

UC Irvine

UC Irvine Previously Published Works

Title

Variational Algorithms for Marginal MAP

Permalink

<https://escholarship.org/uc/item/5qh421tn>

Authors

Liu, Q
Ihler, A

Publication Date

2013-10-01

Copyright Information

This work is made available under the terms of a Creative Commons Attribution License, available at <https://creativecommons.org/licenses/by/4.0/>

Peer reviewed

Variational Algorithms for Marginal MAP

Qiang Liu

Alexander Ihler

*Donald Bren School of Information and Computer Sciences
University of California, Irvine
Irvine, CA, 92697-3425, USA*

QLIU1@UCI.EDU

IHLER@ICS.UCI.EDU

Editor: Amir Globerson

Abstract

The marginal maximum *a posteriori* probability (MAP) estimation problem, which calculates the mode of the marginal posterior distribution of a subset of variables with the remaining variables marginalized, is an important inference problem in many models, such as those with hidden variables or uncertain parameters. Unfortunately, marginal MAP can be NP-hard even on trees, and has attracted less attention in the literature compared to the joint MAP (maximization) and marginalization problems. We derive a general dual representation for marginal MAP that naturally integrates the marginalization and maximization operations into a joint variational optimization problem, making it possible to easily extend most or all variational-based algorithms to marginal MAP. In particular, we derive a set of “mixed-product” message passing algorithms for marginal MAP, whose form is a hybrid of max-product, sum-product and a novel “argmax-product” message updates. We also derive a class of convergent algorithms based on proximal point methods, including one that transforms the marginal MAP problem into a sequence of standard marginalization problems. Theoretically, we provide guarantees under which our algorithms give globally or locally optimal solutions, and provide novel upper bounds on the optimal objectives. Empirically, we demonstrate that our algorithms significantly outperform the existing approaches, including a state-of-the-art algorithm based on local search methods.

Keywords: graphical models, message passing, belief propagation, variational methods, maximum *a posteriori*, marginal-MAP, hidden variable models

1. Introduction

Graphical models such as Bayesian networks and Markov random fields provide a powerful framework for reasoning about conditional dependency structures over many variables, and have found wide application in many areas including error correcting codes, computer vision, and computational biology (Wainwright and Jordan, 2008; Koller and Friedman, 2009). Given a graphical model, which may be estimated from empirical data or constructed by domain expertise, the term *inference* refers generically to answering probabilistic queries about the model, such as computing marginal probabilities or maximum *a posteriori* estimates. Although these inference tasks are NP-hard in the worst case, recent algorithmic advances, including the development of variational methods and the family of algorithms collectively called belief propagation, provide approximate or exact solutions for these problems in many practical circumstances.

In this work we will focus on three common types of inference tasks. The first involves *maximization* or *max-inference* tasks, sometimes called maximum *a posteriori* (MAP) or most probable

explanation (MPE) tasks, which look for a mode of the joint probability. The second are *sum-inference* tasks, which include calculating the marginal probabilities or the normalization constant of the distribution (corresponding to the probability of evidence in a Bayesian network). Finally, the main focus of this work is on *marginal MAP*, a type of *mixed-inference* problem that seeks a partial configuration of variables that maximizes those variables’ marginal probability, with the remaining variables summed out.¹ Marginal MAP plays an essential role in many practical scenarios where there exist hidden variables or uncertain parameters. For example, a marginal MAP problem can arise as a MAP problem on models with hidden variables whose predictions are not of interest, or as a robust optimization variant of MAP with some unknown or noisily observed parameters marginalized w.r.t. a prior distribution. It can be also treated as a special case of the more complicated frameworks of stochastic programming (Birge and Louveaux, 1997) or decision networks (Howard and Matheson, 2005; Liu and Ihler, 2012).

These three types of inference tasks are listed in order of increasing difficulty: max-inference is NP-complete, while sum-inference is #P-complete, and mixed-inference is NP^{PP}-complete (Park and Darwiche, 2004; De Campos, 2011). Practically speaking, max-inference tasks have a host of efficient algorithms such as loopy max-product BP, tree-reweighted BP, and dual decomposition (see, e.g., Koller and Friedman, 2009; Sontag et al., 2011). Sum-inference is more difficult than max-inference: for example there are models, such as those with binary attractive pairwise potentials, on which sum-inference is #P-complete but max-inference is tractable (Greig et al., 1989; Jerrum and Sinclair, 1993).

Mixed-inference is even much harder than either max- or sum- inference problems alone: marginal MAP can be NP-hard even on tree structured graphs, as illustrated in the example by Koller and Friedman (2009) in Figure 1. The difficulty arises in part because the max and sum operators do not commute, causing the feasible elimination orders to have much higher induced width than for sum- or max-inference. Viewed another way, the marginalization step may destroy the dependency structure of the original graphical model, making the subsequent maximization step far more challenging. Probably for these reasons, there is much less work on marginal MAP than that on joint MAP or marginalization, despite its importance to many practical problems. In practice, it is common to over-use the simpler joint MAP or marginalization even when marginal MAP would be more appropriate. This may cause serious problems, as we illustrate in Example 1 and our empirical results in Section 9.

1.1 Contributions

We reformulate the mixed-inference problem to a joint maximization problem as a free energy objective that extends the well-known log-partition function duality form, making it possible to easily extend essentially arbitrary variational algorithms to marginal MAP. In particular, we propose a novel “mixed-product” BP algorithm that is a hybrid of max-product, sum-product, and a special “argmax-product” message updates, as well as a convergent proximal point algorithm that works by iteratively solving pure (or annealed) marginalization tasks. We also present junction graph BP variants of our algorithms, that work on models with higher order cliques. We also discuss mean field methods and highlight their connection to the expectation-maximization (EM) algorithm. We give theoretical guarantees on the global and local optimality of our algorithms for cases when the

1. In some literature (e.g., Park and Darwiche, 2004), marginal MAP is simply referred to as MAP, and the joint MAP problem is called MPE.

sum variables form tree structured subgraphs. Our numerical experiments show that our methods can provide significantly better solutions than existing algorithms, including a similar hybrid message passing algorithm by Jiang et al. (2011) and a state-of-the-art algorithm based on local search methods. A preliminary version of this work has appeared in Liu and Ihler (2011b).

1.2 Related Work

Expectation-maximization (EM) or variational EM provide one straightforward approach for marginal MAP, by viewing the sum nodes as hidden variables and the max nodes as parameters to be estimated; however, EM is prone to getting stuck at sub-optimal configurations. We show that EM can be treated as a special case of our framework when a mean field-like approximation is applied. Other classical state-of-the-art approaches include local search methods (e.g., Park and Darwiche, 2004), Markov chain Monte Carlo methods (e.g., Doucet et al., 2002; Yuan et al., 2004), and variational elimination based methods (e.g., Dechter and Rish, 2003; Mauá and de Campos, 2012). Jiang et al. (2011) recently proposed a hybrid message passing algorithm that has a similar form to our mixed-product BP algorithm, but without theoretical guarantees; we show in Section 5.3 that Jiang et al. (2011) can be viewed as an approximation of the marginal MAP problem that exchanges the order of sum and max operators. Another message-passing-style algorithm was proposed very recently in Altarelli et al. (2011) for general multi-stage stochastic optimization problems based on survey propagation, which again does not have optimality guarantees and has a relatively more complicated form. Finally, Ibrahimi et al. (2011) introduces a robust max-product belief propagation for solving a related worst-case robust optimization problem, where the hidden variables are minimized instead of marginalized. To the best of our knowledge, our work is the first general variational framework for marginal MAP, and provides the first strong optimality guarantees.

We begin in Section 2 by introducing background information on graphical models and variational inference. We then introduce a novel variational dual representation for marginal MAP in Section 3, and propose analogues of the Bethe and tree-reweighted approximations for marginal MAP in Section 4. A class of “mixed-product” message passing algorithms is proposed and analyzed in Section 5 and convergent alternatives are proposed in Section 6 based on proximal point methods. We then discuss the EM algorithm and its connection to our framework in Section 7, and provide an extension of our algorithms to junction graphs in Section 8. Finally, we present numerical results in Section 9 and conclude the paper in Section 10.

2. Background

We give an overview of different inference problems on graphical models, and introduce the variational framework as applied to max- and sum- inference problems.

2.1 Graphical Models

Let $x = \{x_1, x_2, \dots, x_n\}$ be a random vector in a discrete space $\mathcal{X} = \mathcal{X}_1 \times \dots \times \mathcal{X}_n$. Let $V = \{1, \dots, n\}$. For an index set $\alpha \subseteq V$, denote by x_α the sub-vector $\{x_i : i \in \alpha\}$, and similarly, $\mathcal{X}_\alpha$ the cross product of $\{\mathcal{X}_i : i \in \alpha\}$. A graphical model defines a factorized probability on x ,

$$p(x) = \frac{1}{Z(\psi)} \prod_{\alpha \in I} \psi_\alpha(x_\alpha) \quad \text{or} \quad p(x; \theta) = \exp\left[\sum_{\alpha \in I} \theta_\alpha(x_\alpha) - \Phi(\theta)\right],$$

where I is a set of subsets of variable indexes, $\psi_\alpha: \mathcal{X}_\alpha \rightarrow \mathbb{R}^+$ is called a factor function, and $\theta_\alpha(x_\alpha) = \log \psi_\alpha(x_\alpha)$. Since the x_i are discrete, the functions ψ and θ are tables; by alternatively viewing θ as a vector, it is interpreted as the natural parameter in an overcomplete, exponential family representation. Let ψ and θ be the joint vector of all ψ_α and θ_α respectively, for example, $\theta = \{\theta_\alpha(x_\alpha): \alpha \in I, x_\alpha \in \mathcal{X}_\alpha\}$. The normalization constant $Z(\psi)$, called *partition function*, normalizes the probability to sum to one, and $\Phi(\theta) := \log Z(\psi)$ is called the log-partition function,

$$\Phi(\theta) = \log \sum_{x \in \mathcal{X}} \exp[\theta(x)],$$

where we define $\theta(x) = \sum_{\alpha \in I} \theta_\alpha(x_\alpha)$ to be the joint potential function that maps from $\mathcal{X}$ to $\mathbb{R}$. The factorization structure of $p(x)$ can be represented by an undirected graph $G = (V, E)$, where each node $i \in V$ maps to a variable x_i , and each edge $(ij) \in E$ corresponds to two variables x_i and x_j that coappear in some factor function ψ_α , that is, $\{i, j\} \subseteq \alpha$. The set I is then a set of cliques (fully connected subgraphs) of G . For the purpose of illustration, we mainly restrict our scope on the set of pairwise models, on which I is the set of nodes and edges, that is, $I = E \cup V$. However, we show how to extend our algorithms to models with higher order cliques in Section 8.

2.2 Sum-Inference Problems and Variational Approximation

Sum-inference is the task of marginalizing (summing out) variables in the model, for example, calculating the marginal probabilities of single variables, or the normalization constant Z ,

$$p(x_i) = \sum_{x_{V \setminus \{i\}}} \exp[\theta(x) - \Phi(\theta)], \quad \Phi(\theta) = \log \sum_x \exp[\theta(x)].$$

Unfortunately, the problem is generally #P-complete, and the straightforward calculation requires summing over an exponential number of terms. Variational methods are a class of approximation algorithms that transform the marginalization problem into a continuous optimization problem, which is then typically solved approximately.

2.2.1 MARGINAL POLYTOPE

The marginal polytope is a key concept in variational inference. We define the *marginal polytope* $\mathbb{M}$ to be the set of local marginal probabilities $\tau = \{\tau_\alpha(x_\alpha): \alpha \in I\}$ that are extensible to a valid joint distribution, that is,

$$\mathbb{M} = \{\tau : \exists \text{ joint distribution } q(x), \text{ s.t. } \tau_\alpha(x_\alpha) = \sum_{x_{V \setminus \alpha}} q(x) \text{ for } \forall \alpha \in I\}.$$

Denote by $Q[\tau]$ the set of joint distributions whose marginals are consistent with $\tau \in \mathbb{M}$; by the principle of maximum entropy (Jaynes, 1957), there exists a unique distribution in $Q[\tau]$ that has maximum entropy and follows the exponential family form for some θ .² With an abuse of notation, we denote these unique global distributions by $\tau(x)$, and we do not distinguish $\tau(x)$ and τ when it is clear from the context.

2. In the case that $p(x)$ has zero elements, the maximum entropy distribution is still unique and satisfies the exponential family form, but the corresponding θ has negative infinite values (Jaynes, 1957).

2.2.2 LOG-PARTITION FUNCTION DUALITY

A key result to many variational methods is that the log-partition function $\Phi(\theta)$ is a convex function of θ and can be rewritten into a convex dual form,

$$\Phi(\theta) = \max_{\tau \in \mathbb{M}} \{ \langle \theta, \tau \rangle + H(\tau) \}, \quad (1)$$

where $\langle \theta, \tau \rangle = \sum_{\alpha} \sum_{x_{\alpha}} \theta_{\alpha}(x_{\alpha}) \tau_{\alpha}(x_{\alpha})$ is the vectorized inner product, and $H(\tau)$ is the entropy of the corresponding global distribution $\tau(x)$, that is, $H(\tau) = -\sum_x \tau(x) \log \tau(x)$. The unique maximum τ^* of (1) exactly equals the marginals of the original distribution $p(x; \theta)$, that is, $\tau^*(x) = p(x; \theta)$. We call $F_{\text{sum}}(\tau, \theta) = \langle \theta, \tau \rangle + H(\tau)$ the sum-inference free energy (although technically the *negative* free energy).

The dual form (1) transforms the marginalization problem into a continuous optimization, but does not make it any easier: the marginal polytope $\mathbb{M}$ is defined by an exponential number of linear constraints, and the entropy term in the objective function is as difficult to calculate as the log-partition function. However, (1) provides a framework for deriving efficient approximate inference algorithms by approximating both the marginal polytope and the entropy (Wainwright and Jordan, 2008).

2.2.3 BP-LIKE METHODS

Many approximation methods replace $\mathbb{M}$ with the *locally consistent polytope* $\mathbb{L}$; in pairwise models, it is the set of singleton and pairwise “pseudo-marginals” $\{\tau_i(x_i) : i \in V\}$ and $\{\tau_{ij}(x_i, x_j) : (ij) \in E\}$ that are consistent on their intersections, that is,

$$\mathbb{L} = \{ \tau_i, \tau_{ij} : \sum_{x_i} \tau_{ij}(x_i, x_j) = \tau_j(x_j), \sum_{x_i} \tau_i(x_i) = 1, \tau_{ij}(x_i, x_j) \geq 0 \}. \quad (2)$$

Since not all such pseudo-marginals have valid global distributions, it is easy to see that $\mathbb{L}$ is an outer bound of $\mathbb{M}$, that is, $\mathbb{M} \subseteq \mathbb{L}$. Note that this means there may not exist a global distribution $\tau(x)$ for τ in $\mathbb{L}$.

