Bounds on marginal probability distributions

Joris Mooij
MPI for Biological Cybernetics
Tübingen, Germany
`joris.mooij@tuebingen.mpg.de`

Bert Kappen
Department of Biophysics
Radboud University Nijmegen, the Netherlands
`b.kappen@science.ru.nl`

Abstract

We propose a novel bound on single-variable marginal probability distributions in factor graphs with discrete variables. The bound is obtained by propagating local bounds (convex sets of probability distributions) over a subtree of the factor graph, rooted in the variable of interest. By construction, the method not only bounds the exact marginal probability distribution of a variable, but also its approximate Belief Propagation marginal ("belief"). Thus, apart from providing a practical means to calculate bounds on marginals, our contribution also lies in providing a better understanding of the error made by Belief Propagation. We show that our bound outperforms the state-of-the-art on some inference problems arising in medical diagnosis.

1 Introduction

Graphical models are used in many different fields. A fundamental problem in the application of graphical models is that exact inference is NP-hard [1]. In recent years, much research has focused on approximate inference techniques, such as sampling methods and deterministic approximation methods, e.g., Belief Propagation (BP) [2]. Although the approximations obtained by these methods can be very accurate, there are only few useful guarantees on the error of the approximation, and often it is not known (without comparing with the intractable exact solution) how accurate an approximate result is. Thus it is desirable to calculate, in addition to the approximate results, tight bounds on the approximation error.

There exist various methods to bound the BP error [3, 4, 5, 6], which can be used, in conjunction with the results of BP, to calculate bounds on the exact marginals. Furthermore, upper bounds on the partition sum, e.g., [7, 8], can be combined with lower bounds on the partition sum, such as the well-known mean field bound or higher-order lower bounds [9], to obtain bounds on marginals. Finally, a method called Bound Propagation [10] directly calculates bounds on marginals. However, most of these bounds (with the exception of [3, 10]) have only been formulated for the special case of pairwise interactions, which limits their applicability, excluding for example the interesting class of Bayesian networks.

In this contribution we describe a novel bound on exact single-variable marginals in factor graphs which is not limited to pairwise interactions. The original motivation for this work was to better understand and quantify the BP error. This has led to bounds which are at the same time bounds for the exact single-variable marginals as well as for the BP beliefs. A particularly nice feature of our bounds is that their computational cost is relatively low, provided that the number of possible values of each variable in the factor graph is small. On the other hand, the computational complexity is exponential in the number of possible values of the variables, which limits application to factor graphs in which each variable has a low number of possible values. On these factor graphs however, our bound can significantly outperform existing methods, either in terms of accuracy or in terms of computation time (or both). We illustrate this on two toy problems and on real-world problems arising in medical diagnosis.

The basic idea underlying our method is that we recursively propagate bounds over a particular *subtree* of the factor graph. The propagation rules are similar to those of Belief Propagation; however, instead of propagating messages, we propagate *convex sets* of messages. This can be done in such a way that the final "beliefs" at the root node of the subtree are convex sets which contain the exact marginal of the root node (and, by construction, also its BP belief). In the next section, we describe our method in more detail. Due to space constraints, we have omitted the proofs and other technical details; these are provided in a technical report [11], which also reports additional experimental results and presents an extension that uses self-avoiding-walk trees instead of subtrees (inspired by [6]).

2 Theory

2.1 Preliminaries

Factorizing probability distributions. Let $\mathcal{V} := \{1, \ldots, N\}$ and consider N discrete random variables $(x_i)_{i \in \mathcal{V}}$. Each variable x_i takes values in a discrete domain $\mathcal{X}_i$. We will use the following multi-index notation: let $A = \{i_1, i_2, \ldots, i_m\} \subseteq \mathcal{V}$ with $i_1 < i_2 < \ldots i_m$; we write $\mathcal{X}_A := \mathcal{X}_{i_1} \times \mathcal{X}_{i_2} \times \cdots \times \mathcal{X}_{i_m}$ and for any family $(Y_i)_{i \in B}$ with $A \subseteq B \subseteq \mathcal{V}$, we write $Y_A := (Y_{i_1}, Y_{i_2}, \ldots, Y_{i_m})$. We consider a probability distribution over $x = (x_1, \ldots, x_N) \in \mathcal{X}_\mathcal{V}$ that can be written as a product of factors $(\psi_I)_{I \in \mathcal{F}}$:

$$\mathbb{P}(x) = \frac{1}{Z} \prod_{I \in \mathcal{F}} \psi_I(x_{N_I}), \qquad \text{where } Z = \sum_{x \in \mathcal{X}_\mathcal{V}} \prod_{I \in \mathcal{F}} \psi_I(x_{N_I}). \tag{1}$$

