Journal of Machine Learning Research 6 (2005) 905-936 Submitted 11/04; Revised 4/05; Published 5/05

Loopy Belief Propagation: Convergence and Effects of Message Errors

Alexander T. Ihler IHLER@ALUM.MIT.EDU
Donald Bren School of Information and Computer Science, U.C. Irvine
Irvine, CA 92697 USA

John W. Fisher III FISHER@CSAIL.MIT.EDU

Alan S. Willsky WILLSKY@MIT.EDU
Massachusetts Institute of Technology
Cambridge, MA 02139, USA

Editor: Max Chickering

Abstract

Belief propagation (BP) is an increasingly popular method of performing approximate inference on arbitrary graphical models. At times, even further approximations are required, whether due to quantization of the messages or model parameters, from other simplified message or model representations, or from stochastic approximation methods. The introduction of such errors into the BP message computations has the potential to affect the solution obtained adversely. We analyze the effect resulting from message approximation under two particular measures of error, and show bounds on the accumulation of errors in the system. This analysis leads to convergence conditions for traditional BP message passing, and both strict bounds and estimates of the resulting error in systems of approximate BP message passing.

Keywords: Belief propagation, sum-product, convergence, approximate inference, quantization

1. Introduction

Graphical models and message-passing algorithms defined on graphs comprise a growing field of research. In particular, the *belief propagation* (or sum-product) algorithm has become a popular means of solving inference problems exactly or approximately. One part of its appeal lies in its optimality for tree-structured graphical models (models which contain no loops). However, its is also widely applied to graphical models with cycles. In these cases it may not converge, and if it does its solution is approximate; however in practice these approximations are often good. Recently, some additional justifications for loopy belief propagation have been developed, including a handful of convergence results for graphs with cycles (Weiss, 2000; Tatikonda and Jordan, 2002; Heskes, 2004).

The approximate nature of loopy belief propagation is often a more than acceptable price for performing efficient inference; in fact, it is sometimes desirable to make *additional* approximations. There may be a number of reasons for this—for example, when exact message representation is computationally intractable, the messages may be approximated stochastically (Koller et al., 1999) or deterministically by discarding low-likelihood states (Coughlan and Ferreira, 2002). For belief propagation involving continuous, non-Gaussian potentials, some form of approximation is required to obtain a finite parameterization for the messages (Sudderth et al., 2003; Isard, 2003; Minka, 2001). Additionally, simplification of complex graphical models through edge removal, quantization of the potential functions, or other forms of distributional approximation may be considered in this framework. Finally, one may wish to approximate the messages and reduce their representation size for another reason—to decrease the communications required for distributed inference applications. In distributed message passing, one may approximate the transmitted message to reduce its representational cost (Ihler et al., 2004a), or discard it entirely if it is deemed "sufficiently similar" to the previously sent version (Chen et al., 2004). Through such means one may significantly reduce the amount of communications required.

Given that message approximation may be desirable, we would like to know what effect the errors introduced have on our overall solution. In order to characterize the approximation effects in graphs with cycles, we analyze the deviation from the solution given by "exact" loopy belief propagation (*not*, as is typically considered, the deviation of loopy BP from the true marginal distributions). As a byproduct of this analysis, we also obtain some results on the convergence of loopy belief propagation.

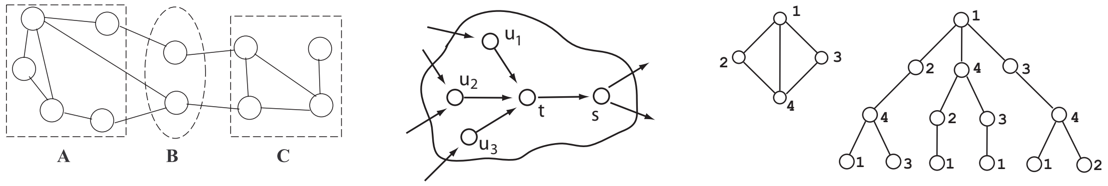
We begin in Section 2 by briefly reviewing the relevant details of graphical models and belief propagation. Section 4 then examines the consequences of measuring a message error by its dynamic range. In particular, we explain the utility of this measure and its behavior with respect to the operations of belief propagation. This allows us to derive conditions for the convergence of traditional loopy belief propagation, and bounds on the distance between any pair of BP fixed points (Sections 5.1–5.2), and these results are easily extended to many approximate forms of BP (Section 5.3). If the errors introduced are independent, as is a typical assumption in, for example, quantization analysis (Gersho and Gray, 1991; Willsky, 1978), tighter estimates of the resulting error can be obtained (Section 5.5).

It is also instructive to examine other measures of message error, in particular ones which emphasize more average-case (as opposed to pointwise or worst-case) differences. To this end, we consider a KL-divergence based measure in Section 6. While the analysis of the KL-divergence measure is considerably more difficult and does not lead to strict guarantees, it serves to give some intuition into the behavior of perturbed BP under an average-case difference measure.

2. Graphical Models

Graphical models (Lauritzen, 1996; Kschischang et al., 2001) provide a convenient means of representing conditional independence relations among large numbers of random variables. Specifically, each node s in an undirected graph is associated with a random variable x_s, while the set of edges $\mathcal{E}$ is used to describe the conditional dependency structure of the variables through *graph separation*. If every path between two sets A and C passes through another set B [see Figure 1(a)], the sets of variables $\mathbf{x}_A = \{x_s : s \in A\}$ and $\mathbf{x}_C = \{x_s : s \in C\}$ must be independent given the values of $\mathbf{x}_B = \{x_s : s \in B\}$. Thus, the distribution $p(\mathbf{x}_A, \mathbf{x}_B, \mathbf{x}_C)$ can be written in the form $p(\mathbf{x}_B)p(\mathbf{x}_A|\mathbf{x}_B)p(\mathbf{x}_C|\mathbf{x}_B)$.

It can be shown that a distribution $p(\mathbf{x})$ is consistent with (i.e., satisfies the conditional independence relations specified by) an undirected graph if it factors into a product of



(a) (b) (c)

Figure 1: (a) Graphical models describe statistical dependency; here, the sets A and C are independent given B. (b) BP propagates information from t and its neighbors u_i is to s by a simple message-passing procedure; this procedure is exact on a tree, but approximate in graphs with cycles. (c) For a graph with cycles, one may show an equivalence between n iterations of loopy BP and the depth-n computation tree [shown here for $n = 3$ and rooted at node 1; example from Tatikonda and Jordan (2002)].

potential functions ψ defined on the cliques (fully-connected subsets) of the graph, and that the converse is also true if $p(\mathbf{x})$ is strictly positive (Clifford, 1990). For convenience, we confine our attention to graphical models with at most pairwise potential functions, so that the distribution factors according to

$$p(\mathbf{x}) = \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_s \psi_s(x_s)$$

This is a typical assumption for belief propagation, and can be taken without incurring any real loss of generality since a graphical model with higher-order potential functions may always be converted to a graphical model with only pairwise potential functions through a process of variable augmentation, though this may also increase the nodes' state dimension undesirably; see, for example, Weiss (2000).

2.1 Belief Propagation

The goal of belief propagation (BP) (Pearl, 1988), also called the sum-product algorithm, is to compute the marginal distribution $p(x_t)$ at each node t. BP takes the form of a message-passing algorithm between nodes, expressed in terms of an update to the outgoing message at iteration i from each node t to each neighbor s in terms of the previous iteration's incoming messages from t's neighbors Γ_t [see Figure 1(b)],

$$m^i_{ts}(x_s) \propto \int \psi_{ts}(x_t, x_s)\psi_t(x_t) \prod_{u\in\Gamma_t\setminus s} m^{i-1}_{ut}(x_t)dx_t \tag{1}$$

Typically each message is normalized so as to integrate to unity (and we assume that such normalization is possible). For discrete-valued random variables, of course, the integral is replaced by a summation. At any iteration, one may calculate the *belief* at node t by

$$M^i_t(x_t) \propto \psi_t(x_t) \prod_{u\in\Gamma_t} m^i_{ut}(x_t) \tag{2}$$

For tree-structured graphical models, belief propagation can be used to efficiently perform exact marginalization. Specifically, the iteration (1) converges in a finite number of iterations (at most the length of the longest path in the graph), after which the belief (2) equals the correct marginal $p(x_t)$. However, as observed by Pearl (1988), one may also apply belief propagation to arbitrary graphical models by following the same *local* message passing rules at each node and ignoring the presence of cycles in the graph; this procedure is typically referred to as "loopy" BP.

For loopy BP, the sequence of messages defined by (1) is not guaranteed to converge to a fixed point after any number of iterations. Under relatively mild conditions, one may guarantee the existence of fixed points (Yedidia et al., 2004). However, they may not be unique, nor are the results exact [the belief M_t^i does not converge to the true marginal $p(x_t)$]. In practice however the procedure often arrives at a reasonable set of approximations to the correct marginal distributions.

2.2 Computation Trees

It is sometimes convenient to think of loopy BP in terms of its *computation tree*. Tatikonda and Jordan (2002) showed that the effect of n iterations of loopy BP at any particular node s is equivalent to exact inference on a tree-structured 'unrolling" of the graph from s. A small graph, and its associated 4-level computation tree rooted at node `1`, are shown in Figure 1(c).

The computation tree with depth n consists of all length-n paths emanating from s in the original graph which do not immediately backtrack (though they may eventually repeat nodes).[1] We draw the computation tree as consisting of a number of *levels*, corresponding to each node in the tree's distance from the root, with the root node at level 0 and the leaf nodes at level n. Each level may contain multiple replicas of each node, and thus there are potentially many replicas of each node in the graph. The root node s has replicas of all neighbors Γ_s in the original graph as children, while all other nodes have replicas of all neighbors except their parent node as children.

Each edge in the computation tree corresponds to both an edge in the original graph *and* an iteration in the BP message-passing algorithm. Specifically, assume an equivalent initialization of both the loopy graph and computation tree—i.e., the initial messages m_{ut}^0 in the loopy graph are taken as inputs to the leaf nodes. Then, the upward messages from level n to level $n-1$ match the messages m_{ut}^1 in the first iteration of loopy BP, and more generally, a upward message m_{ut}^i on the computation tree which originates from a node u on level $n-i+1$ to its parent node t on level $n-i$ is identical to the message from node u to node t in the i^{th} iteration of loopy BP (out of n total iterations) on the original graph. Thus, the incoming messages to the root node (level 0) correspond to the messages in the n^{th} iteration of loopy BP.

2.3 Message Approximations

Let us now consider the concept of *approximate* BP messages. We begin by assuming that the "true" messages $m_{ts}(x_s)$ are some fixed point of BP, so that $m_{ts}^i = m_{ts}^{i+1}$. We may ask what happens when these messages are perturbed by some (perhaps small) error function $e_{ts}(x_s)$. Although there are certainly other possibilities, the fact that BP messages

1. Thus in Figure 1(c), the computation tree includes the sequence `1` $-$ `2` $-$ `4` $-$ `1`, but not the sequence `1` $-$ `2` $-$ `4` $-$ `2`.

are combined by taking their product makes it natural to consider multiplicative message deviations (or additive in the log-domain):

$$\hat{m}^i_{ts}(x_s) = m_{ts}(x_s) e^i_{ts}(x_s)$$

To facilitate our analysis, we split the message update operation (1) into two parts. In the first, we focus on the message *products*

$$\hat{M}^i_{ts}(x_t) \propto \psi_t(x_t) \prod_{u \in \Gamma_t \setminus s} \hat{m}^i_{ut}(x_t) \qquad\qquad \hat{M}^i_t(x_t) \propto \psi_t(x_t) \prod_{u \in \Gamma_t} \hat{m}^i_{ut}(x_t) \tag{3}$$

where the proportionality constant is chosen to normalize $\hat{M}$. The second operation, then, is the message *convolution*

$$\hat{m}^{i+1}_{ts}(x_s) \propto \int \psi_{ts}(x_s, x_t) \hat{M}^i_{ts}(x_t) dx_t \tag{4}$$

where again $\hat{M}$ is a normalized message or product of messages.

In this paper, we use the convention that lowercase quantities ($m_{ts}, e_{ts}, \ldots$) refer to messages and message errors, while uppercase ones ($M_{ts}, E_{ts}, M_t, \ldots$) refer to their products—at node t, the product of all incoming messages and the local potential is denoted $M_t(x_t)$, its approximation $\hat{M}_t(x_t) = M_t(x_t)E_t(x_t)$, with similar definitions for M_{ts}, $\hat{M}_{ts}$, and E_{ts}.

3. Overview of Results

To orient the reader, we lay out the order and general results which are obtained in this paper. We begin in Section 4 by examining a *dynamic range* measure $d\,(e)$ of the variability of a message error $e(x)$ (or more generally of any function) and show how this measure behaves with respect to the BP equations (1) and (2). Specifically, we show in Section 4.2 that the measure $\log d\,(e)$ is sub-additive with respect to the product operation (3), and contractive with respect to the convolution operation (4).

