Focusing Generalizations of Belief Propagation on Targeted Queries

Arthur Choi and **Adnan Darwiche**
Computer Science Department
University of California, Los Angeles
Los Angeles, CA 90095
{*aychoi,darwiche*}*@cs.ucla.edu*

Abstract

A recent formalization of Iterative Belief Propagation (IBP) has shown that it can be understood as an exact inference algorithm on an approximate model that results from deleting every model edge. This formalization has led to (1) new realizations of Generalized Belief Propagation (GBP) in which edges are recovered incrementally to improve approximation quality, and (2) edge-recovery heuristics that are motivated by improving the approximation quality of all node marginals in a graphical model. In this paper, we propose new edge-recovery heuristics, which are focused on improving the approximations of targeted node marginals. The new heuristics are based on newly-identified properties of edge deletion, and in turn IBP, which guarantee the exactness of edge deletion in simple and idealized cases. These properties also suggest new improvements to IBP approximations which are based on performing edge-by-edge corrections on targeted marginals, which are less costly than improvements based on edge recovery.

Introduction

Among approximation algorithms for reasoning in probabilistic graphical models, iterative belief propagation (IBP), also known as loopy belief propagation (Pearl 1988; Murphy, Weiss, and Jordan 1999) has been extremely influential in certain classes of applications. For instance, IBP has spawned approaches capable of solving particularly difficult instances of the satisfiability problem (Braunstein, Mézard, and Zecchina 2005), and has shown to be an effective approach to a variety of computer vision tasks (Szeliski et al. 2006), particularly in stereo vision. Its biggest impact has been in the field of information theory, where revolutionary algorithms for decoding error-correcting codes have shown to be instances of iterative belief propagation in a Bayesian network (Frey and MacKay 1997; McEliece, MacKay, and Cheng 1998).

The successes of IBP as an approximate inference algorithm spurred many improvements and generalizations, including the family of Generalized Belief Propagation (GBP) algorithms (Yedidia, Freeman, and Weiss 2005). More recently, we proposed a special class of GBP approximations, called ED-BP (Choi and Darwiche 2006), which characterized IBP as an algorithm in a fully-disconnected approximation of the original network, found by deleting every edge from the original. By *recovering* edges back into the approximation, we can seek more structured, and hopefully more accurate, approximations. This leads to a spectrum of approximations, with IBP on one end (when every edge is deleted) to exact inference on the other (when every is recovered).

Identifying good instances in this spectrum is vital to the success of this and other similar approaches to approximate inference, as no single instance will likely be effective for all possible queries. Indeed, we proposed in (Choi and Darwiche 2006) a mutual information heuristic for ED-BP in Bayesian networks that is sensitive to both the network parametrization and to the observations at hand, and is further based on a global property of edge deletion that guarantees exact marginals for every variable in the model. This approach can provide good approximations (even exact) for many variables, but may still provide only poor approximations for others.

One must then ask if this is the ideal approach, particularly when one is interested in a particular query variable. Indeed, from query-to-query, ones focus may change from one sub-model to another, while varying observations may render different parts of the model irrelevant. Ideally, one would like to target the approximation so as to maximize the accuracy of the probabilities one is truly interested in, giving less weight to those parts of the model that are only weakly relevant to the query at hand.

We propose here a focused approach to edge recovery in ED-BP approximations that is query-sensitive, and further targets the variables of interest. It is based on new conditions that are sufficient for the exactness of a particular variable's marginals, in the simplified case where a single edge is deleted. The resulting analysis suggests a new heuristic for *focused edge-recovery*, which is empirically more effective at improving the approximation of a targeted variable of interest, compared to a heuristic based on the network as a whole. Our analysis further leads to a notion of *marginal corrections*, that allows a targeted approximation to be improved further when edge recovery becomes infeasible. We illustrate the benefit of a focused approach to approximating marginals, experimentally.

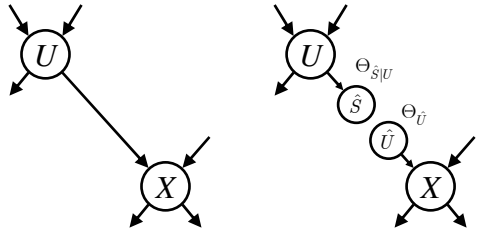

Figure 1: Deleting an edge $U \rightarrow X$ by adding a clone $\hat{U}$ of U and binary evidence variable $\hat{S}$.

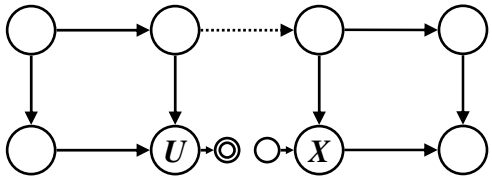

Figure 2: In the absence of the dotted edge, the deleted edge $U \rightarrow X$ split the network into two and ED-BP yields exact marginals. In the presence of the dotted edge, ED-BP parameters may be overcompensating.

Edge Deletion and Belief Propagation

We review here ED-BP, the edge deletion approach to approximate inference, introduced in (Choi and Darwiche 2006), that gives rise to a class of Generalized Belief Propagation (GBP) algorithms (Yedidia, Freeman, and Weiss 2005).

Let $U \rightarrow X$ be an edge in a Bayesian network $\mathcal{N}$ that we wish to delete. When we do so however, we wish also to introduce two auxiliary variables to compensate for the dependencies lost due to edge deletion; see Figure 1.