The free energy remains intractable (and is not even well-defined) in $\mathbb{L}$. We typically approximate the free energy by a combination of singleton and pairwise entropies, which only requires knowing τ_i and τ_{ij} . For example, the Bethe free energy approximation (Yedidia et al., 2003) is

$$H(\tau) \approx \sum_{i \in V} H_i(\tau) - \sum_{(ij) \in E} I_{ij}(\tau), \quad \Phi(\theta) \approx \max_{\tau \in \mathbb{L}} \{ \langle \theta, \tau \rangle + \sum_{i \in V} H_i(\tau) - \sum_{(ij) \in E} I_{ij}(\tau) \}, \quad (3)$$

where $H_i(\tau)$ is the entropy of $\tau_i(x_i)$ and $I_{ij}(\tau)$ the mutual information of x_i and x_j , that is,

$$H_i(\tau) = -\sum_{x_i} \tau_i(x_i) \log \tau_i(x_i), \quad I_{ij}(\tau) = \sum_{x_i, x_j} \tau_{ij}(x_i, x_j) \log \frac{\tau_{ij}(x_i, x_j)}{\tau_i(x_i) \tau_j(x_j)}.$$

We sometimes abbreviate $H_i(\tau)$ and $I_{ij}(\tau)$ into H_i and I_{ij} for convenience. The well-known loopy belief propagation (BP) algorithm of Pearl (1988) can be interpreted as a fixed point algorithm to optimize the Bethe free energy in (3) on the locally consistent polytope $\mathbb{L}$ (Yedidia et al., 2003). Unfortunately, the Bethe free energy is a non-concave function of τ , causing (3) to be a non-convex

optimization. The tree reweighted (TRW) free energy is a convex surrogate of the Bethe free energy (Wainwright et al., 2005a),

$$\Phi(\theta) \approx \max_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{i \in V} H_i(\tau) - \sum_{(ij) \in E} \rho_{ij} I_{ij}(\tau) \right\}, \quad (4)$$

where $\{\rho_{ij} : (ij) \in E\}$ is a set of positive edge appearance probabilities obtained from a weighted collection of spanning trees of G (see Wainwright et al. (2005a) and Section 4.2 for the detailed definition). The TRW approximation in (4) is a convex optimization problem, and is guaranteed to give an upper bound of the true log-partition function. A message passing algorithm similar to loopy BP, called tree reweighted BP, can be derived as a fixed point algorithm for solving the convex optimization in (4).

2.2.4 MEAN-FIELD-BASED METHODS

Mean-field-based methods are another set of approximate inference algorithms, which work by restricting $\mathbb{M}$ to a set of tractable distributions, on which both the marginal polytope and the joint entropy are tractable. Precisely, let $\mathbb{M}_{mf}$ be a subset of $\mathbb{M}$ that corresponds to a set of tractable distributions, for example, the set of fully factored distributions, $\mathbb{M}_{mf} = \{\tau \in \mathbb{M} : \tau(x) = \prod_{i \in V} \tau_i(x_i)\}$. Note that the joint entropy $H(\tau)$ for any $\tau \in \mathbb{M}_{mf}$ decomposes to the sum of singleton entropies $H_i(\tau)$ of the marginal distributions $\tau_i(x_i)$. This method then approximates the log-partition function (1) by

$$\max_{\tau \in \mathbb{M}_{mf}} \left\{ \langle \theta, \tau \rangle + \sum_{i \in V} H_i(\tau) \right\}, \quad (5)$$

which is guaranteed to give a lower bound of the log-partition function. Unfortunately, mean field methods usually lead to non-convex optimization problems, because $\mathbb{M}_{mf}$ is often a non-convex set. In practice, block coordinate descent methods can be adopted to find the local optima of (5).

2.3 Max-Inference Problems

Combinatorial maximization (max-inference), or maximum *a posteriori* (MAP), problems are the tasks of finding a mode of the joint probability. That is,

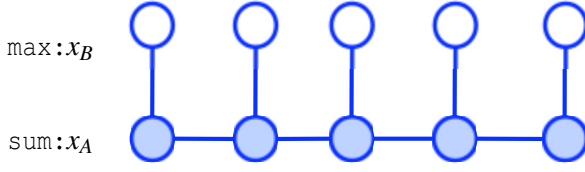
$$\Phi_{\infty}(\theta) = \max_x \theta(x), \quad x^* = \arg \max_x \theta(x),$$

where x^* is a MAP configuration and $\Phi_{\infty}(\theta)$ the optimal energy value. This problem can be reformed into a linear program,

$$\Phi_{\infty}(\theta) = \max_{\tau \in \mathbb{M}} \langle \theta, \tau \rangle, \quad (6)$$

which attains its maximum when $\tau^*(x) = 1(x = x^*)$, where $1(\cdot)$ is the Kronecker delta function, defined as $1(t) = 1$ if condition t is true, and zero otherwise. If there are multiple MAP solutions, say $\{x^{*k} : k = 1, \dots, K\}$, then any convex combination $\sum_k c_k 1(x = x^{*k})$ with $\sum_k c_k = 1, c_i \geq 0$ leads to a maximum of (6).

The problem in (6) remains NP-hard, because the marginal polytope $\mathbb{M}$ includes exponentially many inequality constraints. Most variational methods for MAP (e.g., Wainwright et al., 2005b; Werner, 2007) can be interpreted as relaxing $\mathbb{M}$ to the locally consistent polytop $\mathbb{L}$, yielding a linear relaxation of the original integer programming problem. Note that (6) differs from (1) only by its lack of an entropy term; in the next section, we generalize this similarity to marginal MAP.



Marginal MAP:

$$\begin{aligned} x_B^* &= \arg \max_{x_B} p(x_B) \\ &= \arg \max_{x_B} \sum_{x_A} p(x). \end{aligned}$$

Figure 1: An example from Koller and Friedman (2009) in which a marginal MAP query on a tree requires exponential time complexity. The marginalization over x_A destroys the conditional dependency structure in the marginal distribution $p(x_B)$, causing an intractable maximization problem over x_B . The exact variable elimination method, which sequentially marginalizes the sum nodes and then maximizes the max nodes, has time complexity of $O(\exp(n))$, where n is the length of the chain.

2.4 Marginal MAP Problems

Marginal MAP is simply a hybrid of the max- and sum- inference tasks. Let A be a subset of nodes V , and $B = V \setminus A$ be the complement of A . The marginal MAP problem seeks a partial configuration x_B^* that has the maximum marginal probability $p(x_B) = \sum_{x_A} p(x)$, where A is the set of sum nodes to be marginalized out, and B the max nodes to be optimized. We call this a type of “mixed-inference” problem, since it involves more than one type of variable elimination operator. To facilitate developing our duality results, we formulate marginal MAP in terms of the exponential family representation,

$$\Phi_{AB}(\theta) = \max_{x_B} Q(x_B; \theta), \quad \text{where } Q(x_B; \theta) = \log \sum_{x_A} \exp[\theta(x)], \quad (7)$$

where the maximum point x_B^* of $Q(x_B; \theta)$ is the marginal MAP solution. Although similar to max- and sum-inference, marginal MAP is significantly harder than either of them. A classic example is shown in Figure 1, where marginal MAP is NP-hard even on a tree structured graph (Koller and Friedman, 2009). The main difficulty arises because the max and sum operators do not commute, which restricts feasible elimination orders to those with *all* the sum nodes eliminated before *any* max nodes. In the worst case, marginalizing the sum nodes x_A may destroy any conditional independence among the max nodes x_B , making it difficult to represent or optimize $Q(x_B; \theta)$, even when the sum part alone is tractable (such as when the nodes in A form a tree).

Despite its computational difficulty, marginal MAP plays an essential role in many practical scenarios. The marginal MAP configuration x_B^* in (7) is Bayes optimal in the sense that it minimizes the expected error on B , $\mathbb{E}[1(x_B^* = x_B)]$, where $\mathbb{E}[\cdot]$ denotes the expectation under distribution $p(x; \theta)$. Here, the variables x_A are not included in the error criterion, for example because they are “nuisance” hidden variables of no direct interest, or unobserved or inaccurately measured model parameters. In contrast, the joint MAP configuration x^* minimizes the joint error $\mathbb{E}[1(x^* = x)]$, but gives no guarantees on the partial error $\mathbb{E}[1(x_B^* = x_B)]$. In practice, perhaps because of the wide availability of efficient algorithms for joint MAP, researchers tend to over-use joint MAP even in cases where marginal MAP would be more appropriate. The following toy example shows that this seemingly reasonable approach can sometimes cause serious problems.

Example 1 (Weather Dilemma). Denote by $x_b \in \{\text{rainy}, \text{sunny}\}$ the weather condition of Irvine, and $x_a \in \{\text{walk}, \text{drive}\}$ whether Alice drives or walks to the school depending on the weather con-

dition. Assume the probabilities of x_b and x_a are

$p(x_b) :$	rainy	0.4	$p(x_a x_b) :$		walk	drive
	sunny	0.6		rainy	1/8	7/8
				sunny	1/2	1/2

The task is to calculate the most likely weather condition of Irvine, which is obviously sunny according to $p(x_b)$. The marginal MAP, $x_b^* = \arg \max_{x_b} p(x_b) = \text{sunny}$, gives the correct answer. However, the full MAP estimator, $[x_a^*, x_b^*] = \arg \max p(x_a, x_b) = [\text{drive}, \text{rainy}]$, gives answer $x_b^* = \text{rainy}$ (by dropping the x_a^* component), which is obviously wrong. Paradoxically, if $p(x_a|x_b)$ is changed (say, corresponding to a different person), the solution returned by full MAP could be different.

In the above example, since no evidence on x_a is observed, the conditional probability $p(x_a|x_b)$ does not provide useful information for x_b , but instead provides misleading information when it is incorporated in the full MAP estimator. The marginal MAP, on the other hand, eliminates the influence of the irrelevant $p(x_a|x_b)$ by marginalizing (or averaging) x_a . In general, the marginal MAP and full MAP can differ significantly when the uncertainty in the hidden variables changes as a function of x_B .

3. A Dual Representation for Marginal MAP

In this section, we present our main result, a dual representation of the marginal MAP problem (7). Our dual representation generalizes that of sum-inference in (1) and max-inference in (6), and provides a unified framework for solving marginal MAP problems.

Theorem 2. *The marginal MAP energy $\Phi_{AB}(\theta)$ in (7) has a dual representation,*

$$\Phi_{AB}(\theta) = \max_{\tau \in \mathbb{M}} \{ \langle \theta, \tau \rangle + H_{A|B}(\tau) \}, \quad (8)$$

where $H_{A|B}(\tau)$ is a conditional entropy, $H_{A|B}(\tau) = -\sum_x \tau(x) \log \tau(x_A|x_B)$. If $Q(x_B; \theta)$ has a unique maximum x_B^* , the maximum point τ^* of (8) is also unique, satisfying $\tau^*(x) = \tau^*(x_B) \tau^*(x_A|x_B)$, where $\tau^*(x_B) = 1(x_B = x_B^*)$ and $\tau^*(x_A|x_B) = p(x_A|x_B; \theta)^3$.

Proof. For any $\tau \in \mathbb{M}$ and its corresponding global distribution $\tau(x)$, consider the conditional KL divergence between $\tau(x_A|x_B)$ and $p(x_A|x_B; \theta)$,

$$\begin{aligned} D_{\text{KL}}[\tau(x_A|x_B) || p(x_A|x_B; \theta)] &= \sum_x \tau(x) \log \frac{\tau(x_A|x_B)}{p(x_A|x_B; \theta)} \\ &= -H_{A|B}(\tau) - \mathbb{E}_{\tau}[\log p(x_A|x_B; \theta)] \\ &= -H_{A|B}(\tau) - \mathbb{E}_{\tau}[\theta(x)] + \mathbb{E}_{\tau}[Q(x_B; \theta)] \geq 0, \end{aligned}$$

where $H_{A|B}(\tau)$ is the conditional entropy on $\tau(x)$; the equality on the last line holds because $p(x_A|x_B; \theta) = \exp(\theta(x) - Q(x_B; \theta))$; the last inequality follows from the nonnegativity of KL divergence, and is tight if and only if $\tau(x_A|x_B) = p(x_A|x_B; \theta)$ for all x_A and x_B that $\tau(x_B) \neq 0$. Therefore, we have for any $\tau(x)$,

$$\Phi_{AB}(\theta) = \max_{x_B} Q(x_B; \theta) \geq \mathbb{E}_{\tau}[Q(x_B; \theta)] \geq \mathbb{E}_{\tau}[\theta(x)] + H_{A|B}(\tau).$$

3. Since $\tau(x_B) = 0$ if $x_B \neq x_B^*$, we do not necessarily need to define $\tau^*(x_A|x_B)$ for $x_B \neq x_B^*$.

Problem Type	Primal Form	Dual Form
Max-Inference	$\log \max_x \exp(\theta(x))$	$\max_{\tau \in \mathbb{M}} \{\langle \theta, \tau \rangle\}$
Sum-Inference	$\log \sum_x \exp(\theta(x))$	$\max_{\tau \in \mathbb{M}} \{\langle \theta, \tau \rangle + H(\tau)\}$
Marginal MAP	$\log \max_{x_B} \sum_{x_A} \exp(\theta(x))$	$\max_{\tau \in \mathbb{M}} \{\langle \theta, \tau \rangle + H_{A B}(\tau)\}$

Table 1: The primal and dual forms of the three inference types. The dual forms of sum-inference and max-inference are well known; the form for marginal MAP is a contribution of this work. Intuitively, the max vs. sum operators in the primal form determine the conditioning set of the conditional entropy term in the dual form.

It is easy to show that the two inequality signs are tight if and only if $\tau(x)$ equals $\tau^*(x)$ as defined above. Substituting $\mathbb{E}_\tau[\theta(x)] = \langle \theta, \tau \rangle$ completes the proof. $\square$

Remark 1. If $Q(x_B; \theta)$ has multiple maxima $\{x_B^{*k}\}$, each corresponding to a distribution $\tau^{*k}(x) = 1(x_B = x_B^*)p(x_A|x_B; \theta)$, then the set of maximum points of (8) is the convex hull of $\{\tau^{*k}\}$.

Remark 2. Theorem 2 naturally integrates the marginalization and maximization sub-problems into one joint optimization problem, providing a novel and efficient treatment for marginal MAP beyond the traditional approaches that treat the marginalization sub-problem as a sub-routine of the maximization problem. As we show in Section 5, this enables us to derive efficient “mixed-product” message passing algorithms that simultaneously takes marginalization and maximization steps, avoiding expensive and possibly wasteful inner loop steps in the marginalization sub-routine.

Remark 3. Since we have $H_{A|B}(\tau) = H(\tau) - H_B(\tau)$ by the entropic chain rule (Cover and Thomas, 2006), the objective function in (8) can be view as a “truncated” free energy,

$$F_{mix}(\tau, \theta) := \langle \theta, \tau \rangle + H_{A|B}(\tau) = F_{sum}(\tau, \theta) - H_B(\tau),$$

where the entropy $H_B(\tau)$ of the max nodes x_B are removed from the regular sum-inference free energy $F_{sum}(\tau, \theta) = \langle \theta, \tau \rangle + H(\tau)$. Theorem 2 generalizes the dual form of both sum-inference (1) and max-inference (6), since it reduces to those forms when the max set B is empty or all nodes, respectively. Table 1 shows all three forms together for comparison. Intuitively, since the entropy $H_B(\tau)$ is removed from the objective, the optimal marginal $\tau^*(x_B)$ tends to have lower entropy and its probability mass concentrates on the optimal configurations $\{x_B^*\}$. Alternatively, the $\tau^*(x)$ can be interpreted as the marginals obtained by clamping the value of x_B at x_B^* on the distribution $p(x; \theta)$, that is, $\tau^*(x) = p(x|x_B = x_B^*; \theta)$.

Remark 4. Unfortunately, subtracting the $H_B(\tau)$ term causes some subtle difficulties. First, $H_B(\tau)$ (and hence $F_{mix}(\tau, \theta)$) may be intractable to calculate even when the joint entropy $H(\tau)$ is tractable, because the marginal distribution $p(x_B) = \sum_{x_A} p(x)$ does not necessarily inherit the conditional dependency structure of the joint distribution. Therefore, the dual optimization in (8) may be intractable even on a tree, reflecting the intrinsic difficulty of marginal MAP compared to full MAP or marginalization. Interestingly, we show in the sequel that a certificate of optimality can still be obtained on general tree graphs in some cases.

Secondly, the conditional entropy $H_{A|B}(\tau)$ (and hence $F_{mix}(\tau, \theta)$) is concave, but not strictly concave, with respect to τ . This creates additional difficulty when optimizing (8), since many iterative optimization algorithms, such as coordinate descent, can lose their typical convergence or optimality guarantees when the objective function is not strongly convex.

3.1 Smoothed Approximation

To sidestep the issue of non-strictly convexity, we introduce a smoothed approximation of $F_{mix}(\tau, \theta)$ that “adds back” part of the missing $H_B(\tau)$ term,

$$F_{mix}^\varepsilon(\tau, \theta) = \langle \theta, \tau \rangle + H_{A|B}(\tau) + \varepsilon H_B(\tau),$$

where ε is a small positive constant. Similar smoothing techniques have also been applied to solve the standard MAP problem; see, for example, Hazan and Shashua (2010); Meshi et al. (2012). We show in the following theorem that this smoothed dual approximation is closely connected to a direct approximation in the primal domain.