Applying these results to traditional belief propagation results in a new sufficient condition for BP convergence (Section 5.1), specifically

$$\max_{s,t} \sum_{u \in \Gamma_t \setminus s} \frac{d\,(\psi_{ut})^2 - 1}{d\,(\psi_{ut})^2 + 1} < 1; \tag{5}$$

and this condition may be further improved in many cases. The condition (5) can be shown to be slightly stronger than the sufficient condition given in Tatikonda and Jordan (2002), and empirically appears to be stronger than that of Heskes (2004). In experiments, the condition appears to be tight (exactly predicting uniqueness or non-uniqueness of fixed points) for at least some problems, such as binary–valued random variables with attractive potentials. More importantly, however, the *method* in which it is derived allows us to generalize to many other situations:

1. Using the same methodology, we may demonstrate that any two BP fixed points must be within a ball of a calculable diameter; the condition (5) is equivalent to this diameter being zero (Section 5.2).

2. Both the diameter of the bounding ball and the convergence criterion (5) are easily improved for graphical models with irregular geometry or potential strengths, leading to better conditions on graphs which are more "tree-like" (Section 5.3).

3. The same analysis may also be applied to the case of quantized or otherwise approximated messages and models (potential functions), yielding bounds on the resulting error (Section 5.4).

4. If we regard the message errors as a stochastic process, a similar analysis with a few additional, intuitive assumptions gives alternate, tighter estimates (though not necessarily bounds) of performance (Section 5.5).

Finally, in Section 6 we perform the same analysis for a less strict measure of message error [i.e., disagreement between a message $m(x)$ and its approximation $\hat{m}(x)$], namely the Kullback-Leibler divergence. This analysis shows that, while failing to provide strict bounds in several key ways, one is still able to obtain some intuition into the behavior of approximate message passing under an average-case difference measure.

In the next few sections, we first describe the dynamic range measure and discuss some of its salient properties (Section 4). We then apply these properties to analyze the behavior of loopy belief propagation (Section 5). Almost all proofs are given in an in-line fashion, as they frequently serve to give intuition into the method and meaning of each result.

4. Dynamic Range Measure

In order to discuss the effects and propagation of errors, we first require a measure of the difference between two messages. In this section, we examine the following measure on $e_{ts}(x_s)$: let $d\,(e_{ts})$ denote the function's *dynamic range*[2], specifically

$$d\,(e_{ts}) = \sup_{a,b} \sqrt{e_{ts}(a)/e_{ts}(b)} \tag{6}$$

Then, we have that $m_{ts} \equiv \hat{m}_{ts}$ (i.e., the pointwise equality condition $m_{ts}(x) = \hat{m}_{ts}(x) \forall x$) if and only if $\log d\,(e_{ts}) = 0$. Figure 2 shows an example of $m(x)$ and $\hat{m}(x)$ along with their associated error $e(x)$.

4.1 Motivation

We begin with a brief motivation for this choice of error measure. It has a number of desirable features; for example, it is directly related to the pointwise log error between the two distributions.

Lemma 1. *The dynamic range measure* (6) *may be equivalently defined by*

$$\log d\,(e_{ts}) = \inf_{\alpha} \sup_{x} |\log \alpha m_{ts}(x) - \log \hat{m}_{ts}(x)| = \inf_{\alpha} \sup_{x} |\log \alpha - \log e_{ts}(x)|$$

2. This measure has also been independently investigated to provide a stability analysis for the max-product algorithm in Bayes' nets (acyclic, directed graphical models) (Chan and Darwiche, 2005). While similar in some ways, the analysis for acyclic graphs is considerably simpler; loopy graphs require demonstrating a rate of contraction, which we show is possible for the sum-product algorithm (Theorem 8).

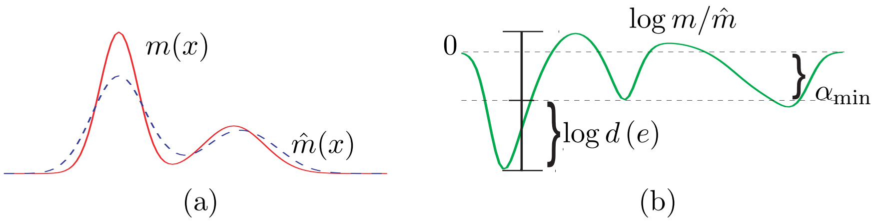


Figure 2: (a) A message $m(x)$ and an example approximation $\hat{m}(x)$; (b) their log-ratio $\log m(x)/\hat{m}(x)$, and the error measure $\log d\,(e)$.

Proof. The minimum is given by $\log \alpha = \frac{1}{2}(\sup_a \log e_{ts}(a) + \inf_b \log e_{ts}(b))$, and thus the right-hand side is equal to $\frac{1}{2}(\sup_a \log e_{ts}(a) - \inf_b \log e_{ts}(b))$, or $\frac{1}{2}(\sup_{a,b} \log e_{ts}(a)/e_{ts}(b))$, which by definition is $\log d\,(e_{ts})$. $\square$

The scalar α serves the purpose of "zero-centering" the function $\log e_{ts}(x)$ and making the measure invariant to simple rescaling. This invariance reflects the fact that the scale factor for BP messages is essentially arbitrary, defining a class of equivalent messages. Although the scale factor cannot be completely ignored, it takes on the role of a nuisance parameter. The inclusion of α in the definition of Lemma 1 acts to select particular elements of the equivalence classes (with respect to rescaling) from which to measure distance—specifically, choosing the closest such messages in a log-error sense. The log-error, dynamic range, and the minimizing α are depicted in Figure 2.

Lemma 1 allows the dynamic range measure to be related directly to an approximation error in the log-domain when both messages are normalized to integrate to unity, using the following theorem:

Theorem 2. *The dynamic range measure can be used to bound the log-approximation error:*

$$|\log m_{ts}(x) - \log \hat{m}_{ts}(x)| \le 2 \log d\,(e_{ts}) \qquad \forall x$$

Proof. We first consider the magnitude of $\log \alpha$:

$$\begin{aligned}
\forall x, \qquad & \left|\log \frac{\alpha m_{ts}(x)}{\hat{m}_{ts}(x)}\right| \le \log d\,(e_{ts}) \\
\Rightarrow \qquad & \frac{1}{d\,(e_{ts})} \le \frac{\alpha m_{ts}(x)}{\hat{m}_{ts}(x)} \le d\,(e_{ts}) \\
\Rightarrow \qquad & \int \hat{m}_{ts}(x) dx \frac{1}{d\,(e_{ts})} \le \alpha \int m_{ts}(x) dx \le \int \hat{m}_{ts}(x) dx d\,(e_{ts})
\end{aligned}$$

and since the messages are normalized, $|\log \alpha| \le \log d\,(e_{ts})$. Then by the triangle inequality,

$$|\log m_{ts}(x) - \log \hat{m}_{ts}(x)| \le |\log \alpha m_{ts}(x) - \log \hat{m}_{ts}(x)| + |\log \alpha| \le 2 \log d\,(e_{ts})\,. \qquad \square$$

In this light, our analysis of message approximation (Section 5.4) may be equivalently regarded as a statement about the required quantization level for an accurate implementation of loopy belief propagation. Interestingly, it may also be related to a floating-point precision on $m_{ts}(x)$.

Lemma 3. *Let $\hat{m}_{ts}(x)$ be an F-bit mantissa floating-point approximation to $m_{ts}(x)$. Then,* $\log d\,(e_{ts}) \le 2^{-F} + \mathcal{O}(2^{-2F})$.

Proof. For an F-bit mantissa, we have $|m_{ts}(x) - \hat{m}_{ts}(x)| < 2^{-F} \cdot 2^{\lfloor \log_2 m_{ts}(x) \rfloor} \le 2^{-F} \cdot m_{ts}(x)$. Then, using the Taylor expansion of $\log\left[1 + (\frac{\hat{m}}{m} - 1)\right] \approx (\frac{\hat{m}}{m} - 1)$ we have that

$$\begin{aligned}
\log d\,(e_{ts}) &\le \sup_x \left| \log \frac{\hat{m}(x)}{m(x)} \right| \\
&\le \sup_x \frac{\hat{m}(x) - m(x)}{m(x)} + \mathcal{O}\left(\left(\sup_x \frac{\hat{m}(x) - m(x)}{m(x)} \right)^2 \right) \\
&\le 2^{-F} + \mathcal{O}\left(2^{-2F}\right)
\end{aligned}$$ □

Thus our measure of error is, to first order, similar to the typical measure of precision in floating-point implementations of belief propagation on microprocessors. We may also relate $d\,(e)$ to other measures of interest, such as the Kullback-Leibler (KL) divergence:

Lemma 4. *The KL-divergence satisfies the inequality* $D(m_{ts} \| \hat{m}_{ts}) \le 2 \log d\,(e_{ts})$

Proof. By Theorem 2, we have

$$D(m_{ts} \| \hat{m}_{ts}) = \int m_{ts}(x) \log \frac{m_{ts}(x)}{\hat{m}_{ts}(x)} dx \le \int m_{ts}(x) \left(2 \log d\,(e_{ts})\right) dx = 2 \log d\,(e_{ts})$$ □

Finally, a bound on the dynamic range or the absolute log-error can also be used to develop confidence intervals for the maximum and median of the distribution.

Lemma 5. *Let $\hat{m}(x)$ be an approximation of $m(x)$ with $\log d\,(\hat{m}/m) \le \epsilon$, so that*

$$\hat{m}^+(x) = \exp(2\epsilon)\hat{m}(x) \qquad\qquad \hat{m}^-(x) = \exp(-2\epsilon)\hat{m}(x)$$

are upper and lower pointwise bounds on $m(x)$, respectively. Then we have a confidence region on the maximum of $m(x)$ given by

$$\arg\max_x m(x) \in \{x : \hat{m}^+(x) \ge \max_y \hat{m}^-(y)\}$$

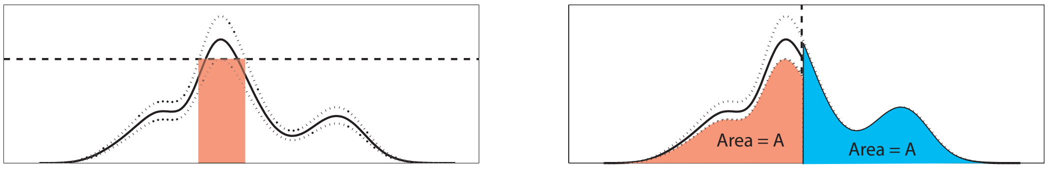
and an upper bound μ on the median of $m(x)$, i.e.,

$$\int_{-\infty}^{\mu} m(x) \ge \int_{\mu}^{\infty} m(x) \qquad \textit{where} \qquad \int_{-\infty}^{\mu} \hat{m}^-(x) = \int_{\mu}^{\infty} \hat{m}^+(x)$$

with a similar lower bound.

Proof. The definitions of $\hat{m}^+$ and $\hat{m}^-$ follow from Theorem 2. Given these bounds, the maximum value of $m(x)$ must be larger than the maximum value of $\hat{m}^-(x)$, and this is only possible at locations x for which $\hat{m}^+(x)$ is also greater than the maximum of $\hat{m}^-$. Similarly, the left integral of $m(x)$ ($-\infty$ to μ) must be larger than the integral of $\hat{m}^-(x)$, while the right integral (μ to ∞) must be smaller than for $\hat{m}^+(x)$. Thus the median of $m(x)$ must be less than μ. □

These bounds and confidence intervals are illustrated in Figure 3: given the approximate message $\hat{m}$ (solid black), a bound on the error yields $\hat{m}^+(x)$ and $\hat{m}^-(x)$ (dotted lines), which yield confidence regions on the maximum and median values of $m(x)$.



Confidence Region on Maximum
(a)

(Right boundary of) Conf. Region on Median
(b)

Figure 3: Using the error measure (6) to find confidence regions on maximum and median locations of a distribution. The distribution estimate $\hat{m}(x)$ is shown in solid black, with $|\log m(x)/\hat{m}(x)| \leq \frac{1}{4}$ bounds shown as dotted lines. Then, the maximum value of $m(x)$ must lie above the shaded region, and the median value is less than the dashed vertical line; a similar computation gives a lower bound.

4.2 Additivity and Error Contraction

We now turn to the properties of our dynamic range measure with respect to the operations of belief propagation. First, we consider the error resulting from taking the product (3) of a number of incoming approximate messages.

Theorem 6. *The log of the dynamic range measure is sub-additive:*

$$\log d\left(E^i_{ts}\right) \leq \sum_{u \in \Gamma_t \setminus s} \log d\left(e^i_{ut}\right) \qquad\qquad \log d\left(E^i_{t}\right) \leq \sum_{u \in \Gamma_t} \log d\left(e^i_{ut}\right)$$

Proof. We show the left-hand sub-additivity statement; the right follows from a similar argument. By definition, we have

$$\log d\left(E^i_{ts}\right) = \log d\left(\hat{M}^i_{ts}/M^i_{ts}\right) = \frac{1}{2}\log \sup_{a,b} \prod e^i_{ut}(a) / \prod e^i_{ut}(b)$$

Increasing the number of degrees of freedom gives

$$\leq \frac{1}{2}\log \prod \sup_{a_u,b_u} e^i_{ut}(a_u)/e^i_{ut}(b_u) = \sum \log d\left(e^i_{ut}(x)\right) \qquad \square$$

Theorem 6 allows us to bound the error resulting from a combination of the incoming approximations from two different neighbors of the node t. It is also important that $\log d\,(e)$ satisfy the triangle inequality, so that the application of two successive approximations results in an error which is bounded by the sum of their respective errors.

