On the Use of Variational Inference for Learning Discrete Graphical Models

Eunho Yang EUNHO@CS.UTEXAS.EDU
Pradeep Ravikumar PRADEEPR@CS.UTEXAS.EDU
Department of Computer Science, The University of Texas, Austin, TX 78712, USA

Abstract

We study the general class of estimators for graphical model structure based on optimizing ℓ_1-regularized approximate log-likelihood, where the approximate likelihood uses tractable variational approximations of the partition function. We provide a message-passing algorithm that *directly* computes the ℓ_1 regularized approximate MLE. Further, in the case of certain reweighted entropy approximations to the partition function, we show that surprisingly the ℓ_1 regularized approximate MLE estimator has a *closed-form*, so that we would no longer need to run through many iterations of approximate inference and message-passing. Lastly, we analyze this general class of estimators for graph structure recovery, or its *sparsistency*, and show that it is indeed sparsistent under certain conditions.

1. Introduction

A Markov random field (MRF) over a p-dimensional discrete random vector $X = (X_1, X_2, \ldots, X_p)$ is specified by an undirected graph $G = (V, E)$, with vertex set $V = \{1, 2, \ldots, p\}$ – one for each variable – and edge set $E \subset V \times V$. The structure of this graph encodes conditional independence assumptions among subsets of the variables. In structure learning, the task is to estimate this underlying graph from n independent and identically distributed samples $\{x^{(1)}, x^{(2)}, \ldots, x^{(n)}\}$.

Recent results on the discrete graphical model structure learning problem have built on a natural connection between such structure learning and *parameter estimation*: indeed, learning the graph structure is equivalent to learning the parameters of a fully saturated graphical model under the assumption of a *sparse* set of parameters, where parameters underlying non-edges are equal to zero. Such parameter estimation in turn hinges crucially on graphical model *inference*: since it involves solving optimization problems that require computing quantities such as the partition function, or the marginals.

Thus, structure learning when reduced to sparse parameter estimation hinges on two components: sparsity recovering regularization methods, and methods for approximate inference in graphical models. While methods that have been proposed in this class of approaches use state of the art sparsity recovery methods (ℓ_1-regularization and the like), their approximate inference components are far from the state of the art. For instance, Ravikumar et al. (2010) use node-wise logistic regressions to estimate node-neighborhoods, which can be thought of as employing a pseudo-likelihood approximation (with asymmetric edge parameters) of the partition function in the log-likelihood. Lee et al. (2007) compute approximate estimates of the gradient using Belief Propagation (Pearl, 1988), which can be thought of as using a Bethe entropy approximation of the partition function. The state of the art in approximate inference on the other hand involves convex variational approximations of the entropy of the graphical model, and dual decompositions involving tractable subcomponents of the graphical model.

We thus have a gap: between state of the art in inference and the use of inference in state of the art in structure learning methods. Towards this, we study a *general* class of estimators for graphical model structure that use tractable approximations of the partition function and ℓ_1-regularization. The resulting optimization problem can be solved naturally using composite gradient descent, since the gradients of the log-likelihood take the form of marginals which can be

Appearing in *Proceedings of the 28^{th} International Conference on Machine Learning*, Bellevue, WA, USA, 2011.
Copyright 2011 by the author(s)/owner(s).

approximated using the approximate inference procedure corresponding to the partition function approximation. However this involves performing graphical model inference for each of the gradient steps, which could be expensive. To address this, we provide a message-passing algorithm that *directly* computes the solution that optimizes the ℓ_1 regularized approximate log-likelihood. Further, in the case of certain reweighted entropy approximations to the partition function such as the tree-reweighted approximation, we show that surprisingly the ℓ_1 regularized approximate MLE estimator has a *closed-form*, so that we would no longer need to run through many iterations of approximate inference and message-passing. Lastly, we analyze this general class of estimators for graph structure recovery, or its *sparsistency*. For such estimators, one might imagine that even though the approximate inference is tight with respect to the partition function or the marginals, the corresponding approximate MLE need not even be consistent; see (Wainwright, 2006) for instance for the case of a weakly regularized approximate MLE. However, note that this is also the case with high-dimensional sparse parameter estimation where typical MLE estimators are not consistent, *unless* one carefully chooses the magnitude of ℓ_1 regularization (Candes & Tao, 2007; Tropp, 2006). Indeed, even for our general class of ℓ_1 regularized *approximate* log-likelihood estimators, we show that under certain conditions on the egde weights, the methods do succeed in recovering the graph structure. Indeed, the development in this paper raises the research agenda of tuning approximate inference procedures to structure learning by developing partition function approximations that would impose the weakest conditions on the parameters.

2. Review, Setup and Notation

2.1. Markov Random Fields

Let $X = (X_1, \ldots, X_p)$ be a random vector, each variable X_i taking values in a discrete set $\mathcal{X}$ of cardinality m. Let $G = (V, E)$ denote a graph with p nodes, corresponding to the p variables $\{X_1, \ldots, X_p\}$. A pairwise Markov random field over $X = (X_1, \ldots, X_p)$ is then specified by nodewise and pairwise functions $\theta_s : \mathcal{X} \mapsto \mathbb{R}$ for all $s \in V$, and $\theta_{st} : \mathcal{X} \times \mathcal{X} \mapsto \mathbb{R}$ for all $(s,t) \in E$, as

$$\mathbb{P}(x) \propto \exp\Big\{ \sum_{s\in V} \theta_s(x_s) + \sum_{(s,t)\in E} \theta_{st}(x_s, x_t). \Big\} \quad (1)$$