Definition 1 (Edge Deletion) *Let $U \rightarrow X$ be an edge in a Bayesian network $\mathcal{N}$. We say that the edge $U \rightarrow X$ is deleted when it results in a network that is obtained from $\mathcal{N}$ by: (1) removing the edge $U \rightarrow X$ from the graph; (2) introducing a clone $\hat{U}$ that replaces U as a parent of X; and (3) introducing an instantiated variable $\hat{S}$ that replaces X as a child of U.*

The deletion of an edge $U \rightarrow X$ thus introduces new parameters into the network: we must specify for the clone variable $\hat{U}$ the parameters $\theta_{\hat{u}}$, and further for the variable $\hat{S}$ the parameters $\theta_{\hat{s}|u}$. Since $\hat{S}$ is instantiated, say to state $\hat{s}$, we only need to specify two vectors of size $|U|$, which we shall refer to as *edge parameters.*

Given a network $\mathcal{N}$ and evidence $\mathbf{e}$, our proposal is then to approximate this network with another $\mathcal{N}'$ that results from deleting some number of edges $U \rightarrow X$ as defined in Definition 1. Moreover, when performing inference on network $\mathcal{N}'$, we will condition on the augmented evidence $\mathbf{e}'$ composed of the original evidence $\mathbf{e}$ and each piece of auxiliary evidence $\hat{s}$ introduced when deleting edges. More formally, if Pr and Pr' are the distributions induced by networks $\mathcal{N}$ and $\mathcal{N}'$, respectively, we will use the conditional distribution $Pr'(.|\mathbf{e}')$ to approximate $Pr(.|\mathbf{e})$.

Before we can use a network $\mathcal{N}'$ to approximate queries in $\mathcal{N}$, we must first specify the values of our edge parameters. The parameters $\theta_{\hat{u}}$ for the clone $\hat{U}$ should compensate for the lost influence that parent U had on child X. The parameters $\theta_{\hat{s}|u}$ for the instantiated variable $\hat{S}$ should compensate for the loss of evidential information that variable U received through its child X. In the simplified scenario where deleting a single edge $U \rightarrow X$ splits a network into two disconnected subnetworks, we can in fact identify edge parameters that compensate for the deleted edge precisely.

Condition 1 *Let $\mathcal{N}$ be a Bayesian network and $\mathcal{N}'$ be the result of deleting edges $U \rightarrow X$ from $\mathcal{N}$. Edge parameters of an* ED-BP *approximation satisfy the following conditions:*

$$\begin{aligned} Pr'(U = u \mid \mathbf{e}') &= Pr'(\hat{U} = \hat{u} \mid \mathbf{e}'), \\ Pr'(U = u \mid \mathbf{e}' \setminus \hat{s}) &= Pr'(\hat{U} = \hat{u}) \end{aligned}$$

for all values $u = \hat{u}$.

These conditions state first that the parent U and its clone $\hat{U}$ should have the same posterior marignals. Given this, it further states that the strength of soft evidence $\hat{s}$ on U is the same as the strength of all evidence $\mathbf{e}'$ on $\hat{U}$.

If an ED-BP network $\mathcal{N}'$ was the result of deleting a single edge that split the network into two independent subnetworks, then Condition 1 guarantees that marginals in each subnetwork are exact. In general, when many edges are deleted, a network $\mathcal{N}'$ satisfying Condition 1 may still be a useful approximation; see Figure 2. Condition 1 is further equivalent to local conditions on edge parameters:

$$\theta_{\hat{u}} = Pr'(u \mid \mathbf{e}' \setminus \hat{s}) \quad (1)$$

$$\theta_{\hat{s}|u} = Pr'(\mathbf{e}' \mid \hat{u}). \quad (2)$$

These local conditions can also be used as update equations in an iterative fixed-point procedure (also called ED-BP) to search for parameters satisfying Condition 1. Starting with an initial approximation $\mathcal{N}'_0$ at iteration 0 (say, with uniform parameters), we can compute edge parameters $\theta^t_{\hat{u}}$ and $\theta^t_{\hat{s}|u}$ for an iteration $t > 0$ by performing exact inference in the simplified network $\mathcal{N}'_{t-1}$. We can repeat this process until all edge parameters converge (if ever) to a fixed point satisfying Equations 1 and 2, and thus Condition 1.

Finally, message passing by IBP in $\mathcal{N}$ corresponds to edge parametrization by ED-BP in $\mathcal{N}'$, in the degenerate case where every edge has been deleted in $\mathcal{N}'$. Moreover, the IBP approximations of node marginals in $\mathcal{N}$ correspond precisely to the marginals $Pr'(X|\mathbf{e}')$ computed exactly in $\mathcal{N}'$. This correspondence to IBP continues to hold, in fact, for all polytree approximations $\mathcal{N}'$; cf. (Wainwright, Jaakkola, and Willsky 2003). When $\mathcal{N}'$ is multiply-connected (i.e., has undirected cycles), then ED-BP induces a class of GBP approximations.[1] Therefore, by choosing edges to delete, we implicitly choose also the structure of a GBP approximation.

[1] An ED-BP approximation $\mathcal{N}'$ corresponds to an instance of GBP run with a particular joingraph (Choi and Darwiche 2006; Aji and McEliece 2001; Dechter, Kask, and Mateescu 2002).

Deleting a Single Edge

Suppose now we are given a Bayesian network $\mathcal{N}$, and we are interested in approximating the marginal distribution of a target variable Q, by computing it in an ED-BP approximation $\mathcal{N}'$. For simplicity, suppose that Q is binary, i.e., Q takes on either the state q or the state $\bar{q}$.[2] For variable Q and state q, we may simply refer to the event $Q = q$ as q, and hence refer to $Pr(Q = q|\mathbf{e})$ as simply $Pr(q|\mathbf{e})$.

