Exact MAP estimates by (hyper)tree agreement

Martin J. Wainwright, Tommi S. Jaakkola, and Alan S. Willsky
Department of Electrical Engineering and Computer Science
Massachusetts Institute of Technology
Cambridge, MA 02139
{*mjwain,tommi,willsky*}*@mit.edu*

Abstract

We describe a method that computes (provably) exact MAP estimates for a subclass of problems on graphs with cycles. The basic idea is to represent the original problem on the graph with cycles as a convex combination of tree-structured problems. A convexity argument then guarantees that the optimal value of the original problem (i.e., the log probability of the MAP assignment) is upper bounded by 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 the MAP configuration for the original problem. Next we develop a tree-reweighted max-product algorithm for attempting to find convex combinations of tree-structured problems that share a common optimum. We prove that these message-passing updates always have a fixed point, and give necessary and sufficient conditions for the fixed point to yield the exact MAP estimate. An attractive feature of our analysis is that it generalizes naturally to convex combinations of hypertree-structured distributions.

1 Introduction

Integer programming problems arise in various fields, including machine learning, statistical physics, communication, and error-correcting coding. In many cases, such problems can be formulated in terms of *graphical models* [1, 2], in which the cost function corresponds to a graph-structured probability distribution, and the problem of interest is to find the *maximum a posteriori* (MAP) configuration.

In previous work [3], 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 apply similar ideas to upper bound the log probability of the MAP configuration. As we show, this upper bound is met with equality whenever there is a configuration that is optimal for all trees, in which case it must also be a MAP configuration for the original problem. The work described here also makes connections with the max-product algorithm [e.g., 4, 5, 6], a well-known method for attempting to compute the MAP configuration, one which is exact for trees but approximate for graphs with cycles. In the context of coding problems, Frey and Koetter [5] developed an attenuated version of max-product, which is guaranteed to find the MAP codeword if it converges. One contribution of this paper is

to develop a *tree-reweighted max-product algorithm* that attempts to find a collection of tree-structured problems that share a common optimum. This algorithm, though bearing conceptual similarities to attenuated max-product [5], differs in certain key ways.

The remainder of this paper is organized as follows. The next two subsections 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 tight (i.e., met with equality). In Section 3, we develop the tree-reweighted max-product for finding a convex combination of trees that yields a tight bound. We prove that the algorithm always has at least one fixed point; moreover, under a key uniqueness assumption, the configuration specified by a fixed point must be MAP optimal. We also illustrate how the algorithm may fail if the uniqueness condition is not satisfied. We conclude in Section 4 with pointers to related work, and extensions of the current work.

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. It is convenient to allow these weights to take values in the extended reals $\mathbb{R}^* = \mathbb{R} \cup \{-\infty\}$. The exponential family determined by $\boldsymbol{\phi}$ is the collection $p(\mathbf{x}; \theta) \propto \exp\{\sum_{\alpha \in \mathcal{I}} \theta_\alpha \phi_\alpha(\mathbf{x})\}$ 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 $\boldsymbol{\phi} = \{x_s \mid s \in V\} \cup \{x_s x_t \mid (s,t) \in E\}$. Here the index set $\mathcal{I} = V \cup E$. In most of our analysis, we use an *overcomplete* representation, in which there are linear dependencies among the potentials $\{\phi_\alpha\}$. In particular, we use indicator functions as potentials:

$$\phi_{s;j}(x_s) = \delta(x_s = j), \quad s \in V;\ j \in \mathcal{X}_s \tag{1a}$$

$$\phi_{st;jk}(x_s, x_t) = \delta(x_s = j, x_t = k),\ (s,t) \in E;\ (j,k) \in \mathcal{X}_s \times \mathcal{X}_t \tag{1b}$$

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

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

Note that the function $\mathcal{F}(\theta)$ is the maximum of a collection of linear functions, and hence is convex [7] as a function 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. 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 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. The notation $\theta(\mathcal{T})$ specifies those subelements of Θ corresponding to spanning tree $\mathcal{T}$.

