Convergence Analysis of Reweighted Sum-Product Algorithms

Tanya Roosta* Martin J. Wainwright*,† Shankar Sastry*

Department of Electrical and Computer Sciences*, and
Department of Statistics†
UC Berkeley, Berkeley, CA, 94720
{roosta,wainwrig,sastry}@eecs.berkeley.edu

Abstract

Markov random fields are designed to represent structured dependencies among large collections of random variables, and are well-suited to capture the structure of real-world signals. Many fundamental tasks in signal processing (e.g., smoothing, denoising, segmentation etc.) require efficient methods for computing (approximate) marginal probabilities over subsets of nodes in the graph. The marginalization problem, though solvable in linear time for graphs without cycles, is computationally intractable for general graphs with cycles. This intractability motivates the use of approximate "message-passing" algorithms. This paper studies the convergence and stability properties of the family of *reweighted sum-product algorithms*, a generalization of the widely-used sum-product or belief propagation algorithm, in which messages are adjusted with graph-dependent weights. For homogeneous models, we provide a complete characterization of the potential settings and message weightings that guarantee uniqueness of fixed points, and convergence of the updates. For more general inhomogeneous models, we derive a set of sufficient conditions that ensure convergence, and provide bounds on convergence rates. The experimental simulations on various classes of graphs validate our theoretical results.

Keywords: Markov random fields; graphical models; belief propagation; sum-product algorithm; convergence analysis; approximate marginalization.

1 Introduction

Graphical models provide a powerful framework for capturing the complex statistical dependencies exhibited by real-world signals. Accordingly, they play a central role in many applications, including statistical signal and image processing [1–3], error-control coding [4], computer vision [5], and computational biology [6]. A core problem common to applications in all of these domains is the *marginalization problem*—namely, to compute marginal distributions over local subsets of random variables. For graphical models without cycles, including Markov chains and trees (see Figure 1(a) and (b)), the marginalization problem is exactly solvable in linear-time via the sum-product algorithm, which operates in a distributed manner by passing "messages" between nodes in the graph. This sum-product framework includes many well-known algorithms as special cases, among them the α-β or forward-backward algorithm [1] for Markov chains, the peeling algorithm in bioinformatics, and the Kalman filter; see the review articles [2–4] for further background on the sum-product algorithm and its uses.

Although Markov chains/trees are tremendously useful, many classes of real-world signals are best captured by graphical models with cycles. (For instance, the lattice or grid-structured model in Figure 1(c) is widely used in computer vision and statistical image processing.) At least in principle, the nodes in any such graph with cycles can be clustered into "supernodes", thereby converting the original graph into junction tree form [7], to which the sum-product algorithm can be applied to obtain exact results. However, the cluster sizes required by this junction tree formulation—and hence the computational complexity of the sum-product algorithm—grow *exponentially* in the treewidth of the graph. For many classes of graphs, among them the lattice model in Figure 1(c), the treewidth grows in an unbounded manner with graph size, so that the junction tree approach rapidly becomes infeasible. Indeed, the marginalization problem is known to be computationally intractable [8, 9] for general graphical models.

This difficulty motivates the use of efficient algorithms for computing *approximations* to the marginal probabilities. In fact, one of the most successful approximate methods is based on applying the sum-product updates to the graphs with cycles. Convergence and correctness, though guaranteed for tree-structured graphs, are no longer ensured when the underlying graph has cycles. Nonetheless, this "loopy" form of the sum-product algorithm has proven very successful in many applications [2–5]. However, there remain a variety of theoretical questions concerning the use of sum-product and related message-passing algorithms for approximate marginalization. It is well known that the standard form of sum-product message-passing is not guaranteed to converge, and in fact may have multiple fixed points in certain regimes. Recent work has shed some light on the fixed points and convergence properties of the ordinary sum-product algorithm. Yedidia et al. [10] showed that sum-product fixed points correspond to local minima of an optimization problem known as the Bethe variational principle. Tatikonda and Jordan [11] established an elegant connection between the convergence of the ordinary sum-product algorithm and the uniqueness of Gibbs measures on the associated computation tree, and provided several sufficient conditions for convergence. Wainwright et al. [12] showed that the sum-product algorithm can be understood as seeking an alternative reparameterization of the distribution, and used this to characterize the error in the approximation. Heskes [13] discussed convergence and its relation to stability properties of the Bethe variational problem. Other researchers [14, 15] have used contraction arguments to provide sharper sufficient conditions for convergence of the standard sum-product algorithm. Finally, several groups [16–18] have proposed modified algorithms for solving the Bethe variational problem with convergence guarantees, albeit at the price of increased complexity.

In this paper, we study the broader class of *reweighted sum-product* algorithms [19–22], including the ordinary sum-product algorithm as a special case, in which messages are adjusted by edge-based weights determined by the graph structure. For suitable choices of these weights, the reweighted sum-product algorithm is known to have a unique fixed point for any graph and any interaction potentials [19]. An additional desirable property of reweighted sum-product is that the message-passing updates tend to be more stable, as confirmed by experimental investigation [19, 21, 22]. This algorithmic stability should be contrasted with the ordinary sum-product algorithm, which can be highly unstable due to phase transitions in the Bethe variational problem [10, 11]. Despite these encouraging empirical results, current theoretical understanding of the stability and convergence properties of reweighted message-passing remains incomplete.

The main contributions of this paper are a number of theoretical results characterizing the convergence properties of reweighted sum-product algorithms, including the ordinary sum-product updates as a special case. Beginning with the simple case of homogeneous binary models, we provide sharp guarantees for convergence, and prove that there always exists a choice of edge weights for which the associated reweighted sum-product algorithm converges. We then analyze more general inhomogeneous models, both for binary variables and the general multinomial model, and provide sufficient conditions for convergence of reweighted algorithms. Relative to past work, a notable feature of our analysis is that it incorporates the benefits of making observations, whether partial or noisy, of the underlying random variables in the Markov random field to which message-passing is applied. Intuitively, the convergence of message-passing algorithms should be function of both the strength of the interactions between random variables, as well as the local observations, which tend to counteract the interaction terms. Indeed, when specialized to the ordinary sum-product algorithm, our results provide a strengthening of the best previously known convergence guarantees for sum-product [14, 15]. As we show empirically, the benefits of incorporating observations into convergence analysis can be substantial, particularly in the regimes most relevant to applications.

The remainder of this paper is organized as follows. In Section 2, we provide basic background on graphical models (with cycles), and the class of reweighted sum-product algorithms that we study. Section 3 provides convergence analysis for binary models, which we then extend to general discrete models in Section 4. In Section 5, we describe experimental results that illustrate our findings, and we conclude in Section 6.

2 Background

In this section we provide some background on Markov random fields, and message-passing algorithms, including the reweighted sum-product that is the focus of this paper.

2.1 Graphical models

Undirected graphical models, also known as Markov random fields, are based on associating a collection of random variables $X = \{X_1, \ldots, X_n\}$ with the vertices of a graph. More precisely, an undirected graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ are vertices, and $E \subset V \times V$ are edges joining pairs of vertices. Each random variable X_i is associated with node $i \in V$, and the edges in the graph (or more precisely, the absences of edges) encode Markov properties of the random vector X. These Markov properties are captured by a particular factorization of the probability distribution p of the random vector X, which is guaranteed to break into a product of local functions on the cliques of the graph. (A graph clique is a subset C of vertices that are all joined by edges.)

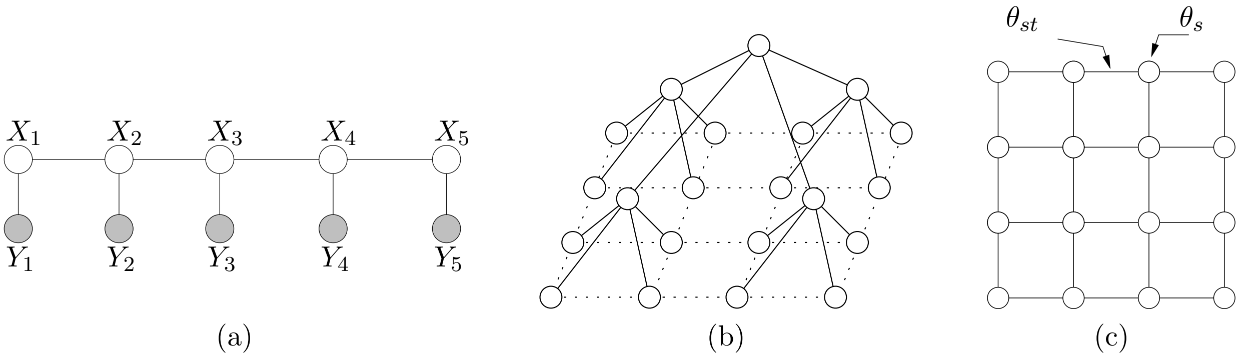


Figure 1. Examples of graphical models. (a) A hidden Markov chain model (with noisy observations Y_s of each hidden X_s), on which the marginalization problem is solved by the forward-backward algorithm. (b) Marginalization can also be performed in linear time on a tree (graph without cycles), as widely used in multi-resolution signal processing [2]. (c) A lattice-based model frequently used in image processing [23], for which the marginalization problem is intractable in general.

