MAP estimation via agreement on (hyper)trees: Message-passing and linear programming approaches

Martin Wainwright* martinw@eecs.berkeley.edu
Tommi Jaakkola† tommi@ai.mit.edu
Alan Willsky† willsky@mit.edu

Electrical Engineering & CS*
UC Berkeley, CA 94720

Electrical Engineering & CS†
MIT, Cambridge, MA 02139

Abstract

We develop an approach for computing provably exact *maximum a posteriori* (MAP) configurations for a subclass of problems on graphs with cycles. By decomposing the original problem into a convex combination of tree-structured problems, we obtain an upper bound on the optimal value of the original problem (i.e., the log probability of the MAP assignment) in terms of the combined optimal values of the tree problems. We prove that this upper bound is met with equality if and only if the tree problems share an optimal configuration in common. An important implication is that any such shared configuration must also be a MAP configuration for the original problem. Next we present and analyze two methods for attempting to obtain tight upper bounds: (a) a *tree-reweighted message-passing algorithm* that is related to but distinct from the max-product (min-sum) algorithm; and (b) a *tree-relaxed linear program* (LP), which is derived from the Lagrangian dual of the upper bounds. Finally, we discuss the conditions that govern when the relaxation is tight, in which case the MAP configuration can be obtained. The analysis described here generalizes naturally to convex combinations of hypertree-structured distributions.

1 Introduction

Integer programming problems arise in various fields, including communication theory, statistical physics, and error-correcting coding. Many such problems can be formulated in terms of graphical models [3], in which the cost function corresponds to a graph-structured probability distribution, and the goal is to find the *maximum a posteriori* (MAP) configuration.

In previous work [9], we have shown how to use convex combinations of tree-structured distributions in order to upper bound the log partition function. In this paper, we show that similar ideas can be used to derive upper bounds on the log probability of the MAP assignment. Moreover, in cases when the bound is tight (i.e., met with equality), the MAP configuration itself can be obtained. We propose and analyze two methods for obtaining a tight bound, and hence the exact MAP assignment. We begin with the goal of finding a collection of tree-structured problems that share a common optimum, from which we are led to a *tree-reweighted* set of message-passing updates. The resulting algorithm, though similar to both the standard max-product algorithm [e.g., 7, 10] and the attenuated max-product updates proposed by Frey and Koetter [8], differs in certain key ways. Our second approach is to exploit the convexity of the upper bounds, and hence apply the theory of Lagrangian duality. We show that optimal upper bounds can be obtained by solving a linear program (LP), one which follows by applying a so-called *tree relaxation*. In this way, our work establishes a connection between two approaches to integer programming: approximate dynamic programming methods using message-passing, and LP-based relaxations.

Work supported in part by ODDR&E MURI Grant DAAD19-00-1-0466 through the ARO; by ONR N00014-00-1-0089; and by the AFOSR F49620-00-1-0362.

The following two subsections of the paper provide background on exponential families and convex combinations. In Section 2, we introduce the basic form of the upper bounds on the log probability of the MAP assignment, and then develop necessary and sufficient conditions for it to be met with equality. In Section 3, we develop tree-reweighted max-product algorithms for finding a convex combination of trees that can yield a tight upper bound. An algorithm in this family always has at least one fixed point which, if it satisfies a key uniqueness condition, specifies a provably MAP-optimal configuration. In Section 4, we derive the Lagrangian dual of our upper bounds, and show that it is a linear program (LP) that can be interpreted as enforcing a set of tree constraints. In Section 5, we explore the conditions that govern the tightness of the resulting bounds. We conclude in Section 6 with discussion of related work, and extensions to the analysis presented here.

1.1 Notation and set-up

Consider an undirected (simple) graph $G = (V, E)$. For each vertex $s \in V$, let x_s be a random variable taking values in the discrete space $\mathcal{X}_s = \{0, 1, \ldots, m_s - 1\}$. We use the letters j, k to denote particular elements of the sample space $\mathcal{X}_s$. The overall random vector $\mathbf{x} = \{x_s \mid s \in V\}$ takes values in the Cartesian product space $\mathcal{X}^N = \mathcal{X}_1 \times \cdots \times \mathcal{X}_N$, where $N = |V|$. We make use of the following exponential representation of a graph-structured distribution $p(\mathbf{x})$. For some index set $\mathcal{I}$, we let $\boldsymbol{\phi} = \{\phi_\alpha \mid \alpha \in \mathcal{I}\}$ denote a collection of potential functions defined on the cliques of G, and let $\theta = \{\theta_\alpha \mid \alpha \in \mathcal{I}\}$ be a vector of weights on these potential functions. The exponential family determined by $\boldsymbol{\phi}$ is the collection $p(\mathbf{x}; \theta) \propto \exp\big\{\sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(\mathbf{x})\big\}$ of Gibbs distributions.

In a *minimal* exponential representation, the functions $\{\phi_\alpha\}$ are linearly independent. For example, one minimal representation of a binary process (i.e., $\mathcal{X}_s = \{0, 1\}$ for all $s \in V$) using pairwise potential functions is the usual Ising model, in which the collection of potentials $\boldsymbol{\phi} = \{x_s \mid s \in V\} \;\cup\; \{\, x_s x_t \mid (s,t) \in E\}$. Here the index set $\mathcal{I} = V \cup E$. In most of our analysis, we use an *overcomplete* representation, with indicator functions as potentials:

$$
\begin{aligned}
\phi_{s;j}(x_s) &= \delta(x_s = j), && s \in V;\; j \in \mathcal{X}_s && \text{(1a)}\\
\phi_{st;jk}(x_s, x_t) &= \delta(x_s = j, x_t = k), && (s,t) \in E;\; (j,k) \in \mathcal{X}_s \times \mathcal{X}_t && \text{(1b)}
\end{aligned}
$$

In this case, the index set $\mathcal{I}$ consists of the union of $\mathcal{I}(V) = \{\, (s;j) \mid s \in V;\; j \in \mathcal{X}_s\}$ with the edge indices $\mathcal{I}(E) = \{\, (st;jk) \mid (s,t) \in E;\; (j,k) \in \mathcal{X}_s \times \mathcal{X}_t \,\}$.

