Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudo-moment matching

Martin J. Wainwright
Dept. of EECS, UC Berkeley
Berkeley, CA 94705
martinw@eecs.berkeley.edu

Tommi S. Jaakkola
Dept. of EECS, MIT
Cambridge, MA 02139
tommi@ai.mit.edu

Alan S. Willsky
Dept. of EECS, MIT
Cambridge, MA 02139
willsky@mit.edu

Abstract

In previous work [10], we presented a class of upper bounds on the log partition function of an arbitrary undirected graphical model based on solving a convex variational problem. Here we develop a class of local message-passing algorithms, which we call *tree-reweighted belief propagation*, for efficiently computing the value of these upper bounds, as well as the associated pseudomarginals. We also consider the uses of our bounds for the problem of maximum likelihood (ML) parameter estimation. For a completely observed model, our analysis gives rise to a concave lower bound on the log likelihood of the data. Maximizing this lower bound yields an approximate ML estimate which, in analogy to the moment-matching of exact ML estimation, can be interpreted in terms of *pseudo-moment-matching*. We present preliminary results illustrating the behavior of this approximate ML estimator.

1 Introduction

Associated with any undirected graphical model is a log partition function. This quantity plays a fundamental role in various contexts, including approximate inference [4], maximum likelihood parameter estimation [5], and large deviations analysis [3]. For a general undirected model, exact computation of this partition function is intractable; therefore, developing approximations and bounds is an important problem.

In previous work [10], we presented a new class of upper bounds on the log partition function of an arbitrary undirected graphical model. These bounds are based on approximating the original distribution by a convex combination of tractable distributions (e.g., tree-structured distributions), and then exploiting the convexity of the log partition function to obtain upper bounds. Through a Lagrangian dual reformulation, we showed that the tightest form of such an upper bound can be obtained by solving a convex variational problem which, in the case of combining tree-structured distributions, can be viewed as a "convexified" form of the Bethe problem [11].

One contribution of this paper is the presentation of local message-passing algorithms, analogous to but distinct from belief propagation [11], for efficiently computing the optimal value of these upper bounds, as well as pseudomarginals that can be used as approximations to the marginals. This class of algorithms, which we refer to as *tree-reweighted belief propagation*, are the sum-product version of the tree-reweighted max-product updates analyzed in our related work [9]. We then consider the use of our bounds for approximate maximum likelihood (ML) parameter estimation. For a completely observed model, our methods give rise to a concave lower bound on the log likelihood of the data; consequently, an approximate ML estimate can be obtained by maximizing this lower bound. We show that, in analogy to the well-known moment-matching properties of exact ML estimates [1], the global maximum of this lower bound is obtained by performing a type of *pseudo-moment matching*. We illustrate the behavior of this approximate ML estimator in application to some simple problems, and compare it to a heuristic method based on ordinary belief propagation.

The remainder of this paper is organized in the following manner. We begin in Section 2 by introducing the formalism of graphical models, and more specifically the exponential representations that are central to our analysis. In Section 3, we first describe the variational problem that underlies our upper bounds [10], and then present tree-reweighted belief propagation algorithms for solving it. In Section 4, we propose and analyze a technique for approximate ML estimation based on our upper bounds. In Section 5, we present some preliminary results of experiments on this approximate ML estimator, and point out open questions to be addressed. We conclude in Section 6 with a discussion.

2 Notation and background

We begin with the notation and background necessary for subsequent development.

Graphical models: An undirected graph $G = (V, E)$ consists of a set of $N = |V|$ nodes, joined by edges $(s,t) \in E$. For each $s \in V$, we let $\Gamma(s)$ denote the set of its neighbors. A graph clique S is a fully-connected subset of the vertex set (i.e., for all $s, t \in S$, we have $(s,t) \in E$). To define an undirected graphical model or Markov random field, we place at each node $s \in V$ a random variable x_s taking values in the discrete space $\mathcal{X}_s = \{0, 1, \ldots, m_s - 1\}$. We let $\mathbf{x} = \{x_s \mid s \in V\}$ be a random vector taking values in the Cartesian product space $\mathcal{X}^N = \mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_N$.

Exponential representations: Our work makes use of exponential representations for undirected graphical models, which have been studied extensively in applied probability theory and statistics [e.g., 1, 8]. For some index set $\mathcal{I}$, we let $\boldsymbol{\phi} = \{\phi_\alpha \mid \alpha \in \mathcal{I}\}$ denote a collection of potential functions defined on the cliques of G, and let $\theta = \{\theta_\alpha \mid \alpha \in \mathcal{I}\}$ be a vector of weights on these potential functions. The exponential family determined by $\boldsymbol{\phi}$ is the following collection of log-linear models:

$$p(\mathbf{x}; \theta) = \exp\Big\{ \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(\mathbf{x}) - \Phi(\theta) \Big\} \tag{1a}$$

$$\Phi(\theta) = \log \sum_{\mathbf{x} \in \mathcal{X}^N} \exp\Big\{ \sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(\mathbf{x}) \Big\} \tag{1b}$$

Here $\Phi(\theta)$ is the *log partition function* that serves to normalize the distribution. This function plays an important role in various contexts, including approximate inference, parameter estimation, and large deviations analysis.

Much of our analysis uses an *overcomplete* exponential representation, in which there are linear dependencies among the potentials $\{\phi_\alpha\}$. In particular, we use indicator functions as potentials:

$$\phi_{s;j}(x_s) = \delta_{s;j}(x_s), \quad s \in V;\ j \in \mathcal{X}_s$$
$$\phi_{st;jk}(x_s, x_t) = \delta_{st;jk}(x_s, x_t),\ (s,t) \in E;\ j,k \in \mathcal{X}_s \times \mathcal{X}_t$$

Here $\delta_{s;j}(x_s)$ takes the value one when $x_s = j$, and zero otherwise. In this case, the index set $\mathcal{I}$ consists of the union of $\mathcal{I}(V) = \{\, (s;j) \mid s \in V;\ j \in \mathcal{X}\}$ with the edge indices $\mathcal{I}(E) = \{\, (st;jk) \mid (s,t) \in E;\ j,k \in \mathcal{X}\,\}$.

