CONVERGENCE ANALYSIS OF REWEIGHTED SUM-PRODUCT ALGORITHMS

Tanya Roosta[1] *and Martin J. Wainwright*[1,2]

Department of Electrical Engineering and Computer Sciences[1]
Department of Statistics[2]
University of California, Berkeley

ABSTRACT

Many signal processing applications of graphical models require efficient methods for computing (approximate) marginal probabilities over subsets of nodes in the graph. The intractability of this marginalization problem for general graphs with cycles motivates the use of approximate message-passing algorithms, including the sum-product algorithm and variants thereof. This paper studies the convergence and stability properties of the family of *reweighted sum-product algorithms*, a generalization of the standard updates in which messages are adjusted with graph-dependent weights. For homogenous 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 estimates of rates. These theoretical results are complemented with experimental simulations on various classes of graphs.

Index terms: Graphical model; Markov random field; belief propagation, sum-product algorithm; message-passing; approximate inference.

1. INTRODUCTION

Graphical models provide a powerful framework for capturing the complex statistical dependencies exhibited by various classes of real-world signals [1, 2]. A fundamental problem common to any signal processing application of a graphical model is that of computing marginal probabilities over subsets of nodes. This *marginalization problem*, though solvable in linear-time for tree-structured models, is computationally intractable for more general graphs with cycles. This difficulty motivates the use of efficient algorithms for computing approximate marginal probabilities in graphical models with cycles. A popular class of algorithms, including the sum-product algorithm [3] and extensions thereof [4], is based on passing "messages" between nodes in the graph. While computationally efficient, the standard form of sum-product message-passing is not guaranteed to converge, and in fact may have multiple fixed points.

Recent work has shed some light on the convergence properties of the ordinary sum-product algorithm. Tatikonda and Jordan [5] connect sum-product convergence to uniqueness of Gibbs measures on the computation tree. These results have been extended in follow-up work by other researchers [6, 7, 8]. At a high level, this line of research establishes that for sufficiently weak dependencies among the random variables in the graphical model, the sum-product updates have a unique fixed point, and will converge at a geometric rate. However, the sum-product algorithm is routinely applied to graphical models for which these theoretical guarantees are not applicable.

The family of *reweighted sum-product* algorithms [9, 10, 11] is a broader class of message-passing algorithms, in which messages are adjusted by edge-based weights determined by the graph structure. It includes the ordinary sum-product algorithm as a special case, in which all the weights are unity. For suitable choices of these weights, it can be shown [9] that reweighted sum-product—in sharp contrast to the ordinary updates— always has a unique fixed point for *any* graph and any dependency strength. An additional benefit of convexity is that the message-passing updates tend to be more stable, as confirmed by experimental investigation [11, 9, 12]. However, the convergence properties of reweighted message-passing have not yet been fully understood. Accordingly, the main contribution of this paper is convergence analysis of the family of reweighted sum-product algorithms.

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. In Section 3, we begin by stating our main results, including a discussion of how they are related to previous results on the ordinary sum-product algorithm. We then turn to the proofs of these claims. Section 4 provides experimental results to illustrate and support our experimental findings, and we conclude in Section 5 with a summary and directions for future work.

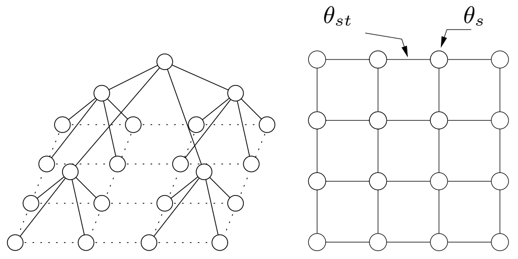
2. BACKGROUND

We begin by providing background on graphical models, and (reweighted) sum-product message-passing.

2.1. Graphical models

There exists a variety of graphical formalisms, including directed, undirected and factor graphs. Here we focus on *Markov random fields*, defined by an undirected graph G with vertices $V = \{1, \ldots, n\}$ and edge set E. Associated with each vertex $s \in V$ is a random variable X_s, taking values in some space $\mathcal{X}$. The random vector $X = (X_1, \ldots, X_n)$ is said to be Markov with respect to the graph if its distribution decomposes into a product of terms over the cliques of G. (A clique $C \subset V$ of a graph G is a fully-connected subset of vertices.)

