Approximations with Reweighted Generalized Belief Propagation

Wim Wiegerinck
SNN, Radboud University Nijmegen
Geert Grooteplein 21, 6525 EZ, Nijmegen, The Netherlands
W.Wiegerinck@science.ru.nl

Abstract

In (Wainwright et al., 2002) a new general class of upper bounds on the log partition function of arbitrary undirected graphical models has been developed. This bound is constructed by taking convex combinations of tractable distributions. The experimental results published so far concentrates on combinations of tree-structured distributions leading to a convexified Bethe free energy, which is minimized by the tree-reweighted belief propagation algorithm. One of the favorable properties of this class of approximations is that increasing the complexity of the approximation is guaranteed to increase the precision. The lack of this guarantee is notorious in standard generalized belief propagation. We increase the complexity of the approximating distributions by taking combinations of junction trees, leading to a convexified Kikuchi free energy, which is minimized by reweighted generalized belief propagation. Experimental results for Ising grids as well as for fully connected Ising models are presented illustrating advantages and disadvantages of the reweighting method in approximate inference.

1 INTRODUCTION

Probabilistic graphical models such as Bayesian networks and Markov random fields are powerful tools for learning and reasoning in domains with uncertainty. Unfortunately, exact inference is intractable in large, complex graphs. Therefore approximate inference methods are of great importance. An approximation method that recently received much attention is loopy belief propagation (BP) (Pearl, 1988). Although the algorithm is not guaranteed to converge, it often gives surprisingly accurate results (Murphy et al., 1999). In (Yedidia et al., 2001), it has been shown that fixed points of loopy BP are actually extrema of the Bethe free energy, which can be considered as a two-node approximation of the exact free energy of the system. By considering the Kikuchi free energy, which is an approximate free energy based on larger clusters of nodes, the more advanced generalized belief propagation (GBP) algorithm has been derived (Yedidia et al., 2001). This algorithm can be viewed as an interpolation between loopy BP and the junction tree algorithm (Lauritzen and Spiegelhalter, 1988; Jensen, 1996). The relation between (G)BP and the approximate Bethe/Kikuchi free energies motivated several researchers to design double-loop algorithms for explicit minimization of the Bethe/Kikuchi free energy with a guaranteed convergence to a (local) optimum (Teh and Welling, 2002; Yuille, 2002; Heskes et al., 2003).

Increasing the cluster-size in GBP often improves the accuracy of the approximation. Unfortunately, this is not guaranteed. A notorious counter example is the fully connected Ising model with pair and triplet approximations (Kappen and Wiegerinck, 2002). Even with moderate interaction strength, the triplet approximation is much worse than the pair approximation. A related problem is that the quality of different GBP approximations cannot be compared by comparing their Kikuchi free energy values. A lower Kikuchi free energy does not imply a better approximation.

A method that is closely related to (G)BP, but derived in a completely different way is the convexified free energy approximation (Wainwright et al., 2002; Wainwright et al., 2003; Wainwright and Jordan, 2003). This approximation is derived to provide an upper bound of the log partition function. The parameters of the exact model are represented as the average of parameters of tractable models. By the convexity of the log partition function, it is upper-bounded by the average of the log partition functions of the tractable models. The optimization of this upper bound then

give rise to a approximate Bethe/Kikuchi-like free energy, in which the cluster beliefs are parameters to be optimized. The advantage of this approach is that this approximate free energy is convex. So it is relatively easy to minimize, and it is guaranteed to have only a single minimum. Another advantage of this method is that a nested sequence of approximations with increasing complexity leads to tighter approximations of the free energy. So, for example, in a fully connected Ising model, the convexified approximation using triplets must be more accurate in free energy than with using pairs. One may expect (or hope) that the increase of precision in free energy is reflected in an increase of precision of other quantities, such as node marginals. Experimental results published so-far only involved approximations with trees, leading to a convexified Bethe free energy, which is optimized by tree-reweighted BP. The optimized pseudo-marginals in these approximations are pair-marginals.

The main contribution of this paper is an experimental study of convexified approximations with increasingly complex clusters. First, we review the convexified free energy approximation of an arbitrary discrete probability distributions (Wainwright et al., 2002; Wainwright et al., 2003), and the use of convex combinations of junction-trees leading to convexified Kikuchi free energy, as outlined earlier in (Wainwright and Jordan, 2003). We present reweighted GBP (i.e., RGBP) to minimize the convexified free energy. This algorithm is a straightforward generalization of the message-free GBP algorithm presented in (Heskes and Zoeter, 2003). We consider Ising grids and fully connected Ising models for which we construct nested convexified pair and cluster approximations of the free energy and of the cluster marginals. These approximations are experimentally compared with each other and with the corresponding standard Bethe/Kikuchi approximations optimized with convergent double loop algorithms (Heskes et al., 2003).

2 EXACT MODEL, PARTITION FUNCTION AND FREE ENERGY

We consider a distribution over discrete variables $\mathbf{x}$ with potential $\psi(\mathbf{x})$,

$$P(\mathbf{x}) = \frac{1}{Z} \exp(\psi(\mathbf{x})) \,. \tag{1}$$

The normalization constant,

$$Z(\psi) = \sum_{\mathbf{x}} \exp(\psi(\mathbf{x})) \tag{2}$$

is known as the partition function. For later reference, we also define the variational free energy of the system,

$$F(\hat{P}) = -\sum_{\mathbf{x}} \hat{P}(\mathbf{x})\psi(\mathbf{x}) + \sum_{\mathbf{x}} \hat{P}(\mathbf{x}) \log \hat{P}(\mathbf{x}), \tag{3}$$

Minimizing the free energy with respect to $\hat{P}$ returns the distribution P in (1). The value of the free energy at its minimum is

$$F(P) = -\log Z \,. \tag{4}$$

2.1 JUNCTION TREES

Now we consider distributions over sets of nodes, $\mathbf{x} = (x_1, \ldots, x_n)$, and potentials that factorize into overlapping cluster potentials,

$$\psi(\mathbf{x}) = \sum_{\alpha} \psi_\alpha(\mathbf{x}_\alpha) \tag{5}$$

with $\alpha \subset \{1, \ldots, n\}$, and $\mathbf{x}_\alpha$ the state vector restricted to the the variables in α. The clusters and potentials are not uniquely defined. For instance, clusters can be merged into bigger clusters $\alpha'' = \alpha \cup \alpha'$ with $\psi_{\alpha''} = \psi_\alpha + \psi_{\alpha'}$. For convenience, we assume that clusters α are not contained in each other. This can be achieved by merging subclusters with their supersets.

In general the distribution $P(\mathbf{x})$ will be intractable. An exception is formed by models in which the (possibly merged) clusters can be organized into a junction tree with small maximal cluster size.

A junction tree (Lauritzen and Spiegelhalter, 1988; Jensen, 1996) is a hyper-tree of clusters $\alpha \in C$ (or actually: a forest of hyper-trees), which has the properties that for any pair of clusters α and α' with nonempty overlap $\alpha \cap \alpha'$: (1) There is a path in the cluster tree that connects α and α' and (2) All the clusters κ in the path connecting α and α' should contain their intersection: $\alpha \cap \alpha' \subset \kappa$ (running intersection property). The links between adjacent clusters in the hyper-tree are labeled with separators. They consists of the intersections $\delta = \alpha \cap \alpha'$ of the adjacent clusters. The number of times that a subcluster δ appears in the hyper-tree is n_δ.

A probabilistic model in which the cluster set $C = \{\alpha\}$ can be organized as a junction tree, can be factorized into a product of probabilities on the cliques C and separators S,

$$P(\mathbf{x}) = \frac{\prod_{\alpha \in \mathcal{C}} P(\mathbf{x}_\alpha)}{\prod_{\delta \in \mathcal{S}} P(\mathbf{x}_\delta)^{n_\delta}} \equiv \prod_{\gamma \in \Gamma} P(\mathbf{x}_\gamma)^{k_\gamma} \tag{6}$$

where we defined $\Gamma = C \cup S$, and the counting numbers $k_\gamma = 1$ for $\gamma \in C$ and $k_\gamma = -n_\gamma$ for $\gamma \in S$. The free

energy can be expressed as

$$F(P) = -\sum_{\alpha\in\mathcal{C}}\sum_{\mathbf{x}_\alpha}\psi_\alpha(\mathbf{x}_\alpha)P(\mathbf{x}_\alpha) + \sum_{\gamma\in\Gamma}k_\gamma\sum_{\mathbf{x}_\gamma}P(\mathbf{x}_\gamma)\log P(\mathbf{x}_\gamma) \quad (7)$$

By defining $P_\gamma(\mathbf{x}_\gamma) \equiv P(\mathbf{x}_\gamma)$, $\gamma \in \Gamma$ the minimization with respect to P can be conveniently be reformulated as a minimization with respect to the independent cluster marginals $\{P_\gamma(\mathbf{x}_\gamma)\}$ under constraint that they are consistent on their overlaps $\gamma \cap \gamma'$.

3 AN UPPER BOUND OF THE PARTITION FUNCTION

From now on, we assume that $P(\mathbf{x})$ is intractable. In this section the goal is to upper-bound $\log Z(\psi)$.

If we have a set of (tractable) distributions $P^{(\mathcal{T})}(\mathbf{x})$, with potentials $\phi^{(\mathcal{T})}(\mathbf{x})$, and a weight $\mu(\mathcal{T}) \geq 0$ for each of of the distributions, with $\sum_{\mathcal{T}}\mu(\mathcal{T}) = 1$, such that the original potential ψ is the weighted sum of potentials $\phi^{(\mathcal{T})}$

$$\sum_{\mathcal{T}}\mu(\mathcal{T})\phi^{(\mathcal{T})}(\mathbf{x}) = \psi(\mathbf{x}) \quad (8)$$

then the log partition function $\log Z(\psi)$ is upper bounded by

$$\log Z(\psi) = \log Z(\sum_{\mathcal{T}}\mu(\mathcal{T})\phi^{(\mathcal{T})}) \quad (9)$$

$$\leq \sum_{\mathcal{T}}\mu(\mathcal{T})\log Z(\phi^{(\mathcal{T})}) \quad (10)$$

This results from the convexity of $\log Z(\psi)$, which follows from the fact that the matrix of second derivatives,

$$\frac{\partial^2\log Z(\psi)}{\partial\psi(\mathbf{x})\partial\psi(\mathbf{y})} = P(\mathbf{x},\mathbf{y}) - P(\mathbf{x})P(\mathbf{y}) = \sum_{\mathbf{z}}P(\mathbf{z})(\delta_{\mathbf{zx}} - P(\mathbf{x}))(\delta_{\mathbf{zy}} - P(\mathbf{y})) \quad (11)$$

is positive semi-definite .

3.1 A CONVEX COMBINATION OF JUNCTION TREES

Now we take $P^{(\mathcal{T})}$ to be junction trees with cluster sets $C(\mathcal{T})$ (such that the maximal cluster size is small enough) and cluster potentials $\phi_\beta^{(\mathcal{T})}$. According to (8) these potentials should satisfy

$$\sum_{\mathcal{T}}\mu(\mathcal{T})\sum_{\beta\in C(\mathcal{T})}\phi_\beta^{(\mathcal{T})}(\mathbf{x}_\beta) = \psi(\mathbf{x}) \quad (12)$$

For convenience, we assume that each cluster α is a member of at least one cluster set $C(\mathcal{T})$, and that the clusters α are not smaller than the clusters β of the junction trees (i.e., $\alpha \subseteq \beta \Rightarrow \alpha = \beta$). This can be achieved by merging of clusters α.

To optimize the upper bound of the log partition function with respect to the cluster potentials $\{\phi_\beta^{(\mathcal{T})}\}$ for fixed $\{\mu(\mathcal{T})\}$, we construct the Lagrangian

$$\mathcal{L}(\{\phi^{(\mathcal{T})}\}, Q) = \sum_{\mathcal{T}}\mu(\mathcal{T})\log Z(\phi^{(\mathcal{T})}) + \sum_{\mathbf{x}}Q(\mathbf{x})\left[\psi(\mathbf{x}) - \sum_{\mathcal{T}}\mu(\mathcal{T})\sum_{\beta\in C(\mathcal{T})}\phi_\beta^{(\mathcal{T})}(\mathbf{x}_\beta)\right] \quad (13)$$

where we introduced Lagrange multipliers $Q(\mathbf{x})$ for the constraints (12). Differentiation with respect to $\phi_\beta^{(\mathcal{T})}(\mathbf{x}_\beta)$, using

$$\frac{\partial\log Z(\phi^{(\mathcal{T})})}{\partial\phi_\beta^{(\mathcal{T})}(\mathbf{x}_\beta)} = P^{(\mathcal{T})}(\mathbf{x}_\beta) \quad (14)$$

and setting to zero yields for $\beta \in C(\mathcal{T})$,

$$P^{(\mathcal{T})}(\mathbf{x}_\beta) = \sum_{\mathbf{x}\setminus\mathbf{x}_\beta}Q(\mathbf{x}) \equiv Q(\mathbf{x}_\beta) \quad (15)$$

Apparently, $Q(\mathbf{x})$ is a 'pseudo-probability' with 'pseudo-marginals' on the clusters of the junction trees that are equal to the cluster marginals of these junction trees. This implies that in the optimal $\{\phi\}$ all the cluster probabilities of the different junction trees are consistent on their overlaps.

$$P^{(\mathcal{T})}(\mathbf{x}_{\beta\cap\beta'}) = P^{(\mathcal{T}')}(\mathbf{x}_{\beta\cap\beta'}) = Q(\mathbf{x}_{\beta\cap\beta'}) \quad (16)$$

for any pair of clusters $\beta \in C(\mathcal{T})$ and $\beta' \in C(\mathcal{T}')$. We reformulate the $\log Z(\phi^{(\mathcal{T})})$ as $\min F(\{P_\gamma^{(\mathcal{T})}\})$. Using the fact that the optimal $\{P_\gamma^{(\mathcal{T})}\}$ are equal to the pseudo-marginals $Q(\mathbf{x}_\gamma)$, and using (12) we can add up of all the free energies

$$\sum_{\mathcal{T}}\mu(\mathcal{T})F^{(\mathcal{T})}(\{Q(\mathbf{x}_\gamma)\}) = \left[-\sum_{\alpha}\psi_\alpha(\mathbf{x}_\alpha)Q(\mathbf{x}_\alpha) + \sum_{\mathcal{T}}\mu(\mathcal{T})\sum_{\gamma\in\Gamma(\mathcal{T})}k_\gamma^{(\mathcal{T})}Q(\mathbf{x}_\gamma)\log Q(\mathbf{x}_\gamma)\right] \quad (17)$$

Introducing the counting numbers

$$c_\gamma = \sum_{\mathcal{T}}\mu(\mathcal{T})k_\gamma^{(\mathcal{T})} \quad (18)$$

and writing $Q(X_\gamma) = Q_\gamma(X_\gamma)$ as independent pseudo-marginals we obtain the convexified Kikuchi free en-

Algorithm 1 Message-free RGBP.

initialize $Q_\alpha(\mathbf{x}_\alpha)$ and $Q_\delta(\mathbf{x}_\delta)$ as in (21)
repeat
  for all inner clusters δ **do**
    update $Q_\delta(x_\delta) \leftarrow Q_\delta^{\text{new}}(x_\delta)$ as in (23)
    for all outer clusters $\alpha \supset \delta$ **do**
      update $Q_\alpha(\mathbf{x}_\alpha) \leftarrow Q_\alpha^{\text{new}}(\mathbf{x}_\alpha)$ as in (24)
    end for
  end for
until convergence
return $Q_\alpha(\mathbf{x}_\alpha)$ and $Q_\delta(x_\delta)$

ergy

$$F_{\text{approx}} = -\sum_\alpha \sum_{\mathbf{x}_\alpha} \psi_\alpha(\mathbf{x}_\alpha) Q_\alpha(\mathbf{x}_\alpha) + \sum_\gamma \sum_{\mathbf{x}_\gamma} c_\gamma Q_\gamma(\mathbf{x}_\gamma) \log Q_\gamma(\mathbf{x}_\gamma) \quad (19)$$

which should be minimized under the constraint that the Q_γ's are consistent on their overlaps. The convexity follows by construction from the fact that F_{approx} is a convex combination of exact free energies which are convex.

3.2 RELATION WITH THE KIKUCHI FREE ENERGY

The Kikuchi free energy (Yedidia et al., 2001; Heskes et al., 2003) has the same functional form as in (19). The difference is in the counting numbers. In the Kikuchi free energy, the starting point is a set of outer clusters, typically coinciding with the clusters α of the original models (possibly after merging). The inner clusters δ are formed by taking all intersections of the outer clusters. The counting numbers follow from the recursive Moebius formula $c_\delta = 1 - \sum_{\gamma \supset \delta} c_\gamma$, with $c_\alpha = 1$ for all outer clusters. The standard Kikuchi free energy need not to be convex.

3.3 REWEIGHTED GENERALIZED BELIEF PROPAGATION

Due to the similarity between the Kikuchi free energy and its convexified version, the GBP algorithm for minimizing the Kikuchi free energy is easily generalized to reweighted generalized belief propagation (RGBP) for the convexified Kikuchi free energy. Here we describe a message-free form of RGBP. It is based on the message-free GBP presented in (Heskes and Zoeter, 2003)). The only difference is that we now allow $c_\alpha \neq 1$ for the outer clusters.

We divide the clusters $\gamma \in \cup_{\mathcal{T}} \Gamma(\mathcal{T})$ into 'outer clusters' coinciding with clusters α in the original model, and 'inner clusters' δ which are subclusters of α. The approximate free energy (19) can (loosely) be interpreted as the free-energy of a "junction tree-like pseudo-probability distribution",

$$\tilde{Q}(\mathbf{x}) = \frac{\prod_\alpha Q_\alpha(\mathbf{x}_\alpha)^{c_\alpha}}{\prod_\delta Q_\delta(x_\delta)^{-c_\delta}} \quad (20)$$

We start by initializing the inner and outer cluster marginals

$$Q_\alpha(\mathbf{x}_\alpha) \propto \exp(\frac{\psi_\alpha(\mathbf{x}_\alpha)}{c_\alpha}) \quad \text{and} \quad Q_\delta(\mathbf{x}_\delta) \propto 1\,. \quad (21)$$

so that $tQ(\mathbf{x})$ is proportional to the target distribution,

$$\tilde{Q}(\mathbf{x}) \propto \exp(\sum_\alpha \psi(\mathbf{x}_\alpha)) \quad (22)$$

Next we repeatedly update the inner and outer cluster pseudo-marginals that re-arrange information in the cluster marginals to make them mutually consistent while keeping (22) satisfied. The update rule for the inner clusters δ is

$$Q_\delta^{\text{new}}(\mathbf{x}_\delta) \propto Q_\delta(\mathbf{x}_\delta)^{\frac{c_\delta}{m_\delta + c_\delta}} \bar{Q}_\delta(\mathbf{x}_\delta)^{\frac{m_\delta}{m_\delta + c_\delta}} \quad (23)$$

with

$$m_\delta = \sum_{\alpha \supset \delta} c_\alpha \quad \text{and} \quad \bar{Q}_\delta(\mathbf{x}_\delta) \propto \left[\prod_{\alpha \supset \delta} Q_\alpha(\mathbf{x}_\delta)^{c_\alpha} \right]^{\frac{1}{m_\delta}}$$

The update rule for the outer clusters α is

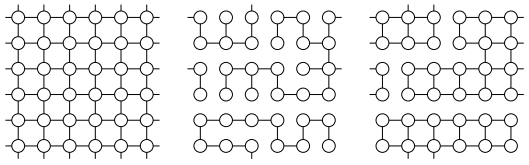
$$Q_\alpha^{\text{new}}(\mathbf{x}_\alpha) \propto Q_\alpha(\mathbf{x}_\alpha) \frac{Q_\delta^{\text{new}}(\mathbf{x}_\delta)}{Q_\alpha(\mathbf{x}_\delta)}\,. \quad (24)$$

The final RGBP algorithm is summarized in Algorithm 1.

In analogy with (Heskes and Zoeter, 2003) it can be shown that fixed points of Algorithm 1 is an extremum (and hence a minimum) of the convexified Kikuchi free energy. In our simulations, the algorithm converged without damping. To our knowledge, however, this is not guaranteed. If needed, a damping term can be introduced similar to the one in (Heskes and Zoeter, 2003).

4 COUNTING NUMBERS IN SOME REGULAR GRAPHS

For a graph with a random structure, the construction of the Kikuchi free energy is straightforward, given the choice of the clusters. The reason is that the counting numbers follow straightforwardly from the Moebius formula. In the convexified case, the computation of the counting numbers is generally much more



(a) (b) (c)

Figure 1: Left: full grid $T_{6\times 6}$ with periodic boundary conditions. Middle: spanning tree. Right: covering junction tree.

difficult (Wainwright et al., 2002). Since we are interested in the performance of nested convexified cluster approximations, without having a general algorithm for computing counting numbers, we restrict ourself to approximations of models with regular graphs, namely an Ising grid with periodic boundary conditions and a fully connected Ising model.

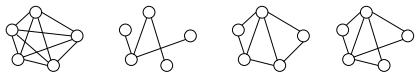
In (Wainwright et al., 2002), a convex combination of all spanning trees in the graph is taken, leading to a convexified Bethe free energy $F_{\rm cB}$. In this approximation the outer clusters are the pairs of connected nodes in the Ising model. In (Wainwright et al., 2002), the distribution $\mu(\mathcal{T})$ is optimized as well. Here we fix $\mu(\mathcal{T})$ to be uniform. Under this simplifying condition, and making use of the symmetries in the models that we consider, we can construct convexified Kikuchi free energies $F_{\rm cK}$ that will provide strictly tighter bounds

$$\min F \geq \min F_{\rm cK} \geq \min F_{\rm cB}. \qquad (25)$$

4.1 ISING GRID

The first Ising model that we consider is a grid of $2n \times 2n$ nodes with periodic boundary conditions (torus). Each node is connected to four neighbors. We denote this graph as $T_{2n\times 2n}$ (see Fig. 1(a)).

We consider pair approximations and 2×2 cluster approximations. To write down $F_{\rm cK}$, we have to compute the counting numbers c_γ. For this we draw a random junction tree with clusters and subclusters $\Gamma(\mathcal{T})$. For each type of cluster, namely singleton (type $\gamma(1)$), pair (type $\gamma(2)$), and 2×2 cluster (type $\gamma(4)$), we compute the sum of the counting numbers $k(i) = \sum_{\gamma \text{ is type } \gamma(i)} k_\gamma$ and divide the result by the total number $n(i)$ of clusters of type $\gamma(i)$ in the original graph. In the $T_{2n\times 2n}$ graph, these are $n(1) = 4n^2$, $n(2) = 8n^2$, and $n(4) = 4n^2$. The resulting counting number for clusters of type i is $c(i) = k(i)/n(i)$, which can be substituted in (19) for the c_γ's with γ of type i.



(a) (b) (c) (d)

Figure 2: Left: fully connected K_5. Middle: spanning tree. Right: two covering junction trees.

If we consider a spanning tree (see Fig. 1(b)), we find $k(2) = 4n^2 - 1$ and $k(1) = -(4n^2 - 2)$. As a result we find for $F_{\rm cB}$ the counting numbers

$$c(2) = \frac{4n^2-1}{8n^2} \quad \text{and} \quad c(1) = -\frac{2n^2-1}{2n^2} \qquad (26)$$

For our cluster approach we construct a set of junction trees as follows: take n non-overlapping horizontal strips of $2n-1$ type-4 clusters and connect these vertically by $n-1$ additional type-4 clusters (see Fig. 1(c)). Shifting and rotating this procedure leads to a homogeneous set. We find $k(4) = 2n^2-1$, $k(2) = -(2n^2-2)$, and $k(1) = 0$. So, with this choice of junction trees the counting numbers for $F_{\rm cK}$ are

$$c(4) = \frac{2n^2-1}{4n^2} \quad \text{and} \quad c(2) = -\frac{n^2-1}{4n^2} \qquad (27)$$

If we now go back to the spanning trees and restrict the spanning trees to those that are contained in the junction trees as constructed above (which is the case in Fig. 1(b)), we find that the resulting counting numbers are the same as in (26). From this we conclude that the approximations are nested and (25) holds.

4.2 FULLY CONNECTED ISING MODEL

We denote the fully connected Ising model with n nodes as K_n (see Fig. 2(a)). We consider pair and triplet approximations. The counting numbers are needed for singletons (type $\gamma(1)$), pairs (type $\gamma(2)$), and triplets (type $\gamma(3)$). In the graph K_n, $n(1) = n$, $n(2) = n(n-1)/2$, and $n(3) = n(n-1)(n-2)/6$.