In this paper, we focus on discrete (multinomial) random variables $X_s \in \mathcal{X} := \{0, 1, \ldots, m-1\}$ with distribution specified according to a pairwise Markov random field. Any such model has a probability distribution of the form

$$p(x;\theta) \quad \propto \quad \exp\left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\}. \tag{1}$$

Here the quantities θ_s and θ_{st} are *potential functions* that depend only on the value $X_s = x_s$, and the pair values $(X_s, X_t) = (x_s, x_t)$ respectively. Otherwise stated, each singleton potential θ_s is a real-valued function of $\mathcal{X} = \{0, 1, \ldots, m\}$, whose values can be represented as an m-vector, whereas each edge potential θ_{st} is a real-valued mapping on the Cartesian product $\mathcal{X} \times \mathcal{X}$, whose values can be represented as a $m \times m$ matrix. For the discrete Markov random fields that we consider, the assumption of pairwise interactions only entails no loss of generality (see Yedidia et al. [10] for details).

With this set-up, the *marginalization problem* is to compute the singleton marginal distributions $p(x_s;\theta) = \sum_{x_t, t \neq s} p(x;\theta)$, and possibly higher-order marginal distributions (e.g., $p(x_s, x_t;\theta)$) as well. Note that if viewed naively, the summation defining $p(x_s;\theta)$ involves an exponentially growing number of terms (m^{n-1} to be precise).

2.2 Sum-product algorithms

The sum-product algorithm is an iterative algorithm for computing either exact marginals (on trees), or approximate marginals (for graphs with cycles). It operates in a distributed manner, with nodes in the graph exchanging statistical information via a sequence of “message-passing” updates. For tree-structured graphical models, the updates can be derived as a form of non-serial dynamic programming, and are guaranteed to converge and compute the correct marginal distributions at each node. However, the updates are routinely applied to more general graphs with cycles, which is the application of interest in this paper. Here we describe the more general family of reweighted sum-product algorithms, which include the ordinary sum-product updates as a particular case.

In any sum-product algorithm, one message is passed in each direction of every edge (s, t) in the graph. The message from node t to node s, denoted by $M_{ts}(x_s)$, is a function of the possible states $x_s \in \{0, 1, \ldots, m-1\}$ at node s. Consequently, in the discrete case, the message can be represented by an m-vector of possible function values. The family of reweighted sum-product algorithms is parameterized by a set of *edge weights*, with $\rho_{st} \in (0, 1]$ associated with edge (s, t). Various choices of these edge weights have been proposed [19, 21, 22], and have different theoretical properties. The simplest case of all—namely, setting $\rho_{st} = 1$ for all edges—recovers the ordinary sum-product algorithm. Given some fixed set of edge weights $\rho_{st} \in (0, 1]$, the reweighted sum-product updates are given by the recursion

$$M_{ts}(x_s) \quad \leftarrow \quad \sum_{x'_t} \exp\left\{\frac{\theta_{st}(x_s, x'_t)}{\rho_{st}} + \theta_t(x'_t)\right\} \frac{\prod\limits_{u \in N(t)\backslash s} [M_{ut}(x'_t)]^{\rho_{ut}}}{[M_{st}(x'_t)]^{\rho_{st}}}, \tag{2}$$

where $N(t) := \{s \in V \mid (s, t) \in E\}$ denotes the neighbors of node t in the graph. Typically, the message vector M_{ts} is normalized to unity after each iteration (i.e., $\sum_{x_s} M_{ts}(x_s) = 1$). Once the updates converge to some message fixed point M^*, then the fixed point can be used to compute (approximate) marginal probabilities τ_s at each node via

$$\tau_s(x_s) \quad \propto \quad \exp\{\theta_s(x_s)\} \prod_{t \in N(s)} [M^*_{ts}(x_s)]^{\rho_{st}} \tag{3}$$

When the ordinary updates($\rho_{st} = 1$) are applied to a tree-structured graph, it can be shown by induction that the algorithm converges after a finite number of steps. Moreover, a calculation using Bayes' rule shows that $\tau_s(x_s)$, computed via equation (3), is equal to the desired marginal probability $p(x_s; \theta)$. However, the sum-product algorithm is routinely applied to graphs with cycles, in which case the message updates (2) are not guaranteed to converge, and the quantities $\tau_s(x_s)$ represent approximations to the true marginal distributions. Our focus in this paper is to determine conditions under which the reweighted sum-product message updates (2) are guaranteed to converge.

3 Convergence Analysis

In this section, we describe and provide proofs of our main results on the convergence properties of the reweighted sum-product updates (2) when the messages belong to a binary state space, which we represent as $\mathcal{X} = \{-1, 1\}$. In this special case, the general MRF distribution (1) can be simplified into the Ising model form

$$p(x;\theta) \quad \propto \quad \exp\left\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \right\}, \tag{4}$$

so that the model is parameterized[1] by a single real number θ_s for each node, and a single real number θ_{st} for each edge.

3.1 Convergence for binary homogeneous models

We begin by stating and proving some convergence conditions for a particularly simple class of models: homogeneous models on d-regular graphs. A graph is d-regular if each vertex has exactly d neighbors. Examples include single cycles ($d = 2$), and lattice models with toroidal boundary conditions ($d = 4$). In a homogeneous model, the edge weights θ_{st} are equal to a common value θ_{ed}, and similarly the node parameters θ_s are all equal to a common value θ_{vx}.

In order to state our convergence result, we first define, for any real numbers u and ϕ, the function

$$G(u;\phi) \quad = \quad \frac{\exp(\phi+u)}{1+\exp(\phi+u)} - \frac{\exp(u)}{\exp(\phi)+\exp(u)}. \tag{5}$$

Note that for any fixed $\phi \in \mathbb{R}$, the function $G(\cdot\ \phi)$ is bounded in absolute value by $|G(0;\phi)|$.

Proposition 1. *For any homogeneous binary model on a d-regular graph with arbitrary choice of* $(\theta_{\text{vx}}, \theta_{\text{ed}})$*, the reweighted updates* (2) *have a unique fixed point and converge as long as* $R < 1$*, where* $\delta := 2|\theta_{\text{vx}}| - 2|\rho d - 1| \, \frac{|\theta_{\text{ed}}|}{\rho}$*, and*

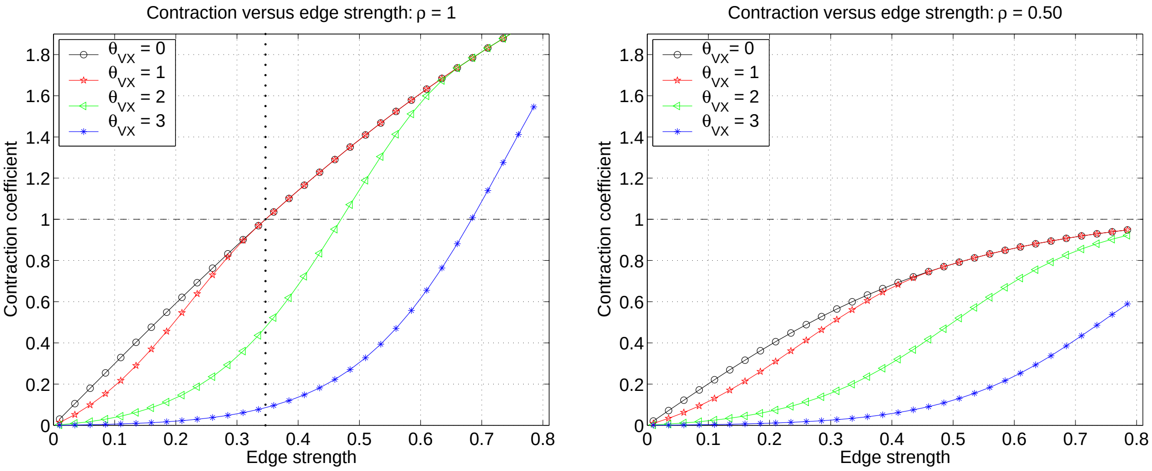
$$R_d(\theta_{\text{vx}}, \theta_{\text{ed}}; \rho) \quad := \quad \begin{cases} |\rho d - 1| \; G\left(0;\ \frac{2|\theta_{\text{ed}}|}{\rho}\right) & \textit{if } \delta \leq 0 \\ |\rho d - 1| \, G\left(\delta;\ \frac{2|\theta_{\text{ed}}|}{\rho}\right) & \textit{otherwise.} \end{cases}$$

Moreover, if $\rho \leq 2/d$*, then* $R < 1$ *for all finite choices of* $(\theta_{\text{vx}}, \theta_{\text{ed}})$*, so that the reweighted updates converge for any problem.*

Remarks: Consider the choice of edge weight $\rho = 1$, corresponding to the standard sum-product algorithm. If the graph is a single cycle ($d = 2$), Proposition 1 shows that the standard sum-product algorithm always converges, consistent with previous work on the single cycle case [11,25]. For more general graphs with $d > 2$, convergence depends on the relation between the observation strength θ_{vx} and the edge strengths θ_{ed}. For the case $d = 4$, corresponding for instance to a lattice model with toroidal boundary as in Figure 1(c), Figure 2(a) provides a plot of the coefficient $R_4(\theta_{\text{vx}}, \theta_{\text{ed}}; 1)$ as

[1] This assumption is valid, because the distribution (1) does not change if we replace $\theta_s(x_s)$ with $\widetilde{\theta}_s(x_s) := \theta_s(x_s) - \theta_s(-1)$, with a similar calculation for the edges. See [24] for details.

a function of the edge strength θ_{vx}, for different choices of the observation potential θ_{vx}. The curve marked with squares corresponds to $\theta_{\text{vx}} = 0$. Observe that it crosses the threshold $R_4 = 1$ from convergence to non-convergence at the critical value $\text{arctanh}(\frac{1}{3}) \approx 0.3466$, corresponding classical



(a)

(b)

Figure 2. Plots of the contraction coefficient $R_4(\theta_{\text{vx}}, \theta_{\text{ed}}; \rho)$ versus the edge strength θ_{ed}. Each curve corresponds to a different choice of the observation potential θ_{vx}. (a) For $\rho = 1$, the updates reduces to the standard sum-product algorithm; note that the transition from convergence to non-convergence occurs at $\theta^*_{\text{ed}} \approx 0.3466$ in the case of no observations ($\theta_{\text{vx}} = 0$). (b) Corresponding plots for reweighted sum-product with $\rho = 0.50$. Since $\rho d = (0.50)4 = 2$, the contraction coefficient is always less than one in this case, as predicted by Proposition 1.

result due to Bethe [26], and also confirmed in other analyses of standard sum-product [11, 14, 15]. The other curves correspond to non-zero observation potentials ($\theta_{\text{vx}} \in \{1, 2, 3\}$) respectively. Here it is interesting to note with $\theta_{\text{vx}} > 0$, Proposition 1 reveals that the standard sum-product algorithm continues to converge well beyond the classical breakdown point without observations ($\theta^*_{\text{ed}} \approx 0.3466$).