Of interest to us is the *maximum a posteriori* configuration $\widehat{\mathbf{x}}_{\mathrm{MAP}} = \arg\max_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}; \theta)$. Equivalently, we can express this MAP configuration as the solution of the integer program $\mathcal{F}(\theta) = \max_{\mathbf{x} \in \mathcal{X}^N} J(\mathbf{x}; \theta)$, where

$$
J(\mathbf{x}; \theta) \;=\; \theta \cdot \boldsymbol{\phi}(\mathbf{x}) \;=\; \sum_{s \in V} \sum_{j \in \mathcal{X}_s} \theta_{s;j} \phi_{s;j}(x_s) + \sum_{(s,t) \in E} \;\sum_{(j,k) \in \mathcal{X}_s \times \mathcal{X}_t} \theta_{st;jk} \phi_{st;jk}(x_s, x_t) \quad (2)
$$

As the maximum of a collection of linear functions, the function $\mathcal{F}(\theta)$ is convex [1] in terms of θ, which is a key property for our subsequent development.

1.2 Convex combinations of trees

Let $\widehat{\theta}$ be a particular parameter vector for which we are interested in computing $\mathcal{F}(\widehat{\theta})$. In this section, we show how to exploit the convexity of $\mathcal{F}$ in order to derive upper bounds. Let $\mathcal{T}$ denote a particular spanning tree of G, and let $\mathfrak{T} = \mathfrak{T}(G)$ denote the set of all spanning trees.[1] For each spanning tree $\mathcal{T} \in \mathfrak{T}$, let $\theta(\mathcal{T})$ be an exponential parameter vector of the same dimension as θ that respects the structure of $\mathcal{T}$. To be explicit, if $\mathcal{T}$ is defined by an edge set $E(\mathcal{T}) \subset E$, then $\theta(\mathcal{T})$ must have zeros in all elements[2] corresponding to edges not in $E(\mathcal{T})$. For compactness in notation, let $\Theta \triangleq \{\theta(\mathcal{T}) \mid \mathcal{T} \in \mathfrak{T}\}$ denote the full collection of tree-structured exponential parameter vectors.

In order to define a convex combination, we require a probability distribution over spanning trees $\vec{\boldsymbol{\mu}} \triangleq \{\ \mu(\mathcal{T}),\ \mathcal{T} \in \mathfrak{T} \mid \mu(\mathcal{T}) \geq 0\ \sum_{\mathcal{T}} \mu(\mathcal{T}) = 1\}$. For any distribution $\vec{\boldsymbol{\mu}}$, we define its *support*, denoted by $\mathrm{supp}(\vec{\boldsymbol{\mu}})$, to be the set of trees to which it assigns strictly positive probability. In the sequel, we will also be interested in the probability $\mu_e = \Pr_{\vec{\boldsymbol{\mu}}}\{e \in \mathcal{T}\}$ that a given edge $e \in E$ appears in a spanning tree $\mathcal{T}$ chosen randomly under $\vec{\boldsymbol{\mu}}$. It can be shown [9] that the vector $\boldsymbol{\mu_e}$ of these edge appearance probabilities must belong to the so-called *spanning tree polytope* [2]. Throughout this paper, we will assume that $\mu_e > 0$ for all $e \in E$ (i.e., each edge belongs to at least one tree in $\mathrm{supp}(\vec{\boldsymbol{\mu}})$); we call any such $\vec{\boldsymbol{\mu}}$ a *valid* distribution. A *convex combination* of exponential parameter vectors is defined via the weighted sum $\sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T})\theta(\mathcal{T})$, which we denote compactly as $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})]$. For a given $\widehat{\theta}$, of particular importance are the collection of *admissible pairs* $\mathcal{A}(\widehat{\theta}) \triangleq \big\{(\Theta; \vec{\boldsymbol{\mu}})\ \big|\ \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] = \widehat{\theta}\big\}$. For any valid distribution $\vec{\boldsymbol{\mu}}$, it can be seen that there exist pairs $(\Theta; \vec{\boldsymbol{\mu}}) \in \mathcal{A}(\widehat{\theta})$.

Example 1 (Single cycle). To illustrate these definitions, consider a binary vector $\mathbf{x} \in \{0, 1\}^4$ on a 4-node cycle, with the distribution in the minimal Ising form $p(\mathbf{x}; \widehat{\theta}) = \exp\{x_1 x_2 + x_2 x_3 + x_3 x_4 + x_4 x_1 - \Phi(\widehat{\theta})\}$. That is, the target distribution is specified by the minimal parameter $\widehat{\theta} = [0\ 0\ 0\ 0\ \ 1\ 1\ 1\ 1]$, where the zeros represent the fact that $\widehat{\theta}_s = 0$ for all $s \in V$. The four possible spanning trees $\mathfrak{T} = \{\mathcal{T}_i \mid i = 1, \ldots, 4\ \}$ on a single cycle on four nodes are

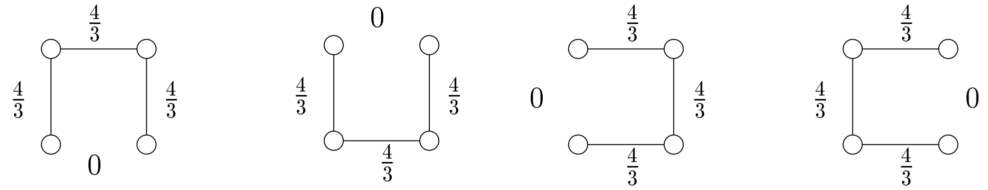


Figure 1. A convex combination of four distributions $p(\mathbf{x}; \theta(\mathcal{T}_i))$, each defined by a spanning tree $\mathcal{T}_i$, is used to approximate the target distribution $p(\mathbf{x}; \widehat{\theta})$ on the single-cycle graph.

illustrated in Figure 1. We define a set of associated exponential parameters $\Theta = \{\theta(\mathcal{T}_i)\}$ as follows:

$$\begin{aligned} \theta(\mathcal{T}_1) &= (4/3)\ \big[0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 1\ \ 0\big] & \theta(\mathcal{T}_3) &= (4/3)\ \big[0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 0\ \ 1\ \ 1\big] \\ \theta(\mathcal{T}_2) &= (4/3)\ \big[0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 0\ \ 1\big] & \theta(\mathcal{T}_4) &= (4/3)\ \big[0\ \ 0\ \ 0\ \ 0\ \ 0\ \ 1\ \ 1\ \ 1\big] \end{aligned}$$

Finally, choosing the uniform distribution $\{\mu(\mathcal{T}_i) = 1/4 \mid i = 1, \ldots 4\}$ over trees, we have $\mu_e = 3/4$ for each edge, and moreover, $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] = \widehat{\theta}$ so that $(\Theta; \vec{\boldsymbol{\mu}}) \in \mathcal{A}(\widehat{\theta})$.

