta). \tag{4.2}$$

This estimator has played a central role in statistics as it is consistent, asymptotically normal, and efficient, among other desirable properties. However, applying maximum likelihood estimation to an MRF is generally intractable since computing the value of $\log \mathscr{L}^{ML}$ and its derivative require evaluating the partition function, or an expectation over the model respectively. Both of these values involve a sum over exponentially many terms.

To surmount this difficulty it is common to approximate $p(\mathbf{x} \mid \theta)$ as a product over more tractable terms.

This approach is known as *composite likelihood* and leads to an objective of the form

$$\mathscr{L}^{CL}(\theta) = \prod_{n=1}^{N} \prod_{i=1}^{I} f^i(\mathbf{x}_n, \theta^i) \tag{4.3}$$

where θ^i denote the (possibly shared) parameters of each composite likelihood factor f^i. Composite likelihood estimators are and both well studied and widely applied [Cox, 1988, Mardia et al., 2009, Liang and Jordan, 2008, Dillon and Lebanon, 2010, Marlin et al., 2010, Asuncion et al., 2010, Okabayashi et al., 2011, Bradley and Guestrin, 2012, Nowozin, 2013]. In practice the f^i terms are chosen to be easy to compute, and are typically local functions, depending only on some local region of the underlying graph $\mathscr{G}$.

An early and influential variant of composite likelihood is *pseudo-likelihood* (PL) [Besag, 1974], where $f^i(\mathbf{x}, \theta^i)$ is chosen to be the conditional distribution of x_i given its neighbours,

$$\mathscr{L}^{PL}(\theta) = \prod_{n=1}^{N} \prod_{m=1}^{M} p(x_{mn} \,|\, \mathbf{x}_{\mathscr{N}(x_m)n}, \theta^m) \tag{4.4}$$

Since the joint distribution has a Markov structure with respect to the graph $\mathscr{G}$, the conditional distribution for x_m depends only on its neighbours, namely $\mathbf{x}_{\mathscr{N}(x_m)}$. In general more efficient composite likelihood estimators can be obtained by blocking, *i.e.* choosing the $f^i(\mathbf{x}, \theta^i)$ to be conditional or marginal likelihoods over blocks of variables, which may be allowed to overlap.

Composite likelihood estimators are often divided into conditional and marginal variants, depending on whether the $f^i(\mathbf{x}, \theta^i)$ are formed from conditional or marginal likelihoods. In machine learning the conditional variant is quite popular [Liang and Jordan, 2008, Dillon and Lebanon, 2010, Marlin et al., 2010, Marlin and de Freitas, 2011, Bradley and Guestrin, 2012] while the marginal variant has received less attention. In statistics, both the marginal and conditional variants of composite likelihood are well studied (see the comprehensive review of Varin et al. [2011]).

An unfortunate difficulty with composite likelihood is that the estimators cannot be computed in parallel, since elements of θ are often shared between the different factors. For a fixed value of θ the terms of $\log \mathscr{L}^{CL}$ decouple over data and over blocks of the decomposition; however, if θ is not fixed then the terms remain coupled.

4.1.1 Consensus estimation

Seeking greater parallelism, researchers have investigated methods for decoupling the sub-problems in composite likelihood. This leads to the class of *consensus estimators*, which perform parameter estimation independently in each composite likelihood factor. This approach results in parameters that are shared between factors being estimated multiple times, and a final consensus step is required to force agreement between the solutions from separate sub-problems [Wiesel and Hero III, 2012, Liu and Ihler, 2012].

Centralized estimators enforce sub-problem agreement throughout the estimation process, requiring many rounds of communication in a distributed setting. Consensus estimators allow sub-problems to dis-

agree during optimization, enforcing agreement as a post-processing step which requires only a single round of communication.

Liu and Ihler [2012] approach distributed composite likelihood by optimizing each term separately

$$\hat{\theta}^i_{\beta_i} = \text{argmax}_{\theta_{\beta_i}} \left(\prod_{n=1}^{N} f^i(\mathbf{x}_{\mathscr{A}^i,n}, \theta_{\beta_i}) \right), \tag{4.5}$$

where $\mathscr{A}^i$ denotes the group of variables associated with block i, and θ_{β_i} is the corresponding set of parameters. In this setting the sets $\beta_i \subseteq \mathscr{V}$ are allowed to overlap, but the optimisations are carried out independently, so multiple estimates for overlapping parameters are obtained. Following Liu and Ihler we have used the notation $\theta^i = \theta_{\beta_i}$ to make this interdependence between factors explicit.

The analysis of this setting proceeds by embedding each local estimator $\hat{\theta}^i_{\beta_i}$ into a degenerate estimator $\hat{\theta}^i$ for the global parameter vector θ by setting $\hat{\theta}^i_c = 0$ for $c \notin \beta_i$. The degenerate estimators are combined into a single non-degenerate global estimate using different consensus operators, *e.g.* weighted averages of the $\hat{\theta}^i$.

The analysis of Liu and Ihler assumes that for each sub-problem i and for each $c \in \beta_i$

$$(\hat{\theta}^i_{\beta_i})_c \xrightarrow{p} \theta_c \tag{4.6}$$

i.e., that each local estimate for the parameter of clique c is a consistent estimator of the corresponding clique parameter in the joint distribution. This assumption does not hold in general, and one of the contributions of this chapter is to give a general condition under which this assumption holds.

The analysis of Liu and Ihler [2012] considers the case where the local estimators in Equation 4.5 are arbitrary M-estimators [van der Vaart, 1998], however their experiments address only the case of pseudo-likelihood. In Section 4.3 we prove that the factorization used by pseudo-likelihood satisfies Equation 4.6, explaining the good results in their experiments.

4.1.2 Distributed estimation

Consensus estimation dramatically increases the parallelism of composite likelihood estimates by relaxing the requirements on enforcing agreement between coupled sub-problems. In Chapter 3 we have shown that if the composite likelihood factorization is constructed correctly then consistent parameter estimates can be obtained without requiring a consensus step.

In the LAP algorithm described in Chapter 3, the domain of each composite likelihood factor (which is called the *auxiliary MRF*) is constructed by surrounding each maximal clique q with the variables in its *1-neighbourhood*

$$A_q = \bigcup_{c \cap q \neq \emptyset} c$$

which contains all of the variables of q itself as well as the variables with at least one neighbour in q; see Figure 4.1 for an example. For MRFs of low degree the sets A_q are small, and consequently maximum likelihood estimates for parameters of MRFs over these sets can be obtained efficiently. The parametric form of each factor in LAP is chosen to coincide with the marginal distribution over A_q. The factorisation of Chapter 3 is essentially the same as in Equation 4.5, but the domain of each term is carefully selected, and the LAP theorems are proved only for the case where

$$f^i(\mathbf{x}_{A_q}, \theta_{\beta_q}) = p(\mathbf{x}_{A_q}, \theta_{\beta_q}).$$

As in consensus estimation, parameter estimation in LAP is performed separately and in parallel for each term; however, agreement between sub-problems is handled differently. Instead of combining parameter estimates from different sub-problems, LAP designates a specific sub-problem as authoritative for each parameter (in particular the sub-problem with domain A_q is authoritative for the parameter θ_q). The global solution is constructed by collecting parameters from each sub-problem for which it is authoritative and discarding the rest.

4.2 Strong LAP argument

In this section we present the Strong LAP Condition, which provides a general condition under which the convergence of Equation 4.6 holds. This turns out to be intimately connected to the structure of the underlying graph.

Definition 16 (Relative Path Connectivity)**.** *Let $G = (S,E)$ be an undirected graph, and let A be a given subset of S. We say that two nodes $i, j \in A$ are* path connected with respect to $S \setminus A$ *if there exists a path $P = \{i, s_1, s_2, \ldots, s_n, j\} \neq \{i, j\}$ with none of the $s_k \in A$. Otherwise, we say that i, j are path disconnected with respect to $S \setminus A$.*

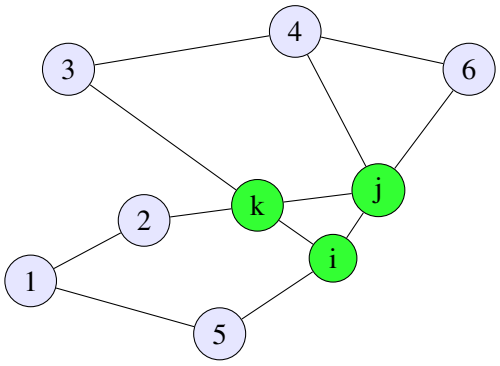


Figure 4.2: Illustrating the concept of relative path connectivity. Here, $A = \{i, j, k\}$. While (k, j) are path connected via $\{3,4\}$ and (k,i) are path connected via $\{2,1,5\}$, the pair (i, j) are path disconnected with respect to $S \setminus A$.

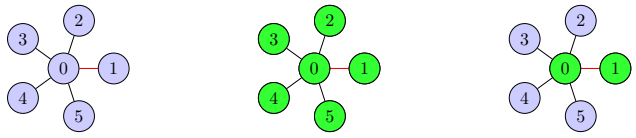


Figure 4.3: Figures (a)-(c) Illustrating the difference between LAP and Strong LAP. (a) Shows a star graph with q highlighted in red. (b) Shows A_q required by LAP. (c) Shows an alternative neighbourhood allowed by Strong LAP. Thus, if the root node is a response variable and the leafs are covariates, Strong LAP states we can estimate each parameter separately and consistently.

For a given $A \subseteq V$ we partition the clique system of G into two parts, $\mathscr{C}_A^{in}$ that contains all of the cliques that are a subset of A, and $\mathscr{C}_A^{out} = \mathscr{C} \setminus \mathscr{C}_A^{in}$ that contains the remaining cliques of G. Using this notation we can write the marginal distribution over $\mathbf{x}_A$ as

$$p(\mathbf{x}_A \,|\, \theta) = \frac{1}{Z(\theta)} \exp(-\sum_{c \in \mathscr{C}_A^{in}} E(\mathbf{x}_c \,|\, \theta_c)) \int \exp(-\sum_{c \in \mathscr{C}_A^{out}} E(\mathbf{x}_c \,|\, \theta_c)) d\mathbf{x}_{S \setminus A}. \tag{4.7}$$

Up to a normalization constant, $\int \exp(-\sum_{c \in \mathscr{C}_{\mathscr{A}}^{out}} E(\mathbf{x}_c \,|\, \theta_c)) d\mathbf{x}_{S \setminus A}$ induces a Gibbs density (and therefore an MRF) on A, which we refer to as the *induced MRF*.

For example, as illustrated in Figure 4.1 centre-left, the induced MRF involves all the cliques over the nodes 4, 5 and 9. By the Hammersley-Clifford theorem this MRF has a corresponding graph which we refer to as the *induced graph* and denote G_A. Note that the induced graph does not have the same structure as the marginal, it contains only edges which are created by summing over $\mathbf{x}_{S \setminus A}$.

Remark 17. *To work in the general case, we assume throughout that that if an MRF contains the path $\{i, j, k\}$ then integration over j creates the edge (i, k) in the marginal.*

In other words, if

$$g(x_i, x_k) = \int \exp(E_1(x_i, x_j)) \exp(E_1(x_j, x_k)) dx_j$$

it can not be factorized into functions of x_i and x_k

$$g(x_i, x_k) \neq f_1(x_i) f_2(x_k)$$

.