Convex duality: The following well-known properties of Φ are central to our analysis:

Lemma 1. *(a) For all indices $\alpha \in \mathcal{I}$, we have*

$$\frac{\partial \Phi(\theta)}{\partial \theta_\alpha} = \mathbb{E}_\theta[\phi_\alpha] = \sum_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}; \theta) \phi_\alpha(\mathbf{x})$$

(b) Moreover, the second derivative is given by an element of the Fisher information matrix — namely:

$$\frac{\partial^2 \Phi(\theta)}{\partial \theta_\alpha \, \partial \theta_\beta} = \mathbb{E}_\theta[\phi_\alpha \phi_\beta] - \mathbb{E}_\theta[\phi_\alpha]\mathbb{E}_\theta[\phi_\beta]$$

so that the log partition function Φ is convex as a function of θ.

A second set of parameters, related to the exponential parameters θ by Legendre duality [1, 6, 8], are obtained by taking expectations of the potential functions as follows:

$$P_\alpha = \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})] \tag{2}$$

These *mean parameters* are sufficient statistics, in that they completely specify the distribution $p(\mathbf{x}; \theta)$. In the case of an overcomplete representation with indicator functions, the dual variables correspond to the values of particular marginal distributions of the distribution $p(\mathbf{x}; \theta)$. For example, when $\alpha = (s;j)$, then we have $P_{s;j} = \mathbb{E}_\theta[\delta_{s;j}(x_s)] = p(x_s = j; \theta)$.

3 Tree-reweighted belief propagation and optimal upper bounds

In this section, we present the variational problem that yields upper bounds on the log partition function [10]. We then develop a family of tree-reweighted belief propagation algorithms designed to solve this optimization problem.

3.1 Basic set-up

Here we provide the notation and background to state our upper bounds.

Edge appearance probabilities: Let $\mathfrak{T} = \mathfrak{T}(G)$ denote the set of all spanning trees of G. We consider a probability distribution $\vec{\boldsymbol{\mu}}$ over the set of spanning trees — that is, a collection of non-negative numbers

$$\vec{\boldsymbol{\mu}} \triangleq \{\, \mu(\mathcal{T}),\ \mathcal{T} \in \mathfrak{T} \mid \mu(\mathcal{T}) \geq 0 \,\} \tag{3}$$

such that $\sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T}) = 1$. Of particular interest in the sequel is the probability $\mu_e = \Pr_{\vec{\boldsymbol{\mu}}}\{e \in \mathcal{T}\}$ that a given edge $e \in E$ appears in a spanning tree $\mathcal{T}$ chosen randomly under $\vec{\boldsymbol{\mu}}$. We let $\boldsymbol{\mu_e} = \{\mu_e \mid e \in E\}$ represent a vector of these *edge appearance probabilities*. It can be shown [8] that these edge appearance vectors must belong to the so-called *spanning tree polytope*, denoted by $\mathbb{T}(G)$. See Figure 1 for an illustration.

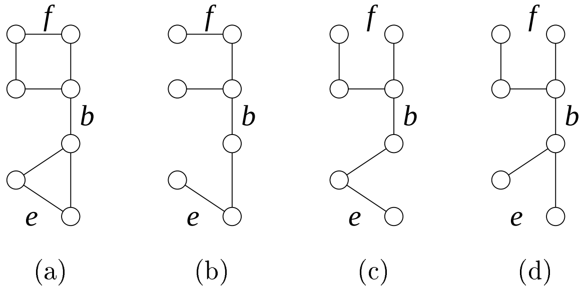

Figure 1. Illustration of the spanning tree polytope $\mathbb{T}(G)$. Original graph is shown in panel (a). Probability 1/3 is assigned to each of the three spanning trees $\{ \mathcal{T}_i \mid i = 1,2,3 \}$ shown in panels (b)–(d). Edge b is a so-called bridge in G, meaning that it must appear in any spanning tree (i.e., $\mu_b = 1$). Edges e and f appear in two and one of the spanning trees respectively, which gives rise to edge appearance probabilities $\mu_e = 2/3$ and $\mu_f = 1/3$.

Tree-consistent pseudomarginals: The constraint set associated with our variational formulation [10] is the set of so-called pseudomarginals that satisfy certain tree-consistency constraints. To be precise, for each node $s \in V$, let $T_s = \{T_{s;j} \mid j \in \mathcal{X}_s\}$ be a non-negative pseudomarginal vector with $m_s = |\mathcal{X}_s|$ elements; similarly, for each edge $(s,t) \in E$, let $T_{st} = \{T_{st;jk} \mid (j,k) \in \mathcal{X}_s \times \mathcal{X}_t\}$ be a non-negative pseudomarginal vector with $m_s \times m_t$ elements. On occasion, we will also use the notation $T_s(x_s)$ to refer to the function that takes the value $T_{s;j}$ when $x_s = j$; the joint function $T_{st}(x_s, x_t)$ is defined similarly. We let $\mathbf{T} = \{T_s,\ s \in V\} \cup \{ T_{st},\ (s,t) \in E \}$ denote the full collection of pseudomarginals on nodes and edges. This set of pseudomarginals is required to satisfy a set of local normalization and marginalization constraints; in particular, we require that they are elements of the set

$$\mathrm{TREE}(G) \triangleq \Big\{ \ \mathbf{T} \ \Big| \ \sum_{k \in \mathcal{X}_t} T_{st;jk} = T_{s;j}, \quad \sum_{j \in \mathcal{X}_s} T_{s;j} = 1 \ \Big\}$$

Our choice of notation is motivated by the fact that if G is a tree, then $\mathrm{TREE}(G)$ is a complete description of the set of valid (single node and edge) marginal distributions.