2 Basic form of upper bounds

We now describe the basic form of the upper bounds to be studied. With the set-up of the previous section, an upper bound follows by applying Jensen's inequality [1] to a pair $(\Theta; \vec{\boldsymbol{\mu}}) \in \mathcal{A}(\widehat{\theta})$,

[1] Although we focus on spanning trees, our analysis applies more generally to acyclic subgraphs.

[2] Given an edge e belonging to two trees $\mathcal{T}_1$ and $\mathcal{T}_2$, the quantity $\theta_e(\mathcal{T}_1)$ can be different than $\theta_e(\mathcal{T}_2)$.

thereby yielding:

$$\mathcal{F}(\widehat{\theta}) \;\leq\; \sum_{\mathcal{T}} \mu(\mathcal{T})\mathcal{F}(\theta(\mathcal{T})) \;=\; \sum_{\mathcal{T}} \mu(\mathcal{T}) \max_{\mathbf{x}\in\mathcal{X}^N} \big\{\theta(\mathcal{T})\cdot\boldsymbol{\phi}(\mathbf{x})\big\} \tag{3}$$

We wish to understand when the upper bound (3) is *tight* — that is, met with equality. For any exponential parameter vector $\theta(\mathcal{T})$, we define the collection of its optimal configurations:

$$\mathrm{OPT}(\theta(\mathcal{T})) \;=\; \{\mathbf{x}\in\mathcal{X}^N \mid \theta(\mathcal{T})\cdot\boldsymbol{\phi}(\mathbf{x}') \;\leq \theta(\mathcal{T})\cdot\boldsymbol{\phi}(\mathbf{x}) \;\text{ for all }\; \mathbf{x}'\in\mathcal{X}^N\} \tag{4}$$

With this notation, we have:

Proposition 1 (Tightness of bound). *The bound of equation* (3) *is tight if and only if the intersection* $\mathrm{OPT}(\Theta) \triangleq \cap_{\mathcal{T}\in\mathrm{supp}(\vec{\boldsymbol{\mu}})} \mathrm{OPT}(\theta(\mathcal{T}))$ *is non-empty. (I.e., there exists a configuration* $\widehat{\mathbf{x}}\in\mathcal{X}^N$ *that for each* $\mathcal{T}\in\mathrm{supp}(\vec{\boldsymbol{\mu}})$ *achieves the maximum defining* $\mathcal{F}(\theta(\mathcal{T}))$*.)*

Proof. Consider some pair $(\Theta;\vec{\boldsymbol{\mu}})\in\mathcal{A}(\widehat{\theta})$. Let $\widehat{\mathbf{x}}$ be a configuration that attains the maximum defining $\mathcal{F}(\widehat{\theta})$. We write the difference of the RHS and the LHS of equation (3) as follows:

$$\begin{aligned} 0 \;\leq\; \Big[\sum_{\mathcal{T}}\mu(\mathcal{T})\mathcal{F}(\theta(\mathcal{T}))\Big] - \mathcal{F}(\widehat{\theta}) \;&=\; \Big[\sum_{\mathcal{T}}\mu(\mathcal{T})\mathcal{F}(\theta(\mathcal{T}))\Big] - \widehat{\theta}\cdot\boldsymbol{\phi}(\widehat{\mathbf{x}}) \\ &=\; \sum_{\mathcal{T}}\mu(\mathcal{T})\big[\mathcal{F}(\theta(\mathcal{T})) - \theta(\mathcal{T})\cdot\boldsymbol{\phi}(\widehat{\mathbf{x}})\big] \end{aligned}$$

Now for each $\mathcal{T}\in\mathrm{supp}(\vec{\boldsymbol{\mu}})$, the term $\mathcal{F}(\theta(\mathcal{T})) - \theta(\mathcal{T})\cdot\boldsymbol{\phi}(\widehat{\mathbf{x}})$ is non-negative, and equal to zero only when $\widehat{\mathbf{x}}$ belongs to $\mathrm{OPT}(\theta(\mathcal{T}))$. Therefore, the bound is met with equality if and only if $\widehat{\mathbf{x}}$ achieves the maximum defining $\mathcal{F}(\theta(\mathcal{T}))$ for all trees $\mathcal{T}\in\mathrm{supp}(\vec{\boldsymbol{\mu}})$. □

We now consider two different strategies for obtaining tight upper bounds, and hence the MAP configuration.

Admissibility and mutual agreement Suppose that for a given spanning tree distribution $\vec{\boldsymbol{\mu}}$, we find a collection of exponential parameters $\Theta^* = \{\theta^*(\mathcal{T})\}$ such that:
(a) Admissibility: The pair $(\Theta^*;\vec{\boldsymbol{\mu}})$ satisfies $\sum_{\mathcal{T}}\mu(\mathcal{T})\theta^*(\mathcal{T}) = \widehat{\theta}$.
(b) Mutual agreement: The intersection $\cap_{\mathcal{T}}\mathrm{OPT}(\theta^*(\mathcal{T}))$ of configurations optimal for all tree problems is non-empty.
Given a collection Θ^* satisfying these two properties, then Proposition 1 guarantees that all configurations in the (non-empty) intersection $\cap_{\mathcal{T}}\mathrm{OPT}(\theta^*(\mathcal{T}))$ achieve the maximum defining $\mathcal{F}(\widehat{\theta})$. Accordingly, in Section 3, we present an iterative message-passing technique for attempting to find such a collection Θ^*.

Lagrangian duality A second approach is to minimize the upper bound in equation (3) as a function of the collection of exponential parameters Θ. In particular, we consider the constrained optimization problem:

$$\min_{\Theta\,\mathrm{s.t}\,\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})]=\widehat{\theta}} \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\mathcal{F}(\theta(\mathcal{T}))] \;=\; \min_{\Theta\,\mathrm{s.t}\,\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})]=\widehat{\theta}} \Big\{\sum_{\mathcal{T}}\mu(\mathcal{T})\max_{\mathbf{x}}\big[\theta(\mathcal{T})\cdot\boldsymbol{\phi}(\mathbf{x})\big]\Big\} \tag{5}$$

Observe that the cost function, as a linear combination of the convex functions $\mathcal{F}(\theta(\mathcal{T}))$, is also convex as a function of Θ, and the constraints are linear in Θ. Therefore, as we show in Section 4, the problem of finding tight upper bounds can also be tackled via the Lagrangian dual [1] of equation (5).