Theorem 3. *Let ε be a positive constant, and $Q(x_B; \theta)$ as defined in (7). Define*

$$\Phi_{AB}^\varepsilon(\theta) = \log \left\{ \left[\sum_{x_B} \exp(Q(x_B; \theta)) \right]^{1/\varepsilon} \right\},$$

then we have

$$\Phi_{AB}^\varepsilon(\theta) = \max_{\tau \in \mathbb{M}} \{ \langle \theta, \tau \rangle + H_{A|B}(\tau) + \varepsilon H_B(\tau) \}.$$

In addition, we have

$$\lim_{\varepsilon \rightarrow 0^+} \Phi_{AB}^\varepsilon(\theta) = \Phi_{AB}(\theta),$$

where $\varepsilon \rightarrow 0^+$ denotes approaching zero from the positive side.

Proof. The proof is similar to that of Theorem 2, but exploits the non-negativity of a weighted sum of two KL divergence terms,

$$D_{KL}[\tau(x_A|x_B) || p(x_A|x_B; \theta)] + \varepsilon D_{KL}[\tau(x_B) || p(x_B)].$$

The remaining part follows directly from the standard zero temperature limit formula,

$$\lim_{\varepsilon \rightarrow 0^+} \left[\sum_x f(x)^{1/\varepsilon} \right]^\varepsilon = \max_x f(x), \quad (9)$$

where $f(x)$ is any function with positive values. □

4. Variational Approximations for Marginal MAP

Theorem 2 transforms the marginal MAP problem into a variational form, but obviously does not decrease its computational hardness. Fortunately, many well-established variational techniques for sum- and max-inference can be extended to apply to (8), opening a new door for deriving novel approximate algorithms for marginal MAP. In the spirit of Wainwright and Jordan (2008), one can either relax $\mathbb{M}$ to a simpler outer bound like $\mathbb{L}$ and replace $F_{mix}(\tau, \theta)$ by some tractable form to give

algorithms similar to loopy BP or TRW BP, or restrict $\mathbb{M}$ to a tractable subset like $\mathbb{M}_{mf}$ to give mean-field-like algorithms. In the sequel, we demonstrate several such approximation schemes, mainly focusing on the BP-like methods with pairwise free energies. We will briefly discuss mean-field-like methods when we connect to EM in section 7, and derive an extension to junction graphs that exploits higher order approximations in Section 8. Our framework can be easily adopted to take advantage of other, more advanced variational techniques, like those using higher order cliques (e.g., Yedidia et al., 2005; Globerson and Jaakkola, 2007; Liu and Ihler, 2011a; Hazan et al., 2012) or more advanced optimization methods like dual decomposition (Sontag et al., 2011) or alternating direction method of multipliers (Boyd et al., 2010).

We start by characterizing the graph structure on which marginal MAP is tractable.

Definition 4.1. *We call G an A - B tree if there exists a partial order on the node set $V = A \cup B$, satisfying*

- 1) **Tree-order.** *For any $i \in V$, there is at most one other node $j \in V$ (called its parent), such that $j \prec i$ and $(ij) \in E$;*
- 2) **A-B Consistency.** *For any $a \in A$ and $b \in B$, we have $b \prec a$.*

We call such a partial order an A - B tree-order of G .

For further notation, let $G_A = (A, E_A)$ be the subgraph induced by nodes in A , that is, $E_A = \{(ij) \in E : i \in A, j \in A\}$, and similarly for $G_B = (B, E_B)$. Let $\partial_{AB} = \{(ij) \in E : i \in A, j \in B\}$ be the edges that join sets A and B .

Obviously, marginal MAP on an A - B tree can be tractably solved by sequentially eliminating the variables along the A - B tree-order (see, e.g., Koller and Friedman, 2009). We show that its dual optimization is also tractable in this case.

Lemma 4. *If G is an A - B tree, then*

- 1) *The locally consistent polytope equals the marginal polytope, that is, $\mathbb{M} = \mathbb{L}$.*
- 2) *The conditional entropy has a pairwise decomposition,*

$$H_{A|B}(\tau) = \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_A \cup \partial_{AB}} I_{ij}(\tau). \quad (10)$$

Proof. 1) The fact that $\mathbb{M} = \mathbb{L}$ on trees is a standard result; see Wainwright and Jordan (2008) for details.

2) Because G is an A - B tree, both $p(x)$ and $p(x_B)$ have tree structured conditional dependency. We then have (see, e.g., Wainwright and Jordan, 2008) that

$$H(\tau) = \sum_{i \in V} H_i(\tau) - \sum_{(ij) \in E} I_{ij}(\tau), \quad \text{and} \quad H_B(\tau) = \sum_{i \in B} H_i(\tau) - \sum_{(ij) \in E_B} I_{ij}(\tau).$$

Equation (10) follows by using the entropic chain rule $H_{A|B}(\tau) = H(\tau) - H_B(\tau)$. $\square$

4.1 Bethe-like Free Energy

Lemma 4 suggests that the free energy of A - B trees can be decomposed into singleton and pairwise terms that are easy to deal with. This is not true for general graphs, but motivates a “Bethe” like approximation,

$$\Phi_{\text{bethe}}(\theta) = \max_{\tau \in \mathbb{L}} F_{\text{bethe}}(\tau, \theta), \quad F_{\text{bethe}}(\tau, \theta) = \langle \theta, \tau \rangle + \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_A \cup \partial_{AB}} I_{ij}(\tau), \quad (11)$$

where $F_{\text{bethe}}(\tau, \theta)$ is a “truncated” Bethe free energy, whose entropy and mutual information terms that involve only max nodes are truncated. If G is an A - B tree, Φ_{bethe} equals the true Φ_{AB} , giving an intuitive justification. In the sequel we give more general theoretical conditions under which this approximation gives the exact solution, and we find empirically that it usually gives surprisingly good solutions in practice. Similar to the regular Bethe approximation, (11) leads to a nonconvex optimization, and we will derive both message passing algorithms and provably convergent algorithms to solve it.

4.2 Tree-reweighted Free Energy

Following the idea of TRW belief propagation (Wainwright et al., 2005a), we construct an approximation of marginal MAP using a convex combination of A - B subtrees (subgraphs of G that are A - B trees). Let $\mathcal{T}_{AB}$ be a collection of A - B subtrees of G . We assign with each $T \in \mathcal{T}_{AB}$ a weight w_T satisfying $w_T \geq 0$ and $\sum_{T \in \mathcal{T}_{AB}} w_T = 1$. For each A - B sub-tree $T = (V, E_T)$, define

$$H_{A|B}(\tau; T) = \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_T \setminus E_B} I_{ij}(\tau).$$

As shown in Wainwright and Jordan (2008), the $H_{A|B}(\tau; T)$ is always a concave function of τ on $\mathbb{L}$, and $H_{A|B}(\tau) \leq H_{A|B}(\tau; T)$ for all $\tau \in \mathbb{M}$ and $T \in \mathcal{T}_{AB}$. More generally, we have $H_{A|B}(\tau) \leq \sum_{T \in \mathcal{T}_{AB}} w_T H_{A|B}(\tau; T)$, which can be transformed to

$$H_{A|B}(\tau) \leq \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_A \cup \partial_{AB}} \rho_{ij} I_{ij}(\tau), \quad (12)$$

where $\rho_{ij} = \sum_{T: (ij) \in E_T} w_T$ are the edge appearance probabilities as defined in Wainwright and Jordan (2008). Replacing $\mathbb{M}$ with $\mathbb{L}$ and $H_{A|B}(\tau)$ with the bound in (12) leads to a TRW-like approximation of marginal MAP,

$$\Phi_{\text{trw}}(\theta) = \max_{\tau \in \mathbb{L}} F_{\text{trw}}(\tau, \theta), \quad F_{\text{trw}}(\tau, \theta) = \langle \theta, \tau \rangle + \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_A \cup \partial_{AB}} \rho_{ij} I_{ij}(\tau). \quad (13)$$

Since $\mathbb{L}$ is an outer bound of $\mathbb{M}$, and F_{trw} is a concave upper bound of the true free energy, we can guarantee that $\Phi_{\text{trw}}(\theta)$ is always an upper bound of $\Phi_{AB}(\theta)$. To our knowledge, this provides the first known convex relaxation for upper bounding marginal MAP. One can also optimize the weights $\{w_T : T \in \mathcal{T}_{AB}\}$ to get the tightest upper bound using methods similar to those used for regular TRW BP (see Wainwright et al., 2005a).

4.3 Global Optimality Guarantees

We show the global optimality guarantees of the above approximations under some circumstances. In this section, we always assume G_A is a tree, and hence the objective function is tractable to

calculate for a given x_B . However, the optimization component remains intractable in this case, because the marginalization step destroys the decomposition structure of the objective function (see Figure 1). It is thus nontrivial to see how the Bethe and TRW approximations behave in this case.

In general, suppose we approximate $\Phi_{AB}(\theta)$ using the following pairwise approximation,

$$\Phi_{tree}(\theta) = \max_{\tau \in \mathbb{L}} \left\{ \langle \theta, \tau \rangle + \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_A} I_{ij}(\tau) - \sum_{(ij) \in \partial_{AB}} \rho_{ij} I_{ij}(\tau) \right\}, \quad (14)$$

where the weights on the sum part, $\{\rho_{ij} : (ij) \in E_A\}$, have been fixed to be ones. This choice makes sure that the sum part is “intact” in the approximation, while the weights on the crossing edges, $\rho_{AB} = \{\rho_{ij} : (ij) \in \partial_{AB}\}$, can take arbitrary values, corresponding to different free energy approximation methods. If $\rho_{ij} = 1$ for $\forall (ij) \in \partial_{AB}$, it is the Bethe free energy; it will correspond to the TRW free energy if $\{\rho_{ij}\}$ are taken to be a set of edge appearance probabilities (which in general have values less than one). The edge appearance probabilities of A - B trees are more restrictive than for the standard trees used in TRW BP. For example, if the max part of a A - B sub-tree is a connected tree, then it can include at most one crossing edge, so in this case ρ_{AB} should satisfy $\sum_{(ij) \in \partial_{AB}} \rho_{ij} = 1$, $\rho_{ij} \geq 0$. Interestingly, we will show in Section 7 that if $\rho_{ij} \rightarrow +\infty$ for $\forall (ij) \in \partial_{AB}$, then Equation (14) is closely related to an EM algorithm.

Theorem 5. *Suppose the sum part G_A is a tree, and we approximate $\Phi_{AB}(\theta)$ using $\Phi_{tree}(\theta)$ defined in (14). Assume that (14) is globally optimized.*

- (i) *We have $\Phi_{tree}(\theta) \geq \Phi_{AB}(\theta)$. If there exists x_B^* such that $Q(x_B^*, \theta) = \Phi_{tree}(\theta)$, we have $\Phi_{tree}(\theta) = \Phi_{AB}(\theta)$, and x_B^* is a globally optimal marginal MAP solution.*
- (ii) *Suppose τ^* is a global maximum of (14), and $\{\tau_i^*(x_i) : i \in B\}$ have integral values, that is, $\tau_i^*(x_i) = 0$ or 1, then $\{x_i^* = \arg \max_{x_i} \tau_i^*(x_i) : i \in B\}$ is a globally optimal solution of the marginal MAP problem (7).*

Proof (sketch). (See appendix for the complete proof.) The fact that the sum part G_A is a tree guarantees the marginalization is exact. Showing (14) is a relaxation of the maximization problem and applying standard relaxation arguments completes the proof. $\square$

Remark. Theorem 5 works for arbitrary values of ρ_{AB} , and suggests a fundamental tradeoff of hardness as ρ_{AB} takes on different values. On the one hand, the value of ρ_{AB} controls the concavity of the objective function in (14) and hence the difficulty of finding a global optimum; small enough ρ_{AB} (as in TRW) can ensure that (14) is a convex optimization, while larger ρ_{AB} (as in Bethe or EM) causes (14) to become non-convex, making it difficult to apply Theorem 5. On the other hand, the value of ρ_{AB} also controls how likely the solution is to be integral—larger ρ_{ij} emphasizes the mutual information terms, forcing the solution towards integral points. Thus the solution of the TRW free energy is less likely to be integral than the Bethe free energy, causing a difficulty in applying Theorem 5 to TRW solutions as well. The TRW approximation ($\sum_{ij} \rho_{ij} = 1$) and EM ($\rho_{ij} \rightarrow +\infty$; see Section 7) reflect two extrema of this tradeoff between concavity and integrality, respectively, while the Bethe approximation ($\rho_{ij} = 1$) appears to represent a reasonable compromise that often gives excellent performance in practice. In Section 5.2, we give a different set of local optimality guarantees that are derived from a reparameterization perspective.

5. Message Passing Algorithms for Marginal MAP

We now derive message-passing-style algorithms to optimize the “truncated” Bethe or TRW free energies in (11) and (13). Instead of optimizing the truncated free energies directly, we leverage the results of Theorem 3 and consider their “annealed” versions,

$$\max_{\tau \in \mathbb{L}} \{ \langle \theta, \tau \rangle + \hat{H}_{A|B}(\tau) + \varepsilon \hat{H}_B(\tau) \},$$

where ε is a positive annealing coefficient (or temperature), and the $\hat{H}_{A|B}(\tau)$ and $\hat{H}_B(\tau)$ are the generic pairwise approximations of $H_{A|B}(\tau)$ and $H_B(\tau)$, respectively. That is,

$$\hat{H}_{A|B}(\tau) = \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_A \cup \partial_{AB}} \rho_{ij} I_{ij}(\tau), \quad \text{and} \quad \hat{H}_B(\tau) = \sum_{i \in B} H_i(\tau) - \sum_{(ij) \in E_B} \rho_{ij} I_{ij}(\tau), \quad (15)$$

where different values of pairwise weights $\{\rho_{ij}\}$ correspond to either the Bethe approximation or the TRW approximation. This yields a generic pairwise free energy optimization problem,

$$\max_{\tau \in \mathbb{L}} \{ \langle \theta, \tau \rangle + \sum_{i \in V} w_i H_i(\tau) - \sum_{(ij) \in E} w_{ij} I_{ij}(\tau) \}, \quad (16)$$

where the weights $\{w_i, w_{ij}\}$ are determined by the temperature ε and $\{\rho_{ij}\}$ via

$$w_i = \begin{cases} 1 & \forall i \in A \\ \varepsilon & \forall i \in B, \end{cases} \quad w_{ij} = \begin{cases} \rho_{ij} & \forall (ij) \in E_A \cup \partial_{AB} \\ \varepsilon \rho_{ij} & \forall (ij) \in E_B. \end{cases} \quad (17)$$

The general framework in (16) provides a unified treatment for approximating sum-inference, max-inference and mixed, marginal MAP problems simply by taking different weights. Specifically,

1. If $w_i = 1$ for all $i \in V$, Equation (16) corresponds to the sum-inference problem and the sum-product BP objectives and algorithms.
2. If $w_i \rightarrow 0^+$ for all $i \in V$ (and the corresponding $w_{ij} \rightarrow 0^+$), Equation (16) corresponds to the max-inference problem and the max-product linear programming objective and algorithms.
3. If $w_i = 1$ for $\forall i \in A$ and $w_i = 0$ for $\forall i \in B$ (and the corresponding $w_{ij} \rightarrow 0^+$), Equation (16) corresponds to the marginal MAP problem; in the sequel, we derive “mixed-product” BP algorithms.

Note the different roles of the singleton and pairwise weights: the singleton weights $\{w_i: i \in V\}$ define the type of inference problem, while the pairwise weights $\{w_{ij}: (ij) \in E\}$ determine the approximation method (e.g., Bethe vs. TRW).

We now derive a message passing algorithm for solving the generic problem (16), using a Lagrange multiplier method similar to Yedidia et al. (2005) or Wainwright et al. (2005a).