In this paper, we largely focus on the case where the variables are binary with $\mathcal{X} = \{-1, +1\}$, where we can rewrite (1) to the Ising model form (Ising, 1925)

$$\mathbb{P}(x) \propto \exp\bigg\{ \sum_{s\in V} \theta_s x_s + \sum_{(s,t)\in E} \theta_{st} x_s x_t \bigg\}, \quad (2)$$

for some set of parameters $\{\theta_s\}$ and $\{\theta_{st}\}$. It will be useful to rewrite (2) as the member of an exponential family, $\mathbb{P}(x) = \exp\left(\langle \theta, \phi \rangle - A(\theta)\right)$, where $\phi(x) = \{\phi_s(x) = x_s;\ \phi_{st}(x) \equiv x_s x_t\}$ are the set of Ising potentials, and $\theta = \{\theta_s;\ \theta_{st}\}$ are the corresponding set of parameters; and $A : \Theta \mapsto \mathbb{R}$ is the log of the normalization constant, also called the log-partition function, $A(\theta) = \log \sum_{x\in\mathcal{X}^p} \exp(\langle \theta, \phi \rangle)$.

2.2. Variational Approximations

A complication for discrete undirected graphical models is that typical inference tasks, even calculation of the log-partition function $A(\theta)$, is computationally intractable. Here, we briefly review variational approximations to the partition function, following the development in Wainwright & Jordan (2008). By properties of exponential families, the moments $\mu(\theta) = \mathbb{E}_\theta(\phi) = \nabla A(\theta)$. Denote the conjugate of the log-partition function by $A^*(\mu) = \sup_{\theta\in\Theta} \langle \theta, \mu \rangle - A(\theta)$. It can be shown that $A^*(\mu)$ is the negative entropy of the graphical model distribution with parameter $\theta = (\nabla A)^{-1}(\mu)$. Consider the set of all possible mean parameters, $\mathcal{M} = \{\mu : \exists \text{ distribution } p \text{ s.t.} \mathbb{E}_p(\phi) = \mu\}$, which is also called the *marginal polytope* of the graphical model. Then, by convex duality, we can write,

$$A(\theta) = \sup_{\mu\in\mathcal{M}} \langle \theta, \mu \rangle - A^*(\mu). \quad (3)$$

(3) thus provides a variational formulation of the log-partition function $A(\theta)$. Following the development in Wainwright & Jordan (2008), it is easier to describe approximations to this log-partition function using so called *overcomplete representations*.

Since $\mathcal{X}$ is discrete, any potential function θ_c can be parameterized as linear combinations of $\{0,1\}$-valued indicator functions. For each $s \in V$ and $j \in \{1, \ldots, m\}$, we can define node-wise indicators, $[x_s = j] = 1$ if $x_s = j$ and equal to 0 otherwise. With this notation, any set of potential functions can then be written as $\theta_s(x_s) = \sum_{j\in[m]} \theta_{s;j}\ [x_s = j]$ for $s \in V$ and $\theta_{st}(x_s, x_t) = \sum_{j,k\in[m]} \theta_{st;jk}\ [x_s, x_t = j, k]$ for $(s,t) \in E$. Thus, (1) can be rewritten as, $\mathbb{P}(x) \propto \exp\Big\{ \sum_{s\in V; j\in[m]} \theta_{s;j} [x_s = j] + \sum_{(s,t)\in E; j,k\in[m]} \theta_{st;jk} [x_s = j,\ x_t = k] \Big\}$, for a set of parameters $\theta := \{\theta_{s;j}, \theta_{st;jk} : s, t \in V;\ (s,t) \in E;\ j, k \in [m]\}$.

Given these sufficient statistics, the mean parameters $\{\mu_{s;j}\}$ and $\{\mu_{st;jk}\}$ are just the node and pairwise marginals. Wainwright & Jordan (2008) then describe variational approximations of the log-partition function $A(\theta)$, as involving approximating the two intractable components in its variational formulation (3) (a) the marginal polytope $\mathcal{M}$, and (b) the graphical model entropy $A^*(\mu)$.

Any variational approximation to the log-partition function (3) can then be written as,

$$B(\theta) = \sup_{\mu\in\mathcal{L}} \langle \theta, \mu \rangle - B^*(\mu), \quad (4)$$

where $\mathcal{L}$ is a tractable bound on the marginal polytope $\mathcal{M}$, and $B^*(\mu)$ is a tractable approximation to the graphical model entropy $A^*(\mu)$. A popular bound L_G of the marginal polytope is given by

$$L_G = \{\mu \Big| \sum_j \mu_{s;j} = 1, \sum_k \mu_{st;jk} = \mu_{s;j}; \quad s,t \in V; j,k \in [m]\}. \tag{5}$$

Popular approximations to the negative entropy use weighted sums of node and edge-entropies. Let $H_s := \sum_{x_s \in \mathcal{X}} \mu_s(x_s) \log \mu_s(x_s)$ and $H_{st}(\mu_{st}) := \sum_{(x_s,x_t)\in\mathcal{X}\times\mathcal{X}} \mu_{st}(x_s,x_t) \log \mu_{st}(x_s,x_t)$ denote the node-based and edge-based negative entropies, respectively, and $I_{st}(\mu_{st}) := H_{st}(\mu_{st}) - H_s(\mu_s) - H_t(\mu_t)$.