Proposition 18. *Let A be a subset of S, and let $i, j \in A$. The edge (i, j) exists in the induced graph G_A if and only if i and j are path connected with respect to $S \setminus A$.*

Proof. If i and j are path connected then there is a path $P = \{i, s_1, s_2, \ldots, s_n, j\} \neq \{i, j\}$ with none of the $s_k \in A$. Integrating over s_k forms an edge (s_{k-1}, s_{k+1}). By induction, integrating over $s_1, \ldots, s_n$ forms the edge (i, j).

If i and j are path disconnected with respect to $S \setminus A$ then integrating over any $s \in S \setminus A$ cannot form the edge (i, j) or i and j would be path connected through the path $\{i, s, j\}$. By induction, if the edge (i, j) is formed by integrating over $s_1, \ldots, s_n$ this implies that i and j are path connected via $\{i, s_1, \ldots, s_n, j\}$, contradicting the assumption. □

Corollary 19. *$B \subseteq A$ is a clique in the induced graph G_A if and only if all pairs of nodes in B are path connected with respect to $S \setminus A$.*

Definition 20 (Strong LAP condition)**.** *Let $G = (S, E)$ be an undirected graph and let $q \in \mathscr{C}$ be a clique of interest. We say that a set A such that $q \subseteq A \subseteq S$ satisfies the strong LAP condition for q if there exist $i, j \in q$ such that i and j are path-disconnected with respect to $S \setminus A$.*

Proposition 21. *Let $G = (S, E)$ be an undirected graph and let $q \in \mathscr{C}$ be a clique of interest. If A_q satisfies the strong LAP condition for q then the joint distribution $p(\mathbf{x}_S \mid \theta)$ and the marginal $p(\mathbf{x}_{A_q} \mid \theta)$ share the same normalized potential for q.*

Proof. If A_q satisfies the Strong LAP Condition for q then by Corollary 19 the induced MRF contains no potential for q. Inspection of Equation 4.7 reveals that the same $E(\mathbf{x}_q \mid \theta_q)$ appears as a potential in both the marginal and the joint distributions. The result follows by uniqueness of the normalized potential representation. □

We now restrict our attention to a set A_q which satisfies the Strong LAP Condition for a clique of interest q. The marginal over $p(\mathbf{x}_{A_q} \mid \theta)$ can be written as in Equation 4.7 in terms of θ, or in terms of auxiliary parameters α

$$p(\mathbf{x}_{A_q} \mid \alpha) = \frac{1}{Z(\alpha)} \exp(-\sum_{c \in \mathscr{C}_q} E(\mathbf{x}_c \mid \alpha_c)) \tag{4.8}$$

Where $\mathscr{C}_q$ is the clique system over the marginal. We will assume both parametrisations are normalised with respect to zero.

Theorem 22 (Strong LAP Argument)**.** *Let q be a clique in G and let $q \subseteq A_q \subseteq S$. Suppose $p(\mathbf{x}_S \mid \theta)$ and $p(\mathbf{x}_{A_q} \mid \theta)$ are parametrised so that their potentials are normalised with respect to zero and the parameters are identifiable with respect to the potentials. If A_q satisfies the Strong LAP Condition for q then $\theta_q = \alpha_q$.*

Proof. From Proposition 21 we know that $p(\mathbf{x}_S \mid \theta)$ and $p(\mathbf{x}_{A_q} \mid \theta)$ share the same clique potential for q. Alternatively we can write the marginal distribution as in Equation 4.8 in terms of auxiliary variables α. By uniqueness, both parametrizations must have the same normalized potentials. Since the potentials are equal, we can match terms between the two parametrizations. In particular since $E(\mathbf{x}_q \mid \theta_q) = E(\mathbf{x}_q \mid \alpha_q)$ we see that $\theta_q = \alpha_q$ by identifiability. □

Figure 4.3 shows the significant complexity reduction achived by the Strong LAP theorm. However, we note that the Strong LAP Condition is sufficient but not necessary. For example see Appendix A.2.

4.2.1 Efficiency and the choice of decomposition

Theorem 22 implies that distributed composite likelihood is consistent for a wide class of decompositions of the joint distribution; however it does not address the issue of statistical efficiency.

This question has been studied empirically in the work of Meng *et. al.* (Meng et al. [2013, 2014]), who introduce a distributed algorithm for Gaussian random fields and consider neighbourhoods of different sizes. Meng *et. al.* find the larger neighbourhoods produce better empirical results and the following theorem confirms this observation.

Theorem 23. *Let A be set of nodes which satisfies the Strong LAP Condition for q. Let $\hat{\theta}_A$ be the ML parameter estimate of the marginal over A. If B is a superset of A, and $\hat{\theta}_B$ is the ML parameter estimate of the marginal over B. Then (asymptotically):*

$$|\theta_q - (\hat{\theta}_B)_q| \leq |\theta_q - (\hat{\theta}_A)_q|.$$

Proof. Suppose that $|\theta_q - (\hat{\theta}_B)_q| > |\theta_q - (\hat{\theta}_A)_q|$. Then the estimates $\hat{\theta}_A$ over the various subsets A of B improve upon the ML estimates of the marginal on B. This contradicts the Cramer-Rao lower bound achieved the by the ML estimate of the marginal on B. □

In general the choice of decomposition implies a trade-off in computational and statistical efficiency. Larger factors are preferable from a statistical efficiency standpoint, but increase computation and decrease the degree of parallelism.

4.3 Conditional LAP

The Strong LAP Argument tells us that if we construct composite likelihood factors using marginal distributions over domains that satisfy the Strong LAP Condition then the LAP algorithm of Chapter 3 remains consistent. In this section we show that more can be achieved.

Once we have satisfied the Strong LAP Condition we know it is acceptable to match parameters between the joint distribution $p(\mathbf{x}_S, |\, \theta)$ and the auxiliary distribution $p(\mathbf{x}_{A_q}, |\, \alpha)$. To obtain a consistent LAP algorithm from this correspondence all that is required is to have a consistent estimate of α_q. In Chapter 3 w.b.ved this by applying maximum likelihood estimation to $p(\mathbf{x}_{A_q}, |\, \alpha)$, but any consistent estimator is valid.

We exploit this fact to show how the Strong LAP Argument can be applied to create a consistent conditional LAP algorithm, where conditional estimation is performed in each auxiliary MRF. This allows us to apply the LAP methodology to a broader class of models. For some models, such as large densely connected graphs, we cannot rely on the marginal LAP algorithm of Chapter 3. For example, for a restricted Boltzmann machine (RBM) (Smolensky [1986]), the 1-neighbourhood of any pairwise clique includes the entire graph. Hence, the complexity of LAP is exponential in the number of cliques. However, it is linear for conditional LAP, without sacrificing consistency.

Theorem 24. *Let q be a clique in G and let $x_j \in q \subseteq A_q \subseteq S$. If A_q satisfies the Strong LAP Condition for q then $p(\mathbf{x}_S|\,\theta)$ and $p(x_j\,|\,\mathbf{x}_{A_q\setminus\{x_j\}},\alpha)$ share the same normalised potential for q.*

Proof. We can write the conditional distribution of x_j given $A_q \setminus \{x_j\}$ as

$$p(x_j\,|\,\mathbf{x}_{A_q\setminus\{x_j\}},\theta) = \frac{p(\mathbf{x}_{A_q}\,|\,\theta)}{\int p(\mathbf{x}_{A_q},|\,\theta)dx_j} \tag{4.9}$$

Both the numerator and the denominator of Equation 4.9 are Gibbs distributions, and can therefore be expressed in terms of potentials over clique systems.

Since A_q satisfies the Strong LAP Condition for q we know that $p(\mathbf{x}_{A_q}|\,\theta)$ and $p(\mathbf{x}_S|\,\theta)$ have the same potential for q. Moreover, the domain of $\int p(\mathbf{x}_{A_q}|\,\theta)dx_j$ does not include q, so it cannot contain a potential for q. We conclude that the potential for q in $p(x_j|\,\mathbf{x}_{A_q\setminus\{x_j\}},\theta)$ must be shared with $p(\mathbf{x}_S|\,\theta)$. □

Remark 25. *There exists a Gibbs representation normalized with respect to zero for $p(x_j\,|\,\mathbf{x}_{A_q\setminus\{x_j\}},\theta)$. Moreover, the clique potential for q is unique in that representation.*

The existence in the above remark is an immediate result of the the existence of normalized representation both for the numerator and denominator of Equation 4.9, and the fact that difference of normalized potentials is a normalized potential. For uniqueness, first note that $p(\mathbf{x}_{A_q}|\,\theta) = p(x_j\,|\,\mathbf{x}_{A_q\setminus\{x_j\}},\theta)p(\mathbf{x}_{A_q\setminus\{x_j\}},\theta)$ The variable x_j is not part of $p(\mathbf{x}_{A_q\setminus\{x_j\}},\theta)$ and hence this distribution does not contain the clique q. Suppose there were two different normalized representations with respect to zero for the conditional $p(x_j\,|\,\mathbf{x}_{A_q\setminus\{x_j\}},\theta)$. This would then imply two normalised representations with respect to zero for the joint, which contradicts the fact that the joint has a unique normalized representation.

We can now proceed as in the original LAP construction from Chapter 3. For a clique of interest q we find a set A_q which satisfies the Strong LAP Condition for q. However, instead of creating an auxiliary parametrization of the marginal we create an auxiliary parametrization of the conditional in Equation 4.9.

$$p(x_j\,|\,\mathbf{x}_{A_q\setminus\{x_j\}},\alpha) = \frac{1}{Z_j(\alpha)}\exp(-\sum_{c\in\mathscr{C}_{A_q}} E(\mathbf{x}_c\,|\,\alpha_c)) \tag{4.10}$$

From Theorem 24 we know that $E(\mathbf{x}_q\,|\,\alpha_q) = E(\mathbf{x}_q\,|\,\theta_q)$. Equality of the parameters is also obtained, provided they are identifiable.

Corollary 26. *If A_q satisfies the Strong LAP Condition for q then any consistent estimator of α_q in $p(x_j\,|\,\mathbf{x}_{A_q\setminus\{x_j\}},\alpha)$ is also a consistent estimator of θ_q in $p(\mathbf{x}_S\,|\,\theta)$.*

The Conditional LAP enables the user to estimate the parameters when the parametric family of the marginal is unknown. However, if the parametric family is known, or even if an extended parametric family is known, the asymptotical estimation of CLAP will not be better then the asymptotical estimation of the marginal (for proof see Appendix A.3).

4.3.1 Connection to distributed Pseudo-Likelihood and composite likelihood

Theorem 24 tells us that if A_q satisfies the Strong LAP Condition for q then to estimate θ_q in $p(\mathbf{x}_S \mid \theta)$ it is sufficient to have an estimate of α_q in $p(x_j \mid \mathbf{x}_{A_q \setminus \{x_j\}}, \alpha)$ for any $x_j \in q$. This tells us that it is sufficient to use pseudo-likelihood-like conditional factors, provided that their domains satisfy the Strong LAP Condition. The following remark completes the connection by telling us that the Strong LAP Condition is satisfied by the specific domains used in the pseudo-likelihood factorisation.