It will be convenient for us to think of the marginal distribution $Pr(Q|\mathbf{e})$, conditioned on evidence $\mathbf{e}$, in terms of the *odds* of an event $Q = q$:

$$O(q \mid \mathbf{e}) = \frac{Pr(q \mid \mathbf{e})}{Pr(\bar{q} \mid \mathbf{e})} = \frac{Pr(q \mid \mathbf{e})}{1 - Pr(q \mid \mathbf{e})}.$$

Similarly, let $O'(q|\mathbf{e}')$ denote the odds of $Q = q$ associated with the approximate distribution $Pr'(Q|\mathbf{e}')$. Note that we can easily recover the probabilities $Pr(q|\mathbf{e})$ and $Pr(\bar{q}|\mathbf{e})$ given the odds $O(q|\mathbf{e})$.

Consider then the simplified case where we have deleted a single edge $U \rightarrow X$. The question now is: under what conditions are the odds $O'(q|\mathbf{e}')$ exact? An ED-BP approximation yields the exact odds for all network variables when deleting the single edge $U \rightarrow X$ splits the network into two (due to Condition 1). This, however, is too strong of an assumption when we are interested in only the odds $O'(q|\mathbf{e}')$, as a particular variable Q may be indifferent to the deletion of the edge $U \rightarrow X$ (say if variable Q and edge $U \rightarrow X$ were already disconnected).

To help us answer this question, we can examine the simplified case where a single edge $U \rightarrow X$ is deleted, where we can in fact express the error of the odds $O'(q|\mathbf{e}')$ in terms of the distribution induced by the approximate network $\mathcal{N}'$.

Lemma 1 *If $\mathcal{N}'$ is an* ED-BP *approximation resulting from the deletion of a single edge $U \rightarrow X$ from a Bayesian network $\mathcal{N}$, then the odds-error $E = O'(q|\mathbf{e}')/O(q|\mathbf{e}')$ is*

$$E = \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \frac{\sum_{u=\hat{u}} Pr'(\bar{q} \mid u\hat{u}, \mathbf{e}')Pr'(u \mid \hat{u}, \mathbf{e}')}{\sum_{u=\hat{u}} Pr'(q \mid u\hat{u}, \mathbf{e}')Pr'(u \mid \hat{u}, \mathbf{e}')}$$
$$= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \frac{\sum_{u=\hat{u}} Pr'(\bar{q} \mid u\hat{u}, \mathbf{e}')Pr'(\hat{u} \mid u, \mathbf{e}')}{\sum_{u=\hat{u}} Pr'(q \mid u\hat{u}, \mathbf{e}')Pr'(\hat{u} \mid u, \mathbf{e}')}.$$

Note that the above two equations differ only in the factors $Pr'(u|\hat{u}, \mathbf{e}') = Pr'(\hat{u}|u, \mathbf{e}')$. The proofs of this lemma, and the subsequent propositions, appear in the appendix.

By specifying the odds-error E in terms of the approximate distribution $Pr'(.|\mathbf{e}')$ only, we can identify conditions where deleting a single edge leads to the exact odds, but in terms of independencies in the *approximate* network. For example, we can find by inspection the independence condition in the following result yields an odds-error of one.

Proposition 1 *If $\mathcal{N}'$ is an* ED-BP *approximation that results from deleting a single edge $U \rightarrow X$ from a Bayesian network $\mathcal{N}$, then $(Q \perp U, \hat{U} \mid \mathbf{e}')$ implies $O'(q|\mathbf{e}') = O(q|\mathbf{e})$.*

[2]The results in this paper can easily be extended to the case of variables Q with arbitrary arity.

Here, $(Q \perp U, \hat{U} \mid \mathbf{e}')$ means that Q is independent of U and $\hat{U}$, given evidence $\mathbf{e}'$. This proposition says, roughly, that if q is independent of the deleted edge $U \rightarrow X$, then the odds $O'(q|\mathbf{e}')$ is exact. This is intuitive and expected. With some simple manipulations, we can also find the following.

Proposition 2 *If $\mathcal{N}'$ is an* ED-BP *approximation that results from deleting a single edge $U \rightarrow X$ from a Bayesian network $\mathcal{N}$, then $(Q, U \perp \hat{U} \mid \mathbf{e}')$ or $(Q, \hat{U} \perp U \mid \mathbf{e}')$ implies $O'(q|\mathbf{e}') = O(q|\mathbf{e})$.*

In the following section, we propose a focused approach to edge recovery that is inspired by these propositions.

Focused Edge-Recovery

We can simplify a model by deleting edges, until exact inference is tractable. In turn, we can exploit the simplified network in order to find more structured, and hopefully more accurate, approximations. Indeed, (Choi and Darwiche 2006) propose an approach that takes a polytree approximation (corresponding to IBP), and recovers into it those edges that are deemed important for an accurate approximation.

Let the mutual information between two disjoint sets of variables $\mathbf{X}$ and $\mathbf{Y}$, which we will compute in a simplified network $\mathcal{N}'$, be defined as follows:

$$MI(\mathbf{X}; \mathbf{Y}|\mathbf{e}') = \sum_{\mathbf{xy}} Pr'(\mathbf{xy}|\mathbf{e}') \log \frac{Pr'(\mathbf{xy}|\mathbf{e}')}{Pr'(\mathbf{x}|\mathbf{e}')Pr'(\mathbf{y}|\mathbf{e}')}.$$

Note that mutual information is non-negative, and zero iff X and Y are independent given $\mathbf{e}'$ (Cover and Thomas 1991). Note further that the above mutual information is defined for a distribution $Pr'(.|\mathbf{e})$ conditioned on evidence $\mathbf{e}$.