The Bethe approximation (Yedidia et al., 2001) to the entropy $A^*(\mu)$ is given by $B^*_{\text{bethe}}(\mu) = \sum_s H_s(\mu_s) - \sum_{st} I_{st}(\mu_{st})$. The tree-reweighted entropy (Wainwright et al., 2003) in turn is given by

$$B^*_{\text{trw}}(\mu) = \sum_s H_s(\mu_s) - \sum_{st} \rho_{st} I_{st}(\mu_{st}), \tag{6}$$

where ρ_{st} are edge-weights that lie in a so-called spanning tree polytope. Heskes (2006); Weiss et al. (2007) also discuss general convex entropic forms, the simplest of which are the weighted forms $B^*_\alpha(\mu) = \sum_{s\in V} \alpha_s H_s(\mu_s) + \sum_{(s,t)\in E} \alpha_{st} H_{st}(\mu_{st})$, for some weights $\{\alpha_s, \alpha_{st} \geq 0\}$.

The approximation given by $B_{\text{bethe}}(\theta) = \sup_{\mu\in L_G} \langle \theta, \mu \rangle - B^*_{\text{bethe}}(\mu)$, underlies belief propagation, while $B_{\text{trw}}(\theta) = \sup_{\mu\in L_G} \langle \theta, \mu \rangle - B^*_{\text{trw}}(\mu)$ yields the tree-reweighted approximation to the log-partition function.

3. Graphical Model Selection

Suppose that we are given a collection $D := \{x^{(1)}, \ldots, x^{(n)}\}$ of n samples, where each p-dimensional vector $x^{(i)} \in \{1,\ldots,m\}^p$ is drawn i.i.d. from a distribution $\mathbb{P}_{\theta^*}$ of the form (2), for parameters θ^* and graph $G = (V, E^*)$ over the p variables. The goal of *graphical model selection* is to infer the edge set E^* of the graphical model defining the probability distribution that generates the samples. Note that the true edge set E^* can also be expressed as a function of the parameters as

$$E^* = \{(s,t) \in V \times V : \theta^*_{st} \neq 0\}. \tag{7}$$

Given the data, $D := \{x^{(1)}, \ldots, x^{(n)}\}$, the ℓ_1 regularized MLE can then be written as the solution of the optimization problem,

$$\widehat{\theta} \in \arg\min_\theta -\langle \theta, \widehat{\phi} \rangle + A(\theta) + \lambda \|\theta\|_{1,E}, \tag{8}$$

where $\widehat{\phi} = \frac{1}{n}\sum_{i=1}^n \phi(x^{(i)})$ is the average of the sufficient statistics, and where $\|\cdot\|_{1,E}$ is the ℓ_1 norm of just the edge-parameters, so that $\|\theta\|_{1,E} = \sum_{s\neq t} |\theta_{st}|$.

The caveat with solving (8) is the intractable computation of the log-partition function $A(\theta)$. We thus consider the following class of M-estimators:

$$\widehat{\theta} \in \arg\min_\theta -\langle \theta, \widehat{\phi} \rangle + B(\theta) + \lambda \|\theta\|_{1,E}, \tag{9}$$

where $B(\theta)$ is a variational approximation to the log-partition function of the form (3) outlined in the previous section. Given the solution $\widehat{\theta}$, we can then estimate the graph structure as $\hat{E} = \{(s,t) : \widehat{\theta}_{st} \neq 0\}$.

4. Optimization Methods

We now consider the task of solving the optimization problem in (9).

4.1. Gradient based methods

Let us first go back to the exact ℓ_1-regularized MLE problem (8), and consider an approximate technique to solve this intractable optimization problem. In particular, as Lee et al. (2007) suggest we could solve (8) using approximate estimates of the gradient, computed using Belief Propagation. One could then perform the following *approximate* composite gradient descent (Nesterov, 2004):

Algorithm 1 Solving (8) Using Approximate Marginals

for $t = 1, 2, \ldots$ **do**

$\theta^{t+1} = \mathcal{S}^{(E)}_{\lambda\eta^t}\left(\theta^t - \eta^t(-\widehat{\phi} + \mu^{approx}(\theta^t))\right).$

end for

where $\mu^{approx}(\theta^t)$ are approximate estimates of marginals.

Here $\mathcal{S}^{(E)}_r$ denotes the soft-thresholding function applied to the edge elements, so that for $\alpha \in E$, $[\mathcal{S}^{(E)}_r(w)]_\alpha = \text{sign}(w_\alpha)\max\{|w_\alpha| - \lambda, 0\}$. Such composite gradient descent has been shown to have at least sublinear convergence provided the step-sizes are chosen appropriately (Nesterov, 2004). Indeed, one way to view these iterates given objective $f(\theta)$ is as minimizing a composite quadratic approximation: $\min_\theta \nabla f(\theta^t).(\theta - \theta^t) + 1/\eta^t \|\theta - \theta^t\|_2^2 + \lambda\|\theta\|_{1,E}$. Vanilla gradient descent on the other hand solves $\min_\theta \nabla f(\theta^t).(\theta - \theta^t) + 1/\eta^t \|\theta - \theta^t\|_2^2$.

Proposition 1. *Suppose that in Algorithm 1 the approximate marginals $\mu^{approx}(\theta)$ satisfy $\mu^{approx}(\theta) = \nabla B(\theta)$, for some approximation $B(\theta)$ to the log-*

partition function $A(\theta)$, so that they are pseudo-moments under the approximation $B(\theta)$. Then the fixed point of Algorithm 1 if any is a local minimum of the optimization problem in (9) with $B(\theta)$ as the log-partition function approximation.

Thus, estimating the marginals using belief propagation as in Lee et al. (2007) corresponds to a Bethe entropy approximation to the partition function $B_{\text{bethe}}(\theta)$.