Proposition 6. *Assuming w_i and w_{ij} are strictly positive, the stationary points of (16) satisfy the fixed point condition of the following message passing update,*

$$\text{Message Update:} \quad m_{i \rightarrow j}(x_j) \leftarrow \left[\sum_{x_i} (\psi_i(x_i) m_{\sim i}(x_i))^{1/w_i} \left(\frac{\Psi_{ij}(x_i, x_j)}{m_{j \rightarrow i}(x_i)} \right)^{1/w_{ij}} \right]^{w_{ij}}, \quad (18)$$

Marginal Decoding:

$$\tau_i(x_i) \propto [\psi_i(x_i) m_{\sim i}(x_i)]^{1/w_i}, \quad \tau_{ij}(x_i, x_j) \propto \tau_i(x_i) \tau_j(x_j) \left[\frac{\Psi_{ij}(x_i, x_j)}{m_{i \rightarrow j}(x_j) m_{j \rightarrow i}(x_i)} \right]^{1/w_{ij}}, \quad (19)$$

Algorithm 1 Annealed BP for Marginal MAP

Define the pairwise weights $\{\rho_{ij} : (ij) \in E\}$, for example, $\rho_{ij} = 1$ for Bethe or valid appearance probabilities for TRW. Initialize the messages $\{m_{i \rightarrow j} : (ij) \in E\}$.

for iteration t **do**

1. Update ε by $\varepsilon = 1/t$, and correspondingly the weights $\{w_i, w_{ij}\}$ by (17).
2. Perform the message passing update in (18) for all edges $(ij) \in E$.

end for

Calculate the singleton beliefs $b_i(x_i)$ and decode the solution x_B^* ,

$$x_i^* = \arg \max_{x_i} b_i(x_i), \quad \forall i \in B, \text{ where } b_i(x_i) \propto \psi_i(x_i) m_{\sim i}(x_i).$$

where $m_{\sim i}(x_i) := \prod_{k \in \partial_i} m_{k \rightarrow i}(x_i)$ is the product of messages sent into node i , and ∂_i is the set of neighboring nodes of i .

Proof (sketch). (See appendix for the complete proof.) Note that (19) is simply the KKT condition of (16), with the log of the message $\log m_{i \rightarrow j}$ being the Lagrange multipliers. Plugging (19) into the local consistency constraints of $\mathbb{L}$ in (2) gives (18). $\square$

The above message update is mostly similar to TRW-BP of Wainwright et al. (2005a), except that it incorporates general singleton weights w_i . The marginal MAP problem can be solved by running (18) with $\{w_i, w_{ij}\}$ defined by (17) and a scheme for choosing the temperature ε , either directly set to be a small constant, or gradually decreased (or annealed) to zero through iterations, for example, by $\varepsilon = 1/t$ where t is the iteration. Algorithm 1 describes the details for the annealing method.

5.1 Mixed-Product Belief Propagation

Directly taking $\varepsilon \rightarrow 0^+$ in message update (18), we can get an interesting “mixed-product” BP algorithm that is a hybrid of the max-product and sum-product message updates, with a novel “argmax-product” message update that is specific to marginal MAP problems. This algorithm is listed in Algorithm 2, and described by the following proposition:

Proposition 7. *As ε approaches zero from the positive side, that is, $\varepsilon \rightarrow 0^+$, the message update (18) reduces to the update in (20)-(22) in Algorithm 2.*

Proof. For messages from $i \in A$ to $j \in A \cup B$, we have $w_i = 1$, $w_{ij} = \rho_{ij}$; the result is obvious. For messages from $i \in B$ to $j \in B$, we have $w_i = \varepsilon$, $w_{ij} = \varepsilon \rho_{ij}$. The result follows from the zero temperature limit formula in (9), by letting $f(x_i) = (\psi_i(x_i) m_{\sim i}(x_i))^{\rho_{ij}} (\frac{\psi_{ij}(x_i, x_j)}{m_{j \rightarrow i}(x_i)})$. For messages from $i \in B$ to $j \in A$, we have $w_i = \varepsilon$, $w_{ij} = \rho_{ij}$. One can show that

$$\lim_{\varepsilon \rightarrow 0^+} \left[\frac{\psi_i(x_i) m_{\sim i}(x_i)}{\max_{x_i} \psi_i(x_i) m_{\sim i}(x_i)} \right]^{1/\varepsilon} = 1(x_i \in X_i^*),$$

where $X_i^* = \arg \max_{x_i} \psi_i(x_i) m_{\sim i}(x_i)$. Plugging this into (18) and dropping the constant term, we get the message update in (22). $\square$

Algorithm 2 Mixed-product Belief Propagation for Marginal MAP

Define the pairwise weights $\{\rho_{ij} : (ij) \in E\}$ and initialize messages $\{m_{i \rightarrow j} : (ij) \in E\}$ as in Algorithm 1.

for iteration t **do**

for edge $(ij) \in E$ **do**

Perform different message updates depending on the node type of the source and destination,

$$\begin{array}{ll} A \rightarrow A \cup B: & m_{i \rightarrow j}(x_j) \leftarrow \left[\sum_{x_i} (\psi_i(x_i) m_{\sim i}(x_i)) \left(\frac{\Psi_{ij}(x_i, x_j)}{m_{j \rightarrow i}(x_i)} \right)^{1/\rho_{ij}} \right]^{\rho_{ij}}, \end{array} \quad (20)$$

(sum-product)

$$\begin{array}{ll} B \rightarrow B: & m_{i \rightarrow j}(x_j) \leftarrow \max_{x_i} (\psi_i(x_i) m_{\sim i}(x_i))^{\rho_{ij}} \left(\frac{\Psi_{ij}(x_i, x_j)}{m_{j \rightarrow i}(x_i)} \right), \end{array} \quad (21)$$

(max-product)

$$\begin{array}{ll} B \rightarrow A: & m_{i \rightarrow j}(x_j) \leftarrow \left[\sum_{x_i \in \mathcal{X}_i^*} (\psi_i(x_i) m_{\sim i}(x_i)) \left(\frac{\Psi_{ij}(x_i, x_j)}{m_{j \rightarrow i}(x_i)} \right)^{1/\rho_{ij}} \right]^{\rho_{ij}}, \end{array} \quad (22)$$

(argmax-product)

where the set $\mathcal{X}_i^* = \arg \max_{x_i} \psi_i(x_i) m_{\sim i}(x_i)$ and $m_{\sim i}(x_i) = \prod_{k \in \partial_i} m_{ki}(x_i)$.

end for

end for

Calculate the singleton beliefs $b_i(x_i)$ and decode the solution x_B^* ,

$$x_i^* = \arg \max_{x_i} b_i(x_i), \quad \forall i \in B, \text{ where } b_i(x_i) \propto \psi_i(x_i) m_{\sim i}(x_i).$$

Algorithm 2 has an intuitive interpretation: the sum-product and max-product messages in (20) and (21) correspond to the marginalization and maximization steps, respectively. The special “argmax-product” messages in (22) serves to synchronize the sum-product and max-product messages—it restricts the max nodes to the currently decoded local marginal MAP solutions $\mathcal{X}_i^* = \arg \max_{x_i} \psi_i(x_i) m_{\sim i}(x_i)$, and passes the posterior beliefs back to the sum part. Note that the summation notation in (22) can be ignored if $\mathcal{X}_i^*$ has only a single optimal state.

One critical feature of our mixed-product BP is that it takes simultaneous movements on the marginalization and maximization sub-problems in a parallel fashion, and is computationally much more efficient than the traditional methods that require fully solving a marginalization sub-problem before taking each maximization step. This advantage is inherited from our general variational framework, which naturally integrates the marginalization and maximization sub-problems into a joint optimization problem.

Interestingly, Algorithm 2 also bears similarity to a recent hybrid message passing method of Jiang et al. (2011), which differs from Algorithm 2 only in replacing the special argmax-product messages (22) with regular max-product messages. We make a detailed comparison of these two algorithms in Section 5.3, and show that it is in fact the argmax-product messages (22) that lends our algorithm several appealing optimality guarantees.

5.2 Reparameterization Interpretation and Local Optimality Guarantees

An important interpretation of the sum-product and max-product BP is the reparameterization viewpoint (Wainwright et al., 2003; Weiss et al., 2007): Message passing updates can be viewed as moving probability mass between local pseudo-marginals (or beliefs), in a way that leaves their product a reparameterization of the original distribution, while ensuring some consistency conditions at the fixed points. Such viewpoints are theoretically important, because they are useful for proving optimality guarantees for the BP algorithms. In this section, we show that the mixed-product BP in Algorithm 2 has a similar reparameterization interpretation, based on which we establish a local optimality guarantee for mixed-product BP.

To start, we define a set of “mixed-beliefs” as

$$b_i(x_i) \propto \Psi_i(x_i) m_{\sim i}(x_i), \quad b_{ij}(x_{ij}) \propto b_i(x_i) b_j(x_j) \left[\frac{\Psi_{ij}(x_i, x_j)}{m_{i \rightarrow j}(x_j) m_{j \rightarrow i}(x_i)} \right]^{1/\rho_{ij}}. \quad (23)$$

The marginal MAP solution should be decoded from $x_i^* \in \arg \max_{x_i} b_i(x_i), \forall i \in B$, as is typical in max-product BP. Note that the above mixed-beliefs $\{b_i, b_{ij}\}$ are different from the local marginals $\{\tau_i, \tau_{ij}\}$ defined in (19), but are rather softened versions of $\{\tau_i, \tau_{ij}\}$. Their relationship is explicitly clarified in the following.

Proposition 8. *The $\{\tau_i, \tau_{ij}\}$ in (19) and the $\{b_i, b_{ij}\}$ in (23) are associated via,*

$$\begin{cases} b_i \propto \tau_i & \forall i \in A, \\ b_i \propto (\tau_i)^\varepsilon & \forall i \in B \end{cases} \quad \begin{cases} b_{ij} \propto b_i b_j (\frac{\tau_{ij}}{\tau_i \tau_j}) & \forall (ij) \in E_A \cup \partial_{AB} \\ b_{ij} \propto b_i b_j (\frac{\tau_{ij}}{\tau_i \tau_j})^\varepsilon & \forall (ij) \in E_B. \end{cases}$$

Proof. Result follows from the simple algebraic transformation between (19) and (23). $\square$

Therefore, as $\varepsilon \rightarrow 0^+$, the $\tau_i (= b_i^{1/\varepsilon})$ for $i \in B$ should concentrate their mass on a deterministic configuration, but b_i may continue to have soft values.

We now show that the mixed-beliefs $\{b_i, b_{ij}\}$ have a reparameterization interpretation.

Theorem 9. *At the fixed point of mixed-product BP in Algorithm 2, the mixed-beliefs defined in (23) satisfy*

Reparameterization:

$$p(x) \propto \prod_{i \in V} b_i(x_i) \prod_{(ij) \in E} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}}. \quad (24)$$

Mixed-consistency:

$$(a) \quad \sum_{x_i} b_{ij}(x_i, x_j) = b_j(x_j), \quad \forall i \in A, j \in A \cup B, \quad (25)$$

$$(b) \quad \max_{x_i} b_{ij}(x_i, x_j) = b_j(x_j), \quad \forall i \in B, j \in B, \quad (26)$$

$$(c) \quad \sum_{x_i \in \arg \max b_i} b_{ij}(x_i, x_j) = b_j(x_j), \quad \forall i \in B, j \in A. \quad (27)$$

Proof. Directly substitute the definition (23) into the message update (20)-(22). $\square$

The three mixed-consistency constraints exactly map to the three types of message updates in Algorithm 2. Constraint (a) and (b) enforces the regular sum- and max- consistency of the sum- and max- product messages in (20) and (21), respectively. Constraint (c) corresponds to the argmax-product message update in (22): it enforces the marginals to be consistent after x_i is assigned to the currently decoded solution, $x_i = \arg \max_{x_i} b_i(x_i) = \arg \max_{x_i} \sum_{x_j} b_{ij}(x_i, x_j)$, corresponding to solving a local marginal MAP problem on $b_{ij}(x_i, x_j)$. It turns out that this special constraint is a crucial ingredient of mixed-product BP, enabling us to prove guarantees on the strong local optimality of the solution.

Some notation is required. Suppose C is a subset of max nodes in B . Let $G_{C \cup A} = (C \cup A, E_{C \cup A})$ be the subgraph of G induced by nodes $C \cup A$, where $E_{C \cup A} = \{(ij) \in E : i, j \in C \cup A\}$. We call $G_{C \cup A}$ a semi- A - B subtree of G if the edges in $E_{C \cup A} \setminus E_B$ form an A - B tree. In other words, $G_{C \cup A}$ is a semi- A - B tree if it is an A - B tree when ignoring any edges entirely within the max set B . See Figure 2 for examples of semi A - B trees.

Following Weiss et al. (2007), we say that a set of weights $\{\rho_{ij}\}$ is *provably convex* if there exist positive constants κ_i and $\kappa_{i \rightarrow j}$, such that $\kappa_i + \sum_{i' \in \partial_i} \kappa_{i' \rightarrow i} = 1$ and $\kappa_{i \rightarrow j} + \kappa_{j \rightarrow i} = \rho_{ij}$. Weiss et al. (2007) shows that if $\{\rho_{ij}\}$ is provably convex, then $H(\tau) = \sum_i H_i(\tau) - \sum_{ij} \rho_{ij} I_{ij}(\tau)$ is a concave function of τ in the locally consistent polytope $\mathbb{L}$.

Theorem 10. *Suppose C is a subset of B such that $G_{C \cup A}$ is a semi- A - B tree, and the weights $\{\rho_{ij}\}$ satisfy*

1. $\rho_{ij} = 1$ for $(ij) \in E_A$;
2. $0 \leq \rho_{ij} \leq 1$ for $(ij) \in E_{C \cup A} \cap \partial_{AB}$;
3. $\{\rho_{ij} : (ij) \in E_{C \cup A} \cap E_B\}$ is provably convex.

At the fixed point of mixed-product BP in Algorithm 2, if the mixed-beliefs on the max nodes $\{b_i, b_{ij} : i, j \in B\}$ defined in (23) all have unique maxima, then there exists a B -configuration x_B^ satisfying $x_i^* = \arg \max b_i$ for $\forall i \in B$ and $(x_i^*, x_j^*) = \arg \max b_{ij}$ for $\forall (ij) \in E_B$, and x_B^* is locally optimal in the sense that $Q(x_B^*; \theta)$ is not smaller than any B -configuration that differs from x_B^* only on C , that is, $Q(x_B^*; \theta) = \max_{x_C} Q([x_C, x_{B \setminus C}^*]; \theta)$.*

Proof (sketch). (See appendix for the complete proof.) The mixed-consistency constraint (c) in (27) and the fact that $G_{C \cup A}$ is a semi- A - B tree enables the summation part to be eliminated away. The remaining part only involves the max nodes, and the method in Weiss et al. (2007) for analyzing standard MAP can be applied. $\square$

Remark. The proof of Theorem 10 relies on transforming the marginal MAP problem to a standard MAP problem by eliminating the summation part. Therefore, variants of Theorem 10 may be derived using other global optimality conditions of convexified belief propagation or linear programming algorithms for MAP, such as those in Werner (2007, 2010); Wainwright et al. (2005b). We leave this to future work.

For $G_{C \cup A}$ to be a semi A - B tree, the sum part G_A must be a tree, which Theorem 10 assumes implicitly. For the hidden Markov chain in Figure 1, Theorem 10 implies only the local optimality up to Hamming distance one (or coordinate-wise optimality), because any semi A - B subtree of G in Figure 1 can contain at most one max node. However, Theorem 10 is in general much stronger, especially when the sum part is not fully connected, or when the max part has interior regions disconnected from the sum part. As examples, see Figure 2(b)-(c).