This work 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. Additional support was provided by an Alfred P. Sloan Foundation Fellowship and NSF grant DMS-0528488 to MJW.



(a) (b)

Fig. 1. Examples of graphical models. (a) A quad-tree model used in multi-resolution signal processing [2]. (b) A lattice-based model used in image processing.

In this paper, we restrict attention to the case of pairwise cliques (an assumption which entails no loss of generality [4] for discrete spaces $\mathcal{X}$), for which the p.m.f. of X decomposes as

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

Here the quantities θ_s and θ_{st} are *potential functions* that depend only on the random variables X_s, and the pair (X_s, X_t) respectively.

Examples of such graphical models commonly used in signal processing include the (hidden) Markov model, and the quad tree model (Fig. 1(a)). In contrast to these graphs without cycles (in which exact calculations are computationally feasible), of primary interest in this paper are graphs with cycles, such as the lattice or grid-based model shown in Figure 1(b).

2.2. Reweighted sum-product message-passing

The sum-product algorithm is an iterative algorithm, in which nodes in the graph exchange 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. Here we describe the more general class of reweighted sum-product algorithms. For each edge, let $\rho_{st} \in [0,1]$ be an associated edge weight. Denoting by $M_{ts}(x_s)$ the message vector passed from node t to node s, the reweighted sum-product update equations (up to normalization) are given by

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

where $N(t)$ denotes the neighbors of node t in the graph. Setting the edge weights $\rho_{st} = 1$ for all edges recovers the standard sum-product updates. When the updates converge, the messages are used to compute (approximate) marginal probabilities τ_s at each node via

$$\tau_s(x_s) \quad \propto \quad \exp\{\theta_s(x_s)\} \prod_{t\in N(s)} [M_{ts}(x_s)]^{\rho_{st}}. \qquad (3)$$

Of interest is under what conditions the message updates (2) are guaranteed to converge.

3. CONVERGENCE ANALYSIS

In this section, we describe and sketch proofs on our main results on the convergence properties of the reweighted sum-product updates (2). For simplicity in exposition in this short report, we restrict our results to the case of binary random variables (i.e., $\mathcal{X} = \{-1, 1\}$), but note that our analysis can be extended to more general spaces. In the binary case, each singleton potential $\theta_s(\cdot)$ can be parameterized by a single real number, which we denote θ_s for convenience. Similarly, the pairwise potential $\theta_{st}(\cdot,\cdot)$ can be parameterized by a single real number θ_{st}.

3.1. Statement of main results

Our convergence analysis is based on establishing that, under suitable conditions, the reweighted updates (2) specify a contractive mapping in the ℓ_∞ norm. The contraction coefficient $K(G;\theta;\rho)$ that emerges from our analysis is defined by

$$\max_{(s,t)\in E} \left\{\sum_{u\in N(t)\backslash s} \rho_{ut}\frac{\exp(\frac{2\theta_{ut}}{\rho_{ut}})-1}{\exp(\frac{2\theta_{ut}}{\rho_{ut}})+1} + (1-\rho_{ts})\frac{\exp(\frac{2\theta_{st}}{\rho_{st}})-1}{\exp(\frac{2\theta_{st}}{\rho_{st}})+1}\right\}. \qquad (4)$$

With these definitions, we have:

Theorem 1. *For an arbitrary pairwise Markov random field, the condition* $K(G;\theta;\rho) < 1$ *is sufficient for convergence of the reweighted sum-product updates* (2)*.*

If we specialize this result to the case of uniform edge weights $\rho_{st} = 1$, which corresponds to the standard sum-product updates, then we recover previous results [6, 8] as a corollary.

It is worth noting that Theorem 1 is a somewhat conservative condition, in that it requires that the message updates be contractive at *every* node of the graph, as opposed to requiring that they be attractive in an average sense. For homogeneous models (in which $\theta_s = \theta$ and $\theta_{st} = \eta$ are constant across nodes and edges, and each node has degree d), we provide a sharpened analysis of convergence properties:

Theorem 2. *For any homogeneous binary model on a* d*-regular graph with arbitrary choice of* (θ,η)*, the reweighted updates have a unique fixed point and converge for all edge weights* ρ *such that* $(\rho d - 1) \leq 1$.