3 Tight bounds via equal max-marginals

We begin by pursuing the former approach; more specifically, we develop an algorithm that attempts to find, for a given spanning tree distribution $\vec{\boldsymbol{\mu}}$, a collection $\Theta^* = \{\theta^*(\mathcal{T})\}$ that satisfies admissibility and mutual agreement. Although this algorithm can be formulated in terms of reparameterization updates [e.g., 10], here we present a set of message-passing updates.

The foundation of our development is the factorization of any tree-structured distribution $p(\mathbf{x}; \theta(\mathcal{T}))$ in terms of its max-marginals [3]. In particular, for each node s, the associated single node max-marginal is defined by the following maximum over all other nodes in the graph:

$$T_s(x_s) \quad = \quad \max_{\{\mathbf{x}' \mid x'_s = x_s\}} p(\mathbf{x}'; \theta(\mathcal{T})) \tag{6}$$

That is, $T_s(x_s)$ is the probability of the most likely configuration under the constraint $x'_s = x_s$. For each edge (s,t), the joint pairwise max-marginal is defined in an analogous manner as the maximum $T_{st}(x_s, x_t) = \max_{\{\mathbf{x}' \mid (x'_s, x'_t) = (x_s, x_t)\}} p(\mathbf{x}'; \theta(\mathcal{T}))$. An important fact [3] is that any tree-structured distribution can be factorized in terms of its max-marginals as follows:

$$p(\mathbf{x}; \theta(\mathcal{T})) \quad \propto \quad \prod_{s \in V} T_s(x_s) \prod_{(s,t) \in E(\mathcal{T})} \frac{T_{st}(x_s, x_t)}{T_s(x_s) T_t(x_t)} \tag{7}$$

One interpretation of the standard max-product algorithm for trees, as shown in our related work [10], is as computing this alternative factorization of the distribution.

Now suppose that for each node $s \in V$, the following uniqueness condition holds:
Uniqueness Condition: *For each $s \in V$, the max-marginal T_s has a unique optimum x^*_s.*
In this case, it can be shown [see 10] that the vector $\mathbf{x}^* = \{x^*_s \mid s \in V\}$ is the MAP configuration for the tree-structured distribution.

3.1 Tree-reweighted max-product

One formulation of the tree-reweighted max-product (TRMP) method is as a message-passing algorithm, with fixed points that specify a collection of tree exponential parameters $\Theta^* = \{\theta^*(\mathcal{T})\}$ that satisfy the admissibility condition. The defining feature of Θ^* is that the associated tree distributions $p(\mathbf{x}; \theta^*(\mathcal{T}))$ all share a common set $\mathbf{T}^* = \{T^*_s, T^*_{st}\}$ of max-marginals. In particular, for a given tree $\mathcal{T}$ with edge set $E(\mathcal{T})$, the distribution $p(\mathbf{x}; \theta^*(\mathcal{T}))$ is specified compactly by the subcollection $\Pi^{\mathcal{T}}(\mathbf{T}^*) \triangleq \{T^*_s \mid s \in V\} \cup \{T^*_{st} \mid (s,t) \in E(\mathcal{T})\}$ as follows:

$$p(\mathbf{x}; \theta^*(\mathcal{T})) \equiv p^{\mathcal{T}}(\mathbf{x}; \mathbf{T}^*) \ \triangleq \ \kappa \prod_{s \in V} T^*_s(x_s) \prod_{(s,t) \in E(\mathcal{T})} \frac{T^*_{st}(x_s, x_t)}{T^*_s(x_s) T^*_t(x_t)} \tag{8}$$

where κ is a constant[3] independent of $\mathbf{x}$. As long as $\mathbf{T}^*$ satisfies the Uniqueness Condition, the configuration $\mathbf{x}^* = \{x^*_s \mid s \in V\}$ must be the MAP configuration for each tree-structured distribution $p(\mathbf{x}; \theta^*(\mathcal{T}))$. This mutual agreement on trees, in conjunction with the admissibility of Θ^*, implies that $\mathbf{x}^*$ is also the MAP configuration for $p(\mathbf{x}; \widehat{\theta})$.

For each valid $\vec{\boldsymbol{\mu}}$ (i.e., satisfying $\mu_e > 0$ for all $e \in E$), we define a TRMP algorithm designed to find the requisite set $\mathbf{T}^*$ of max-marginals via a sequence of message-passing operations. For each edge $(s,t) \in E$, let $M_{ts}(x_s)$ be the message passed from node t to node s. It is a vector of length m_s, with one element for each state $j \in \mathcal{X}_s$. We use $\phi_s(x_s; \widehat{\theta})$ as

[3]We use this notation throughout the paper, where the value of κ may change from line to line.

a shorthand for $\sum_j \widehat{\theta}_{s;j}\phi_{s;j}(x_s)$, with the quantity $\phi_{st}(x_s, x_t; \widehat{\theta}_{st})$ similarly defined. With this notation, we use the messages $\mathbf{M} = \{M_{st}\}$ to specify a set of functions $\mathbf{T} = \{T_s, T_{st}\}$ as follows:

$$T_s(x_s) \;=\; \kappa \, \exp\big(\phi_s(x_s; \widehat{\theta}_s)\big) \prod_{v \in \Gamma(s)} \big[M_{vs}(x_s)\big]^{\mu_{vs}} \tag{9a}$$

$$T_{st}(x_s, x_t) \;=\; \kappa \, \varphi_{st}(x_s, x_t; \widehat{\theta}) \, \frac{\prod_{v \in \Gamma(s)\backslash t} \big[M_{vs}(x_s)\big]^{\mu_{vs}}}{\big[M_{ts}(x_s)\big]^{(1-\mu_{ts})}} \, \frac{\prod_{v \in \Gamma(t)\backslash s} \big[M_{vt}(x_t)\big]^{\mu_{vt}}}{\big[M_{st}(x_t)\big]^{(1-\mu_{st})}} \tag{9b}$$

where $\varphi_{st}(x_s, x_t; \widehat{\theta}) = \exp\big[\frac{1}{\mu_{st}}\phi_{st}(x_s, x_t; \widehat{\theta}) + \phi(x_s; \widehat{\theta}_s) + \phi(x_t; \widehat{\theta}_t)\big]$. The scalars μ_{st} in equations (9a) and (9b) are the edge appearance probabilities associated with the spanning tree distribution $\vec{\boldsymbol{\mu}}$, as defined in Section 1.2.