4.2. Message Passing Updates

In the sequel, we derive a message-passing algorithm for solving the estimator in (9). It will again be useful to consider the overcomplete representation as outlined in Section 2.2. We overload notation and continue to use $\widehat{\phi}$ and θ for the sufficient statistics and parameters. We thus solve the ℓ_1 regularized optimization problem

$$\widehat{\theta} \in \arg\min_{\theta} -\langle\, \theta, \widehat{\phi}\,\rangle + B(\theta) + \lambda\|\theta\|_{1,E}. \tag{10}$$

Note that even though the overcomplete representation is not identifiable, the added ℓ_1-regularization makes the solution unique provided the smooth component is strictly convex. Indeed, for the binary Ising model case, we have the equivalence:

Proposition 2. *Suppose that $\widetilde{\theta}$ is the unique solution of the approximate MLE (9) for the Ising model with regularization penalty 4λ. Then the overcomplete estimator in (10) with regularization penalty λ has a unique solution $\widehat{\theta}$ given by*

$$\widehat{\theta}_{st}(x_s, x_t) = \widetilde{\theta}_{st} x_s x_t.$$

Now, by duality, this overcomplete approximate MLE (10) can be rewritten as

$$\inf_{\theta} \sup_{Z \in C} -\langle\, \theta, \widehat{\phi}\,\rangle + B(\theta) + \langle\, \theta, Z\,\rangle,$$

where $C := \{Z \subseteq \Theta : \|Z\|_\infty \leq \lambda, Z_N = 0\}$, where we use Z_N to denote the coordinates of Z corresponding to node potentials. By strong duality (Boyd & Vandenberghe, 2004), this in turn can be rewritten as

$$\sup_{Z \in C} \inf_{\theta} -\langle\, \theta, \widehat{\phi} - Z\,\rangle + B(\theta) = -\inf_{\mu \in W} B^*(\mu), \tag{11}$$

where $W := \{\mu : \mu \in L_G;\, \mu_N = \phi_N;\, \|\mu_E - \phi_E\|_\infty \leq \lambda\}$, where L_G is the approximation to the marginal polytope $\mathcal{M}$ underlying the variational approximation $B(\theta)$, as outlined in Section 2.2.

We note that this optimization problem is very similar to the typical variational optimization problems for approximation of the partition function, where a linear term and the entropy are optimized with μ constrained to the outer polytope L. These latter optimization problems are typically solved using graph-structured message-passing algorithms, and here, we derive a message passing algorithm that solves the above objective instead, so that it would obtain the ℓ_1-regularized approximate MLE in one shot in contrast to the iterative message passing in the previous section.

Towards this, we use an iterative projection method (Censor & Zenios, 1988), that iteratively projects the primal variables onto individual constraints, while maintaining dual feasibility.

Now, note that the dual in (11) of the ℓ_1 regularized approximate MLE has the following form:

$$\begin{aligned} &\inf_{\mu} B^*(\mu) \\ &s.t.\ \langle\, a_i, \mu\,\rangle = b_i, \quad i = 1, \ldots, m. \\ &\quad l_j \leq \langle\, c_j, \mu\,\rangle \leq u_j, \quad j = 1, \ldots, r, \end{aligned}$$

which has a mix of linear equality and some interval constraints.

In the Supplementary Material, we outline a row-action algorithm for this class of optimization problems that use iterative projections (with corrections to maintain dual-feasibility) onto these interval constraints (Censor & Zenios, 1988). We briefly outline this algorithm below for completeness, but the reader could skip to the next section, where we describe these updates for convex entropic approximations of the partition function.

It will be useful to define the following notation: Given a convex function f, an iterate, x, and a linear constraint $h \equiv \langle\, x, a\,\rangle = b$, suppose we project x onto the hyperplane defining the equality constraint, under the Bregman divergence induced by f. This can be rewritten quite simply as computing y such that:

$$\begin{aligned} \nabla f(y) &= \nabla f(x) - \theta a, \\ \langle\, y, a\,\rangle &= b. \end{aligned}$$

We are interested in the value of θ above; let us denote this by $\Pi_D(f; x, h)$. We can now detail the row-action algorithm:

Let Φ denote the $(m + r) \times p$ matrix with rows as $\{a_j\}_{j=1}^m$ stacked above $\{c_j\}_{j=1}^r$.

Initialization: (μ^0, z^0) such that $\nabla B^*(\mu^0) = -\Phi^T z^0$.

Iterative Step: Given μ^t and z^t, and the current constraint h_j corresponding to the j-th row of Φ, calculate the next primal and dual iterates μ^{t+1} and z^{t+1} as

$$\nabla B^*(\mu^{t+1}) = \nabla B^*(\mu^t) + d^t \Phi_j, \\ z^{t+1} = z^t - d^t e_j,$$

where if the constraint $h_j \equiv \langle a_j, \mu \rangle = b_j$ is a linear equality, then $d^t = \Pi_D(B^*; \mu^t, h_j)$; and if the constraint $h_j \equiv l_j \leq \langle c_j, \mu \rangle \leq u_j$ is an interval constraint, then denoting $h_j^- \equiv \langle c_j, \mu \rangle = l_j$ and $h_j^+ \equiv \langle c_j, \mu \rangle = u_j$, then $d^t = \text{median}(u^t, A_t, B_t)$, where $A_t = \Pi_D(B^*; \mu^t, h_j^-)$ and $B_t = \Pi_D(B^*; \mu^t, h_j^+)$.