We proposed in (Choi and Darwiche 2006) to recover first those edges $U \rightarrow X$ with high mutual information $MI(U; \hat{U}|\mathbf{e}')$. We can understand the rationale behind this heuristic by considering again the case where deleting a single edge $U \rightarrow X$ splits a network into two independent subnetworks, where the mutual information $MI(U; \hat{U}|\mathbf{e}')$ happens to be zero. Since splitting a network into two leads to exact odds for all variables, we may expect that recovering edges with high values of $MI(U; \hat{U}|\mathbf{e}')$ may improve, globally, the quality of the approximation.

In contrast, Propositions 1 and 2 suggest query-specific mutual information heuristics for deciding which edges to recover. First, Proposition 1, which says that the odds of Q are exact when $(Q \perp U, \hat{U} \mid \mathbf{e}')$ in $\mathcal{N}'$, suggests that we should recover those edges with the highest score:

$$s_1 = MI(Q; U\hat{U} \mid \mathbf{e}').$$

Second, Proposition 2, which says the odds of Q are exact when either $(Q, U \perp \hat{U} \mid \mathbf{e}')$ or $(Q, \hat{U} \perp U \mid \mathbf{e}')$ in $\mathcal{N}'$, suggests that we recover those edges with the highest scores:

$$s_2 = MI(QU; \hat{U} \mid \mathbf{e}')$$
$$s_3 = MI(Q\hat{U}; U \mid \mathbf{e}').$$

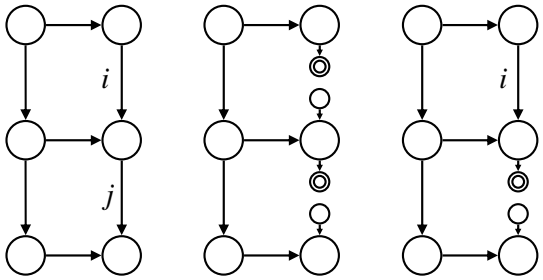

Figure 3: A Bayesian network $\mathcal{N}$ (left), a network $\mathcal{N}'$ where edges i and j have been deleted in $\mathcal{N}$ (center), and a network $\mathcal{N}'_i$ where edge i has been recovered back into $\mathcal{N}'$ (right).

Finally, since if any of the conditions given by Propositions 1 or 2 are satisfied, the odds $O'(q|\mathbf{e}')$ are correct, we take the minimum as a score to rank an edge $U \rightarrow X$ for recovery:

$$s_{ij} = \min\{s_1, s_2, s_3\}.$$

We then prefer to recover those edges $U \rightarrow X$ with the highest scores s_{ij}. Note that although the scores s_1, s_2 and s_3 are symmetrical, the conditions on which they are based on are independent, in the sense that one condition does not imply another. Note also that in the case $Q = U$, the score s_{ij} reduces to $MI(U; \hat{U} \mid \mathbf{e}')$, which is the score given to all edges $U \rightarrow X$ in the (unfocused) edge-recovery heuristic given by (Choi and Darwiche 2006).

Assuming we use a jointree-based algorithm for exact inference in $\mathcal{N}'$, we can use the technique we described in (Choi and Darwiche 2006), to compute all joint marginals required for scoring edges. Let T be the number of tails U in the set of edges deleted. To compute the unfocused edge-recovery scores, requiring marginals $Pr'(U\hat{U}|\mathbf{e}')$, we require only $O(T \max_U |U|)$ jointree propagations. To compute the focused edge-recovery scores, requiring marginals $Pr'(QU\hat{U}|\mathbf{e}')$, we require only $O(|Q| \cdot T \max_U |U|)$ jointree propagations. Assuming that Q is binary, this is the same complexity as the unfocused edge-recovery heuristic.

Odds Correction

Given the error E when deleting a single edge $U \rightarrow X$, we can trivially correct the approximate odds $O'(q|\mathbf{e}')$ to the true odds: $O'(q|\mathbf{e}') \cdot E^{-1} = O(q|\mathbf{e})$. We now ask: can we exploit this fact to improve our approximation $O'(q|\mathbf{e}')$ when more than one edge is deleted?

Say edges $U \xrightarrow{i} X$ deleted in network $\mathcal{N}'$ are labeled with indices i. Since Lemma 1 specifies the odds-error in terms of the approximate distribution $Pr'(.|\mathbf{e}')$ only, we can compute the correction factor E_i^{-1} for a particular edge $U \xrightarrow{i} X$, even when other edges have been deleted. Although this single correction factor is insufficient to rectify the odds $O'(q|\mathbf{e}')$ to the odds $O(q|\mathbf{e})$, it is sufficient to rectify the approximate odds to the exact odds of Q in another network where the single edge has been recovered.

Consider, for example, Figure 3, where we have deleted edges from a network $\mathcal{N}$. Using only the resulting network $\mathcal{N}'$, we can compute the odds $O'_i(q|\mathbf{e}'_i)$ of the hypothetical network $\mathcal{N}'_i$ where the single edge $U \xrightarrow{i} X$ has been recovered into $\mathcal{N}'$.[3] In particular, we can apply the odds-correction $E_i^{-1} = O'_i(q|\mathbf{e}'_i)/O'(q|\mathbf{e}')$ to the odds $O'(q|\mathbf{e}')$. As far as the target approximation $O'(q|\mathbf{e}')$ is concerned, we have effectively recovered the single edge $U \xrightarrow{i} X$ back into the approximation, but without the explicit construction of another network $\mathcal{N}'_i$.