For each tree $\mathcal{T}$, the subcollection $\Pi^{\mathcal{T}}(\mathbf{T})$ can be used to define a tree-structured distribution $p^{\mathcal{T}}(\mathbf{x}; \mathbf{T})$, in a manner analogous to equation (8). By expanding the expectation $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\log p^{\mathcal{T}}(\mathbf{x}; \mathbf{T})]$ and making use of the definitions of T_s and T_{st}, we can prove:

Lemma 1 (Admissibility). *Given any collection $\{T_s, T_{st}\}$ defined by a set of messages $\mathbf{M}$ as in equations* (9a) *and* (9b)*, the convex combination $\sum_{\mathcal{T}} \mu(\mathcal{T}) \log p^{\mathcal{T}}(\mathbf{x}; \mathbf{T})$ is equivalent to $\log p(\mathbf{x}; \widehat{\theta})$ up to an additive constant.*

We now need to ensure that $\mathbf{T} = \{T_s, T_{st}\}$ are a consistent set of max-marginals for each tree-distribution $p^{\mathcal{T}}(\mathbf{x}; \mathbf{T})$. It is sufficient [3, 10] to impose, for each edge (s,t), the local *edgewise consistency* condition

$$\max_{x_t' \in \mathcal{X}_t} T_{st}(x_s, x_t') \;=\; \kappa \, T_s(x_s) \tag{10}$$

In order to do so, we update the messages in the following manner:

Algorithm 1 (Tree reweighted max-product).

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

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

$$M^{n+1}_{ts}(x_s) = \kappa \, \max_{x_t' \in \mathcal{X}_t} \Big\{ \exp\Big(\frac{1}{\mu_{st}}\phi_{st}(x_s, x_t'; \widehat{\theta}_{st}) + \phi_t(x_t'; \widehat{\theta}_t)\Big) \frac{\prod_{v \in \Gamma(t)\backslash s} \big[M^n_{vt}(x_t')\big]^{\mu_{vt}}}{\big[M^n_{st}(x_t')\big]^{(1-\mu_{st})}} \Big\} \tag{11}$$

Using the definitions of T^*_s and T^*_{st}, as well as the message update equation (11), we can show:

Lemma 2 (Edgewise consistency). *Let $\mathbf{M}^*$ be a fixed point of the message update equation* (11)*, and let $\mathbf{T}^* = \{T^*_s, T^*_{st}\}$ be defined via $\mathbf{M}^*$ as in equations* (9a) *and* (9b) *respectively. Then the edgewise consistency condition of equation* (10) *is satisfied.*

The message update equation (11), while similar to both the standard [7, 10] and attenuated [8] max-product updates, differs in some important ways. If G is actually a tree, then we must have $\mu_{st} = 1$ for every edge $(s,t) \in E$, in which case equation (11) is precisely equivalent to the ordinary max-product update. However, if G has cycles, then it is impossible to have

$\mu_{st} = 1$ for every edge $(s,t) \in E$, so that the updates in equation (11) differ in three critical ways. First, the weight $\widehat{\theta}_{st}$ on the potential function ϕ_{st} is scaled by the (inverse of the) edge appearance probability $1/\mu_{st} \geq 1$. Secondly, for each neighbor $v \in \Gamma(t)\backslash s$, the incoming message M_{vt} is scaled by the corresponding edge appearance probability $\mu_{vt} \leq 1$. Lastly, the update of message M_{ts} — that is, from t to s along edge (s,t) — depends on the *reverse direction* message M_{st} from s to t along the same edge. Despite these features, the messages can still be updated in a synchronous manner, as in ordinary max-product [7, 10]. As noted earlier, it is also possible to perform reparameterization updates on trees, analogous to those described elsewhere [10] for standard max-product.

3.2 Analysis of fixed points

In related work [10], we have proved the existence of fixed points for the ordinary max-product algorithm for positive compatibility functions on an arbitrary graph. The same proof can be adapted to show that any TRMP algorithm also has at least one fixed point $\mathbf{M}^*$ under these conditions. Any such fixed point $\mathbf{M}^*$ defines a set of pseudo-max-marginals $\mathbf{T}^*$ via equations (9a) and (9b). We now formalize the key property of such a collection $\mathbf{T}^*$, one which is guaranteed by design of the algorithm:

Theorem 1 (Exact MAP). *If* $\mathbf{T}^*$ *satisfies the Uniqueness Condition, then the configuration* $\mathbf{x}^*$ *with elements* $x_s^* = \arg\max_{x_s' \in \mathcal{X}_s} T_s^*(x_s')$ *is a MAP configuration for* $p(\mathbf{x}; \widehat{\theta})$.

Proof. From Lemma 1, the collection $\Theta^* = \{\theta^*(\mathcal{T})\}$ of tree-structured distributions specified by $\mathbf{T}^*$ (as in equation (8)) is admissible. Lemma 2 implies that the max-marginals $\{T_s^*, T_{st}^*\}$ are edgewise consistent. The Uniqueness Condition guarantees that the tree-structured problems all share the common optimum $\mathbf{x}^*$, which by Proposition 1 must be MAP optimal for $p(\mathbf{x}; \widehat{\theta})$. □

In all of our experiments so far, the message updates of equation (11), if suitably relaxed, have always converged. Here a relaxed update means taking an α-step towards the new (log) message, where $\alpha \in (0, 1]$ is a step-size parameter. To date, we have not proved that they will always converge if relaxed. However, rather than convergence problems, the breakdown of the algorithm appears to stem primarily from failure of the Uniqueness Condition, as we discuss in Section 5.

4 Lagrangian duality and tree relaxation

In this section, we approach the problem of minimizing the upper bound, formulated as in equation (5), via its Lagrangian dual. This dual is a linear program (LP), one which turns out to have a natural interpretation as a *tree relaxation*.

4.1 Linear program over the marginal polytope

Before calculating the dual, we make use of ideas from integer programming [e.g., 2] in order to reformulate the integer program $\mathcal{F}(\theta) = \max_{\mathbf{x}} J(\mathbf{x}; \theta)$ as a linear program. In order to do so, it is useful to consider marginal distributions over subsets of random variables in the full collection $\mathbf{x}$. Given that the collection of potentials $\boldsymbol{\phi}$ consists of functions at single nodes and on edges, of particular relevance are single node marginals over x_s for $s \in V$ and joint pairwise marginals over (x_s, x_t) for $(s,t) \in E$. For a given probability distribution $p(\mathbf{x})$, the associated *single node marginal* P_s at each node s is a vector with $m_s = |\mathcal{X}_s|$ elements, where the j^{th} element is defined as $P_{s;j} = \sum_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}) \delta(x_s = j)$. In a similar way, we define the *joint marginal* distribution $P_{st} = \{P_{st;jk} \mid j \in \mathcal{X}_s,\ k \in \mathcal{X}_t\ \}$ over x_s and x_t, where element (jk) is given by $P_{st;jk} = \sum_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}) \delta(x_s = j) \delta(x_t = k)$.