Remark 27. *Let $q = \{x_1, x_2, .., x_m\}$ be a clique of interest, with 1-neighbourhood $\mathscr{F}_q = q \cup \{\mathscr{N}(x_i)\}_{x_i \in q}$. Then for any $x_j \in q$, the set $q \cup \mathscr{N}(x_j)$ satisfies the Strong LAP Condition for q. Moreover, $q \cup \mathscr{N}(x_j)$ satisfies the Strong LAP Condition for all cliques in the graph that contain x_j.*

Importantly, to estimate every unary clique potential, we need to visit each node in the graph. However, to estimate pairwise clique potentials, visiting all nodes is redundant because the parameters of each pairwise clique are estimated twice. This observation is important because it takes us back to the work of Liu and Ihler (Liu and Ihler [2012]). If a parameter is estimated more than once, it is reasonable from a statistical standpoint to apply the consensus operators of Liu and Ihler to obtain a single consensus estimate. The theory of Liu and Ihler tells us that the consensus estimates are consistent and asymptotically normal, provided Equation 4.6 is satisfied. In turn, the Strong LAP Condition guarantees the convergence of Equation 4.6.

The above remark tell us that the convergence in Equation 4.6 is satisfied for the distributed pseudo-likelihood setting of Liu and Ihler. We can go beyond this and consider either marginal or conditional factorisations over larger groups of variables. Since the asymptotic results of Liu and Ihler (Liu and Ihler [2012]) apply to any distributed composite likelihood estimator where the convergence of Equation 4.6 holds, it follows that any distributed composite likelihood estimator where each factor satisfies the Strong LAP Condition (including LAP and the conditional composite likelihood estimator from Section 4.3) immediately gains asymptotic normality and variance guarantees as a result of their work and ours.

Chapter 5

Applying LAP to Gaussian graphical models and discrete tables

In this chapter, we describe the application of LAP to sparse Gaussian graphical models. In particular, we apply the LAP algorithm to the problem of estimating the inverse covariance of a Gaussian distribution subject to conditional independence constraints, encoded as zeros in the inverse covariance. This is a problem that has attracted a great deal of attention in optimization, machine learning and statistics [Dempster, 1972, Songsiri et al., 2009, Ravikumar et al., 2010].

Subsequently, we discuss how to parametrize discrete probability functions which are given in table form, in order to apply LAP to these models.

While addressing these two popular models, we investigate the issue of trading-off complexity in favour of accuracy, by generalizing the first neighbourhood concept into higher orders. In addition, we discuss the memory allocation attributes of LAP.

5.1 Simple example: The Gaussian distribution

Let $p(\mathbf{x})$ be the density of a zero mean Gaussian distribution. In the Gaussian distribution, the non-zero clique functions are pairwise and unary, and Equations (2.2) and (2.4) take the form

$$p(\mathbf{x};\theta) = \frac{1}{Z}\sum_i\sum_j \theta_{ij}x_i x_j, \tag{5.1}$$

where $\theta_{ij} \neq 0$ if (i,j) is an edge. It is common to represent the PDF of Equation (5.1) using the matrix form

$$p(\mathbf{x}) = \frac{\det(\Sigma)^{-\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}}\exp(-\frac{1}{2}\mathbf{x}^T\Sigma^{-1}\mathbf{x}). \tag{5.2}$$

Where $\frac{\det(\Sigma)^{-\frac{1}{2}}}{(2\pi)^{\frac{d}{2}}}$ is the analytically known partition function (which may not be easy to calculate in practice).

The graph pattern is equivalent to the sparsity pattern of the SPD matrix Σ^{-1}. Σ^{-1} is known as the *precision matrix* or *inverse covariance* since it can be shown that the precision matrix is the inverse of the covariance matrix Σ. It is important to mention that every marginal PDF of a multivariate Gaussian is multivariate Gaussian on its own, with a smaller dimension.

Given a set of realizations $\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_N$ we denote by $\mathbf{D}$ the sample covariance matrix,

$$\mathbf{D} = \frac{1}{N} \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i{}^T .$$

Our goal is to estimate the Σ^{-1} using $\mathbf{D}$.

To this end, we first show that $p(\mathbf{x})$ is natively specified in the normalized potentials form. In particular, Equation (5.2) can be written as

$$p(\mathbf{x}) = \frac{1}{Z} \exp\left(\sum_{i,j} -\frac{\Sigma_{ij}^{-1} x_i x_j}{2} \right). \tag{5.3}$$

The non-zero clique potentials have either one or two variables. That is, the neighbourhood structure satisfies

$$\mathcal{N}_i = \{ j : \Sigma_{ij}^{-1} \neq 0 \}. \tag{5.4}$$

The energy function in Equation (5.3) can be re-written as

$$E(\mathbf{x}_{i,i} \,|\, \theta_{i,i}) = \Sigma_{ii}^{-1} \frac{x_i^2}{2}$$

and

$$E(\mathbf{x}_{i,j} \,|\, \theta_{i,j}) = \Sigma_{ij}^{-1} x_i x_j .$$

Both satisfy the normalized potential definition since

$$\{ x_i = 0 \; or \; x_j = 0 \} \Longrightarrow \Sigma_{ij}^{-1} \frac{x_i x_j}{2} = 0.$$

5.1.1 Gaussian example: Using the first neighbourhood

We demonstrate the computation of the parameters of a specific clique for the zero-mean Gaussian graphical model depicted in Figure 5.1. In this detailed example, we follow the estimation of $\Sigma_{9,10}^{-1}$, which is the parameter associated with the clique potential $\Sigma_{9,10}^{-1} x_9 x_{10}$. The first neighbourhood of the clique $\{9,10\}$ is $\{7,8,9,10\}$. $\{A_q \setminus q\}$ is $\{7,8\}$ and a new edge connecting $\{8,7\}$ was added to the auxiliary graph, as shown in Figure 5.4. Following the second step of the LAP algorithm we have to estimate the marginal PDF which

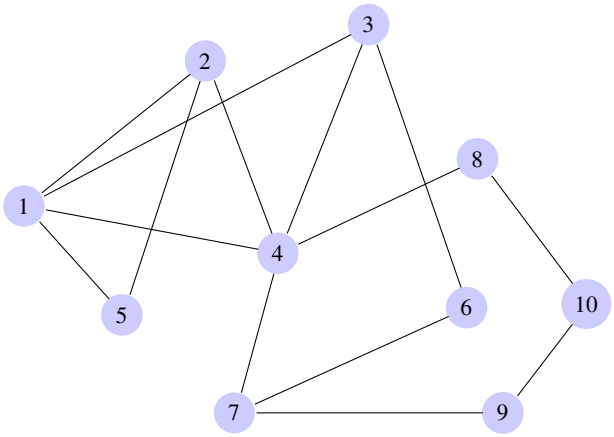


Figure 5.1: A simple sparsity pattern for a Gaussian graphical model. The neighbourhood system described in the graph is compatible with the precision matrix for the multivariate Gaussian precision matrix in Figure 5.2.

$\Sigma^{-1} =$

x	x	x	x	x					
x	x		x	x					
x		x	x		x				
x	x	x	x			x	x		
x	x			x					
		x			x	x			
			x		x	x		x	
			x				x		x
						x		x	x
							x	x	x

Figure 5.2: The precision matrix associated with the graph of Figure 5.1. The symbol $\times$ stands for non-zero entries in the precision matrix, and $\Sigma^{-1}(i,j) \neq 0 \iff (i,j) \in E$.

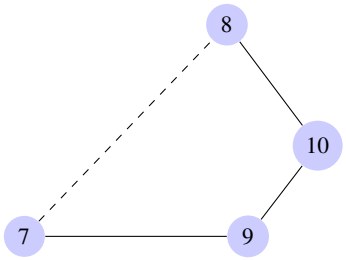


Figure 5.3: The first neighbourhood of the clique $\{9,10\}$ and the auxiliary graph.

is a Gaussian restricted to x_7, x_8, x_9, x_{10} and given by:

$$f_{A_{c_q}}(x_7, x_8, x_9, x_{10}) = \frac{|\Sigma_{marginal}{}^{-\frac{1}{2}}|}{(2\pi)^{\frac{4}{2}}} \exp(-\frac{1}{2}\mathbf{x}_{c_q}{}^T \Sigma^{-1}_{marginal}\mathbf{x}_{c_q}). \tag{5.5}$$

That is, we have to estimate the much smaller $\Sigma^{-1}_{marginal}$ with respect to graph in Figure 5.4 and the sample covariance matrix

$$\begin{bmatrix} D_{7,7} & D_{7,8} & D_{7,9} & D_{7,10} \\ D_{8,7} & D_{8,8} & D_{8,9} & D_{8,10} \\ D_{9,7} & D_{9,8} & D_{9,9} & D_{9,10} \\ D_{10,7} & D_{10,8} & D_{10,9} & D_{10,10} \end{bmatrix}, \tag{5.6}$$

which is a 4×4 projection matrix from the full covariance matrix. (Here, is use the term projection to refer to the process of extracting the relevant columns and rows from the original matrix.) We accomplish the procedure by substituting the value calculated in $\Sigma^{-1}_{marginal}(3,4)$ into $\Sigma_{-1}(9,10)$.

5.1.2 Gaussian example: Using sub-neighbourhood

In this section, we follow the estimation of the same parameter $\Sigma^{-1}_{9,10}$, associated with the energy function $\Sigma^{-1}_{9,10}x_9x_{10}$. However this time we use a domain smaller than the first neighbourhood.

There two valid alternatives for the choice of sub-neighbourhood for the clique $\{9,10\}$: either $\{8,9,10\}$ or $\{7,9,10\}$. Both are shown in Figure 5.1.2, with dashed lines representing the new edges added to the auxiliary graphs because of marginalization.

Here, $\{A_q \setminus q\}$ corresponds to $\{7,8\}$ and a new edge connecting $\{8,7\}$ was added to the auxiliary graph, as shown in Figure 5.4. The second step of the LAP algorithm is similar to the one presented in Section 5.1.1.

Let us, first, restrict our attention to the domain of x_8, x_9, x_{10}, with marginal:

$$f_{A_{c_q}}(x_8, x_9, x_{10}) = \frac{|\Sigma_{marginal}{}^{-\frac{1}{2}}|}{(2\pi)^{\frac{4}{2}}} \exp(-\frac{1}{2}\mathbf{x}_{c_q}{}^T \Sigma^{-1}_{marginal_{10}}\mathbf{x}_{c_q}). \tag{5.7}$$

The notation $marginal_{10}$ stands for the sub neighbourhood which includes

$$\{9,10\} \cup \mathcal{N}_{10}.$$

Note the domain is formed simply by the non-zero terms in the 10^{th} row or column. Our goal is to estimate

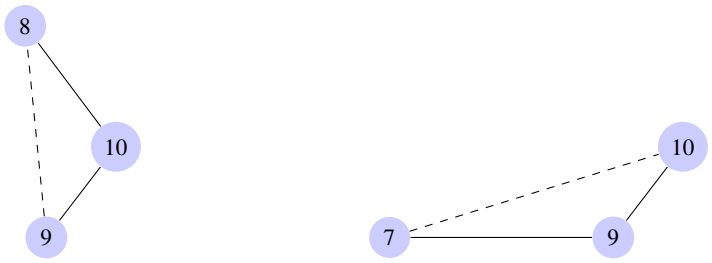


Figure 5.4: Two different alternative sub neighbourhoods for the clique $\{9,10\}$. On the left, we use the union of the neighbours of node 10 and on the right the neighbours of node 9.