This suggests an *edge-correction* approach to approximating the odds $O(q|\mathbf{e})$, where we collect single-edge corrections E_i^{-1}, and apply them to $O'(q|\mathbf{e}')$ as if the corrections were independent. In particular, we propose to accumulate corrections multiplicatively across all edges $U \xrightarrow{i} X$ deleted:

$$O'(q \mid \mathbf{e}') \cdot \prod_{U \xrightarrow{i} X} E_i^{-1} = O'(q \mid \mathbf{e}') \cdot \prod_{U \xrightarrow{i} X} \frac{O'_i(q \mid \mathbf{e}'_i)}{O'(q \mid \mathbf{e}')}. \quad (3)$$

As we shall see in the following section, this simple edge-correction scheme can be effective in improving the accuracy of an approximation for a targeted query $O(q|\mathbf{e})$.

Note that both edge-correction and edge-recovery require the computation of the joint marginals $Pr'(QU\hat{U} \mid \mathbf{e}')$, thus computing edge-recovery scores is computationally equivalent to computing edge-correction factors. In edge recovery, however, edges must then be recovered explicitly into a more connected approximation, where inference can become more difficult. In edge-correction, we only implicitly recover substructure, and is thus a cheaper alternative to improving a targeted odds-approximation.[4]

In Algorithm 1, we summarize the resulting approach to computing odds-approximations, where we perform edge-recovery, followed by edge-correction.

Empirical Results

We evaluate here the effectiveness of a focused approach to edge recovery, in a selection of Bayesian networks.[5] In our experiments, we performed an adaptive edge recovery procedure (Choi and Darwiche 2006), which is described in Steps 1 to 7 of Algorithm 1. We rank edges deleted in Step 4 in three ways: randomly (for reference), by the unfocused mutual information heuristic of (Choi and Darwiche 2006), and by our focused heuristic. We also performed focused edge-recovery, with and without edge-corrections. Our primary concern is the quality of approximations computed in

[3] We assume that when we recover this edge, all other edge parameters remain fixed. Note that network $\mathcal{N}'$ is an ED-BP approximation (satisfying Condition 1), in both cases: (1) as the result of deleting edges in $\mathcal{N}$, and (2) as the result of deleting an edge in $\mathcal{N}'_i$.

[4] In principle, we could have proposed instead to approximate a target marginal $Pr(q|\mathbf{e})$ as a ratio of two approximations $Pr'_q(q, \mathbf{e}')$ and $Pr'(\mathbf{e}')$ computed in two different ED-BP networks $\mathcal{N}'_q$ and $\mathcal{N}'$ (the former using new edge parameters found after conditioning on $Q = q$). Although this approach may be worth exploring, we consider it a nontrivial extension.

[5] Most of the networks used for our evaluation are available at http://www.cs.huji.ac.il/labs/compbio/Repository/. Networks `emdec` and `tcc` are noisy-or networks for diagnosis, courtesy of HRL Laboratories, LLC.

Algorithm 1 EDGE-RECOVERY,EDGE-CORRECTION

input:
- $\mathcal{N}$: a Bayesian network
- $\mathbf{e}$: an instantiation of some variables in network $\mathcal{N}$
- q: target query q for variable Q in $\mathcal{N}$

output: an approximation to the odds $O(q \mid \mathbf{e}')$

main:

1: $\Sigma \leftarrow$ edges to delete to render $\mathcal{N}$ a spanning polytree
2: $\mathcal{N}' \leftarrow$ ED-BP$(\mathcal{N}, \mathbf{e}, \Sigma)$
3: **while** recovery of edges in $\mathcal{N}'$ is feasible **do**
4: rank deleted edges $U \rightarrow X$ (given a heuristic)
5: $\Sigma \leftarrow \Sigma - \{$top k edges with the largest scores$\}$
6: $\mathcal{N}' \leftarrow$ ED-BP$(\mathcal{N}, \mathbf{e}, \Sigma)$
7: **end while**
8: $O'(q \mid \mathbf{e}') \leftarrow$ odds of q computed in $\mathcal{N}'$
9: $\{E_i\} \leftarrow$ odds-errors for edges Σ still deleted
10: **return** odds-correction $O'(q \mid \mathbf{e}') \cdot \prod_i E_i^{-1}$

supporting function: ED-BP$(\mathcal{N}, \mathbf{e}, \Sigma)$: returns ED-BP network for network $\mathcal{N}$, evidence $\mathbf{e}$, with edges Σ deleted

the ED-BP network resulting in Step 6, or if edge-recovery is not feasible, the edge-corrected approximation returned in Step 10.

Each plot that we present corresponds to a particular Bayesian network where edge recovery heuristics are evaluated based on an average of at least 50 problem instances. Each problem instance corresponds to observations $\mathbf{e}$ on all leaves of the original network, whose values are sampled from the original joint distribution (except for networks `emdec` and `tcc`, where we set values on leaves at random as the joint distribution is highly skewed). In all problem instances, IBP converged to within an absolute difference of 10^{-8}, in 200 iterations.