4.2.1. Weighted Entropic Approximation

Recall from Section 2.2 the general weighted free-energy approximation consisting of a weighted sum of negative entropies

$$B^*_\alpha(\mu) = \sum_{s \in V} \alpha_s H_s(\mu_s) + \sum_{(s,t) \in E} \alpha_{st} H_{st}(\mu_{st}), \quad (12)$$

where $H_s := \sum_{x_s \in \mathcal{X}} \mu_s(x_s) \log \mu_s(x_s)$ and $H_{st}(\mu_{st}) := \sum_{(x_s,x_t) \in \mathcal{X} \times \mathcal{X}} \mu_{st}(x_s,x_t) \log \mu_{st}(x_s,x_t)$ denote the node-based and edge-based negative entropies respectively. Now, consider the variational approximation given by,

$$B_\alpha(\theta) = \sup_{\mu \in L_G} \langle \theta, \mu \rangle - B^*_\alpha(\mu),$$

where L_G is the marginal polytope outer bound in (5). The optimization problem in (11) then becomes:

$$\begin{aligned} \inf_\mu & \Big\{ \sum_s \alpha_s \sum_{x_s} \mu_s(x_s) \log \mu_s(x_s) \\ & \quad + \sum_{st} \alpha_{st} \sum_{x_s,x_t} \mu_{st}(x_s,x_t) \log \mu_{st}(x_s,x_t) \Big\} \\ \text{s.t. } & \mu_s(x_s) = \widehat{\phi}_s(x_s), \ \mu_{st}(x_s,x_t) \geq 0, \\ & \sum_{x_t} \mu_{st}(x_s,x_t) = \widehat{\phi}_s(x_s). \\ & \widehat{\phi}_{st}(x_s,x_t) - \lambda \leq \mu_{st}(x_s,x_t) \leq \widehat{\phi}_{st}(x_s,x_t) + \lambda. \end{aligned}$$

Thus, we have a set of equality marginalization constraints, and some box constraints. We can then apply the algorithm template above in the previous section to get the set of iterative updates in Algorithm 2.

4.3. Closed Form Updates

The stationary condition characterizing the solution of (9) is given by

$$-\widehat{\phi} + \widehat{\mu}(\widehat{\theta}) + \lambda \widehat{Z}(\widehat{\theta}) = 0, \quad (14)$$

where $\widehat{\mu} = \nabla B(\widehat{\theta})$, and $\widehat{Z}$ is a subgradient vector, with $\widehat{Z}_N = 0$, and $\widehat{Z}_E \in \partial \|\widehat{\theta}_E\|_1$, so that

Algorithm 2 Weighted Entropic Message Passing

Initialization:

$$\begin{aligned} \mu_s(x_s) &= \widehat{\phi}(x_s), & (13a) \\ \mu^{(0)}_{st}(x_s,x_t) &= \widehat{\phi}(x_s,x_t), & (13b) \\ Z_{st}(x_s,x_t) &= -\alpha_{st}(\log \widehat{\phi}(x_s,x_t) + 1). & (13c) \end{aligned}$$

repeat
  for each edge $(s,t) \in E$ **do**

$$\mu^{(t+1)}_{st}(x_s,x_t) = \mu^{(t)}_{st}(x_s,x_t) \Big(\frac{\widehat{\phi}_s(x_s)}{\sum_{x_t} \mu^{(t)}_{st}(x_s,x_t)} \Big),$$

$$\mu^{(t+1)}_{st}(x_s,x_t) = \mu^{(t+1)}_{st}(x_s,x_t) \Big(\frac{\widehat{\phi}_t(x_t)}{\sum_{x_s} \mu^{(t+1)}_{st}(x_s,x_t)} \Big).$$

$$\Delta_+ = \alpha_{st} \log \frac{(\widehat{\phi}_{st}(x_s,x_t) + \lambda)}{\mu^{(t+1)}_{st}(x_s,x_t)},$$

$$\Delta_- = \alpha_{st} \log \frac{(\widehat{\phi}_{st}(x_s,x_t) - \lambda)}{\mu^{(t+1)}_{st}(x_s,x_t)}.$$

$$C_{st}(x_s,x_t) = \text{median}(Z_{st}(x_s,x_t), \Delta_+, \Delta_-).$$

$$Z_{st}(x_s,x_t) = Z_{st}(x_s,x_t) - C_{st}(x_s,x_t),$$

$$\mu^{(t+1)}_{st}(x_s,x_t) = \mu^{(t+1)}_{st}(x_s,x_t) \exp(C_{st}(x_s,x_t)/\alpha_{st}).$$

  end for
until convergence

(a) if $\hat{\theta}_{st;jk} \neq 0$, then $\widehat{Z}_{st;jk} = \text{sign}(\hat{\theta}_{st;jk})$,

(b) if $\hat{\theta}_{st;jk} = 0$, then $|\widehat{Z}_{st;jk}| \leq 1$,

(c) $\widehat{Z}_{s;j} = 0$. That is, $\widehat{\mu}_{s;j} = \widehat{\phi}_{s;j}$.

Now suppose we use the tree-reweighted entropy approximation (6) as the variational approximation to the log-partition function. An interesting property satisfied by the pseudo-moments $\widehat{\mu}$ of this approximation is a reparameterization condition (Wainwright et al., 2003), so that tuple $(\widehat{\theta}, \widehat{\mu})$ is a valid primal dual pair iff they satisfy

$$\widehat{\theta}_{s;j} = \log \widehat{\mu}_{s;j} + C_{sj} + C_s \quad (15)$$

$$\widehat{\theta}_{st;jk} = \rho_{st} \log \frac{\widehat{\mu}_{st;jk}}{\widehat{\mu}_{s;j} \cdot \widehat{\mu}_{t;k}} - C_{sj} - C_{tk}, \quad (16)$$

for some constants $\{C_s, C_{sj}\}$, and further that $\widehat{\mu}$ lies in the pseudomarginal polytope L_G (5), so that

$$\sum_k \widehat{\mu}_{st;jk} = \widehat{\mu}_{s;j}, \ \widehat{\mu}_{st;jk} \geq 0. \quad (17)$$

Now consider the three cases of of the sign of any edge parameter $\widehat{\theta}_{st;jk}$:

(a) $\widehat{\theta}_{st;jk} > 0$: Then from (14), we get $\widehat{\mu}_{st;jk} = \widehat{\phi}_{st;jk} - \lambda$, and substituting this in (16), we get $\widehat{\phi}_{st;jk} > \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} \exp(C_{tk}) + \lambda$.

