Mach Learn (2008) 72: 205–229
DOI 10.1007/s10994-008-5072-8

New closed-form bounds on the partition function

Dvijotham Krishnamurthy · Soumen Chakrabarti · Subhasis Chaudhuri

Received: 23 June 2008 / Revised: 23 June 2008 / Accepted: 23 June 2008 / Published online: 15 July 2008
Springer Science+Business Media, LLC 2008

Abstract Estimating the partition function is a key but difficult computation in graphical models. One approach is to estimate tractable upper and lower bounds. The piecewise upper bound of Sutton et al. is computed by breaking the graphical model into pieces and approximating the partition function as a product of local normalizing factors for these pieces. The tree reweighted belief propagation algorithm (TRW-BP) by Wainwright et al. gives tighter upper bounds. It optimizes an upper bound expressed in terms of convex combinations of spanning trees of the graph. Recently, Globerson et al. gave a different, convergent iterative dual optimization algorithm TRW-GP for the TRW objective. However, in many practical applications, particularly those that train CRFs with many nodes, TRW-BP and TRW-GP are too slow to be practical. Without changing the algorithm, we prove that TRW-BP converges in a single iteration for associative potentials, and give a closed form for the solution it finds. The closed-form solution obviates the need for complex optimization. We use this result to develop new closed-form upper bounds for MRFs with arbitrary pairwise potentials. Being closed-form, they are much faster to compute than TRW-based bounds. We also prove similar convergence results for loopy belief propagation (LBP) and use it to obtain closed-form solutions to the LBP pseudomarginals and approximation to the partition function for associative potentials. We then use recent results proved by Wainwright et al for binary MRFs to obtain closed-form lower bounds on the partition function. We then develop novel lower bounds for arbitrary associative networks. We report on experiments with synthetic and real-world graphs. Our new upper bounds are considerably tighter than the piecewise bounds in practice. Moreover, we can compute our bounds on several graphs where TRW-BP does not converge. Our novel lower bound, in spite of being closed-form and much faster to compute, outperforms more complicated popular algorithms for computing lower bounds like mean-field on densely connected graphs by wide margins although it does worse on sparsely connected graphs like chains.

Editors: Walter Daelemans, Bart Goethals, Katharina Morik.

D. Krishnamurthy (✉) · S. Chakrabarti · S. Chaudhuri
IIT Bombay, Mumbai, India
e-mail: dvij@cse.iitb.ac.in

Keywords Partition function · Graphical model · Associative potential · Approximate inference · Belief propagation · Variational methods

1 Introduction

Graphical models are indispensable tools for computer vision, information extraction, language processing, bioinformatics, and other applications. The computation of the partition function is a critical computation in graphical models, for evaluating marginal probabilities and for training the parameters of the potential functions. Exact computation of the partition function is computationally intractable except for trees and other restricted families of graphs. Hence, developing tractable approximations to the partition function is a very important problem.

Variational methods (Jordan and Wainwright 2003) are one of the most popular ways of developing principled approximations to the partition function, and can be used to compute rigorous lower and upper bounds on the partition function. The idea behind variational methods is to express the computation of the partition function as an optimization problem (albeit computationally intractable) and relax this optimization problem in various ways to obtain tractable optimization problems that can provide upper and lower bounds on the partition function. The optimization is generally done by algorithms known as message passing algorithms that solve the optimization problem by passing messages between nodes of the graph till convergence is achieved. Two popular algorithms of this kind are loopy belief propagation (LBP) and tree-reweighted belief propagation (TRW-BP). TRW-BP is a message passing algorithm that optimizes an upper bound over convex decompositions of the parameter vector. However, it is not guaranteed to converge. Recently, Globerson and Jaakkola (2007) have proposed a different message-passing algorithm TRW-GP that solves the same optimization problem as TRW-BP and is guaranteed to converge. LBP is an algorithm that is not guaranteed to give bounds on the partition function although it has been used very successfully in practice and has some theoretical justification (Yedidia et al. 2000) on the basis of the Bethe entropy approximation. However, even this algorithm is not guaranteed to converge in general. In spite of being widely used, all these methods prove to be prohibitively expensive for applications that required repeated inference over large graphs. A recent technique proposed to provide approximations that remain tractable even for such applications is the piecewise approximation (PW) (Sutton and McCallum 2005). The piecewise approximation simply breaks the graph into pieces and computes an upper bound on the partition function by taking a product over locally normalizing factors for each edge, thus avoiding any kind of message-passing altogether. Although this is a fairly loose bound on the partition function, the authors in (Sutton and McCallum 2005) successfully apply it to several common NLP tasks.

Our contributions In Sect. 3, we give a sufficient condition on the initialization of the TRW-BP algorithm that guarantees convergence in one iteration for a class of potentials(including associative potentials). As a result of this convergence proof, we obtain a closed form expression for the TRW bound for MRFs with this class of potentials.

The closed form for associative potentials (and a generalization of them described in Sect. 4) is important, because, given an arbitrary potential, we consider a decomposition of it into a convex combination of an associative part, for which we use the closed-form TRW bound and a non-associative residue for which we can use a PW bound, thus obtaining an upper bound on the partition function using its convexity. We optimize over all such decompositions and show that this optimization problem has a closed-form solution as well.

We also prove that our bound is a convex function of the model parameters and also prove bounds on the gap between our new bounds and the piecewise bound. We present a similar convergence result for LBP as well in Sect. 5. Using results from (Sudderth et al. 2008), we obtain closed-form lower bounds on the log partition function for binary MRFs with attractive associative potentials. We then develop novel closed-form lower bounds for arbitrary associative potentials by decomposing a potential into a convex combination of attractive and non-attractive parts and using the closed-form LBP lower bounds on the attractive part and the closed-form TRW bound on the non-attractive part.

In Sect. 6, we report on experiments with real and synthetic graphs: chain graphs used in information extraction and statistical NLP, grid graphs commonly used in computer vision, the social network of papers and authors extracted from CiteSeer (nodes represent papers or authors, edges represent relations like author-wrote-paper and paper-cited-paper), and social networks of actors and movies extracted from the Internet Movie Database (IMDB).

Our new upper bounds on the partition function are about as fast as piecewise, but tighten the piecewise bound by 10–25% for different kinds of graphs across graph sizes. We can quickly compute bounds on graphs on which TRW does not converge, or takes impractical amounts of time.

Our novel lower bound for associative binary MRFs also performs well, beating standard methods like mean-field by large margins on densely connected graphs although it does worse on sparser graphs like chains.

2 Preliminaries

In this section, we introduce notation and review relevant related work that motivate our new analysis.

2.1 Graphical model basics

An undirected graphical model consists of an undirected graph $G = (V, E)$ with potential functions $\phi_C(\mathbf{y}_C)$ (where C is a clique in G) that defines the following distribution over variables associated with nodes of the graph:

$$\Pr(\mathbf{y}) = \frac{1}{Z}\prod_C \Psi_C(\mathbf{y}_C)$$

where $\mathbf{y}$ is a vector of length $|V|$ giving the values of variables and $\mathbf{y}_C$ denotes the subset of variables associated with the clique C. Here, we shall consider only node and pairwise potentials, that is, the cliques C consist only of nodes and edges of the graph. We shall assume that for each $s \in V$, the corresponding variable $\mathbf{y}_s$ takes values in some discrete set $\mathcal{X}_s = \{0, 1, 2, \dots, m_s - 1\}$. We shall assume for convenience that $m_s = m$ for each s. We denote the joint configuration space by $\prod_s \mathcal{X}_s$. In this paper, for notational convenience, we assume that $\mathcal{X}_s = \mathcal{X} \forall s$ so that the space becomes $\mathcal{X}^n$ where $n = |V|$. These popular graphical models are known as discrete Markov random fields (MRFs) with pairwise potential functions. We shall use $m = |\mathcal{X}|$ to denote the number of labels for each node.

The *node potential* $\Psi_s(y_s)$ depends on the state of a single node s. The *edge potential* $\Psi_{st}(y_s, y_t)$ depends on the states of nodes (s, t) across an edge. These can be further parameterized as

$$\Psi_s(y_s) = \exp\left(\sum_{i,\mathcal{X}} [\![y_s = i]\!] \theta_{s;i}\right),$$

$$\Psi_{st}(y_s, y_t) = \exp\left(\sum_{i,j\in\mathcal{X}} [\![y_s = i]\!] [\![y_t = j]\!] \theta_{st;ij}\right)$$

where $[\![y_s = i]\!]$ is an indicator function that takes the value 1 if $y_s = i$ and 0 otherwise. In general, node potentials can be absorbed into edge potentials (we present one way to do this in Sect. 2.2) so in this paper we shall mostly consider models that have only edge potentials (i.e. all node potentials are set to 0). The quantity

$$\sum_{\mathbf{y}\in\mathcal{X}^n} \exp\left(\sum_{s\in V, i\in\mathcal{X}} \theta_{s;i} [\![y_s = i]\!] + \sum_{(s,t)\in E, i,j\in\mathcal{X}} \theta_{st;ij} [\![y_s = i]\!] [\![y_t = j]\!]\right)$$

is called the partition function (denoted by $Z(\Theta)$) of the MRF and plays a central role in parameter estimation in graphical models. We shall use $A(\Theta)$ to denote $\log(Z(\Theta))$ and it is known that this is a convex function of Θ (Jordan and Wainwright 2003).