For each factor index $I \in \mathcal{F}$, there is an associated subset $N_I \subseteq \mathcal{V}$ of variable indices and the factor ψ_I is a nonnegative function $\psi_I : \mathcal{X}_{N_I} \to [0, \infty)$. For a Bayesian network, the factors are (conditional) probability tables. In case of Markov random fields, the factors are often called potentials. In general, the normalizing constant ("partition sum") Z is not known and exact computation of Z is infeasible, due to the fact that the number of terms to be summed is exponential in the number of variables N. Similarly, computing marginal distributions $\mathbb{P}(x_A)$ for subsets of variables $A \subseteq \mathcal{V}$ is intractable in general. In this article, we focus on the task of obtaining rigorous bounds on single-variable marginals $\mathbb{P}(x_i) = \sum_{x_{\mathcal{V} \setminus \{i\}}} \mathbb{P}(x)$.

Factor graphs. We can represent the structure of the probability distribution (1) using a *factor graph* $(\mathcal{V}, \mathcal{F}, \mathcal{E})$. This is a bipartite graph, consisting of *variable nodes* $i \in \mathcal{V}$, *factor nodes* $I \in \mathcal{F}$, and *edges* $e \in \mathcal{E}$, with an edge $\{i, I\}$ between $i \in \mathcal{V}$ and $I \in \mathcal{F}$ if and only if the factor ψ_I depends on x_i (i.e., if $i \in N_I$). We will represent factor nodes visually as rectangles and variable nodes as circles. Figure 1(a) shows a simple example of a factor graph. The set of neighbors of a factor node I is precisely N_I; similarly, we denote the set of neighbors of a variable node i by $N_i := \{I \in \mathcal{F} : i \in N_I\}$. We will assume throughout this article that the factor graph corresponding to (1) is connected.

Convexity. We denote the set of *extreme points* of a convex set $X \subseteq \mathbb{R}^d$ by $\text{ext}(X)$. For a subset $Y \subseteq \mathbb{R}^d$, the *convex hull of* Y is defined as the smallest convex set $X \subseteq \mathbb{R}^d$ with $Y \subseteq X$; we denote the convex hull of Y as $\text{conv}(Y)$.

Measures. For $A \subseteq \mathcal{V}$, define $\mathcal{M}_A := [0, \infty)^{\mathcal{X}_A}$ as the set of nonnegative functions on $\mathcal{X}_A$. Each element of $\mathcal{M}_A$ can be identified with a finite measure on $\mathcal{X}_A$; therefore we will call the elements of $\mathcal{M}_A$ "measures on A". We write $\mathcal{M}_A^* := \mathcal{M}_A \setminus \{0\}$.

Operations on measures. Adding two measures $\Psi, \Phi \in \mathcal{M}_A$ results in the measure $\Psi + \Phi$ in $\mathcal{M}_A$. For $A, B \subseteq \mathcal{V}$, we can multiply a measure on $\mathcal{M}_A$ with a measure on $\mathcal{M}_B$ to obtain a measure on $\mathcal{M}_{A \cup B}$; a special case is multiplication with a scalar. Note that there is a natural embedding of $\mathcal{M}_A$ in $\mathcal{M}_B$ for $A \subseteq B \subseteq \mathcal{V}$ obtained by multiplying a measure $\Psi \in \mathcal{M}_A$ by $\mathbf{1}_{B \setminus A} \in \mathcal{M}_{B \setminus A}$, the constant function with value 1 on $\mathcal{X}_{B \setminus A}$. Another important operation is the partial summation: given $A \subseteq B \subseteq \mathcal{V}$ and $\Psi \in \mathcal{M}_B$, define $\sum_{x_A} \Psi$ to be the measure in $\mathcal{M}_{B \setminus A}$ obtained by summing Ψ over all $x_a \in \mathcal{X}_A$, i.e., $\sum_{x_A} \Psi : x_{B \setminus A} \mapsto \sum_{x_A \in \mathcal{X}_A} \Psi(x_A, x_{B \setminus A})$.

Operations on sets of measures. We will define operations on *sets* of measures by applying the operation on elements of these sets and taking the set of the resulting measures; e.g., if we have two

subsets $\Xi_A \subseteq \mathcal{M}_A$ and $\Xi_B \subseteq \mathcal{M}_B$ for $A, B \subseteq \mathcal{V}$, we define the product of the sets Ξ_A and Ξ_B to be the set of the products of elements of Ξ_A and Ξ_B, i.e., $\Xi_A \Xi_B := \{\Psi_A \Psi_B : \Psi_A \in \Xi_A, \Psi_B \in \Xi_B\}$.