In order to define a convex combination, we require a probability distribution $\vec{\boldsymbol{\mu}}$ over the set of spanning trees — that is, a vector $\vec{\boldsymbol{\mu}} \triangleq \{\, \mu(\mathcal{T}),\ \mathcal{T} \in \mathfrak{T} \mid \mu(\mathcal{T}) \geq 0 \,\}$ such that $\sum_{\mathcal{T} \in \mathfrak{T}} \mu(\mathcal{T}) = 1$. For any distribution $\vec{\boldsymbol{\mu}}$, we define its *support* 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 $\vec{\boldsymbol{\mu}}$. We let $\boldsymbol{\mu_e} = \{\mu_e \mid e \in E\}$ represent a vector of these *edge appearance probabilities*. It can be shown [3] that these edge appearance vectors must belong to the so-called *spanning tree polytope*.

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 $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})]$. Of particular importance are collections of exponential parameters Θ for which there exists a convex combination that is equal to $\widehat{\theta}$ (where $\mathcal{F}(\widehat{\theta})$ is the original problem of interest). Accordingly, we define the set $\mathcal{A}(\widehat{\theta}) \triangleq \left\{(\Theta; \vec{\boldsymbol{\mu}}) \;\middle|\; \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] = \widehat{\theta}\right\}$. As long as every edge appears in at least one tree $\mathcal{T} \in \operatorname{supp}(\vec{\boldsymbol{\mu}})$ (i.e., $\mu_e > 0$ for all $e \in E$), 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 distribution ($\mathcal{X}_s = \{0, 1\}$ for all nodes $s \in V$) defined by a single cycle on 4 nodes. Consider a target 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 tractable class consists of

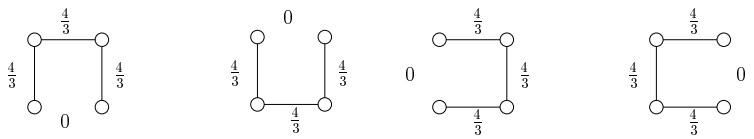


Figure 1. A convex combination of four distributions $p(\mathbf{x}; \theta(\mathcal{T}_i))$, each defined by a spanning tree $\mathcal{T}_i$, is used to approximate the target distribution $p(\mathbf{x}; \widehat{\theta})$ on the single-cycle graph.

the four possible spanning trees $\mathfrak{T} = \{\mathcal{T}_i \mid i = 1, \ldots, 4\}$ on a single cycle on four nodes, illustrated in Figure 1. We define a set of associated exponential parameters $\Theta = \{\theta(\mathcal{T}_i)\}$ as follows:

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

Finally, we choose $\mu(\mathcal{T}_i) = 1/4$ for all $\mathcal{T}_i \in \mathfrak{T}$. With this uniform distribution over trees, we have $\mu_e = 3/4$ for each edge, and moreover, $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\theta(\mathcal{T})] = \widehat{\theta}$. That is, the specified pair $(\Theta; \vec{\boldsymbol{\mu}})$ belongs to $\mathcal{A}(\widehat{\theta})$.

2 Optimal upper bounds

With the set-up of the previous section, the basic form of the upper bounds follows by applying Jensen's inequality [7]. In particular, for any pair $(\Theta; \vec{\boldsymbol{\mu}}) \in \mathcal{A}(\widehat{\theta})$, we have the upper bound $\mathcal{F}(\widehat{\theta}) \leq \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\mathcal{F}(\theta(\mathcal{T}))]$. The goal of this section is to examine this bound, and understand when it is met with equality. In more explicit terms, the upper bound can be written as:

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

Now suppose that there exists an $\widehat{\mathbf{x}} \in \mathcal{X}^N$ that attains the maximum defining $\mathcal{F}(\theta(\mathcal{T}))$ for each tree $\mathcal{T} \in \mathrm{supp}(\vec{\boldsymbol{\mu}})$. In this case, it is clear that the bound (3) is met with equality. An important implication is that the configuration $\widehat{\mathbf{x}}$ also attains the maximum defining $\mathcal{F}(\widehat{\theta})$, so that it is an *optimal* solution to the original problem.