2.2 Piecewise (PW) bound

In piecewise training we compute an approximation to the exact partition function by taking a product over all locally normalized factors. For discrete MRFs with pairwise potential functions and without node potentials,

$$Z_{pw}(\Theta) = \prod_{(s,t)\in E} \left(\sum_{y_s, y_t} \Psi_{st}(y_s, y_t)\right). \tag{1}$$

Even if there are single-node potentials, these can be absorbed into pairwise potentials as follows: let $s \in V$ and let $Nb(s)$ denote the set of neighbors of s in the graph. Then we can modify the pairwise potentials containing s such that $\Psi'_{st}(y_s, y_t) = \Psi_{st}^{w_t}(y_s)$ and $\sum_{t\in Nb(s)} w_t = 1$. We do this for each node and in this manner absorb all node potentials into the pairwise potentials. Assume first that the graph is connected. Construct the BFS tree of any node in r in the graph. For each node $s \neq r$ in the graph, we call $e(s)$ the edge through which it was discovered in the BFS and the node through which it was discovered $n(s)$. Note that this assigns a unique edge to each $s \neq r$. Let the set of all these edges be called E_{BFS}. We also pick any neighbor of r and call it $n(r)$ and call the edge between them $e(r)(\in E_{BFS}$ by definition) Let $f_s(y_s) = \sum_{y_{n(s)}} \Psi_{e(s)}(y_s, y_{n(s)}) \forall s \neq r, n(r)$.

Fact 1 $Z_{pw} \geq Z$.

Proof We have

$$\begin{aligned} Z(\Theta) &= \sum_{\mathbf{y}\in\mathcal{X}^n} \prod_{(s,t)\in E} \Psi_{st}(y_s, y_t) \\ &= \sum_{\mathbf{y}\in\mathcal{X}^n} \left(\prod_{s\in V\backslash r, n(r)} \Psi_{e_s}(y_s, y_{n(s)})\right)\left(\prod_{(s,t)\notin E_{BFS}} \Psi_{st}(y_s, y_t)\right) \end{aligned}$$

We now use $\Psi_{st}(y_s, y_t) \leq \sum_{y_s, y_t\in\mathcal{X}} \Psi_{st}(y_s, y_t) \forall (s,t) \notin E_{BFS}$ so that these can be brought out of the summation. Then we upper bound the terms left inside using $\Psi_{e(s)}(y_s, y_{n(s)}) \leq \sum_{y_{n(s)}} \Psi_{e(s)}(y_s, y_{n(s)})$. This gives us the following upper bound:

$$
\begin{aligned}
Z(\Theta) \le & \left(\sum_{\mathbf{y}\in\mathcal{X}^n} \Psi_{e(r)}(y_r, y_{n(r)}) \prod_{s\in V\backslash\{r,n(r)\}} \left(\sum_{y_{n(s)}} \Psi_{e_s}(y_s, y_{n(s)}) \right) \right) \\
& \times \prod_{(s,t)\notin E_{BFS}} \left(\sum_{y_s, y_t\in\mathcal{X}} \Psi_{st}(y_s, y_t) \right) \\
= & \left(\sum_{\mathbf{y}\in\mathcal{X}^n} \Psi_{e(r)}(y_r, y_{n(r)}) \left(\prod_{s\in V\backslash\{r,n(r)\}} f_s(y_s) \right) \right) \prod_{(s,t)\notin E_{BFS}} \left(\sum_{y_s, y_t\in\mathcal{X}} \Psi_{st}(y_s, y_t) \right) \\
= & \left(\sum_{y_r, y_{n(r)}} \Psi_{e(r)}(y_r, y_{n(r)}) \right) \left(\prod_{s\in V\backslash\{r,n(r)\}} \sum_{y_s} f(y_s) \right) \left(\prod_{(s,t)\notin E_{BFS}} \left(\sum_{y_s, y_t\in\mathcal{X}} \Psi_{st}(y_s, y_t) \right) \right) \\
= & \prod_{(s,t)\in E} \left(\sum_{y_s, y_t\in\mathcal{X}} \Psi_{st}(y_s, y_t) \right). \qquad \square
\end{aligned}
$$

This proof is much simpler than the one presented in (Sutton and McCallum 2005), where they derive the bound using a convex decomposition of the parameter vector and taking the limits as some weights in the convex combination tend to 1.

2.3 Loopy belief propagation (LBP)

Loopy belief propagation (Yedidia et al. 2000) is one of the most popular algorithms used for obtaining approximations to the partition function and marginal probabilities. It is an iterative algorithm that starts with a random initial set of messages $\{M_{st}^0\}$ and updates them as follows:

$$M_{ts}^{n+1}(i) = \alpha \sum_j \exp\{\theta_{st;ij} + \theta_{t;j}\} \prod_{v\in Nb(t)\backslash s} M_{vt}^n(j).$$

When convergence is achieved(that is, the messages remain unchanged after the above updates), to say M^*, an approximation to the marginals is given by

$$
\begin{aligned}
&\mu_{s;i}^* \propto \exp(\theta_{s;i}) \prod_{t\in Nb(s)} M_{ts}^*(i), \\
&\mu_{st;ij}^* \propto \exp(\theta_{st;ij} + \theta_{s;i} + \theta_{t;j}) \prod_{u\in Nb(s)\backslash t} M_{us}^*(i) \prod_{v\in Nb(t)\backslash s} M_{vt}^*(j), \\
&\mu^* = [\mu_{s;i} : i \in \mathcal{X} s \in V] \cup [\mu_{st;ij} : i, j \in \mathcal{X}(s,t) \in E]
\end{aligned}
$$

is called the set of *pseudomarginals*. μ by definition belongs to the LOCAL(G), an outer approximation to the set of marginals realizable from the graphical model, defined as follows:

$$\text{LOCAL(G)} = \tau : \begin{cases} \sum_i \tau_{s;i} = 1, \\ \sum_i \tau_{st;ij} = \tau_{s;i}, \\ \sum_j \tau_{st;ij} = \tau_{t;j}. \end{cases} \tag{2}$$

The LBP algorithm provides an approximation to the log partition function given by:

$$\mathrm{LBP}(\theta) = \langle \mu^*, \theta \rangle + \sum_{s \in V} H_s(\mu_s^*) - \sum_{(s,t) \in E} I_{st}(\mu_{st}^*) \tag{3}$$

where

$$H_s(\mu_s) = -\sum_{i \in \mathcal{X}} \mu_{s;i} \log(\mu_{s;i})$$

is the entropy function and

$$I_{st}(\mu_{st}) = \sum_{i,j \in \mathcal{X}} \mu_{st;ij} \log\left(\frac{\mu_{st;ij}}{\mu_{s;i}\mu_{t;j}}\right)$$

is the mutual information between two random variables with marginals μ_s, μ_t and joint distribution μ_{st} and $\langle ., . \rangle$ denotes the inner product of two vectors.

2.4 Tree reweighted belief propagation (TRW-BP)

The tree reweighted belief propagation algorithm (Wainwright et al. 2005) is very popular for computing tractable upper bounds on the log partition function. It uses the convexity of the partition function to derive upper bounds by decomposing the parameter vector over spanning trees of the graph and optimizes bounds thus obtained over all possible decompositions. Let G be a graph underlying an MRF. Let $ST(G)$ be the set of all spanning trees of G. We use Θ^T to denote a parameter vector for the MRF that respects the structure of T, that is, $\Theta_{st}^T = 0 \; \forall (s,t) \notin E$. The idea behind the TRW upper bound is to write the parameter vector as a convex combination of parameters over trees and then use the convexity of the log partition function. Let $A(\Theta)$ denote the log partition function of the MRF parameterized by the vector Θ. Suppose that we have a probability distribution $\vec{\rho}$ over the set of spanning trees of the graph. Then the TRW upper bound on the log partition function is given by solving the following optimization problem:

$$\min_{\{\Theta^T\}: \sum_{T \in ST(G)} \rho^T \Theta^T = \Theta} \sum_{T \in ST(G)} \rho^T A(\Theta^T). \tag{4}$$

Since the number of spanning trees is exponentially large for several classes of graphs, the optimization is done in the dual where the number of optimization variables is tractable. Strong duality is shown to hold, and the dual is given by

$$\mathrm{TRW}(\theta, \vec{\rho}_e) = \max_{\tau \in \mathrm{LOCAL(G)}} \langle \tau, \Theta \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}). \tag{5}$$

We observe that the optimal solution depends only on the set of edge appearance probabilities (ρ_{st}) and not on the entire probability vector $\vec{\rho}$. Let $\rho_e = \{\rho_{st} : (s,t) \in E\}$ be the vector of edge appearance probabilities. This vector must belong to the *spanning tree polytope* $\mathcal{T}(G)$ of the graph (Wainwright et al. 2005). Thus, for any fixed $\rho_e \in \mathcal{T}(G)$, we can get a upper bound on the log partition function by solving the optimization problem (5). It is possible to compute a valid ρ_e efficiently using the matrix-tree theorem (Wainwright 2002). Given a fixed ρ_e, it is possible to solve this using standard convex optimization techniques (Boyd and Vandenberghe 2004) reasonably efficiently. However, for large graphs, even these techniques become prohibitively expensive and alternatives are required. Wainwright et al propose an iterative message-passing algorithm called Tree Reweighted Belief Propagation (TRW-BP) to solve the convex optimization problem (5). The algorithm iteratively updates messages at each node using information from its neighbors until convergence is achieved. Wainwright et al show that any fixed point of this iterative scheme can be used to compute a stationary point of the Lagrangian of (5), which is also a global maximum due to the concavity of the objective function and convexity of the constraint set. Thus, if and when the iterations converge, they can be used to compute the optimal solution to the problem (5). However, convergence of the iterative scheme is not guaranteed and it is possible for the algorithm to get stuck in cycles. In (Wainwright et al. 2005), the authors also propose a method to optimize the TRW bound with respect to ρ_e as well, using conditional gradient descent and alternating between steps of running the TRW-BP algorithm and solving a maximum spanning tree problem. However, in this paper, we shall assume that we are dealing with a fixed ρ_e throughout. We denote the TRW upper bound for a fixed ρ_e as $\mathrm{TRW}(\theta, \rho_e)$ although we might sometimes let the dependence on ρ_e be implicit.

2.5 Convergent alternatives to TRW-BP

In recent work, Globerson and Jaakkola (2007) propose provably convergent alternatives to TRW-BP for solving the TRW optimization problem (4). Using oriented trees, they derive an alternative dual to the TRW optimization problem that can be expressed as an unconstrained instance of a generalized geometric program and derive a message passing algorithm to optimize the dual. They prove convergence of this new message-passing algorithm, TRW-GP, for arbitrary potentials. However, even TRW-GP does not have any guarantees on the number of iterations required for convergence. Both TRW-BP and TRW-GP are likely to be too expensive for applications involving training large CRFs that require repeated inference.