Note that as a corollary, when for the edge weight $\rho = 1$—the choice corresponding to the standard sum-product algorithm—then we recover the known result that sum-product converges for any single cycle graph ($d = 2$). For $\rho d > 2$, the updates may have multiple fixed points, but this depends on the choice of (θ,η), as we discuss in more detail in the sequel.

3.2. Proof of Theorem 1

We begin by re-writing the message update equation (2) in a form more amenable to analysis. For each edge (s,t), define the log message ratio $z_{ts} = \log\frac{M_{ts}(1)}{M_{ts}(-1)}$. It is equivalent to update these log ratios using the update function:

$$F_{ts}(z) := \log\frac{\exp\left(\frac{\theta_{st}}{\rho_{st}} + \theta_t + \rho_{st}(\sum_{v\in N(t)} z_{vt}) - z_{st}\right) + \exp\left(\frac{-\theta_{st}}{\rho_{st}} - \theta_t\right)}{\exp\left(\frac{-\theta_{st}}{\rho_{st}} + \theta_t + \rho_{st}(\sum_{v\in N(t)} z_{vt}) - z_{st}\right) + \exp\left(\frac{\theta_{st}}{\rho_{st}} - \theta_t\right)}. \qquad (5)$$

We begin with a lemma required in proving the theorem:

Lemma 3. *The partial derivative of F can be bounded as follows: for $u \in N(t)\backslash s$, we have*

$$\frac{\partial F_{ts}}{\partial z_{ut}}(z) \quad \leq \quad \rho_{ut} \; \frac{\exp(\frac{2\theta_{ut}}{\rho_{ut}}) - 1}{\exp|(\frac{2\theta_{ut}}{\rho_{ut}}) + 1} \tag{6}$$

whereas for edge (s,t), we have

$$\frac{\partial F_{ts}}{\partial z_{ut}}(z) \quad \leq \quad (1-\rho_{st}) \; \frac{\exp(\frac{2\theta_{st}}{\rho_{st}}) - 1}{\exp|(\frac{2\theta_{st}}{\rho_{st}}) + 1} \; . \tag{7}$$

The proof, omitted due to space constraints, is based on a Taylor series expansion, and some analysis to bound the second derivative term. Turning to the proof of Theorem 1, it is based on analyzing the evolution of the vector $z \in \mathbb{R}^{|E|}$ of log likelihood ratios associated with edges of the graph.

Lemma 4. *Consider a sequence of iterates $\{z^m\}$ generated by applying the update functions $\{F_{ts}\}$ in parallel to each edge. Let z^* be a fixed point of these updates, and let $\Delta^m = z^m - z^*$ be the difference at iteration m. Then for each edge $(s,t) \in E$, the following inequality holds at each iteration:*

$$|\Delta_{ts}^{m+1}| \quad \leq \quad \underset{u \in N(t)\backslash s}{} \quad \rho_{ut} L_{ut} |\Delta_{ut}^m| + (1-\rho_{ts}) L_{st} |\Delta_{st}^m|. \tag{8}$$

where for edge (u,v), the constant $L_{uv} := \frac{\exp(\frac{2\theta_{uv}}{\rho_{uv}})-1}{\exp|(\frac{2\theta_{uv}}{\rho_{uv}})+1}$.

Proof. Using the facts that $z^{m+1} = F(z^m)$ and $z^* = F(z^*)$ (since z^* is a fixed point), we have for each edge $(t,s) \in E$:

$$\begin{aligned} |\Delta_{ts}^{m+1}| &= |F_{ts}(z^m) - F_{ts}(z^*)| \\ &= \underset{u \in N(t)}{} \quad \frac{\partial F_{ts}}{\partial z_{ut}} \left(\alpha z^m + (1-\alpha) z^*\right) (z_{ut}^m - z_{ut}^*) \\ &\leq \underset{u \in N(t)}{} \quad \frac{\partial F_{ts}}{\partial z_{ut}} \left(\alpha z^m + (1-\alpha) z^*\right) \; |\Delta_{ut}^m| \, , \end{aligned}$$

where $\alpha \in (0,1)$. (In this second equality, we have applied the mean value theorem to F_{ts}.) Now applying our bounds on partial derivatives from Lemma 3, we obtain that