Completely factorized measures. For $A \subseteq \mathcal{V}$, we will define $\mathcal{Q}_A$ to be the set of *completely factorized* measures on A, i.e., $\mathcal{Q}_A := \prod_{a \in A} \mathcal{M}_{\{a\}}$. Note that $\mathcal{M}_A$ is the convex hull of $\mathcal{Q}_A$. Indeed, we can write each measure $\Psi \in \mathcal{M}_A$ as a convex combination of measures in $\mathcal{Q}_A$ which are zero everywhere except at one particular value of their argument. We denote $\mathcal{Q}_A^* := \mathcal{Q}_A \setminus \{0\}$.

Normalized (probability) measures. We denote with $\mathcal{P}_A$ the set of probability measures on A, i.e., $\mathcal{P}_A = \{\Psi \in \mathcal{M}_A : \sum_{x_A} \Psi = 1\}$. The set $\mathcal{P}_A$ is called a *simplex*. Note that a simplex is convex; the simplex $\mathcal{P}_A$ has precisely $\#(\mathcal{X}_A)$ extreme points, each of which corresponds to putting all probability mass on one of the possible values of x_A. We define the *normalization operator* $\mathcal{N}$ which normalizes measures, i.e., for $\Psi \in \mathcal{M}_A^*$ we define $\mathcal{N}\Psi := \frac{1}{Z}\Psi$ with $Z = \sum_{x_A} \Psi$.

Boxes. Let $a, b \in \mathbb{R}^d$ such that $a_\alpha \leq b_\alpha$ for all components $\alpha = 1, \ldots, d$. Then we define the *box* with lower bound a and upper bound b by $\mathcal{B}(a, b) := \{x \in \mathbb{R}^d : a_\alpha \leq x_\alpha \leq b_\alpha$ for all $\alpha = 1, \ldots, d\}$. Note that a box is convex; indeed, its extreme points are the "corners" of which there are 2^d.

Smallest bounding boxes. Let $X \subseteq \mathbb{R}^d$ be bounded. The *smallest bounding box* of X is defined as $\mathcal{B}(X) := \mathcal{B}(a, b)$ where the lower bound a is given by the pointwise infimum of X and the upper bound b is given by the pointwise supremum of X, that is $a_\alpha := \inf\{x_\alpha : x \in X\}$ and $b_\alpha := \sup\{x_\alpha : x \in X\}$ for all $\alpha = 1, \ldots, d$. Note that $\mathcal{B}(X) = \mathcal{B}(\mathrm{conv}(X))$. Therefore, if X is convex, the smallest bounding box for X depends only on the extreme points $\mathrm{ext}(X)$, i.e., $\mathcal{B}(X) = \mathcal{B}(\mathrm{ext}(X))$; this bounding box can be easily calculated if the number of extreme points is not too large.

2.2 The basic tools

To calculate marginals of subsets of variables in some factor graph, several operations performed on measures are relevant: normalization, taking products of measures, and summing over subsets of variables. Here we study the interplay between convexity and these operations. This will turn out to be useful later on, because our bounds make use of convex sets of measures that are propagated over the factor graph.

The interplay between convexity and normalization, taking products and partial summation is described by the following lemma.

Lemma 1 *Let $A \subseteq \mathcal{V}$ and let $\Xi \subseteq \mathcal{M}_A^*$. Then:*

    1. $\mathrm{conv}(\mathcal{N}\Xi) = \mathcal{N}(\mathrm{conv}\,\Xi)$;

    2. for all $B \subseteq \mathcal{V}$, $\Psi \in \mathcal{M}_B$: $\mathrm{conv}(\Psi\Xi) = \Psi(\mathrm{conv}\,\Xi)$;

    3. for all $B \subseteq A$: $\mathrm{conv}\left(\sum_{x_B} \Xi\right) = \sum_{x_B} \mathrm{conv}\,\Xi$.   □

The next lemma concerns the interplay between convexity and taking products; it says that if we take the product of convex sets of measures on different spaces, the resulting set is contained in the convex hull of the product of the extreme points of the convex sets.

Lemma 2 *Let $(A_t)_{t=1,\ldots,T}$ be disjoint subsets of $\mathcal{V}$. For each $t = 1, \ldots, T$, let $\Xi_t \subseteq \mathcal{M}_{A_t}$ be convex with a finite number of extreme points. Then* $\mathrm{conv}\left(\prod_{t=1}^T \Xi_t\right) = \mathrm{conv}\left(\prod_{t=1}^T \mathrm{ext}\,\Xi_t\right)$.   □

The third lemma says that the product of several boxes on the same subset A of variables can be easily calculated: the product of the boxes is again a box, with as lower (upper) bound the product of the lower (upper) bounds of the boxes.