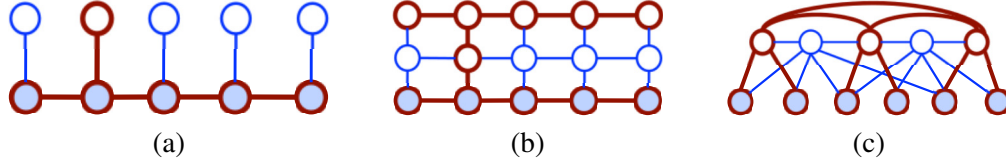


Figure 2: Examples of semi A - B trees. The shaded nodes represent sum nodes, while the unshaded are max nodes. In each graph, a semi A - B tree is labeled by red bold lines. Under the conditions of Theorem 10, the fixed point of mixed-product BP is locally optimal up to jointly perturbing all the max nodes in any semi- A - B subtree of G .

5.3 The Importance of the Argmax-product Message Updates

Jiang et al. (2011) proposed a similar hybrid message passing algorithm, repeated here as Algorithm 3, which differs from our mixed-product BP only in replacing our argmax-product message update (22) with the usual max-product message update (21). We show in this section that this very difference gives Algorithm 3 very different properties, and fewer optimality guarantees, than our mixed-product BP.

Algorithm 3 Hybrid Message Passing by Jiang et al. (2011)

1. Message Update:

$$\begin{aligned}
 A \rightarrow A \cup B: & \quad m_{i \rightarrow j}(x_j) \leftarrow \left[\sum_{x_i} (\psi_i(x_i) m_{\sim i}(x_i)) \left(\frac{\Psi_{ij}(x_i, x_j)}{m_{j \rightarrow i}(x_i)} \right)^{1/\rho_{ij}} \right]^{\rho_{ij}}, \\
 (\text{sum-product}) & \\
 A \rightarrow A \cup B: & \quad m_{i \rightarrow j}(x_j) \leftarrow \max_{x_i} (\psi_i(x_i) m_{\sim i}(x_i))^{\rho_{ij}} \left(\frac{\Psi_{ij}(x_i, x_j)}{m_{j \rightarrow i}(x_i)} \right). \\
 (\text{max-product}) &
 \end{aligned}$$

2. Decoding: $x_i^* = \arg \max_{x_i} b_i(x_i)$ for $\forall i \in B$, where $b_i(x_i) \propto \psi_i(x_i) m_{\sim i}(x_i)$.

Similar to our mixed-product BP, Algorithm 3 also satisfies the reparameterization property in (24) (with beliefs $\{b_i, b_{ij}\}$ defined by (23)); it also satisfies a set of similar, but crucially different, consistency conditions at its fixed points,

$$\begin{aligned}
 \sum_{x_i} b_{ij}(x_i, x_j) &= b_j(x_j), & \forall i \in A, j \in A \cup B, \\
 \max_{x_i} b_{ij}(x_i, x_j) &= b_j(x_j), & \forall i \in B, j \in A \cup B,
 \end{aligned}$$

which exactly map to the max- and sum- product message updates in Algorithm 3.

Despite its striking similarity, Algorithm 3 has very different properties, and does not share the appealing variational interpretation and optimality guarantees that we have demonstrated for mixed-product BP. First, it is unclear whether Algorithm 3 can be interpreted as a fixed point algorithm for maximizing our, or a similar, variational objective function. Second, it does not inherit the same optimality guarantees in Theorem 10, despite its similar reparameterization and consistency conditions. These disadvantages are caused by the miss of the special argmax-product message update and its associated mixed-consistency condition in (27), which was a critical ingredient of the proof of Theorem 10.

More detailed insights into Algorithm 3 and mixed-product BP can be obtained by considering the special case when the full graph G is an undirected tree. We show that in this case, Algorithm 3 can be viewed as optimizing a set of *approximate* objective functions, obtained by rearranging the max and sum operators into orders that require less computational cost, while mixed-product BP attempts to maximize the *exact* objective function by message updates that effectively perform some “asynchronous” coordinate descent steps. In the sequel, we use an illustrative toy example to explain the main ideas.

Example 2. Consider a marginal MAP problem on a four node chain-structured graphical model $x_3 - x_1 - x_2 - x_4$, where the sum and max sets are $A = \{1, 2\}$ and $B = \{3, 4\}$, respectively. We analyze how Algorithm 3 and mixed-product BP in Algorithm 2 perform on this toy example, when both taking Bethe weights ($\rho_{ij} = 1$ for $(ij) \in E$).

Algorithm 3 (Jiang et al. 2011). Since G is a tree, one can show that Algorithm 3 (with Bethe weights) terminates after a full forward and backward iteration (e.g., messages passed along $x_3 \rightarrow x_1 \rightarrow x_2 \rightarrow x_4$ and then $x_4 \rightarrow x_2 \rightarrow x_1 \rightarrow x_3$). By tracking the messages, one can write its final decoded solution in a closed form,

$$x_3^* = \arg \max_{x_3} \sum_{x_1} \sum_{x_2} \max_{x_4} [\exp(\theta(x))], \quad x_4^* = \arg \max_{x_4} \sum_{x_2} \sum_{x_1} \max_{x_3} [\exp(\theta(x))],$$

On the other hand, the true marginal MAP solution is given by,

$$x_3^* = \arg \max_{x_3} \max_{x_4} \sum_{x_1} \sum_{x_2} [\exp(\theta(x))], \quad x_4^* = \arg \max_{x_4} \max_{x_3} \sum_{x_2} \sum_{x_1} [\exp(\theta(x))].$$

Here, Algorithm 3 approximates the exact marginal MAP problem by rearranging the max and sum operators into an elimination order that makes the calculation easier. A similar property holds for the general case when G is undirected tree: Algorithm 3 (with Bethe weights) terminates in a finite number of steps, and its output solution x_i^* effectively maximizes an approximate objective function obtained by reordering the max and sum operators along a tree-order (see Definition 4.1) that is rooted at node i . The performance of the algorithm should be related to the error caused by exchanging the order of max and sum operators. However, exact optimality guarantees are likely difficult to show because it maximizes an inexact objective function. In addition, since each component x_i^* uses a different order of arrangement, and hence maximizes a different surrogate objective function, it is unclear whether the joint B -configuration $x_B^* = \{x_i^* : i \in B\}$ given by Algorithm 3 maximizes a single consistent objective function.

Algorithm 2 (mixed-product). On the other hand, the mixed-product belief propagation in Algorithm 2 may not terminate in a finite number of steps, nor does it necessarily yield a closed form solution when G is an undirected tree. However, Algorithm 2 proceeds in an attempt to optimize the exact objective function. In this toy example, we can show that the true solution is guaranteed to be a fixed point of Algorithm 2. Let $b_3(x_3)$ be the mixed-belief on x_3 at the current iteration, and $x_3^* = \arg \max_{x_3} b_3(x_3)$ its unique maxima. After a message sequence passed from x_3 to x_4 , one can show that $b_4(x_4)$ and x_4^* update to

$$x_4^* = \arg \max_{x_4} b_4(x_4), \quad b_4(x_4) = \sum_{x_2} \sum_{x_1} \exp(\theta([x_3^*, x_{-3}])) = \exp(Q([x_3^*, x_4]; \theta)),$$

where we maximize the exact objective function $Q([x_3, x_4]; \theta)$ with fixed $x_3 = x_3^*$. Therefore, on this toy example, one sweep ($x_3 \rightarrow x_4$ or $x_4 \rightarrow x_3$) of Algorithm 2 is effectively performing a coordinate

Algorithm 4 Proximal Point Algorithm for Marginal MAP (Exact)

 Initialize local marginals τ^0 .

for iteration t **do**

$$\theta^{t+1} = \theta + \lambda^t \log \tau_B^t, \quad (28)$$

$$\tau^{t+1} = \arg \max_{\tau \in \mathbb{M}} \{ \langle \tau, \theta^{t+1} \rangle + H_{A|B}(\tau) + \lambda^t H_B(\tau) \}, \quad (29)$$

end for

 Decoding: $x_i^* = \arg \max_{x_i} \tau_i(x_i)$ for $\forall i \in B$.

descent step, which monotonically improves the true objective function towards a local maximum. In more general models, Algorithm 2 differs from sequential coordinate descent, and does not guarantee monotonic convergence. But, it can be viewed as a “parallel” version of coordinate descent, which ensures the stronger local optimality guarantees shown in Theorem 10.

6. Convergent Algorithms by Proximal Point Methods

An obvious disadvantage of mixed-product BP is its lack of convergence guarantees, even when G is an undirected tree. In this section, we apply a proximal point approach (e.g., Martinet, 1970; Rockafellar, 1976) to derive convergent algorithms that directly optimize our free energy objectives, which take the form of transforming marginal MAP into a sequence of pure (or annealed) sum-inference tasks. Similar methods have been applied to standard sum-inference (Yuille, 2002) and max-inference (Ravikumar et al., 2010).

For the purpose of illustration, we first consider the problem of maximizing the *exact* marginal MAP free energy, $F_{\text{mix}}(\tau, \theta) = \langle \tau, \theta \rangle + H_{A|B}(\tau)$. The proximal point algorithm works by iteratively optimizing a smoothed problem,

$$\tau^{t+1} = \arg \min_{\tau \in \mathbb{M}} \{ -F_{\text{mix}}(\tau, \theta) + \lambda^t D(\tau || \tau^t) \},$$

where τ^t is the solution at iteration t , and λ^t is a positive coefficient. Here, $D(\cdot || \cdot)$ is a distance, called the proximal function, which forces τ^{t+1} to be close to τ^t ; typical choices of $D(\cdot || \cdot)$ are Euclidean or Bregman distances or ψ -divergences (e.g., Teboulle, 1992; Iusem and Teboulle, 1993). Proximal algorithms have nice convergence guarantees: the objective series $\{f(\tau^t)\}$ is guaranteed to be non-increasing at each iteration, and $\{\tau^t\}$ converges to an optimal solution, under some regularity conditions. See, for example, Rockafellar (1976); Tseng and Bertsekas (1993); Iusem and Teboulle (1993). The proximal algorithm is closely related to the majorize-minimize (MM) algorithm (Hunter and Lange, 2004) and the convex-concave procedure (Yuille, 2002).

For our purpose, we take $D(\cdot || \cdot)$ to be a KL divergence between distributions on the max nodes,

$$D(\tau || \tau^t) = \text{KL}(\tau_B(x_B) || \tau_B^t(x_B)) = \sum_{x_B} \tau_B(x_B) \log \frac{\tau_B(x_B)}{\tau_B^t(x_B)}.$$

In this case, the proximal point algorithm reduces to Algorithm 4, which iteratively solves a smoothed free energy objective, with natural parameter θ^t updated at each iteration. Intuitively, the proximal inner loop (28)-(29) essentially “adds back” the truncated entropy term $H_B(\tau)$, while

canceling its effect by adjusting θ in the opposite direction. Typical choices of λ^t include $\lambda^t = 1$ (constant) and $\lambda^t = 1/t$ (harmonic). Note that the proximal approach is distinct from an annealing method, which would require that the annealing coefficient vanish to zero. Interestingly, if we take $\lambda^t = 1$, then the inner maximization problem (29) reduces to the standard log-partition function duality (1), corresponding to a pure marginalization task. This has the interpretation of transforming the marginal MAP problem into a sequence of standard sum-inference problems.

In practice we approximate $H_{A|B}(\tau)$ and $H_B(\tau)$ by pairwise entropy decomposition $\hat{H}_{A|B}(\tau)$ and $\hat{H}_B(\tau)$ in (15), respectively. If $\hat{H}_B(\tau)$ is provably convex in the sense of Weiss et al. (2007), that is, there exist positive constants $\{\kappa_i, \kappa_{i \rightarrow j}\}$ satisfying $\rho_i = \kappa_i + \sum_{k \in \partial_i} \kappa_{k \rightarrow i}$ and $\rho_{ij} = \kappa_{i \rightarrow j} + \kappa_{j \rightarrow i}$ for $i, j \in B$. Then the resulting approximate algorithm can be interpreted as a proximal algorithm that maximizes $\hat{F}_{\text{mix}}(\tau, \theta)$ with proximal function as

$$D_{\text{pair}}(\tau || \tau^t) = \sum_{i \in B} \kappa_i \text{KL}[\tau_i(x_i) || \tau_i^0(x_i)] + \sum_{(ij) \in E_B} \kappa_{i \rightarrow j} \text{KL}[(\tau_{ij}(x_i | x_j) || \tau_{ij}^0(x_i | x_j))].$$

In this case, Algorithm 4 is still a valid proximal algorithm and inherits its convergence guarantees. In practice one uses approximations that are not provably convex. An interesting special case is when both $H_{A|B}(\tau)$ and $H_B(\tau)$ are approximated by a Bethe approximation. This has the effect that the optimization (29) can be solved using standard belief propagation. Although the Bethe form for $H_{A|B}(\tau)$ and $H_B(\tau)$ is provably convex only in some special cases, such as when G is tree structured, we find in practice that this approximation gives very accurate solutions, even on general loopy graphs where its convergence is no longer theoretically guaranteed.

The global convergence guarantees of the proximal point algorithm may also fail if the inner update (29) is not solved exactly. It should also be possible to develop globally convergent algorithms without inner loops using the techniques that have been developed for full marginalization or MAP problems (e.g., Meltzer et al., 2009; Hazan and Shashua, 2010; Jojic et al., 2010; Savchynskyy et al., 2010), but we leave this to future work.

7. Connections to EM

A natural algorithm for solving the marginal MAP problem is to use the expectation-maximization (EM) algorithm, by treating x_A as the hidden variables and x_B as the “parameters” to be maximized. In this section, we show that the EM algorithm can be seen as a coordinate ascent algorithm on a mean field variant of our framework.

We start by introducing a “non-convex” generalization of Theorem 2.

Corollary 11. *Let $\mathbb{M}^o$ be the subset of the marginal polytope $\mathbb{M}$ corresponding to the distributions in which x_B are clamped to some deterministic values, that is,*

$$\mathbb{M}^o = \{\tau \in \mathbb{M} : \exists x_B^* \in X_B, \text{ such that } \tau(x_B) = 1(x_B = x_B^*)\}.$$

Then the dual optimization (8) remains exact if the marginal polytope $\mathbb{M}$ is replaced by any $\mathbb{N}$ satisfying $\mathbb{M}^o \subseteq \mathbb{N} \subseteq \mathbb{M}$, that is,

$$\Phi_{AB} = \max_{\tau \in \mathbb{N}} \{\langle \theta, \tau \rangle + H_{A|B}(\tau)\}.$$

Proof. For an arbitrary marginal MAP solution x_B^* , the τ^* with $\tau^*(x) = p(x | x_B = x_B^*; \theta)$ is an optimum of (8) and satisfies $\tau^* \in \mathbb{M}^o$. Therefore, restricting the optimization on $\mathbb{M}^o$ (or any $\mathbb{N}$) does not change the maximum value of the objective function. $\square$

Remark. Among all $\mathbb{N}$ satisfying $\mathbb{M}^o \subseteq \mathbb{N} \subseteq \mathbb{M}$, the marginal polytope $\mathbb{M}$ is the smallest (and the unique) convex set that includes $\mathbb{M}^o$, that is, it is the convex hull of $\mathbb{M}^o$.

To connect to EM, we define $\mathbb{M}^\times$, the set of distributions in which x_A and x_B are independent, that is, $\mathbb{M}^\times = \{\tau \in \mathbb{M} : \tau(x) = \tau(x_A)\tau(x_B)\}$. Since $\mathbb{M}^o \subset \mathbb{M}^\times \subset \mathbb{M}$, the dual optimization (8) remains exact when restricted to $\mathbb{M}^\times$, that is,

$$\Phi_{AB}(\theta) = \max_{\tau \in \mathbb{M}^\times} \{\langle \theta, \tau \rangle + H_{A|B}(\tau)\} = \max_{\tau \in \mathbb{M}^\times} \{\langle \theta, \tau \rangle + H_A(\tau)\},$$

where the second equality holds because $H_{A|B}(\tau) = H_A(\tau)$ for $\tau \in \mathbb{M}^\times$.