In fact, as we show below, the converse to this statement also holds. More formally, for any exponential parameter vector $\theta(\mathcal{T})$, let $\mathrm{OPT}(\theta(\mathcal{T}))$ be the collection of configurations $\mathbf{x}$ that attain the maximum defining $\mathcal{F}(\theta(\mathcal{T}))$. That is

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

With this notation, the critical property of $\widehat{\mathbf{x}}$ is that it belongs to the intersection $\mathrm{OPT}(\Theta) \triangleq \cap_{\mathcal{T}} \mathrm{OPT}(\theta(\mathcal{T}))$. We thus have the following result:

Proposition 1 (Tightness of bound). *The bound of equation* (3) *is tight if and only if 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 $\mathcal{F}(\theta(\mathcal{T}))$. (That is, $\widehat{\mathbf{x}} \in \mathrm{OPT}(\Theta)$.)*

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 $\mathcal{F}(\widehat{\theta})$. Similarly, for each tree $\mathcal{T} \in \mathrm{supp}(\vec{\boldsymbol{\mu}})$, let $\mathbf{x}^*(\mathcal{T})$ attain the maximum defining $\mathcal{F}(\theta(\mathcal{T}))$; that is, the configuration $\mathbf{x}^*(\mathcal{T})$ belongs to $\mathrm{OPT}(\theta(\mathcal{T}))$. With this notation, we write the difference of the LHS and the RHS of equation (3) as follows:

$$\begin{aligned}
\mathcal{F}(\widehat{\theta}) - \sum_{\mathcal{T}} \mu(\mathcal{T}) \mathcal{F}(\theta(\mathcal{T})) \quad &= \quad \widehat{\theta} \cdot \boldsymbol{\phi}(\widehat{\mathbf{x}}) - \sum_{\mathcal{T}} \mu(\mathcal{T}) \theta(\mathcal{T}) \cdot \boldsymbol{\phi}(\mathbf{x}^*(\mathcal{T})) \\
&= \quad \sum_{\mathcal{T}} \mu(\mathcal{T}) \theta(\mathcal{T}) \cdot \left[\boldsymbol{\phi}(\widehat{\mathbf{x}}) - \boldsymbol{\phi}(\mathbf{x}^*(\mathcal{T})) \right] \\
&\leq \quad 0
\end{aligned}$$

Now for each $\mathcal{T} \in \mathrm{supp}(\vec{\boldsymbol{\mu}})$, the term $\theta(\mathcal{T}) \cdot \left[\boldsymbol{\phi}(\widehat{\mathbf{x}}) - \boldsymbol{\phi}(\mathbf{x}^*(\mathcal{T})) \right]$ is non-positive, 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 $\mathcal{T} \in \mathrm{supp}(\vec{\boldsymbol{\mu}})$. □

Proposition 1 motivates the following strategy: given a spanning tree distribution $\vec{\boldsymbol{\mu}}$, find a collection of exponential parameters $\Theta^* = \{\theta^*(\mathcal{T})\}$ such that the following holds: (a) Feasibility: 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 tree-optimal configurations is non-empty.

If (for a fixed $\vec{\boldsymbol{\mu}}$) we are able to find a collection Θ^* satisfying these two properties, then Proposition 1 guarantees that all configurations in the (non-empty) intersection $\cap_i \mathrm{OPT}(\theta^*(\mathcal{T}))$ achieve the maximum defining $\mathcal{F}(\widehat{\theta})$. As discussed above, assuming that $\vec{\boldsymbol{\mu}}$ assigns strictly positive probability to every edge in the graph, satisfying the feasibility condition is not difficult. It is the second condition of mutual optimality on all trees that poses the challenge.