Let $\mathbf{P} = \{P_s \mid s \in V\ \} \cup \{P_{st} \mid (s,t) \in E\}$ denote the full collection of single node and joint pairwise marginals. Of interest to us is the set of all marginal vectors $\mathbf{P}$ that arise, as in the definitions above, from some underlying distribution $p(\mathbf{x})$. We use $\mathrm{MARG}(G)$ to denote the set of all such *realizable* marginal vectors. It can be seen that this *marginal polytope* is a linear polytope with elements between 0 and 1. It is a useful object, in that it gives rise to an alternative linear programming formulation of the integer program defining $\mathcal{F}(\theta)$:

Lemma 3. *The function $\mathcal{F}(\theta)$ has the alternative representation:*

$$\mathcal{F}(\theta) \quad = \max_{\mathbf{P} \in \mathrm{MARG}(G)} \mathbf{P} \cdot \theta = \max_{\mathbf{P} \in \mathrm{MARG}(G)} \Big\{ \sum_{s \in V} \sum_{j} P_{s;j} \theta_{s;j} + \sum_{(s,t) \in E} \sum_{(j,k)} P_{st;jk} \theta_{st;jk} \Big\} \quad (12)$$

Proof. Any maximization over $\mathbf{x} \in \mathcal{X}^N$ can be rewritten as an equivalent maximization over distributions $p(\mathbf{x})$ — that is, $\max_{\mathbf{x} \in \mathcal{X}^N} J(\mathbf{x}; \theta) = \max_p \Big\{ \sum_{\mathbf{x} \in \mathcal{X}^N} p(\mathbf{x}) J(\mathbf{x}; \theta) \Big\}$. Equation (12) then follows from the linearity of expectation. □

4.2 Tree relaxation

In general, the complexity of solving a LP depends polynomially on (among other factors) the number of constraints needed to characterize the constraint set [see, e.g., 2]. For the LP of equation (12), this constraint set is the marginal polytope $\mathrm{MARG}(G)$, which for a general graph with cycles, is defined by a number of inequalities that is exponential in graph size [4].

The difficulty of exactly characterizing $\mathrm{MARG}(G)$ motivates the idea of imposing a partial set of restrictions, thereby obtaining an outer bound. Since the elements of $\mathbf{P}$ are marginal probabilities, each single node marginal probability must satisfy the *normalization constraint* $\sum_{j \in \mathcal{X}_s} P_{s;j} = 1$. In addition, for each joint marginal P_{st}, we enforce the *marginalization* constraint $\sum_{j \in \mathcal{X}_s} P_{st;jk} = P_{t;k}$ for each $k \in \mathcal{X}_t$. Let $\mathrm{TREE}(G)$ denote the collection of all vectors $\mathbf{Q} = \{Q_s, Q_{st}\}$ with non-negative real-valued elements that satisfy these marginalization and normalization constraints:

$$\mathrm{TREE}(G) \quad = \quad \{\ \mathbf{Q} \ \geq\ \mathbf{0} \mid \sum_{j \in \mathcal{X}_s} Q_{s;j} = 1,\ \ \sum_{j \in \mathcal{X}_s} Q_{st;jk} = Q_{t;k}\ \}. \quad (13)$$

Our choice of notation is motivated by the fact, which follows from the junction tree theorem [3], that $\mathrm{TREE}(\mathcal{T}) = \mathrm{MARG}(\mathcal{T})$ for any tree $\mathcal{T}$. For a general graph with cycles, however, $\mathrm{TREE}(G)$ is a strict superset of $\mathrm{MARG}(G)$.

For this reason, we call any $\mathbf{Q}$ a tree-consistent *pseudomarginal* vector. A natural relaxation of the linear program in equation (12), then, is to replace the constraint set $\mathrm{MARG}(G)$ by the superset $\mathrm{TREE}(G)$. It turns out that the Lagrangian dual of problem (5) is precisely this so-called *tree relaxation*:

Proposition 2 (Lagrangian dual). *The Lagrangian dual to problem* (5) *is given by the tree relaxation:*

$$\mathcal{F}(\widehat{\theta}) \quad \leq \max_{\mathbf{Q} \in \mathrm{TREE}(G)} \mathbf{Q} \cdot \widehat{\theta} \ \triangleq \max_{\mathbf{Q} \in \mathrm{TREE}(G)} \Big\{ \sum_{s \in V} \sum_{j} Q_{s;j} \widehat{\theta}_{s;j} + \sum_{(s,t) \in E} \sum_{(j,k)} Q_{st;jk} \widehat{\theta}_{st;jk} \Big\} \quad (14)$$

Proof. Let $\mathbf{Q}$ be a vector of Lagrange multipliers corresponding to the admissibility constraints $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] = \widehat{\theta}$. With this notation, we form the Lagrangian associated with problem (5):

$$\begin{aligned}\mathcal{L}(\Theta, \mathbf{Q}; \vec{\boldsymbol{\mu}}; \widehat{\theta}) &= \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\mathcal{F}(\theta(\mathcal{T}))] + \mathbf{Q} \cdot \big[\widehat{\theta} - \sum_{\mathcal{T}} \mu(\mathcal{T})\theta(\mathcal{T})\big] \\ &= \sum_{\mathcal{T}} \mu(\mathcal{T})\big[\mathcal{F}(\theta(\mathcal{T})) - \theta(\mathcal{T}) \cdot \Pi^{\mathcal{T}}(\mathbf{Q})\big] + \mathbf{Q} \cdot \widehat{\theta} \qquad (15)\end{aligned}$$

Here $\Pi^{\mathcal{T}}(\mathbf{Q})$ denotes the subcollection[4] $\{Q_s \mid s \in V\} \cup \{Q_{st} \mid (s,t) \in E(\mathcal{T})\}$ of Lagrange multipliers corresponding to tree $\mathcal{T}$. The dual function [1] is defined by taking, for a fixed Lagrange multiplier vector $\mathbf{Q}$, the infimum of the Lagrangian as a function of Θ; its domain corresponds to those $\mathbf{Q}$ for which this infimum is greater than $-\infty$.