Consider then Figure 4, where each row corresponds to a particular Bayesian network. We compare the number of edges recovered (x-axis) versus the accuracy of marginal approximations (y-axis). We measured the error of a marginal by $\max_q |Pr'(q|\mathbf{e}') - Pr(q|\mathbf{e})|$ (we assume in our experiments that q is of arbitrary arity, not just binary). In the left column of Figure 4, we report the average marginal error of a target variable Q, which we picked as the variable with the largest marginal error in an IBP approximation (computed beforehand). On the right column, we report the average marginal error over all unobserved variables, whether it was the target or not. As we move from left-to-right on the x-axis, we move from the case where no edge is recovered (corresponding to IBP) to the case where all edges are recovered (corresponding to exact inference), recovering $\frac{1}{8}$-th of the edges at a time. In networks `barley` and `mildew` however, which are relatively more challenging to evaluate, we recover only 25% of the edges, $\frac{1}{16}$-th of the edges at a time.

In all plots, both recovery heuristics produced on average more accurate marginals than random recovery. On the left column, where we compare average marginal error for the target variable only, focused recovery (FOCUS)

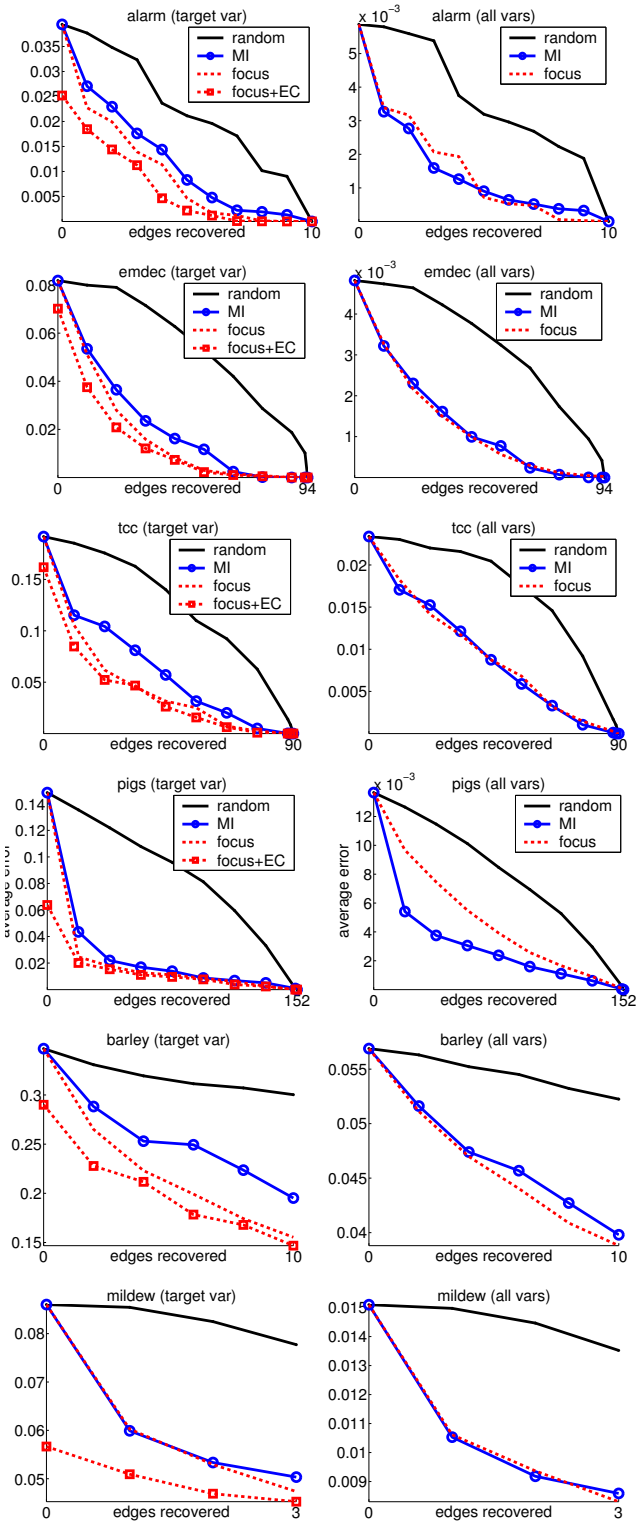

Figure 4: [The Effectiveness of Focused Approximations]. Left: Average marginal error for the target variable Q. Right: Average marginal error, averaged over all variables. In all plots, the y-axis is average error.

produced consistently better approximations than unfocused recovery (MI) for all networks but the mildew network, where they were not well distinguishable. This improvement was modest in some cases, but we believe the separation is sufficient to justify the usage of a focused heuristic over an unfocused one, as FOCUS requires essentially the *same time complexity* as MI (as we discussed when we introduced the focused heuristic). For focused recovery, we further computed edge-corrections (FOCUS+EC) which, as we discussed in the previous section, is computationally equivalent to computing edge recovery scores, yet does not require us to actually recover edges. We see here consistent improvement over focused recovery alone, giving us *another* layer of improvement over unfocused recovery. Moreover, FOCUS+EC yielded significant improvements in approximation quality for the pigs and mildew networks, without recovering *any* edges.

On the right column of Figure 4, where we compared approximation quality based on average error over all variables, we found that, not surprisingly, the unfocused MI heuristic (which targets all variables) typically produced as good, or better, approximations. Indeed, it would be unusual if the focused heuristic consistently led to better marginals for untargeted variables.

Related Work

Several generalizations of belief propagation have been proposed with the increased interest in IBP, perhaps the most notable being the Generalized Belief Propagation (GBP) algorithms (Yedidia, Freeman, and Weiss 2005). Although the structure of a GBP approximation, typically specified via an auxiliary graph, is key to the quality of the approximation, relatively little attention has been spent in identifying good structures, much less those that are focused on a particular variable of interest. While some properties have been suggested as desirable for the structuring of GBP approximations (Yedidia, Freeman, and Weiss 2005; Welling, Minka, and Teh 2005), we are only aware of the region pursuit algorithm (for identifying region graphs) (Welling 2004), and the original edge recovery heuristic (which implicitly identifies a joingraph) (Choi and Darwiche 2006), for systematic and query-specific approaches to identifying structure. Neither, however, targets a variable of interest.