Figure 2 shows the corresponding curves of $R_4(\theta_{\text{vx}}, \theta_{\text{ed}}; 0.50)$, corresponding to the reweighted sum-product algorithm with $\rho = 0.50$. Note that $\rho d = 0.5(4) = 2$, so that as predicted by Proposition 1, the contraction coefficient R_4 remains below 1 for all values of θ_{vx} and θ_{ed}, meaning that the reweighted sum-product algorithm converges *for all* values of the potentials θ_{vx} and θ_{ed}.

Proof of Proposition 1: Given the edge and node homogeneity of the model and the d-regularity of the graph, the message-passing updates can be completely characterized by a single log message $z = \log M(1)/M(-1) \in \mathbb{R}$, and the update

$$F(z; \theta_{\text{vx}}, \theta_{\text{ed}}, \rho) \quad = \quad \log\left[\frac{\exp[\frac{2\theta_{\text{ed}}}{\rho} + (\rho d - 1)z + 2\theta_{\text{vx}}] + 1}{\exp[(\rho d - 1)z + 2\theta_{\text{vx}}] + \exp(\frac{2\theta_{\text{ed}}}{\rho})}\right]. \tag{6}$$

We begin by observing that for any choice of $z \in \mathbb{R}$, we have $|F(z; \theta_{\text{vx}}, \theta_{\text{ed}}, \rho)| \leq 2\frac{\theta_{\text{ed}}}{\rho}$, so that the message z must belong to the *admissible interval* $[-2\frac{\theta_{\text{ed}}}{\rho}, 2\frac{\theta_{\text{ed}}}{\rho}]$. Next we compute and bound the derivative of F over this set of admissible messages. A straightforward calculation yields that $F'(z) = (\rho d - 1) G\left(2\theta_{\text{vx}} + (\rho d - 1)z\,;\, 2\frac{\theta_{\text{ed}}}{\rho}\right)$, where the function G was defined previously in (5). Note that for any fixed $\phi \in \mathbb{R}$, the function $|G(u\,;\phi)|$ achieves its maximum at $u^* = 0$. Consequently, the unconstrained maximum of $|F'(z)|$ is achieved at the point $z^* = -2\theta_{\text{vx}}$ satisfying $2\theta_{\text{vx}} + (\rho d - 1)z^* = 0$, with $F(z^*) = G(0\,; 2\frac{|\theta_{\text{ed}}|}{\rho})$. Otherwise, if $|z^*| > 2\frac{|\theta_{\text{ed}}|}{\rho}$, then the constrained maximum is obtained at the boundary point of the admissible region closest to 0—namely, at the point $2\theta_{\text{vx}} - 2\,\text{sign}(\theta_{\text{vx}})\frac{\theta_{\text{ed}}}{\rho}$. Overall, we conclude that for all admissible messages z, we have

$$\frac{|F'(z)|}{|\rho d - 1|} \quad \leq \quad \begin{cases} G(0; 2\frac{|\theta_{\text{ed}}|}{\rho}) & \text{if } |\theta_{\text{vx}}| \leq \frac{|\theta_{\text{ed}}|}{\rho} \\ |G(2\theta_{\text{vx}} - 2\,\text{sign}(\theta_{\text{vx}})\frac{\theta_{\text{ed}}}{\rho}; \frac{2\theta_{\text{ed}}}{\rho})| \;=\; G(2|\theta_{\text{vx}}| - 2\frac{|\theta_{\text{ed}}|}{\rho}; \frac{2|\theta_{\text{ed}}|}{\rho}) & \text{otherwise,} \end{cases} \tag{7}$$

so that $|F'(z)| < R$ as defined in the statement. Note that if $R < 1$, the update is an iterated contraction, and hence converges [27].

3.2 Extension to binary inhomogeneous models

We now turn to the generalization of the previous result to the case of inhomogeneous models, in which the node parameters θ_s and edge parameters θ_{st} may differ across nodes and edges, respectively. For each *directed* edge $(t \to s)$, define the quantity

$$D_{t\to s}(\theta; \rho) \quad = \quad 2\left\{|\theta_t| - \sum_{u \in N(t)\backslash s} \rho_{ut}|\theta_{ut}| + (1 - \rho_{st})\frac{|\theta_{st}|}{\rho_{st}}\right\} \tag{8}$$

and the weight

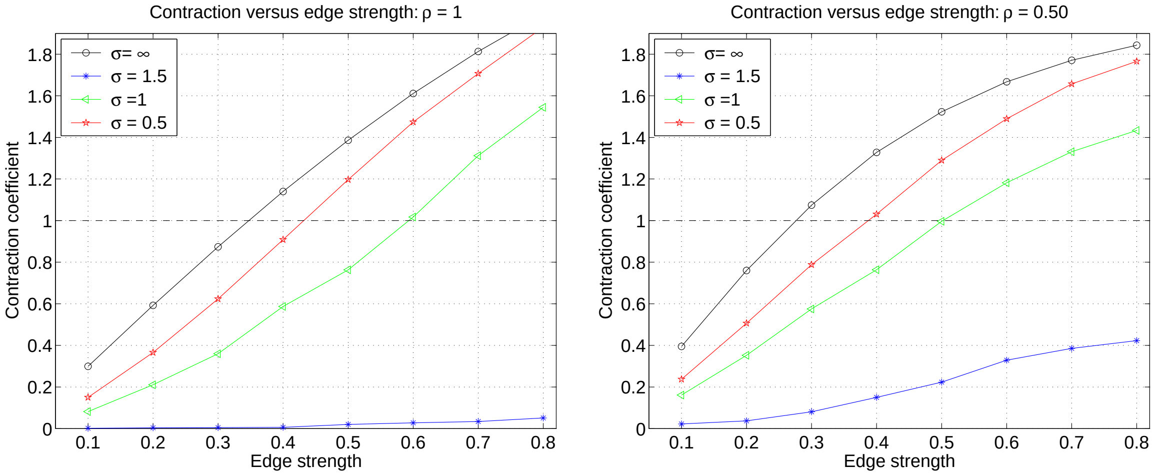
$$L_{t\to s} \quad = \quad \begin{cases} G(0; 2\frac{|\theta_{st}|}{\rho_{st}}) & \text{if } D_{t\to s}(\theta;\rho) \leq 0. \\ G\left(D_{t\to s}(\theta;\rho);\; 2\frac{|\theta_{st}|}{\rho_{st}}\right) & \text{otherwise,} \end{cases} \tag{9}$$

where the function G was previously defined (5). Finally, define a $2|E| \times 2|E|$ matrix $M = M(\theta; \rho)$, with entries indexed by directed edges $(t \to s)$, and of the form

$$M_{(t\to s),(u\to v)} \quad = \quad \begin{cases} \rho_{ut} L_{t\to s} & \text{if } v = t \text{ and } u \neq s \\ (1 - \rho_{st})\; L_{t\to s} & \text{if } v = t \text{ and } u = s \\ 0 & \text{otherwise.} \end{cases} \tag{10}$$

Theorem 2. *For an arbitrary pairwise Markov random field, the reweighted sum-product algorithm converges if the spectral radius of $M(\theta; \rho)$ is less than 1.*

When specialized to the case of uniform edge weights $\rho_{st} = 1$, then Theorem 2 strengthens previous results, due independently to Ihler et al. [14] and Mooij and Kappen [15], on the ordinary sum-product algorithm. This earlier work provided conditions based on matrices that involved only terms of the form $G(0; 2|\theta_{st}|)$, as opposed to the *smaller and observation-dependent* weights $G(D_{t\to s}(\theta;\rho);\ 2\,|\theta_{st}|)$ that our analysis yields once $D_{t\to s}(\theta;\rho) > 0$. As a consequence, Theorem 2 can yield sharper estimates of convergence by incorporating the benefits of having observations. In addition to these consequences for the ordinary sum-product algorithm, Theorem 2 also provides sufficient conditions for convergence of any reweighted sum-product algorithm.



(a)

(b)

Figure 3. Illustration of the benefits of observations. Plots of the contraction coefficient versus the edge strength. Each curve corresponds to a different setting of the noise variance σ^2 as indicated. (a) Ordinary sum-product algorithm $\rho = 1$. Upper-most curve labeled $\sigma^2 = +\infty$ corresponds to the best bounds from previous work [14,15]. (b) Reweighted sum-product algorithm $\rho = 0.50$.