3 Mutual agreement via message-passing

We now develop a message-passing algorithm that attemps to find, for a given spanning tree distribution $\vec{\mu}$, a collection $\Theta^* = \{\theta^*(\mathcal{T})\}$ satisfying both of these properties. Interestingly, the method that we develop here is closely related to the ordinary max-product algorithm, but differs in several key ways.

3.1 Max-marginals

The foundation of our development is the well-known fact [1] that any tree-structured distribution $p(\mathbf{x}; \theta(\mathcal{T}))$ can be factored in terms of its max-marginals. In particular, for each node $s \in V$, the corresponding single node max-marginal is defined as follows:

$$T_s(x_s) \quad = \quad \max_{\{\mathbf{x}' \mid x'_s = x_s\}} p(\mathbf{x}'; \theta(\mathcal{T})) \tag{5}$$

That is, for each $x_s \in \mathcal{X}_s$, $T_s(x_s)$ is the maximum probability over the subset of configurations $\mathbf{x}'$ with element x'_s fixed to x_s. For each edge $(s,t) \in E$, the pairwise max-marginal is defined in an analogous fashion:

$$T_{st}(x_s, x_t) \quad = \quad \max_{\{\mathbf{x}' \mid (x'_s, x'_t) = (x_s, x_t)\}} p(\mathbf{x}'; \theta(\mathcal{T})) \tag{6}$$

With these definitions, the max-marginal tree factorization [1] is given by:

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

One interpretation of the ordinary max-product algorithm for trees, as shown in our previous work [6], is as computing this alternative representation.

Suppose moreover that for each node $s \in V$, the following uniqueness condition holds:
Uniqueness Assumption: *The max-marginal T_s has a unique optimum x^*_s.*
In this case, the vector $\mathbf{x}^* = \{x^*_s \mid s \in V\}$ is the MAP configuration for the tree-structured distribution [see 6].

3.2 Tree-reweighted max-product

With this set-up, the goal of tree-reweighted max-product algorithm is to find a particular collection of tree exponential parameters $\Theta^* = \{\theta^*(\mathcal{T})\}$ that satisfy the feasibility 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 log distribution $\log p(\mathbf{x}; \theta^*(\mathcal{T}))$ 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:

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

where κ is a constant[1] independent of $\mathbf{x}$. As long as $\mathbf{T}^*$ satisfies the Uniqueness Assumption, 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 feasibility of Θ^*, implies that $\mathbf{x}^*$ is also the MAP configuration for $p(\mathbf{x}; \widehat{\theta})$.

For each $\vec{\mu}$ such that $\mu_e > 0$ for all $e \in E$, there exists a tree-reweighted max-product 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

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

from node t to node s. It is a vector of length $m(s)$, with one element for each state $j \in \mathcal{X}_s$. We use the messages $\mathbf{M} = \{M_{st}\}$ to specify a set of functions $\mathbf{T} = \{T_s, T_{st}\}$ as follows:

$$T_s(x_s) \quad = \quad \exp\left(\widehat{\theta}_s \phi_s(x_s)\right) \prod_{u \in \mathcal{N}(s)} \left[M_{us}(x_s)\right]^{\mu_{us}} \tag{9a}$$

$$T_{st}(x_s, x_t) \quad = \quad \varphi_{st}(x_s, x_t; \widehat{\theta}) \, \frac{\prod_{u \in \mathcal{N}(s)\backslash t} \left[M_{us}(x_s)\right]^{\mu_{us}}}{\left[M_{ts}(x_s)\right]^{(1-\mu_{st})}} \, \frac{\prod_{u \in \mathcal{N}(t)\backslash s} \left[M_{ut}(x_t)\right]^{\mu_{ut}}}{\left[M_{st}(x_t)\right]^{(1-\mu_{ts})}} \tag{9b}$$

where $\varphi_{st}(x_s, x_t; \widehat{\theta}) = \exp\left(\frac{\widehat{\theta}_{st}}{\mu_{st}} \phi_{st}(x_s, x_t) + \widehat{\theta}_s \phi(x_s) + \widehat{\theta}_t \phi(x_t)\right)$.