Variational formulation We now present the variational problem that gives rise to upper bounds on the log partition function. We begin by setting up the necessary notation. For each $s \in V$ and pseudomarginal T_s, we define the single node entropy:

$$H_s(T_s) \quad = \quad - \sum_{j \in \mathcal{X}_s} T_{s;j} \log T_{s;j}$$

Similarly, for each $(s,t) \in E$, we define the mutual information between x_s and x_t as measured under the joint pseudomarginal T_{st}:

$$I_{st}(T_{st}) \quad = \quad \sum_{(j,k)} T_{st;jk} \log \frac{T_{st;jk}}{(\sum_{k \in \mathcal{X}_t} T_{st;jk})(\sum_{j \in \mathcal{X}_s} T_{st;jk})}$$

Borrowing terminology from statistical physics [11], we define an "average energy" term as follows:

$$\mathbf{T} \cdot \theta^* \quad = \quad \sum_{s \in V} \sum_{j} T_{s;j} \theta^*_{s;j} + \sum_{(s,t) \in E} \sum_{(j,k)} T_{st;jk} \theta^*_{st;jk}$$

Using this notation, our bounds are based on the following function:

$$\mathcal{F}(\mathbf{T}; \boldsymbol{\mu_e}; \theta^*) \triangleq - \sum_{s \in V} H_s(T_s) + \sum_{(s,t) \in E} \mu_{st} I_{st}(T_{st}) - \mathbf{T} \cdot \theta^*$$

It can be seen that this function is closely related to the Bethe free energy [11]. In fact, suppose that we set $\mu_{st} = 1$ for all edges $(s,t) \in E$, meaning that every edge appears with probability one. In this case, the function $\mathcal{F}(\mathbf{T}; \boldsymbol{\mu_e}; \theta^*)$ is equivalent to the Bethe free energy on the constraint set $\mathrm{TREE}(G)$. However, the choice $\boldsymbol{\mu_e} = \mathbf{1}$ belongs to the spanning tree polytope $\mathbb{T}(G)$ only when the graph G is actually a tree.

In the paper [10], we prove the following result:

Theorem 1. *For all $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, the function $\mathcal{F}(T; \boldsymbol{\mu_e}; \theta^*)$ is a convex in terms of $\mathbf{T}$. Moreover, the log partition function is bounded above by the solution of the following variational problem:*

$$\Phi(\theta^*) \quad \leq \quad - \min_{\mathbf{T} \in \mathrm{TREE}(G)} \mathcal{F}(\mathbf{T}; \boldsymbol{\mu_e}; \theta^*) \qquad (4)$$

The optimal solution $\widehat{\mathbf{T}} = \widehat{\mathbf{T}}(\theta^*)$ to this minimization is unique.

3.2 Tree-reweighted belief propagation

We now present a tree-reweighted belief propagation algorithm designed to find the requisite set $\widehat{\mathbf{T}}$ of pseudomarginals via a sequence of message-passing operations. This algorithm is the sum-product version of the tree-reweighted max-product updates analyzed in our related work [9].

The optimal collection $\widehat{\mathbf{T}}$ of pseudomarginals, as a solution to the constrained optimization problem (4), must belong to $\mathrm{TREE}(G)$. In addition, it can be shown [10] that they are characterized by the following *admissibility* condition:

$$\theta^* \cdot \boldsymbol{\phi}(\mathbf{x}) + C = \sum_{s \in V} \log \widehat{T}(x_s) + \sum_{(s,t) \in E} \mu_{st} \log \frac{\widehat{T}_{st}(x_s, x_t)}{\widehat{T}_s(x_s) \widehat{T}_t(x_t)} \qquad (5)$$

Algorithm 1 (Tree-reweighted belief propagation).

1. Initialize the messages $\mathbf{M}^0 = \{M^0_{st}\}$ with arbitrary positive real numbers.

2. For iterations $n = 0, 1, 2, \ldots$, update the messages as follows:

$$M^{n+1}_{ts}(x_s) \quad = \quad \kappa \sum_{x'_t \in \mathcal{X}_t} \exp\Big(\frac{1}{\mu_{st}}\phi_{st}(x_s, x'_t; \theta^*_{st}) + \phi_t(x'_t; \theta^*_t)\Big) \Big\{ \frac{\prod_{v \in \Gamma(t) \backslash s} \big[M^n_{vt}(x'_t)\big]^{\mu_{vt}}}{\big[M^n_{st}(x'_t)\big]^{(1-\mu_{ts})}} \Big\} \tag{6}$$

Note: The quantity $\phi_s(x_s; \theta^*_s)$ takes the value $\theta^*_{s;j}$ when $x_s = j$. The function $\phi_{st}(x_s, x_t; \theta^*_{st})$ is defined analogously.

Here C is a constant independent of $\mathbf{x}$. If we set $\boldsymbol{\mu_e} = \mathbf{1}$, then equation (5) asserts that the collection $\widehat{\mathbf{T}}$ constitutes a *reparameterization* of the original distribution. (See [8] for more details on reparameterization and invariance properties of BP and related algorithms.) However, as noted before, the choice $\boldsymbol{\mu_e} = \mathbf{1}$ is a valid choice only when the graph itself is tree-structured.