2.6 Certain special classes of potentials

In this section, we describe various classes of potential functions that will appear in the rest of this paper. We first note that a pairwise potential on edge $(s, t) \in E$ can be conveniently represented as an $m \times m$ matrix where the ijth entry is $\theta_{st;ij}$.

Associative potentials Pairwise Potentials are said to be *associative* if they only depend on whether the labels of the neighboring nodes are the same or not. That is,

$$\theta_{st;ij} = \begin{cases} \theta_{st;p} & \text{if } i = j, \\ \theta_{st;n} & \text{if } i \neq j. \end{cases}$$

An example of associative potentials is the homogeneous Ising model $\theta_{st,p} = \theta, \theta_{st;n} = -\theta \forall (s, t) \in E$.

Attractive potentials A binary pairwise potential is said to be *attractive* if $\theta_{st;11} + \theta_{st;00} \geq \theta_{st;01} + \theta_{st;10}$ Given this definition, an associative potential is an *attractive associative* potential if $\theta_{st;p} \geq \theta_{st;n}$.

Generalized associative potentials We call a potential *Generalized Associative* if every row and every column of the matrix representing it is a permutation of the same set of m real numbers. An example of a generalized associative matrix is the matrix

$$\begin{pmatrix} \theta_{st;1} & \theta_{st;2} & \dots & \theta_{st;m} \\ \theta_{st;2} & \theta_{st;3} & \dots & \theta_{st;1} \\ \vdots & \vdots & \vdots & \vdots \\ \theta_{st;m} & \theta_{st;1} & \dots & \theta_{st;m-1} \end{pmatrix}.$$

In this paper, we shall use the above form as representative of generalized associative potentials although other potentials that satisfy the condition would do as well. We call the above form a *cyclic potential.*

TRW closed-form potentials In Sect. 3, we shall prove that TRW(θ, $\vec{\rho}$) has a closed-form expression for potentials satisfying the following condition:

$$\begin{aligned} \frac{\sum_j \exp(\frac{\theta_{st;ij}}{\rho_{st}}) \exp(\theta_{t;j})}{\sum_j \exp(\theta_{t;j})} &= \frac{\sum_i \exp(\frac{\theta_{st;ij}}{\rho_{st}}) \exp(\theta_{s;i})}{\sum_i \exp(\theta_{s;i})} \\ &= \text{const}_{st} \quad \forall (s,t) \in E \end{aligned} \tag{6}$$

where const$_{st}$ is a constant that depends only on the edge (s,t) and not on the labels i, j.

LBP closed-form potentials In Sect. 5, we shall prove that LBP(θ) has a closed-form expression for potentials satisfying the following condition:

$$\begin{aligned} \frac{\sum_j \exp(\theta_{st;ij} + \theta_{t;j})}{\sum_j \exp(\theta_{t;j})} &= \frac{\sum_i \exp(\theta_{st;ij} + \theta_{s;i})}{\sum_i \exp(\theta_{s;i})} \\ &= \text{const}_{st} \quad \forall (s,t) \in E \end{aligned} \tag{7}$$

where const$_{st}$ is a constant that depends only on the edge (s,t) and not on the labels i, j.

Observe that any model that has associative potentials and no node potentials satisfies both the above conditions. More generally, any Generalized Associative potential satisfies these conditions.

3 New analysis of TRW

In this section, we present quick convergence results for the TRW-BP algorithm for certain classes of potentials and using these results we give a closed-form expression for the TRW upper bound for this class of potentials.

3.1 Convergence of TRW

The TRW-BP algorithm (Wainwright et al. 2005) starts with an arbitrary set of initial messages $\{M_{st}^0\}$ and updates them as follows:

$$M_{ts}^{n+1}(i) = \alpha \sum_j \frac{\prod_{v \in Nb(t) \setminus s} [M_{vt}^n(j)]^{\rho_{vt}}}{[M_{st}^n(j)]^{(1-\rho_{ts})}} \exp\left[\frac{\theta_{st;ij}}{\rho_{st}} + \theta_{t;j}\right]$$

where the constant α is a normalizing factor chosen such that the messages sum to one and $Nb(t)$ is the set of neighbors of node t in the graph G. In this section, we shall assume that the initial messages M^0_{st} are uniform, that is, $M^0_{st}(i) = \frac{1}{m} = M^0_{ts}(j) \forall i, j \in \mathcal{X}, (s,t) \in E$. Now, suppose the potentials satisfy condition (6). Then

$$\begin{aligned}
M^1_{ts}(i) &= \alpha \sum_j \frac{\prod_{v \in Nb(t)\backslash s} [M^0_{vt}(j)]^{\rho_{vt}}}{[M^0_{st}(j)]^{(1-\rho_{ts})}} \exp\left[\frac{\theta_{st;ij}}{\rho_{st}} + \theta_{t;j}\right] \\
&= \alpha \sum_j \left(\frac{1}{m}\right)^{(\sum_{v \in Nb(s)} \rho_{vt} - 1)} \exp\left[\frac{\theta_{st;ij}}{\rho_{st}} + \theta_{t;j}\right] \\
&= \alpha \left(\frac{1}{m}\right)^{(\sum_{v \in Nb(s)} \rho_{vt} - 1)} \sum_j \exp\left(\frac{\theta_{st;ij}}{\rho_{st}} + \theta_{t;j}\right) \\
&= \alpha \left(\frac{1}{m}\right)^{(\sum_{v \in Nb(s)} \rho_{vt} - 1)} \left(\sum_j \exp(\theta_{t;j})\right) \mathrm{const}_{st},
\end{aligned}$$

where the final expression was derived using the condition on the potentials (6). The final expression is independent of i, hence the messages are still uniform. Thus, we have shown that TRW converges in one iteration if we start with uniform messages.

Fact 2 *For MRFs with potentials satisfying the condition* (6), *the TRW-BP algorithm converges in a single iteration.*

3.2 Closed-form TRW upper bound

Once the messages converge, say, to $\{M^f_{st}\}$, the optimal solution to (5) is given by (Jordan and Wainwright 2003)

$$\begin{aligned}
\tau_{st;ij} &\propto \exp\left[\frac{\theta_{st;ij}}{\rho_{st}} + \theta_{s;i} + \theta_{t;j}\right] \frac{\prod_{v \in Nb(t)\backslash s} [M^f_{vt}(j)]^{\rho_{vt}}}{[M^f_{st}(j)]^{(1-\rho_{ts})}} \frac{\prod_{v \in Nb(s)\backslash t} [M^f_{vs}(j)]^{\rho_{vs}}}{[M^f_{ts}(j)]^{(1-\rho_{ts})}}, \\
\tau_{s;i} &\propto \exp(\theta_{s;i}) \prod_{v \in \Gamma(s)} [M_{vs}(i)]^{\rho_s}
\end{aligned}$$

where the proportionality constants are determined by the constraint that $\tau \in \mathrm{LOCAL}(G)$. In our case, since the final messages are uniform, the solutions τ are

$$\tau_{st;ij} \propto \exp\left(\frac{\theta_{st;ij}}{\rho_{st}} + \theta_{s;i} + \theta_{t;j}\right) \quad \text{and} \quad \tau_{s;i} \propto \exp(\theta_{s;i}).$$

Thus, from the constraints $\tau \in \mathrm{LOCAL}(G)$, we can determine that the optimal pseudomarginals for the optimization problem (5) are given by

$$\begin{aligned}
\tau_{st;ij} &= \frac{\exp(\frac{\theta_{st;ij}}{\rho_{st}} + \theta_{s;i} + \theta_{t;j})}{\mathrm{const}_{st}(\sum_i \exp(\theta_{s;i}))(\sum_j \exp(\theta_{t;j}))}, \\
\tau_{s;i} &= \frac{\exp(\theta_{s;i})}{\sum_i \exp(\theta_{s;i})}.
\end{aligned} \tag{8}$$

The value of the dual objective function (5) at this set of pseudomarginals gives the value of the TRW bound (since strong duality holds (Wainwright et al. 2005)). This fact can be used to obtain a closed-form expression for $\mathrm{TRW}(\theta, \rho_e)$.

Fact 3 *For MRFs with potentials satisfying* (6), *the TRW bound is given by*

$$\mathrm{TRW}(\theta, \rho_e) = \left(\sum_{s\in V} \log\left(\sum_{i\in\mathcal{X}} \exp(\theta_{s;i})\right)\right) + \left(\sum_{(s,t)\in E} \rho_{st} \log(\mathrm{const}_{st})\right) \tag{9}$$

It can be easily seen that models that have associative potentials and no node potentials ($\theta_{s;i} = 0$) satisfy (6). Using the notation in Sect. 2, we can write the TRW upper bound in this case as

$$\log(m) + \left(\sum_{(s,t)\in E} \rho_{st} \log\left(\exp\left(\frac{\theta_{st;p}}{\rho_{st}}\right) + (m-1)\exp\left(\frac{\theta_{st;n}}{\rho_{st}}\right)\right)\right).$$

3.3 Accuracy of TRW pseudomarginals

In this section, we observe using an example that the TRW pseudomarginals can be an arbitrarily bad approximation to the true marginals. Consider for example the case of a complete graph with 3 binary-valued nodes $s, u, v$and associative potentials on each edge. Let $\theta_{su;p} = \theta_{sv;p} = \beta, \theta_{su;n} = \theta_{sv;n} = -\beta, \theta_{uv;p} = \gamma, \theta_{uv;n} = -\gamma$. Then, a direct calculation shows that

$$\frac{P(x_u = 1, x_v = 1)}{P(x_u = 1, x_v = 0)} = \frac{\exp(\gamma + \beta + \beta) + \exp(\gamma - \beta - \beta)}{\exp(-\gamma + \beta - \beta) + \exp(-\gamma - \beta + \beta)} = \exp(2\gamma)\cosh(2\beta).$$

We exclude Z from the above expressions since it cancels out in the ratio. Assuming a uniform distribution over spanning trees, we get $\rho_{st} = 2/3$ for all edges. The TRW estimate of the ratio $\frac{P(x_s=1,x_t=1)}{P(x_s=1,x_t=0)}$ is then (from (8))

$$\exp\left(\frac{\gamma}{\rho_{st}} - \frac{-\gamma}{\rho_{st}}\right) = \exp(3\gamma).$$

The factor $\cosh(2\beta)\exp(-\gamma)$ can be arbitrarily large or small and hence the TRW approximation can be arbitrarily skewed on either side of the true estimate. This shows that the pseudomarginals obtained from TRW cannot be taken as reliable estimates of the true probabilities in general.