If for some $\mathcal{T} \in \operatorname{supp}(\vec{\boldsymbol{\mu}})$ the vector $\Pi^{\mathcal{T}}(\mathbf{Q})$ does not belong to the linear polytope $\operatorname{MARG}(\mathcal{T})$, then there must exist some vector $\gamma(\mathcal{T})$ and constant β such that (i) $\gamma(\mathcal{T}) \cdot \mathbf{P}(\mathcal{T}) \leq \beta$ for all $\mathbf{P}(\mathcal{T}) \in \operatorname{MARG}(\mathcal{T})$ (with equality attained for at least one $\mathbf{P}(\mathcal{T})$), and (ii) $\gamma(\mathcal{T}) \cdot \Pi^{\mathcal{T}}(\mathbf{Q}) > \beta$. Using $\gamma(\mathcal{T})$ as a valid choice for $\theta(\mathcal{T})$, then we have

$$\mathcal{F}(\gamma(\mathcal{T})) - \gamma(\mathcal{T}) \cdot \Pi^{\mathcal{T}}(\mathbf{Q}) = \beta - \gamma(\mathcal{T}) \cdot \Pi^{\mathcal{T}}(\mathbf{Q}) \; < \; 0$$

By scaling both $\gamma(\mathcal{T})$ and β by a positive number, we can send this term to negative infinity. Thus, the infimum in the Lagrangian will be $-\infty$ unless $\Pi^{\mathcal{T}}(\mathbf{Q}) \in \operatorname{MARG}(\mathcal{T})$ for all trees $\mathcal{T} \in \operatorname{supp}(\vec{\boldsymbol{\mu}})$. Since each edge belongs to at least one tree in $\operatorname{supp}(\vec{\boldsymbol{\mu}})$, it can be seen that these constraints imply that the domain of the dual function is the set $\operatorname{TREE}(G)$.

Now for any $\Pi^{\mathcal{T}}(\mathbf{Q}) \in \operatorname{MARG}(\mathcal{T})$, the definition of $\mathcal{F}(\theta(\mathcal{T}))$ implies that the difference $\mathcal{F}(\theta(\mathcal{T})) - \theta(\mathcal{T}) \cdot \Pi^{\mathcal{T}}(\mathbf{Q})$ is always non-negative, and is equal to zero when $\Pi^{\mathcal{T}}(\mathbf{Q})$ attains the maximum defining $\mathcal{F}(\theta(\mathcal{T}))$. Therefore, taking the infimum of the Lagrangian yields the expression $\mathbf{Q} \cdot \widehat{\theta}$, as in the tree relaxation of equation (14).

□

Remark: Suppose that an optimal solution $\mathbf{Q}^*$ to the dual problem in equation (14) has all integral (0 or 1) components; any such $\mathbf{Q}^*$ correspond to the marginal vector associated with a delta distribution $p(\mathbf{x}) = \delta(\mathbf{x} = \mathbf{x}^*)$, so that in fact, it belongs to the marginal polytope $\operatorname{MARG}(G)$. In this case, the upper bound is tight, and the configuration $\mathbf{x}^*$ is MAP-optimal.

5 Conditions governing tree agreement

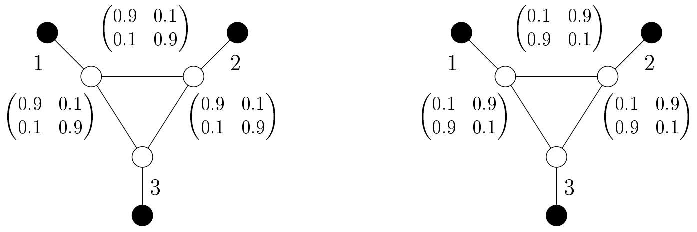
In this section, we consider the conditions that govern whether or not the optimal bound is tight, and the MAP configuration can be obtained. In this context, the tree-reweighted max-product algorithm and the LP tree relaxation approach offer complementary perspectives. Difficulties can arise with the message-passing approach when a fixed point $\mathbf{T}^*$ (defined by a message fixed point $\mathbf{M}^*$) fails to satisfy the Uniqueness Condition. In this case, it may no longer be possible to specify a configuration $\mathbf{x}^*$ that is optimal for each tree-structured problem. More specifically, it can be shown that a configuration $\mathbf{x}^*$ is optimal on each tree if and only if:
Node optimality: *The element x^*_s must achieve $\max_{x'_s} T^*_s(x'_s)$ for every $s \in V$.*
Edge optimality: *The pair (x^*_s, x^*_t) must achieve $\max_{(x'_s, x'_t)} T^*_{st}(x'_s, x'_t)$ for all $(s,t) \in E$.*
When the Uniqueness Condition fails, then depending on the nature of the fixed point $\mathbf{T}^*$, it may or may not be possible to satisfy these optimality conditions for every node and edge. Example 2 illustrates both of these two possibilities.

[4]The dot product $\theta(\mathcal{T}) \cdot \Pi^{\mathcal{T}}(\mathbf{Q})$ in equation (15) involves a minor abuse of notation, since strictly speaking, $\theta(\mathcal{T})$ is the same length as $\widehat{\theta}$ (with $\theta(\mathcal{T})_\alpha = 0$ for indices α not in tree $\mathcal{T}$), whereas $\Pi^{\mathcal{T}}(\mathbf{Q})$ is a shorter vector.

The optimum of a linear program, such as that of equation (14), is always attained at an extreme point of the constraint set [2]. For the tree relaxed polytope $\mathrm{TREE}(G)$, a subset of the extreme points have integral components (0 or 1); each such extreme point corresponds to the marginal vector $\mathbf{P}^*$ defined by a delta distribution $p(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{x}^*)$ that places all its mass on a single configuration. Whenever the optimum (14) is attained at such an integral point, the bound is tight and the underlying configuration $\mathbf{x}^*$ is MAP-optimal. If, on the other hand, the optimum is attained only at a fractional extreme point of $\mathrm{TREE}(G)$, as illustrated in Example 2, then the bound is not tight, nor can the MAP configuration be determined.