$$|\Delta_{ts}^{m+1}| \quad \leq \quad \underset{u \in N(t)\backslash s}{} \quad \rho_{ut} L_{ut} |\Delta_{ut}^m| + (1-\rho_{st}) L_{st} |\Delta_{st}^m|$$

as claimed. □

With this lemma in hand, we have the necessary ingredients to prove Theorem 1. From the error recursion (8), we have

$$\begin{aligned} \|\Delta^{m+1}\|_\infty &\leq \max_{(t,s) \in E} \quad \underset{u \in N(t)\backslash s}{} \quad \rho_{ut} L_{ut} |\Delta_{ut}^m| + (1-\rho) L_{st} |\Delta_{st}^m| \\ &\leq \max_{(t,s) \in E} \quad \underset{u \in N(t)\backslash s}{} \quad \rho_{ut} L_{ut} + (1-\rho) L_{st} \quad \|\Delta^m\|_\infty \\ &= K(G;\theta;\rho) \; \|\Delta^m\|_\infty . \end{aligned}$$

Consequently, if $K < 1$, then the mapping $F : \mathbb{R}^{|E|} \to \mathbb{R}^{|E|}$ is strictly contractive in the ℓ_∞-norm, which establishes the claim by standard fixed point results [13].

3.3. Proofs for homogeneous case

In the homogeneous Ising model, the edge weights θ_{st} are equal to a common value η, and similarly the node parameters θ_s are all equal to a common value θ. Under these assumptions 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

$$F(z;\eta,\theta\rho) \quad = \quad \log \; \frac{\exp(\frac{-\eta}{\rho} - \theta) + \exp(\frac{\eta}{\rho} + (\rho d - 1)z) + \theta}{\exp(\frac{\eta}{\rho} - \theta) + \exp(\frac{-\eta}{\rho} + (\rho d - 1)z) + \theta} \tag{9}$$

We analyze the behavior of the updates $z^{m+1} = F(z^m)$ by suitably controlling the derivative of F with respect to z. A straightforward calculation yields that $F'(z) = (\rho d - 1)(a - b)$ where

$$a \quad = \quad \frac{\exp(\frac{\eta}{\rho} + (\rho d - 1)z + \theta)}{\exp(\frac{-\eta}{\rho} - \theta) + \exp(\frac{\eta}{\rho} + (\rho d - 1)z + \theta)} \tag{10a}$$

$$b \quad = \quad \frac{\exp(\frac{-\eta}{\rho} + (\rho d - 1)z + \theta)}{\exp(\frac{\eta}{\rho} - \theta) + \exp(\frac{-\eta}{\rho} + (\rho d - 1)z + \theta)} \tag{10b}$$

Note that we have $0 < a, b < 1$ and $|a - b| < \max\{a, b\}$, from which we obtain

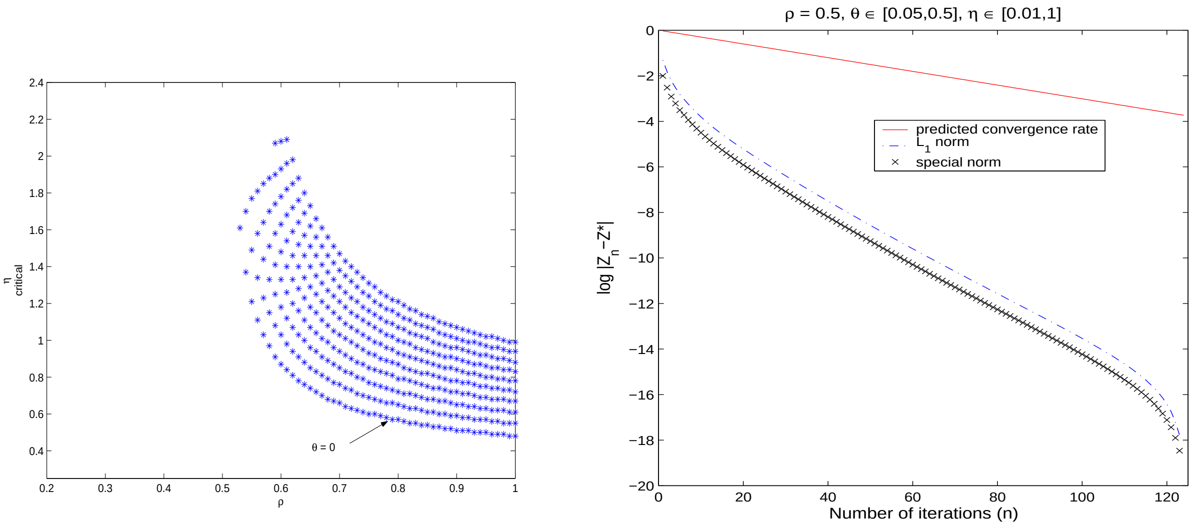
$$|F'(z;\eta,\theta)| \leq |(\rho d - 1)| \max\{a, b\| \leq |(\rho d - 1)|.$$