4 Upper bounds for arbitrary pairwise MRFs

In this section, we derive an upper bound for MRFs with arbitrary potentials by decomposing the pairwise potentials into a part that satisfies equation (6) and a part that does not. In order to obtain closed-form solutions, we consider decompositions into convex combinations of cyclic and non-cyclic parts. We consider cyclic potentials for concreteness although we could use any form that is a Generalized Associative potential as well. The closed-form TRW bound for models where all potentials are of this form is given by (9) by:

$$\log(m) + \sum_{(s,t)\in E} \rho_{st} \log\left(\sum_{i} \exp\left(\frac{\theta_{st;i}}{\rho_{st}}\right)\right). \tag{10}$$

We then use the TRW bound for the cyclic part and the piecewise bound for the non-cyclic part to get an upper bound on the partition function. We can optimize over all such decompositions (keeping the ratio in the convex combination fixed) and show that this optimized bound has a closed-form expression. This gives a tractable upper bound while requiring no extra computation time. We also prove tightness results showing that our bound is not more than $\log(m)$ greater than the piecewise bound, and experimental results show that in practice our bound is almost always tighter. In this section, for notational simplicity, we consider MRFs with only pairwise and no node potentials. In general, node potentials can be absorbed into pairwise potentials so this does not lead to any loss in generality.

Let Θ be the parameter vector associated with the graphical model. We write this as a convex combination of an cyclic and a non-cyclic part as follows:

$$\Theta = p\beta + (1-p)\gamma$$

where β is cyclic, γ is non-cyclic and p $(0 < p < 1)$ is fixed. Now, by the convexity of the log-partition function, we get

$$A(\Theta) \le pA(\beta) + (1-p)A(\gamma)$$

Now, to get a closed form upper bound from this, we use the closed form tree-reweighted upper bound on $A(\beta)$ and the piecewise bound on $A(\gamma)$. Thus we get

$$\begin{aligned} A(\Theta) \le{}& p\left(\sum_{(s,t)\in E} \rho_{st} \log\left(\frac{\sum_i \exp(\frac{\beta_{st;i}}{\rho_{st}})}{m}\right) + n\log m\right) \\ &+ (1-p)\left(\sum_{(s,t)\in E} \log\left(\sum_{i,j} \exp(\gamma_{st;ij})\right)\right). \end{aligned} \tag{11}$$

Please note that for now we are only optimizing over β, γ keeping $p \in (0, 1)$ fixed. From the constraint $\Theta = p\beta + (1-p)\gamma$, we can express γ in terms of β and Θ and convert this into an unconstrained optimization problem over β. This is a convex optimization problem and hence setting gradients with respect to β to 0, we can get the optimal upper bound. Since the bound decomposes additively over edges, we can optimize each β_{st} separately. We can optimize each β_{st} easily by setting derivatives to zero and solving the resulting equations. Doing this gives us the following optimal upper bound over all decompositions:

Fact 4 *The overall optimal upper bound on the log partition function is given by*

$$\mathrm{TRWPW}(\theta, \rho_e, p) = \sum_{(s,t)\in E} (1 + p\rho_{st} - p) \log\left(\sum_i a_{st;i}^{\frac{1-p}{1-p+p\rho_{st}}}\right) + p\log(m) \tag{12}$$

where

$$a_{st;i} = \sum_j \exp\left(\frac{\theta_{st;j(1+(i+j-2)\mathrm{mod}\, m)}}{1-p}\right). \tag{13}$$

We call this the TRWPW bound.

4.1 Convexity of TRWPW

By an argument similar to the one in (Rennie 2005), $a_{st;i}(\theta_{st})$ is a log-convex function of θ_{st} $\forall i$. Since log-convex functions are closed under positive exponentiation, scaling and addition (Lange 2004),

$$\left(\sum_i a_{st;i}(\theta_{st})^{\frac{1-p}{1+p\rho_{st}-p}}\right)$$

is log convex as well. Thus, from (12), we get that

Fact 5 TRWPW(θ, ρ_e, p) *is a convex function of* Θ *for fixed* ρ_e, p.

4.2 Tightness of TRWPW

TRWPW vs PW Let $v_{st} = [a_{st;i}^{1-p} : i \in \mathcal{X}]$. Then, by the standard inequality between norms, $\|v_{st}\|_{\frac{1}{1-q}} \leq \|v_{st}\|_1 \leq m^q \|v_{st}\|_{\frac{1}{1-q}}$. The TRWPW is given by $\sum_{(s,t)\in E} \log(\|v_{st}\|_{\frac{1}{1-p+p\rho_{st}}}) + p\log(m)$. By the above inequality, this is less than $\sum_{st} \log(\|v_{st}\|_1) + p\log(m)$. Let $w_{st;i} = [\exp(\theta_{st;j(1+(i+j-2)\bmod m)}) : j \in \mathcal{X}]$. Then $\|v_{st}\|_1 = \sum_i \|w_{st;i}\|_{\frac{1}{1-p}}$. Thus $\|v_{st}\|_1 \leq \sum_i \|w_{st;i}\|_1 = \sum_{i,j} \exp(\theta_{s;i})$. Combining all the above inequalities, we get

$$\mathrm{TRWPW}(\theta, \rho_e, p) \leq p\log(m) + \sum_{(s,t)\in E} \log\left(\sum_{i,j} \exp(\theta_{st;ij})\right) \leq p\log(m) + \mathrm{PW}(\theta).$$

Equality is achieved above in the limit when all $\theta_{st;ij} \to -\infty$ for all except one pair of (i, j), $\forall (s,t)$. so that all the vectors used above have at most one non-zero component and all norms become equal then. Similarly, using the corresponding lower bounds, we get

$$\mathrm{TRWPW}(\theta, \rho_e, p) \geq \mathrm{PW}(\theta) - (|E| - p|V|)\log(m)$$

with equality being achieved when $\theta_{st;ij} = \theta_{st} \forall i, j$, which corresponds to the uniform distribution over all configurations.

Fact 6 $\mathrm{PW}(\theta) - (|E| - p|V|)\log(m) \leq \mathrm{TRWPW}(\theta, \rho_e, p) \leq \mathrm{PW}(\theta) + p\log(m)$ *and there exist potentials for which equality is achieved on both sides.*

TRWPW vs TRW TRWPW is a continuous and convex (Sect. 4.1) function of θ and in this section, we denote it as TRWPW(θ) making the dependence on ρ_e, p implicit. Hence, it can be represented in terms of its convex conjugate (Borwein and Lewis 2006). Its convex conjugate is given by

$$\mathrm{TRWPW}^*(\mu) = \sup_\theta \langle \mu, \theta \rangle - \mathrm{TRWPW}(\theta).$$

The above optimization problem can be solved easily by taking derivatives with respect to $\theta_{st;ij}$, setting them to 0 and solving the resulting equations to obtain expressions for $\mu_{st;ij}$. Substituting these back into the expression gives us the following formula for the convex conjugate:

$$\mathrm{TRWPW}^*(\mu) = \sum_{(s,t)\in E} -((1-p)H_{st}(\mu_{st}) - p\rho_{st} H_{st}(\mu'_{st})) - p\log(m) \qquad (14)$$

if $\sum_{i,j} \mu_{st;ij} = 1, \mu_{st;ij} \geq 0 \forall (s,t) \in E$ and ∞ otherwise.

Using this, we obtain a representation of TRWPW as follows:

$$\sup_{\mu : \sum_{i,j} \mu_{st;ij} = 1, \mu_{st;ij} \geq 0} \langle \mu, \Theta \rangle + \sum_{(s,t) \in E} ((1-p) H_{st}(\mu_{st}) + p \rho_{st} H_{st}(\mu'_{st})) + p \log(m) \tag{15}$$

where $H_{st}(v) = -\sum_i v_i \log(v_i)$ represents the entropy function and

$$\mu'_{st} = \left[\sum_j \mu_{st;j(1+(i+j-2) \bmod m)} : i \in \mathcal{X} \right]$$

We could augment the vector μ with node marginals $\mu_s : s \in V$ and enforce the constraint that $\sum_i \mu_{st;ij} = \mu_{t;j}$ and $\sum_j \mu_{st;ij} = \mu_{s;i}$ without making any difference to the above optimization problem. Once we do this, we are optimizing over the same constraint as the TRW dual (5), that is, LOCAL(G). We now show the following inequality: $\sum_{(s,t) \in E} (1-p) H(\mu_{st}) + p \rho_{st} H(\mu'_{st}) + \rho_{st} I_{st}(\mu_{st}) - \sum_{s \in V} H_s(\mu_s) \leq 2(1-p)(|E| - |V| + 1) \log(m) + (|V| - 2) \log(m)$ Letting $\rho_s = \sum_{t \in Nb(s)} \rho_{st}$ and using $H_{st}(\mu'_{st}) \leq H_{st}(\mu_{st})$(which follows from the basic properties of entropy) we can manipulate the LHS to get LHS $\leq (1-p) \sum_{(s,t) \in E} (1-\rho_{st}) H_{st}(\mu_{st}) + \sum_{s \in V} (\rho_s - 1) H_s(\mu_s)$ Now using $H_{st}(\mu_{st}) \leq 2 \log(m)$ and $H_s(\mu_s) \leq \log(m)$ and the fact that $\sum_{(s,t) \in E} \rho_{st} = |V| - 1$, we get LHS $\leq 2(1-p)(|E| - |V| + 1) \log(m) + (|V| - 2) \log(m)$ Thus, we have

$$\begin{aligned}
\text{TRWPW}(\Theta) &= p \log(m) + \sup_{\mu \in \text{LOCAL}(G)} \langle \mu, \Theta \rangle + \sum_{(s,t) \in E} (1-p) H_{st}(\mu_{st}) + p \rho_{st} H_{st}(\mu'_{st}) \\
&\leq \sup_{\mu \in \text{LOCAL}(G)} \langle \mu, \Theta \rangle + \sum_s H_s(\mu_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\mu_{st}) \\
&\quad + 2(1-p)(|E| - |V| + 1) \log(m) + (|V| - 2) \log(m) + p \log(m) \\
&= \text{TRW}(\Theta) + 2(1-p)(|E| - |V| + 1) \log(m) + (|V| - 2 + p) \log(m).
\end{aligned}$$