Although $\mathbb{M}^\times$ is no longer a convex set, it is natural to consider a coordinate update that alternately optimizes $\tau(x_A)$ and $\tau(x_B)$,

$$\begin{aligned} \text{Updating sum part : } \quad \tau_A^{t+1} &\leftarrow \operatorname{argmax}_{\tau_A \in \mathbb{M}_A} \{\langle \mathbb{E}_{\tau_B^t}(\theta), \tau_A \rangle + H_A(\tau_A)\}, \\ \text{Updating max part : } \quad \tau_B^{t+1} &\leftarrow \operatorname{argmax}_{\tau_B \in \mathbb{M}_B} \langle \mathbb{E}_{\tau_A^{t+1}}(\theta), \tau_B \rangle, \end{aligned} \tag{30}$$

where $\mathbb{M}_A$ and $\mathbb{M}_B$ are the marginal polytopes over x_A and x_B , respectively. Note that the sum and max step each happen to be the dual of a sum-inference and max-inference problem, respectively. If we go back to the primal, and update the primal configuration x_B instead of τ_B , (30) can be rewritten into

$$\begin{aligned} \text{E step : } \quad \tau_A^{t+1}(x_A) &\leftarrow p(x_A | x_B^t; \theta), \\ \text{M step : } \quad x_B^{t+1} &\leftarrow \operatorname{argmax}_{x_B} \mathbb{E}_{\tau_A^{t+1}}(\theta), \end{aligned}$$

which is exactly the EM update, viewing x_B as parameters and x_A as hidden variables. Similar connections between EM and the coordinate ascent method on variational objectives has been discussed in Neal and Hinton (1998) and Wainwright and Jordan (2008).

When the E-step or M-step are intractable, one can insert various approximations. In particular, approximating $\mathbb{M}_A$ by a mean-field inner bound $\mathbb{M}_A^{mf}$ leads to variational EM. An interesting observation is obtained by using a Bethe approximation (3) to solve the E-step and a linear relaxation to solve the M-step; in this case, the EM-like update is equivalent to solving

$$\max_{\tau \in \mathbb{L}^\times} \left\{ \langle \theta, \tau \rangle + \sum_{i \in A} H_i(\tau) - \sum_{(ij) \in E_A} I_{ij}(\tau) \right\}, \tag{31}$$

where $\mathbb{L}^\times$ is the subset of $\mathbb{L}$ in which $\tau_{ij}(x_i, x_j) = \tau_i(x_i)\tau_j(x_j)$ for $(ij) \in \partial_{AB}$. Equivalently, $\mathbb{L}^\times$ is the subset of $\mathbb{L}$ in which $I_{ij}(\tau) = 0$ for $(ij) \in \partial_{AB}$. Therefore, (31) can be treated as a special case of (14) by taking $\rho_{ij} \rightarrow +\infty$, forcing the solution τ^* to fall into $\mathbb{L}^\times$. As we discussed in Section 4.3, EM represents an extreme of the tradeoff between convexity and integrality implied by Theorem 5, which strongly encourages vertex solutions by sacrificing convexity, and hence is likely to become stuck in local optima.

8. Junction Graph Belief Propagation for Marginal MAP

In the above, we have restricted the discussion to pairwise models and pairwise entropy approximations, mainly for the purpose of clarity. In this section, we extend our algorithms to leverage

higher order cliques, based on the junction graph representation (Mateescu et al., 2010; Koller and Friedman, 2009). Other higher order methods, like generalized BP (Yedidia et al., 2005) or their convex variants (Wainwright et al., 2005a; Wiegnerinck, 2005), can be derived similarly.

For notation, a cluster graph is a graph of subsets of variables (called clusters). Formally, it is a triple $(\mathcal{G}, \mathcal{C}, \mathcal{S})$, where $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ is an undirected graph, with each node $k \in \mathcal{V}$ associated with a cluster $c_k \in \mathcal{C}$, and each edge $(kl) \in \mathcal{E}$ with a subset $s_{kl} \in \mathcal{S}$ (called separators) satisfying $s_{kl} \subseteq c_k \cap c_l$. We assume that $\mathcal{C}$ subsumes the index set I , that is, for any $\alpha \in I$, we can assign it with a $c_k \in \mathcal{C}$, denoted $c[\alpha]$, such that $\alpha \subseteq c_k$. In this case, we can reparameterize $\theta = \{\theta_\alpha : \alpha \in I\}$ into $\theta = \{\theta_{c_k} : k \in \mathcal{V}\}$ by taking $\theta_{c_k} = \sum_{\alpha: c[\alpha]=c_k} \theta_\alpha$, without changing the distribution. Therefore, we simply

assume $\mathcal{C} = I$ in this paper without loss of generality. A cluster graph is called a *junction graph* if it satisfies the *running intersection property*—for each $i \in V$, the induced sub-graph consisting of the clusters and separators that include i is a connected tree. A junction graph is a junction tree if $\mathcal{G}$ is a tree.

To approximate the variational dual form, we first replace $\mathbb{M}$ with a higher order locally consistent polytope $\mathbb{L}(\mathcal{G})$, which is the set of local marginals $\tau = \{\tau_{c_k}, \tau_{s_{kl}} : k \in \mathcal{V}, (kl) \in \mathcal{E}\}$ that are consistent on the intersections of the clusters and separators, that is,

$$\mathbb{L}(\mathcal{G}) = \{\tau : \sum_{x_{c_k \setminus s_{kl}}} \tau_{c_k}(x_{c_k}) = \tau_{s_{kl}}(x_{s_{kl}}), \tau_{c_k}(x_{c_k}) \geq 0, \text{ for } \forall k \in \mathcal{V}, (kl) \in \mathcal{E}\}.$$

Clearly, we have $\mathbb{M} \subseteq \mathbb{L}(\mathcal{G})$ and that $\mathbb{L}(\mathcal{G})$ is tighter than the pairwise polytope $\mathbb{L}$ we used previously.

We then approximate the joint entropy term by a linear combination of the entropies over the clusters and separators,

$$H(\tau) \approx \sum_{k \in \mathcal{V}} H_{c_k}(\tau) - \sum_{(kl) \in \mathcal{E}} H_{s_{kl}}(\tau),$$

where $H_{c_k}(\tau)$ and $H_{s_{kl}}(\tau)$ are the entropy of the local marginals τ_{c_k} and $\tau_{s_{kl}}$, respectively. Further, we approximate $H_B(\tau)$ by a slightly more restrictive entropy decomposition,

$$H_B(\tau) \approx \sum_{k \in \mathcal{V}} H_{\pi_k}(\tau),$$

where $\{\pi_k : k \in \mathcal{V}\}$ is a non-overlapping partition of the max nodes B satisfying $\pi_k \subseteq c_k$ for $\forall k \in \mathcal{V}$. In other words, π represents an assignment of each max node $x_b \in B$ into a cluster k with $x_b \in \pi_k$. Let $\mathcal{B}$ be the set of clusters $k \in \mathcal{V}$ for which $\pi_k \neq \emptyset$, and call $\mathcal{B}$ the *max-clusters*; correspondingly, call $\mathcal{A} = \mathcal{V} \setminus \mathcal{B}$ the *sum-clusters*. See Figure 3 for an example.

Overall, the marginal MAP dual form in (8) is approximated by

$$\max_{\tau \in \mathbb{L}(\mathcal{G})} \left\{ \langle \theta, \tau \rangle + \sum_{k \in \mathcal{A}} H_{c_k}(\tau) + \sum_{k \in \mathcal{B}} H_{c_k|\pi_k}(\tau) - \sum_{(kl) \in \mathcal{E}} H_{s_{kl}}(\tau) \right\} \quad (32)$$

where $H_{c_k|\pi_k}(\tau) = H_{c_k}(\tau) - H_{\pi_k}(\tau)$. Optimizing (32) using a method similar to the derivation of mixed-product BP in Algorithm 2, we obtain a “mixed-product” junction graph belief propagation, given in Algorithm 5.

Similarly to our mixed-product BP in Algorithm 2, Algorithm 5 also admits an intuitive reparameterization interpretation and a strong local optimality guarantee. Algorithm 5 can be seen as

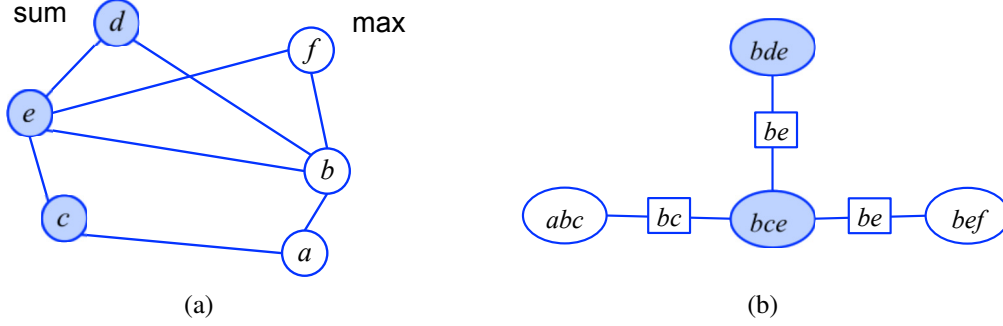


Figure 3: (a) An example of marginal MAP problem, where d, c, e are sum nodes (shaded) and a, b, f are max nodes. (b) A junction graph of (a). Selecting a partitioning of max nodes, $\pi_{bde} = \pi_{bef} = \emptyset$, $\pi_{abc} = \{a, b\}$, and $\pi_{bef} = \{f\}$, results in $\{bde\}, \{bce\}$ being sum clusters (shaded) and $\{abc\}, \{bef\}$ being max clusters.

Algorithm 5 Mixed-product Junction Graph BP

1. Passing messages between clusters on the junction graph until convergence:

$$\begin{aligned}
 \mathcal{A} &\rightarrow \mathcal{A} \cup \mathcal{B}: & m_{k \rightarrow l}(x_{s_{kl}}) &\propto \sum_{x_{c_k} \setminus s_{kl}} \psi_{c_k}(x_{c_k}) m_{\sim k \setminus l}(x_{c_k}), \\
 (\text{sum-product}) & & & \\
 \mathcal{B} &\rightarrow \mathcal{A} \cup \mathcal{B}: & m_{k \rightarrow l}(x_{s_{kl}}) &\propto \sum_{x_{c_k} \setminus s_{kl}} (\psi_{c_k}(x_{c_k}) m_{\sim k \setminus l}(x_{c_k})) \cdot 1[x_{\pi_k} \in \mathcal{X}_{\pi_k}^*], \\
 (\text{argmax-product}) & & & \\
 \text{where } \mathcal{X}_{\pi_k}^* &= \arg \max_{x_{\pi_k}} \sum_{x_{c_k} \setminus \pi_k} b_k(x_{c_k}), \\
 b_k(x_{c_k}) &= \psi_{c_k}(x_{c_k}) \prod_{k' \in \mathcal{N}(k)} m_{k' \rightarrow k}(x_{s_{k'l_k}}) \quad \text{and} \quad m_{\sim k \setminus l}(x_{c_k}) = \prod_{k' \in \mathcal{N}(k) \setminus \{l\}} m_{k' \rightarrow k}(x_{s_{k'l_k}}).
 \end{aligned}$$

2. Decoding: $x_{\pi_k}^* = \arg \max_{x_{\pi_k}} \sum_{x_{c_k} \setminus \pi_k} b_k(x_{c_k})$ for $\forall k \in \mathcal{B}$.
-

a special case of a more general junction graph BP algorithm derived in Liu and Ihler (2012) for solving maximum expected utility tasks in decision networks. For more details, we refer the reader to that work.

9. Experiments

We illustrate our algorithms on both simulated models and more realistic diagnostic Bayesian networks taken from the UAI08 inference challenge. We show that our Bethe approximation algorithms perform best among all the tested algorithms, including Jiang et al. (2011)’s hybrid message passing and a state-of-the-art local search algorithm (Park and Darwiche, 2004).

We implement our mixed-product BP in Algorithm 2 with Bethe weights (mix-product (Bethe)), the regular sum-product BP (sum-product), max-product BP (max-product) and Jiang et al. (2011)’s hybrid message passing (with Bethe weights) in Algorithm 3 (Jiang’s method),

where the solutions are all extracted by maximizing the singleton marginals of the max nodes. For all these algorithms, we run a maximum of 50 iterations; in case they fail to converge, we run 100 additional iterations with a damping coefficient of 0.1. We initialize all these algorithms with 5 random initializations and pick the best solution; for `mix-product (Bethe)` and Jiang’s method, we run an additional trial initialized using the sum-product messages, which was reported to perform well in Park and Darwiche (2004) and Jiang et al. (2011). We also run the proximal point version of mixed-product BP with Bethe weights (`Proximal (Bethe)`), which is Algorithm 4 with both $H_{A|B}(\tau)$ and $H_B(\tau)$ approximated by Bethe approximations.

We also implement the TRW approximation, but only using the convergent proximal point algorithm, because the TRW upper bounds are valid only when the algorithms converge. The TRW weights of $\hat{H}_{A|B}$ are constructed by first (randomly) selecting spanning trees of G_A , and then augmenting each spanning tree with one uniformly selected edge in ∂_{AB} ; the TRW weights of $\hat{H}_B(\tau)$ are constructed to be provably convex, using the method of TRW-S in Kolmogorov (2006). We run all the proximal point algorithms for a maximum of 100 iterations, with a maximum of 5 iterations of weighted message passing updates (18)-(19) for the inner loops (with 5 additional damping with 0.1 damping coefficient).

In addition, we compare our algorithms with SamIam, which is a state-of-the-art implementation of the local search algorithm for marginal MAP (Park and Darwiche, 2004); we use its default Taboo search method with a maximum of 500 searching steps, and report the best results among 5 trials with random initializations, and one additional trial initialized by its default method (which sequentially initializes x_i by maximizing $p(x_i|x_{\text{pa}_i})$ along some predefined order).

We also implement an EM algorithm, whose expectation and maximization steps are approximated by sum-product and max-product BP, respectively. We run EM with 5 random initializations and one initialization by sum-product marginals, and pick the best solution.

9.1 Simulated Models

We consider pairwise models over discrete random variables taking values in $\{-1, 0, +1\}^n$,

$$p(x) \propto \exp \left[\sum_i \theta_i(x_i) + \sum_{(ij) \in E} \theta_{ij}(x_i, x_j) \right].$$

The value tables of θ_i and θ_{ij} are randomly generated from normal distribution, $\theta_i(k) \sim \text{Normal}(0, 0.01)$, $\theta_{ij}(k, l) \sim \text{Normal}(0, \sigma^2)$, where σ controls the strength of coupling. Our results are averaged on 1000 randomly generated sets of parameters.

We consider different choices of graph structures and max / sum node patterns:

1. *Hidden Markov chain* with 20 nodes, as shown in Figure 1.
2. *Latent tree models*. We generate random trees of size 50, by finding the minimum spanning trees of random symmetric matrices with elements drawn from $\text{Uniform}([0, 1])$. We take the leaf nodes to be max nodes, and the non-leaf nodes to be sum nodes. See Figure 5(a) for a typical example.
3. 10×10 *Grid* with max and sum nodes distributed in two opposite chess board patterns shown in Figure 6(a) and Figure 7(a), respectively. In Figure 6(a), the sum part is a loopy graph, and the max part is a (fully disconnected) tree; in Figure 7(a), the max and sum parts are flipped.

The results on the hidden Markov chain are shown in Figure 4, where we plot in panel (a) different algorithms’ percentages of obtaining the globally optimal solutions among 1000 random trials, and in panel (b) their relative energy errors defined by $Q(\hat{x}_B; \theta) - Q(x_B^*; \theta)$, where $\hat{x}_B$ is the solution returned by the algorithms, and x_B^* is the true optimum.

The results of the latent tree models and the two types of 2D grids are shown in Figure 5, Figure 6 and Figure 7, respectively. Since the globally optimal solution x_B^* is not tractable to calculate in these cases, we report the approximate relative error defined by $Q(\hat{x}_B; \theta) - Q(\tilde{x}_B; \theta)$, where $\tilde{x}_B$ is the best solution we found across all algorithms.

9.2 Diagnostic Bayesian Networks

We also test our algorithms on two diagnostic Bayesian networks taken from the UAI08 Inference Challenge, where we construct marginal MAP problems by randomly selecting varying percentages of nodes to be max nodes. Since these models are not pairwise, we implement the junction graph versions of `mix-product` (Bethe) and `proximal` (Bethe) shown in Section 8. Figure 8 shows the approximate relative errors of our algorithms and `local search` (SamIam) as the percentage of the max nodes varies.