We now specify a collection of pseudomarginals $\mathbf{T}$ via a set of messages, such that the admissibility condition (5) is always satisfied. For each edge $(s,t) \in E$, let $M_{ts}(x_s)$ be the message passed from node t to node s. It is a vector of length m_s, with one element for each state $j \in \mathcal{X}_s$. We use the messages $\mathbf{M} = \{M_{st}\}$ to specify a set of functions $\mathbf{T} = \{T_s, T_{st}\}$ as follows:

$$T_s(x_s) = \kappa\ \exp\big(\phi_s(x_s; \theta^*_s)\big) \prod_{v \in \Gamma(s)} \big[M_{vs}(x_s)\big]^{\mu_{vs}} \tag{7}$$

$$T_{st}(x_s, x_t) = \kappa\ \varphi_{st}(x_s, x_t; \theta^*)\ \frac{\prod_{v \in \Gamma(s) \backslash t} \big[M_{vs}(x_s)\big]^{\mu_{vs}}}{\big[M_{ts}(x_s)\big]^{(1-\mu_{st})}} \times \frac{\prod_{v \in \Gamma(t) \backslash s} \big[M_{vt}(x_t)\big]^{\mu_{vt}}}{\big[M_{st}(x_t)\big]^{(1-\mu_{ts})}} \tag{8}$$

Here κ denotes a constant chosen so as to ensure that the normalization conditions (e.g., $\sum_{x'_s} T_s(x'_s) = 1$) are satisfied. Moreover, $\varphi_{st}(x_s, x_t; \theta^*)$ is a compact notation for the quantity

$$\exp\big(\frac{1}{\mu_{st}}\phi_{st}(x_s, x_t; \theta^*_{st}) + \phi_s(x_s; \theta^*_s) + \phi_t(x_t; \theta^*_t)\big)$$

This construction ensures that the admissibility condition is satisfied:

Lemma 2 (Admissibility). *Given any collection $\mathbf{T} = \{T_s, T_{st}\}$ defined by a set of messages $\mathbf{M}$ as in equations* (7) *and* (8)*, the admissibility condition* (5) *is satisfied.*

Proof. First of all, by the definitions of equations (7) and (8), we have for each $(s,t) \in E$:

$$\mu_{st} \log \frac{T_{st}(x_s, x_t)}{T_s(x_s) T_t(x_t)} \;=\; \phi_{st}(x_s, x_t; \theta^*_{st}) - \mu_{st} \log[M_{st}(x_t) M_{ts}(x_s)] + C$$

Secondly, equation (7) gives for each node $s \in V$:

$$\log T_s(x_s) = \phi_s(x_s; \theta^*_s) + \sum_{u \in \Gamma(s)} \mu_{us} \log M_{us}(x_s) + C$$

Summing together a copy of the first equation for each edge $(s,t) \in E$ and a copy of the second equation for each node $s \in V$ yields the statement of equation (5). □

We now need to ensure that $\widehat{\mathbf{T}}$ satisfies the local consistency constraints defining membership in TREE(G) – namely, $\sum_{x'_t} T_{st}(x_s, x'_t) = T_s(x_s)$. In order to do so, we update the messages according to Algorithm 1. It can shown with some elementary calculations that any fixed point of the message update equation (6), as with fixed points of ordinary belief propagation, satisfies the following property:

Lemma 3 (Tree consistency). *Suppose that we use a fixed point of the message update equation* (6) *to specify a collection of pseudomarginals $\widehat{\mathbf{T}}$ as in equation* (7) *and* (8)*. Then the marginalization condition $\sum_{x'_t} \widehat{T}_{st}(x_s, x'_t) = \widehat{T}_s(x_s)$ is satisfied.*

In fact, since the optimization problem (4) has a unique global minimum, the message update equation (6) always has a unique fixed point. By Lemma 3, the pseudomarginal $\widehat{\mathbf{T}}$ specified by this fixed point belongs to TREE(G); by Lemma 2, it satisfies the admissibility condition (5). Therefore, it corresponds to the global minimum of the minimization problem (4).

Observe that Algorithm 1, although quite similar to the standard belief propagation updates [11], differs in some key ways. First of all, the weights $\theta_{st;jk}$ corresponding to edge (s,t) are all rescaled by $1/\mu_{st}$. Secondly, all the messages M_{us} for nodes $u \in \Gamma(s)\backslash t$ are exponentiated by the corresponding edge appearance μ_{us}. Lastly, the message M_{st} running in the *reverse direction* on edge (s,t) is involved in updating M_{ts}. Despite these differences, it is still possible to perform the message updates in a parallel fashion, as in ordinary belief propagation. It is also possible to perform updates over spanning trees, as in the tree reparameterization approach to BP [8]. In practice, we find that the updates of Algorithm 1 converge when suitably relaxed,[1] although the convergence rate can be slower than that of ordinary BP. However, we do not have a proof of convergence.

4 Approximate ML estimation

An important feature of ML parameter estimation, as well as other problems (e.g., large deviations analysis), is that the solution is specified by *moment-matching*. To illustrate this notion, suppose that we are given an IID sequence $\mathbf{Y} = \{\mathbf{y}^1, \ldots, \mathbf{y}^M\}$ of data from some unknown model $p(\mathbf{x};\theta)$. The log likelihood $L(\theta)$ of the data $\mathbf{Y}$ is given by:

$$L(\theta) \;=\; \frac{1}{M}\sum_{k=1}^{M} \log p(\mathbf{y}^k;\theta) \;=\; \sum_{\alpha\in\mathcal{I}} \theta_\alpha \bar{P}_\alpha - \Phi(\theta) \tag{9}$$

where $\bar{P}_\alpha \triangleq \frac{1}{M}\sum_{k=1}^{M}\phi_\alpha(\mathbf{y}^k)$ is the empirical marginal of ϕ_α under the data. Taking derivatives of $L(\theta)$ using Lemma 1, we find that the maximum likelihood solution θ^{ML} satisfies $\mathbb{E}_{\theta^{ML}}[\phi_\alpha] = \bar{P}_\alpha$. That is, the optimal distribution $p(\mathbf{x};\theta^{ML})$ has its moments matched to the empirical averages $\bar{P}$. In this section, we show how our upper bounds on the partition function can be used to develop a method for approximate ML estimation that, in analogy to this exact moment-matching, performs a type of pseudo-moment matching.