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 the following:

Lemma 1 (Feasibility). *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 [1, 6] to impose, for each edge (s, t), the following local consistency condition:

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

In order to achieve this local consistency, we update the messages in the following manner:

Algorithm 1 (Tree reweighted max-product).
For any $\boldsymbol{\mu_e} \in \mathbb{T}(G)$ with $\mu_e > 0$ for all $e \in E$: 1. Initialize the messages $\mathbf{M}^0 = \{M_{st}^0\}$ with arbitrary positive real numbers.

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

$$M_{ts}^{n+1}(x_s) = \max_{x_t' \in \mathcal{X}_t} \left\{ \exp\left(\frac{\widehat{\theta}_{st}}{\mu_{st}} \phi_{st}(x_s, x_t') + \widehat{\theta}_t \phi_t(x_t')\right) \frac{\prod_{u \in \mathcal{N}(t)\backslash s} \left[M_{ut}(x_t)\right]^{\mu_{ut}}}{\left[M_{st}(x_t)\right]^{(1-\mu_{ts})}} \right\} \tag{11}$$

Using the definitions of T_s^* and T_{st}^*, as well as the message update equation (11), the following result can be proved:

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 of equation* (10) *is satisfied.*

The message update equation (11) is closely related to the message updates of ordinary max-product algorithm [4, 6]. Indeed, 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 from ordinary max-product in some key ways. First of all, 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 $u \in \mathcal{N}(t)\backslash s$, the incoming message M_{ut} is scaled by the corresponding edge appearance probability $\mu_{ut} \leq 1$. Third of all, 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. This dependence is in sharp contrast with the ordinary max-product updates.

3.3 Analysis of fixed points

In previous work [6], 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 the tree-reweighted max-sum algorithm also has at least one fixed point $\mathbf{M}^*$. 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 Assumption, 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. For each tree $\mathcal{T}$ with edge set $E(\mathcal{T})$, the fixed point $\mathbf{T}^*$ defines a tree-structured distribution $p(\mathbf{x}; \theta^*(\mathcal{T}))$ via equation (8). By Lemma 2, the elements of $\mathbf{T}$ are edgewise-consistent. Therefore, by the equivalence of edgewise and global consistency for trees, the subcollection $\Pi^{\mathcal{T}}(\mathbf{T}^*) = \{T_s^* \mid s \in V\} \cup \{T_{st}^* \mid (s,t) \in E(\mathcal{T})\}$ are exact max-marginals for the tree-structured distribution $p(\mathbf{x}; \theta^*(\mathcal{T}))$. As a consequence, the configuration $\mathbf{x}^*$ must belong to $\mathrm{OPT}(\theta^*(\mathcal{T}))$ for each tree $\mathcal{T}$. That is, the condition of *mutual agreement* is satisfied. By Lemma 1, the convex combination $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\log p(\mathbf{x}; \theta^*(\mathcal{T}))]$ is equal to $\log p(\mathbf{x}; \widehat{\theta})$, so that *feasibility* is satisfied. By Proposition 1, we conclude that $\mathbf{x}^*$ is a MAP configuration for $p(\mathbf{x}; \widehat{\theta})$. □

3.4 Failures of tree-reweighted max-product

In all of our experiments so far, the message updates of equation (11), if suitably relaxed, have always converged, although typically more slowly than the ordinary max-product (or sum-product) updates.[2] Rather than convergence problems, the breakdown of the algorithm appears to stem primarily from failure of the Uniqueness Assumption. If this assumption is not satisfied, we are no longer guaranteed that the mutual agreement condition is satisfied (i.e., $\mathrm{OPT}(\Theta^*) = \cap_{\mathcal{T}} \mathrm{OPT}(\theta^*(\mathcal{T}))$ may be empty). Indeed, a configuration $\mathbf{x}^*$ belongs to $\mathrm{OPT}(\Theta^*)$ if and only if the following conditions hold:
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 $(s,t) \in E$.*
For a given fixed point $\mathbf{T}^*$ that fails the Uniqueness Assumption, it may or may not be possible to satisfy these conditions, as the following example illustrates.