(b) Similarly, if $\widehat{\theta}_{st;jk} < 0$, then $\widehat{\mu}_{st;jk} = \widehat{\phi}_{st;jk} + \lambda$, which entails that: $\widehat{\phi}_{st;jk} < \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} \exp(C_{tk}) - \lambda$.

(c) Finally, if $\widehat{\theta}_{st;jk} = 0$, then $|\widehat{\phi}_{st;jk} - \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} exp(C_{tk})| < \lambda$, and where $\widehat{\mu}_{st;jk} = \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} \exp(C_{tk})$.

4.3.1. Explicit Construction

Given any constants $\{C_s, C_{sj}\}$, suppose we write out a tuple $(\widehat{\mu}, \widehat{\theta}, \widehat{Z})$ as follows:

(a) If $\widehat{\phi}_{st;jk} > \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} \exp(C_{tk}) + \lambda$, then $\widehat{\mu}_{st;jk} = \widehat{\phi}_{st;jk} - \lambda$, $\widehat{Z}_{st;jk} = 1$;

(b) If $\widehat{\phi}_{st;jk} < \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} \exp(C_{tk}) - \lambda$, then $\widehat{\mu}_{st;jk} = \widehat{\phi}_{st;jk} + \lambda$, $\widehat{Z}_{st;jk} = -1$;

(c) Otherwise, $\widehat{\mu}_{st;jk} = \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} \exp(C_{tk})$, and $\widehat{Z}_{st;jk} = \widehat{\phi}_{st;jk} - \widehat{\phi}_{sj} \exp(C_{sj})\widehat{\phi}_{tk} \exp(C_{tk})$.

(d) Set $\widehat{\theta}$ from (15),(16).

Note that the tuple $(\widehat{\theta}, \widehat{\mu})$ satisfies the stationary condition (14) by construction. The subgradient condition $\widehat{Z} \in \partial\|\widehat{\theta}\|_1$ also holds by construction. Thus, this is a valid tuple and $\widehat{\theta}$ is the solution of (14) provided the resulting $\widehat{\mu} \in L$. The goal then is to derive constants $\{C_s, C_{sj}\}$ so that the $\widehat{\mu}$ constructed as above lies in the polytope L_G.

While this is difficult to do in general, we show that when the variables are binary, the appropriate constants $\{C_s, C_{s;j}\}$ can be written out explicitly.

Proposition 3. *When the variables are binary, the solution $\widehat{\theta}$ of (9) can be computed in closed form. Specifically, setting $\{C_s = 0, C_{sj} = 0\}$ in the construction above yields solution $\widehat{\theta}$, with pseudomarginal $\widehat{\mu}$.*

The proof consists of showing that with the constants set to zero, the pseudomarginal $\widehat{\mu}$ resulting from the construction satisfies $\widehat{\mu} \in L$; and is provided in detail in the Supplementary Material.

As noted in Proposition 2, even though we have an overcomplete parameterization, the resulting solution has the Ising form, and in particular is equivalent to solving a corresponding ℓ_1-regularized approximate MLE in (9) for the Ising model. Further, the solution $\widehat{\theta}$ is available in closed-form, which thus leads to this very simple estimator for the structure of graph:

$$\widehat{E} = \{(s,t) : \exists j,k;\ |\widehat{\phi}_{st;jk} - \widehat{\phi}_{s;j} \cdot \widehat{\phi}_{t;k}| > \lambda\}. \quad (18)$$

Similar algorithms based on correlations have been proposed elsewhere, see Montanari & Pereira (2009) for instance for the case of a homogeneous Ising model, where the edge parameters are all equal. Here, we obtain the very interesting connection between the specific correlation based edge-detection algorithm above and the tree-reweighted entropy based approximate MLE. This thus opens up new avenues for analyzing such methods, see for instance our sparsistency analysis in Section 5.

Discussion. Note that the closed form estimator in (18) requires time scaling as $O(p^2)$, whereas other state of the art methods such as (Ravikumar et al., 2010) require time that could scale as $O(p^5)$. As we will see in the experiments section in spite of this, their graph structure recovery performance is nonetheless comparable.

5. Sparsistency

In this section, we show that any structure learning estimator obtained as a solution of (9) given any partition function approximation $B(\theta)$ is *sparsistent* under certain conditions. While the analysis builds on standard tools such as the dual-witness technique from (Wainwright, 2009), here we face multiple subtleties: the objective function in (9) does not arise from the likelihood of the data, which has the consequence that the gradient of the objective need not be small at the true parameter. Moreover, the number of non-zero elements here equal the number of edges which scale at least linearly with the number of nodes; so that we needed additional tools such as Brouwers fixed point theorem (Ortega & Rheinboldt, 2000). Lastly, the sparsistency theorem holds not just for one, but a whole family of estimators obtained as solutions of (9).