We begin by defining a function via the minimization specified by Theorem 1 as follows:

$$\mathcal{H}(\theta) \;=\; -\min_{\mathbf{T}\in\mathrm{TREE}(G)} \mathcal{F}(\mathbf{T};\boldsymbol{\mu_e};\theta) \tag{10}$$

Here we view $\boldsymbol{\mu_e} \in \mathbb{T}(G)$ as a fixed parameter, so that $\mathcal{H}$ is a function only of θ. By Theorem 1, the function $\mathcal{H}$ gives an upper bound on the log partition function. In addition, it turns out to have the following desirable properties:

Theorem 2. *The function $\mathcal{H}$ is convex (strictly so in a minimal representation), and differentiable in terms of θ. Moreover, for each $\alpha \in \mathcal{I}$, the partial derivative is given by the pseudomarginal vector:*

$$\frac{\partial\mathcal{H}}{\partial\theta_\alpha}(\theta) \;=\; \widehat{T}_\alpha(\theta) \tag{11}$$

where $\widehat{\mathbf{T}}(\theta) = \arg\min_{\mathbf{T}\in\mathrm{TREE}(G)} \mathcal{F}(\mathbf{T};\boldsymbol{\mu_e};\theta^*)$.

Proof. Since $\mathcal{H}$ is the negative of the minimum of a collection of functions that are linear in θ, it is convex [2]. For a fixed $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, the function $\mathcal{F}(\mathbf{T};\boldsymbol{\mu_e};\theta)$ is differentiable in terms of $(\mathbf{T},\theta)$; moreover, it is linear (and hence concave) in θ for each fixed $\mathbf{T}$, and the constraint set $\mathrm{TREE}(G)$ is convex and compact. For each fixed θ, the minimum of $\mathcal{F}(\mathbf{T};\boldsymbol{\mu_e};\theta)$ over $\mathbf{T}$ is attained at a unique point in $\mathrm{TREE}(G)$. Therefore, the problem satisfies the hypotheses of Danskin's theorem [2], implying that $\mathcal{H}$ is differentiable, with derivatives given by the pseudomarginals $\frac{\partial\mathcal{H}}{\partial\theta_\alpha}(\theta) = -\frac{\partial\mathcal{F}}{\partial\theta_\alpha}(\widehat{\mathbf{T}};\boldsymbol{\mu_e};\theta) = \widehat{T}_\alpha(\theta)$ as claimed.

To establish strict convexity in a minimal representation, we use γ to denote the minimal analog of θ. For each tree $\mathcal{T}$, let $\gamma(\mathcal{T})$ denote an arbitrary tree-structured exponential parameter (i.e., $\gamma_\alpha(\mathcal{T}) = 0$ for all $\alpha \notin \mathcal{I}(\mathcal{T})$, where $\mathcal{I}(\mathcal{T})$ are the indices corresponding to tree $\mathcal{T}$). In our previous work [10, 8], we showed that the optimal solution in Theorem 1 is defined by a particular set of tree-structured exponential parameters $\{\widehat{\gamma}(\mathcal{T})\}$ such that (i) the tree-structured distributions $p(\mathbf{x};\widehat{\gamma}(\mathcal{T}))$ all share a common set of marginals $\widehat{\mathbf{T}}(\gamma)$ (i.e., given by the optimum in Theorem 1); (ii) we have the equivalence $\sum_{\mathcal{T}} \mu(\mathcal{T})\widehat{\gamma}(\mathcal{T}) = \gamma$; and lastly (iii) we have the relation $\mathcal{H}(\gamma) = \sum_{\mathcal{T}}\mu(\mathcal{T})\Phi(\widehat{\gamma}(\mathcal{T}))$. Given two bounded minimal parameter vectors that are distinct (i.e., $\gamma^1 \neq \gamma^2$), consider the associated tree parameters $\{\widehat{\gamma}^1(\mathcal{T})\}$ and $\{\widehat{\gamma}^2(\mathcal{T})\}$. For each tree, we use the convexity of Φ, property (i) and Lemma 1(a) to write:

$$\Phi(\widehat{\gamma}^1(\mathcal{T})) \geq \Phi(\widehat{\gamma}^2(\mathcal{T})) + \sum_{\alpha\in\mathcal{I}(\mathcal{T})} \widehat{T}_\alpha(\gamma^2)\,\left[\widehat{\gamma}^1(\mathcal{T}) - \widehat{\gamma}^2(\mathcal{T})\right]_\alpha$$

From property (ii) and the fact $\gamma^1 \neq \gamma^2$, we must have $\widehat{\gamma}^1(\mathcal{T}) \neq \widehat{\gamma}^2(\mathcal{T})$ for at least one tree $\mathcal{T}$, whence by the strict convexity of Φ in a minimal representation this inequality is strict. Thus, the inequality remains strict in the weighted sum:

$$\sum_{\mathcal{T}} \mu(\mathcal{T})\Phi(\widehat{\gamma}^1(\mathcal{T})) > \sum_{\mathcal{T}} \mu(\mathcal{T})\Phi(\widehat{\gamma}^2(\mathcal{T})) + \sum_{\alpha\in\mathcal{I}} \widehat{T}_\alpha(\gamma^2) \sum_{\mathcal{T}} \mu(\mathcal{T})\big[\widehat{\gamma}^1(\mathcal{T}) - \widehat{\gamma}^2(\mathcal{T})\big]_\alpha \tag{12}$$

where we have used the fact that $\widehat{T}_\alpha(\gamma^2)$ is the same for all trees such that $\alpha \in \mathcal{I}(\mathcal{T})$ to move the sum over

[1] By relaxation, we mean performing updates of the form $(1-\alpha)\log M^n_{ts} + \alpha \log M^{n+1}_{ts}$, where $\alpha \in (0,1]$ is a step size.