In order to illustrate the benefits of including observations in the convergence analysis, we conducted experiments on grid-structured graphical models in which a binary random vector, with a prior distribution of the form (4), is observed in Gaussian noise (see Section 5.1 for the complete details of the experimental set-up). Figure 3 provides summary illustrations of our findings, for the ordinary sum-product ($\rho = 1$) in panel (a), and reweighted sum-product ($\rho = 0.50$) in panel (b). Each plot shows the contraction coefficient predicted by Theorem 2 as a function of an edge strength parameter. Different curves show the effect of varying the noise variance σ^2 specifying the signal-to-noise ratio in the observation model (see Section 5 for the complete details). The extreme case $\sigma^2 = +\infty$ corresponds to the case of no observations. Notice how the contraction coefficient steadily decreases as the observations become more informative, both for the ordinary and reweighted sum-product algorithms.

3.3 Proof of Theorem 2

We begin by establishing a useful auxiliary result that plays a key role in this proof, as well as other proofs in the sequel:

Lemma 3. *For real numbers ϕ and u, define the function*

$$H(u;\phi) \quad = \quad \log \frac{\exp(\phi + u) + 1}{\exp(u) + \exp(\phi)}. \tag{11}$$

For each fixed ϕ, we have $\sup_{u \in \mathbb{R}} |H(u;\phi)| \leq |\phi|$.

Proof. Computing the derivative of H with respect to u, we have

$$H'(u;\phi) \quad = \quad \frac{\exp(\phi+u)}{1+\exp(\phi+u)} - \frac{\exp(u)}{\exp(u)+\exp(\phi)} \quad = \quad \frac{\exp(u)\left\{\exp(2\phi)-1\right\}}{\left[\exp(u)+\exp(\phi)\right]\left[1+\exp(\phi+u)\right]},$$

so that H is strictly increasing if $\phi > 0$ and strictly decreasing if $\phi < 0$. Consequently, the supremum is obtained by taking $u \rightarrow \pm\infty$, and is equal to $|\phi|$ as claimed. □

With this lemma in hand, we begin by re-writing the message update (2) in a form more amenable to analysis. For each directed edge $(t \rightarrow s)$, define the log message ratio $z_{t \rightarrow s} = \log \frac{M_{t \rightarrow s}(1)}{M_{t \rightarrow s}(-1)}$. From the standard form of the updates, a few lines of algebra show that it is equivalent to update these log ratios via

$$F_{t \rightarrow s}(z) := \log \frac{\exp\left[\frac{2\theta_{st}}{\rho_{st}} + 2\theta_t + \sum\limits_{v \in N(t) \backslash s} \rho_{vt} z_{v \rightarrow t} + (1-\rho_{st}) z_{s \rightarrow t}\right] + 1}{\exp\left[2\theta_t + \sum\limits_{v \in N(t) \backslash s} \rho_{vt} z_{v \rightarrow t} + (1-\rho_{st}) z_{s \rightarrow t}\right] + \exp\left[\frac{2\theta_{st}}{\rho_{st}}\right]}. \tag{12}$$

A key property of the message update function $F_{t \rightarrow s}$ is that it can be written as a function H of the form (11), with $\phi = 2\frac{\theta_{st}}{\rho_{st}}$ and $u = 2\theta_t + \sum\limits_{v \in N(t) \backslash s} \rho_{vt} z_{vt} + (1-\rho_{st}) z_{st}$. Consequently, if we apply Lemma 3, we may conclude that $|F_{t \rightarrow s}(z)| \leq 2\frac{|\theta_{st}|}{\rho_{st}}$ for all $z \in \mathbb{R}$, and consequently that $|z_{ts}^n| \leq 2\frac{|\theta_{st}|}{\rho_{st}}$ for all iterations $n \geq 1$. Consequently, we may assume that message vector z^n for all iterations $n \geq 1$ belongs to the box of admissible messages defined by

$$\mathbb{B}(\theta;\rho) \quad := \quad \left\{ z \in \mathbb{R}^{2|E|} \;\mid\; |z_{t \rightarrow s}| \leq 2\frac{|\theta_{st}|}{\rho_{st}} \qquad \text{for all edges } (t \rightarrow s) \right\}. \tag{13}$$

We now bound the derivative of the message-update equation over this set of admissible messages:

Lemma 4. *For all $z \in \mathbb{B}(\theta; \rho)$, the elements of $\nabla F_{t\to s}(z)$ are bounded as*

$$\left|\frac{\partial F_{t\to s}}{\partial z_{u\to t}}(z)\right| \leq \rho_{ut} L_{t\to s} \quad \forall \quad u \in N(t)\backslash s, \qquad \textit{and} \qquad \left|\frac{\partial F_{t\to s}}{\partial z_{s\to t}}(z)\right| \;\leq\; (1-\rho_{st}) L_{t\to s}, \tag{14}$$

where the directed weights $L_{t\to s}$ were defined previously (9)*. All other gradient elements are zero.*

See Appendix A for the proof. In order to exploit Lemma 4, for any iteration $n \geq 2$, let us use the mean-value theorem to write

$$z^{n+1}_{st} - z^{n}_{st} \;=\; F_{t\to s}(z^n) - F_{t\to s}(z^{n-1}) \;=\; \nabla F_{t\to s}(z^\lambda)^T (z^n - z^{n-1}), \tag{15}$$

where $z^\lambda = \lambda z^n + (1-\lambda) z^{n-1}$ for some $\lambda \in (0,1)$. Since z^n and z^{n-1} both belong to the convex set $\mathbb{B}(\theta;\rho)$, so does the convex combination z^λ, and we can apply Lemma 4. Starting from equation (15), we have

$$\begin{aligned} \left|z^{n+1}_{t\to s} - z^{n}_{t\to s}\right| \;&\leq\; \left|\nabla F_{t\to s}(z^\lambda)\right|^T \left|z^n - z^{n-1}\right| \\ &\leq\; \left(\sum_{u\in N(t)\backslash s} \rho_{ut} L_{t\to s} |z^n_{u\to t} - z^{n-1}_{u\to t}|\right) + (1-\rho_{st}) L_{t\to s} |z^n_{s\to t} - z^{n-1}_{s\to t}|. \end{aligned} \tag{16}$$

Since this bound holds for each directed edge, we have established that the vector of message differences obeys $|z^{n+1} - z^n| \leq M(\theta,\rho)\,|z^n - z^{n-1}|$, where the non-negative matrix $M = M(\theta,\rho)$ was defined previously. By standard results on non-negative matrix recursions [28], if the spectral radius of M is less than 1, then the sequence $|z^n - z^{n-1}|$ converges to zero. Thus, the sequence $\{z^n\}$ is a Cauchy sequence, and so must converge.

3.4 Explicit conditions for convergence

A drawback of Theorem 2 is that it requires computing the spectral radius of the $2|E| \times 2|E|$ matrix M, which can be a non-trivial computation for large problems. Accordingly, we now specify some corollaries that are sufficient to ensure convergence of the reweighted sum-product algorithm. As in the work of Mooij and Kappen [15], the first two conditions follow by upper bounding the spectral norm by standard matrix norms. Conditions (c) and (d) are refinements that require further work.

Corollary 5. *Convergence of reweighted sum-product is guaranteed by any of the following conditions:*

(a) Row sum condition:

$$\max_{(t\to s)} \left(\sum_{u\in N(t)\backslash s} \rho_{ut} + (1-\rho_{st})\right) L_{t\to s} \;<\; 1. \tag{17}$$

(b) Column sum condition:

$$\max_{(t\to s)} C_{t\to s} \;=\; \max_{(t\to s)} \left\{ \rho_{ts}\big(\sum_{u\in N(t)\backslash s} L_{u\to t}\big) + (1-\rho_{ts}) L_{s\to t} \right\} \;<\; 1. \tag{18}$$

(c) Reweighted norm condition:

$$K(\theta) := \max_{(t\to s)} \left\{ \big(\sum_{u\in N(t)\backslash s} \rho_{ut} L_{u\to t}\big) + (1-\rho_{ts}) L_{s\to t} \right\} \;<\; 1. \tag{19}$$

(d) Pairwise neighborhood condition: the quantity

$$\min_{\lambda\in[0,1]} \max_{(t\to s)} \left\{ \rho_{ts}\big(\sum_{w\in N(t)\backslash s} L_{w\to t}\big) + (1-\rho_{ts}) L_{s\to t} \right\}^{\lambda} \max_{u\in N(t)} \left\{ \rho_{ts}\big(\sum_{v\in N(u)\backslash t} L_{v\to u}\big) + (1-\rho_{tu}) L_{t\to u} \right\}^{1-\lambda}$$

is less than one.

Remarks: To put these results in perspective, if we specialize to $\rho_{st} = 1$ for all edges and use the *weaker version* of the weights $L_{t\to s}$ that ignore the effects of observations, then the ℓ_∞-norm condition (18) is equivalent to earlier results on the ordinary sum-product algorithm [14, 15]. In addition, one may observe that for the ordinary sum-product algorithm (where $\rho_{ab} = 1$ for all edges), condition (19) is equivalent to the ℓ_∞-condition (18). However, for the general reweighted algorithm, these two conditions are distinct.

Proof. Conditions (a) and (b) follows immediately from the fact that the spectral norm of M is upper bounded by any other matrix norm [28]. It remains to prove conditions (c) and (d) in the corollary statement.