$\Sigma^{-1}_{marginal_{10}}$ with respect to the graph in Figure 5.4 and the sample covariance matrix

$$\begin{bmatrix} D_{8,8} & D_{8,9} & D_{8,10} \\ D_{9,8} & D_{9,9} & D_{9,10} \\ D_{10,8} & D_{10,9} & D_{10,10} \end{bmatrix}, \tag{5.8}$$

which is a 3×3 projection matrix from the full covariance matrix. We accomplish the procedure by substituting the value calculated in $\Sigma^{-1}_{marginal}(2,3)$ into $\Sigma_{-1}(9,10)$.

Alternatively, we can chose the domain of x_7, x_9, x_{10}, with marginal:

$$f_{A_{c_q}}(x_7, x_9, x_{10}) = \frac{|\Sigma_{marginal}{}^{-\frac{1}{2}}|}{(2\pi)^{\frac{4}{2}}} \exp(-\frac{1}{2}\mathbf{x}_{c_q}{}^T \Sigma^{-1}_{marginal_9} \mathbf{x}_{c_q}). \tag{5.9}$$

That is, our goal is now to estimate the much smaller $\Sigma^{-1}_{marginal}$ with respect to graph in Figure 5.4 and the sample covariance matrix

$$\begin{bmatrix} D_{7,7} & D_{7,9} & D_{7,10} \\ D_{9,7} & D_{9,9} & D_{9,10} \\ D_{10,7} & D_{10,9} & D_{10,10} \end{bmatrix}, \tag{5.10}$$

which is a different 3×3 projection matrix from the full covariance matrix. This time we accomplish our goal by substituting the value calculated in $\Sigma^{-1}_{marginal}(2,3)$ into $\Sigma_{-1}(9,10)$.

The two alternative estimators explain the need for averaging in distributed estimators. For big data applications, the sub neighbourhoods are preferable. If one is looking to invert an SPD matrix, where the desired inverse matrix has a known sparse pattern, it is sufficient to use LAP with sub neighbourhoods. The benefit not only stems from inverting a 3×3 matrix instead of a 4×4 matrix, but also from the fact that the same sub neighbourhoods hold for several different parameters. Hence, averaging the estimates of the same parameter for different sub neighbourhoods improves statistical efficiency.

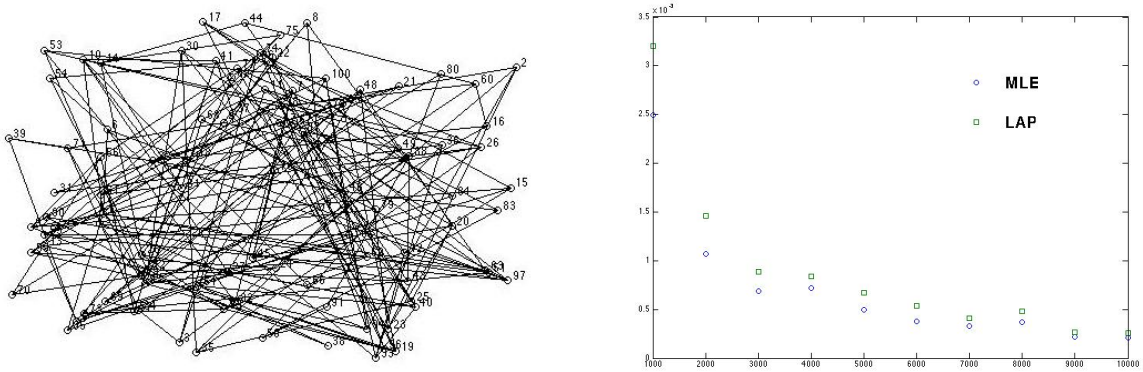


Figure 5.5: A non trivial Gaussian graphical model (left) and the relative estimation error for LAP and MLE as a function of the number of data (right).

5.2 LAP for tables

Section 5.1 may be thought of as "ideal" from some perspectives. The PDF is taken from a well known parametric family, where every marginal distribution is a multivariate Gaussian. In this section, we assume the PDF is a discrete distribution with no known parametric family. The graph structure $G(S,E)$ and data are the only sources of information. However, we also know that the PDF is described as a probability table.

Any general discrete distribution may be expressed *table* form. We will show (by construction) that if the table contains only positive probabilities ($\forall \mathbf{x}\ \ p(\mathbf{x}) > 0$) then it can be expressed as an exponential distribution, in which the energy function consists of polynomials. Since $x_i \in \{0,1\}$ then for any k, $x_i^k = x_i$. Hence, the energy function consists multiplicative terms of x_i of degree 1, such as $\alpha x_1 x_4 x_5$. The terms in the energy function are linear combinations of the subsets of $\{x_1,...,x_k\}$. This is reasonable since for every k, the number of vectors in the discrete distribution

$$f(\mathbf{x});\ \ \mathbf{x} \in \{0,1\}^k$$

is 2^k and the number of subsets of $\{x_1,...,x_k\}$ is also 2^k. Hence, $f(\mathbf{x})$ is spanned by a Gibbs distribution with a polynomial energy function, made of the subsets of $\{x_1,...,x_k\}$, with density:

$$f(\mathbf{x}) = \frac{1}{Z}\exp(\sum_{s\subset\{1..k\}} \theta_s x_{i_1}..x_{i_s}). \tag{5.11}$$

For a given positive table $p(\mathbf{x})$, it is clear that the partition function Z is equal to $p(0)$ (see Figure 5.2),

$$Z = \frac{1}{p(0)}.$$

x_1	x_2	x_3	$p(x)$
0	0	0	$\frac{1}{Z}$
0	0	1	$\frac{1}{Z}\exp(c)$
0	1	0	$\frac{1}{Z}\exp(b)$
0	1	1	$\frac{1}{Z}\exp(c+b+e) \iff f(x) = \frac{1}{Z}\exp(ax_1+bx_2+cx_3+dx_1x_2+ex_2x_3+fx_1x_3+gx_1x_2x_3)$
1	0	0	$\frac{1}{Z}\exp(a)$
1	0	1	$\frac{1}{Z}\exp(a+c+f)$
1	1	0	$\frac{1}{Z}\exp(a+b+d)$
1	1	1	$\frac{1}{Z}\exp(a+b+c+d+e+f+g)$

Figure 5.6: A discrete probability distribution in table form and in full exponential form for $x \in \{0,1\}^3$.

Next we search for vectors $\mathbf{x}$ with sum of entries equal to 1, $\sum(x_i) = 1$:

$$\theta_{(1,0,0,..0)} = log(p(1,0,0,...0) - log(Z))$$

$$\theta_{(0,1,0,..0)} = log(p(0,1,0,...0) - log(Z))$$

.

.

$$\theta_{(0,0,..0,1)} = log(p(0,0,...0,1) - log(Z)).$$

Then we search for vectors with sum of entries equal to 2, $\sum(x_i) = 2$,

$$\theta_{(1,1,0,..0)} = log(p(1,1,0,...0) - log(p(1,0,..0)) - log(p(0,1,0..0)) - log(Z))$$

and so on.

Next, we ask "how can we learn the table of probabilities?". The trivial way is to do it by counting over the realizations:

$$p(x_k = b) = \frac{1}{N}\sum_{n=1}^{N} \delta(x_{n,k} = b). \tag{5.12}$$

However, this may yield zero probabilities, which cannot be represented by exponential distributions. Hence, we form the table in a slightly different way.

Choose $\varepsilon > 0$ and

$$p(x_k = b) = \frac{1}{N+\varepsilon}(\varepsilon + \sum_{n=1}^{N} \delta(x_{n,k} = b)). \tag{5.13}$$

Note that the estimator is consistent, since the full PDF $p(\mathbf{x})$ is non zero for any $\mathbf{x}$ and the weight of any ε goes to zero as $N \to \infty$.

In summary, for each clique potential, the algorithm for tables proceeds as follows:

- Find the first neighbourhood (or any domain that satisfies the Strong LAP condition).
- Use the realizations to form the table of probabilities for the marginal neighbourhood.
- convert the table into full exponential form, and pull out the terms related to the clique of interest.

5.3 Improving the accuracy of LAP

LAP is a cost effective alternative to ML estimation, but not identical to ML for finite sample sizes. However, one can increase the computational cost of LAP in order to improve its estimation accuracy.

Recall that theorem 22 holds not only for domains that satisfy the Strong LAP condition. Clearly, if A_q satisfies the condition, then for any larger subset B that contains A_q the condition holds. We saw the that first neighbourhood is not the minimal sub-graph for which the marginal PDF shares the same normalized potential with the full PDF, and there is computational benefit for choosing smaller domains.

On the other hand, one may take bigger domains than the first neighbourhood. This lead us to the idea of the second neighbourhood, and, recursively, to higher order neighbourhoods.

Let us first generalize the definition of the first neighbourhood, as given in (3.1),

Definition 28. *The k-neighbourhood of a subset q, denoted A_q^k, is the first neighbourhood of the subset A_q^{k-1}.*

Alternatively,

Definition 29. *The k-neighbourhood of a subset q, denoted A_q^k, is the collection of nodes with distance of k or less edges from q.*

As proven in Theorem 23 Estimating the marginal over higher neighbourhoods will yield results closer to ML. Enlarging the marginal domain to the maximal set (*i.e.*, to the entire graph) will simply coincide with the ML itself. In other words, LAP can be understood as a series of nested estimators of increasing statistical efficiency and decreasing computational efficiency.

To illustrate the effect of neighbourhood size, we consider the Gaussian graphical model of Figure 5.7 and two clique neighbourhoods as illustrated in the same figure. Figure 5.8 shows the relative errors obtained using these neighbourhoods. Clearly higher order neighbourhoods perform better, but it is not clear how important they are as both LAP estimators are very close to ML.

5.4 Memory allocation for LAP

LAP has low memory requirements. If each node has a bounded number of p neighbours, then the first neighbourhood involves no more than p^2 nodes and if working with the minimal sub neighbourhoods, only p nodes are needed in each local estimator.

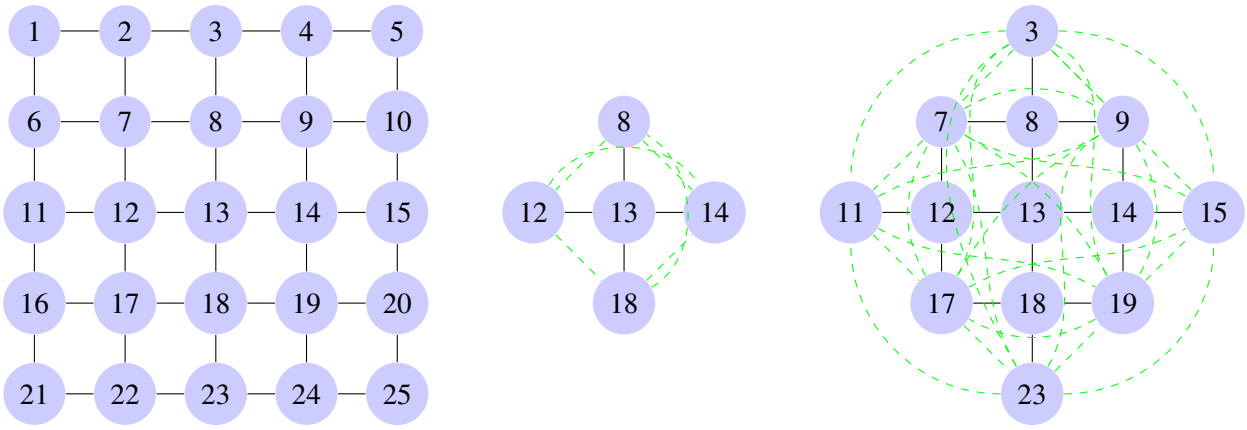


Figure 5.7: On the left, a 5×5 lattice MRF with vertical and horizontal neighbours. The middle graph is the first neighbourhood for the unary clique $\{13\}$. The figure on the right shows the second neighbourhood for the unary clique $\{13\}$. New edges introduced by marginalization are depicted with dashed green lines.

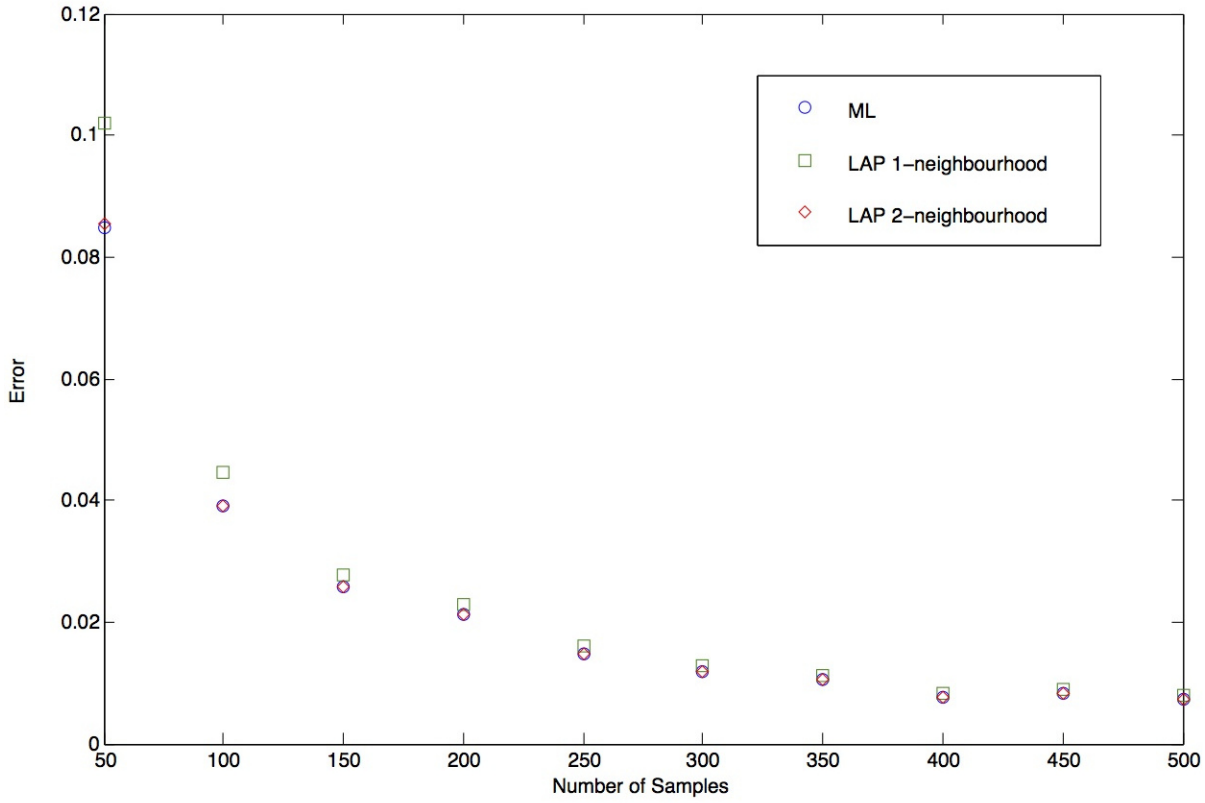


Figure 5.8: Relative error of parameter estimates compared to maximum likelihood for 1-neighbourhood LAP and 2-neighbourhood LAP. The full graph contained 300 nodes and the PDF is a multivariable Gaussian.

Consider the sparse inverse covariance estimation problem. If there are n nodes, working with the full PDF requires allocating memory for n^2 cells. The complexity of inverting the covariance matrix is n^3. One can apply LAP to each row, with each local estimator using the nodes with the non-zero values in the row. Thus the maximal memory allocation needed is of the order p^2. This may play a significant role in large systems where $n >> p$.

Chapter 6

i LAP: applying LAP to inverse problems

Large-scale inverse problems arise in a multitude of applications, including weather forecasting, medical imaging, and oil exploration. In these big data domains, the high-dimensionality of the models that must be recovered from data creates substantial computational challenges, see for example [Roosta-Khorasani et al., 2014, Herrmann et al., 2012, Ascher and Haber, 2004, Haber and Ascher, 2001, Scheichl et al., 2013] and the references therein.

The unknown model often represents some physical property of the system under investigation. The relationship between the model and the data is often assumed known and encoded in the form of a *forward operator*, which predicts data for a given model.

The model is typically obtained by computing the *Maximum a Posteriori* (MAP) estimate. If the model is high-dimensional, computing this estimate is computationally expensive and demands large amounts of memory.

In the present chapter, drawing upon the strong LAP argument, we offer an alternative approach for solving inverse problems. In particular, we introduce an efficient parallel algorithm, named *i*LAP, which appropriately divides the large problem into smaller sub-problems of much lower dimension. This process of localization offers substantial advantages in terms of computational efficiency and memory allocation.

6.1 Inverse problems

In inverse problems, the physical system under investigation is modelled as follows:

$$\mathbf{A}(\mathbf{m}) + \boldsymbol{\varepsilon} = \mathbf{d}. \tag{6.1}$$

Here, $\mathbf{m} \in \mathbb{R}^n$ is the unknown model, $\mathbf{A} : \mathbb{R}^n \to \mathbb{R}^k$ is the known (linear or non-linear) forward operator that is assumed to be sparse in our setting, $\mathbf{d} \in \mathbb{R}^k$ is the data and $\boldsymbol{\varepsilon} \in \mathbb{R}^k$ is the noise, which is assumed to be zero mean Gaussian with known diagonal covariance matrix Σ_d,

$$\boldsymbol{\varepsilon} \sim N(0, \Sigma_d). \tag{6.2}$$

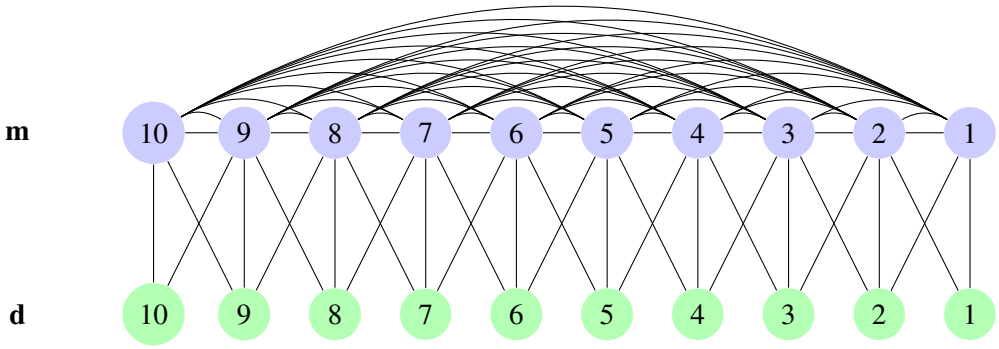


Figure 6.1: The graphical representation of the MRF in which both $\mathbf{m}$ and $\mathbf{b}$ are variables

The inverse problem of recovering $\mathbf{m}$ given the data $\mathbf{d}$ and the known parameters $\theta = (\mathbf{A}, \Sigma_d^{-1})$ is ill-posed in most applications of interest [Vogel, 2002, Kaipio and Somersalo, 2005]. This forces us to introduce a regularizer or prior $\pi(\mathbf{m})$, which by Bayes rule yields the following posterior distribution:

$$p(\mathbf{m}|\mathbf{d}) \propto p(\mathbf{d}|\mathbf{m})\pi(\mathbf{m}). \tag{6.3}$$

We refer to $p(\mathbf{d}|\mathbf{m})$ as the conditional distribution. Its form follows from the model specification:

$$p(\mathbf{d}|\mathbf{m};\theta) = \frac{1}{Z}\exp\left(-\frac{1}{2}(\mathbf{Am}-\mathbf{d})^\top \Sigma_d^{-1}(\mathbf{Am}-\mathbf{d})\right) \tag{6.4}$$

Our goal is to compute the MAP estimate of the model:

$$\hat{\mathbf{m}}_{MAP} = \arg\max_{\mathbf{m}} \left(\exp\left(-\frac{1}{2}(\mathbf{Am}-\mathbf{d})^\top \Sigma_d^{-1}(\mathbf{Am}-\mathbf{d})\right)\pi(\mathbf{m})\right). \tag{6.5}$$

6.2 Localizing inverse problems

Suppose the forward operator is sparse. For example, consider the model illustrated in Figure 6.1. For each component of the data $\mathbf{d}_i$, let $\{\mathbf{m}\}_i$ denote the group of elements of the model that influence $\mathbf{d}_i$ directly. In our example, to generate d_8, we only need to know the components $\{\mathbf{m}\}_8 = \{m_7, m_8, m_9\}$ of the model.

Since we are assuming that $\mathbf{A}$ is sparse, we have that $|\{\mathbf{m}\}_i| << |\mathbf{m}|$. Moreover since the observation noise is such that the entries of $\mathbf{d}$ are conditionally independent, we can express the joint model as an MRF:

$$p(\mathbf{m},\mathbf{d}) = \pi(\mathbf{m})p(d_1|\{\mathbf{m}\}_1)p(d_2|\{\mathbf{m}\}_2)...p(d_k|\{\mathbf{m}\}_k). \tag{6.6}$$

Our goal is to replace the global objective

$$p(\mathbf{m},\mathbf{d};\theta) = \pi(\mathbf{m})p(\mathbf{d}|\mathbf{m};\theta)$$

with a set of *local low-dimensional objectives*,

$$p(\tilde{\mathbf{m}}, \tilde{\mathbf{d}}; \tilde{\theta}) = \pi(\tilde{\mathbf{m}}) p(\tilde{\mathbf{d}} | \tilde{\mathbf{m}}; \tilde{\theta})$$

that can be solved easily and independently.

The problem with this strategy is that we do not know the values of the local parameters $\tilde{\theta}$ or the expression for the local prior $\pi(\tilde{\mathbf{m}})$ in general.

However, using the Strong LAP results (Theorems 22 and 24), we will be able to construct local objectives for which $\tilde{\theta}_i = \theta_i$ for i associated with m_i. Once the local parameters are known, this divide-and-conquer strategy will enable us to compute each of the model components in a fully parallel manner. Using this distributed estimation approach, we will be able to recover the global MAP estimates provided the prior also decomposes.

6.2.1 Localizing the conditional distribution

Let $\mathbf{m}_i \in \mathbf{m}$ be the entry of interest. To appeal to the Strong LAP Condition, we need to construct an appropriate neighbourhood for the local model involving $\mathbf{m}_i$. This construction is illustrated in Figure 6.2.

Definition 30. *The 1-hop data of* $\mathbf{m}_i$ *is the set of all entries in* $\mathbf{d}$ *at distance 1 from* $\mathbf{m}_i$ *in the graphical model representation.*

Definition 31. *The 1-blanket of* $\mathbf{m}_i$ *is the set of all nodes in the graphical model at distance 1 or less from the 1-hop data of* $\mathbf{m}_i$.

Proposition 32. *Let* $\{\tilde{\mathbf{d}}, \tilde{\mathbf{m}}\}$ *be the 1-blanket of* $\mathbf{m}_i$. *Then,* $p(\tilde{\mathbf{d}} | \tilde{\mathbf{m}}; \tilde{\theta})$ *inherits* $\tilde{\theta}$ *from* θ.

Proof. By construction, $\tilde{\mathbf{d}}$ is the 1-hop data of $\mathbf{m}_i$. By conditional independence, there are no edges in the joint graphical representation connecting two data nodes. By definition, the 1-blanket is the 1-neighbourhood of the parameters associated with the edges connecting $\mathbf{m}_i$ and $\tilde{\mathbf{d}}$. Hence, the Strong LAP Condition is satisfied for the parameters connecting $\mathbf{m}_i$ and the 1-hop data. The result follows by the Strong LAP Theorem (Theorem 22). □

6.2.2 Localizing the prior

The previous subsection provided us with conditions under which the parameters of the local conditionals are the same as the parameters of the global conditional distribution. In this subsection, we focus on the prior distribution. Let $\pi(\tilde{\mathbf{m}})$ be the marginal of the global prior $\pi(\mathbf{m})$.

In some tractable cases, it is feasible to obtain an analytical expression for the local prior $\pi(\tilde{\mathbf{m}})$ by marginalization. For example, this is true if the prior is Gaussian.

In some applications, the prior is not expressed in terms of a function, but rather in terms of a set of N samples $\{\mathbf{m}\}_{i=1}^{N}$. Here the local prior is straightforwardly obtained by discarding the samples of model components not associated with the marginal.

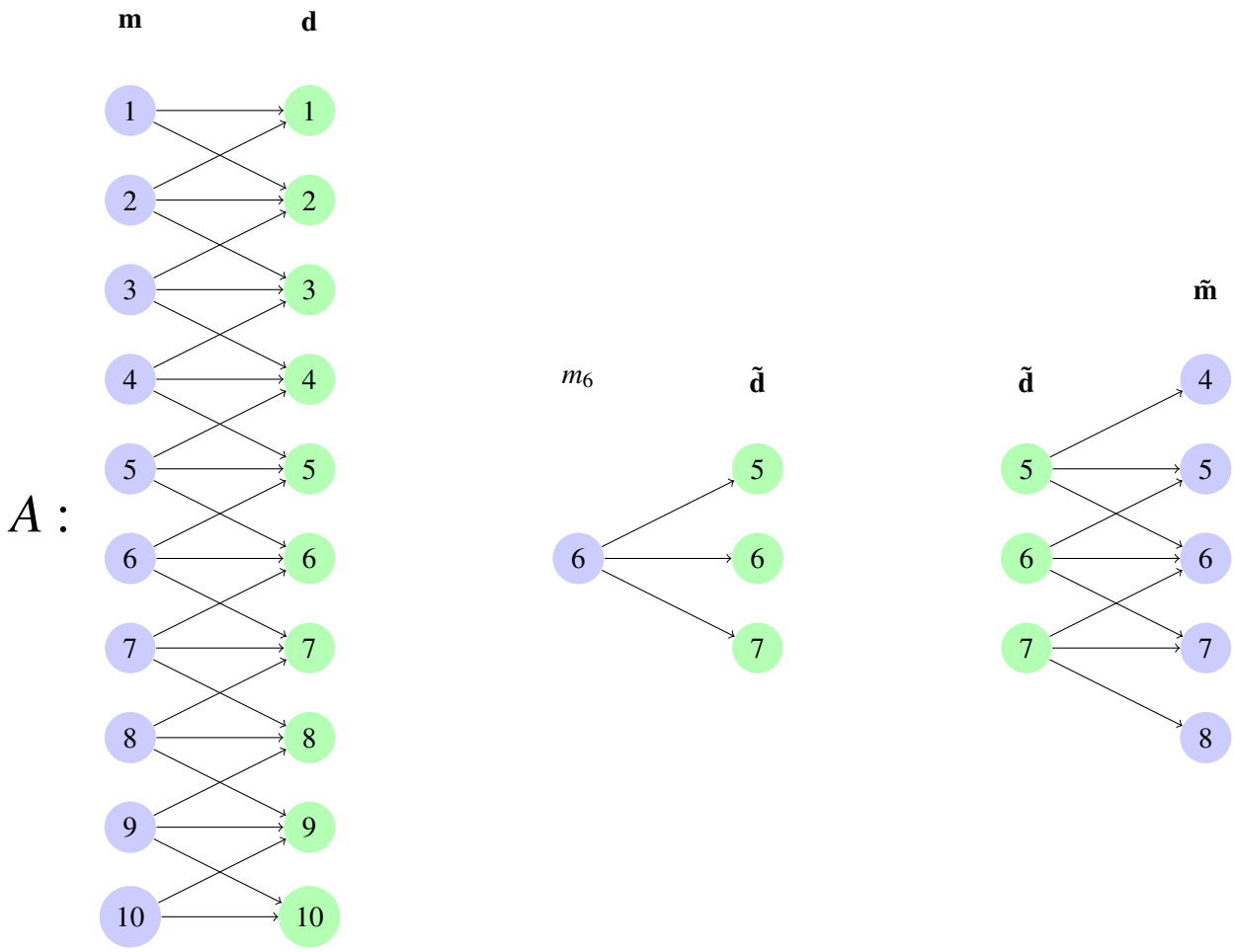


Figure 6.2: Construction of the local models. Suppose we are interested in estimating m_6. Then, $\tilde{\mathbf{d}} = \{d_5, d_6, d_7\}$ are the 1-hop data of m_6. $\tilde{\mathbf{m}} = \{m_4, m_5, m_6, m_7, m_8\}$ are the components of the model that affect $\tilde{\mathbf{d}}$ directly. The 1-blanket consisting of $\tilde{\mathbf{d}}$ and $\tilde{\mathbf{m}}$ constitutes the 1-neighbourhood of the parameters θ_{65}, θ_{66} and θ_{67}.

In the above two cases, if the conditional distributions factorises, the posterior also factorises and we are able to recover the global MAP estimates by computing local MAP estimates.

In general, the prior simply encodes sparseness and smoothness assumptions. For example, one may use the ℓ_1 norm to construct the prior when the model is assumed to be sparse. In this situation, we can no longer localize the prior and hence we cannot guarantee that the local estimates coincide with the global MAP estimates. However, denoising and deblurring experiments, using the ℓ_1 norm on the local model, will show that the *i*LAP approach cab be effective even in these situations.

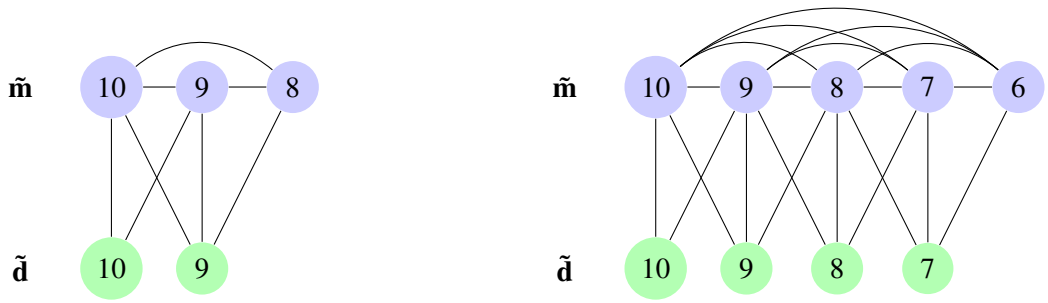


Figure 6.3: The 1-blanket (left) and 2-blanket (right) of $\mathbf{m}_{10}$.

6.3 The *i*LAP algorithm

Following the construction of the 1-neighbourhood of the parameters of the conditional distribution, the *i*LAP algorithm is as follows:

Algorithm 2 *i*LAP (1-blanket)

Input: Forward operator $\mathbf{A}$ and observation $\mathbf{d}$
for $m_i \in \mathbf{m}$ **do**
  Find $\tilde{\mathbf{d}} \subseteq \mathbf{d}$; the 1-hop data of $\mathbf{m}_i$.
  Find $\{\tilde{\mathbf{m}}, \tilde{\mathbf{d}}\}$; the 1-blanket of $\mathbf{m}_i$.
  Find the local prior as described in Section 6.2.2.
  Construct the local objective function: $\pi(\tilde{\mathbf{m}})p(\tilde{\mathbf{d}}|\tilde{\mathbf{m}})$.
  Compute the local MAP estimate of m_i by maximizing the local objective.
end for

Remark 33. *For any* $\breve{\mathbf{m}}$*, s.t.*

$$\tilde{\mathbf{m}} \subset \breve{\mathbf{m}}$$

one can define the augmented local inverse problem by $\pi(\breve{\mathbf{m}})p(\tilde{\mathbf{d}}|\tilde{\mathbf{m}})$ *because*

$$p(\tilde{\mathbf{d}}|\tilde{\mathbf{m}}) = p(\tilde{\mathbf{d}}|\breve{\mathbf{m}}).$$

The choice of a correct local objective is not unique and one can consider larger neighbourhoods. In particular, we can define a 2-blanket as follows.

Definition 34. *The 2-hop data of* $\mathbf{m}_i$ *is the set of all entries in* $\mathbf{d}$ *with distance 1 from the 1-blanket* $\mathbf{m}_i$ *in the graphical model.*

Definition 35. *The 2-blanket of* $\mathbf{m}_i$ *is the set of all nodes in the graphical model at distance 1 or less from the 2-hop data of* $\mathbf{m}_i$*.*

Proceeding in a similar fashion, we can generalize to the k-blanket of $\mathbf{m}_i$. Figure 6.3 depicts the 1-blanket and the 2-blanket of $\mathbf{m}_{10}$ for the model shown in Figure 6.1.

As a result of the previous observation, we can select larger subsets of $\mathbf{m}$, as opposed to a single entry, and perform local MAP estimation in blocks. We follow this strategy in the experiments presented in the following section.

6.4 Image deblurring example using the DCT and wavelet transforms

Figure 6.4: MAP estimation using *i*LAP (1-blanket) and 4×4 blocks. Top left: true model, top right: data, middle left: full inverse reconstruction with DCT transform, middle right: *i*LAP with DCT transform, bottom left: full inverse reconstruction with wavelet transform, bottom right: *i*LAP with wavelet transform.

We consider the problem of recovering a 128×128 image $\mathbf{m}$ that has been corrupted by a global blurring

operator and noise.

Specifically, we generate the corrupted image by convolving it with a Gaussian blurring kernel with standard deviation 1.2 in the spatial domain. Subsequently, 2% white Gaussian noise is added to the blurred image.

In order to recover the image, the dense blurring operator is made sparse by by setting to zero all entries of $\mathbf{A}$ that fall below 10^{-4}. The recovery is done by basis pursuit denoising with an ℓ_1 regularizer:

$$\min \|\mathbf{f}\|_1 \quad \text{subject to} \quad \|\mathbf{A}\mathbf{H}^{-1}\mathbf{f} - \mathbf{d}\|_2^2 \leq \alpha, \tag{6.7}$$

where $\mathbf{f} = \mathbf{H}\mathbf{m}$ is the transformed model and $\mathbf{H}$ is the transformation (DCT or wavelet bases). Both $\mathbf{A}$ and $\mathbf{H}$ are of size $16,384 \times 16,384$. We refer the reader to Mallat [2009] or Chan and Shen [2005] for a detailed discussion regarding DCT and wavelet transforms.