$\mathcal{T}$ inside in the second term on the RHS. Finally, using properties (ii) and (iii), we can re-write equation (12) as $\mathcal{H}(\gamma^1) > \mathcal{H}(\gamma^2) + \sum_{\alpha \in \mathcal{I}} \widehat{T}_\alpha(\gamma^2)[\gamma^1 - \gamma^2]_\alpha$. Since $\frac{\partial \mathcal{H}}{\partial \gamma_\alpha}(\gamma) = \widehat{T}_\alpha(\gamma)$, this establishes the strict convexity. $\square$

Another way to understand the nature of $\mathcal{H}$ is on the basis of conjugate duality [6, 8]. First, let us define the function:

$$\mathcal{G}(\mathbf{T}; \boldsymbol{\mu_e}) = -\sum_{s \in V} H_s(T_s) + \sum_{(s,t) \in E} \mu_{st} I_{st}(T_{st}) \quad (13)$$

By Theorem 1, this function is convex in terms of $\mathbf{T}$. We use it to define $\mathcal{H}$ in an alternative but equivalent manner as follows:

$$\mathcal{H}(\theta) = \max_{\mathbf{T} \in \mathrm{TREE}(G)} \{\mathbf{T} \cdot \theta - \mathcal{G}(\mathbf{T}; \boldsymbol{\mu_e})\} \quad (14)$$

Equation (14) shows that for each fixed $\boldsymbol{\mu_e}$, $\mathcal{H}$ is the *conjugate dual* of $\mathcal{G}$.

We now use $\mathcal{H}$ to form the following approximation to the log likelihood:

$$\widetilde{L}(\theta) \triangleq \sum_{\alpha \in \mathcal{I}} \theta_\alpha \bar{P}_\alpha - \mathcal{H}(\theta) \quad (15)$$

Note that Theorems 1 and 2, in conjunction, guarantee that $\widetilde{L}(\theta)$ is a *concave lower bound* on the exact log likelihood $L(\theta)$ of equation (9). It is therefore reasonable to consider maximizing $\widetilde{L}(\theta)$ as a proxy to the exact log likelihood. Using Theorem 2, we see that the approximate ML solution $\widehat{\theta}$ thus obtained is specified by the pseudo-moment matching conditions $\widehat{T}_\alpha(\widehat{\theta}) = \bar{P}_\alpha$ for all indices $\alpha \in \mathcal{I}$.

For a given set of empirical marginals $\{\bar{P}_\alpha\}$, this pseudo-moment matching characterization enables us to specify the maximizing argument $\widehat{\theta}$ of $\widetilde{L}(\theta)$. In fact, using the admissibility of equation (5), we see that $\widehat{\theta}$ can be specified, for each $s \in V$ and edge (s,t), as follows:

$$\widehat{\theta}_{s;j} = \log \bar{P}_s(x_s = j) \quad (16a)$$

$$\widehat{\theta}_{st;jk} = \mu_{st} \log \frac{\bar{P}_{st}(x_s = j, x_t = k)}{\bar{P}_s(x_s = j)\, \bar{P}_t(x_t = k)} \quad (16b)$$

Running tree-reweighted BP on the problem $p(\mathbf{x}; \widehat{\theta})$ will, by construction, yield a set of pseudomarginals $\widehat{T}_\alpha$ equal to the empirical marginals $\bar{P}_\alpha$. Therefore, the appropriate pseudomarginal matching is ensured.

Of course, one can also imagine performing a similar kind of pseudomarginal matching using the ordinary BP approximation. That is, we could form another approximate ML estimate θ^{BP} via equation (16) with the choice $\mu_{st} = 1$ for all edges (s,t). By the tree reparameterization characterization of BP fixed points [8], this estimation method has the property that if BP is run on the problem $p(\mathbf{x}; \theta^{BP})$, then one of its (possibly many) fixed points are the correct empirical marginals $\bar{P}_\alpha$. Unlike the tree-reweighted analog, however, it is not guaranteed to maximize a lower bound on the log likelihood. Moreover, the experiments to follow suggest that it is less stable than the tree-reweighted approximate ML estimator.

5 Experiments

Suppose that we are given a set of data $\mathbf{Y}$ with associated empirical marginals $\{\bar{P}_\alpha\}$. Using a fixed set of edge appearance probabilities $\boldsymbol{\mu_e} \in \mathbb{T}(G)$, we can then estimate $\widehat{\theta}$ from equation (16); this yields an estimated distribution $p(\mathbf{x}; \widehat{\theta})$, for which applying the tree-reweighted BP algorithm will compute the "correct" marginals (i.e., the empirical marginals). In this

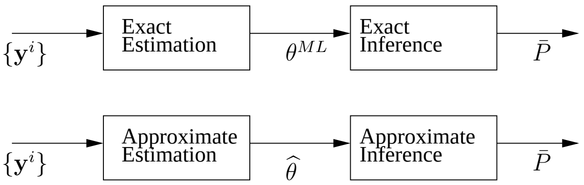

Figure 2. Block diagram of inference and parameter estimation. The combination of approximate estimation and inference is matched to the exact combination at a single point.

sense, as illustrated in Figure 2, the combination of tree-reweighted approximate ML estimation and inference is functionally equivalent to performing exact inference with the exact ML distribution. Of course, this exact relation holds only at a single point; of interest, then, is its robustness to perturbations in the problem.