Let us first study the stationary condition characterizing the solution of (9):

$$-\widehat{\phi} + \widehat{\mu} + \lambda\widehat{Z} = 0. \quad (19)$$

We will now introduce some notation to simplify this condition. Let $\mu^* = \nabla A(\theta^*)$ be the true marginals and let $\bar{\mu} = \nabla B(\theta^*)$ be the "true" variational pseudo-marginals. Denote $W_1 = \widehat{\phi} - \mu^*$ and $W_2 = \mu^* - \bar{\mu}$, and further that $W = W_1 + W_2$. Then, the stationary condition can be rewritten as,

$$\widehat{\mu} - \bar{\mu} - W + \lambda Z = 0. \tag{20}$$

We define the second-order Taylor's expansion remainder of ∇B around θ^* as,

$$R(\Delta; \theta^*) = \nabla B(\theta^* + \Delta) - \nabla B(\theta^*) - \nabla^2 B(\theta^*)\Delta.$$

Denoting $\widehat{\Delta} = \widehat{\theta} - \theta^*$, and noting that $\widehat{\mu} = \nabla B(\widehat{\theta})$ and $\bar{\mu} = \nabla B(\theta^*)$ we can rewrite (20) as

$$\nabla^2 B(\theta^*)\widehat{\Delta} + R(\widehat{\Delta}) - W + \lambda Z = 0. \tag{21}$$

We can now state our assumptions. We will assume that the difference between the true marginals $\mu^* = \nabla A(\theta^*)$ and the "true" variational pseudomarginals $\bar{\mu} = \nabla B(\theta^*)$ is bounded so that

Assumption 1. $\kappa_B := \|\nabla B(\theta^*) - \nabla A(\theta^*)\|_\infty < 1$.

Thus, $\|W_2\|_\infty = \kappa_B$ is controlled.

We will also assume a mild regularity condition on the remainder:

Assumption 2. *For all sparse and bounded Δ where $\|\Delta\|_\infty \leq \lambda$, and $\Delta_{S^c} = 0$, the second-order remainder is bounded by,*

$$\|R(\Delta; \theta^*)\|_\infty \leq \kappa_R \|\Delta\|_\infty^2, \tag{22}$$

for some $\kappa_R > 0$.

We note that such conditions have been considered in other analyses of ℓ_1-regularized non-linear objectives. For instance, in the case where the objective is the log-likelihood of a Gaussian distribution with true covariance matrix Σ^*, Ravikumar et al. (2008) showed that the second-order remainder is bounded as in the assumption for $\kappa_R = 3d/2 \,|\!|\!|\Sigma^*|\!|\!|_\infty^3$. Note that since both θ^* and Δ have support restricted to S, and κ_R only depends on the support size and is independent of the ambient dimension.

Lastly, let $Q = \nabla^2 B(\theta^*)$ denote the Hessian of $B(\theta)$ at the true parameters θ^*. Let S denote the support of the the true parameters θ^* and let S^c denote its complement. Thus, $S = V \cup \{(s,t) : \theta^*_{st} \neq 0\}$, and $S^c = \{(s,t) : \theta^*_{st} = 0\}$. Let Q_{AB} denote the submatrix of Q with rows indexed by A and columns indexed by B. We then assume the following incoherence assumption:

Assumption 3. $|\!|\!|Q_{SS}^{-1} Q_{S^cS}|\!|\!|_\infty \leq 1 - \alpha$, *for some constant $\alpha > 0$.*

Theorem 1. *Consider a graphical model distribution with parameters θ^* satisfying assumptions 1,2,3. Suppose we solve (9) by setting the regularization parameter λ as $\lambda \geq \frac{4}{\alpha}(\sqrt{\frac{\log p}{n}} + \kappa_B)$, and where the sample size n scales as $n \geq \left(\alpha^2/(32\kappa_Q^2\kappa_R) - \kappa_B\right)^2 \log p$. Then with probability greater than $1 - \exp(-c\log p)$ for some constant $c > 0$, we have:*

(a) the estimate $\widehat{\theta}$ from (9) satisfies the elementwise ℓ_∞ bound: $\|\widehat{\theta} - \theta^*\| \leq \min\{1/(2\kappa_Q\kappa_R), 4\kappa_Q\lambda\}$,

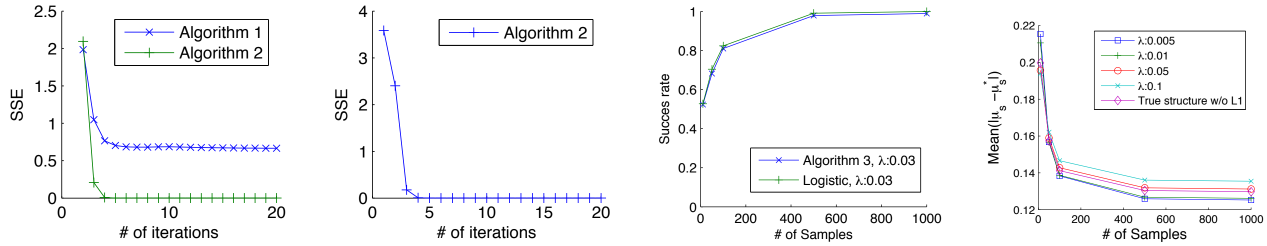
(b) it specifies an edgeset $E(\widehat{\theta})$ that has no false inclusions (i.e. $E(\widehat{\theta}) \subseteq E(\theta^*)$), and moreover includes all edges (s,t) such that, $|\theta^*_{st}| > \min\{1/(2\kappa_Q\kappa_R), 4\kappa_Q\lambda\}$.

The detailed proof is provided in the Supplementary Material.