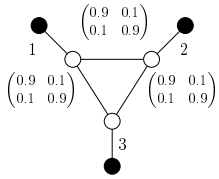
Example 2. Consider the single cycle on three vertices, as illustrated in Figure 2. We define a distribution $p(\mathbf{x}; \widehat{\theta})$ in an indirect manner, by first defining a set of pseudo-max-marginals $\mathbf{T}^*$ in panel (a). Here $\beta \in [0,1]$ is a parameter to be specified. Observe that the symmetry of this construction ensures that $\mathbf{T}^*$ satisfies the edgewise consistency condition (10) for any $\beta \in [0,1]$. 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 define the underlying distribution via $\log p(\mathbf{x}; \widehat{\theta}) \triangleq \mathbb{E}_{\vec{\boldsymbol{\mu}}}[\log p^{\mathcal{T}}(\mathbf{x}; \mathbf{T}^*)]$ where $\vec{\boldsymbol{\mu}}$ is the uniform distribution (weight $(1/3)$ on each tree).

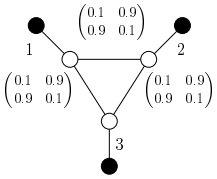
In the case $\beta > 0.5$, illustrated in panel (b), it can be seen that two configurations — namely $[0\ 0\ 0]$ and $[1\ 1\ 1]$ — satisfy the node and edgewise optimality conditions. Therefore, each of these configurations are global maxima for the cost function $\mathbb{E}_{\vec{\boldsymbol{\mu}}}[\log p(\mathbf{x}; \mathbf{T}^*)]$. 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 $(s,t) \in E$. This is clearly impossible, so that the fixed point $\mathbf{T}^*$ cannot be used to specify a MAP assignment.

[2] To date, we have not been able to prove that they will always converge if relaxed.

$$T_s^* = \begin{bmatrix} 1 & 1 \end{bmatrix}$$

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





(a) (b) (c)

Figure 2. Cases where the Uniqueness Assumption fails. (a) Specification of pseudo-max-marginals $\mathbf{T}^*$. Underlying problem is $\log p(\mathbf{x}; \widehat{\theta}) = \mathbb{E}_{\widetilde{\boldsymbol{\mu}}}[\log p^{\mathcal{T}}(\mathbf{x}; \mathbf{T}^*)]$. (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.

4 Discussion

This paper demonstrated the utility of convex combinations of tree-structured distributions in upper bounding the log probability of the MAP configuration. We developed a family of tree-reweighted max-product algorithms for computing optimal upper bounds. In certain cases, the optimal upper bound is met with equality, and hence yields an exact MAP configuration for the original problem on the graph with cycles. An important open question is to characterize the range of problems for which the upper bound is tight. For problems involving a binary-valued random vector, we have isolated a class of problems for which the upper bound is guaranteed to be tight.

We have also investigated the Lagrangian dual associated with the upper bound (3). The dual form of the upper bound entails solving a linear program (LP) that is closely related to the zero temperature limit of the convexified Bethe free energy analyzed in our previous work [3]. We have also established a link between this dual formulation, and a new LP-based turbo decoding method recently proposed by Feldman and Karger [8].

Finally, the analysis and upper bounds of this paper can be extended in a straightforward manner to hypertrees of width ≥ 2. In this context, hypertree-reweighted forms of generalized max-product updates [see 6] can again be used to find optimal upper bounds, which (when they are tight) again yield exact MAP configurations.

References

[1] 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.

[2] M. Jordan. *Learning in graphical models*. MIT Press, Cambridge, MA, 1999.

[3] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. In *UAI*, volume 18, August 2002.

[4] 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.

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

[6] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. Tree consistency and bounds on the performance of the max-product algorithm and its generalizations. LIDS Tech. report, June 2002.

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

[8] J. Feldman and D. Karger. Decoding turbo-like codes in polynomial time with provably good error-correcting performing via linear programming. Technical report, LCS, MIT, April 2002.