Fact 7 *The difference between TRWPW and TRW is bounded above by*

$$2(1-p)(|E| - |V| + 1) \log(m) + (|V| - 2) \log(m) + p \log(m).$$

Given these bounds, it does appear that p close to 1 is likely to give us tighter upper bounds and we have observed this experimentally as well. However, taking $p \to 1$ requires taking a limit that results in a non-differentiable expression involving max functions. This limits the utility of the bound when used as an approximation while training CRFs (one of the key applications of the piecewise bound). Also, there are cases where $p = 1$ does not give us the optimal upper bound (although these are rare).

5 Closed-form lower bounds on the partition function

In this section, we observe that LBP converges in a single iteration for potentials satisfying (7). We have seen in Sect. 2.6 that Generalized Associative potentials always satisfy this condition. We then use this to obtain closed-form solutions to the LBP approximation to the partition function and the LBP pseudomarginals. In (Sudderth et al. 2008), the authors prove that LBP gives a lower bound on the log partition function for binary MRFs with attractive potentials under certain conditions. We use the above results to obtain a closed-form lower bound for attractive associative potentials. We then use a decomposition approach similar to the previous section to obtain closed-form lower bounds for arbitrary potentials. We also use the new lower bounds to obtain error bounds on the accuracy of the TRW and LBP approximation to the partition function in these cases.

5.1 Convergence of belief propagation

In this section, we observe that loopy belief propagation converges in a single iteration for MRFs with potentials that satisfy the following condition (7). The proof closely parallels the proof of convergence of TRW-BP. If we start with uniform initial messages, it is easy to show(using the form of the belief propagation updates) that after updates the messages still remain uniform and hence convergence is achieved. The final marginals are given as:

$$\tau_{st;ij} = \frac{\exp(\theta_{st;ij} + \theta_{s;i} + \theta_{t;j})}{\text{const}_{st}(\sum_i \exp(\theta_{s;i}))(\sum_j \exp(\theta_{t;j}))},$$

$$\tau_{s;i} = \frac{\exp(\theta_{s;i})}{\sum_i \exp(\theta_{s;i})}.$$

Using the above equations and (3), we obtain

Fact 8 *The* LBP *approximation to the partition function for MRFs with potentials satisfying* (7) *is given by*

$$\text{LBP}(\theta) = \left(\sum_{s\in V} \log\left(\sum_{i\in\mathcal{X}} \exp(\theta_{s;i})\right)\right) + \left(\sum_{(s,t)\in E} \rho_{st} \log(\text{const}_{st})\right). \tag{16}$$

5.2 Closed-form Bethe variational bounds for attractive associative MRFs

In (Sudderth et al. 2008), Wainwright et al prove that loopy belief propagation gives a lower bound on the partition function for binary MRFs with attractive potentials. In this section, we use this result and the convergence result proved above to derive closed form lower bounds for binary random fields with associative potentials satisfying the following condition: $\theta_{st;p} \geq \theta_{st;n}$ and when the final node marginals produced by LBP satisfy $\tau_{s;1} \leq 1/2 \forall s$. Using the above result, we can see that for the case of associative potentials, belief propagation converges to the following marginals:

$$\tau_{st;00} = \theta_{st;11} = \frac{\exp(\theta_{st;p})}{2(\exp(\theta_{st;p}) + \exp(\theta_{st;n}))},$$

$$\tau_{st;10} = \theta_{st;01} = \frac{\exp(\theta_{st;n})}{2(\exp(\theta_{st;p}) + \exp(\theta_{st;n}))},$$

$$\tau_{s;i} = \frac{1}{2}.$$

Thus, if all the potentials satisfy $\theta_{st;p} \geq \theta_{st;n}$, all conditions required in (Sudderth et al. 2008) are satisfied and we get a closed-form lower bound on the partition function.

Fact 9 *The* log *partition function of an attractive associative binary pairwise MRF is bounded below by*:

$$\mathrm{LBP}(\Theta) \geq \sum_{(s,t)\in E} \log\left(\frac{\exp(\theta_{st;p}) + \exp(\theta_{st;n})}{2}\right) + |V|\log(2). \tag{17}$$

5.3 Upper bounds on the error for TRW and Bethe variational bounds

Since we now have both upper and lower bounds on the partition function for the case of attractive associative potentials, we can derive bounds on the error introduced by using these as approximations to the true partition function. $|\mathrm{TRW}(\Theta, \rho_e) - A(\Theta)| \leq \mathrm{TRW}(\Theta, \rho_e) - \mathrm{LBP}(\Theta)$. Manipulating the RHS, we can show that it is equal to

$$(|E| - |V| + 1)\log(2) + \sum_{(s,t)\in E} \rho_{st}\log\left(\cosh\left(\frac{\theta_{st;p} - \theta_{st;n}}{2\rho_{st}}\right)\right) - \log\left(\cosh\left(\frac{\theta_{st;p} - \theta_{st;n}}{2}\right)\right).$$

Fact 10 *The error between the* TRW *upper bound/Bethe Variational lower bound and the true log partition function for the case of binary pairwise MRFs with attractive associative potentials is bounded above by*

$$(|E| - |V| + 1)\log(2) + \sum_{(s,t)\in E} \rho_{st}\log\left(\cosh\left(\frac{\theta_{st}}{\rho_{st}}\right)\right) - \log(\cosh(\theta_{st}))$$

where $\theta_{st} = \frac{\theta_{st;p} - \theta_{st;n}}{2}$.

5.4 Closed-form lower-bounds for binary MRFs with arbitrary associative potentials

In the last section, we obtained closed-form lower bounds for binary MRFs with attractive associative potentials. In this section, we use the above result to obtain new closed-form lower bound with arbitrary associative potentials. By the convexity of the partition function, $pA(\Theta) + (1-p)A(\gamma) \geq A(\beta)$ where $p\Theta + (1-p)\gamma = \beta$. We now restrict β to be an attractive associative potential so that we can use the closed form LBP lower bound on $A(\beta)$. Since Θ itself is associative, β becomes the difference of two associative potentials and hence is associative and we can use the TRW upper bound on $A(\beta)$. We can then obtain a lower bound on the partition function as follows: $A(\Theta) \geq \frac{LBP(\beta) - (1-p)TRW(\gamma)}{p}$. We then maximize this bound with respect to β, γ subject to the constraint $\beta = p\Theta + (1-p)\gamma$ and β is attractive. This optimization problem can be written as follows:

$$\max_{\beta:\beta_{st;p}\geq\beta_{st;n}\forall(s,t)\in E} \frac{1}{p}(|V| - |E| - p)\log(2) + \frac{1}{p}\sum_{(s,t)\in E} \log(\exp(\beta_{st;p}) + \exp(\beta_{st;n}))$$
$$- (1-p)\rho_{st}\log\left(\exp\left(\frac{\beta_{st;p} - p\theta_{st;p}}{(1-p)\rho_{st}}\right) + \exp\left(\frac{\beta_{st;n} - p\theta_{st;n}}{(1-p)\rho_{st}}\right)\right).$$

It is easy to see that the optimization can be done independently for each edge. Doing so (by setting derivatives to zero and solving the resulting equations) results in the following lower bound:

$$\mathrm{BPTRW}(\theta, \rho_e, p) = \frac{1}{p}(|V| - |E| - p)\log(2) + \sum_{(s,t)\in E} [\![\theta_{st;p} \geq \theta_{st;n}]\!]\frac{1}{p}(1 - \rho_{st}(1-p))$$

$$\begin{aligned} &\times \log\left(\exp\left(\frac{p\theta_{st;p}}{1-\rho_{st}(1-p)}\right)+\exp\left(\frac{p\theta_{st;n}}{1-\rho_{st}(1-p)}\right)\right) \\ &+ [\![\theta_{st;p} < \theta_{st;n}]\!]\frac{1}{p}\Bigg(\log(2) \\ &- (1-p)\rho_{st}\log\left(\exp\left(\frac{-p\theta_{st;p}}{\rho_{st}(1-p)}\right)+\exp\left(\frac{-p\theta_{st;n}}{\rho_{st}(1-p)}\right)\right)\Bigg). \end{aligned} \tag{18}$$

6 Experiments

We compare our upper bounds to PW and TRW bounds using synthetic and two types of real-world graphs. We use the implementation of TRW-BP made available by Talya Meltzer at http://www.cs.huji.ac.il/~talyam/inference.html. The edge potentials are generated independently from a uniform distribution on the interval $[-5, 5]$ for all the graphs. Unless explicitly stated otherwise, all experiments are performed on graphs with binary-valued nodes. For the experiments involving TRWPW or BPTRW, we choose p by performing a coarse grid search with step 0.1 over the interval $[0.1, 0.9]$.

For TRW, TRWPW and BPTRW, we use as $\vec{\rho}$ the uniform distribution over spanning trees. It is possible to optimize $\vec{\rho}$ over all possible distributions using, e.g., conditional gradient descent (Wainwright et al. 2005). However, it does not appear that there is a closed-form solution for this optimization problem in general.