Theorem 7. *The log of the dynamic range measure satisfies the triangle inequality:*

$$\log d\left(e_1 e_2\right) \leq \log d\left(e_1\right) + \log d\left(e_2\right)$$

Proof. This follows from the same argument as Theorem 6. $\square$

We may also derive a minimum rate of contraction occurring with the convolution operation (4). We characterize the strength of the potential ψ_{ts} by extending the definition of the dynamic range measure:

$$d\left(\psi_{ts}\right)^2 = \sup_{a,b,c,d} \frac{\psi_{ts}(a,b)}{\psi_{ts}(c,d)} \tag{7}$$

When this quantity is finite, it represents a minimum rate of *mixing* for the potential, and thus causes a contraction on the error. This fact is exhibited in the following theorem:

Theorem 8. *When $d\left(\psi_{ts}\right)$ is finite, the dynamic range measure satisfies a rate of contraction:*

$$d\left(e_{ts}^{i+1}\right) \le \frac{d\left(\psi_{ts}\right)^2 d\left(E_{ts}^i\right) + 1}{d\left(\psi_{ts}\right)^2 + d\left(E_{ts}^i\right)}. \tag{8}$$

Proof. See Appendix A. □

Two limits are of interest. First, if we examine the limit as the potential strength $d(\psi)$ grows, we see that the error cannot increase due to convolution with the pairwise potential ψ. Similarly, if the potential strength is finite, the outgoing error cannot be arbitrarily large (independent of the size of the incoming error).

Corollary 9. *The outgoing message error $d\left(e_{ts}\right)$ is bounded by*

$$d\left(e_{ts}^{i+1}\right) \le d\left(E_{ts}^i\right) \qquad\qquad d\left(e_{ts}^{i+1}\right) \le d\left(\psi_{ts}\right)^2$$

Proof. Let $d\left(\psi_{ts}\right)$ or $d\left(E_{ts}^i\right)$ tend to infinity in Theorem 8. □

The contractive bound (8) is shown in Figure 4, along with the two simpler bounds of Corollary 9, shown as straight lines. Moreover, we may evaluate the asymptotic behavior by considering the derivative

$$\frac{\partial}{\partial d\left(E\right)} \left. \frac{d\left(\psi\right)^2 d\left(E\right) + 1}{d\left(E\right) + d\left(\psi\right)^2} \right|_{d(E)\to 1} = \frac{d\left(\psi\right)^2 - 1}{d\left(\psi\right)^2 + 1} = \tanh(\log d\left(\psi\right))$$

The limits of this bound are quite intuitive: for $\log d\left(\psi\right) = 0$ (independence of x_t and x_s), this derivative is zero; increasing the error in incoming messages m_{ut}^i has no effect on the error in m_{ts}^{i+1}. For $d\left(\psi\right) \to \infty$, the derivative approaches unity, indicating that for very large $d\left(\psi\right)$ (strong potentials) the propagated error can be nearly unchanged.

We may apply these bounds to investigate the behavior of BP in graphs with cycles. We begin by examining loopy belief propagation with exact messages, using the previous results to derive a new sufficient condition for BP convergence to a unique fixed point. When this condition is not satisfied, we instead obtain a bound on the relative distances between any two fixed points of the loopy BP equations. This allows us to consider the effect of introducing additional errors into the messages passed at each iteration, showing sufficient conditions for this operation to converge, and a bound on the resulting error from exact loopy BP.

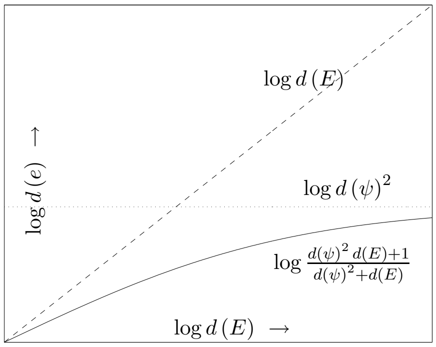


Figure 4: Three bounds on the error output $d(e)$ as a function of the error on the product of incoming messages $d(E)$.

5. Applying Dynamic Range to Graphs with Cycles

In this section, we apply the framework developed in Section 4, along with the computation tree formalism of Tatikonda and Jordan (2002), to derive results on the behavior of traditional belief propagation (in which messages and potentials are represented exactly). We then use the same methodology to analyze the behavior of loopy BP for quantized or otherwise approximated messages and potential functions.

5.1 Convergence of Loopy Belief Propagation

The work of Tatikonda and Jordan (2002) showed that the convergence and fixed points of loopy BP may be considered in terms of a Gibbs measure on the graph's computation tree. In particular, this led to the result that loopy BP is guaranteed to converge if the graph satisfies Dobrushin's condition (Georgii, 1988). Dobrushin's condition is a global measure, and difficult to verify; given in Tatikonda and Jordan (2002) is the easier to check sufficient condition (often called Simon's condition),

Theorem 10 (Simon's condition). *Loopy belief propagation is guaranteed to converge if*

$$\max_t \sum_{u \in \Gamma_t} \log d(\psi_{ut}) < 1 \tag{9}$$

where $d(\psi)$ is defined as in (7).

Proof. See Tatikonda and Jordan (2002). □

Using the previous section's analysis, we obtain the following, stronger condition, and (after the proof) show analytically how the two are related.

Theorem 11 (BP convergence). *Loopy belief propagation is guaranteed to converge if*

$$\max_{(s,t)\in\mathcal{E}} \sum_{u \in \Gamma_t \setminus s} \frac{d(\psi_{ut})^2 - 1}{d(\psi_{ut})^2 + 1} < 1 \tag{10}$$

Proof. By induction. Let the "true" messages m_{ts} be any fixed point of BP, and consider the incoming error observed by a node t at level $n-1$ of the computation tree (corresponding to the first iteration of BP), and having parent node s. Suppose that the total incoming error $\log d\left(E_{ts}^1\right)$ is bounded above by some constant $\log \epsilon^1$ for all $(t,s) \in \mathcal{E}$. Note that this is trivially true (for any n) for the constant $\log \epsilon^1 = \max_t \sum_{u \in \Gamma_t} \log d\left(\psi_{ut}\right)^2$, since the error on any message m_{ut} is bounded above by $d\left(\psi_{ut}\right)^2$.

Now, assume that $\log d\left(E_{ut}^i\right) \leq \log \epsilon^i$ for all $(u,t) \in \mathcal{E}$. Theorem 8 bounds the maximum log-error $\log d\left(E_{ts}^{i+1}\right)$ at any replica of node t with parent s, where s is on level $n-i$ of the tree (which corresponds to the i^{th} iteration of loopy BP) by

$$\log d\left(E_{ts}^{i+1}\right) \leq g_{ts}(\log \epsilon^i) = G_{ts}(\epsilon^i) = \sum_{u \in \Gamma_t \setminus s} \log \frac{d\left(\psi_{ut}\right)^2 \epsilon^i + 1}{d\left(\psi_{ut}\right)^2 + \epsilon^i} \tag{11}$$

We observe a contraction of the error between iterations i and $i+1$ if the bound $g_{ts}(\log \epsilon^i)$ is smaller than $\log \epsilon^i$ for every $(t,s) \in \mathcal{E}$, and asymptotically achieve $\log \epsilon^i \to 0$ if this is the case for any value of $\epsilon^i > 1$.

Defining $z = \log \epsilon$, we may equivalently show $g_{ts}(z) < z$ for all $z > 0$. This can be guaranteed by the conditions $g_{ts}(0) = 0$, $g'_{ts}(0) < 1$, and $g''_{ts}(z) \leq 0$ for each t, s. The first is easy to verify, as is the last (term by term) using the identity $g''_{ts}(z) = \epsilon^2 G''_{ts}(\epsilon) + \epsilon G'_{ts}(\epsilon)$; the second ($g'_{ts}(0) < 1$) can be rewritten to give the convergence condition (10). $\square$

We may relate Theorem 11 to Simon's condition by expanding the set $\Gamma_t \setminus s$ to the larger set Γ_t, and observing that $\log x \geq \frac{x^2-1}{x^2+1}$ for all $x \geq 1$ with equality as $x \to 1$. Doing so, we see that Simon's condition is sufficient to guarantee Theorem 11, but that Theorem 11 may be true (implying convergence) when Simon's condition is not satisfied. The improvement over Simon's condition becomes negligible for highly-connected systems with weak potentials, but can be significant for graphs with low connectivity. For example, if the graph consists of a single loop then each node t has at most two neighbors. In this case, the contraction (11) tells us that the outgoing message in either direction is *always* as close or closer to the BP fixed point than the incoming message. Thus we easily obtain the result of Weiss (2000), that (for finite-strength potentials) BP always converges to a unique fixed point on graphs containing a single loop. Simon's condition, on the other hand, is too loose to demonstrate this fact. The form of the condition in Theorem 11 is also similar to a result shown for binary spin models; see Georgii (1988) for details.

However, both Theorem 10 and Theorem 11 depend only on the pairwise potentials $\psi_{st}(x_s, x_t)$, and not on the single-node potentials $\psi_s(x_s)$, $\psi_t(x_t)$. As noted by Heskes (Heskes, 2004), this leaves a degree of freedom to which the single-node potentials may be chosen so as to minimize the (apparent) strength of the pairwise potentials. Thus, (9) can be improved slightly by writing

$$\max_t \sum_{u \in \Gamma_t} \min_{\psi_u, \psi_t} \log d\left(\frac{\psi_{ut}}{\psi_u \psi_t}\right) < 1 \tag{12}$$

and similarly for (10) by writing

$$\max_{(s,t)\in\mathcal{E}} \sum_{u\in\Gamma_t\setminus s} \min_{\psi_u,\psi_t} \frac{d\left(\frac{\psi_{ut}}{\psi_u\psi_t}\right)^2 - 1}{d\left(\frac{\psi_{ut}}{\psi_u\psi_t}\right)^2 + 1} < 1. \tag{13}$$

To evaluate this quantity, one may also observe that

$$\min_{\psi_u,\psi_t} d\left(\frac{\psi_{ut}}{\psi_u\psi_t}\right)^4 = \sup_{a,b,c,d} \frac{\psi_{ts}(a,b)}{\psi_{ts}(a,d)}\frac{\psi_{ts}(c,d)}{\psi_{ts}(c,b)}.$$

In general we shall ignore this subtlety and simply write our results in terms of $d\,(\psi)$, as given in (9) and (10). For binary random variables, it is easy to see that the minimum–strength ψ_{ut} has the form

$$\psi_{ut} = \begin{bmatrix} \eta & 1-\eta \\ 1-\eta & \eta \end{bmatrix},$$

and that when the potentials are of this form (such as in the examples of this section) the two conditions are completely equivalent.

We provide a more empirical comparison between our condition, Simon's condition, and the recent work of Heskes (2004) shortly. Similarly to Heskes (2004), we shall see that it is possible to use the graph geometry to improve our bound (Section 5.3); but perhaps more importantly (and in contrast to both other methods), when the condition is *not* satisfied, we still obtain useful information about the relationship between any pair of fixed points (Section 5.2), allowing its extension to quantized or otherwise distorted versions of belief propagation (Section 5.4).

5.2 Distance of multiple fixed points

Theorem 11 may be extended to provide not only a sufficient condition for a unique BP fixed point, but an upper bound on distance between the beliefs generated by successive BP updates and any BP fixed point. Specifically, the proof of Theorem 11 relied on demonstrating a bound $\log \epsilon^i$ on the distance from some arbitrarily chosen fixed point $\{M_t\}$ at iteration i. When this bound decreases to zero, we may conclude that only one fixed point exists. However, even should it decrease only to some positive constant, it still provides information about the distance between any iteration's belief and the fixed point. Moreover, applying this bound to another, different fixed point $\{\tilde{M}_t\}$ tells us that all fixed points of loopy BP must lie within a sphere of a given diameter [as measured by $\log d\left(M_t/\tilde{M}_t\right)$]. These statements are made precise in the following two theorems:

Theorem 12 (BP distance bound). *Let $\{M_t\}$ be any fixed point of loopy BP. Then, after $n > 1$ iterations of loopy BP resulting in beliefs $\{\hat{M}_t^n\}$, for any node t and for all x*

$$\log d\left(M_t/\hat{M}_t^n\right) \le \sum_{u\in\Gamma_t} \log \frac{d\,(\psi_{ut})^2\,\epsilon^{n-1} + 1}{d\,(\psi_{ut})^2 + \epsilon^{n-1}}$$

where ϵ^i *is given by* $\epsilon^1 = \max_{s,t} d\left(\psi_{st}\right)^2$ *and*

$$\log \epsilon^{i+1} = \max_{(s,t)\in\mathcal{E}} \sum_{u\in\Gamma_t\setminus s} \log \frac{d\left(\psi_{ut}\right)^2 \epsilon^i + 1}{d\left(\psi_{ut}\right)^2 + \epsilon^i}$$

Proof. The result follows directly from the proof of Theorem 11. □

We may thus infer a distance bound between any two BP fixed points:

Theorem 13 (Fixed-point distance bound). *Let* $\{M_t\}$, $\{\tilde{M}_t\}$ *be the beliefs of any two fixed points of loopy BP. Then, for any node* t *and for all* x

$$|\log M_t(x)/\tilde{M}_t(x)| \le 2\log d\left(M_t/\tilde{M}_t\right) \le 2\sum_{u\in\Gamma_t} \log \frac{d\left(\psi_{ut}\right)^2 \epsilon + 1}{d\left(\psi_{ut}\right)^2 + \epsilon} \tag{14}$$

where ϵ *is the largest value satisfying*

$$\log \epsilon = \max_{(s,t)\in\mathcal{E}} G_{ts}(\epsilon) = \max_{(s,t)\in\mathcal{E}} \sum_{u\in\Gamma_t\setminus s} \log \frac{d\left(\psi_{ut}\right)^2 \epsilon + 1}{d\left(\psi_{ut}\right)^2 + \epsilon} \tag{15}$$

Proof. The inequality $|\log M_t(x)/\tilde{M}_t(x)| \le 2\log d\left(M_t/\tilde{M}_t\right)$ follows from Theorem 2. The rest follows from Theorem 12—taking the "approximate" messages to be any other fixed point of loopy BP, we see that the error cannot decrease over any number of iterations. However, by the same argument given in Theorem 11, $g''_{ts}(z) < 0$, and for z sufficiently large, $g_{ts}(z) < z$. Thus (15) has at most one solution greater than unity, and $\epsilon^{i+1} < \epsilon^i$ for all i with $\epsilon^i \to \epsilon$ as $i \to \infty$. Letting the number of iterations $i \to \infty$, we see that the message "errors" $\log d\left(M_{ts}/\tilde{M}_{ts}\right)$ must be at most ϵ, and thus the difference in M_t (the belief of the root node of the computation tree) must satisfy (14). □

Thus, if the value of $\log \epsilon$ is small (the sufficient condition of Theorem 11 is nearly satisfied) then although we cannot guarantee convergence to a unique fixed point, we can still make a strong statement: that the set of fixed points are all mutually close (in a log-error sense), and reside within a ball of diameter described by (14). Moreover, even though it is possible that loopy BP does not converge, and thus even after infinite time the messages may not correspond to *any* fixed point of the BP equations, we are guaranteed by Theorem 12 that the resulting belief estimates *will* asymptotically approach the same bounding ball [achieving distance at most (14) from *all* fixed points].

5.3 Path-counting

If we are willing to put a bit more effort into our bound-computation, we may be able to improve it further, since the bounds derived using computation trees are very much "worst-case" bounds. In particular, the proof of Theorem 11 assumes that, as a message error propagates through the graph, repeated convolution with *only* the strongest set of potentials is possible. But often even if the worst potentials are quite strong, every cycle which contains them may also contain several weaker potentials. Using an iterative algorithm much like belief propagation itself, we may obtain a more globally aware estimate of how errors can propagate through the graph.

Theorem 14 (Non-uniform distance bound). *Let $\{M_t\}$ be any fixed point belief of loopy BP. Then, after $n \geq 1$ iterations of loopy BP resulting in beliefs $\{\hat{M}_t^n\}$, for any node t and for all x*

$$|\log M_t(x)/\hat{M}_t(x)| \leq 2 \log d\left(M_t/\hat{M}_t^n\right) \leq 2 \sum_{u \in \Gamma_t} \log \upsilon_{ut}^n$$

where υ_{ut}^i is defined by the iteration

$$\log \upsilon_{ts}^{i+1} = \log \frac{d\left(\psi_{ts}\right)^2 \epsilon_{ts}^i + 1}{d\left(\psi_{ts}\right)^2 + \epsilon_{ts}^i} \qquad\qquad \log \epsilon_{ts}^i = \sum_{u \in \Gamma_t \setminus s} \log \upsilon_{ut}^i \tag{16}$$

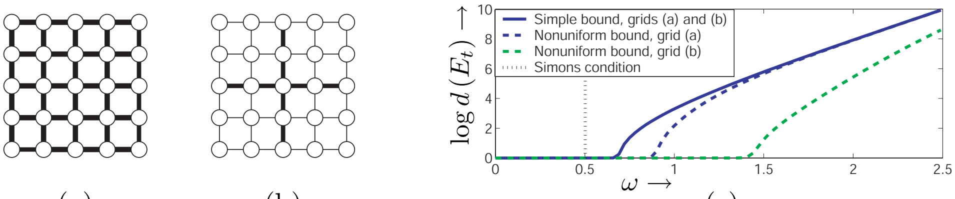
with initial condition $\upsilon_{ut}^1 = d\left(\psi_{ut}\right)^2$.

Proof. Again we consider the error $\log d\left(E_{ts}^i\right)$ incoming to node t with parent s, where t is at level $n-i+1$ of the computation tree. Using the same arguments as Theorem 11 it is easy to show by induction that the error products $\log d\left(E_{ts}^i\right)$ are bounded above by ϵ_{ts}^i, and the individual message errors $\log d\left(e_{ts}^i\right)$ are bounded above by υ_{ts}^i, and . Then, by additivity we obtain the stated bound on $d\left(E_t^n\right)$ at the root node. $\square$

The iteration defined in Theorem 14 can also be interpreted as a (scalar) message-passing procedure, or may be performed offline. As before, if this procedure results in $\log \epsilon_{ts} \to 0$ for all $(t,s) \in \mathcal{E}$ we are guaranteed that there is a unique fixed point for loopy BP; if not, we again obtain a bound on the distance between any two fixed-point beliefs. When the graph is perfectly symmetric (every node has identical neighbors and potential strengths), this yields the same bound as Theorem 12; however, if the potential strengths are inhomogeneous Theorem 14 provides a strictly better bound on loopy BP convergence and errors.

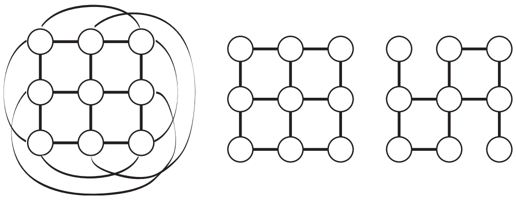
This situation is illustrated in Figure 5—we specify two different graphical models defined on a 5×5 grid in terms of their potential strengths $\log d\left(\psi\right)^2$, and compute bounds on the dynamic range $d\left(M_t/\tilde{M}_t\right)$ of any two fixed point beliefs M_t, $\tilde{M}_t$ for each model. (Note that, while potential strength does not completely specify the graphical model, it is sufficient for all the bounds considered here.) One grid (a) has equal-strength potentials $\log d\left(\psi\right)^2 = \omega$, while the other has many weaker potentials $(\omega/2)$. The worst-case bounds are the same (since both have a node with four strong neighbors), shown as the solid curve in (c). However, the dashed curves show the estimate of (16), which improves only slightly for the strongly coupled graph (a) but considerably for the weaker graph (b). All three bounds give considerably more information than Simon's condition (dotted vertical line).

Having shown how our bound may be improved for irregular graph geometry, we may now compare our bounds to two other known uniqueness conditions (Tatikonda and Jordan, 2002; Heskes, 2004). Simon's condition can be related analytically, as described in Section 5.1. On the other hand, the recent work of Heskes (2004) takes a very different approach to uniqueness based on analysis of the minima of the Bethe free energy, which directly correspond to stable fixed points of BP (Yedidia et al., 2004). This leads to an alternate sufficient condition for uniqueness. As observed in Heskes (2004) it is unclear whether a unique fixed point necessarily implies convergence of loopy BP. In contrast, our



(a) (b) (c)

Figure 5: (a-b) Two small (5×5) grids. In (a), the potentials ψ are all of equal strength ($\log d\,(\psi)^2 = \omega$), while in (b) several potentials (thin lines) are weaker ($\log d\,(\psi)^2 = .5\omega$). The methods described may be used to compute bounds (c) on the distance $d\,(E_t)$ between any two fixed point beliefs as a function of potential strength ω.



(a) (b) (c)

Method	(a)	(b)	(c)
Simon's condition	.62	.62	.62
Heskes' condition	.55	.58	.65
This work	.67	.79	.88
Empirical	.67	.79	.88

η_{crit}

Figure 6: Comparison of various uniqueness bounds: for binary potentials parameterized by η, we find the predicted η_{crit} at which loopy BP can no longer be guaranteed to be unique. For these simple problems, the η_{crit} at which the trivial (correct) solution becomes unstable may be found empirically. Examples and empirical values of η_{crit} from Heskes (2004).

approach gives a sufficient condition for the convergence of BP to a unique solution, which implies uniqueness of the fixed point.

Showing an analytic relation between all three approaches does not appear straightforward; to give some intuition, we show the three example binary graphs compared in Heskes (2004), whose structures are shown in Figure 6(a-c) and whose potentials are parameterized by a scalar $\eta > .5$, namely

$$\psi = \begin{bmatrix} \eta & 1-\eta \\ 1-\eta & \eta \end{bmatrix} \tag{17}$$

(so that $d\,(\psi)^2 = \frac{\eta}{1-\eta}$). The trivial solution $M_t = [.5; .5]$ is always a fixed point, but may not be stable; the precise η_{crit} at which this fixed point becomes unstable (implying the existence of other, stable fixed points) can be found empirically for each case (Heskes, 2004); the same values may also be found algebraically by imposing symmetry requirements on the messages (Yedidia et al., 2004). This value may then be compared to the uniqueness bounds of Tatikonda and Jordan (2002), the bound of Heskes (2004), and this work; these are shown in Figure 6.

Notice that our bound is always better than Simon's condition, though for the perfectly symmetric graph the margin is not large (and decreases further with increased connectivity, for example a cubic lattice). Additionally, in all three examples our method appears to outperform that of Heskes (2004), though without analytic comparison it is unclear whether this is always the case. In fact, for these simple binary examples, our bound appears to be tight.

However, our method also allows us to make statements about the results of loopy BP after finite numbers of iterations, up to some finite degree of numerical precision in the final results. For example, we may also find the value of η below which BP will attain a particular precision, say $\log d\left(M_t/\hat{M}_t^n\right) < 10^{-3}$ in at least $n = 100$ iterations [obtaining the values $\{.66, .77, .85\}$ for the grids in Figure 6(a), (b), and (c), respectively].

5.4 Introducing intentional message errors and censoring

As discussed in the introduction, we may wish to introduce or allow *additional* errors in our messages at each stage, in order to improve the computational or communication efficiency of the algorithm. This may be the result of an actual distortion imposed on the message (perhaps to decrease its complexity, for example quantization), or the result of censoring the message update (reusing the message from the previous iteration) when the two are sufficiently similar. Errors may also arise from quantization or other approximation of the potential functions. Such additional errors may be easily incorporated into our framework.

Theorem 15. *If at every iteration of loopy BP, each message is further approximated in such a way as to guarantee that the additional distortion has maximum dynamic range at most δ, then for any fixed point beliefs $\{M_t\}$, after $n \geq 1$ iterations of loopy BP resulting in beliefs $\{\hat{M}_t^n\}$ we have*

$$\log d\left(M_t/\hat{M}_t^n\right) \leq \sum_{u \in \Gamma_t} \log \upsilon_{ut}^n$$

where υ_{ut}^i is defined by the iteration

$$\log \upsilon_{ts}^{i+1} = \log \frac{d\left(\psi_{ts}\right)^2 \epsilon_{ts}^i + 1}{d\left(\psi_{ts}\right)^2 + \epsilon_{ts}^i} + \log \delta \qquad\qquad \log \epsilon_{ts}^i = \sum_{u \in \Gamma_t \setminus s} \log \upsilon_{ut}^i$$

with initial condition $\upsilon_{ut}^1 = \delta\, d\left(\psi_{ut}\right)^2$.

Proof. Using the same logic as Theorems 12 and 14, apply additivity of the log dynamic range measure to the additional distortion $\log \delta$ introduced to each message. $\square$

As with Theorem 14, a simpler bound can also be derived (similar to Theorem 12). Either gives a bound on the maximum total distortion from any true fixed point which will be incurred by quantized or censored belief propagation. Note that (except on tree-structured graphs) this does *not* bound the error from the true marginal distributions, only from the loopy BP fixed points.

It is also possible to interpret the additional error as arising from an approximation to the correct single-node and pairwise potentials ψ_t, ψ_{ts}.