If we consider a spanning tree (see Fig. 2(b)), we find $k(2) = n - 1$ and $k(1) = -(n-2)$. The resulting counting numbers for $F_{\rm cB}$ are

$$c(2) = \frac{2}{n} \quad \text{and} \quad c(1) = -\frac{n-2}{n} \qquad (28)$$

For the triplet approximation we consider junction trees that have clusters of three nodes, and separators of two nodes (see Fig. 2(c,d)). For such type of junction trees we find $k(3) = n - 2$, $k(2) = -(n-3)$,

and $k(1) = 0$. So we find for F_{cK} the counting numbers

$$c(3) = \frac{6}{n(n-1)} \quad \text{and} \quad c(2) = -\frac{2(n-3)}{n(n-1)} \tag{29}$$

Each spanning tree is contained in a junction tree. Therefore the approximations are again nested and (25) holds.

5 SIMULATIONS

We apply (R)(G)BP to the Ising models described in the previous section. The prefix "G" (i.e, "generalized"), implies the use of outer-clusters larger than two, namely the 2×2 for the grids and the triplets for the fully connected models. The prefix "R" (i.e., "reweighted"), implies the use of counting numbers c_γ from the convexified free energy as described in the previous section rather than the standard ones obtained by the Moebius relation. For R(G)BP we used the RGBP algorithm as described earlier in this paper. For (G)BP we used the double-loop algorithm described in (Heskes et al., 2003).

5.1 EXPERIMENTAL SET-UP

We considered Ising grids $T_{6\times6}$ and $T_{8\times8}$ as well as the fully connected models K_9 and K_{12}. The variables in the models are binary $x_i = \pm 1$. We choose $\psi(\mathbf{x}_{ij}) = w_{ij}x_ix_j + \theta_i/n_ix_i + \theta_j/n_jx_j)$ with n_i the number of neighbors of node i. The external fields are generated according $\theta_i \sim \mathcal{N}(0, 0.01)$ (in both type of graphs) and couplings according to $J_{ij} \sim \mathcal{N}(0, \frac{\beta}{2})$ for the torus and $J_{ij} \sim \mathcal{N}(0, \frac{\beta}{\sqrt{N}})$, with N the number of nodes in the graph, for the fully connected model. We consider eight scalings $\beta = [0.2, 0.5, 0.75, 1, 1.5, 2, 5, 10]$. With each scaling, we generated 5 models. For each model realization we ran simulations with RBP, RGBP, BP and GBP. In all runs, we computed the exact minimal free energy $F_{\text{exact}} = -\log Z$, the exact edge probabilities $P(\mathbf{x}_{ij})$, the approximating free energy F_{approx} according to (19), and the approximating pseudo-marginals on the edges $Q(\mathbf{x}_{ij})$.

5.2 RESULTS

In figure 3 we plotted for the four models the maximum absolute deviation (MAD) of edge probabilities

$$\text{MAD} = \max_{(i,j)\in E} \max_{x_{ij}} |Q(\mathbf{x}_{ij}) - P(\mathbf{x}_{ij})|, \tag{30}$$

the relative error in free energy

$$\epsilon = \frac{F_{\text{approx}} - F_{\text{exact}}}{F_{\text{exact}}}, \tag{31}$$

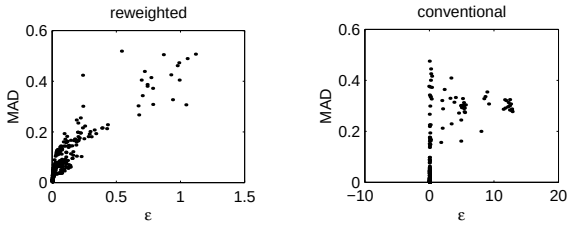


Figure 4: Scatter plot of performance in MAD as a function of relative error ϵ for reweighted approximations (RBP and RGBP, (left)) and conventional approximations (BP and GBP, (right)).

and the absolute values of these relative errors $|\epsilon|$. Plotted are the means and standard deviations of these quantities as function of interaction strength β.

On the grids $T_{6\times6}$ and $T_{8\times8}$, in both the conventional approximations (BP, GBP) as in the reweighted approximations (RBP,RGBP), increasing clustersize consistently improves the result: GBP outperforms BP, and RGBP outperforms RBP. In addition, we see that the conventional approximations (BP, GBP) outperform their reweighted counterparts (RBP, RGBP) both in the MAD and ϵ.

With the fully connected models K_9 and K_{12}, the results show a completely different picture for the conventional approximations: GBP performs only well for small β. If β is of order one, or larger, the GBP approximation collapses and BP outperforms GBP. (The large error bars in GBP for $\beta \approx 1$ are due to the fact that for some model realizations GBP performed very well, and for others very bad). On the other hand, adding complexity in the reweighted approximation always improves the result: RGBP is always better than RBP. The improvement is remarkably constant, almost independent of β, and independent of whether the approximation is good or bad.

To investigate the relation between ϵ and MAD of the two different classes of approximations, we pooled all the simulation results (i.e. of all the runs for the four models with all the settings of β) into two groups: one for the reweighted approximations (RBP and RGBP) and one for the conventional approximations (BP and GBP). In figure 4 we made scatter plots of each pool by plotting the MAD versus ϵ for each simulation run. In these plots we see that the relation between ϵ and MAD is much stronger in the the reweighted approximations than in the conventional approximations.

Furthermore, we investigated the effect of increasing the cluster size in each of the approximation classes. For each model realization, we compared the errors ϵ and MAD for the pair approximations BP and RBP

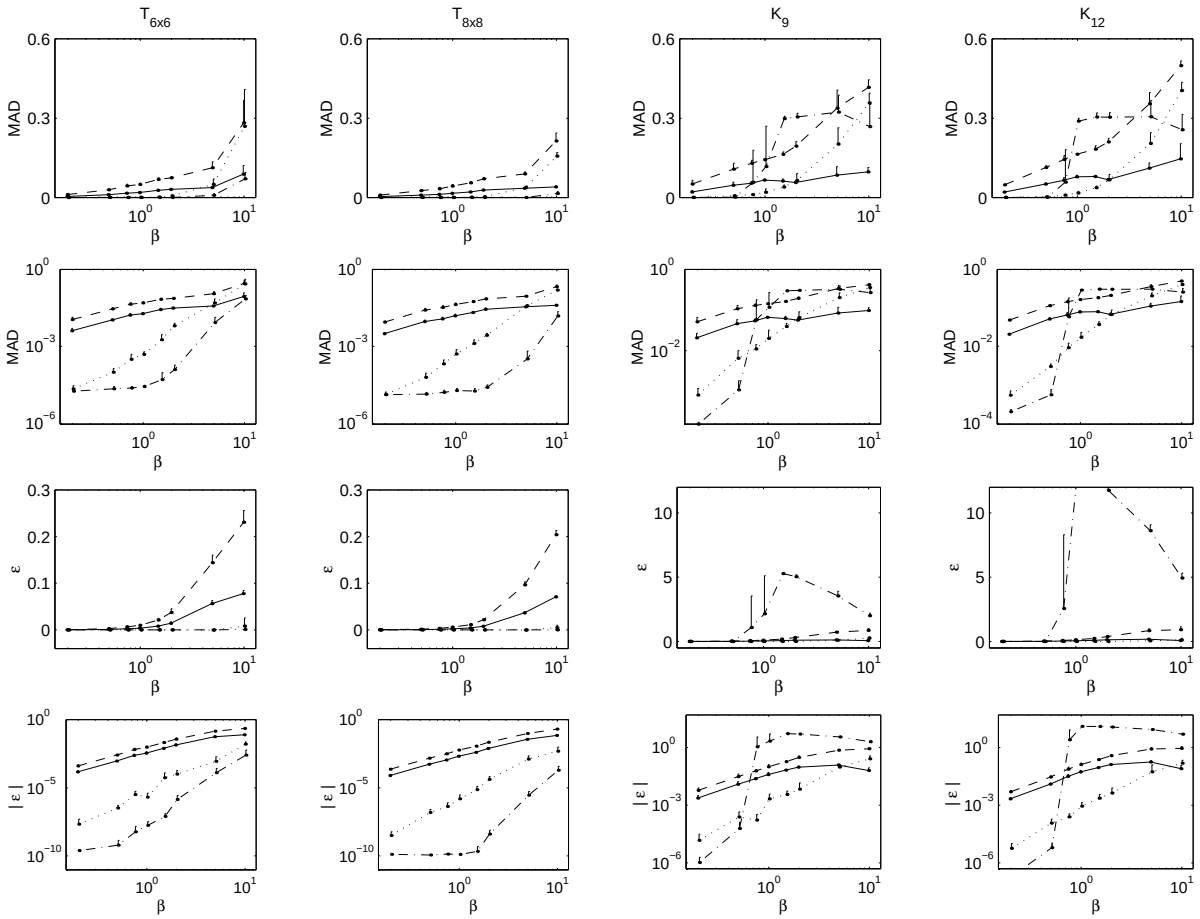


Figure 3: Columns from left to right: (a) Results for $T_{6\times6}$, (2 D grid with $6 \times 6 = 36$ nodes, periodic boundary conditions); (b) Results for $T_{8\times8}$. (c) Results for K_9 (fully connected model with 9 nodes). (d) Results for K_{12}. First row: Maximal absolute deviation (MAD) for edge probabilities $p(x_{ij})$ Second row:MAD, in log scale. Third row: Relative error in free energy $\epsilon = (F_{\text{approx}} - F_{\text{exact}})/F_{\text{exact}}$ in linear scale. Fourth row: absolute values of relative errors $|\epsilon|$ in log scale. Results are plotted for RGBP (full line), RBP (dashed), GBP (dash-dotted) and BP (dotted). Only upper parts of errorbars are drawn for visual clarity. (Large errorbars in K_9 and K_{12} are in GBP results).

with the cluster approximation in the same approximating class, GBP and RGBP respectively. The scatter plots in figure 5 clearly show that in the reweighted approximation class the performance with larger clusters improves in all model realizations. For the approximation in F, this improvement is theoretically guaranteed. The improvement in both ϵ and MAD is surprisingly constant. In the conventional approximation class, the improvement clearly depends on the regime. In the easy regime (small ϵ, small MAD), larger clusters improves the results. In the hard regime (large ϵ, large MAD), increasing cluster size may actually do harm. Furthermore, there are exceptions: sometimes BP improves upon GBP even in the easy regime.