In most experiments, we report the value of the given bound minus the mean-field bound for the same graphical model so that the quantity reported is scale-free(invariant to multiplication of potential functions by a constant).

6.1 Arbitrary pairwise potentials

In this section, we report on experiments with arbitrary pairwise potentials.

Effect of p We do not consider the problem of optimizing over the TRWPW wrt p in this paper: It does not appear that this problem has a closed-form solution and the problem is probably not even convex. The dependence of the TRWPW bound is complicated and does not seem amenable to standard optimization techniques. However, we observe empirically (Fig. 1) that we get better bounds with larger ps most of the time in accordance with the error bounds in fact 7. However, there are cases when this is not true. Also, at $p = 1$ we need to take a limit, and the resulting expression is not differentiable in general, which limits its utility for training. So we limit ourselves to $p = 0.9$.

Chains These are cycle-less graphs used very frequently in natural language processing and information extraction. TRW is exact on these graphs, so this tells us how close TRWPW is to the exact answer. We present results comparing TRWPW and PW with TRW on these graphs in Fig. 2. TRWPW does not do much better than the piecewise although the bounds get better as the number of nodes increase.

Grid graphs These graphs occur naturally in modeling physical systems, and are also extensively used in computer vision to model images and interactions between neighboring pixels. The typical computer vision task is region segmentation. We consider $n \times n$ square lattices with n ranging from 3 to 45. The results are plotted in Fig. 3. TRWPW reduces the gap between TRW and PW by about 15–20%.

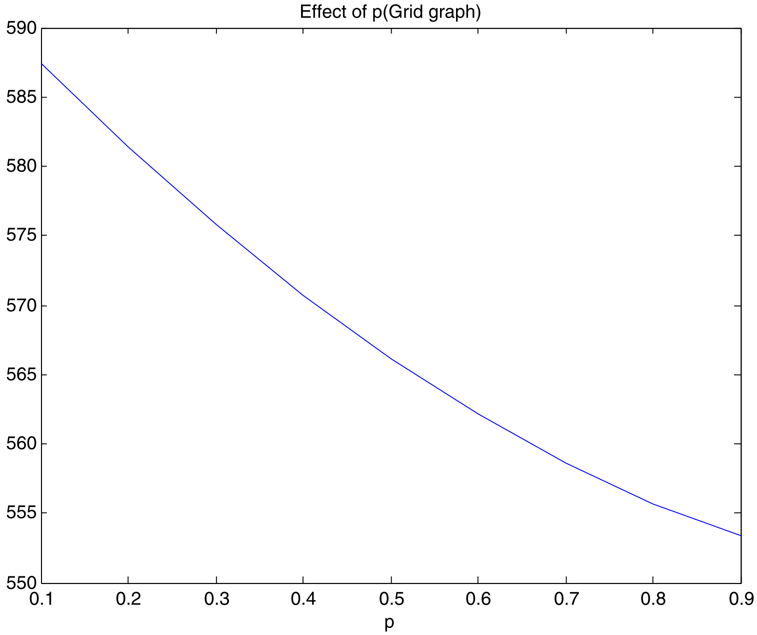

Fig. 1 Effect of *p* on bounds

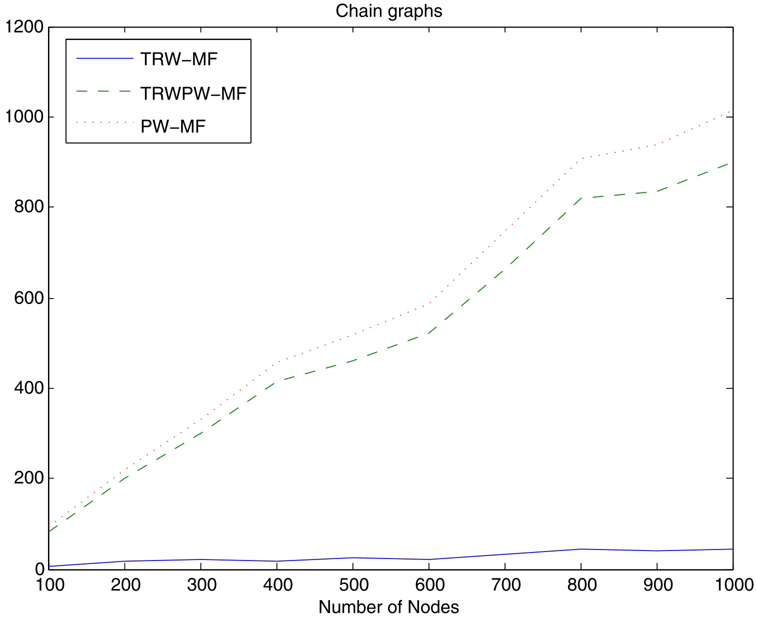

Fig. 2 TRWPW vs. TRW and PW on chain graphs

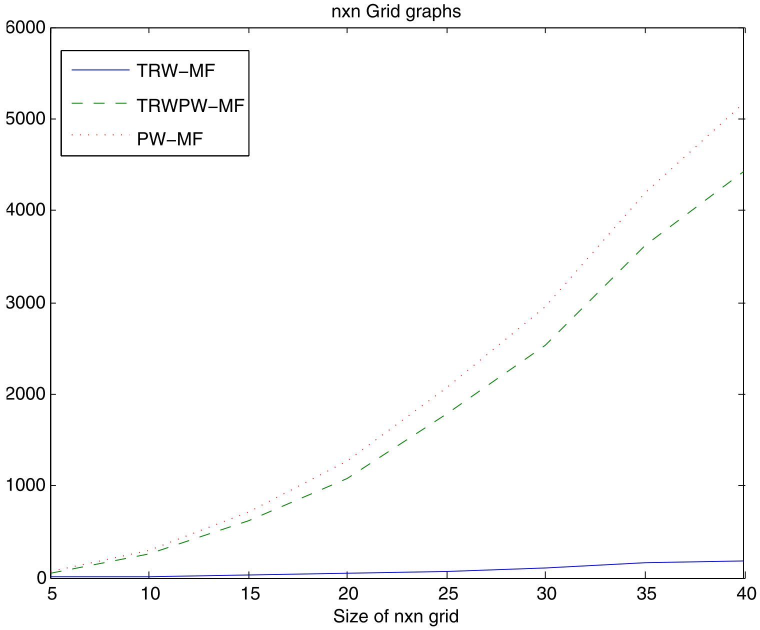

Fig. 3 TRWPW vs. TRW and PW on grid graphs

Complete graphs For stress-testing we also tried complete graphs. On complete graphs TRW has considerable trouble, often not converging after 2000 iterations. At the time of writing, an implementation of the convergent alternative to TRW (Globerson and Jaakkola 2007) was not available. Therefore we can present comparisons of our bounds with only PW. The results are plotted in Fig. 4.

Citation networks Starting from author and paper seed nodes, we performed breadth-first traversals to collect several neighborhood graphs from CITESEER. Probabilistic graphical models have been frequently used to label nodes in social networks (Lu and Getoor 2003). In CITESEER, for examples, one may wish to use a probabilistic graphical model for labeling papers about object-oriented databases apart from relational databases. Another motivation from Web search is to label host nodes as spam-prone or not, given the Web's link graph. These are both associative Markov networks. Our sample CITESEER subgraphs had 150–200 nodes and 400–500 edges.

Another important parameter to vary while testing bounds is the number of possible node labels m. In case of Web spam, there may be only two labels, but if labels represent topics of papers, there can be many. In Fig. 5 we varied the number of labels and measured the gap between PW and TRWPW. For these graphs, TRW again frequently failed to converge in 2000 iterations. TRWPW levels off quickly with increasing m, while PW continues to rise, with some slowdown. The number of terms in the piecewise bound increases quadratically and since we are using independent and randomly generated potentials, it is reasonable to expect the PW bound to grow approximately quadratically. Thus, when we take the log of this, we get a function with decreasing slope.

IMDB actor-movie graphs Explicit social networks are increasingly common outside academic citation. We collected a graph-structured version of the IMDB database

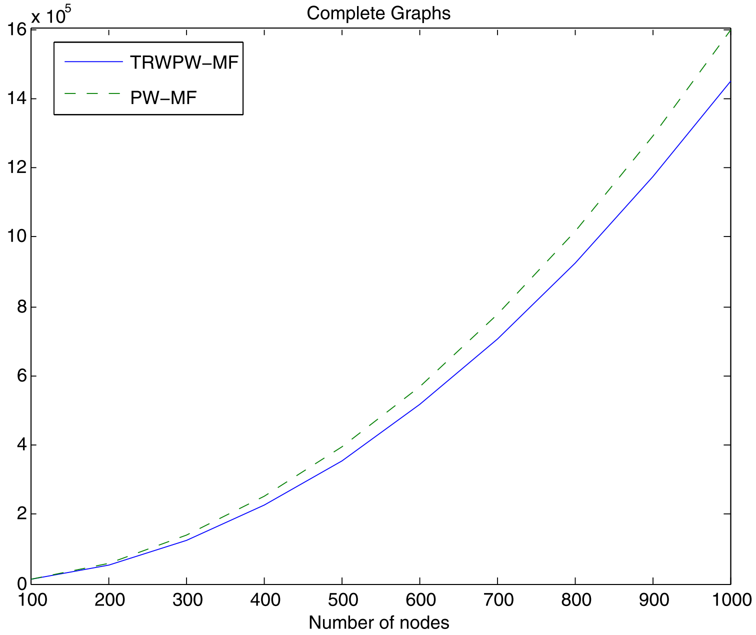

Fig. 4 TRWPW vs. PW on complete graphs (TRW frequently failed to converge in 2000 iterations)