Theorem 16. *Suppose that* $\{M_t\}$ *are a fixed point of loopy BP on a graph defined by potentials* ψ_{ts} *and* ψ_t*, and let* $\{\hat{M}_t^n\}$ *be the beliefs of* n *iterations of loopy BP performed on a graph with potentials* $\hat{\psi}_{ts}$ *and* $\hat{\psi}_t$*, where* $d\left(\hat{\psi}_{ts}/\psi_{ts}\right) \le \delta_1$ *and* $d\left(\hat{\psi}_t/\psi_t\right) \le \delta_2$*. Then,*

$$\log d\left(M_t/\hat{M}_t^n\right) \le \sum_{u\in\Gamma_t} \log \upsilon_{ut}^n + \log \delta_2$$

where υ_{ut}^i *is defined by the iteration*

$$\log \upsilon_{ts}^{i+1} = \log \frac{d\left(\psi_{ts}\right)^2 \epsilon_{ts}^i + 1}{d\left(\psi_{ts}\right)^2 + \epsilon_{ts}^i} + \log \delta_1 \qquad\qquad \log \epsilon_{ts}^i = \log \delta_2 + \sum_{u\in\Gamma_t\setminus s} \log \upsilon_{ut}^i$$

with initial condition $\upsilon_{ut}^1 = \delta_1\, d\left(\psi_{ut}\right)^2$*.*

Proof. We first extend the contraction result given in Appendix A by applying the inequality

$$\frac{\int \psi(x_t,a)\frac{\hat{\psi}(x_t,a)}{\psi(x_t,a)} M(x_t)E(x_t)dx_t}{\int \psi(x_t,b)\frac{\hat{\psi}(x_t,b)}{\psi(x_t,b)} M(x_t)E(x_t)dx_t} \le \frac{\int \psi(x_t,a)M(x_t)E(x_t)dx_t}{\int \psi(x_t,b)M(x_t)E(x_t)dx_t} \cdot d\left(\hat{\psi}/\psi\right)^2$$

Then, proceeding similarly to Theorem 15 yields the definition of υ_{ts}^i, and including the additional errors $\log \delta_2$ in each message product (resulting from the product with $\hat{\psi}_t$ rather than ψ_t) gives the definition of ϵ_{ts}^i. ∎

Incorrect models $\hat{\psi}$ may arise when the exact graph potentials have been estimated or quantized; Theorem 16 gives us the means to interpret the (worst-case) overall effects of using an approximate model. As an example, let us again consider the model depicted in Figure 6(b). Suppose that we are given *quantized* versions of the pairwise potentials, $\hat{\psi}$, specified by the value (rounded to two decimal places) $\eta = .65$. Then, the true potential ψ has $\eta \in .65 \pm .005$, and thus is within $\delta_1 \approx 1.022 = \frac{(.35)(.655)}{(.345)(.65)}$ of the known approximation $\hat{\psi}$. Applying the recursion of Theorem 16 allows us to conclude that the solution obtained using the approximate model $\hat{\psi}$ and true model ψ are within $\log d\,(e) \le .36$, or alternatively that the beliefs found using the approximate model are correct to within a multiplicative factor of about 1.43. The same $\hat{\psi}$, with η assumed correct to three decimal places, gives a bound $\log d\,(e) \le .04$, or multiplicative factor of 1.04.

5.5 Stochastic Analysis

Unfortunately, the bounds given by Theorem 16 are often pessimistic compared to actual performance. We may use a similar analysis, coupled with the assumption of uncorrelated message errors, to obtain a more realistic estimate (though no longer a strict bound) on the resulting error.

Proposition 17. *Suppose that the errors* $\log e_{ts}$ *are random and uncorrelated, so that at each iteration* i*, for* $s \neq u$ *and any* x*,* $E\left[\log e_{st}^i(x) \cdot \log e_{ut}^i(x)\right] = 0$*, and that at each* *iteration of loopy BP, the additional error (in the log domain) imposed on each message is uncorrelated with variance at most* $(\log \delta)^2$. *Then,*

$$E\left[\left(\log d\left(E_t^i\right)\right)^2\right] \leq \sum_{u \in \Gamma_t} \left(\sigma_{ut}^i\right)^2 \tag{18}$$

where $\sigma_{ts}^1 = \log d\left(\psi_{ts}\right)^2$ *and*

$$\left(\sigma_{ts}^{i+1}\right)^2 = \left(\log \frac{d\left(\psi_{ts}\right)^2 \lambda_{ts}^i + 1}{d\left(\psi_{ts}\right)^2 + \lambda_{ts}^i}\right)^2 + (\log \delta)^2 \qquad \left(\log \lambda_{ts}^i\right)^2 = \sum_{u \in \Gamma_t \setminus s} \left(\sigma_{ut}^i\right)^2$$

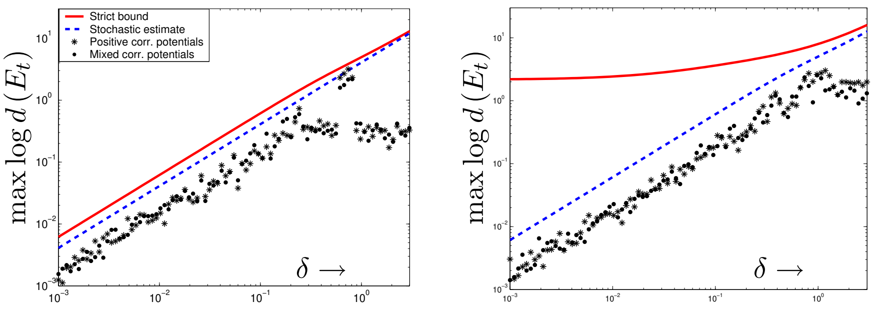
Proof. Let us define the (nuisance) scale factor $\alpha_{ts}^i = \arg\min_\alpha \sup_x |\log \alpha e_{ts}^i(x)|$ for each error e_{ts}^i, and let $\zeta_{ts}^i(x) = \log \alpha_{ts}^i e_{ts}^i(x)$. Now, we model the error function $\zeta_{ts}^i(x)$ (for each x) as a random variable with mean zero, and bound the standard deviation of $\zeta_{ts}^i(x)$ by σ_{ts}^i at each iteration i; under the assumption that the errors in any two incoming messages are uncorrelated, we may assert additivity of their variances. Thus the variance of $\sum_{\Gamma_t \setminus s} \zeta_{ut}^i(x)$ is bounded by $(\log \lambda_{ts}^i)^2$. The contraction of Theorem 8 is a non-linear relationship; we estimate its effect on the error variance using a simple sigma-point quadrature ("unscented") approximation (Julier and Uhlmann, 1996), in which the standard deviation σ_{ts}^{i+1} is estimated by applying Theorem 8's nonlinear contraction to the standard deviation of the error on the incoming product $(\log \lambda_{ts}^i)$. □

The assumption of uncorrelated errors is clearly questionable, since propagation around loops may couple the incoming message errors. However, similar assumptions have yielded useful analysis of quantization effects in assessing the behavior and stability of digital filters (Willsky, 1978). It is often the case that empirically, such systems behave similarly to the predictions made by assuming uncorrelated errors. Indeed, we shall see that in our simulations, the assumption of uncorrelated errors provides a good estimate of performance.

Given the bound (18) on the variance of $\log d\,(E)$, we may apply a Chebyshev-like argument to provide probabilistic guarantees on the magnitude of errors $\log d\,(E)$ observed in practice. In our experiments (Section 5.6), the 2σ distance was almost always larger than the observed error. The probabilistic bound derived using (18) is typically much smaller than the bound of Theorem 15 due to the strictly sub-additive relationship between the standard deviations. However, the underlying assumption of uncorrelated errors makes the estimate obtained using (18) unsuitable for deriving strict convergence guarantees.

5.6 Experiments

We demonstrate the dynamic range error bounds for quantized messages with a set of Monte Carlo trials. In particular, for each trial we construct a binary–valued 5×5 grid with uniform potential strengths, which are either (1) all positively correlated, or (2) randomly chosen to be positively or negatively correlated (equally likely); we also assign random single-node potentials to each variable x_s. We then run a quantized version of BP for $n = 100$ iterations from the same initial conditions, rounding each log-message to discrete values separated by $2 \log \delta$ (ensuring that the newly introduced error satisfies $d\,(e) \leq \delta$). Figure 7 shows the maximum belief error in each of 100 trials of this procedure for various values of δ.



(a) $\log d(\psi)^2 = .25$ (b) $\log d(\psi)^2 = 1$

Figure 7: Maximum belief errors incurred as a function of the quantization error. The scatterplot indicates the maximum error measured in the graph for each of 200 Monte Carlo runs; this is strictly bounded above by Theorem 15, solid, and bounded with high probability (assuming uncorrelated errors) by Proposition 17, dashed.

Also shown are two performance estimators—the *bound* on belief error developed in Section 5.4, and the 2σ estimate computed assuming uncorrelated message errors as in Section 5.5. As can be seen, the stochastic estimate is a much tighter, more accurate assessment of error, but it does not possess the same strong theoretical guarantees. Since [as observed for digital filtering applications (Willsky, 1978)] the errors introduced by quantization are typically close to independent, the assumptions underlying the stochastic estimate are reasonable, and empirically we observe that the estimate and actual errors behave similarly.

6. KL-Divergence Measures

Although the dynamic range measure introduced in Section 4 leads to a number of strong guarantees, its performance criterion may be unnecessarily (and undesirably) strict. Specifically, it provides a *pointwise* guarantee, that m and $\hat{m}$ are close for every possible state x. For continuous-valued states, this is an extremely difficult criterion to meet—for instance, it requires that the messages' tails match almost exactly. In contrast, typical measures of the difference between two distributions operate by an average (mean squared error or mean absolute error) or weighted average (Kullback-Leibler divergence) evaluation. To address this, let us consider applying a measure such as the Kullback-Leibler (KL) divergence,

$$D(p\|\hat{p}) = \int p(x) \log \frac{p(x)}{\hat{p}(x)} dx$$

The pointwise guarantees of Section 4 are necessary to bound performance even in the case of "unlikely" events. More specifically, the tails of a message approximation can become important if two parts of the graph strongly disagree, in which case the tails of each message are the only overlap of significant likelihood. One way to discount this possibility is to consider the graph potentials themselves (in particular, the single node potentials ψ_t) as a realization of random variables which "typically" agree, then apply a probabilistic measure to estimate the typical performance. From this viewpoint, since a strong disagreement between parts of the graph is unlikely we will be able to relax our error measure in the message tails.

Unfortunately, many of the properties which we relied on for analysis of the dynamic range measure do not strictly hold for a KL-divergence measure of error, resulting in an *approximation*, rather than a bound, on performance. In Appendix B, we give a detailed analysis of each property, showing the ways in which each aspect can break down and discussing the reasonability of simple approximations. In this section, we apply these approximations to develop a KL-divergence based estimate of error.

6.1 Local Observations and Parameterization

To make this notion concrete, let us consider a graphical model in which the single-node potential functions are specified in terms of a set of observation variables $\mathbf{y} = \{y_t\}$; in this section we will examine the average (expected) behavior of BP over multiple realizations of the observation variables $\mathbf{y}$. We further assume that both the prior $p(\mathbf{x})$ and likelihood $p(\mathbf{y}|\mathbf{x})$ exhibit conditional independence structure, expressed as a graphical model. Specifically, we assume throughout this section that the observation likelihood factors as

$$p(\mathbf{y}|\mathbf{x}) = \prod_t p(y_t|x_t) \tag{19}$$

in other words, that each observation variable y_t is *local* to (conditionally independent given) one of the x_t. As for the prior model $p(\mathbf{x})$, for the moment we confine our attention to tree-structured distributions, for which one may write (Wainwright et al., 2003)

$$p(\mathbf{x}) = \prod_{(s,t)\in\mathcal{E}} \frac{p(x_s, x_t)}{p(x_s)p(x_t)} \prod_s p(x_s) \tag{20}$$

The expressions (19)-(20) give rise to a convenient parameterization of the joint distribution, expressed as

$$p(\mathbf{x}, \mathbf{y}) \propto \prod_{(s,t)\in\mathcal{E}} \psi_{st}(x_s, x_t) \prod_s \psi_s^x(x_s)\psi_s^y(x_s) \tag{21}$$

where

$$\psi_{st}(x_s, x_t) = \frac{p(x_s, x_t)}{p(x_s)p(x_t)} \qquad \text{and} \qquad \psi_s^x(x_s) = p(x_s) \qquad , \qquad \psi_s^y(x_s) = p(y_s|x_s). \tag{22}$$

Our goal is to compute the posterior marginal distributions $p(x_s|\mathbf{y})$ at each node s; for the tree-structured distribution (21) this can be performed exactly and efficiently by BP. As discussed in the previous section, we treat the $\{y_t\}$ as random variables; thus almost all quantities in this graph are themselves random variables (as they are dependent on the y_t), so that the single node observation potentials $\psi_s^y(x_s)$, messages $m_{st}(x_t)$, *etc.* are random functions of their argument x_s. The potentials due to the prior (ψ_{st} and ψ_s^x), however, are not random variables as they do not depend on any of the observations y_t.

For models of the form (21)-(22), the (unique) BP message fixed point consists of normalized versions of the likelihood functions $m_{ts}(x_s) \propto p(\mathbf{y}_{ts}|x_s)$, where $\mathbf{y}_{ts}$ denotes the set of all observations $\{y_u\}$ such that t separates u from s. In this section it is also convenient to perform a *prior-weighted* normalization of the messages m_{ts}, so that $\int p(x_s)m_{ts}(x_s) = 1$ (as opposed to $\int m_{ts}(x_s) = 1$ as assumed previously); we again assume this prior-weighted normalization is always possible (this is trivially the case for discrete-valued states $\mathbf{x}$). Then, for a tree-structured graph, the prior-weight normalized fixed-point message from t to s is precisely

$$m_{ts}(x_s) = p(\mathbf{y}_{ts}|x_s)/p(\mathbf{y}_{ts}) \tag{23}$$

and the products of incoming messages to t, as defined in Section 2.3, are equal to

$$M_{ts}(x_t) = p(x_t|\mathbf{y}_{ts}) \qquad\qquad M_t(x_t) = p(x_t|\mathbf{y}).$$

We may now apply a *posterior-weighted log-error* measure, defined by

$$\mathcal{D}(m_{ut}\|\hat{m}_{ut}) = \int p(x_t|\mathbf{y}) \log \frac{m_{ut}(x_t)}{\hat{m}_{ut}(x_t)} dx_t; \tag{24}$$

and may relate (24) to the Kullback-Leibler divergence.

