Understanding the Bethe Approximation: When and How can it go Wrong?

Adrian Weller
Columbia University
New York NY 10027
adrian@cs.columbia.edu

Kui Tang
Columbia University
New York NY 10027
kt2384@cs.columbia.edu

David Sontag
New York University
New York NY 10012
dsontag@cs.nyu.edu

Tony Jebara
Columbia University
New York NY 10027
jebara@cs.columbia.edu

Abstract

Belief propagation is a remarkably effective tool for inference, even when applied to networks with cycles. It may be viewed as a way to seek the minimum of the Bethe free energy, though with no convergence guarantee in general. A variational perspective shows that, compared to exact inference, this minimization employs two forms of approximation: (i) the true entropy is approximated by the Bethe entropy, and (ii) the minimization is performed over a relaxation of the marginal polytope termed the local polytope. Here we explore when and how the Bethe approximation can fail for binary pairwise models by examining each aspect of the approximation, deriving results both analytically and with new experimental methods.

1 INTRODUCTION

Graphical models are a central tool in machine learning. However, the task of inferring the marginal distribution of a subset of variables, termed *marginal inference*, is NP-hard (Cooper, 1990), even to approximate (Dagum and Luby, 1993), and the closely related problem of computing the normalizing partition function is #P-hard (Valiant, 1979). Hence, much work has focused on finding efficient approximate methods. The sum-product message-passing algorithm termed belief propagation is guaranteed to return exact solutions if the underlying topology is a tree. Further, when applied to models with cycles, known as loopy belief propagation (LBP), the method is popular and often strikingly accurate (McEliece et al., 1998; Murphy et al., 1999).

A variational perspective shows that the true partition function and marginal distributions may be obtained by minimizing the true free energy over the marginal polytope. The standard Bethe approximation instead minimizes the Bethe free energy, which incorporates the Bethe pairwise approximation to the true entropy, over a relaxed pseudo-marginal set termed the local polytope. A fascinating link to LBP was shown (Yedidia et al., 2001), in that fixed points of LBP correspond to stationary points of the Bethe free energy $\mathcal{F}$. Further, stable fixed points of LBP correspond to minima of $\mathcal{F}$ (Heskes, 2003). Werner (2010) demonstrated a further equivalence to stationary points of an alternate function on the space of homogeneous reparameterizations.

In general, LBP may converge only to a local optimum or not converge at all. Various sufficient conditions have been derived for the uniqueness of stationary points (Mooij and Kappen, 2007; Watanabe, 2011), though convergence is often still not guaranteed (Heskes, 2004). Convergent methods based on analyzing derivatives of the Bethe free energy (Welling and Teh, 2001) and double-loop techniques (Heskes et al., 2003) have been developed. Recently, algorithms have been devised that are guaranteed to return an approximately stationary point (Shin, 2012) or a point with value ϵ-close to the optimum (Weller and Jebara, 2013a).

However, there is still much to learn about when and why the Bethe approximation performs well or badly. We shall explore both aspects of the approximation in this paper. Interestingly, sometimes they have opposing effects such that together, the result is better than with just one (see §4 for an example). We shall examine minima of the Bethe free energy over three different polytopes: marginal, local and cycle (see §2 for definitions). For experiments, we explore two methods, dual decomposition and Frank-Wolfe, which may be of independent interest. To provide another benchmark and isolate the entropy component, we also examine the tree-reweighted (TRW) approximation (Wainwright et al., 2005). Sometimes we shall focus on models where all edges are *attractive*, that is neighboring variables are pulled toward the same value; in this case it is known that the Bethe approximation is a lower bound for the true partition function (Ruozzi, 2012).

Questions we shall address include:

- In attractive models, why does the Bethe approximation perform well for the partition function but, when local potentials are low and coupling high, poorly for

marginals?

- In models with both attractive and repulsive edges, for low couplings, the Bethe approximation performs much better than TRW, yet as coupling increases, this advantage disappears. Can this be repaired by tightening the relaxation of the marginal polytope?

- Does tightening the relaxation of the marginal polytope always improve the Bethe approximation? In particular, is this true for attractive models?

This paper is organized as follows. Notation and preliminary results are presented in §2. In §3-4 we derive instructive analytic results, first focusing on the simplest topology that is not a tree, i.e. a single cycle. Already we observe interesting effects from both the entropy and polytope approximations. For example, even for attractive models, the Bethe optimum may lie outside the marginal polytope and tightening the relaxation leads to a worse approximation to the partition function. In §5 we examine more densely connected topologies, demonstrating a dramatic phase transition in attractive models as a consequence of the entropy approximation that leads to poor singleton marginals. Experiments are described in §6, where we examine test cases. Conclusions are discussed in §7. Related work is discussed throughout the text. An Appendix with technical details and proofs is attached in the Supplement.

2 NOTATION AND PRELIMINARIES

Throughout this paper, we restrict attention to binary pairwise Markov random fields (MRFs). We consider a model with n variables $X_1, \ldots, X_n \in \mathbb{B} = \{0, 1\}$ and graph topology $(\mathcal{V}, \mathcal{E})$; that is $\mathcal{V}$ contains nodes $\{1, \ldots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains an edge for each pairwise relationship. