Linear and Parallel Learning of Markov Random Fields

Yariv Dror Mizrahi[1] YARIV@MATH.UBC.CA
Misha Denil[2] MISHA.DENIL@CS.OX.AC.UK
Nando de Freitas[1,2,3] NANDO@CS.OX.AC.UK
[1]University of British Columbia, Canada
[2]University of Oxford, United Kingdom
[3]Canadian Institute for Advanced Research, CIFAR NCAP Program

Abstract

We introduce a new embarrassingly parallel parameter learning algorithm for Markov random fields which is efficient for a large class of practical models. Our algorithm parallelizes naturally over cliques and, for graphs of bounded degree, its complexity is linear in the number of cliques. Unlike its competitors, our algorithm is fully parallel and for log-linear models it is also data efficient, requiring only the local sufficient statistics of the data to estimate parameters.

1. Introduction

Markov Random Fields (MRFs), also known as undirected probabilistic graphical models, are ubiquitous structured probability models that have significantly impacted a large number of fields, including computer vision (Li, 2001; Szeliski et al., 2008), computational photography and graphics (Agarwala et al., 2004), computational neuroscience (Ackley et al., 1985), bioinformatics (Yanover et al., 2007), sensor networks (Liu & Ihler, 2012), social networks (Strauss & Ikeda, 1990), Markov logic (Richardson & Domingos, 2006), natural language processing (Lafferty et al., 2001; Sutton & McCallum, 2012) and statistical physics (Kindermann & Snell, 1980). As pointed out in Wainwright & Jordan (2008) there are also many applications in 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 (Kindermann & Snell, 1980; Lauritzen, 1996; Bremaud, 2001; Koller & Friedman, 2009; Murphy, 2012).

Despite the great success and impact of these models, fitting them to data remains a formidable challenge. Although the log-likelihood is typically convex in the parameters, the gradient of these models is intractable.

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. The intractability of exact maximum likelihood has prompted the introduction of many approximate methods of parameter estimation (Besag, 1975; Hinton, 2000; Hyvärinen, 2005; Marlin et al., 2010; Varin et al., 2011; Marlin & de Freitas, 2011; Swersky et al., 2011).

An important class of approximate method for this problem are stochastic approximation methods, which approximate the model term by drawing samples from the model distribution, typically via 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. Replacing the joint distribution with a product of conditionals eliminates the

Proceedings of the 31st International Conference on Machine Learning, Beijing, China, 2014. JMLR: W&CP volume 32.

model term from the gradient of the pseudo-likelihood objective, which circumvents the model inefficiency of maximum likelihood estimation. However, pseudo-likelihood is not data efficient, since the conditional distributions often depend on the actual data and the current value of the parameters. We return to this issue in more detail in Section 2.3.

Applying pseudo likelihood in a distributed setting is also 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 & Hero III, 2012; Liu & Ihler, 2012).

In this paper 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, our algorithm is embarrassingly parallel and can be implemented in a distributed setting without modification. Our algorithm replaces 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. Our 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 our 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 similar method 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 revising this paper, 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.

2. Model Specification and Objectives

We are interested in estimating the parameter vector $\boldsymbol{\theta}$ of a positive distribution $p(\mathbf{x} \,|\, \boldsymbol{\theta}) > 0$ that satisfies the Markov properties of an undirected graph G. That is, a distribution that can be represented as a product of factors, one per maximal clique,

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

where $\mathcal{C}$ is the set of maximal cliques of G, $\psi_c(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c) \geq 0$ is the *potential function* or *factor* associated with the variables in clique c, and $Z(\boldsymbol{\theta})$ is the *partition function*: $Z(\boldsymbol{\theta}) = \sum_{\mathbf{x}} \prod_{c \in \mathcal{C}} \psi_c(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c)$. In such models we often use exponential functions to represent the potentials, $\psi_c(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c) = \exp(-E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c))$, where $E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c) \in \mathbb{R}$ is called the *energy*, which we will assume is chosen so that the parameters are identifiable. The resulting joint distribution can then be written as a *Gibbs distribution*

$$p(\mathbf{x} \,|\, \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(-\sum_{c} E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c)).$$

When the energy is a linear function of the parameters, i.e. $E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c) = -\boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{x}_c)$ where $\boldsymbol{\phi}_c(\mathbf{x}_c)$ is a feature vector derived from the values of the variables $\mathbf{x}_c$, we have a *maximum entropy* 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.

Notation: We 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.

2.1. Maximum Likelihood

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

Consider the fully-observed maximum entropy model

$$p(\mathbf{x} \mid \boldsymbol{\theta}) = \frac{1}{Z(\boldsymbol{\theta})} \exp(\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{x})) \tag{2}$$

where c indexes the maximal cliques. The scaled log-likelihood is given by

$$\begin{aligned}\ell(\boldsymbol{\theta}) &= \frac{1}{N}\sum_{n=1}^{N} \log p(\mathbf{x}_n \mid \boldsymbol{\theta}) \\ &= \frac{1}{N}\sum_{n=1}^{N}\left[\sum_c \boldsymbol{\theta}_c^T \boldsymbol{\phi}_c(\mathbf{x}_n) - \log Z(\boldsymbol{\theta})\right]\end{aligned}$$

which is a convex function of $\boldsymbol{\theta}$.

The derivative for the parameters of a particular clique, q, is given by

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}_q} = \frac{1}{N}\sum_{n=1}^{N}\left[\boldsymbol{\phi}_q(\mathbf{x}_n) - \frac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_q}\right] \quad , \tag{3}$$

where

$$\frac{\partial \log Z(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}_q} = \mathbb{E}\left[\boldsymbol{\phi}_q(\mathbf{x}) \mid \boldsymbol{\theta}\right] = \sum_{\mathbf{x}} \boldsymbol{\phi}_q(\mathbf{x}) p(\mathbf{x} \mid \boldsymbol{\theta}). \tag{4}$$

Equation (4) is the expectation of the feature $\boldsymbol{\phi}_q(\mathbf{x})$ over the model distribution. For many models of interest this quantity is intractable.

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

$$\frac{\partial \ell}{\partial \boldsymbol{\theta}_q} = \frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}_q(\mathbf{x}_n) - \mathbb{E}\left[\boldsymbol{\phi}_q(\mathbf{x}) \mid \boldsymbol{\theta}\right] \quad . \tag{5}$$

At the optimum these two terms will be equal and the empirical distribution of the features will match the model predictions.

2.2. Maximum Pseudo-Likelihood

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

$$\ell^{PL}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{n=1}^{N}\sum_{m=1}^{M} \log p(x_{mn} \mid \mathbf{x}_{-mn}, \boldsymbol{\theta}) \tag{6}$$

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 neighbors of node m.) In the binary, log-linear case, the gradient of this objective can be expressed in contrastive form,

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

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).

2.3. Model and Data Efficiency

There are two terms in the gradient of Equation 5. 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, $\mathbb{E}\left[\boldsymbol{\phi}_q(\mathbf{x}) \mid \boldsymbol{\theta}\right]$, 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 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 $p(\bar{x}_{mn}^m \mid \mathbf{x}_{-mn}, \boldsymbol{\theta})$. 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. 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

Algorithm 1 LAP

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

the sub-problems have been solved, it combines the solutions to each sub-problem together into a solution to the full problem.

For a fixed clique q we define its *1-neighbourhood*

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

to contain all of the variables of q itself as well as the variables with at least one neighbour in q.

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 $\boldsymbol{\theta}_q$ 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 $\boldsymbol{\alpha}$ of the auxiliary MRF by maximum likelihood, the relevant derivative is

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

This approach is data efficient, since the sufficient statistics $\frac{1}{N}\sum_{n=1}^{N} \boldsymbol{\phi}_q(\mathbf{x}_{A_q n})$ can be easily precomputed. 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 $\mathbb{E}\left[\boldsymbol{\phi}_q(\mathbf{x}_{A_q})|\boldsymbol{\alpha}\right]$ can be easily computed when the number of variables in A_q is small. To illustrate this point, consider the models shown in Figure 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 & de Freitas, 2011), this cost is acceptable.

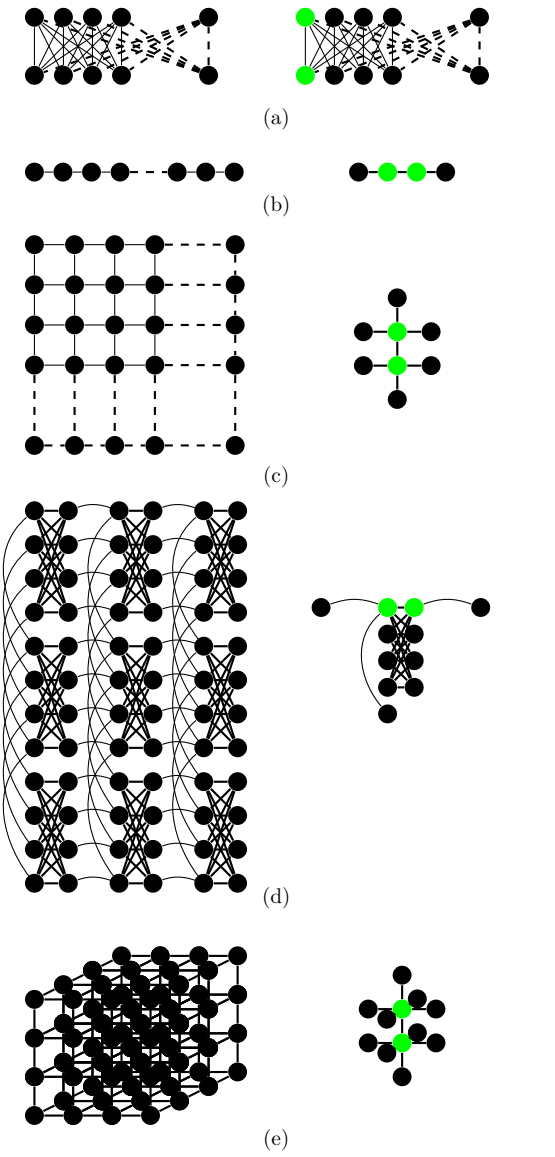

Figure 1. The left column shows several popular MRFs: (a) a restricted Boltzmann machine (RBM), (b) a chain graph, (c) a 2-D Ising grid, (d) a Chimera $3 \times 3 \times 4$ lattice, and (e) a 3-D Ising lattice. The right hand side shows the corresponding 1-neighborhoods A_q for cliques of interest (in green). Models (b) to (e) have small 1-neighborhoods and can learned efficiently with the LAP algorithm.

3.1. 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).

4. 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 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.

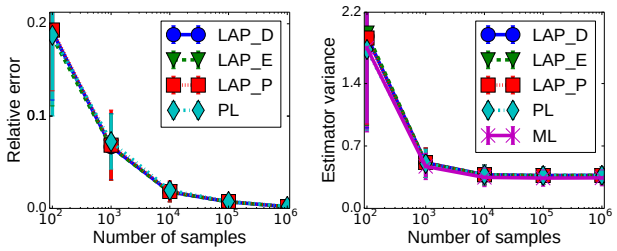

Figure 2. **Left:** Relative error of parameter estimates compared to maximum likelihood for LAP and pseudo-likelihood on a 4×4 Ising grid. Error bars show the standard deviation over several runs. **Right:** Variance of the parameter estimates for each algorithm.

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-neighborhoods, and the results are shown in Figure 2. The second is a $4 \times 4 \times 4$ Ising lattice with 6-neighborhoods, which is shown in Figure 3. Finally, we also consider a Chimera $3 \times 3 \times 3$ model, with results shown in Figure 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

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

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

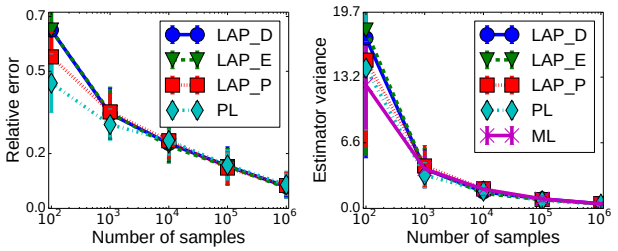

Figure 3. **Left:** Relative error of parameter estimates compared to maximum likelihood for LAP and pseudo-likelihood on a $4 \times 4 \times 4$ Ising lattice. Error bars show the standard deviation over several runs. **Right:** Variance of the parameter estimates for each algorithm.

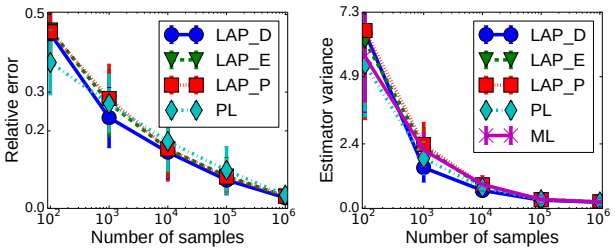

Figure 4. **Left:** Relative error of parameter estimates compared to ML for LAP and pseudo-likelihood on a Chimera $3 \times 3 \times 3$ model. Error bars show the standard deviation over several runs. **Right:** Variance of the parameter estimates for each algorithm.

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.

5. Theory

In this section show that matching parameters in the joint and the marginal distributions is valid, provided the parametrisations 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 $\sum_c E(\mathbf{x}_c|\boldsymbol{\theta}_c)$. The Hammersley-Clifford theorem (Hammersley & Clifford, 1971) 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 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 1. *A Gibbs potential $\{E(\mathbf{x}_c|\boldsymbol{\theta}_c)\}_{c\in\mathcal{C}}$ is said to be normalized with respect to zero if $E(\mathbf{x}_c|\boldsymbol{\theta}_c) = 0$ 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 2. [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.*

5.1. The LAP Argument

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

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

have the same form as the marginal distribution on A_q (with clique system $\mathcal{C}_q$) parametrised so that the potentials are normalized with respect to zero.

We can obtain the marginal from the joint in the following way

$$\begin{aligned} p(\mathbf{x}_{A_q} \,|\, \boldsymbol{\theta}) &= \sum_{\mathbf{x}_{S\setminus A_q}} p(\mathbf{x}_S \,|\, \boldsymbol{\theta}) \\ &= \frac{1}{Z(\boldsymbol{\theta})} \sum_{\mathbf{x}_{S\setminus A_q}} \exp(-\sum_{c\in\mathcal{C}} E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_c)) \\ &= \frac{1}{Z(\boldsymbol{\theta})} \exp(-E(\mathbf{x}_q \,|\, \boldsymbol{\theta}_q) - \sum_{c\in\mathcal{C}_q\setminus\{q\}} E(\mathbf{x}_c \,|\, \boldsymbol{\theta}_{S\setminus q})) \end{aligned}$$

Proposition 3. *If the parametrisations of $p(\mathbf{x}_S \mid \boldsymbol{\theta})$ and $p(\mathbf{x}_{A_q} \mid \boldsymbol{\alpha})$ are chosen to be normalized with respect to zero, and if the parameters are identifiable with respect to the potentials, then $\boldsymbol{\theta}_q = \boldsymbol{\alpha}_q$.*

Proof. The terms $E(\mathbf{x}_q \mid \boldsymbol{\theta}_q)$ and $E(\mathbf{x}_q \mid \boldsymbol{\alpha}_q)$ appear as separate factors in $p(\mathbf{x}_{A_q} \mid \boldsymbol{\theta})$ and $p(\mathbf{x}_{A_q} \mid \boldsymbol{\alpha})$ respectively. By existence and uniqueness of the normalized potentials (Theorem 2), we have

$$E(\mathbf{x}_q \mid \boldsymbol{\alpha}_q) = E(\mathbf{x}_q \mid \boldsymbol{\theta}_q)$$

which implies that $\boldsymbol{\theta}_q = \boldsymbol{\alpha}_q$ if the parameters are identifiable. □

5.2. Consistency of LAP

Let $\boldsymbol{\theta}^\star$ be the true vector of parameters taken from the unknown generating distribution $p(\mathbf{x}_S \mid \boldsymbol{\theta}^\star)$ parametrized such that the potentials are normalized with respect to zero. Suppose we have N samples drawn *iid* from this distribution. Let $\hat{\boldsymbol{\theta}}^{ML}$ be the ML estimate of $\boldsymbol{\theta}$ given the data and let $\hat{\boldsymbol{\alpha}}^{ML}$ the corresponding ML estimate for the auxiliary MRF with true parameters $\boldsymbol{\alpha}^\star$.

Proposition 4. *If the true marginal distributions are contained in the class of auxiliary MRFs, we have for all q that $\hat{\boldsymbol{\alpha}}_q^{ML} \to \boldsymbol{\theta}_q^\star$ as $N \to \infty$.*

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

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

By the lap argument (Proposition 3), we know that $\boldsymbol{\alpha}_q^\star = \boldsymbol{\theta}_q^\star$. Since ML is consistent under smoothness and identifiability assumptions (for example, see Fienberg & Rinaldo (2012)), we also have $\hat{\boldsymbol{\alpha}}^{ML} \to \boldsymbol{\alpha}^\star$, so

$$\hat{\boldsymbol{\alpha}}_q^{ML} \to \boldsymbol{\theta}_q^\star$$

□

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.

5.3. Relationship to ML

Here we prove that, under certain (strong) assumptions, 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 $\mathcal{C}$, and let $q \in \mathcal{C}$ be a particular clique of interest.

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

Lemma 5. *If a distribution $\hat{p}(\mathbf{x}_S)$ satisfies that for each $c \in \mathcal{C}$*

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

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 6. *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

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

and

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

For an arbitrary clique $c \in \mathcal{C}$, either $c \subset S \setminus q$ or $c \subset A_q$, and we see that $\hat{p}(x_c) = \tilde{p}(x_c)$ by further marginalizing one of the above expressions. This shows that our expression for $\hat{p}(\mathbf{x}_S)$ satisfies the criteria of Lemma 5, and is therefore an ML estimate for $\tilde{p}(\mathbf{x}_S)$. □

Suppose we have a family of distributions $\mathcal{F}$ on $\mathbf{x}_S$ which satisfy the Markov properties of the MRF, and suppose that $\hat{p}(\mathbf{x}_S) \in \mathcal{F}$ where $\hat{p}(\mathbf{x}_S)$ is defined as in Proposition 6.

Define the auxiliary family $\mathcal{F}_q$ associated with the clique q as follows.

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

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

Proposition 7. *The auxiliary family $\mathcal{F}_q$ 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 $\mathcal{F}_q$.*

Proof. Recall that $\hat{p}(\mathbf{x}_S)$ from Proposition 6 is in $\mathcal{F}$ by assumption. Thus,

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

is in $\mathcal{F}_q$ by definition. That $\hat{p}(\mathbf{x}_{A_q}) \in \mathcal{F}_q$ is an ML estimate follows since the log likelihood gradient in Equation 5 is zero when the model and empirical distributions are equal. □

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

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

for some domain of parameters $\boldsymbol{\Theta}$, where

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

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

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

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

for some domain of parameters $\boldsymbol{\Psi}$. We will again suppose that this parametrisation 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 $\mathcal{F}$ and $\mathcal{F}_q$, respectively. Since we have chosen the parametrisations of these families to be normalized we also have unique ML parameters $\hat{\boldsymbol{\theta}} \in \boldsymbol{\Theta}$ and $\hat{\boldsymbol{\alpha}} \in \boldsymbol{\Psi}$ such that $p(\mathbf{x}_S \,|\, \hat{\boldsymbol{\theta}}) \in \mathcal{F}(\boldsymbol{\Theta})$ is an ML estimate for $\tilde{p}(\mathbf{x}_S)$ and $p(\mathbf{x}_{A_q} \,|\, \hat{\boldsymbol{\alpha}}) \in \mathcal{F}(\boldsymbol{\Psi})$ is an ML estimate for $\tilde{p}(\mathbf{x}_{A_q})$.

We can now prove the main result of this section.

Theorem 8. *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,*

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

Proof. We have the following sequence of equalities:

$$p(\mathbf{x}_S \,|\, \hat{\boldsymbol{\theta}}) \overset{(1)}{=} \hat{p}(\mathbf{x}_S) \overset{(2)}{=} \frac{\tilde{p}(\mathbf{x}_{A_q})\tilde{p}(\mathbf{x}_{S\setminus q})}{\tilde{p}(\mathbf{x}_{A_q\setminus q})}$$

$$\overset{(3)}{=} \frac{\hat{p}(\mathbf{x}_{A_q})\tilde{p}(\mathbf{x}_{S\setminus q})}{\tilde{p}(\mathbf{x}_{A_q\setminus q})} \overset{(4)}{=} \frac{p(\mathbf{x}_{A_q} \,|\, \hat{\boldsymbol{\alpha}})\tilde{p}(\mathbf{x}_{S\setminus q})}{\tilde{p}(\mathbf{x}_{A_q\setminus q})}$$

The first equality follows from the parametrisation of $\mathcal{F}$, the second follows from Proposition 6, the third from Proposition 7 and the fourth follows from the parametrisation of $\mathcal{F}_q$. The theorem is proved by summing both sides of the equality over $\mathbf{x}_{S\setminus A_q}$. □

Applying the LAP argument (Proposition 3) to Theorem 8 we see that $\hat{\boldsymbol{\theta}}_q = \hat{\boldsymbol{\alpha}}_q$.

Remark: The assumption that $\hat{p}(\mathbf{x}_S) \in \mathcal{F}$ 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.

6. Conclusion

We have presented a distributed learning algorithm for practical MRFs, where the parameters of each clique can be estimated in different machines. The algorithm is also data efficient in log-linear models, since the estimation of each clique parameter only requires access to local sufficient statistics of the data. Not only are the statistics local to the 1-neighborhoods of each clique, but they can also be precomputed.

Our experiments indicate that the LAP estimators behave similarly to pseudo-likelihood and maximum likelihood for large sample sizes. However, these alternative estimators do not enjoy the same data and model efficiencies as LAP. Finally, we proved that the proposed estimator is consistent.

This work opens up many directions for future work, including the application of LAP to model selection problems, models with latent variables, and models with tied parameters. Since LAP is fully parallel, our experiments focused on the question of statistical efficiency. However, implementations on distributed computing platforms, such as Apache Spark/Hadoop, would be very valuable. A further addition to the theory would be the derivation of PAC bounds to improve our understanding of the sampling complexity of these estimators.

Acknowledgements

We would like to thank Alexandre Bouchard-Côté, Paul Fearnhead, Joel Friedman, Eldad Haber, Fred Roosta, Luis Tenorio and the anonymous reviewers for greatly helping us to improve this work. We thank NSERC for financial support.

References

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

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

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

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

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

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

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

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

Hammersley, J. M. and Clifford, P. Markov fields on finite graphs and lattices. *Unpublished manuscript*, 1971.

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

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

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

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

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

Lafferty, J. D., McCallum, A., and Pereira, F. C. N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*, pp. 282–289, 2001.

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

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

Liu, Q. and Ihler, A. Distributed parameter estimation via pseudo-likelihood. In *ICML*, 2012.

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

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

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

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

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

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

Richardson, M. and Domingos, P. Markov logic networks. *Machine Learning*, 62(1-2):107–136, 2006.

Strauss, D. and Ikeda, M. Pseudolikelihood estimation for social networks. *Journal of the American Statistical Association*, 85(409):204–212, 1990.

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

Swersky, K., Ranzato, M.A., Buchman, D., Marlin, B., and Freitas, N. On autoencoders and score matching for energy based models. In *ICML*, pp. 1201–1208, 2011.

Szeliski, R., Zabih, R., Scharstein, D., Veksler, O., Kolmogorov, V., Agarwala, Aseem, Tappen, M., and Rother, C. 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.

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

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

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

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

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