9.3 Insights

Across all the experiments, we find that `mix-product` (Bethe), `proximal` (Bethe) and `local search` (SamIam) significantly outperform all the other algorithms, while `proximal` (Bethe) outperforms the two others in some circumstances. In the hidden Markov chain example in Figure 4, these three algorithms almost always (with probability $\geq 99\%$) find the globally optimal solutions. However, the performance of SamIam tends to degenerate when the max part has loopy dependency structures (see Figure 7), or when the number of max nodes is large (see Figure 8), both of which make it difficult to explore the solution space by local search. On the other hand, `mix-product` (Bethe) tends to degenerate as the coupling strength σ increases (see Figure 7), probably because its convergence gets worse as σ increases.

We note that our TRW approximation gives much less accurate solutions than the other algorithms, but is able to provide an upper bound on the optimal energy. Similar phenomena have been observed for TRW-BP in standard max- and sum- inference.

The hybrid message passing of Jiang et al. (2011) is significantly worse than `mix-product` (Bethe), `proximal` (Bethe) and `local search` (SamIam), but is otherwise the best among the remaining algorithms. EM performs similarly to (or sometimes worse than) Jiang’s method.

The regular max-product BP and sum-product BP are among the worst of the tested algorithms, indicating the danger of approximating mixed-inference by pure max- or sum- inference. Interestingly, the performances of max-product BP and sum-product BP have opposite trends: In Figure 4, Figure 5 and Figure 6, where the max parts are fully disconnected and the sum parts are connected and loopy, max-product BP usually performs worse than sum-product BP, but gets better as the coupling strength σ increases; sum-product BP, on the other hand, tends to degenerate as σ increases. In Figure 7, where the max / sum pattern is reversed (resulting in a larger, loopier max subgraph), max-product BP performs better than sum-product BP.

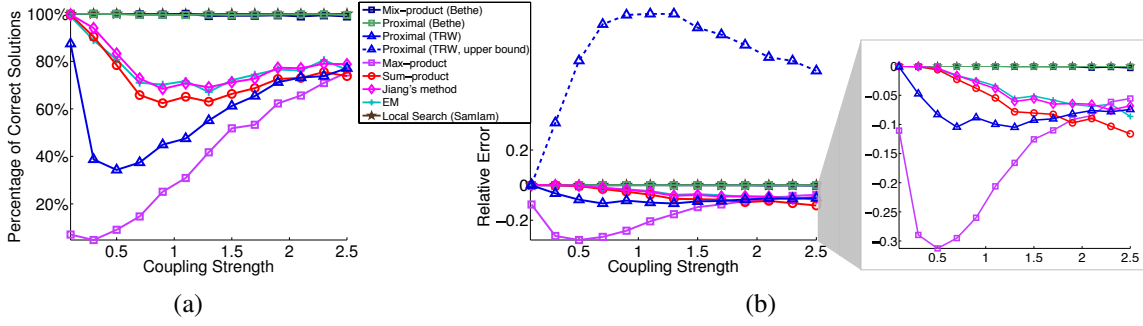


Figure 4: Results on the hidden Markov chain in Figure 1 (best viewed in color). (a) different algorithms’ probabilities of obtaining the globally optimal solution among 1000 random trials. Mix-product (Bethe), Proximal (Bethe) and Local Search (SamIam) almost always (with probability $\geq 99\%$) find the optimal solution. (b) The relative energy errors of the different algorithms, and the upper bounds obtained by Proximal (TRW) as a function of coupling strength σ .

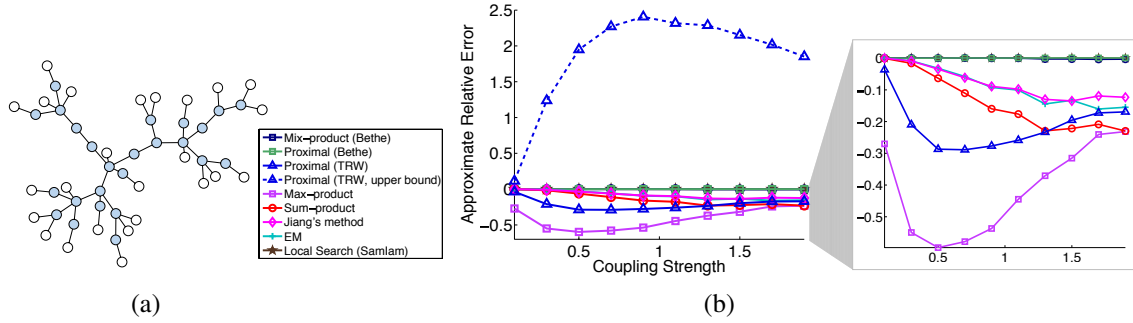


Figure 5: (a) A typical latent tree model, whose leaf nodes are taken to be max nodes (white) and non-leaf nodes to be sum nodes (shaded). (b) The approximate relative energy errors of different algorithms, and the upper bound obtained by Proximal (TRW) as a function of coupling strength σ .

10. Conclusion and Further Directions

We have presented a general variational framework for solving marginal MAP problems approximately, opening new doors for developing efficient algorithms. In particular, we show that our proposed “mixed-product” BP admits appealing theoretical properties and performs well in practice.

Potential future directions include improving the performance of the truncated TRW approximation by optimizing weights, deriving optimality conditions that may be applicable even when the sum component does not form a tree, studying the convergent properties of mixed-product BP, and leveraging our results to learn hidden variable models for data.

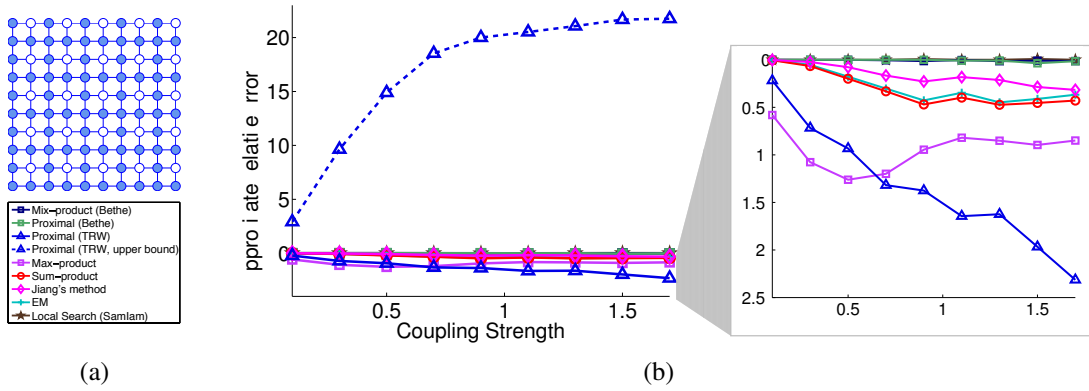


Figure 6: (a) A marginal MAP problem defined on a 10×10 Ising grid, with shaded sum nodes and unshaded max nodes; note that the sum part is a loopy graph, while max part is fully disconnected. (b) The approximate relative errors of different algorithms and the upper bound obtained by Proximal (TRW) as a function of coupling strength σ .

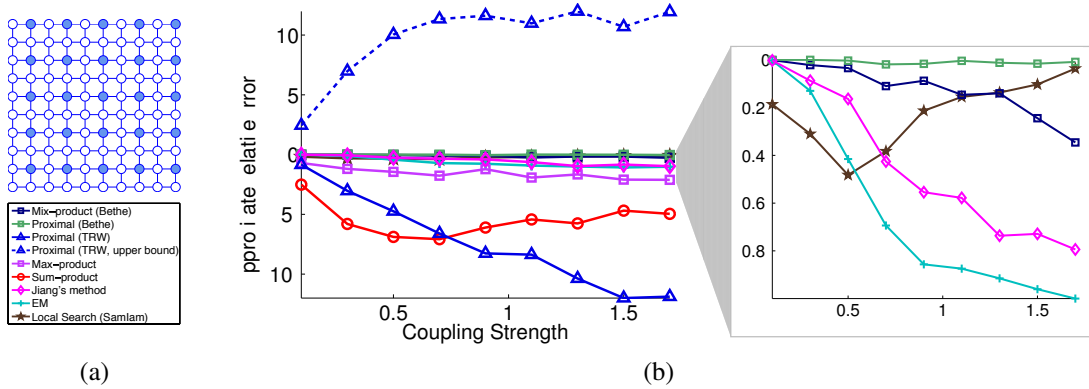


Figure 7: (a) A marginal MAP problem defined on a 10×10 Ising grid, but with max / sum part exactly opposite to that in Figure 6; note that the max part is loopy, while the sum part is fully disconnected in this case. (b) The approximate relative errors of different algorithms and the upper bound obtained by Proximal (TRW) as a function of coupling strength σ .

Acknowledgments

We thank Arthur Choi for providing help on SamIam. This work was supported in part by the National Science Foundation (awards IIS-1065618 and IIS-1254071), and a Microsoft Research Ph.D Fellowship.

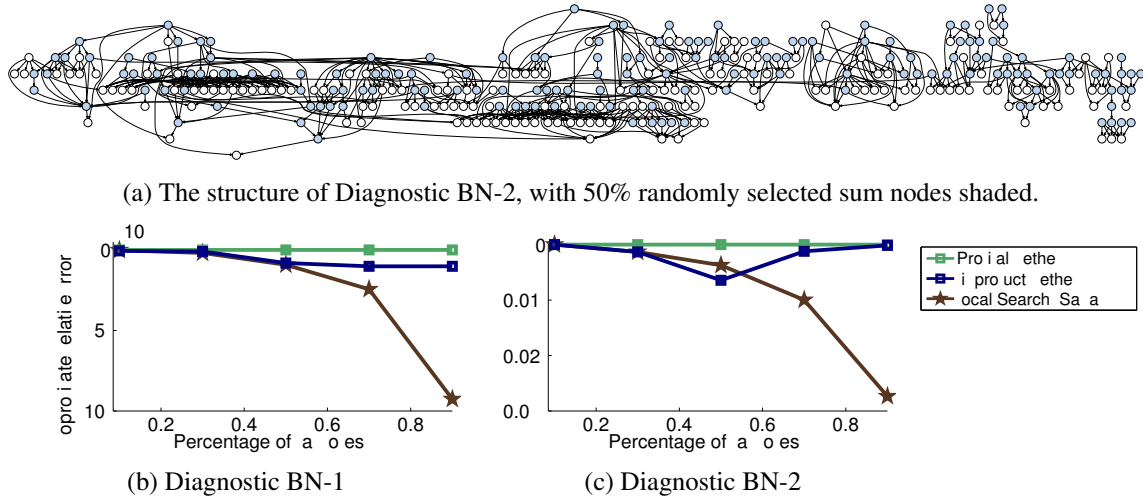


Figure 8: The results on two diagnostic Bayesian networks (BNs) in the UAI08 inference challenge. (a) The Diagnostic BN-2 network. (b)-(c) The performances of algorithms on the two BNs as a function of the percentage of max nodes. The local search method tends to degenerate when the number of max nodes is large, making it difficult to search over the solution space. Results are averaged over 100 random trials.

Appendix A. Proof of Proposition 6

Proof. The Lagrangian of (16) with the local consistency constraint of $\mathbb{L}$ in (2) is

$$\langle \theta, \tau \rangle + \sum_{i \in V} [w_i H_i(\tau) + \lambda_i^0 \sum_{x_i} \tau_i(x_i)] - \sum_{(ij) \in E} [w_{ij} I_{ij}(\tau) + \sum_{x_j} \lambda_{i \rightarrow j}(x_j) \sum_{x_i} (\tau_{ij}(x_i, x_j) - \tau_j(x_j))],$$

where $\{\lambda_i^0 : i \in V\}$ and $\{\lambda_{j \rightarrow i}(x_i) : (ij) \in E, x_i \in \mathcal{X}_i\}$ are the Lagrange multipliers. Recall that

$$\langle \theta, \tau \rangle = \sum_{i \in V} \theta_i(x_i) \tau_i(x_i) + \sum_{(ij) \in E} \theta_{ij}(x_i, x_j) \tau_{ij}(x_i, x_j),$$

$$H_i(\tau) = - \sum_{x_i} \tau_i(x_i) \log \tau_i(x_i),$$

$$I_{ij}(\tau) = \sum_{x_i, x_j} \tau_{ij}(x_i, x_j) \log \frac{\tau_{ij}(x_i, x_j)}{\sum_{x_i} \tau_{ij}(x_i, x_j) \sum_{x_j} \tau_{ij}(x_i, x_j)}.$$

Taking the derivative of the Lagrangian w.r.t. $\tau_i(x_i)$ and $\tau_{ij}(x_i, x_j)$, we have

$$\theta_i(x_i) - w_i \log \tau_i(x_i) + \sum_{j \in \partial_i} \lambda_{j \rightarrow i}(x_i) = \text{const}, \quad (33)$$

$$\theta_{ij}(x_i, x_j) - w_{ij} \log \frac{\tau_{ij}(x_i, x_j)}{\tau_i(x_i) \tau_j(x_j)} + \lambda_{i \rightarrow j}(x_j) + \lambda_{j \rightarrow i}(x_i) = \text{const}, \quad (34)$$

where we used the local consistency condition that $\sum_{x_j} \tau_{ij}(x_i, x_j) = \tau_i(x_i)$. By defining $m_{i \rightarrow j}(x_j) = \exp(\lambda_{i \rightarrow j}(x_j))$, we obtain (19) directly from (33)-(34).

Plugging (19) into the constraint that $\sum_{x_j} \tau_{ij}(x_i, x_j) = \tau_i(x_i)$ gives (18). $\square$

Appendix B. Proof of Theorem 5

Proof. (i). For $\tau \in \mathbb{M}^o$, the objective function in (14) equals

$$\begin{aligned} F_{tree}(\tau, \theta) &= \langle \theta, \tau \rangle + \sum_{i \in V} H_i(\tau) - \sum_{(ij) \in E_A} I_{ij}(\tau) - \sum_{(ij) \in \partial_{AB}} \rho_{ij} I_{ij}(\tau) \\ &= \langle \theta, \tau \rangle + \sum_{i \in V} H_i(\tau) - \sum_{(ij) \in E_A} I_{ij}(\tau) \end{aligned} \quad (35)$$

$$\begin{aligned} &= \langle \theta, \tau \rangle + H_{A|B}(\tau) \\ &= F_{mix}(\tau, \theta), \end{aligned} \quad (36)$$

where the equality in (35) is because $I_{ij}(\tau) = 0$ if $\forall (ij) \in \partial_{AB}$, and the equality in (36) is because the sum part G_A is a tree and we have the tree decomposition $H_{A|B} = \sum_{i \in V} H_i(\tau) - \sum_{(ij) \in E_A} I_{ij}(\tau)$. Therefore we have

$$\Phi_{tree}(\theta) = \max_{\tau \in \mathbb{L}} F_{tree}(\tau, \theta) \geq \max_{\tau \in \mathbb{M}^o} F_{tree}(\tau, \theta) = \max_{\tau \in \mathbb{M}^o} F_{mix}(\tau, \theta) = \Phi_{AB}(\theta), \quad (37)$$

where the inequality is because $\mathbb{M}^o \subset \mathbb{M} \subset \mathbb{L}$.

If there exists x_B^* such that $Q(x_B^*; \theta) = \Phi_{tree}(\theta)$, then we have

$$Q(x_B^*; \theta) = \Phi_{tree}(\theta) \geq \Phi_{AB}(\theta) = \max_{x_B} Q(x_B; \theta).$$

This proves that x_B^* is a globally optimal marginal MAP solution.