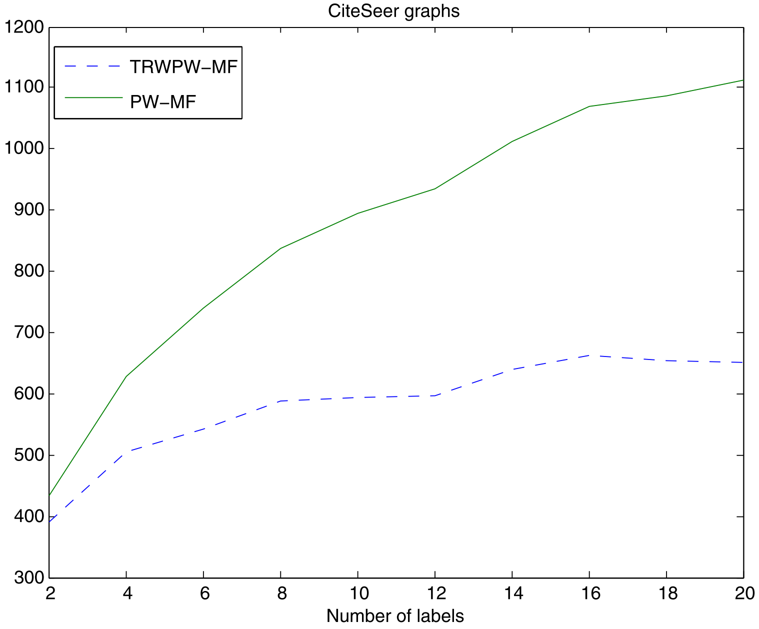

Fig. 5 TRWPW vs. PW on densely connected neighborhoods of CiteSeer, grown from seed nodes

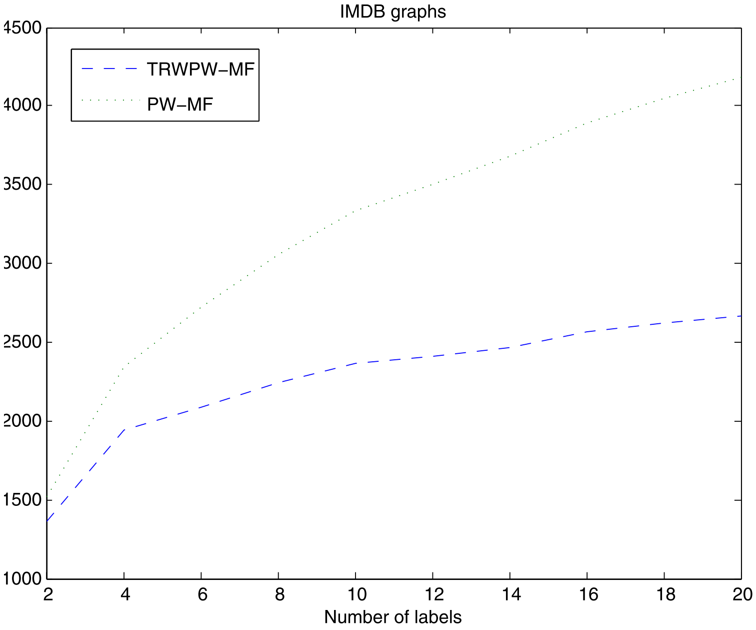

Fig. 6 TRWPW vs. PW on network communities around actor nodes in the IMDB movie database

(http://imdb.com) with nodes representing movies and actors, and bipartite edges representing the relation “actor acted-in movie”. Then, as with CITESEER, we started breadth-first traversals from popular seed nodes like *Sean Connery* to collect other actors and movies in their social network neighborhood. We observe that TRWPW is tighter by 25–50% for these graphs.

Figure 6 compares PW and TRWPW for one such graph. The results are very similar to CITESEER.

Effect of associativity Another factor that clearly influences TRWPW is the extent of associativity of the potential. To study this effect, we took a fixed associative potential β and a fixed non-associative potential γ(each generated with entries randomly chosen from the interval $[-5, 5]$), and built different convex combinations of the form $\Theta = \alpha\beta + (1-\alpha)\gamma$, for different values of α. We keep p, β, γ fixed in this experiment in order to observe just the effect of associativity on the quality of bounds. As expected, as we make the overall potential closer to associative (by increasing α), TRWPW gets better than PW relative to TRW. The results are plotted in Fig. 7.

TRW convergence In contrast to our one-shot bound computation, Fig. 8 shows that TRW can need a large number of iterations to converge. The ratio of TRW to our running times is the same order as the number of iterations needed by TRW (i.e., often two orders of magnitude or more).

6.2 Associative potentials

In this section, we present results comparing the BPTRW lower bound with the standard mean-field lower bound on various kinds of graphs with binary nodes and associative poten-

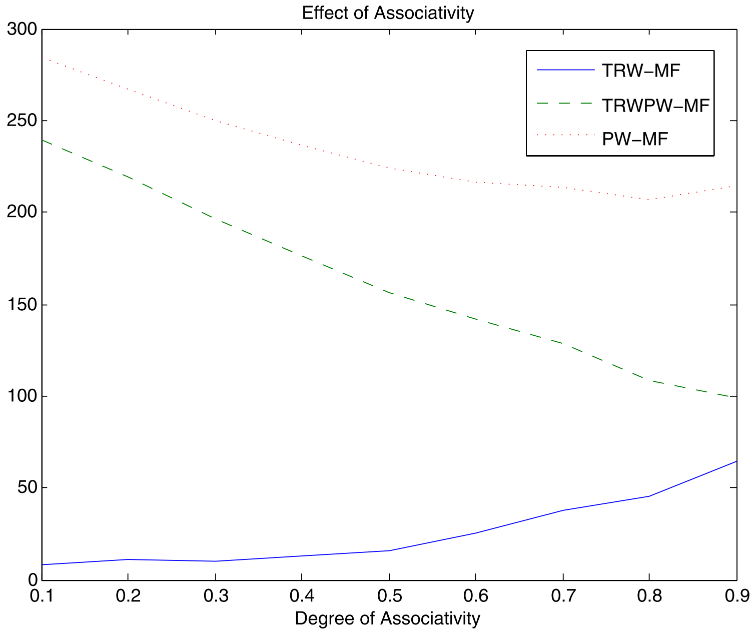

Fig. 7 Effect of associativity of potential on quality of TRW, PW and TRWPW

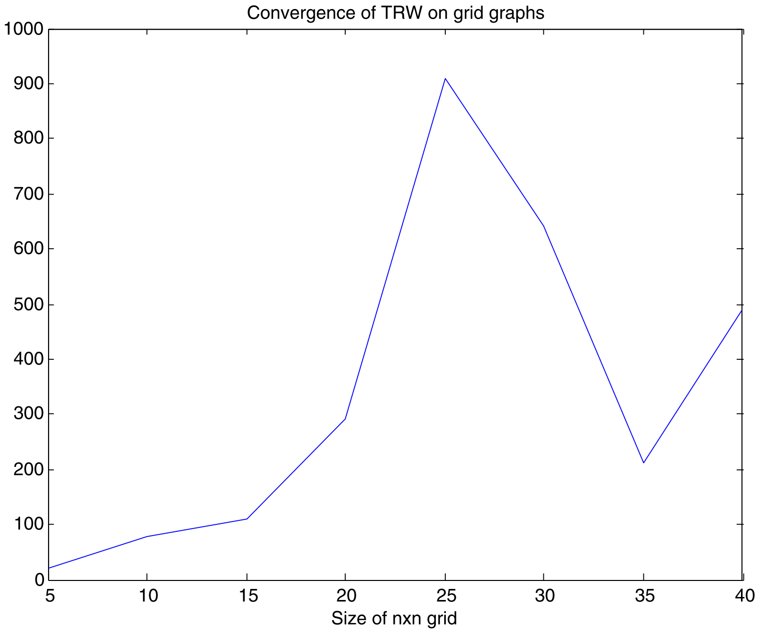

Fig. 8 Iterations needed for TRW convergence on a grid graph

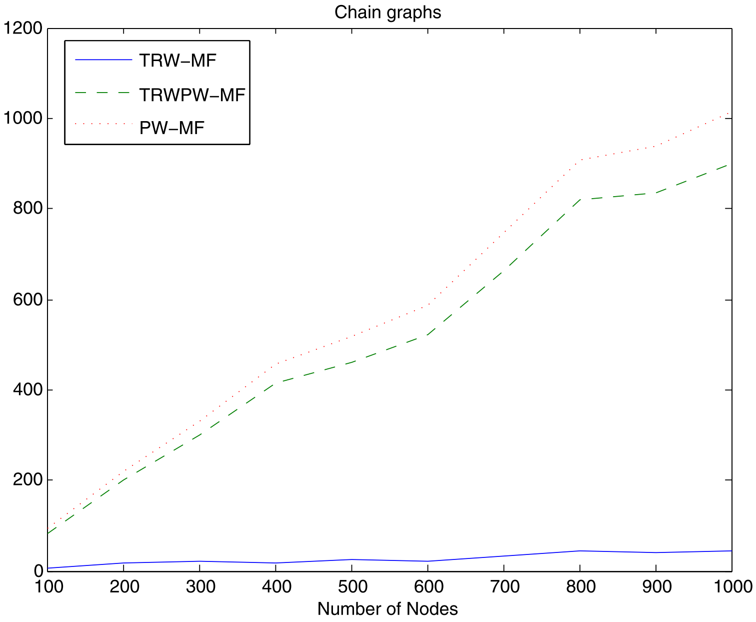

Fig. 9 TRW, BPTRW and mean-field on chain graphs

tials. We plot the values of TRW-MF as well on each graph in order to give an idea of how far these bounds are from the actual partition function (since TRW is an upper bound, the gap between it and a lower bound is an upper bound on the gap between the lower bound and the true partition function). The experiments show that BPTRW outperforms mean-field by fairly large margins on densely connected graphs although it does worse on sparse graphs like chains.

Chain graphs On these graphs, mean-field does much better than BPTRW and BPTRW in fact gets worse as the number of nodes in the graph increase. Since TRW is exact on these graphs, the bounds show that both mean-field and BPTRW are fairly poor approximations in this case (Fig. 9).