6 DISCUSSION

Finding accurate approximations of graphical models such as Bayesian networks is crucial if their application to large scale problems is to be realized. Generalized belief propagation (GBP) is nowadays considered as one of the most powerful approximation methods. The method is flexible in the sense that there is a tradeoff in computational complexity and cluster-size. Unfortunately, increasing the cluster size does not guarantee to improve accuracy, and sometimes even deteriorate results. For this reason we are interested in alternative approximate methods that *do* provide a guarantee of improvement (at least in the free energy). Besides the convexified free energy approach, which is studied in this paper, another method that provide this guar-

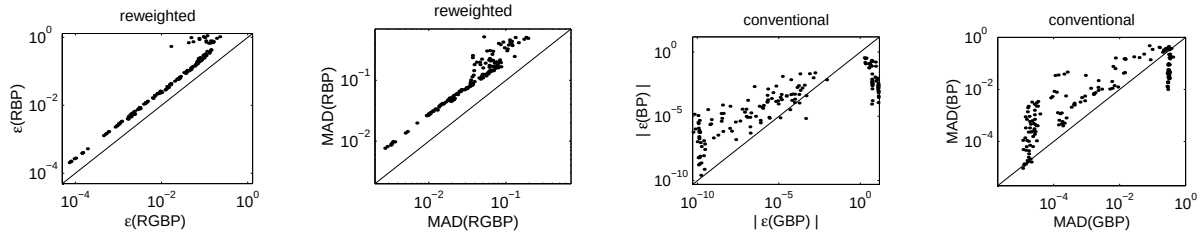


Figure 5: Performances (ϵ and MAD) of pair-approximation ((R)BP) versus performance of cluster-approximations (R)GBP.

antee is the structural mean field (SMF) approximation (Wiegerinck, 2000). The convexified free energy approach, however, has several appealing advantages. Unlike SMF theory, all the edges in the target distribution need (by construction) to be covered. This not only circumvents the problem in SMF of which edges to keep and which to delete, but it also suggests more powerful approximation. The additional fact that no results about its performance with cluster size larger than two have been published (as far as we know) motivated us to further investigate this method.

The experimental results with RGBP (reweighted generalized belief propagation -the counterpart of RBP for the convexified approach) suggest that in 'easy' problems where (G)BP performs well, the reweighted approximations do not provide a competing alternative. However, in 'hard' problems, it might be worthwhile to consider RGBP with larger clusters as an alternative. The method seems to be more robust in such problems.

There is, however, an important open problem, not addressed in this paper, but mentioned earlier in (Wainwright et al., 2002), which is: how to find – or even better: optimize the counting numbers in RGBP. In this paper, we computed them (suboptimally - since μ was taken constant) by hand, which was possible thanks to the symmetries in the model. An automatic procedure, however, is crucial if the RGBP is to be applied in real world problems with graphical models of arbitrary structure

Acknowledgments

This research is supported by the Dutch Technology Foundation STW. Thanks to Kees Albers for sharing his double-loop software.

References

Heskes, T., Albers, K., and Kappen, H. J. (2003). Approximate inference and constraint optimisation. In *UAI 19*, pages 313–320.

Heskes, T. and Zoeter, O. (2003). Generalized belief propagation for approximate inference in hybrid Bayesian networks. In *AISTATS 9*.

Jensen, F. (1996). *An introduction to Bayesian networks.* UCL Press.

Kappen, H. J. and Wiegerinck, W. (2002). Novel iteration schemes for the cluster variation method. In *NIPS 14*, pages 415–420.

Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilties on graphical structures and their application to expert systems (with discussion). *J. Royal Statistical Society Series B*, 50:157–224.

Murphy, K. P., Weiss, Y., and Jordan, M. I. (1999). Loopy belief propagation for approximate inference: An empirical study. In *UAI 15*, pages 467–475.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc.

Teh, Y. and Welling, M. (2002). The unified propagation and scaling algorithm. In *NIPS 14*, pages 953–960.

Wainwright, M. J., Jaakkola, T., and Willsky, A. S. (2002). A new class of upper bounds on the log partition function. In *UAI 18*, pages 536–543.

Wainwright, M. J., Jaakkola, T., and Willsky, A. S. (2003). Tree-reweighted belief propagation and approximate ML estimation by pseudo-moment matching. In *AISTATS 9*.

Wainwright, M. J. and Jordan, M. I. (2003). Graphical models, exponential families, and variational inference. Technical Report 649, UC Berkeley, Dept. of Statistics.

Wiegerinck, W. (2000). Variational approximations between mean field theory and the junction tree algorithm. In *UAI 16*, pages 626–633.

Yedidia, J. S., Freeman, W. T., and Weiss, Y. (2001). Generalized belief propagation. In *NIPS 13*, pages 689–695.

Yuille, A. L. (2002). CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural computation*, 14:1691–1722.