(c) Defining the vector $\Delta^n = |z^n - z^{n-1}|$ of successive changes, from equation (16), we have

$$\Delta^{n+1}_{t\to s} \;\leq\; L_{t\to s} \left\{ \sum_{u\in N(t)\backslash s} \rho_{ut} \Delta^n_{u\to t} + (1-\rho_{st}) \Delta^n_{s\to t} \right\} \tag{20}$$

The previous step of the updates yields a similar equation—namely

$$\Delta^{n}_{u\to t} \;\leq\; L_{u\to t} \left\{ \sum_{v\in N(u)\backslash t} \rho_{vu} \Delta^{n-1}_{v\to u} + (1-\rho_{ut}) \Delta^{n-1}_{t\to u} \right\}. \tag{21}$$

Now let us define a norm on Δ by $\|\Delta\|_* = \max_{(t\to s)\in E} \Big\{ \sum_{u\in N(t)\backslash s} \rho_{ut} |\Delta_{u\to t}| + (1-\rho_{st}) |\Delta_{s\to t}|) \Big\}$. With this notation, the bound (21) implies that $\Delta^n_{u\to t} \leq L_{u\to t} \|\Delta^{n-1}\|_*$. Substituting this bound into equation (20) yields that

$$\Delta^{n+1}_{t\to s} \;\leq\; L_{t\to s}\left\{\sum_{u\in N(t)\backslash s}\rho_{ut}L_{u\to t} + (1-\rho_{st})L_{s\to t}\right\}\|\Delta^{n-1}\|_{*} \;\leq K(\theta)\|\Delta^{n-1}\|_{*}.$$

For any edge $(s \to u)$, summing weighted versions of this equation over all neighbors of s yields that

$$\begin{aligned}\sum_{v\in N(s)\backslash u}\rho_{vs}\Delta^{n+1}_{v\to s} + (1-\rho_{us})\Delta^{n+1}_{u\to s} \;&\leq\; \left\{\sum_{v\in N(s)\backslash u}\rho_{vs}L_{v\to s} + (1-\rho_{us})L_{u\to s}\right\} K(\theta)\|\Delta^{n-1}\|_{*} \\ &\leq\; K^2(\theta)\,\|\Delta^{n-1}\|_{*}.\end{aligned}$$

Finally, since the edge $(s \to u)$ was arbitrary, we can maximize over it, which proves that $\|\Delta^{n+1}\|_{*} \leq K^2(\theta)\,\|\Delta^{n-1}\|_{*}$. Therefore, if $K(\theta) < 1$, the updates are an iterated contraction in the $\|\cdot\|_{*}$ norm, and hence converge by standard contraction results [27].

(d) Given a non-negative matrix A, let $C_\alpha(A)$ denote the column sum indexed by some element α. In general, it is known [28] that for any $\lambda \in [0,1]$, the spectral radius of A is upper bounded by the quantity $\max_{\alpha,\beta}[C_\alpha(A)]^\lambda\,[C_\beta(A)]^{1-\lambda}$, where α and β range over all column indices. A more refined result due to Kolotilina [29, 30] asserts that if A is a sparse matrix, then one need only optimize over column index pairs α, β such that $A_{\alpha\beta} \neq 0$. For our problem, the matrix M is indexed by directed edges $(s \to t)$, and $M_{(t\to s),(u\to v)}$ is non-zero only if $v = t$. Consequently, we can reduce the maximization over column sums to maximizing over directed edge pairs $(t \to s)$ with $(u \to t)$, which yields the stated claim.

□

4 Convergence for General Discrete Models

In this section, we describe how our results generalize to multinomial random variables, with the variable X_s at each node s taking a total of $m \geq 2$ states in the space $\mathcal{X} = \{0, 1, \ldots, m-1\}$. Given our Markov assumptions, the distribution takes the factorized form

$$p(x;\theta) \;\propto\; \exp\left\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s, x_t)\right\}, \tag{22}$$

where each $\theta_s(\cdot)$ is a vector of m numbers, and each $\theta_{st}(\cdot\,,\cdot)$ is a $m \times m$ matrix of numbers.

In our analysis, it will be convenient to work with an alternative parameter vector $\widetilde{\theta}$ that represents the same Markov random field as $p(x;\theta)$, given by

$$\begin{aligned} \widetilde{\theta}_{st}(x_s, x_t) \quad &:= \quad \theta_{st}(x_s, x_t) - \theta_{st}(x_s, 0) - \theta_{st}(0, x_t) + \theta_{st}(0,0), \text{ and} && \text{(23a)} \\ \widetilde{\theta}_s(x_s) \quad &= \quad \theta_s(x_s) + \sum_{t \in N(s)} \left[\theta_{st}(x_s, 0) - \theta_{st}(0,0)\right]. && \text{(23b)} \end{aligned}$$

This set of functions $\widetilde{\theta}$ is a different parameterization of the distribution $p(x;\theta)$ because

$$\sum_{s \in V} \widetilde{\theta}_s(x_s) + \sum_{(s,t) \in E} \widetilde{\theta}_{st}(x_s, x_t) \quad = \quad \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) + C,$$

where C is a constant independent of x. Moreover, note that $\widetilde{\theta}_s(0) = 0$ for all nodes $s \in V$, and $\theta_{st}(x_s, 0) = \theta_{st}(0, x_t) = 0$ for all $x_s, x_t \in \{0, 1, \ldots, m-1\}$.

4.1 Convergence Theorem and Some Consequences

In order to state a result about convergence for multinomial Markov random fields, we require some preliminary definitions. For each directed edge $(t \to s)$ and states i, $k \in \{1, \ldots, m-1\}$, define functions of the vector $\vec{v} = (v(1), \ldots, v(m-1))$ as follows

$$\begin{aligned} \psi_{t \to s}(\vec{v}; k, i) \quad &:= \quad \frac{1}{2}\left| -\beta_{t \to s}(\vec{v}; i, k) - \alpha_{t \to s}(\vec{v}; i, k) + v(k) + \widetilde{\theta}_t(k) + \frac{\widetilde{\theta}_t(k,i)}{\rho_{st}} \right|, \text{ and} && \text{(24a)} \\ \phi_{t \to s}(\vec{v}; k, i) \quad &:= \quad \frac{1}{2}\left| \beta_{t \to s}(\vec{v}; k, i) - \alpha_{t \to s}(\vec{v}; k, i) + \frac{\widetilde{\theta}_{t \to s}(k,i)}{\rho_{st}} \right|, && \text{(24b)} \end{aligned}$$

where

$$\begin{aligned} \alpha_{t \to s}(\vec{v}; k, i) \quad &:= \quad \log\left(1 + \sum_{x_t \neq 0, k} \exp\left\{ \frac{\widetilde{\theta}_{ts}(i, x_t)}{\rho_{st}} + v(x_t) + \widetilde{\theta}_t(k) \right\}\right) && \text{(25a)} \\ \beta_{t \to s}(\vec{v}; k, i) \quad &:= \quad \log\left(1 + \sum_{x_t \neq 0, k} \exp(v(x_t) + \widetilde{\theta}_t(k))\right). && \text{(25b)} \end{aligned}$$

With these definitions, we now specify, for each directed edge $(t \to s)$, the following non-negative weight

$$L_{t \to s} \quad := \quad \max_{i,k \in \{1, \ldots, m\}} \max_{\vec{v} \in \mathbb{B}_{ts}(\widetilde{\theta}; \rho)} \left| G\left(\psi_{t \to s}(\vec{v}; i, k); \phi_{t \to s}(\vec{v}; i, k)\right)\right|, \tag{26}$$

where the function G was defined previously (5) and the box of admissible vectors is given by

$$\mathbb{B}_{ts}(\theta;\rho) \quad := \quad \left\{ \vec{v} \in \mathbb{R}^{m-1} \;\mid\; |v(k)| \leq \sum_{u \in N(t)\backslash s} \max_j \frac{|\widetilde{\theta}_{u\to t}(j,k)|}{\rho_{ut}} + (1-\rho_{st}) \max_j \frac{|\widetilde{\theta}_{s\to t}(j,k)|}{\rho_{st}} \right\}. \quad (27)$$

Finally, using the choice of weights $L_{t\to s}$ in equation (26), we define the $2|E| \times 2|E|$ matrix $M = M(L)$ as before (see equation (10)).

Theorem 6. *The reweighted sum-product algorithm converges if the spectral radius of M is less than one.*

Despite its notational complexity, Theorem 6 is simply a natural generalization of our earlier results for binary variables. When $m = 2$, note that the functions $\beta_{t\to s}$ and $\alpha_{t\to s}$ are identically zero (since there are no states other than $k = 1$ and 0), so that the form of ϕ and ψ simplify substantially. Moreover, as in our earlier development on the binary case, when specialized to $\rho_{st} = 1$, Theorem 6 provides a strengthening of previous results [14, 15] on the ordinary sum-product algorithm. In particular, we now show how these previous results can be recovered from Theorem 6 by ignoring the box constraints (27):

Corollary 7. *The reweighted sum-product algorithm converges if*

$$\max_{t\to s} \sum_{u \in N(t)\backslash s} \rho_{u\to t} W_{u\to t} + (1-\rho_{s\to t}) W_{s\to t} \quad < \quad 1, \qquad (28)$$

where $W_{u\to t} = \tanh\left(\frac{1}{4\rho_{ut}} \max_{i\neq j} \max_{\ell\neq k} |\theta_{ts}(\ell,i) - \theta_{ts}(\ell,j) - \theta_{ts}(k,i) + \theta_{ts}(k,j)|\right)$.

Proof. We begin by proving that $L_{u\to t} \leq W_{u\to t}$. First of all, ignoring the box constraints (27), then certainly

$$\begin{aligned} L_{t\to s} &\leq \max_{i,k \in \{1,\ldots,m\}} \max_{\vec{v} \in \mathbb{R}^{m-1}} |G(\psi_{t\to s}(\vec{v};i,k); \phi_{t\to s}(\vec{v};i,k))| \\ &\leq \max_{i,k \in \{1,\ldots,m\}} \max_{\vec{v} \in \mathbb{R}^{m-1}} |G(0; \phi_{t\to s}(\vec{v};i,k))| \\ &= \max_{i,k \in \{1,\ldots,m\}} \max_{\vec{v} \in \mathbb{R}^{m-1}} \tanh\left(\frac{1}{2}|\phi_{t\to s}(\vec{v};i,k)|\right), \end{aligned}$$

since for any fixed ϕ, the function $G(u^*;\phi)$ is maximized at $u^* = 0$, and $G(0;|\phi|) = \tanh(|\phi|/2)$. Due to the monotonicity of $G(0;|\phi|)$ in ϕ, it now suffices to maximize the absolute value of $|\phi_{t\to s}(\vec{v};i,k)|$.