The above denoising objective for a single image states that the model is sparse in the transformed domain (frequency or wavelet domain). This is a reasonable assumption in many applications. We solve the optimization problem using the `spgl1` package [van den Berg and Friedlander, 2009, 2011].

For the DCT, we considered 4×4 blocks of pixels. For this size of local estimates and the truncated Gaussian kernel of radius 5, the 1-blankets and 2-blankets were augment to square patches of size 24×24 and 44×44 respectively. They could be made smaller since the kernel is circular, but by making the square they become considerably easier to code. For the wavelet transform, we had to consider blocks of size 8×8 with 1-blankets of radius 32.

The recovery results are shown in Figure 6.4. In this particular example, the local approach appears to do better than the global approach. This however may not necessarily generalize to other images.

In Figure 6.5 we show the relative error for the global and *i*LAP algorithms using the Wavelet transform as a function of the regularization coefficient α. The relative error is calculated as follows:

$$Error = \frac{||\mathbf{m}^* - \mathbf{m}||}{||\mathbf{m}^*||},$$

where $\mathbf{m}^*$ is the true image and $\mathbf{m}$ is the deblurred image. While the optimal range of the coefficient varies, the relative error of *i*LAP is similar to the one of the global MAP estimator.

Finally, Figure 6.5 compares the results for the 1-blanket and 2-blanket estimators using the DCT. There appears to be no gain for using larger neighbourhoods in this case. This may be attributed to the sparsity prior.

It is both interesting and reassuring that similar ideas, but solely in the context of image deblurring, have been studied since the seminal work in Trussell and Hunt [1978]. Our *i*LAP approach provides a more general framework for understanding these approximations and for developing more powerful algorithms.

If, for example, $\mathbf{m}$ contain 10^6 entries (in the case of 1000×1000 pixels), solving each entry at a time requires 10^6 local solvers, while solving in 10×10 blocks requires only 10^4 solvers.

In Figure 6.6 we present the comparison of a 4×4 blocks 1-blanket reconstruction and a 4×4 blocks

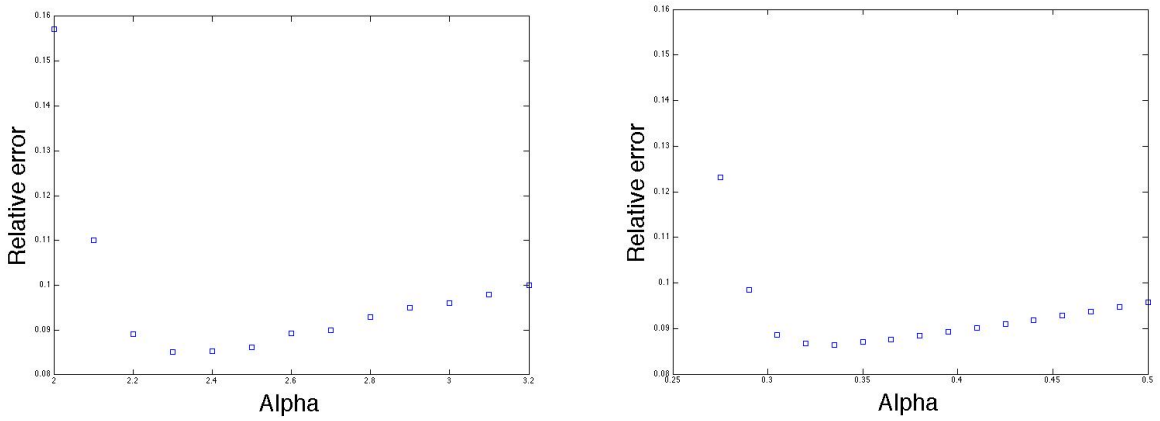


Figure 6.5: Relative error of global MAP estimate (left) and *i*LAP distributed estimates (right) using the wavelet transform. While the values of the regularization coefficient are different, the relative errors at the optimum values are very close.

Figure 6.6: Recovery using *i*LAP with the first-blanket (left) and the second-blanket (right) with blocks of size 4×4 and the DCT.

2-blanket reconstruction, for the same problem presented in Section 6.4.

Chapter 7

Concluding remarks and future work

The key to this work was the somehow under-appreciated theorem on the existence and uniqueness of the normalized Gibbs representation of an MRF.

We attacked an important question: Under what conditions do marginal MRFs have the same potentials as the full MRF? The answer to this question is the Strong LAP Theorem.

The LAP algorithm for distributed learning in MRFs, both in its marginal and conditional forms, is just a way to exploit this theoretical result. *i*LAP for inverse problems is another example of the utility of the theory advanced in this thesis. We believe many more applications abound.

However, there remain several open questions, mainly regarding *i*LAP. For example, should a bigger blanket improve the estimation accuracy? How can we generalize *i*LAP to the non-independent noise or dense operator scenarios?

The following subsections discuss a few additional areas that merit further research.

7.1 Corrupted data

We discussed parameter estimation using data that was sampled from the true PDF. An exciting generalization for LAP will be to consider cases in which the data is corrupted. Naturally we would like to consider first the case where the data is *locally* corrupted. That is, in each sample noise is added to a small number of unknown entries. Filtering the noise (that is, for each sample learning for which entries the noise was added, and ignoring these) may be a complicated task when handling the entire sample, and one can benefit from reducing the dimension.

Let $\mathbf{x}_1..\mathbf{x}_q$ be q locally corrupted samples taken from the PDF $p(\mathbf{x}|\theta)$. Using LAP, for each parameter we find a local domain, project the samples $\mathbf{x}_1..\mathbf{x}_q$ on that domain, then estimate the marginal PDF. Filtering the sample projections may be a more reasonable task, since we can consider each projection as *"corrupted"* or *"not corrupted"*.

7.2 Structure learning

Assume the set of samples is taken from the true PDF, but that the graph structure is not known. We wish to estimate the graph structure using the sparsity assumption. In other words, each node has only few neighbours but these neighbours are not known.

The main difficulty lays in the fact that even if the number of direct neighbours for each node is limited, the total number of combinations is intractable. Using LAP, we would like to explore new local approaches, which may simplify the problem. The Basis for such approach have to be a *local* structure learning. That is, for each node we estimate the potentials locally, rating different local structures independently, and then combine these local structures into a global one.

7.3 Tied parameters

We ignored the possibility that the entries of the parameters vector θ may dependent on each other. If, for example, the same parameter appears in several different clique potentials. In such case, one may naively take the following two steps:

- Ignore the relation between the entries, by assuming that the PDF is given by

$$f(x;\theta) = \frac{1}{Z(\theta)} exp(\sum_{c \in \mathscr{C}} V_C(x_c;\theta_c)). \tag{7.1}$$

  where each entry is independent from the other. Use LAP to find the estimate $\hat{\theta}$.

- Estimate each entry θ_i by averaging the relevant estimates in $\hat{\theta}$.

Clearly, this leads to a consistent estimator since it is known that

$$\hat{\theta} \xrightarrow{n \to \infty} \theta_{true} \tag{7.2}$$

for each of the local estimators.

However, equally simple averaging is not an optimal method. An open question is: What is the best way to choose averaging weights? Regardless, it is clear that LAP should be applied to domains with tied parameters, including conditional random fields, to assess its merits. The one thing that LAP has going for itself is that it is data efficient, model efficient, and consistent. Therefore, in the age of big data, it should become an important player in the field of parameter learning.

Bibliography

D. H. Ackley, G. Hinton, and T. Sejnowski. A learning algorithm for Boltzmann machines. *Cognitive Science*, 9:147–169, 1985.

A. Agarwala, M. Dontcheva, M. Agrawala, S. Drucker, A. Colburn, B. Curless, D. Salesin, and M. Cohen. Interactive digital photomontage. In *ACM SIGGRAPH*, pages 294–302, 2004.

U. Ascher and E. Haber. Computational methods for large distributed parameter estimation problems with possible discontinuities. In *Proceedings of the Symposium on Inverse Problems, Design and Optimization*, pages 201–208, 2004.

A. Asuncion, Q. Liu, A. Ihler, and P. Smyth. Learning with blocks: Composite likelihood and contrastive divergence. In *Artificial Intelligence and Statistics*, pages 33–40, 2010.

J. Besag. Spatial interaction and the statistical analysis of lattice systems. *Journal of the Royal Statistical Society, Series B*, 36:192–236, 1974.

J. Besag. Statistical analysis of non-lattice data. *Journal of the Royal Statistical Society. Series D*, 24(3): 179–195, 1975.

J. Besag. Comments on Conditionally Specified Distributions: An Introduction by B. Arnold, E. Castillo and J. Sarabia. *Statistical Science*, 16(3):249–265, 2001.

Y. Boykov and O. Veksler. Graph cuts in vision and graphics: Theories and applications. In N. Paragios, Y. Chen, and O. Faugeras, editors, *Handbook of Mathematical Models in Computer Vision*, chapter 5, pages 79–96. Springer, 2006.

J. Bradley. *Learning Large-Scale Conditional Random Fields*. PhD thesis, Machine Learning Department, Carnegie-Mellon University, 2013.

J. K. Bradley and C. Guestrin. Sample complexity of composite likelihood. In *Artificial Intelligence and Statistics*, pages 136–160, 2012.

A. Braunstein, M. Mezard, and R. Zecchina. Survey propagation: An algorithm for satisfiability. *Random Structures and Algorithms*, (2):201–226, 2005.

P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Springer-Verlag, 2001.

D. Buchman, M. W. Schmidt, S. Mohamed, D. Poole, and N. de Freitas. On sparse, spectral and other parameterizations of binary probabilistic models. *Journal of Machine Learning Research - Proceedings Track*, 22:173–181, 2012.

T. Chan and J. Shen. *Image Processing and Analysis: Variational, PDE, Wavelet and Stochastic Methods*. SIAM, 2005.

X. Chen, B. Neubert, Y. Xu, O. Deussen, and S. B. Kang. Sketch-based tree modeling using Markov random fields. In *ACM SIGGRAPH Asia*, pages 1–9, 2008.

B. Cox. Composite likelihood methods. *Contemporary Mathematics*, 80:221–239, 1988.

A. P. Dempster. Covariance selection. *Biometrics*, 28(1):157–175, 1972.

M. Denil and N. de Freitas. Toward the implementation of a quantum RBM. In *NIPS Deep Learning and Unsupervised Feature Learning Workshop*, 2011.

J. V. Dillon and G. Lebanon. Stochastic composite likelihood. *Journal of Machine Learning Research*, 11: 2597–2633, 2010.

S. E. Fienberg and A. Rinaldo. Maximum likelihood estimation in log-linear models. *The Annals of Statistics*, 40(2):996–1023, 2012.

R. A. Fisher. On the mathematical foundations of theoretical statistics. *Philosophical Transactions of the Royal Society of London Series A*, 222:309–368, 1922.

M. Galley. A skip-chain conditional random field for ranking meeting utterances by importance. In *Empirical Methods in Natural Language Processing*, pages 364–372, 2006.

D. Griffeath. Introduction to random fields. In *Denumerable Markov Chains*, volume 40 of *Graduate Texts in Mathematics*, pages 425–458. Springer, 1976.

E. Haber and U. Ascher. Preconditioned all-at-one methods for large, sparse parameter estimation problems. *Inverse Problems*, 17:1847–1864, 2001.

J. M. Hammersley and P. Clifford. Markov fields on finite graphs and lattices. 1971.

F. J. Herrmann, M. P. Friedlander, and O. Yilmaz. Fighting the curse of dimensionality: compressive sensing in exploration seismology. *IEEE Signal Processing Magazine*, 29:88–100, 2012.

G. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, 14(8): 1771–1800, 2000.

J. J. Hopfield. Neurons with graded response have collective computational properties like those of two-state neurons. *Proceedings of the National Academy of Sciences*, 81(10):3088–3092, 1984.

A. Hyvärinen. Estimation of non-normalized statistical models using score matching. *Journal of Machine Learning Research*, 6:695–709, 2005.

A. Hyvärinen. Connections between score matching, contrastive divergence, and pseudolikelihood for continuous-valued variables. *IEEE Transactions on Neural Networks*, 18(5):1529–1531, 2007.

M. I. Jordan. *An introduction to probabilistic graphical models*. 2002.

J. Kaipio and E. Somersalo. *Statistical and computational inverse problems*. Springer, 2005.

R. Kindermann and J. L. Snell. *Markov Random Fields and their Applications*. American Mathematical Society, 1980.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, 2009.

J. D. Lafferty, A. McCallum, and F. C. N. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *International Conference on Machine Learning*, pages 282–289, 2001.

S. Lauritzen. *Graphical models*. Oxford University Press, 1996.

S. Z. Li. *Markov random field modeling in image analysis*. Springer-Verlag, 2001.

P. Liang and M. I. Jordan. An asymptotic analysis of generative, discriminative, and pseudolikelihood estimators. In *International Conference on Machine Learning*, pages 584–591, 2008.

Q. Liu and A. Ihler. Distributed parameter estimation via pseudo-likelihood. In *International Conference on Machine Learning*, 2012.

D. MacKay. *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 2003.

S. Mallat. *A Wavelet Tour of Signal Processing: the Sparse Way*. Academic Press, 2009.

K. V. Mardia, J. T. Kent, G. Hughes, and C. C. Taylor. Maximum likelihood estimation using composite likelihoods for closed exponential families. *Biometrika*, 96(4):975–982, 2009.

E. Marinari, G. Parisi, and J. Ruiz-Lorenzo. Numerical simulations of spin glass systems. *Spin Glasses and Random Fields*, pages 59–98, 1997.

B. Marlin and N. de Freitas. Asymptotic efficiency of deterministic estimators for discrete energy-based models: Ratio matching and pseudolikelihood. In *Uncertainty in Artificial Intelligence*, pages 497–505, 2011.

B. Marlin, K. Swersky, B. Chen, and N. de Freitas. Inductive principles for restricted Boltzmann machine learning. In *Artificial Intelligence and Statistics*, pages 509–516, 2010.

Z. Meng, D. Wei, A. Wiesel, and A. O. Hero III. Distributed learning of Gaussian graphical models via marginal likelihoods. In *Artificial Intelligence and Statistics*, pages 39–47, 2013.

Z. Meng, D. Wei, A. Wiesel, and A. O. Hero III. Marginal likelihoods for distributed parameter estimation of Gaussian graphical models. Technical report, arXiv:1303.4756, 2014.

K. P. Murphy. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

S. Nowozin. Constructing composite likelihoods in general random fields. In *ICML Workshop on Inferning: Interactions between Inference and Learning*, 2013.

S. Okabayashi, L. Johnson, and C. Geyer. Extending pseudo-likelihood for Potts models. *Statistica Sinica*, 21(1):331–347, 2011.

P. Ravikumar, M. J. Wainwright, and J. D. Lafferty. High-dimensional Ising model selection using ℓ_1-regularized logistic regression. *Annals of Statistics*, 38(3):1287–1319, 2010.

F. Roosta-Khorasani, K. van den Doel, and U. Ascher. Stochastic algorithms for inverse problems involving PDEs and many measurements. *SIAM Journal of Scientific Computing*, 2014.

R. Scheichl, S. Kindermann, M. A. Freitag, and M. Cullen. *Large Scale Inverse Problems: Computational Methods and Applications in the Earth Sciences*. De Gruyter, 2013.

P. Smolensky. Information processing in dynamical systems: foundations of harmony theory. *Parallel distributed processing: explorations in the microstructure of cognition*, 1:194–281, 1986.

J. Songsiri, J. Dahl, and L. Vandenberghe. Maximum-likelihood estimation of autoregressive models with conditional independence constraints. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 1701–1704, 2009.

C. Sutton and A. McCallum. An introduction to conditional random fields. *Foundations and Trends in Machine Learning*, 4(4):267–373, 2012.

K. Swersky, M. Ranzato, D. Buchman, B. Marlin, and N. Freitas. On autoencoders and score matching for energy based models. In *International Conference on Machine Learning*, pages 1201–1208, 2011.

R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rother. A comparative study of energy minimization methods for Markov random fields with smoothness-based priors. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(6):1068–1080, 2008.

H. Trussell and B. Hunt. Sectioned methods for image restoration. *IEEE Transactions on Acoustics, Speech and Signal Processing*, 26(2):157–164, 1978.

E. van den Berg and M. Friedlander. Probing the pareto frontier for basis pursuit solutions. *SIAM Journal on Scientific Computing*, 31(2):890–912, 2009.

E. van den Berg and M. Friedlander. Sparse optimization with least-squares constraints. *SIAM Journal on Optimization*, 21(4):1201–1229, 2011.

A. W. van der Vaart. *Asymptotic statistics*. Cambridge University Press, 1998.

C. Varin, N. Reid, and D. Firth. An overview of composite likelihood methods. *Statistica Sinica*, 21:5–42, 2011.

C. Vogel. *Computational methods for inverse problems*. SIAM, 2002.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

L. Wasserman. *All of Statistics*. Springer, 2004.

A. Wiesel and A. Hero III. Distributed covariance estimation in Gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(1):211–220, 2012.

C. Yanover, O. Schueler-Furman, and Y. Weiss. Minimizing and learning energy functions for side-chain prediction. In T. Speed and H. Huang, editors, *Research in Computational Molecular Biology*, volume 4453 of *Lecture Notes in Computer Science*, pages 381–395. Springer, 2007.

L. Younes. Parametric inference for imperfectly observed Gibbsian fields. *Probability Theory and Related Fields*, 82(4):625–645, 1989.

Appendix A

A.1 Equivalent definitions of the 1-neighbourhood

For a given clique, q, define its *1- neighborhood* A_q by :

$$\widehat{A}_q = \cup c, \quad \forall c \in \mathscr{C} \ \ s.t. \ \ c \cap q \neq \emptyset \tag{A.1}$$

as the union of all cliques with non empty intersection. Or, alternatively, as the union of all nodes with distance one or less from q

$$\tilde{A}_q = q \cup \mathscr{N}_i, \ \ \forall i \in q. \tag{A.2}$$

We show here the Equivalence of these two definitions.

Proof. Clearly $q \subseteq \widehat{A}_q$ and $q \subseteq \tilde{A}_q$.
$\Rightarrow$ Let $i \in \tilde{A}_q$, and $i \notin q$. Then, $i \in \mathscr{N}_j$ for $j \in q$ and the subset $\{i, j\}$ is a clique with non empty intersection with q. Therefore $i \in \widehat{A}_q$.

$\Leftarrow$ Let $i \in \widehat{A}_q$. Then $i \in c_i$, where c_i is a clique with non empty intersection with q,

$$\exists j \in q \cap c_i,$$

c_i is a clique, then $i \in \mathscr{N}_j$ and $i \in \tilde{A}_q$.

□

A.2 Strong LAP condition is sufficient but not necessary

In this subsection we prove by example that the strong LAP condition is not a necessary condition. Let the graph be as in Figure A.1. Our interest lays in the parametric estimation for the clique $\{1,2,3\}$.

The marginal PDF over the domain $\{1,2,3\}$ is achieved by integrating over the nodes i, j and k and will not satisfies the Strong lap condition with respect to the clique $\{1,2,3\}$, as all the possible edges in the clique are formed in the induced MRF. The edge $\{1,2\}$ by integration over i, the edge $\{2,3\}$ by integration over j and the edge $\{1,3\}$ by integration over k. However, the induced MRF consist only pairwise cliques,

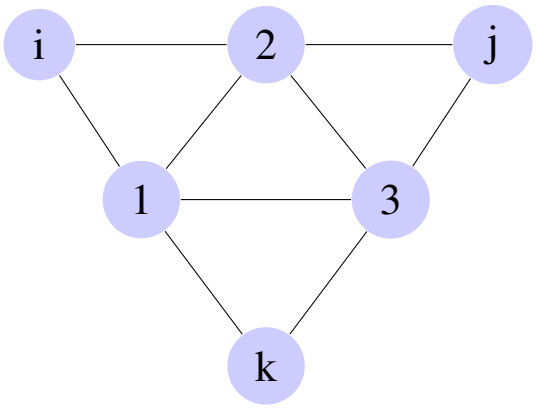


Figure A.1: A simple graph. Our interest is in the clique $\{1,2,3\}$

and tough the clique $\{1,2,3\}$ is formed in the graph, the clique potential will not be formed. Hence, the marginal over $\{1,2,3\}$ and the full PDF shares the same potential over the clique of interest.

A.3 Conditional estimator can not be better than marginal estimator

Let $p(\mathbf{x};\theta)$ be the full MRF, let q be the clique of interest and A_q be the marginal domain. We assume one can not find the exact marginal parametric family, i.e., the exact solution of

$$p_{A_q}(\mathbf{x}_{A_q}) = \int p(\mathbf{x});\theta) d\mathbf{x}_{S\setminus A_q}$$

is not known. In such cases we say that the parametric family is not integrable. In particular we note that the marginal is combined from Energy potentials which appeared originally in the full PDF and are all known, and the induced Energy potentials which are not known.

On the other hand, the conditional $p(x_q|x_{A_q\setminus q})$ is known and can be directly derived from the full PDF. Moreover, we have shown that the full PDF and the conditional shares the same Energy potential over the clique of interest, by that observation we derived the CLAP algorithm.

The marginal PDF can be written as

$$p_{A_q}(\mathbf{x}_{A_q}) = p(\mathbf{x}_{A_q\setminus q};\alpha)p(x_q|x_{A_q\setminus q}).$$

Again, we assume the exact parametric form of $p(\mathbf{x}_{A_q\setminus q};\alpha)$ cannot be derived directly from $p(\mathbf{x};\theta)$.

However, by expanding the parametric family and introducing more parameters one may find larger family, $p(\mathbf{x}_{A_q\setminus q};\beta)$ that will contain $p(\mathbf{x}_{A_q\setminus q};\alpha)$.

In such a case, the marginal model is taken from the larger family:

$$p_{A_q}(\mathbf{x}_{A_q}) = p(\mathbf{x}_{A_q\setminus q};\beta)p(x_q|x_{A_q\setminus q}).$$

Now, there may be two different approaches. One may either use CLAP (which is not data efficient) or estimate the marginal (for which the model is not optimal).

Proposition 36. *Asymptotically, the estimation of $E_q(\mathbf{x}_q;\theta_q)$ derived from the conditional estimation can not be better then the the estimation of $E_q(\mathbf{x}_q;\theta_q)$ derived from the marginal estimation.*

Proof. If the estimation of θ_q will be better the estimation of the marginal, one would be able to improve the ML estimator for $p_{A_q}(\mathbf{x}_{A_q}) = p(\mathbf{x}_{A_q\setminus q};\beta)p(x_q|x_{A_q\setminus q})$. This is in contradiction to the Cramer-rao bound achieved by the MLE. □