(ii). Because $\tau_i^*(x_i)$ for $\forall i \in B$ are deterministic, and the sum part G_A is a tree, we have that $\tau^* \in \mathbb{M}^o$. Therefore the inequality in (37) is tight, and we can conclude the proof by using Corollary 11. $\square$

Appendix C. Proof of Theorem 10

Proof. By Theorem 9, the beliefs $\{b_i, b_{ij}\}$ should satisfy the reparameterization property in (24) and the consistency conditions in (25)-(27). Without loss of generality, we assume $\{b_i, b_{ij}\}$ are normalized such that $\sum_{x_i} b_i(x_i) = 1$ for $i \in A$ and $\max_{x_i} b_i(x_i) = 1$ for $i \in B$.

I) For simplicity, we first prove the case of $C = B$, when $G = G_{C \cup A}$ itself is a semi A - B tree, and the theorem implies that x_B^* is a global optimum. By the reparameterization condition, we have

$$p(x) = \hat{p}_B(x_B) \hat{p}_{A|B}(x),$$

where

$$\hat{p}_B(x_B) = \prod_{i \in B} b_i(x_i) \prod_{(ij) \in E_B} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}}, \quad (38)$$

$$\hat{p}_{A|B}(x) = \prod_{i \in A} b_i(x_i) \prod_{(ij) \in E_A} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}} \prod_{(ij) \in \partial_{AB}} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}}. \quad (39)$$

Note we have

$$p(x_B) = \sum_{x_A} p(x) = \sum_{x_A} \hat{p}_B(x_B) \hat{p}_{A|B}(x) = \hat{p}_B(x_B) \sum_{x_A} \hat{p}_{A|B}(x).$$

We just need to show that x_B^* maximizes $\hat{p}_B(x_B)$ and $\sum_{x_A} \hat{p}_{A|B}(x)$, respectively.

First, since $\hat{p}_B(x_B)$ involves only the max nodes, a standard MAP analysis applies. Because the max part of the beliefs, $\{b_i, b_{ij} : (ij) \in E_B\}$, satisfy the standard max-consistency conditions, and the corresponding TRW weights $\{\rho_{ij} : (ij) \in E_B\}$ are provably convex by assumption, we establish that x_B^* is the MAP solution of $\hat{p}_B(x_B)$ by Theorem 1 of Weiss et al. (2007).

Secondly, to show that x_B^* also maximizes $\hat{p}_{A|B}(x)$ requires the combination of the mixed-consistency and sum-consistency conditions. Since G is a semi A - B tree, we denote by π_i the unique parent node of i ($\pi_i = \emptyset$ if i is a root). In addition, let ∂_A be the subset of A whose parent nodes are in B , that is, $\partial_A = \{i \in A : \pi_i \in B\}$. Equation (39) can be reformed into

$$\hat{p}_{A|B}(x) = \prod_{i \in A \setminus \partial_A} \frac{b_{i,\pi_i}(x_i, x_{\pi_i})}{b_{\pi_i}(x_{\pi_i})} \prod_{i \in \partial_A} \left[\frac{b_{i,\pi_i}(x_i, x_{\pi_i})}{b_{\pi_i}(x_{\pi_i})} \right]^{\rho_{i,\pi_i}} \left[b_i(x_i) \right]^{1-\rho_{i,\pi_i}},$$

where we used the fact that $\rho_{ij} = 1$ for $(ij) \in E_A$. Therefore, we have for any $x_B \in \mathcal{X}_B$,

$$\begin{aligned} \sum_{x_A} \hat{p}_{A|B}(x) &= \sum_{x_A} \left\{ \prod_{i \in A \setminus \partial_A} \frac{b_{i,\pi_i}(x_i, x_{\pi_i})}{b_{\pi_i}(x_{\pi_i})} \prod_{i \in \partial_A} \left[\frac{b_{i,\pi_i}(x_i, x_{\pi_i})}{b_{\pi_i}(x_{\pi_i})} \right]^{\rho_{i,\pi_i}} \left[b_i(x_i) \right]^{1-\rho_{i,\pi_i}} \right\} \\ &= \prod_{i \in \partial_A} \sum_{x_i} \left[\frac{b_{i,\pi_i}(x_i, x_{\pi_i})}{b_{\pi_i}(x_{\pi_i})} \right]^{\rho_{i,\pi_i}} \left[b_i(x_i) \right]^{1-\rho_{i,\pi_i}} \end{aligned} \quad (40)$$

$$\leq \prod_{i \in \partial_A} \left[\sum_{x_i} \frac{b_{i,\pi_i}(x_i, x_{\pi_i})}{b_{\pi_i}(x_{\pi_i})} \right]^{\rho_{i,\pi_i}} \left[\sum_{x_i} b_i(x_i) \right]^{1-\rho_{i,\pi_i}} \quad (41)$$

$$= 1, \quad (42)$$

where the equality in (40) eliminates (by summation) all the interior nodes in A . The inequality in (41) follows from Hölder's inequality. Finally, the equality in (42) holds because all the sum part of beliefs $\{b_i, b_{ij} : (ij) \in E_A\}$ satisfies the sum-consistency (25).

On the other hand, for any $(i, \pi_i) \in \partial_{AB}$, because $x_{\pi_i}^* = \arg \max_{x_{\pi_i}} b_{\pi_i}(x_{\pi_i})$, we have $b_{i,\pi_i}(x_i, x_{\pi_i}^*) = b_i(x_i)$ by the mixed-consistency condition (27). Therefore,

$$\begin{aligned} \sum_{x_A} \hat{p}_{A|B}([x_A, x_B^*]) &= \prod_{i \in \partial_A} \sum_{x_i} \left[\frac{b_{i,\pi_i}(x_i, x_{\pi_i}^*)}{b_{\pi_i}(x_{\pi_i}^*)} \right]^{\rho_{i,\pi_i}} \left[b_i(x_i) \right]^{1-\rho_{i,\pi_i}} \\ &= \prod_{i \in \partial_A} \left[\frac{1}{b_{\pi_i}(x_{\pi_i}^*)} \right]^{\rho_{i,\pi_i}} \sum_{x_i} b_i(x_i) \\ &= 1. \end{aligned} \quad (43)$$

Combining (42) and (43), we have $\sum_{x_A} \hat{p}_{A|B}(x) \leq \sum_{x_A} \hat{p}_{A|B}([x_A, x_B^*]) = 1$ for any $x_B \in \mathcal{X}_B$, that is, x_B^* maximizes $\sum_{x_A} \hat{p}_{A|B}(x)$. This finishes the proof for the case $C = B$.

II) In the case of $C \neq B$, let $D = B \setminus C$. We decompose $p(x)$ into

$$p(x) = \hat{p}_B([x_C, x_D]) \hat{p}_{A|C}([x_A, x_C]) \hat{r}_{AD}([x_A, x_D])$$

where $\hat{p}_B(x_B)$ and $\hat{p}_{A|B}(x)$ are defined similarly to (38) and (39),

$$\begin{aligned}\hat{p}_B(x_B) &= \prod_{i \in B} b_i(x_i) \prod_{(ij) \in E_B} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}}, \\ \hat{p}_{A|C}([x_A, x_C]) &= \prod_{i \in A} b_i(x_i) \prod_{(ij) \in E_A} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}} \prod_{(ij) \in \partial_{AC}} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}},\end{aligned}$$

where π_i is the parent node of i in the semi A - B tree $G_{A \cup C}$ and ∂_{AC} is set of edges across A and C , that is, $\partial_{AC} = \{(ij) \in E : i \in A, j \in C\}$. The term $\hat{r}_{AD}(x)$ is defined as

$$\hat{r}_{AD}([x_A, x_D]) = \prod_{(ij) \in \partial_{AD}} \left[\frac{b_{ij}(x_i, x_j)}{b_i(x_i) b_j(x_j)} \right]^{\rho_{ij}},$$

where similarly ∂_{AD} is the set of edges across A and D .

Because $x_j^* = \arg \max_{x_j} b_j(x_j)$ for $j \in D$, we have $b_{ij}(x_i, x_j^*) = b_i(x_i)$ for $(ij) \in \partial_{AD}$, $j \in D$ by the mixed-consistency condition in (27). Therefore, one can show that $\hat{r}_{AD}([x_A, x_D^*]) = 1$, and hence

$$p([x_A, x_C, x_D^*]) = \hat{p}_B([x_C, x_D^*]) \hat{p}_{A|C}([x_A, x_C]).$$

The remainder of the proof is similar to that for the case $C = B$: by the analysis in Weiss et al. (2007), it follows that $x_C^* \in \arg \max_{x_C} p([x_C, x_D^*])$, and we have previously shown that $x_C^* \in \arg \max_{x_C} \sum_{x_A} \hat{p}_{A|C}([x_A, x_C])$. This establishes that x_C^* maximizes

$$\sum_{x_A} p([x_A, x_C, x_D^*]) = p([x_C, x_D^*]) \sum_{x_A} \hat{p}_{A|C}([x_A, x_C]),$$

which concludes the proof. $\square$

References

- F. Altarelli, A. Braunstein, A. Ramezanpour, and R. Zecchina. Stochastic optimization by message passing. *Journal of Statistical Mechanics: Theory and Experiment*, 2011(11):P11009, 2011.
- J.R. Birge and F. Louveaux. *Introduction to Stochastic Programming*. Springer Verlag, 1997.
- S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning*, 2010.
- T.M. Cover and J.A. Thomas. *Elements of Information Theory*. John Wiley & Sons, INC., 2nd edition, 2006.
- C.P. De Campos. New complexity results for MAP in Bayesian networks. In *Proceedings of the 22nd International Joint Conference on Artificial Intelligence (IJCAI-11)*, pages 2100–2106, 2011.
- R. Dechter and I. Rish. Mini-buckets: A general scheme for bounded inference. *Journal of the ACM (JACM)*, 50(2):107–153, 2003.

- A. Doucet, S.J. Godsill, and C.P. Robert. Marginal maximum a posteriori estimation using Markov chain Monte Carlo. *Statistics and Computing*, 12(1), January 2002.
- A. Globerson and T.S. Jaakkola. Approximate inference using conditional entropy decompositions. In *Proceedings of the Eleventh International Conference on Artificial Intelligence and Statistics (AISTATS-07)*, 2007.
- D.M. Greig, B.T. Porteous, and A.H. Seheult. Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 271–279, 1989.
- T. Hazan and A. Shashua. Norm-product belief propagation: Primal-dual message-passing for approximate inference. *Information Theory, IEEE Transactions on*, 56(12):6294–6316, 2010.
- T. Hazan, J. Peng, and A. Shashua. Tightening fractional covering upper bounds on the partition function for high-order region graphs. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-12)*, 2012.
- R.A. Howard and J.E. Matheson. Influence diagrams. *Decision Analysis*, 2(3):127–143, 2005.
- D.R. Hunter and K. Lange. A tutorial on MM algorithms. *The American Statistician*, 1(58), February 2004.
- M. Ibrahimi, A. Javanmard, Y. Kanoria, and A. Montanari. Robust max-product belief propagation. In *Signals, Systems and Computers (ASILOMAR), 2011 Conference Record of the Forty Fifth Asilomar Conference on*, pages 43–49. IEEE, 2011.
- A. Iusem and M. Teboulle. On the convergence rate of entropic proximal optimization methods. *Computational and Applied Mathematics*, 12:153–168, 1993.
- E.T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, May 1957.
- M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the ising model. *SIAM Journal on Computing*, 22(5):1087–1116, 1993.
- J. Jiang, P. Rai, and H. Daumé III. Message-passing for approximate map inference with latent variables. In *Advances in Neural Information Processing Systems (NIPS-11)*, 2011.
- V. Jojic, S. Gould, and D. Koller. Accelerated dual decomposition for MAP inference. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, 2010.
- D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT press, 2009.
- V. Kolmogorov. Convergent tree-reweighted message passing for energy minimization. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 28(10):1568–1583, oct. 2006.
- Q. Liu and A. Ihler. Bounding the partition function using Hölder’s inequality. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*, pages 849–856, June 2011a.

- Q. Liu and A. Ihler. Variational algorithms for marginal MAP. In *Proceedings of the 27th Conference on Uncertainty in Artificial Intelligence (UAI-11)*. 2011b.
- Q. Liu and A. Ihler. Belief propagation for structured decision making. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-12)*. August 2012.
- B. Martinet. Régularisation d'inéquations variationnelles par approximations successives. *Revue Française d'Informatique et de Recherche Opérationnelle*, 4:154–158, 1970.
- R. Mateescu, K. Kask, V. Gogate, and R. Dechter. Join-graph propagation algorithms. *Journal of Artificial Intelligence Research*, 37(1):279–328, 2010.
- D.D. Mauá and C.P. de Campos. Anytime marginal maximum a posteriori inference. In *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*, 2012.
- T. Meltzer, A. Globerson, and Y. Weiss. Convergent message passing algorithms: a unifying view. In *Proceedings of the 25th Conference on Uncertainty in Artificial Intelligence (UAI-09)*, 2009.
- O.P. Meshi, T. Jaakkola, and A. Globerson. Convergence rate analysis of MAP coordinate minimization algorithms. In *Advances in Neural Information Processing Systems (NIPS-12)*, 2012.
- R. Neal and G.E. Hinton. A view of the EM algorithm that justifies incremental, sparse, and other variants. In M. Jordan, editor, *Learning in Graphical Models*, pages 355–368. Kluwer, 1998.
- J. Park and A. Darwiche. Complexity results and approximation strategies for MAP explanations. *Journal of Artificial Intelligence Research*, 21:101–133, 2004.
- J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.
- P. Ravikumar, A. Agarwal, and M.J. Wainwright. Message-passing for graph-structured linear programs: Proximal projections, convergence, and rounding schemes. *Journal of Machine Learning Research*, 11:1043–1080, Mar 2010.
- R.T. Rockafellar. Monotone operators and the proximal point algorithm. *SIAM Journal on Control and Optimization*, 14(5):877, 1976.
- B. Savchynskyy, S. Schmidt, J.H. Kappes, and C. Schnörr. Efficient MRF energy minimization via adaptive diminishing smoothing. In *Proceedings of the 28th Conference on Uncertainty in Artificial Intelligence (UAI-12)*, 2010.
- D. Sontag, A. Globerson, and T.S. Jaakkola. Introduction to dual decomposition for inference. In *Optimization for Machine Learning*. MIT Press, 2011.
- M. Teboulle. Entropic proximal mappings with applications to nonlinear programming. *Mathematics of Operations Research*, 17(3):pp. 670–690, 1992.
- P. Tseng and D.P. Bertsekas. On the convergence of the exponential multiplier method for convex programming. *Mathematical Programming*, 60(1):1–19, 1993.

- M.J. Wainwright and M. Jordan. Graphical models, exponential families, and variational inference. *Foundation and Trends in Machine Learning*, 1(1-2):1–305, 2008.
- M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *Information Theory, IEEE Transactions on*, 45: 1120–1146, 2003.
- M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. A new class of upper bounds on the log partition function. *Information Theory, IEEE Transactions on*, 51(7):2313–2335, July 2005a.
- M.J. Wainwright, T.S. Jaakkola, and A.S. Willsky. MAP estimation via agreement on trees: message-passing and linear programming. *Information Theory, IEEE Transactions on*, 51(11): 3697–3717, 2005b.
- Y. Weiss, C. Yanover, and T. Meltzer. MAP estimation, linear programming and belief propagation with convex free energies. In *Proceedings of the 23rd Conference on Uncertainty in Artificial Intelligence (UAI-07)*, 2007.
- T. Werner. A linear programming approach to max-sum problem: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(7):1165–1179, 2007.
- T. Werner. Revisiting the linear programming relaxation approach to Gibbs energy minimization and weighted constraint satisfaction. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 32(8):1474–1488, 2010.
- W. Wiegerinck. Approximations with reweighted generalized belief propagation. In *In Proceedings of the 9th International Conference on Artificial Intelligence and Statistics (AISTATS-05)*, 2005.
- J.S. Yedidia, W.T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. *Exploring Artificial Intelligence in the New Millennium*, 8:236–239, 2003.
- J.S. Yedidia, W.T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized BP algorithms. *Information Theory, IEEE Transactions on*, 51, July 2005.
- C. Yuan, T.C. Lu, and M.J. Druzdzal. Annealed MAP. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence (UAI-04)*, pages 628–635, 2004.
- A. L. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14(7):1691–1722, July 2002.