Since ϕ is defined in terms of β and α, we first bound their difference. In one direction, we have

$$\begin{aligned}\alpha_{t\to s}(\vec{v};i,k)-\beta_{t\to s}(\vec{v};i;k) &= \log\frac{1+\sum_{x_t\neq 0,k}\exp\left\{\frac{\widetilde{\theta}_{t\to s}(x_t,i)}{\rho_{st}}+v(x_t)+\widetilde{\theta}_t(x_t)\right\}}{1+\sum_{x_t\neq 0,k}\exp\left\{v(x_t)+\widetilde{\theta}_t(x_t)\right\}}\\ &\geq \min_{x_t\neq k,0}\frac{\widetilde{\theta}_{ts}(x_t,i)}{\rho_{st}},\end{aligned}$$

and hence

$$\phi_{t\to s}(\vec{v};k,i) \quad\leq\quad \frac{1}{2\rho_{st}}\max_{x_t\neq k,0}\left\{\widetilde{\theta}_{ts}(k,i)-\widetilde{\theta}_{ts}(x_t,i)\right\}. \tag{29}$$

In the other direction, we have $\beta_{t\to s}(\vec{v};i,k)-\alpha_{t\to s}(\vec{v};i;k)\geq -\max_{x_t\neq k,0}\frac{\widetilde{\theta}_{ts}(x_t,i)}{\rho_{st}}$, and hence

$$\phi_{t\to s}(\vec{v};i,k) \quad\geq\quad -\frac{1}{2\rho_{st}}\max_{x_t\neq k,0}\left\{\widetilde{\theta}_{ts}(x_t,i)-\widetilde{\theta}_{st}(i,k)\right\}. \tag{30}$$

Combining equations (29) and (30), we conclude that

$$\begin{aligned}\max_{i,k\neq 0}\max_{\vec{v}\in\mathbb{R}^{m-1}}|\phi_{t\to s}(\vec{v};i,k)| &\leq \frac{1}{2\rho_{st}}\max_{i,k}\max_{\ell\neq k,0}\left|\widetilde{\theta}_{ts}(\ell,i)-\widetilde{\theta}_{ts}(k,i)\right|\\ &= \frac{1}{2\rho_{st}}\max_{i\neq 0}\max_{\ell\neq k}|\theta_{ts}(\ell,i)-\theta_{ts}(\ell,0)-\theta_{ts}(k,i)+\theta_{ts}(k,0)|\end{aligned}$$

Therefore, we have proved that $L_{t\to s}\leq W_{t\to s}$, where $W_{t\to s}$ was defined in the corollary statement. Consequently, if we define a matrix $M(W)$ using the weights W, we have $M(L)\leq M(W)$ in an elementwise sense, and therefore, the spectral radius of $M(W)$ is an upper bound on the spectral radius of $M(L)$ (see Bertsekas and Tsitsiklis [28]). □

If the graphical model has very weak observations (uniform potentials θ_s), then Theorem 6 provides little benefit over Corollary 7. However, as with the earlier results on binary models (see Figure 3), the benefits are substantial when the model has stronger observations, as would be common in applications. We provide the proof of Theorem 6 in the appendix.

5 Experimental Results

In this section, we present the results of experimental simulations to illustrate and support our theoretical findings.

5.1 Dependence on Signal-to-Noise Ratio

We begin by describing the experimental set-up used to generate the plots in Figure 3, which illustrate the effect of increased signal-to-noise ratio (SNR) on convergence bounds. In these sim-

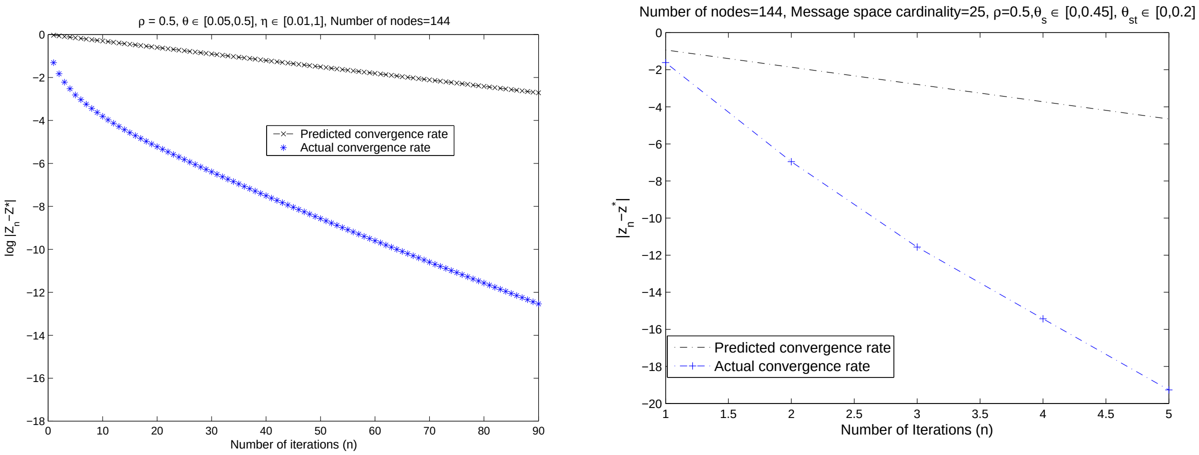


Figure 4. Convergence rates of the reweighted sum-product algorithm as compared to the rate predicted by reweighted norm condition 19. (a) Binary state spaces ($m = 2$). (b) Higher-order spaces ($m = 25$).

ulations, the random vector $X \in \{-1, +1\}^n$ is posited to have a prior distribution $p(x; \theta)$ of the form (4), with the edge parameters θ_{st} set uniformly to some fixed number θ_{ed}, and symmetric node potentials $\theta_s = 0$. Now suppose that the we make a noisy observation of the random vector X, say of the form

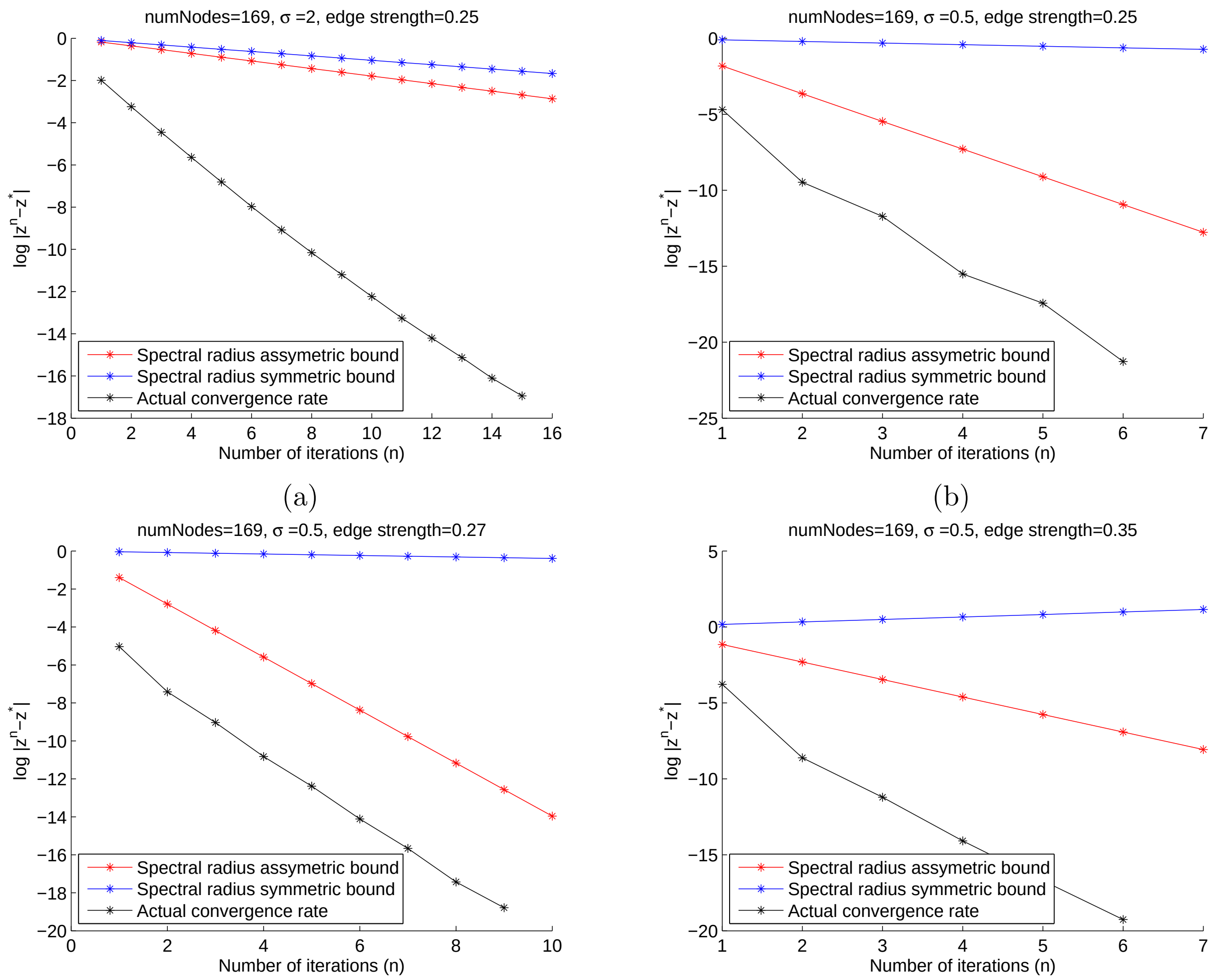
$$Y_s \quad = \quad X_s + W_s, \qquad \text{where } W_s \sim N(0, \sigma^2), \tag{31}$$

so that we have a conditional distribution of the form $p(y_s \,|\, x_s) \propto \exp(-\frac{1}{2\sigma^2}(y_s - x_s)^2)$. We then examined the convergence behavior of both ordinary and reweighted sum-product algorithms for the posterior distribution $p(x \mid y) \propto p(x; \theta) \prod_{s=1}^{n} p(y_s \mid x_s)$.