(Rosales and Jaakkola 2005) also considers a notion of focused approximations, in the sense that variable (bucket) elimination is a query-specific form of exact inference (Zhang and Poole 1996; Dechter 1996). Indeed, it is similar in spirit to mini-bucket elimination (Dechter and Rish 2003), where simplifications are performed during the process of variable (bucket) elimination. Our proposal for edge-correction is similar in spirit to other methods based on recovering (implicitly) substructure, including sequential fitting (Frey et al. 2000) (which, unlike edge-correction, is sensitive to the ordering of individual corrections), and expectation propagation (EP) (Minka 2001) (which, unlike edge-correction, could be considered a form of iterative correction). EP also characterizes IBP as a disconnected approximation, and is another distinguished class of GBP approximations (Welling, Minka, and Teh 2005).

While specific to ED-BP approximations, we expect that a focused edge recovery approach can also be applied to targeting approximations in other reasoning algorithms, particularly those that can be formulated as exact inference in simplified models. These include, as we have shown here, IBP and some of its generalizations (Yedidia, Freeman, and Weiss 2005), but also mean-field variational methods (Jordan et al. 1999; Wiegerinck 2000; Geiger, Meek, and Wexler 2006) and mini-bucket approximations (Dechter and Rish 2003; Choi, Chavira, and Darwiche 2007).

Conclusion

We have proposed here an approach to approximate inference that is based on focusing an approximation to a targeted variable of interest. We proposed a focused edge recovery approach, that begins with a polytree approximation, labels each deleted edge with a query-specific score, and recovers into it those edges with the highest score. The focused edge recovery scores are based on conditions that we identified that led to an exact odds-approximation for a targeted variable, in the case where a single edge is deleted. We further identified the odds-error in this simplified case, which led also to an edge-correction scheme that is cheaper than edge recovery for focused approximations. Experimentally, we found that both approaches to focused approximations are more effective than an unfocused approach.

Acknowledgments

This work has been partially supported by Air Force grant #FA9550-05-1-0075 and by NSF grant #IIS-0713166.

Proofs

Proof of Lemma 1 It suffices to show that

$$Pr(q \mid \mathbf{e}) \propto \sum_{u=\hat{u}} Pr'(q \mid u\hat{u}, \mathbf{e}')Pr'(u \mid \hat{u}, \mathbf{e}'). \quad (4)$$

First, note that if an edge $U \rightarrow \hat{U}$, representing an equivalence constraint, is recovered into $\mathcal{N}'$ (dropping only auxiliary variable $\hat{S}$) the probability of evidence is equivalent to that of the original network $\mathcal{N}$. Assuming this recovery:

$$Pr(\mathbf{e}) = \sum_{u=\hat{u}} Pr(u\hat{u}, \mathbf{e}) = \sum_{u=\hat{u}} \frac{\partial Pr(\mathbf{e})}{\partial \theta_{\hat{u}|u}} = \sum_{u=\hat{u}} \frac{\partial^2 Pr'(\mathbf{e}')}{\partial \theta_{\hat{u}} \partial \theta_{\hat{s}|u}}$$
$$= \sum_{u=\hat{u}} \frac{Pr'(u\hat{u}, \mathbf{e}')}{\theta_{\hat{u}}\theta_{\hat{s}|u}} \propto \sum_{u=\hat{u}} \frac{Pr'(u\hat{u}, \mathbf{e}')}{Pr'(\hat{u}, \mathbf{e}')}. \quad (5)$$

The last equality follows from the fact that:

$$\theta_{\hat{u}}\theta_{\hat{s}|u} \propto \theta_{\hat{u}} \frac{\partial Pr'(\mathbf{e}')}{\partial \theta_{\hat{u}}} = Pr'(\hat{u}, \mathbf{e}'),$$

using ED-BP fixed-point condition $\theta_{\hat{s}|u} = \partial Pr'(\mathbf{e}')/\partial\theta_{\hat{u}}$, which is equivalent to the one given in (Choi and Darwiche 2006). Conditioning on $Q = q$ as evidence leads to the desired result. The second equality of Lemma 1 follows from Equation 5 by noting that $Pr'(\hat{u}|\mathbf{e}') = Pr'(u|\mathbf{e}')$ in an ED-BP approximation; see Condition 1. □

Proof of Proposition 1 Assuming $(Q \perp U, \hat{U} \mid \mathbf{e}')$,

$$\begin{aligned} E &= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \frac{\sum_{u=\hat{u}} Pr'(\bar{q} \mid u\hat{u}, \mathbf{e}') Pr'(u \mid \hat{u}, \mathbf{e}')}{\sum_{u=\hat{u}} Pr'(q \mid u\hat{u}, \mathbf{e}') Pr'(u \mid \hat{u}, \mathbf{e}')} \\ &= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \frac{\sum_{u=\hat{u}} Pr'(\bar{q} \mid \mathbf{e}') Pr'(u \mid \hat{u}, \mathbf{e}')}{\sum_{u=\hat{u}} Pr'(q \mid \mathbf{e}') Pr'(u \mid \hat{u}, \mathbf{e}')} \\ &= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \frac{Pr'(\bar{q} \mid \mathbf{e}') \sum_{u=\hat{u}} Pr'(u \mid \hat{u}, \mathbf{e}')}{Pr'(q \mid \mathbf{e}') \sum_{u=\hat{u}} Pr'(u \mid \hat{u}, \mathbf{e}')} \end{aligned}$$