Lemma 18. *On a tree-structured graph, the error measure $\mathcal{D}(M_t, \hat{M}_t)$ is equivalent to the KL-divergence of the true and estimated posterior distributions at node t:*

$$\mathcal{D}(M_t\|\hat{M}_t) = D(p(x_t|\mathbf{y})\|\hat{p}(x_t|\mathbf{y}))$$

Proof. This follows directly from the definitions of $\mathcal{D}$, and the fact that on a tree, the unique fixed point has beliefs $M_t(x_t) = p(x_t|\mathbf{y})$. □

Again, the error $\mathcal{D}(m_{ut}\|\hat{m}_{ut})$ is a function of the observations $\mathbf{y}$, both explicitly through the term $p(x_t|\mathbf{y})$ and implicitly through the message $m_{ut}(x_t)$, and is thus also a random variable. Although the definition of $\mathcal{D}(m_{ut}\|\hat{m}_{ut})$ involves the *global* observation $\mathbf{y}$ and thus cannot be calculated at node u without additional (non-local) information, we will primarily be interested in the expected value of these errors over many realizations $\mathbf{y}$, which is a function only of the distribution. Specifically, we can see that in expectation over the data $\mathbf{y}$, it is simply

$$E\left[\mathcal{D}(m_{ut}\|\hat{m}_{ut})\right] = E\left[\int p(x_t)m_{ut}(x_t) \log \frac{m_{ut}(x_t)}{\hat{m}_{ut}(x_t)} dx_t\right]. \tag{25}$$

One nice consequence of the choice of potential functions (22) is the locality of prior information. Specifically, if *no* observations $\mathbf{y}$ are available, and only prior information is present, the BP messages are trivially constant $[m_{ut}(x) = 1\ \forall x]$. This ensures that any message approximations affect only the data likelihood, and not the prior $p(x_t)$; this is similar to the motivation of Paskin and Guestrin (2004), in which an additional message-passing procedure is used to create this parameterization.

Finally, two special cases are of note. First, if x_s is discrete-valued and the prior distribution $p(x_s)$ constant (uniform), the expected message distortion with prior-normalized messages, $E[\mathcal{D}(m\|\hat{m})]$, and the KL-divergence of traditionally normalized messages behave equivalently, i.e.,

$$E\left[\mathcal{D}(m_{ts}\|\hat{m}_{ts})\right] = E\left[D\left(\frac{m_{ts}}{\int m_{ts}}\|\frac{\hat{m}_{ts}}{\int \hat{m}_{ts}}\right)\right]$$

where we have abused the notation of KL-divergence slightly to apply it to the normalized likelihood $m_{ts}/\int m_{ts}$. This interpretation leads to the same message-censoring criterion used in Chen et al. (2004).

Secondly, when the state x_s is a discrete-valued random variable taking on one of M possible values, a straightforward uniform quantization of the value of $p(x_s)m(x_s)$ results in a bound on the divergence (25). Specifically, we have the following lemma:

Lemma 19. *For an M-ary discrete variable x, the quantization*

$$p(x)m(x) \to \{\epsilon, 3\epsilon, \ldots, 1-\epsilon\}$$

results in an expected divergence bounded by

$$E\left[\mathcal{D}(m(x)\|\hat{m}(x))\right] \leq (2\log 2 + M)M\epsilon + \mathcal{O}(M^3\epsilon^2)$$

Proof. Define $\mu(x) = p(x)m(x)$, and $\bar{\mu}(x) \in \{\epsilon, 3\epsilon, \ldots, 1-\epsilon\}$ (for each x) to be its quantized value. Then, the prior-normalized approximation $\hat{m}(x)$ satisfies

$$p(x)\hat{m}(x) = \bar{\mu}(x) \;/\; \sum_x \bar{\mu}(x) = \bar{\mu}(x)/C$$

where $C \in [1 - M\epsilon, 1 + M\epsilon]$. The expected divergence

$$\begin{aligned} E\left[\mathcal{D}(m(x)\|\hat{m}(x))\right] &= \sum_x p(x)m(x)\log\frac{m(x)}{\hat{m}(x)} \\ &\leq \sum_x \mu(x)\log\frac{\mu(x)}{\bar{\mu}(x)} + \sum_x |\log C| \end{aligned}$$

The first sum is at its maximum for $\mu(x) = 2\epsilon$ and $\bar{\mu}(x) = \epsilon$, which results in the value $\sum_x (2\log 2)\epsilon$. Applying the Taylor expansion of the log, the second sum $\sum |\log C|$ is bounded above by $M^2\epsilon + \mathcal{O}(M^3\epsilon^2)$. ☐

Thus, for example, for uniform quantization of a message with binary–valued state x, fidelity up to two significant digits ($\epsilon = .005$) results in an error $\mathcal{D}$ which, on average, is less than .034.

We now state the approximations which will take the place of the fundamental properties used in the preceding sections, specifically versions of the triangle inequality, sub-additivity, and contraction. Although these properties do *not* hold in general, in practice useful estimates are obtained by making approximations corresponding to each property and following the same development used in the preceding sections. (In fact, experimentally these estimates still appear quite conservative.) A more detailed analysis of each property, along with justification for the approximation applied, is given in Appendix B.

6.2 Approximations

Three properties of the dynamic range described in Section 4 are important in the error analysis of Section 5—a form of the triangle inequality, enabling the accumulation of errors in successive approximations to be bounded by the sum of the individual errors, a form of sub-additivity, enabling the accumulation of errors in the message product operation to be bounded by the sum of incoming errors, and a rate of contraction due to convolution with each pairwise potential. We assume the following three properties for the expected error; see Appendix B for a more detailed discussion.

Approximation 20 (Triangle Inequality). *For a true BP fixed-point message m_{ut} and two approximations $\hat{m}_{ut}$, $\tilde{m}_{ut}$, we assume*

$$\mathcal{D}(m_{ut}\|\tilde{m}_{ut}) \leq \mathcal{D}(m_{ut}\|\hat{m}_{ut}) + \mathcal{D}(\hat{m}_{ut}\|\tilde{m}_{ut}) \tag{26}$$

Comment. This is not strictly true for arbitrary $\hat{m}$, $\tilde{m}$, since the KL-divergence (and thus $\mathcal{D}$) does not satisfy the triangle inequality.

Approximation 21 (Sub-additivity). *For true BP fixed-point messages $\{m_{ut}\}$ and approximations $\{\hat{m}_{ut}\}$, we assume*

$$\mathcal{D}(M_{ts}\|\hat{M}_{ts}) \leq \sum_{u\in\Gamma_t\setminus s} \mathcal{D}(m_{ut}\|\hat{m}_{ut}) \tag{27}$$

Approximation 22 (Contraction). *For a true BP fixed-point message product M_{ts} and approximation $\hat{M}_{ts}$, we assume*

$$\mathcal{D}(m_{ts}\|\hat{m}_{ts}) \leq (1-\gamma_{ts})\mathcal{D}(M_{ts}\|\hat{M}_{ts}) \tag{28}$$

where

$$\gamma_{ts} = \min_{a,b} \int \min\left[\rho(x_s, x_t = a)\,,\, \rho(x_s, x_t = b)\right] dx_s \qquad \rho(x_s, x_t) = \frac{\psi_{ts}(x_s, x_t)\psi_s^x(x_s)}{\int \psi_{ts}(x_s, x_t)\psi_s^x(x_s)dx_s}$$

Comment. For tree-structured graphical models with the parametrization described by (21)-(22), $\rho(x_s, x_t) = p(x_s|x_t)$, and γ_{ts} corresponds to the rate of contraction described by Boyen and Koller (1998).

6.3 Steady-state errors

Applying these approximations to graphs with cycles, and following the same development used for constructing the strict bounds of Section 5, we find the following estimates of steady-state error. Note that, other than those outlined in the previous section (and described in Appendix B), this development involves no additional approximations.

Approximation 23. *After $n \geq 1$ iterations of loopy BP subject to additional errors at each iteration of magnitude (measured by $\mathcal{D}$) bounded above by some constant δ, with initial messages $\{m_{tu}^0\}$ satisfying $\mathcal{D}(m_{tu}\|m_{tu}^0)$ less than some constant C, results in an expected*

KL-divergence between a true BP fixed point $\{M_t\}$ and the approximation $\{\hat{M}_t^n\}$ bounded by

$$E_{\mathbf{y}}\left[D(M_t \| \hat{M}_t^n)\right] = E_{\mathbf{y}}\left[\mathcal{D}(\mathcal{M}_t \| \hat{\mathcal{M}}_t^n)\right] \leq \sum_{u \in \Gamma_t} ((1-\gamma_{ut})\epsilon_{ut}^{n-1} + \delta)$$

where $\epsilon_{ts}^0 = C$ and

$$\epsilon_{ts}^i = \sum_{u \in \Gamma_t \setminus s} ((1-\gamma_{ut})\epsilon_{ut}^{i-1} + \delta)$$

Comment. The argument proceeds similarly to that of Theorem 15. Let ϵ_{ts}^i bound the quantity $\mathcal{D}(\mathcal{M}_{ts} \| \hat{\mathcal{M}}_{ts}^i)$ at each iteration i, and apply Approximations 20-22.

We refer to the estimate described in Approximation 23 as a "bound-approximation", in order to differentiate it from the stochastic error estimate presented next.

Just as a stochastic analysis of message error gave a tighter estimate for the pointwise difference measure, we may obtain an alternate Chebyshev-like "bound" by assuming that the message perturbations are uncorrelated (already an assumption of the KL additivity analysis) and that we require only an estimate which exceeds the expected error with high probability.

Approximation 24. *Under the same assumptions as Approximation 23, but describing the error in terms of its variance and assuming that these errors are uncorrelated gives the estimate*

$$E\left[\mathcal{D}(\mathcal{M}_t \| \hat{\mathcal{M}}_t^n)^2\right] \leq \sum_{u \in \Gamma_t} (\sigma_{ut}^{n-1})^2$$

where $(\sigma_{ts}^0)^2 = C$ and

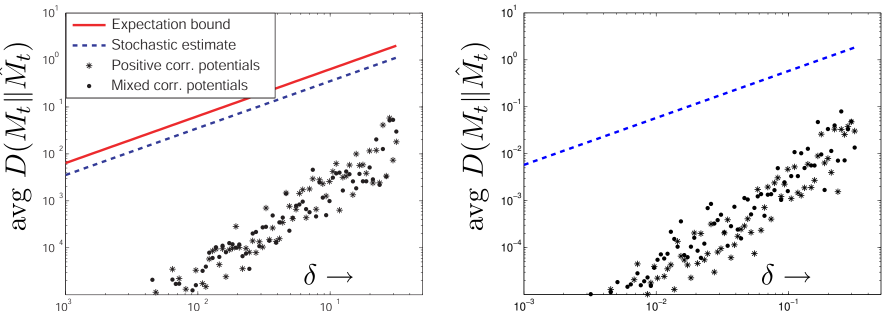
$$(\sigma_{ts}^i)^2 = \sum_{u \in \Gamma_t \setminus s} ((1-\gamma_{ut})\sigma_{ut}^{i-1})^2 + \delta^2$$

Comment. The argument proceeds similarly to Proposition 17, by induction on the claim that $(\sigma_{ut}^i)^2$ bounds the variance at each iteration i. This again applies Theorem 29 ignoring any effects due to loops, as well as the assumption that the message errors are uncorrelated (implying additivity of the variances of each incoming message). As in Section 5.5, we take the 2σ value as our performance estimate.

6.4 Experiments

Once again, we demonstrate the utility of these two estimates on the same uniform grids used in Section 5.6. Specifically, we generate 200 example realizations of a 5×5 binary grid and its observation potentials (100 with strictly attractive potentials and 100 with mixed potentials), and compare a quantized version of loopy BP with the solution obtained by exact loopy BP, as a function of KL-divergence bound δ incurred by the quantization level ϵ (see Lemma 18).

Figure 8(a) shows the maximum KL-divergence from the correct fixed point resulting in each Monte Carlo trial for a grid with relatively weak potentials (in which loopy BP is analytically guaranteed to converge). As can be seen, both the bound (solid) and stochastic estimate (dashed) still provide conservative estimates of the expected error. In Figure 8(b)



(a) $\log d(\psi)^2 = .25$ (b) $\log d(\psi)^2 = 1$

Figure 8: KL-divergence of the beliefs as a function of the added message error δ. The scatterplots indicates the average error measured in the graph for each of 200 Monte Carlo runs, along with the expected divergence bound (solid) and 2σ stochastic estimate (dashed). For stronger potentials, the upper bound may be trivially infinite; in this example the stochastic estimate still gives a reasonable gauge of performance.

.

we repeat the same analysis but with stronger pairwise potentials (for which convergence to a unique solution is not guaranteed but typically occurs in practice). In this case, the bound-based estimate of KL-divergence is trivially infinite—its linear rate of contraction is insufficient to overcome the accumulation rate. However, the greater sub-additivity in the stochastic estimate leads to the non-trivial curve shown (dashed), which still provides a reasonable (and still conservative) estimate of the performance in practice.

7. Conclusions and Future Directions

We have described a framework for the analysis of belief propagation stemming from the view that the message at each iteration is some noisy or erroneous version of some true BP fixed point. By measuring and bounding the error at each iteration, we may analyze the behavior of various forms of BP and test for convergence to the ideal fixed-point messages, or bound the total error from any such fixed point.