From this bound, we conclude that if $0 \leq (\rho d - 1) < 1$, then $|F'(z)| < 1$ for all $z \in \mathbb{R}$. From this fact, it follows that the update is a contraction on $\mathbb{R}$, and hence has a unique fixed point [13]. In the boundary case when $(\rho d - 1) = 1$, the fixed point equation (9) has only one valid root. Moreover, the derivative $F'(z)$ remains strictly less than one for all finite η and θ, so that we are again guaranteed uniqueness and convergence of the updates. (We omit the proofs of these claims due to space constraints.)

Finally, when $(\rho d - 1) > 1$, the update equation (9) may have more than one fixed point, depending on the choice of η and θ. Indeed, for a fixed setting of the node potential θ, it is possible to plot the critical value of η for which the second fixed point appears. For a d-regular graph with $d = 4$, Figure 2(a) provides a number of different curves, each corresponding to a different θ, showing how this critical value of η changes as the edge weight paramets is decreased from $\rho = 1$ (corresponding to the ordinary sum-product algorithm) to the critical value $\rho = 1/2$ (where $\rho d - 1 = 1$). Note that as one corner case, we recover the classical result that for $\theta = 0$, multiple fixed points appear in the ordinary sum-product algorithm as soon as $\eta \geq \eta_{\text{crit}} \approx 0.3466$.

4. EXPERIMENTAL RESULTS

In this section, we present the results of experimental simulations to illustrate and support our theoretical findings. The simulations were applied to the Ising model, obtaining by the potential function settings $\theta_s(x_s) = \theta_s x_s$ and $\theta_{st}(x_s, x_t) = \theta_{st} x_s x_t$ in equation (1). The numbers parameterizing the node potentials, θ_s, were chosen uniformly from $[0.05, 0.5]$, and the edge potentials, θ_{st}, were chosen uniformly from $[0.01, 1]$. The simulations were carried out for different values of the edge weights ρ in the reweighted sum-product algorithm; however, so as to appropriately narrow the space, we restricted attention to the case where the value of ρ is the same for all edges. Due to space constraints, here we show only results for the lattice with $n = 144$ nodes, and $\rho = 0.5$. Figure 2(b) compares the convergence rate as predicted by Theorem 1 vs. the true convergence



(a) (b)

Fig. 2. (a) Plots of the appearance of multiple fixed points versus η and ρ. Each curve shows, for a fixed node potential θ, the critical value of $\eta_{\rm crit}$ at which multiple fixed points occur, for edge weights ranging from $\rho = 1$ (ordinary sum-product) down to $\rho = 1/2$. (b) The rate of convergence of the reweighted sum-product algorithm as compared to the rate predicted by Theorem 1.

rate. We have plotted $\log|z_m - z^*|_1$ vs. the number of iterations m. In this setting, z^* is the fixed point of the reweighted sum-product algorithm, and z_m the message vector at the nth iteration. As illustrated by Figure 2(b), the true convergence rate is faster than the predicted value by Theorem 1; this finding both validates our result, and reveals that our analysis appears to be overly conservative (as discussed earlier).

5. CONCLUSION

Many signal processing applications make use of graphical models. This 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 homogenous 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.

6. REFERENCES

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

[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] F.R. Kschischang, B.J. Frey, and H.-A. Loeliger, “Factor graphs and the sum-product algorithm,” *IEEE Trans. Info. Theory*, vol. 47, no. 2, pp. 498–519, February 2001.

[4] 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.

[5] 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.

[6] 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.

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

[8] 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.

[9] 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.

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

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

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

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