which is simply one. □

Proof of Proposition 2 Assuming $(Q, U \perp \hat{U} \mid \mathbf{e}')$,

$$\begin{aligned} E &= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \cdot \frac{\sum_{u=\hat{u}} Pr'(\bar{q} \mid u\hat{u}, \mathbf{e}') Pr'(u \mid \hat{u}, \mathbf{e}')}{\sum_{u=\hat{u}} Pr'(q \mid u\hat{u}, \mathbf{e}') Pr'(u \mid \hat{u}, \mathbf{e}')} \\ &= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \cdot \frac{\sum_{u=\hat{u}} Pr'(\bar{q}u \mid \hat{u}, \mathbf{e}')}{\sum_{u=\hat{u}} Pr'(qu \mid \hat{u}, \mathbf{e}')} \\ &= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \cdot \frac{\sum_{u} Pr'(\bar{q}u \mid \mathbf{e}')}{\sum_{u} Pr'(qu \mid \mathbf{e}')} \\ &= \frac{Pr'(q \mid \mathbf{e}')}{Pr'(\bar{q} \mid \mathbf{e}')} \cdot \frac{Pr'(\bar{q} \mid \mathbf{e}')}{Pr'(q \mid \mathbf{e}')} = 1. \end{aligned}$$

Assuming $(Q, \hat{U} \perp U \mid \mathbf{e}')$, and using the second equality of Lemma 1, we can find similarly that again $E = 1$. □

References

Aji, S. M., and McEliece, R. J. 2001. The generalized distributive law and free energy minimization. In *Proceedings of the 39th Allerton Conference on Communication, Control and Computing*, 672–681.

Braunstein, A.; Mézard, M.; and Zecchina, R. 2005. Survey propagation: An algorithm for satisfiability. *Random Struct. Algorithms* 27(2):201–226.

Choi, A., and Darwiche, A. 2006. An edge deletion semantics for belief propagation and its practical impact on approximation quality. In *Proceedings of the 21st National Conference on Artificial Intelligence (AAAI)*, 1107–1114.

Choi, A.; Chavira, M.; and Darwiche, A. 2007. Node splitting: A scheme for generating upper bounds in Bayesian networks. In *UAI*, 57–66.

Cover, T. M., and Thomas, J. A. 1991. *Elements of information theory*. Wiley-Interscience.

Dechter, R., and Rish, I. 2003. Mini-buckets: A general scheme for bounded inference. *J. ACM* 50(2):107–153.

Dechter, R.; Kask, K.; and Mateescu, R. 2002. Iterative join-graph propagation. In *UAI*, 128–136.

Dechter, R. 1996. Bucket elimination: A unifying framework for probabilistic inference. In *UAI*, 211–219.

Frey, B. J., and MacKay, D. J. C. 1997. A revolution: Belief propagation in graphs with cycles. In *Advances in Neural Information Processing Systems (NIPS)*, 479–485.

Frey, B. J.; Patrascu, R.; Jaakkola, T.; and Moran, J. 2000. Sequentially fitting "inclusive" trees for inference in noisy-or networks. In *Advances in Neural Information Processing Systems (NIPS)*, 493–499.

Geiger, D.; Meek, C.; and Wexler, Y. 2006. A variational inference procedure allowing internal structure for overlapping clusters and deterministic constraints. *J. Artif. Intell. Res. (JAIR)* 27:1–23.

Jordan, M. I.; Ghahramani, Z.; Jaakkola, T.; and Saul, L. K. 1999. An introduction to variational methods for graphical models. *Machine Learning* 37(2):183–233.

McEliece, R. J.; MacKay, D. J. C.; and Cheng, J.-F. 1998. Turbo decoding as an instance of Pearl's "belief propagation" algorithm. *IEEE Journal on Selected Areas in Communications* 16(2):140–152.

Minka, T. P. 2001. *A family of algorithms for approximate Bayesian inference*. Ph.D. Dissertation, MIT.

Murphy, K. P.; Weiss, Y.; and Jordan, M. I. 1999. Loopy belief propagation for approximate inference: An empirical study. In *UAI*, 467–475.

Pearl, J. 1988. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., San Mateo, California.

Rosales, R., and Jaakkola, T. 2005. Focused inference. In *AISTATS*.

Szeliski, R.; Zabih, R.; Scharstein, D.; Veksler, O.; Kolmogorov, V.; Agarwala, A.; Tappen, M. F.; and Rother, C. 2006. A comparative study of energy minimization methods for Markov random fields. In *ECCV (2)*, 16–29.

Wainwright, M. J.; Jaakkola, T.; and Willsky, A. S. 2003. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on Information Theory* 49(5):1120–1146.

Welling, M.; Minka, T. P.; and Teh, Y. W. 2005. Structured region graphs: morphing EP into GBP. In *UAI*, 609–614.

Welling, M. 2004. On the choice of regions for generalized belief propagation. In *UAI*, 585–592. Arlington, Virginia: AUAI Press.

Wiegerinck, W. 2000. Variational approximations between mean field theory and the junction tree algorithm. In *UAI*, 626–633.

Yedidia, J.; Freeman, W.; and Weiss, Y. 2005. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Transactions on Information Theory* 51(7):2282–2312.

Zhang, N. L., and Poole, D. 1996. Exploiting causal independence in bayesian network inference. *Journal of Artificial Intelligence Research* 5:301–328.