Grid graphs On grid graphs, mean-field and BPTRW perform comparably with BPTRW doing slightly better most of the time (Fig. 10).

Complete graphs On complete graphs, BPTRWdoes much better than mean-field and gets better relative to mean-field as the number of nodes increases (Fig. 11).

Variance of potentials In this experiment, we generate potentials for a non-uniform Potts MRF of the following form:

$$P(\mathbf{x}) \propto \exp\left(\sum_{(s,t)\in E} \theta_{st} x_s x_t\right)$$

where $x_s \in \{-1, 1\}\ \forall s \in V$. Such a model is both binary and associative and hence we can use all the bounds on the partition function: BPTRW, TRWPW, TRW, PW and mean-field.

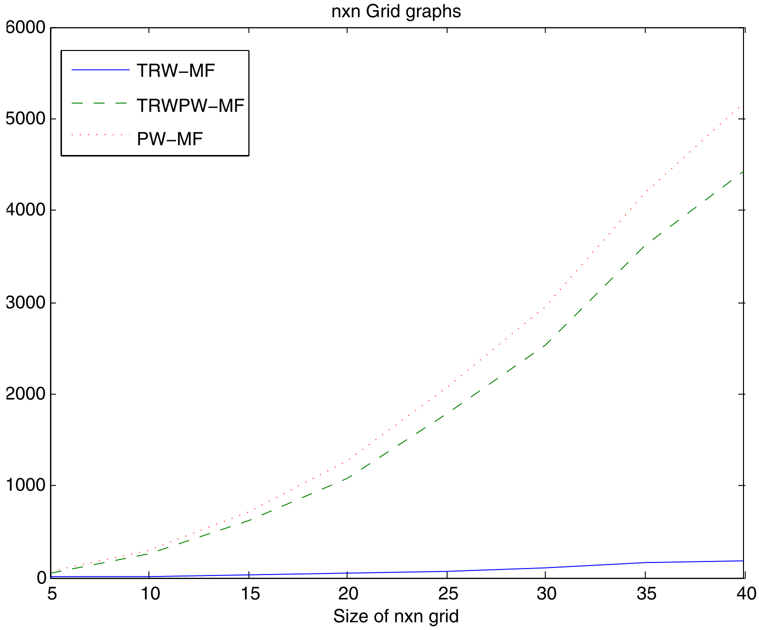

Fig. 10 TRW, BPTRW and mean-field on grid graphs

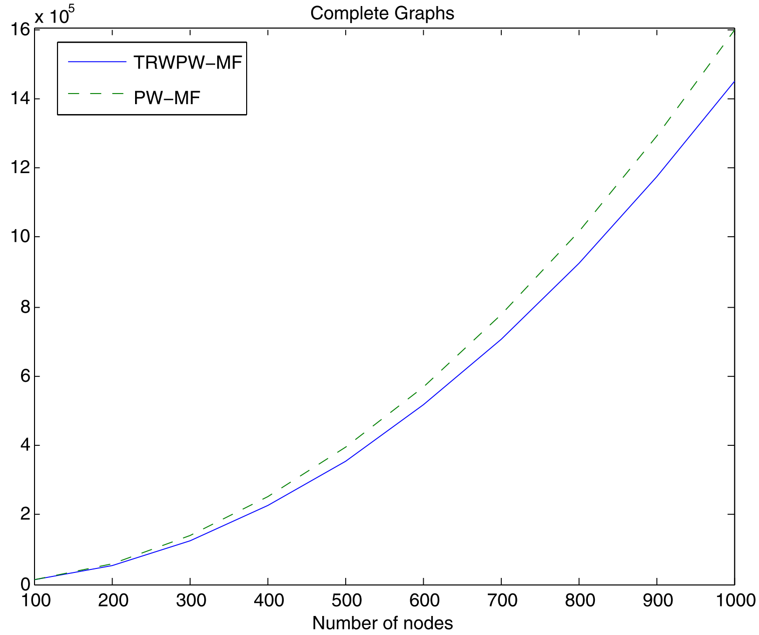

Fig. 11 TRW, BPTRW and mean-field on complete graphs

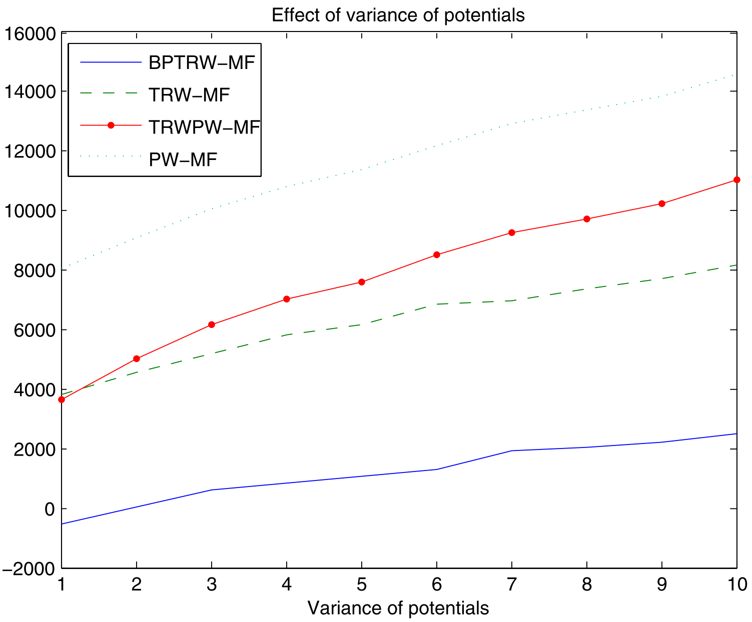

Fig. 12 Effect of variance of Θ on TRW, BPTRW, TRWPW, PW

Each θ_{st} is generated independently from a Gaussian distribution with mean zero. We use a complete graph of 100 nodes and examine the effect of the variance of the potential-generating Gaussian on the quality of the bounds. As the variance increases, we observe that BPTRW does better than mean-field. We also observe that TRWPW gets closer to PW than TRW as the variance increases from 1 to 10 (Fig. 12).

7 Conclusion

In this paper, we have proved convergence of the TRW-BP algorithm in one iteration for the case of Generalized Associative potentials. We have also developed closed-form upper bounds on the partition function for general pairwise MRFs. There are several important implications of these results: For the case of Generalized Associative potentials, we have seen that the pseudomarginals produced by the TRW-BP algorithm can be a bad approximation to the true marginals (Sect. 3). The closed-form of marginals can help to characterize when the TRW pseudomarginals are accurate. We have also observed that similar convergence results can be shown for loopy belief propagation (Sect. 5) and these results could again help in characterizing when loopy belief propagation gives good results. We have also developed closed-form lower bounds for the binary case: this may help in analyzing when the Bethe-variational bounds (Sudderth et al. 2008) outperforms other popular lower bounds like mean-field for the case of attractive associative potentials. We have developed the BP-TRW bounds as closed-form lower bounds for associative binary potentials: in doing this we utilized the closed-form solutions to TRW and the Bethe variational bound. It may be possible to generalize this approach to arbitrary pairwise binary fields by using a similar decomposition technique with a more complex optimization procedure. Another interesting area of possible further work is on computing bounds on the marginals and event probabilities. Ravikumar and Lafferty (2004) propose a framework through which variational bounds on the log partition function can be used to obtain bounds on marginal probabilities and general event probabilities. The closed-form bounds we have developed here (particularly for associative potentials) could be used to significantly speed up the complex optimization procedures currently required for computing these bounds (Ravikumar and Lafferty 2004) and perhaps even obtain closed-form bounds on event probabilities for some special cases. The closed-form bounds may also have implications on the optimization procedures for the tree edge appearance probabilities in TRW: it might be possible to exploit the existence of a closed-form function for a fixed ρ to develop better optimization procedures for obtaining the optimal upper bound. These closed-form bounds may also have significant impacts on problems arising in Bayesian inference: The form of the functions could be used to obtain fast(perhaps even closed-form) approximations/bounds on posterior likelihoods(using both upper and lower bounds) and outperform currently popular methods like variational Bayes.

References

Borwein, J., & Lewis, A. S. (2006). *Convex analysis and nonlinear optimization*. Berlin: Springer.

Boyd, S., & Vandenberghe, L. (2004). *Convex optimization*. Cambridge: Cambridge University Press.

Globerson, A., & Jaakkola, T. (2007). Convergent propagation algorithms via oriented trees. In *Proceedings of the twenty-second conference on uncertainty in AI (UAI)*, Vancouver, Canada, July 2007.

Jordan, M., & Wainwright, M. (2003). *Graphical models, exponential families and variational inference* (Technical Report 649). Department of Statistics, U.C. Berkeley.

Lange, K. (2004). *Optimization*. Berlin: Springer.

Lu, Q., & Getoor, L. (2003). Link-based classification. In *ICML* (pp. 496–503).

Ravikumar, P., & Lafferty, J. (2004). Variational Chernoff bounds for graphical models. In *UAI conference*.

Rennie, J. D. M. (2005). A class of convex functions. http://people.csail.mit.edu/jrennie/writing, May 2005.

Sudderth, E., Wainwright, M., & Willsky, A. (2008). Loop series and Bethe variational bounds in attractive graphical models. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems 20* (pp. 1425–1432). Cambridge: MIT Press.

Sutton, C., & McCallum, D. (2005). Piecewise training for undirected models. In *Proceedings of the twenty-second conference on uncertainty in AI (UAI)*, Toronto, Canada, July 2005.

Wainwright, M. J. (2002). Stochastic processes on graphs with cycles: geometric and variational approaches. PhD thesis, Massachusetts Institute of Technology, Supervisors A. S. Willsky and T. S. Jaakkola.

Wainwright, M. J., Jaakkola, T. S., & Willsky, A. S. (2005). A new class of upper bounds on the log partition function. *IEEE Transactions on Information Theory*, *51*(7), 2313–2335.

Yedidia, J. S., Freeman, W. T., & Weiss, Y. I. (2000). Generalized belief propagation. In *NIPS* (pp. 689–695).