6. Experiments

We now briefly illustrate our results on 25 node Ising models (2). Further details on our experimental settings, as well as more exhaustive simulations are provided in the appendix. Figure 1 (a) compares our message-passing updates (Algorithm 2) to gradient descent for tree-reweighted entropic approximation, on Ising models with four-nearest neighbor lattice graph-structure. Here, we plot the ℓ_2 deviation of the pseudomarginals to the optimum against iterations: it can be seen that our message-passing updates (Algorithm 2) converge very fast. Figure 1(b) plots the convergence of our message-passing updates for a more general weighted entropic approximation. Figure 1 shows the edge-recovery rate of the TRW-approx estimator that we have available in closed form. It has comparable performance to the state of the art method (Ravikumar et al., 2010) that uses nodewise ℓ_1 regularized logistic regressions, which is impressive considering that our estimator in this case has a simple closed-form solution. In Figure 1(d) we follow (Wainwright, 2006). We compare two parameter estimators: our TRW-approx estimator, and an oracle estimator that knows the true graph structure and estimates the parameters using the tree-reweighted entropy approximation to the log-partition function (called pseudo-moment matching in Wainwright et al. (2003)). We then plot the ℓ_2 error in moment estimates after perturbing this parameter estimate. This is a surrogate metric for gauging the use of the estimated model for prediction; we see that our estimator is very close to the oracle estimator.

Summary. We investigate a whole class of estimators that recover graphical model structure by minimizing ℓ_1-regularized surrogate log-likelihoods based on variational approximations to the partition function. As we note in the introduction, many state of the art methods fall into



(a) 4-NN, Mixed, $\omega = 0.5$ (b) 4-NN, Positive, $\omega = 0.25$ (c) 4-NN, Mixed, $\omega = 0.25$ (d) 4-NN, Mixed, $\omega = 0.5$

Figure 1. The four panels illustrate the following four experiments:
(a) Convergence Rate: Gradient Descent (Alg. 1) vs Our Mesg. Passing updates (Alg. 2) for tree-reweighted entropic approx.
(b) Our Mesg. Passing updates (Alg. 2) for weighted Entropic approximation with $\alpha_s = \alpha_{st} = 1$.
(c) Structure Recovery: TRW-approx estimator (in closed form, Alg. 3) vs Nodewise logistic regressions of (Ravikumar et al., 2010).
(d) Prediction (Error in moments after perturbing parameter estimate): TRW-approx estimator vs Oracle parameter estimator that knows the true graph structure.

this category. For this general setting, we provide (a) a general message passing algorithm for *directly* solving the resulting ℓ_1 regularized optimization problems (in contrast to iterative calls to a separate approximate inference procedure), and (b) sparsistency results for this entire class of estimators. Our study also revealed that in special cases, the resulting estimator is available in closed form.

References

Boyd, S. and Vandenberghe, L. *Convex optimization*. Cambridge University Press, Cambridge, UK, 2004.

Candes, E. and Tao, T. The Dantzig selector: Statistical estimation when p is much larger than n. *Annals of Statistics*, 35(6):2313–2351, 2007.

Censor, Y. and Zenios, S. A. *Parallel Optimization: Theory, Algorithms, and Applications*. Numerical Mathematics and Scientific Computation. Oxford University Press, 1988.

Heskes, T. Convexity arguments for efficient minimization of the bethe and kikuchi free energies. *J. Artif. Intell. Res. (JAIR)*, 26:153–190, 2006.

Ising, E. Beitrag zur theorie der ferromagnetismus. *Zeitschrift für Physik*, 31:253–258, 1925.

Lee, S.-I., Ganapathi, V., and Koller, D. Efficient structure learning of markov networks using l1-regularization. In *NIPS 19*, 2007.

Montanari, A. and Pereira, J. A. Which graphical models are difficult to learn. In *Neural Information Processing Systems (NIPS)*, 2009.

Nesterov, Y. Introductory lectures on convex optimization: a basic course. 2004.

Ortega, J. M. and Rheinboldt, W. C. *Iterative solution of nonlinear equations in several variables*. Classics in applied mathematics. SIAM, New York, 2000.

Pearl, J. *Probabilistic reasoning in intelligent systems*. Morgan Kaufman, San Mateo, 1988.

Ravikumar, P., Wainwright, M. J., Raskutti, G., and Yu, B. High-dimensional covariance estimation by minimizing ℓ_1-penalized log-determinant divergence. Technical Report 767, Dept. of Stat., UC Berkeley, 2008.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. High-dimensional ising model selection using ℓ_1-regularized logistic regression. *Ann. Stat.*, 38(3):1287–1319, 2010.

Tropp, J. A. Just relax: Convex programming methods for identifying sparse signals. *IEEE Trans. Info. Theory*, 51 (3):1030–1051, March 2006.

Wainwright, M. J. Estimating the "wrong" graphical model: Benefits in the computation-limited regime. *JMLR*, 7:1829–1859, September 2006.

Wainwright, M. J. Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ_1-constrained quadratic programming (lasso). *IEEE Transactions on Information Theory*, 55:2183–2202, 2009.

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.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudomoment matching. In *AISTATS*, 2003.

Wainwright, M. J., Jaakkola, T. S., and Willsky, A. S. A new class of upper bounds on the log partition function. *IEEE Trans. Info. Theory*, 51(7):2313–2335, July 2005.

Weiss, Y., Yanover, C., and Meltzer, T. MAP estimation, linear programming, and belief propagation with convex free energies. In *UAI*, 2007.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. Generalized belief propagation. In *NIPS 13*, pp. 689–695, 2001.