Lemma 3 *Let $A \subseteq \mathcal{V}$ and for each $t = 1, \ldots, T$, let $\underline{\Psi}_t, \overline{\Psi}_t \in \mathcal{M}_A$ such that $\underline{\Psi}_t \leq \overline{\Psi}_t$. Then* $\prod_{t=1}^T \mathcal{B}\left(\underline{\Psi}_t, \overline{\Psi}_t\right) = \mathcal{B}\left(\prod_{t=1}^T \underline{\Psi}_t, \prod_{t=1}^T \overline{\Psi}_t\right)$.   □

We are now ready to state the basic lemma. It basically says that one can bound the marginal of a variable by replacing a factor depending on some other variables by a product of single-variable
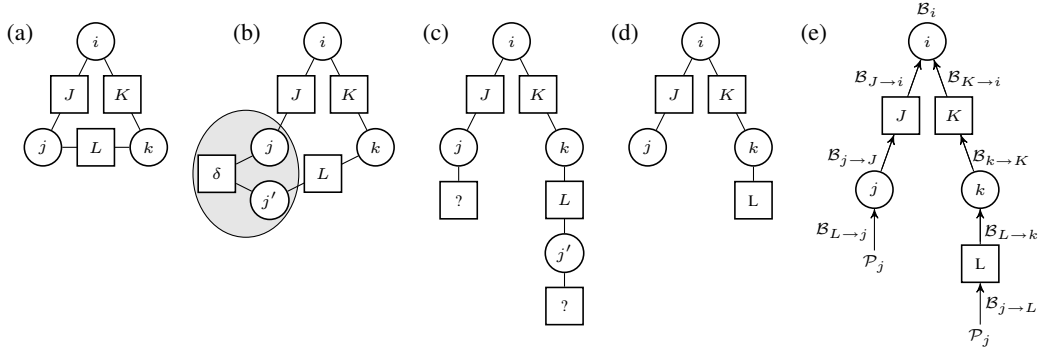
Figure 1: (a) Example factor graph with three variable nodes (i, j, k) and three factor nodes (J, K, L), with probability distribution $\mathbb{P}(x_i, x_j, x_k) = \frac{1}{Z}\psi_J(x_i, x_j)\psi_K(x_i, x_k)\psi_L(x_j, x_k)$; (b) Cloning node j by adding a new variable j' and a factor $\psi_\delta(x_j, x_{j'}) = \delta_{x_j}(x_{j'})$; (c) Illustration of the bound on $\mathbb{P}(x_i)$ based on (b): "what can we say about the range of $\mathbb{P}(x_i)$ when the factors corresponding to the nodes marked with question marks are arbitrary?"; (d) Subtree of the factor graph; (e) Propagating convex sets of measures (boxes or simplices) on the subtree (d), leading to a bound $\mathcal{B}_i$ on the marginal probability of x_i in $\mathcal{G}$.

factors and bounding the result. This can be exploited to simplify the computational complexity of bounding the marginal. An example of its use will be given in the next subsection.

Lemma 4 *Let $A, B, C \subseteq \mathcal{V}$ be mutually disjoint subsets of variables. Let $\Psi \in \mathcal{M}_{A \cup B \cup C}$ such that for each $x_C \in \mathcal{X}_C$, $\sum_{x_{A \cup B}} \Psi > 0$. Then:*

$$\mathcal{B}\left(\mathcal{N}\left(\sum_{x_B, x_C} \Psi \mathcal{M}_C^*\right)\right) = \mathcal{B}\left(\mathcal{N}\left(\sum_{x_B, x_C} \Psi \mathcal{Q}_C^*\right)\right).$$

Proof. Note that $\mathcal{M}_C^*$ is the convex hull of $\mathcal{Q}_C^*$ and apply Lemma 1. $\qquad\square$

The positivity condition is a technical condition, which in our experience is fulfilled for many practically relevant factor graphs.

2.3 A simple example

Before proceeding to our main result, we first illustrate for a simple case how the basic lemma can be employed to obtain computationally tractable bounds on marginals. We derive a bound for the marginal of the variable x_i in the factor graph in Figure 1(a). We start by cloning the variable x_j, i.e., adding a new variable $x_{j'}$ that is constrained to take the same value as x_j. In terms of the factor graph, we add a variable node j' and a factor node δ, connected to variable nodes j and j', with corresponding factor $\psi_\delta(x_j, x_{j'}) := \delta_{x_j}(x_{j'})$; see also Figure 1(b). Clearly, the marginal of x_i satisfies:

$$\mathbb{P}(x_i) = \mathcal{N}\left(\sum_{x_j}\sum_{x_k} \psi_J \psi_K \psi_L\right) = \mathcal{N}\left(\sum_{x_j}\sum_{x_{j'}}\sum_{x_k} \psi_J \psi_K \psi_L \delta_{x_j}(x_{j'})\right)$$

where it should be noted that the first occurrence of ψ_L is shorthand for $\psi_L(x_j, x_k)$, but the second occurrence is shorthand for $\psi_L(x_{j'}, x_k)$. Noting that $\psi_\delta \in \mathcal{M}_{\{j,j'\}}^*$ and applying the basic lemma with $A = \{i\}$, $B = \{k\}$, $C = \{j, j'\}$ and $\Psi = \psi_J \psi_K \psi_L$ yields:

$$\mathbb{P}(x_i) \in \mathcal{N}\left(\sum_{x_j}\sum_{x_{j'}}\sum_{x_k} \psi_J \psi_K \psi_L \mathcal{M}_{\{j,j'\}}^*\right) \in \mathcal{BN}\left(\sum_{x_j}\sum_{x_{j'}}\sum_{x_k} \psi_J \psi_K \psi_L \mathcal{Q}_{\{j,j'\}}^*\right).$$

Applying the distributive law, we obtain (see also Figure 1(c)):

$$\mathbb{P}(x_i) \in \mathcal{BN}\left(\left(\sum_{x_j} \psi_J \mathcal{M}_{\{j\}}^*\right)\left(\sum_{x_k} \psi_K \sum_{x_{j'}} \psi_L \mathcal{M}_{\{j'\}}^*\right)\right),$$

which we relax to

$$\mathbb{P}(x_i) \in \mathcal{BN}\left(\mathcal{BN}\left(\sum_{x_j} \psi_J \mathcal{P}_{\{j\}}\right) \mathcal{BN}\left(\sum_{x_k} \psi_K \mathcal{BN}\left(\sum_{x_{j'}} \psi_L \mathcal{P}_{\{j'\}}\right)\right)\right).$$

Now it may seem that this smallest bounding box would be difficult to compute. Fortunately, we only need to compute the extreme points of these sets because of convexity. Since smallest bounding boxes only depend on extreme points, we conclude that

$$\mathbb{P}(x_i) \in \mathcal{BN}\left(\mathcal{BN}\left(\sum_{x_j} \psi_J \mathrm{ext}\, \mathcal{P}_{\{j\}}\right) \mathcal{BN}\left(\sum_{x_k} \psi_K \mathcal{BN}\left(\sum_{x_{j'}} \psi_L \mathrm{ext}\, \mathcal{P}_{\{j'\}}\right)\right)\right).$$

which can be calculated efficiently if the number of possible values of each variable is small.

2.4 The main result

The example in the previous subsection can be generalized as follows. First, one chooses a particular subtree of the factor graph, rooted in the variable for which one wants to calculate a bound on its marginal. Then, one propagates messages (which are either bounding boxes or simplices) over this subtree, from the leaf nodes towards the root node. The update equations resemble those of Belief Propagation. The resulting "belief" at the root node is a box that bounds the exact marginal of the root node. The choice of the subtree is arbitrary; different choices lead to different bounds in general. We now describe this "box propagation" algorithm in more detail.

Definition 5 *Let $(\mathcal{V}, \mathcal{F}, \mathcal{E})$ be a factor graph. We call the bipartite graph (V, F, E) a subtree of $(\mathcal{V}, \mathcal{F}, \mathcal{E})$ with root i if $i \in V \subseteq \mathcal{V}$, $F \subseteq \mathcal{F}$, $E \subseteq \mathcal{E}$ such that (V, F, E) is a tree with root i and for all $\{j, J\} \in E$, $j \in V$ and $J \in F$ (i.e., there are no "loose edges").*[1] *We denote the parent of $j \in V$ according to (V, F, E) by $\mathrm{par}(j)$ and similarly, we denote the parent of $J \in F$ by $\mathrm{par}(J)$.*

An illustration of a possible subtree of the factor graph in Figure 1(a) is the one shown in Figure 1(d). The bound that we will obtain using this subtree corresponds to the example described in the previous subsection.