To be more specific, suppose, for instance, that we receive a new set of noisy observations $\mathbf{z}$. Taking the assumption that the components of $\mathbf{z}$ are conditionally independent given $\mathbf{x}$, the measurement model has the form $p(\mathbf{z} \,|\, \mathbf{x}) = \prod_{s \in V} p(z_s \,|\, x_s)$. On one hand, we can combine this measurement model with the ML distribution $p(\mathbf{x}; \theta)$ so as to obtain the "true" posterior distribution $p(\mathbf{x} \,|\, \mathbf{z}; \theta) \propto p(\mathbf{x}; \theta) p(\mathbf{z} \,|\, \mathbf{x})$, for which we can then (at least in principle) compute the exact marginal distributions. An alternative and computationally tractable approach is to combine the measurements with the approximate ML model $p(\mathbf{x}; \widehat{\theta})$ so as to form an approximate posterior $p(\mathbf{x} \,|\, \mathbf{z}; \widehat{\theta})$, and then compute approximations to its marginals using the tree-reweighted BP algorithm. The interesting question is

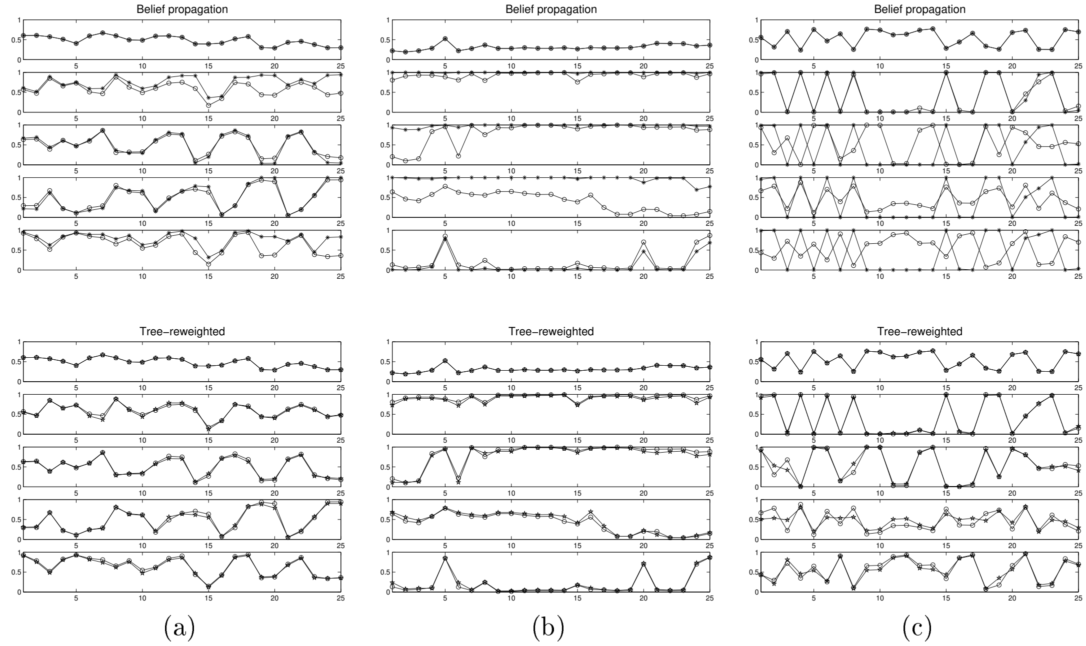

Figure 3. Behavior of BP-based method for parameter estimation (top plot in each column) versus TRW-based method (bottom plot in each column). Each of the five panels in each plot shows the marginal probability of $x_s = 1$ (exact marginal $P_{s;1}$ in open circles, and approximate marginal $\widehat{T}_{s;1}$ with stars) versus node number s. Top panel in each plot shows the marginals estimated from original data without any new data. By construction, both BP and TRW marginals are matched to the exact ones. Remaining four panels in each subplot shows four trials with new data added. (a) For weak attractive potentials ($\sigma_{edge} = 0.2$), both methods perform well. (b) For stronger attractive potentials ($\sigma_{edge} = 0.6$), BP-based method becomes overly sensitive to effect of new data, while TRW-based method remains stable. (c) For strong mixed potentials ($\sigma_{edge} = 0.6$), BP-based method is erratic, whereas TRW-method is still reasonable.

assessing how close these approximate marginals are to the exact marginals of the ML posterior. Accordingly, this section describes the results of experiments designed to test how the combination of approximate tree-reweighted ML estimation and inference perform as additional data is collected. We also compared the performance of the tree-reweighted combination to the heuristic BP analog described above.

We performed experiments for a binary Ising model of spins $\{-1, +1\}$:

$$p(\mathbf{x};\theta) \quad = \quad \exp\Big\{\sum_{s\in V}\theta_s x_s + \sum_{(s,t)\in E}\theta_{st}x_s x_t - \Phi(\theta)\Big\}$$

The parameter vector θ was chosen randomly from various ensembles in the following manner. In all cases, we chose the single node parameters randomly as $\theta_s \sim \mathcal{N}(0, \sigma^2_{node})$. In the *mixed* condition, we chose edge weights $\theta_{st} \sim \mathcal{N}(0, \sigma^2_{edge})$ independently for each edge. In the *attractive* condition, we set the edge weights $\theta_{st} = |a_{st}|$ independently for each edge, where $a_{st} \sim \mathcal{N}(0, \sigma^2_{edge})$. For all the experiments reported here, we preserved the ratio $\sigma_{node} = 0.5\,\sigma_{edge}$, and varied the choice of σ_{edge}.