The results in Figure 3 were obtained from a grid with $n = 100$ nodes, and by varying the observation noise σ^2 from $\sigma^2 = +\infty$ corresponding to $SNR = 0$, down to $\sigma^2 = 0.5$. For any fixed setting of σ^2, each curve plots the average of the spectral radius bound from Theorem 2 over 20 trials versus the edge strength parameter θ_{ed}. Note how the convergence guarantees are substantially strengthened, relative to the $SNR = 0$ case, as the benefits of observations are incorporated.

5.2 Convergence rates

We now turn to a comparison of the empirical convergence rates of the reweighted sum-product algorithm to the theoretical upper bounds provided by the inductive norm (19) in Corollary 7. We have performed a large number of simulations for different values of number of nodes, edge weights ρ, node potentials, edge potentials, and message space size. Figure 4 shows a few plots that are representative of our findings, for binary state spaces (panel (a)) and higher order state spaces (panel (b)). The numbers parameterizing the node potentials, θ_s, and the edge potentials, θ_{st}, are shown on the corresponding plots. As shown in these plots, the convergence rates predicted by



(a)

(b)

(c)

(d)

Figure 5. Empirical rates of convergence of the reweighted sum-product algorithm as compared to the rate predicted by the symmetric and asymmetric bounds from Theorem 2.

Corollary 7 are consistent with the empirical performance, but tend to be overly conservative.

Figure 5 compares the convergence rates predicted by Theorem 2 to the empirical rates in both the *symmetric* and *asymmetric* settings. The symmetric case corresponds to computing the weights $L_{t\to s}$ while ignoring the observation potentials, so that the overall matrix M is symmetric in the edges (i.e., $L_{t\to s} = L_{s\to t}$). The asymmetric case explicitly incorporates the observation potentials, and leads to bounds that are as good or better than the symmetric case. Figure 5 illustrates the benefits of including observations in the convergence analysis. Perhaps most strikingly, panel (d) both shows a case where the symmetric bound predicts divergence of the algorithm, whereas the asymmetric bound predicts convergence.

6 Discussion

Many applications of graphical models require efficient methods for computing (approximate) marginal probabilities over subsets of nodes in the graph. For general graphs, the problem of marginalization becomes intractable due to the existence of cycles in the graph. This motivates the use of approximate message-passing algorithms, including the sum-product algorithm and its variants. In this paper, we studied the convergence and stability properties of the family of reweighted sum-product algorithms. For homogeneous models, we provided a complete characterization of the potential settings and message weightings that guarantee uniqueness of fixed points, and convergence of the updates. For more general inhomogeneous models, we derived a set of sufficient conditions that ensure convergence, and provide estimates of rates. We provided simulation results to complement the theoretical results presented.

Even though we have shown the benefits for convergence bounds of including observation potentials, as with past work, all of the conditions provided are still somewhat conservative. The reason is that the condition requires that the message updates be contractive at every node and every update of the graph, as opposed to requiring that they be attractive in some suitably averaged sense. An interesting direction would be to derive sharper "average-case" conditions for message-passing convergence.

Acknowledgments

The work of TR and SS was supported in part by TRUST (The Team for Research in Ubiquitous Secure Technology), which receives support NSF grant CCF-0424422, and the following organizations: Cisco, ESCHER, HP, IBM, Intel, Microsoft, ORNL, Qualcomm, Pirelli, Sun and Symantec. The work of MJW was supported by NSF grants DMS-0528488 and NSF CAREEER CCF-0545862.

A Proof of Lemma 4

Setting $\Delta_{ts} = \sum_{u \in N(t)\backslash s} \rho_{ut} z_{u\to t} + (1-\rho_{st}) z_{s\to t}$, we compute via chain rule

$$\frac{\partial F_{t\to s}}{\partial z_{u\to t}}(z) \quad = \quad \begin{cases} \rho_{ut}\frac{\partial F_{t\to s}}{\partial \Delta_{ts}} & \text{for } u \in N(t)\backslash s, \\ \rho_{ut}\frac{\partial F_{t\to s}}{\partial \Delta_{ts}} & \text{for } u = s, \end{cases}$$

so that it suffices to upper bound $|\frac{\partial F_{t\to s}}{\partial \Delta_{ts}}|$. Computing this partial derivative from the message update (12) yields

$$\begin{aligned}\frac{\partial F_{t\to s}}{\partial \Delta_{ts}} &= \frac{\exp\left[\frac{2\theta_{st}}{\rho_{st}} + 2\theta_t + \Delta_{ts}\right]}{1+\exp\left[\frac{2\theta_{st}}{\rho_{st}} + 2\theta_t + \Delta_{ts}\right]} - \frac{\exp\left[2\theta_t + \Delta_{ts}\right]}{\exp\left[2\theta_t + \Delta_{ts}\right] + \exp\left[\frac{2\theta_{st}}{\rho_{st}}\right]} \\ &= G\left(2\theta_t + \Delta_{ts}; \frac{2\theta_{st}}{\rho_{st}}\right),\end{aligned}$$

where the function G was previously defined (5). Since the message vector z^n must belong to the box (13) of admissible messages, the vector Δ_{ts} must satisfy the bound

$$|\Delta_{ts}| \;\leq\; \sum_{u\in N(t)\backslash s} 2\rho_{ut}|\theta_{ut}| + 2(1-\rho_{st})\frac{|\theta_{st}|}{\rho_{st}} \;| := \; U_{st}$$

For any fixed ϕ, the function $|G(u;\phi)|$ achieves its maximal value $|G(0;\phi)| = G(0;|\phi|)$ at $u^* = 0$. Noting that by its definition (8), we have $D_{t\to s}(\theta;\rho) = 2|\theta_t| - U_{ts}$, we conclude that

$$\begin{aligned}|\frac{\partial F_{t\to s}}{\partial \Delta_{ts}}| &\leq \max_{|\Delta_{ts}|\leq U_{ts}} \left|G\left(2\theta_t + \Delta_{ts}; \frac{2\theta_{st}}{\rho_{st}}\right)\right| \\ &= \begin{cases} |G(0; \frac{2|\theta_{st}|}{\rho_{st}})| & \text{if } D_{t\to s}(\theta;\rho) \leq 0 \\ G\left(D_{t\to s}(\theta;\rho);\; 2\frac{|\theta_{st}|}{\rho_{st}}\right) & \text{otherwise.}\end{cases},\end{aligned}$$

B Proof of Theorem 6

We begin by parameterizing the reweighted sum-product messages in terms of the log ratios $z_{st}(i) := \log\frac{M_{st}(i)}{M_{st}(0)}$. For each $i \in \{1,\ldots,m-1\}$, the message updates (2) can be re-written, following some straightforward algebra, in terms of these log messages and the modified potentials $\widetilde{\theta}$ as

$$F_{t\to s}(z) \;=\; \log\frac{1+\sum_{x_t\neq 0}\exp\left\{\frac{\widetilde{\theta}_{ts}(i,x_t)}{\rho_{st}} + \widetilde{\theta}_t(x_t) + \Delta_{ts}(x_t)\right\}}{1+\sum_{x_t\neq 0}\exp\left\{\widetilde{\theta}_t(x_t) + \Delta_{ts}(x_t)\right\}}, \tag{32}$$

where $\Delta_{ts}(x_t) := (1-\rho_{st})\, z_{st}(x_t) + \sum_{v\in N(t)\backslash s}\rho_{vt}z_{vt}(x_t)$. Analogously to the proof of Lemma 3, we have $|z_{ts}(i)| \leq \max_{x_t}|\frac{\widetilde{\theta}_{ts}(i,x_t)}{\rho_{st}}|$. Consequently, each vector $\Delta_{ts}(x_t)$ must belong the admissible box (27).