In the following, we will use the topology of the *original* factor graph $(\mathcal{V}, \mathcal{F}, \mathcal{E})$ whenever we refer to neighbors of variables or factors. Each edge of the subtree will carry one message, oriented such that it "flows" towards the root node. In addition, we define messages entering the subtree for all "missing" edges in the subtree (see also Figure 1(e)). Because of the bipartite character of the factor graph, we can distinguish two types of messages: messages $\mathcal{B}_{J \to j} \subseteq \mathcal{M}_j$ sent to a variable $j \in V$ from a neighboring factor $J \in N_j$, and messages $\mathcal{B}_{j \to J} \subseteq \mathcal{M}_j$ sent to a factor $J \in F$ from a neighboring variable $j \in N_J$. The messages entering the subtree are all defined to be simplices; more precisely, we define the incoming messages

$$\begin{aligned} \mathcal{B}_{j \to J} &= \mathcal{P}_j \qquad \text{for all } J \in F, \{j, J\} \in \mathcal{E} \setminus E \\ \mathcal{B}_{J \to j} &= \mathcal{P}_j \qquad \text{for all } j \in V, \{j, J\} \in \mathcal{E} \setminus E. \end{aligned}$$

We propagate messages towards the root i of the tree using the following update rules (note the similarity with the BP update rules). The message sent from a variable $j \in V$ to its parent $J = \mathrm{par}(j) \in F$ is defined as

$$\mathcal{B}_{j \to J} = \begin{cases} \displaystyle\prod_{K \in N_j \setminus J} \mathcal{B}_{K \to j} & \text{if all incoming } \mathcal{B}_{K \to j} \text{ are boxes} \\ \mathcal{P}_j & \text{if at least one of the } \mathcal{B}_{K \to j} \text{ is the simplex } \mathcal{P}_j, \end{cases}$$

where the product of the boxes can be calculated using Lemma 3. The message sent from a factor $J \in F$ to its parent $k = \mathrm{par}(J) \in V$ is defined as

$$\mathcal{B}_{J \to k} = \mathcal{BN}\left(\sum_{x_{N_J \setminus k}} \psi_J \prod_{l \in N_J \setminus k} \mathcal{B}_{l \to J}\right) = \mathcal{BN}\left(\sum_{x_{N_J \setminus k}} \psi_J \prod_{l \in N_J \setminus k} \mathrm{ext}\, \mathcal{B}_{l \to J}\right), \qquad (2)$$

---

[1] Note that this corresponds to the notion of subtree of a bipartite graph; for a subtree of a factor graph, one sometimes imposes the additional constraint that for all factors $J \in F$, all its connecting edges $\{J, j\}$ with $j \in N_J$ have to be in E; here we do not impose this additional constraint.

where the second equality follows from Lemmas 1 and 2. The final "belief" $\mathcal{B}_i$ at the root node i is calculated by

$$
\mathcal{B}_i = \begin{cases} \mathcal{BN}\left(\displaystyle\prod_{K \in N_i} \mathcal{B}_{K \to i} \right) & \text{if all incoming } \mathcal{B}_{K \to i} \text{ are boxes} \\ \mathcal{P}_i & \text{if at least one of the } \mathcal{B}_{K \to i} \text{ is the simplex } \mathcal{P}_i. \end{cases}
$$

We can now formulate our main result, which gives a rigorous bound on the exact single-variable marginal of the root node:

Theorem 6 *Let $(\mathcal{V}, \mathcal{F}, \mathcal{E})$ be a factor graph with corresponding probability distribution (1). Let $i \in \mathcal{V}$ and (V, F, E) be a subtree of $(\mathcal{V}, \mathcal{F}, \mathcal{E})$ with root $i \in V$. Apply the "box propagation" algorithm described above to calculate the final "belief" $\mathcal{B}_i$ on the root node i. Then $\mathbb{P}(x_i) \in \mathcal{B}_i$.*

Proof sketch The first step consists in extending the subtree such that each factor node has the right number of neighboring variables by cloning the missing variables. The second step consists of applying the basic lemma where the set C consists of all the variable nodes of the subtree which have connecting edges in $\mathcal{E} \setminus E$, together with all the cloned variable nodes. Then we apply the distributive law, which can be done because the extended subtree has no cycles. Finally, we relax the bound by adding additional normalizations and smallest bounding boxes at each factor node in the subtree. It should now be clear that the "box propagation" algorithm described above precisely calculates the smallest bounding box at the root node i that corresponds to this procedure. $\square$

Because each subtree of the orginal factor graph is also a subtree of the *computation tree for i* [12], the bounds on the (exact) marginals that we just derived are at the same time bounds on the approximate Belief Propagation marginals (beliefs):

Corollary 7 *In the situation described in Theorem 6, the final bounding box $\mathcal{B}_i$ also bounds the (approximate) Belief Propagation marginal of the root node i, i.e., $\mathbb{P}_{BP}(x_i) \in \mathcal{B}_i$.* $\square$

2.5 Related work

We briefly discuss the relationship of our bound to previous work. More details are provided in [11].