In order to do so, we introduced a measure of the pointwise dynamic range, which represents a strong condition on the agreement between two messages; after showing its utility for common inference tasks such as MAP estimation and its transference to other common measures of error, we showed that under this measure the influence of message errors is both sub-additive and measurably contractive. These facts led to conditions under which traditional belief propagation may be shown to converge to a unique fixed point, and more generally a bound on the distance between any two fixed points. Furthermore, it enabled analysis of quantized, stochastic, or other approximate forms of belief propagation, yielding conditions under which they may be guaranteed to converge to some unique region, as well as bounds on the ensuing error over exact BP. If we further assume that the message perturbations are uncorrelated, we obtain an alternate, tighter estimate of the resulting error.

The second measure considered an average case error similar to the Kullback-Liebler divergence, in expectation over the possible realizations of observations within the graph. While this gives no guarantees about any particular realization, the difference measure itself is able to be much less strict (allowing poor approximations in the distribution tails, for example). Analysis of this case is substantially more difficult and leads to approximations rather than guarantees, but explains some of the observed similarities in behavior among the two forms of perturbed BP. Simulations indicate that these estimates remain sufficiently accurate to be useful in practice.

Further analysis of the propagation of message errors has the potential to give an improved understanding of when and why BP converges (or fails to converge), and potentially also the role of the message schedule in determining the performance. Additionally, there are many other possible measures of the deviation between two messages, any of which may be able to provide an alternative set of bounds and estimates on performance of BP using either exact or approximate messages.

8. Acknowledgements

The authors would like to thank Tom Heskes, Martin Wainwright, Erik Sudderth, and Lei Chen for many helpful discussions. Thanks also to the anonymous reviewers of JMLR for their insightful comments, and for suggesting the improvement of equation (13). This research was supported in part by AFOSR grant F49620-00-0362 and by ODDR&E MURI through ARO grant DAAD19-00-0466. Portions of this work have appeared as a conference paper (Ihler et al., 2004b).

Appendix A. Proof of Theorem 8

Because all quantities in this section refer to the pair (t, s), we suppress the subscripts. The error measure $d(e)$ is given by

$$d(e)^2 = d(\hat{m}/m)^2 = \max_{a,b} \frac{\int \psi(x_t, a) M(x_t) E(x_t) dx_t}{\int \psi(x_t, a) M(x_t) dx_t} \cdot \frac{\int \psi(x_t, b) M(x_t) dx_t}{\int \psi(x_t, b) M(x_t) E(x_t) dx_t} \tag{29}$$

subject to a few constraints: positivity of the messages and potential functions, normalization of the message product M, and the definitions of $d(E)$ and $d(\psi)$. In order to analyze the maximum possible value of $d(e)$ for any functions ψ, M, and E, we make repeated use of the following property:

Lemma 25. *For f_1, f_2, g_1, g_2 all positive,*

$$\frac{f_1 + f_2}{g_1 + g_2} \le \max\left[\frac{f_1}{g_1}, \frac{f_2}{g_2}\right]$$

Proof. Assume without loss of generality that $f_1/g_1 \ge f_2/g_2$. Then we have $f_1/g_1 \ge f_2/g_2 \Rightarrow f_1 g_2 \ge f_2 g_1 \Rightarrow f_1 g_1 + f_1 g_2 \ge f_1 g_1 + f_2 g_1 \Rightarrow \frac{f_1}{g_1} \ge \frac{f_1+f_2}{g_1+g_2}$. ☐

This fact, extended to more general sums, may be applied directly to (29) to prove Corollary 9. However, a more careful application leads to the result of Theorem 8. The following lemma will assist us:

Lemma 26. *The maximum of $d(e)$ with respect to $\psi(x_t,a)$, $\psi(x_t,b)$, and $E(x_t)$ is attained at some extremum of their feasible function space. Specifically,*

$$\psi(x,a) = 1 + (d(\psi)^2 - 1)\chi_A(x) \qquad E(x) = 1 + (d(E)^2 - 1)\chi_E(x)$$
$$\psi(x,b) = 1 + (d(\psi)^2 - 1)\chi_B(x)$$

where χ_A, χ_B, and χ_E are indicator functions taking on only values 0 and 1.

Proof. We simply show the result for $\psi(x,a)$; the proofs for $\psi(x,b)$ and $E(x)$ are similar. First, observe that without loss of generality we may scale $\psi(x,a)$ so that its minimum value is 1. Now consider a convex combination of any two possible functions: let $\psi(x_t,a) = \alpha_1\psi_1(x_t,a) + \alpha_2\psi_2(x_t,a)$ with $\alpha_1 \geq 0$, $\alpha_2 \geq 0$, and $\alpha_1 + \alpha_2 = 1$. Then, applying Lemma 25 to the left-hand term of (29) we have

$$\frac{\alpha_1 \int \psi_1(x_t,a)M(x_t)E(x_t)dx_t + \alpha_2 \int \psi_2(x_t,a)M(x_t)E(x_t)dx_t}{\alpha_1 \int \psi_1(x_t,a)M(x_t)dx_t + \alpha_2 \int \psi_2(x_t,a)M(x_t)dx_t}$$
$$\leq \max\left[\frac{\int \psi_1(x_t,a)M(x_t)E(x_t)dx_t}{\int \psi_1(x_t,a)M(x_t)dx_t}\ ,\ \frac{\int \psi_2(x_t,a)M(x_t)E(x_t)dx_t}{\int \psi_2(x_t,a)M(x_t)dx_t}\right] \qquad (30)$$

Thus, $d(e)$ is maximized by taking whichever of ψ_1, ψ_2 results in the largest value—an extremum. It remains only to describe the form of such a function extremum. Any potential $\psi(x,a)$ may be considered to be the convex combination of functions of the form $\left(d(\psi)^2 - 1\right)\chi(x) + 1$, where χ takes on values $\{0,1\}$. This can be seen by the construction

$$\psi(x,a) = \int_0^1 \left(d(\psi)^2 - 1\right)\chi_m^y(x,a) + 1\ dy$$
$$\text{where} \qquad \chi_m^y(x,a) = \begin{cases} 1 & \psi(x,a) \geq 1 + (d(\psi)^2 - 1)\,y \\ 0 & \text{otherwise.} \end{cases}$$

Thus, the maximum value of $d(e)$ will be attained by a potential equal to one of these functions. □

Applying Lemma 26, we define the shorthand

$$M_A = \int M(x)\chi_A(x) \qquad M_B = \int M(x)\chi_B(x) \qquad M_E = \int M(x)\chi_E(x)$$
$$M_{AE} = \int M(x)\chi_A(x)\chi_E(x) \qquad M_{BE} = \int M(x)\chi_B(x)\chi_E(x)$$
$$\alpha = d(\psi)^2 - 1 \qquad \beta = d(E)^2 - 1$$

giving

$$d(e)^2 \leq \max_M \frac{1 + \alpha M_A + \beta M_E + \alpha\beta M_{AE}}{1 + \alpha M_B + \beta M_E + \alpha\beta M_{BE}} \cdot \frac{1 + \alpha M_B}{1 + \alpha M_A}$$

Using the same argument outlined by Equation 30, one may argue that the scalars M_{AE}, M_{BE}, M_A, and M_B must also be extremum of their constraint sets. Noticing that M_{AE} should be large and M_{BE} small, we may summarize the constraints by

$$0 \le M_A,\ M_B,\ M_E \le 1 \qquad M_{AE} \le \min[M_A,\ M_E] \qquad M_{BE} \ge \max[0,\ M_E - (1 - M_B)]$$

(where the last constraint arises from the fact that $M_E + M_B - M_{BE} \le 1$). We then consider each possible case: $M_A \le M_E$, $M_A \ge M_E$, ... In each case, we find that the maximum is found at the extrema $M_{AE} = M_A = M_E$ and $M_E = 1 - M_B$. This gives

$$d\,(e)^2 \le \max_M \frac{1 + (\alpha + \beta + \alpha\beta)M_E}{1 + \alpha + (\beta - \alpha)M_E} \cdot \frac{1 + \alpha - \alpha M_E}{1 + \alpha M_E}$$

The maximum with respect to M_E (whose optimum is not an extreme point) is given by taking the derivative and setting it to zero. This procedure gives a quadratic equation; solving and selecting the positive solution gives $M_E = \frac{1}{\beta}(\sqrt{\beta + 1} - 1)$. Finally, plugging in, simplifying, and taking the square root yields

$$d\,(e) \le \frac{d\,(\psi)^2\, d\,(E) + 1}{d\,(\psi)^2 + d\,(E)}$$

□

Appendix B. Properties of the Expected Divergence

We begin by examining the properties of the expected divergence (25) on tree-structured graphical models parameterized by (21)-(22); we discuss the application of these results to graphs with cycles in Appendix B.4. Recall that, for tree-structured models described by (21)-(22), the prior-weight normalized messages of the (unique) fixed point are equivalent to

$$m_{ut}(x_t) = p(\mathbf{y}_{ut}|x_t)/p(\mathbf{y}_{ut}).$$

and that the message products are given by

$$M_{ts}(x_t) = p(x_t|\mathbf{y}_{ts})M_t(x_t) = p(x_t|\mathbf{y})$$

Furthermore, let us define the *approximate* messages $\hat{m}_{ut}(x)$ in terms of some approximate likelihood function, i.e., $\hat{m}_{ut}(x) = \hat{p}(\mathbf{y}_{ut}|x_t)/\hat{p}(\mathbf{y}_{ut})$ where $\hat{p}(\mathbf{y}_{ut}) = \int \hat{p}(\mathbf{y}_{ut}|x_t)p(x_t)dx_t$. We may then examine each of the three properties in turn: the triangle inequality, additivity, and contraction.

B.1 Triangle Inequality

Kullback-Leibler divergence is not a true distance, and in general, it does not satisfy the triangle inequality. However, the following generalization does hold:

Theorem 27. *For a tree-structured graphical model parameterized as in* (21)-(22), *and given the true BP message $m_{ut}(x_t)$ and two approximations $\hat{m}_{ut}(x_t)$, $\tilde{m}_{ut}(x_t)$, suppose that $m_{ut}(x_t) \le c_{ut}\hat{m}_{ut}(x_t)\ \forall x_t$. Then,*

$$\mathcal{D}(m_{ut}\|\tilde{m}_{ut}) \le \mathcal{D}(m_{ut}\|\hat{m}_{ut}) + c_{ut}\mathcal{D}(\hat{m}_{ut}\|\tilde{m}_{ut})$$

and furthermore, if $\hat{m}_{ut}(x_t) \le c^*_{ut}\tilde{m}_{ut}(x_t)\ \forall x_t$, then $m_{ut}(x_t) \le c_{ut}c^*_{ut}\tilde{m}_{ut}(x_t)\ \forall x_t$.

Comment. Since $m, \hat{m}$ are prior-weight normalized ($\int p(x)m(x) = \int p(x)\hat{m}(x) = 1$), for a strictly positive prior $p(x)$ we see that $c_{ut} \geq 1$, with equality if and only if $m_{ut}(x) = \hat{m}_{ut}(x)\ \forall x$. However, this is often quite conservative and Approximation 20 ($c_{ut} = 1$) is sufficient to estimate the resulting error. Moreover, we shall see that the constants $\{c_{ut}\}$ are also affected by the product operation, described next.

B.2 Near-Additivity

For BP fixed-point messages $\{m_{ut}(x_t)\}$, approximated by the messages $\{\hat{m}_{ut}(x_t)\}$, the resulting error is not quite guaranteed to be sub-additive, but is almost so.