Example 2. Consider the 3-node cycle illustrated in Figure 2. We define a distribution $p(\mathbf{x}; \widehat{\theta})$ on the binary vector $\mathbf{x} \in \{0, 1\}^3$ in an indirect manner, by first defining a set of pseudo-max-marginals $\mathbf{T}^*$ in panel (a). Observe that this construction ensures that for any $\beta \in [0, 1]$, the collection $\mathbf{T}^*$ satisfies the edgewise consistency condition of Lemma 2; however, the Uniqueness Condition fails at every node. For each of the three spanning trees of this graph, the collection $\mathbf{T}^*$ defines a tree-structured distribution $p^{\mathcal{T}}(\mathbf{x}; \mathbf{T}^*)$ as in equation (8). We use these tree-structured distributions to define the underlying distribution on the single cycle via $\log p(\mathbf{x}; \widehat{\theta}) \triangleq \mathbb{E}_{\vec{\mu}}[\log p^{\mathcal{T}}(\mathbf{x}; \mathbf{T}^*)] + C$, where the distribution $\vec{\mu}$ places weight $1/3$ on each tree.

$$
\begin{aligned}
T_s^* &= \begin{bmatrix} 1 & 1 \end{bmatrix} \\
T_{st}^* &= \begin{bmatrix} \beta & 1-\beta \\ 1-\beta & \beta \end{bmatrix}
\end{aligned}
$$



(a) (b) (c)

Figure 2. Failure of tree agreement. (a) Specification of pseudo-max-marginals $\mathbf{T}^*$. Underlying problem is $\log p(\mathbf{x}; \widehat{\theta}) = \mathbb{E}_{\vec{\mu}}[\log p^{\mathcal{T}}(\mathbf{x}; \mathbf{T}^*)]$. (b) Case $\beta > 0.5$ for which both $[0\ 0\ 0]$ and $[1\ 1\ 1]$ are node and edgewise optimal. (c) Case $\beta < 0.5$ for which there are no configurations that are node and edgewise optimal on the full graph.

In the case $\beta > 0.5$, illustrated in panel (b), it can be seen that two configurations — namely $\mathbf{0} = [0\ 0\ 0]$ and $\mathbf{1} = [1\ 1\ 1]$ — satisfy the node and edgewise optimality conditions. Therefore, each of these configurations are MAP assignments for $p(\mathbf{x}; \widehat{\theta})$. With reference to the tree-relaxed LP, each of the marginal vectors $\mathbf{P}$ corresponding to the delta functions $p(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{0})$ and $p(\mathbf{x}) = \delta(\mathbf{x} - \mathbf{1})$ are optimal for the LP in equation (14). On the other hand, when $\beta < 0.5$, as illustrated in panel (c), any configuration $\mathbf{x}^*$ that is edgewise optimal for all three edges must satisfy $x_s^* \neq x_t^*$ for all $(s, t) \in E$. It is clearly impossible to satisfy this condition for all three edges on the cycle, so that the fixed point $\mathbf{T}^*$ cannot be used to specify a MAP assignment. In the LP context, the optimum of equation (14) is attained at the pseudomarginal with fractional elements $Q_s = [0.5 \quad 0.5]'$ for all $s \in V$, and $Q_{st} = [0 \quad 0.5;\ 0.5 \quad 0]$ for all $(s, t) \in E$. Here the $(j, k)^{th}$ element of the matrix Q_{st} represents the marginal value $Q_{st;jk}$. It can be seen that this pseudomarginal $\mathbf{Q}$ defined in this way belongs to $\mathrm{TREE}(G)$, but it does not belong to $\mathrm{MARG}(G)$. As a consequence, the upper bound is loose, and moreover the MAP configuration cannot be determined.

Of course, Example 2 has been deliberately constructed so as to break down the relaxation. It should be noted that as shown in our related work [10], the standard max-product algorithm can also break down when the Uniqueness Condition is not satisfied.

6 Discussion

In this paper, we demonstrated the utility of convex combinations of tree-structured distributions in upper bounding the log probability of the MAP configuration. In addition, we proposed two approaches for obtaining optimal upper bounds of this form. First of all, we developed a family of *tree-reweighted* max-product (TRMP) algorithms that reparameterize a collection of tree-structured distributions in terms of a common set of max-marginals. If these max-marginals satisfy a key uniqueness condition, then the upper bound is tight, and the MAP configuration can be obtained. It remains to develop a complete understanding of the convergence behavior of this algorithm, and its relation to the LP relaxation. Secondly, we showed how to optimize the upper bounds via their Lagrangian dual, which turns out to be a *tree-relaxed* linear program (LP). In fact, this LP can also be obtained by examining the zero temperature limit of the convexified Bethe free energy analyzed in our previous work [9]. We can identify a class of binary problems with pairwise interactions for which the relaxation is guaranteed to be tight; an open problem is to characterize this class more generally. A potentially valuable link is that between the tree relaxation described here and a recently proposed flow-based LP relaxation for turbo decoding, as discussed in other work [5, 6].

Finally, the analysis and upper bounds of this paper can be extended in a straightforward manner to hypertrees of higher width. (A tree has width one, whereas hypertrees of higher width are a natural generalization of trees [e.g., 3, 10].) In this context, hypertree-reweighted forms of generalized max-product updates [see 10] can be used, yielding upper bounds that (when tight) again specify exact MAP configurations. On the linear programming side, this leads to a hierarchy of LP relaxations for the original integer program based on hypertree constraints.

References

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

[2] D. Bertsimas and J. Tsitsikilis. *Introduction to linear optimization*. Athena Scientific, Belmont, MA, 1997.

[3] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probablistic Networks and Expert Systems*. Statistics for Engineering and Information Science. Springer-Verlag, 1999.

[4] M. Deza and M. Laurent. *Geometry of cuts and metric embeddings*. Springer-Verlag, New York, 1997.

[5] J. Feldman and D. Karger. Decoding turbo-like codes in polynomial time with provably good error-correcting performing via linear programming. In *Symp. Found. Comp. Science (FOCS)*, November 2002. To appear.

[6] J. Feldman, D. R. Karger, and M. J. Wainwright. Linear programming-based decoding of turbo codes and its relation to iterative approaches. In *Proc. Allerton Conf. Communication, Control and Computing*, October 2002. To appear.

[7] W. T. Freeman and Y. Weiss. On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. *IEEE Trans. Info. Theory*, 47:736–744, 2001.

[8] B. J. Frey and R. Koetter. Exact inference using the attenuated max-product algorithm. In *Advanced mean field methods: Theory and Practice*. MIT Press, 2000.

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

[10] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree consistency and bounds on the max-product algorithm and its generalizations. LIDS Tech. report, MIT; Available online at http://ssg.mit.edu/group/mjwain/mjwain.shtml, July 2002.