The bound in [6] is related to the bound we present here; however, the bound in [6] differs from ours in that it (i) goes deeper into the computation tree by propagating bounds over *self-avoiding-walk (SAW) trees* instead of mere subtrees, (ii) uses a different parameterization of the propagated bounds and a different update rule, and (iii), it is only formulated for the special case of factors depending on two variables, while it is not entirely obvious how to extend the result to more general factor graphs.

Another method to obtain bounds on exact marginals is "Bound Propagation" [10]. The idea underlying Bound Propagation is very similar to the one employed in this work, with one crucial difference. For a variable $i \in \mathcal{V}$, we define the sets $\Delta i := \bigcup N_i$ (consisting of all variables that appear in some factor in which i participates) and $\partial i := \Delta i \setminus \{i\}$ (the *Markov blanket* of i). Whereas our method uses a *cavity* approach, using as basis equation

$$
\mathbb{P}(x_i) \propto \sum_{x_{\partial i}} \left(\prod_{I \in N_i} \psi_I \right) \mathbb{P}^{\setminus i}(x_{\partial i}), \qquad \mathbb{P}^{\setminus i}(x_{\partial i}) \propto \sum_{x_{\mathcal{V} \setminus \Delta i}} \prod_{I \in \mathcal{F} \setminus N_i} \psi_I
$$

and bound the quantity $\mathbb{P}(x_i)$ by optimizing over $\mathbb{P}^{\setminus i}(x_{\partial i})$, the basis equation employed by Bound Propagation is $\mathbb{P}(x_i) = \sum_{x_{\partial i}} \mathbb{P}(x_i \mid x_{\partial i}) \mathbb{P}(x_{\partial i})$ and the optimization is over $\mathbb{P}(x_{\partial i})$. Unlike in our case, the computational complexity of Bound Propagation is exponential in the size of the Markov blanket, because of the required calculation of the conditional distribution $\mathbb{P}(x_i \mid x_{\partial i})$. On the other hand, the advantage of this approach is that a bound on $\mathbb{P}(x_j)$ for $j \in \partial i$ is also a bound on $\mathbb{P}(x_{\partial i})$, which in turn gives rise to a bound on $\mathbb{P}(x_i)$. In this way, bounds can propagate through the graphical model, eventually yielding a new (tighter) bound on $\mathbb{P}(x_{\partial i})$. Although the iteration can result in rather tight bounds, the main disadvantage of Bound Propagation is its computational cost: it is exponential in the Markov blanket and often many iterations are needed for the bounds to become tight.
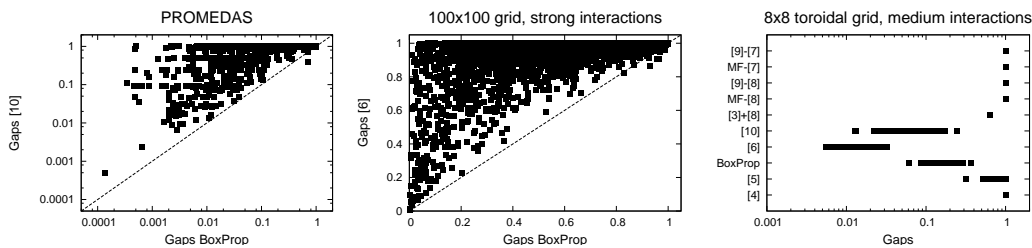
Figure 2: Comparisons of various methods on different factor graphs: PROMEDAS (left), a large grid with strong interactions (middle) and a small grid with medium-strength interactions (right).

3 Experiments

In this section, we present only few empirical results due to space constraints. More details and additional experimental results are given in [11]. We have compared different methods for calculating bounds on single-variable marginals; for each method and each variable, we calculated the *gap* (tightness) of the bound, which we defined as the ℓ_∞ distance between the upper and lower bound of the bounding box. We have investigated three different types of factor graphs; the results are shown in Figure 2. The factor graphs used for our experiments are provided as supplementary material to the electronic version of this article at `books.nips.cc`. We also plan to release the source code of several methods as part of a new release of the approximate inference library libDAI [13]. For our method, we chose the subtrees in a breadth-first manner.

First, we applied our bound on simulated PROMEDAS patient cases [14]. These factor graphs have binary variables and singleton, pairwise and triple interactions (containing zeros). We generated nine different random instances. For each instance, we calculated bounds for each "finding" variable in that instance using our method ("BOXPROP") and the method in [10]. Note that the tightness of both bounds varies widely depending on the instance and on the variable of interest. Our bound was tighter than the bound from [10] for all but one out of 1270 variables. Furthermore, whereas [10] had only finished on 7 out of 9 instances after running for $75000\,\mathrm{s}$ (after which we decided to abort the calculation on the remaining two instances), our method only needed $51\,\mathrm{s}$ to calculate all nine instances.

We also compared our method with the method described in [6] on a large grid of 100×100 binary (± 1-valued) variables with strong interactions. Note that this is an intractable problem for exact inference methods. The single-variable factors were chosen as $\exp(\theta_i x_i)$ with $\theta_i \sim \mathcal{N}(0,1)$, the pair factors were $\exp(\theta_{ij} x_i x_j)$ with $\theta_{ij} \sim \mathcal{N}(0,1)$. We truncated the subtree to 400 nodes and the SAW tree to 10^5 nodes. Note that our method yields the tightest bound for almost all variables.

Finally, we compared our method with several other methods referred to in Section 1 on a small 8×8 grid with medium-strength interactions (similarly chosen as for the large grid, but with $\theta_i \sim \mathcal{N}(0, 0.2^2)$ and $\theta_{ij} \sim \mathcal{N}(0, 0.2^2)$). The small size of the grid was necessary because some methods would need several days to handle a large grid. In this case, the method by [6] yields the tightest bounds, followed by [10], and our method gets a third place. Note that many methods return completely uninformative bounds in this case.

4 Conclusion and discussion

We have described a novel bound on exact single-variable marginals, which is at the same time a bound on the (approximate) Belief Propagation marginals. Contrary to many other existing bounds, it is formulated for the general case of factor graphs with discrete variables and factors depending on an arbitrary number of variables. The bound is calculated by propagating convex sets of measures over a subtree of the factor graph, with update equations resembling those of BP. For variables with a limited number of possible values, the bounds can be computed efficiently. We have compared our bounds with existing methods and conclude that our method belongs to the best methods, but that it is difficult to say in general which method will yield the tightest bounds for a given variable in a

specific factor graph. Our method could be further improved by optimizing over the choice of the subtree.

Although our bounds are a step forward in quantifying the error of Belief Propagation, the actual error made by BP is often at least one order of magnitude lower than the tightness of these bounds. This is due to the fact that (loopy) BP cycles information through loops in the factor graph; this cycling apparently often improves the results. The interesting and still unanswered question is why it makes sense to cycle information in this way and whether this error reduction effect can be quantified.

Acknowledgments

References

[1] G.F. Cooper. The computational complexity of probabilistic inferences. *Artificial Intelligence*, 42(2-3):393–405, March 1990.

[2] J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference*. Morgan Kaufmann, San Francisco, CA, 1988.

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

[4] S. C. Tatikonda. Convergence of the sum-product algorithm. In *Proceedings 2003 IEEE Information Theory Workshop*, pages 222–225, April 2003.

[5] Nobuyuki Taga and Shigeru Mase. Error bounds between marginal probabilities and beliefs of loopy belief propagation algorithm. In *MICAI*, pages 186–196, 2006.

[6] A. Ihler. Accuracy bounds for belief propagation. In *Proceedings of the 23th Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, July 2007.

[7] T. S. Jaakkola and M. Jordan. Recursive algorithms for approximating probabilities in graphical models. In *Proc. Conf. Neural Information Processing Systems (NIPS 9)*, pages 487–493, Denver, CO, 1996.

[8] M. J. Wainwright, T. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, 51:2313–2335, July 2005.

[9] M. A. R. Leisink and H. J. Kappen. A tighter bound for graphical models. In Lawrence K. Saul, Yair Weiss, and Léon Bottou, editors, *Advances in Neural Information Processing Systems 13 (NIPS*2000)*, pages 266–272, Cambridge, MA, 2001. MIT Press.

[10] M. Leisink and B. Kappen. Bound propagation. *Journal of Artificial Intelligence Research*, 19:139–154, 2003.

[11] J. M. Mooij and H. J. Kappen. Novel bounds on marginal probabilities. *arXiv.org*, arXiv:0801.3797 [math.PR], January 2008. Submitted to Journal of Machine Learning Research.

[12] S. C. Tatikonda and M. I. Jordan. Loopy belief propagation and Gibbs measures. In *Proc. of the 18th Annual Conf. on Uncertainty in Artificial Intelligence (UAI-02)*, pages 493–500, San Francisco, CA, 2002. Morgan Kaufmann Publishers.

[13] J. M. Mooij. libDAI: A free/open source C++ library for discrete approximate inference methods, 2008. http://mloss.org/software/view/77/.

[14] B. Wemmenhove, J. M. Mooij, W. Wiegerinck, M. Leisink, H. J. Kappen, and J. P. Neijt. Inference in the Promedas medical expert system. In *Proceedings of the 11th Conference on Artificial Intelligence in Medicine (AIME 2007)*, volume 4594 of *Lecture Notes in Computer Science*, pages 456–460. Springer, 2007.