As in our earlier proof, we now seek to bound the partial derivatives of the message update $z_{ts} \leftarrow F_{t\to s}(z)$. By chain rule, we have

$$\frac{\partial z_{ts}}{\partial z_{vt}} \;=\; \begin{cases}\rho_{vt}\frac{\partial z_{ts}}{\partial \Delta ts} & \text{if } v \in N(t)\backslash s \\ (1-\rho_{st})\frac{\partial z_{ts}}{\partial \Delta ts} & \text{if } v = s.\end{cases}$$

so that it suffices to bound $\frac{\partial z_{ts}}{\partial \Delta ts}$. Computing the partial derivative of component $x_s = i$ with respect to message index $x_t = k$ yields

$$\frac{\partial z_{ts}(i)}{\partial \Delta_{st}(k)} \quad = \quad \frac{\exp\left\{\frac{\widetilde{\theta}_{ts}(i,k)}{\rho_{st}} + \widetilde{\theta}_t(x_t) + \Delta_{ts}(k)\right\}}{1 + \sum_{x_t \neq 0} \exp\left\{\frac{\widetilde{\theta}_{ts}(i,x_t)}{\rho_{st}} + \widetilde{\theta}_t(x_t) + \Delta_{ts}(x_t)\right\}} - \frac{\exp\left\{\widetilde{\theta}_t(x_t) + \Delta_{ts}(k)\right\}}{1 + \sum_{x_t \neq 0} \exp\left\{\widetilde{\theta}_t(x_t) + \Delta_{ts}(x_t)\right\}}$$

Isolating the term involving $x_t = k$, we have

$$\begin{aligned}\frac{\partial z_{st}(i)}{\partial \Delta_{st}(k)} = &\frac{\exp\left\{\frac{\widetilde{\theta}_{ts}(i,k)}{\rho_{st}} + \widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}}{1 + \sum_{x_t \neq 0,k} \exp\left\{\frac{\widetilde{\theta}_{ts}(i,x_t)}{\rho_{st}} + \widetilde{\theta}_t(x_t) + \Delta_{ts}(x_t)\right\} + \exp\left\{\frac{\widetilde{\theta}_{ts}(i,k)}{\rho_{st}} + \widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}} \\ &- \frac{\exp\left\{\widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}}{1 + \sum_{x_t \neq 0,k} \exp\left\{\widetilde{\theta}_t(x_t) + \Delta_{ts}(x_t)\right\} + \exp\left\{\widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}}.\end{aligned}$$

Further simplifying

$$\frac{\partial z_{st}(i)}{\partial \Delta_{st}(k)} \quad = \quad \frac{\exp\left\{\frac{\widetilde{\theta}_{ts}(i,k)}{\rho_{st}} - \alpha_{t\to s}(i,k) + \widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}}{1 + \exp\left\{\frac{\widetilde{\theta}_{ts}(i,k)}{\rho_{st}} - \alpha_{t\to s}(i,k) + \widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}} - \frac{\exp\left\{-\beta_{t\to s}(i,k) + \widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}}{1 + \exp\left\{-\beta_{t\to s}(i,k) + \widetilde{\theta}_t(k) + \Delta_{ts}(k)\right\}} \tag{33}$$

where $\alpha_{t\to s}(\Delta; i, k)$ and $\beta_{t\to s}(\Delta; i, k)$ were previously defined.

Setting $v = -\beta_{t\to s}(i,k) + \widetilde{\theta}_t(k) + \Delta_{ts}(k)$ and $\varphi = \frac{\widetilde{\theta}_{ts}(i,k)}{\rho_{st}} - \alpha_{t\to s}(i,k) + \beta_{t\to s}(i,k)$, we have

$$\begin{aligned}\frac{\partial z_{st}(i)}{\partial \Delta_{st}(k)} &= \frac{\exp(\varphi + v)}{1 + \exp(\varphi + v)} - \frac{\exp(v)}{1 + \exp(v)} \\ &= \frac{\exp(\varphi + v)}{1 + \exp(\varphi + v)} - \frac{\exp(v + \frac{\varphi}{2})}{\exp(\frac{\varphi}{2}) + \exp(v + \frac{\varphi}{2})} \;=\; G(v + \frac{\varphi}{2}; \frac{\varphi}{2}),\end{aligned}$$

where G was previously defined (5). Using the monotonicity properties of G, we have

$$\left|\frac{\partial z_{st}(i)}{\partial \Delta_{st}(k)}\right| \quad \leq \quad G\left(\left|v + \frac{\varphi}{2}\right|; \left|\frac{\varphi}{2}\right|\right). \tag{34}$$

The claim follows by noting that as defined, we have

$$\begin{aligned}\psi_{t\to s}(\vec{v}; k, i) &= \left|v + \frac{\varphi}{2}\right| = \frac{1}{2}\left|-\beta_{t\to s}(\vec{v}; i, k) - \alpha_{t\to s}(\vec{v}; i, k) + v(k) + \widetilde{\theta}_t(k) + \frac{\widetilde{\theta}_t(k,i)}{\rho_{st}}\right|, \text{ and} \\ \phi_{t\to s}(\vec{v}; k, i) &= \left|\frac{\varphi}{2}\right| = \frac{1}{2}\left|\beta_{t\to s}(\vec{v}; k, i) - \alpha_{t\to s}(\vec{v}; k, i) + \frac{\widetilde{\theta}_{t\to s}(k,i)}{\rho_{st}}\right|.\end{aligned}$$

References

[1] L. R. Rabiner and B. H. Juang, *Fundamentals of speech recognition*, Prentice Hall, Englewood Cliffs, N.J., 1993.

[2] A. S. Willsky, “Multiresolution Markov models for signal and image processing,” *Proceedings of the IEEE*, vol. 90, no. 8, pp. 1396–1458, 2002.

[3] H. A. Loeliger, “An introduction to factor graphs,” *IEEE Signal Processing Magazine*, vol. 21, pp. 28–41, 2004.

[4] F. Kschischang, “Codes defined on graphs,” *IEEE Signal Processing Magazine*, vol. 41, pp. 118–125, August 2003.

[5] W. T. Freeman, E. C. Pasztor, and O. T. Carmichael, “Learning low-level vision,” in *International Journal of Computer Vision*, October 2000.

[6] R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, Eds., *Biological Sequence Analysis*, Cambridge University Press, Cambridge, 1998.

[7] S. L. Lauritzen and D. J. Spiegelhalter, “Local computations with probabilities on graphical structures and their application to expert systems (with discussion),” *Journal of the Royal Statistical Society B*, vol. 50, pp. 155–224, January 1988.

[8] G. Cooper, “The computational complexity of probabilistic inference using Bayesian belief networks,” *Artificial Intelligence*, vol. 42, pp. 393–405, 1990.

[9] P. Dagum and M. Luby, “Approximate probabilistic reasoning in Bayesian belief networks is NP-hard,” *Artificial Intelligence*, vol. 60, pp. 141–153, 1993.

[10] J.S. Yedidia, W. T. Freeman, and Y. Weiss, “Constructing free energy approximations and generalized belief propagation algorithms,” *IEEE Trans. Info. Theory*, vol. 51, no. 7, pp. 2282–2312, July 2005.

[11] S. Tatikonda and M. I. Jordan, “Loopy belief propagation and Gibbs measures,” in *Proc. Uncertainty in Artificial Intelligence*, August 2002, vol. 18, pp. 493–500.

[12] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, “Tree-based reparameterization framework for analysis of sum-product and related algorithms,” *IEEE Trans. Info. Theory*, vol. 49, no. 5, pp. 1120–1146, May 2003.

[13] Tom Heskes, “On the uniqueness of loopy belief propagation fixed points,” *Neural Computation*, vol. 16, no. 11, 2004.

[14] A. T. Ihler, J. W. Fisher III, and A. S. Wilsky, "Loopy belief propagation: Convergence and effects of message errors," *Journal of Machine Learning Research*, vol. 6, pp. 905–936, 2005.

[15] J. M. Mooij and H. J. Kappen, "Sufficient conditions for convergence of loopy belief propagation," Tech. Rep. arxiv:cs.IT:0504030, University of Nijmegen, April 2005, Submitted to IEEE Trans. Info. Theory.

[16] A. Yuille, "CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation," *Neural Computation*, vol. 14, pp. 1691–1722, 2002.

[17] M. Welling and Y. Teh, "Belief optimization: A stable alternative to loopy belief propagation," in *Uncertainty in Artificial Intelligence*, July 2001.

[18] T. Heskes, K. Albers, and B. Kappen, "Approximate inference and constrained optimization," in *Uncertainty in Artificial Intelligence*, July 2003, vol. 13, pp. 313–320.

[19] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "A new class of upper bounds on the log partition function," *IEEE Trans. Info. Theory*, vol. 51, no. 7, pp. 2313–2335, July 2005.

[20] W. Wiegerinck and T. Heskes, "Fractional belief propagation," in *NIPS*, 2002, vol. 12, pp. 438–445.

[21] W. Wiegerinck, "Approximations with reweighted generalized belief propagation," in *Workshop on Artificial Intelligence and Statistics*, January 2005.

[22] A. Levin and Y. Weiss, "Learning to combine bottom-up and top-down segmentation," in *European Conference on Computer Vision (ECCV)*, June 2006.

[23] S. Geman and D. Geman, "Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images," *IEEE Trans. PAMI*, vol. 6, pp. 721–741, 1984.

[24] M. J. Wainwright and M. I. Jordan, "A variational principle for graphical models," in *New Directions in Statistical Signal Processing*. MIT Press, Cambridge, MA, 2006.

[25] Y. Weiss, "Correctness of local probability propagation in graphical models with loops," *Neural Computation*, vol. 12, pp. 1–41, 2000.

[26] H. A. Bethe, "Statistics theory of superlattices," *Proc. Royal Soc. London, Series A*, vol. 150, no. 871, pp. 552–575, 1935.

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

[28] D. P. Bertsekas and J. N. Tsitsiklis, *Parallel and Distributed Computation: Numerical Methods*, Athena Scientific, Boston, MA, 1997.

[29] L. Yu. Kolotilina, "Bounds and inequalities for the perron root of a nonnegative matrix: II. circuit bounds and inequalities," *Journal of Mathematical Sciences*, vol. 127, 2005.

[30] L. Yu. Kolotilina, "Bounds for the singular values of a matrix involving its sparsity pattern," *Journal of Mathematical Sciences*, vol. 137, 2006.