Given a randomly chosen $p(\mathbf{x};\theta)$, we computed the exact marginals $\bar{P}_\alpha$ of $p(\mathbf{x};\theta)$, and used them as input to the approximate parameter estimator $\widehat{\theta}$ specified in equation (16) for a fixed set of edge appearance probabilities $\boldsymbol{\mu_e} \in \mathbb{T}(G)$. This procedure yields an approximation $p(\mathbf{x};\widehat{\theta})$ to the true ML distribution $p(\mathbf{x};\theta)$. We performed the same procedure with $\mu_{st} = 1$ so as to obtain a BP-based estimate θ^{BP}. (Recall that $\boldsymbol{\mu_e} = \mathbf{1}$ is not a valid choice, from the perspective of the tree-reweighting, unless the graph itself is a tree).

We then perturbed the original problem by adding new noisy observations at the single nodes. In the exponential domain, the effect of this new set of (randomly specified) noisy observations can be modeled by the addition of a random vector δ. We chose a perturbation vector δ with components $\delta_{st} = 0$ for all edges $(s,t) \in E$, and made independent random choices $\delta_s \sim \mathcal{N}(0, \sigma^2_{pert})$ for each node. For all experiments reported here, we fixed the standard deviation $\sigma_{pert} = 0.5$.

In the exponential representation, the "true" posterior is given by $p(\mathbf{x};\theta + \delta)$. As a measure of ground truth, we computed the exact single node marginals P_s of this true posterior using the junction tree algorithm on the grid. We also formed the approximate posterior $p(\mathbf{x};\widehat{\theta} + \delta)$, and computed approximations $\widehat{T}_s$ to its single node marginal distributions using the tree-reweighted BP (Algorithm 1). We performed similar calculations using the BP-based approximate posterior $p(\mathbf{x}; \theta^{BP} + \delta)$, where we used the ordinary BP algorithm to compute approximations to its marginals.

Figure 3 shows the results of a number of experiments on a 5×5 grid. Each of the six plots shows five trials for one experiment and one method; plots in the top and bottom rows correspond, respectively, to the BP-based method and the tree-reweighted method as applied to same problem. Within each plot, the first trial shown in the top panel corresponds to the non-perturbed case ($\delta = \mathbf{0}$). For all of these trials, we see that the approximate marginals computed by either method are equivalent to the exact empirical marginals, as they should be. The remaining four panels show different trials with independent choices of δ as described above. For an original distribution $p(\mathbf{x}; \theta)$ with relatively weak couplings (column (a)), both methods perform well. However, in general, for problems with stronger weights (columns (b) and (c)), the behavior of the BP-based method can be erratic, whereas the tree-reweighted technique appears more stable.

Given the preliminary nature of these experimental results, there remain a number of open questions. One important issue is a better understanding of the bias in the approximation (i.e., difference between exact marginals of $p(\mathbf{x}; \theta^{ML} + \delta)$, and the approximations computed from $\widehat{\theta} + \delta$ using tree-reweighted BP). At least to first order, this bias is determined by the difference between the Fisher information matrix and the Hessian of $\mathcal{H}$. Secondly, in the simple experiments described here, the approximate ML estimator was given as input the exact marginals, which can be viewed as the limit of infinite data. It would be interesting to explore the behavior of the approximate ML estimator given only a finite sample of data. Moreover, the work described here treated the case of perfectly observed data. It remains to be seen if similar ideas, perhaps in conjunction with other variational techniques [e.g., 5], are useful in the partially observed context. Lastly, we have not yet explored the choice of the spanning tree probabilities $\boldsymbol{\mu_e}$, and its effect on approximate parameter estimation.

6 Discussion

Building on our previous work [8, 9, 10], we developed a family of tree-reweighted belief propagation algorithms for computing an upper bound on the log partition function, as well as the associated minimizing arguments (pseudomarginals). In the case of complete observations, we showed how to use this upper bound to derive a concave lower bound on the log likelihood. This lower bound has a unique global maximum, which can be obtained by performing pseudo-moment matching. We provided the results of some preliminary experiments to illustrate its behavior. In fact, the method presented here is a particular type of M-estimator, so that known techniques [7] can be used to analyze its asymptotic behavior.

Finally, it is worthwhile to note the connection between the approximate methods described in this paper, and the classical Legendre duality between the negative entropy function and the log partition function [1, 6]. In particular, the approximation methods in this paper are based on a parallel Legendre duality between the function $\mathcal{G}$ defined in equation (13), which gives a lower bound on the negative entropy, and the function $\mathcal{H}$ in equation (10), which gives an upper bound on the log partition function. As with the one-to-one Legendre mapping between exponential parameters and moments [1, 6, 8], our functions $\mathcal{H}$ and $\mathcal{G}$ induce a one-to-one mapping between approximate exponential parameters and moments.

Acknowledgements: Thanks to Max Welling and Yee Whye Teh for code for inference on grids. Work supported by ODDR&E MURI Grant DAAD19-00-1-0466 through ARO; by ONR Grant N00014-00-1-0089; and by AFOSR Grant F49620-00-1-0362.

References

[1] S. Amari. Differential geometry of curved exponential families — curvatures and information loss. *Annals of Statistics*, 10(2):357–385, 1982.

[2] D. Bertsekas. *Nonlinear programming.* Athena Scientific, Belmont, MA, 1995.

[3] T. Cover and J. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

[4] T. S. Jaakkola and M. Jordan. Computing upper and lower bounds on likelihoods in intractable networks. In *UAI*, pages 340–348, 1996.

[5] M. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. MIT Press, 1999.

[6] G. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[7] A. W. van der Vaart. *Asymptotic statistics.* Cambridge University Press, Cambridge, UK, 1998.

[8] M. J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches.* PhD thesis, MIT, January 2002.

[9] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Exact MAP estimates by (hyper)tree agreement. In *NIPS*, volume 15, December 2002.

[10] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. In *Proc. Uncertainty in Artificial Intelligence*, volume 18, pages 536–543, August 2002.

[11] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS 13*, pages 689–695. MIT Press, 2001.