Theorem 28. *The expected error $E[\mathcal{D}(M_t \| \hat{M}_t)]$ between the true and approximate beliefs is nearly sub-additive; specifically,*

$$E\left[\mathcal{D}(M_t \| \hat{M}_t)\right] \leq \sum_{u \in \Gamma_t} E\left[\mathcal{D}(m_{ut} \| \hat{m}_{ut})\right] + \left(\hat{I} - I\right) \tag{31}$$

$$\text{where} \qquad I = E\left[\log\ p(\mathbf{y}) / \prod_{u \in \Gamma_t} p(\mathbf{y}_{ut})\right] \qquad \text{and} \qquad \hat{I} = E\left[\log\ \hat{p}(\mathbf{y}) / \prod_{u \in \Gamma_t} \hat{p}(\mathbf{y}_{ut})\right]$$

Moreover, if $m_{ut}(x_t) \leq c_{ut}\hat{m}_{ut}(x_t)$ for all x_t and for each $u \in \Gamma_t$, then

$$M_t(x_t) \leq \prod_{u \in \Gamma_t} c_{ut} C_t^* \hat{M}_t(x_t) \qquad\qquad C_t^* = \frac{\hat{p}(\mathbf{y})}{\prod_{u \in \Gamma_t} \hat{p}(\mathbf{y}_{ut})} \frac{\prod_{u \in \Gamma_t} p(\mathbf{y}_{ut})}{p(\mathbf{y})} \tag{32}$$

Proof. By definition we have

$$E[\mathcal{D}(M_t \| \hat{M}_t)] = E\left[\int p(x_t, \mathbf{y}) \log \frac{M_t(x_t)}{\hat{M}_t(x_t)} dx_t\right] = E\left[\int p(x_t|\mathbf{y}) \log \frac{p(x_t)}{p(x_t)} \frac{p(\mathbf{y}|x_t)}{\hat{p}(\mathbf{y}|x_t)} \frac{\hat{p}(\mathbf{y})}{p(\mathbf{y})} dx_t\right]$$

Using the Markov property of (21) to factor $p(\mathbf{y}|x_t)$, we have

$$= E\left[\int p(x_t|\mathbf{y}) \sum_{u \in \Gamma_t} \log \frac{p(\mathbf{y}_{ut}|x_t)}{\hat{p}(\mathbf{y}_{ut}|x_t)} + p(x_t|\mathbf{y}) \log \frac{\hat{p}(\mathbf{y})}{p(\mathbf{y})} dx_t\right]$$

and, applying the identity $m_{ut}(x_t) = p(\mathbf{y}_{ut}|x_t)/p(\mathbf{y}_{ut})$ gives

$$\begin{aligned} &= \sum_{u \in \Gamma_t} E\left[\int p(x_t|\mathbf{y}) \log \frac{m_{ut}(x_t)}{\hat{m}_{ut}(x_t)}\right] + E\left[\log \frac{\hat{p}(\mathbf{y})}{\prod_u \hat{p}(\mathbf{y}_{ut})} \frac{\prod_u p(\mathbf{y}_{ut})}{p(\mathbf{y})}\right] dx_t \\ &= \sum_{u \in \Gamma_t} E\left[\mathcal{D}(m_{ut} \| \hat{m}_{ut})\right] + (\hat{I} - I) \end{aligned}$$

where $\hat{I}$, I are as defined. Here, I is the mutual information (the divergence from independence) of the variables $\{\mathbf{y}_{ut}\}_{u \in \Gamma_t}$. Equation (32) follows from a similar argument. $\square$

Unfortunately, it is *not* the case that the quantity $\hat{I} - I$ must necessarily be less than or equal to zero. To see how it may be positive, consider the following example. Let $x = [x_a, x_b]$ be a two-dimensional binary random variable, and let y_a and y_b be observations of the specified dimension of x. Then, if y_a and y_b are independent ($I = 0$), the true messages $m_a(x)$ and $m_b(x)$ have a regular structure; in particular, m_a and m_b have the forms $[p_1 p_2 p_1 p_2]$ and $[p_3 p_3 p_4 p_4]$ for some $p_1, \ldots, p_4$. However, we have placed no such requirements on the message *errors* $\hat{m}/m$; they have the potentially arbitrary forms $e_a = [e_1 e_2 e_3 e_4]$, *etc.*. If either message error e_a, e_b does *not* have the same structure as m_a, m_b respectively (even if they are random and independent), then $\hat{I}$ will in general not be zero. This creates the *appearance* of information between y_a and y_b, and the KL-divergence will not be strictly sub-additive.

However, this is not a typical situation. One may argue that in most problems of interest, the information I between observations is non-zero, and the types of message perturbations [particularly random errors, such as appear in stochastic versions of BP (Sudderth et al., 2003; Isard, 2003; Koller et al., 1999)] tend to degrade this information on average. Thus, is is reasonable to assume that $\hat{I} \leq I$.

A similar quantity defines the multiplicative constant C_t^* in (32). When $C_t^* \leq 1$, it acts to reduce the constant which bounds M_t by $\hat{M}_t$; if this occurs "typically", it lends additional support for Approximation (20). Moreover, if $E[C_t^*] \leq 1$, then by Jensen's inequality, we have $\hat{I} - I \leq 0$, ensuring sub-additivity as well.

B.3 Contraction

Analysis of the contraction of expected KL-divergence is also non-trivial; however, the work of Boyen and Koller (1998) has already considered this problem in some depth for the specific case of directed Markov chains (in which additivity issues do not arise) and projection-based approximations (for which KL-divergence does satisfy a form of the triangle inequality). We may directly apply their findings to construct Approximation 22.

Theorem 29. *On a tree-structured graphical model parameterized as in* (21)-(22)*, the error measure* $\mathcal{D}(M, \hat{M})$ *satisfies the inequality*

$$E\left[\mathcal{D}(m_{ts} \| \hat{m}_{ts})\right] \leq (1 - \gamma_{ts}) E\left[\mathcal{D}(M_{ts} \| \hat{M}_{ts})\right]$$

$$\text{where} \qquad \gamma_{ts} = \min_{a,b} \int \min\left[p(x_s | x_t = a)\,,\, p(x_s | x_t = b)\right] dx_s$$

Proof. For a detailed development, see Boyen and Koller (1998); we merely sketch the proof here. First, note that

$$\begin{aligned} E\left[\mathcal{D}(m_{ts} \| \hat{m}_{ts})\right] &= E\left[\int p(x_s | \mathbf{y}) \log \frac{p(\mathbf{y}_{ts} | x_s)}{p(\mathbf{y}_{ts})} \frac{\hat{p}(\mathbf{y}_{ts})}{\hat{p}(\mathbf{y}_{ts} | x_s)}\right] \\ &= E\left[\int p(x_s | \mathbf{y}_{ts}) \log \frac{p(x_s | \mathbf{y}_{ts})}{\hat{p}(x_s | \mathbf{y}_{ts})}\right] \\ &= E\left[D(p(x_s | \mathbf{y}_{ts}) \| \hat{p}(x_s | \mathbf{y}_{ts}))\right] \end{aligned}$$

[which is the quantity considered in Boyen and Koller (1998)], and further that $p(x_s | \mathbf{y}_{ts}) = \int p(x_s | x_t) p(x_t | \mathbf{y}_{ts}) dx_t$. By constructing two valid conditional distributions $f_1(x_s | x_t)$ and

$f_2(x_s|x_t)$ such that f_1 has the form $f_1(x_s|x_t) = f_1(x_s)$ (independence of x_s, x_t), and

$$p(x_s|x_t) = \gamma_{ts} f_1(x_s|x_t) + (1 - \gamma_{ts} f_2(x_s|x_t)$$

one may use the convexity of KL-divergence to show

$$D(p(x_s|\mathbf{y}_{ts}) \| \hat{p}(x_s|\mathbf{y}_{ts})) \leq \gamma_{ts} D(f_1 * p(x_t|\mathbf{y}_{ts}) \| f_1 * \hat{p}(x_t|\mathbf{y}_{ts})) + \\ (1 - \gamma_{ts}) D(f_2 * p(x_t|\mathbf{y}_{ts}) \| f_2 * \hat{p}(x_t|\mathbf{y}_{ts}))$$

where "$*$" denotes convolution, i.e., $f_1 * p(x_t|\mathbf{y}_{ts}) = \int f_1(x_s|x_t) p(x_t|\mathbf{y}_{ts}) dx_t$. Since the conditional f_1 induces independence between x_s and x_t, the first divergence term is zero, and since f_2 is a valid conditional distribution, the second divergence term is less than $D(p(x_t|\mathbf{y}_{ts}) \| \hat{p}(x_t|\mathbf{y}_{ts}))$ [see Cover and Thomas (1991)]. Thus we have a minimum rate of contraction of $(1 - \gamma_{ts})$. $\square$

It is worth noting that Theorem 29 gives a *linear* contraction rate. While this makes for simpler recurrence relations than the nonlinear contraction found in Section 4.2, it has the disadvantage that, if the rate of error addition exceeds the rate of contraction it may result in a trivial (infinite) bound. Theorem 29 is the best contraction rate currently known for arbitrary conditional distributions, although certain special cases (such as binary–valued random variables) appear to admit stronger contractions.

B.4 Graphs with Cycles

The analysis and discussion of each property (Appendices B.1–B.3) also relied on assuming a tree-structured graphical model, and using the direct relationship between messages and likelihood functions for the parameterization (21)-(22). However, for BP on general graphs, this parameterization is not valid.

One way to generalize this choice is given by the re-parameterization around some fixed point of loopy BP on the graphical model of the prior. If the original potentials $\tilde{\psi}_{st}, \tilde{\psi}_s^x$ specify the prior distribution [cf. (22)],

$$p(\mathbf{x}) \propto \prod_{(s,t)\in\mathcal{E}} \tilde{\psi}_{st}(x_s, x_t) \tilde{\prod}_s \psi_s^x(x_s) \tag{33}$$

then given a BP fixed point $\{\tilde{M}_{st}, \tilde{M}_s\}$ of (33), we may choose a new parameterization of the same prior ψ_{st}, ψ_s^x given by

$$\psi_{st}(x_s, x_t) = \frac{\tilde{M}_{st}(x_s) \tilde{M}_{ts}(x_t) \tilde{\psi}_{st}(x_s, x_t)}{\tilde{M}_s(x_s) \tilde{M}_t(x_t)} \qquad \text{and} \qquad \psi_s^x(x_s) = \tilde{M}_s(x_s) \tag{34}$$

This parameterization ensures that uninformative messages [$m_{ut}(x_t) = 1 \ \forall x_t$] comprise a fixed point for the graphical model of $p(\mathbf{x})$ as described by the new potentials $\{\psi_{st}, \psi_s\}$. For a tree-structured graphical model, this recovers the parameterization given by (22).

However, the messages of loopy BP are no longer precisely equal to the likelihood functions $m(x) = p(\mathbf{y}|x)/p(\mathbf{y})$, and thus the expectation applied in Theorem 28 is no longer consistent with the messages themselves. Additionally, the additivity and contraction statements were developed under the assumption that the observed data $\mathbf{y}$ along different branches of the tree are conditionally *independent*; in graphs with cycles, this is not the case. In the computation tree formalism, instead of being conditionally independent, the observations $\mathbf{y}$ actually *repeat* throughout the tree.

However, the assumption of independence is precisely the same assumption applied by loopy belief propagation itself to perform tractable approximate inference. Thus, for problems in which loopy BP is well-behaved and results in answers similar to the true posterior distributions, we may expect our estimates of belief error to be similarly incorrect but near to the true divergence.

In short, all three properties required for a strict analysis of the propagation of errors in BP fail, in one sense or another, for graphs with cycles. However, for many situations of practical interest, they are quite close to the real average-case behavior. Thus we may expect that our approximations give rise to reasonable estimates of the total error incurred by approximate loopy BP, an intuition which appears to be borne out in our simulations (Section 6.4).

References

Xavier Boyen and Daphne Koller. Tractable inference for complex stochastic processes. In *Uncertainty in Artificial Intelligence*, pages 33–42, 1998.

Hei Chan and Adnan Darwiche. A distance measure for bounding probabilistic belief change. *International Journal of Approximate Reasoning*, 38(2):149–174, Feb 2005.

L. Chen, M. Wainwright, M. Cetin, and A. Willsky. Data association based on optimization in graphical models with application to sensor networks. Submitted to *Mathematical and Computer Modeling*, 2004.

P. Clifford. Markov random fields in statistics. In G. R. Grimmett and D. J. A. Welsh, editors, *Disorder in Physical Systems*, pages 19–32. Oxford University Press, Oxford, 1990.

J. M. Coughlan and S. J. Ferreira. Finding deformable shapes using loopy belief propagation. In *European Conference on Computer Vision 7*, May 2002.

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, New York, 1991.

Hans-Otto Georgii. *Gibbs measures and phase transitions*. Studies in Mathematics. de Gruyter, Berlin / New York, 1988.

A. Gersho and R. M. Gray. *Vector quantization and signal compression*. Kluwer, Boston, 1991.

T. Heskes. On the uniqueness of loopy belief propagation fixed points. *Neural Computation*, 16(11): 2379–2413, 2004.

A. T. Ihler, J. W. Fisher III, and A. S. Willsky. Communication-constrained inference. Technical Report 2601, MIT, Laboratory for Information and Decision Systems, 2004a.

A. T. Ihler, J. W. Fisher III, and A. S. Willsky. Message errors in belief propagation. In *Neural Information Processing Systems*, 2004b.

M. Isard. PAMPAS: Real–valued graphical models for computer vision. In *IEEE Computer Vision and Pattern Recognition*, 2003.

S. Julier and J. Uhlmann. A general method for approximating nonlinear transformations of probability distributions. Technical report, RRG, Dept. of Eng. Science, Univ. of Oxford, 1996.

D. Koller, U. Lerner, and D. Angelov. A general algorithm for approximate inference and its application to hybrid Bayes nets. In *Uncertainty in Artificial Intelligence 15*, pages 324–333, 1999.

F. Kschischang, B. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Transactions on Information Theory*, 47(2):498–519, February 2001.

S. L. Lauritzen. *Graphical Models.* Oxford University Press, Oxford, 1996.

T. Minka. Expecatation propagation for approximate bayesian inference. In *Uncertainty in Artificial Intelligence*, 2001.

M. A. Paskin and C. E. Guestrin. Robust probabilistic inference in distributed systems. In *Uncertainty in Artificial Intelligence 20*, 2004.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems.* Morgan Kaufman, San Mateo, 1988.

E. B. Sudderth, A. T. Ihler, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *IEEE Computer Vision and Pattern Recognition*, 2003.

S. Tatikonda and M. Jordan. Loopy belief propagation and gibbs measures. In *Uncertainty in Artificial Intelligence*, 2002.

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree–based reparameterization analysis of sum–product and its generalizations. *IEEE Transactions on Information Theory*, 49(5), May 2003.

Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1), 2000.

A. Willsky. Relationships between digital signal processing and control and estimation theory. *Proceedings of the IEEE*, 66(9):996–1017, September 1978.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. Technical Report 2004-040, MERL, May 2004.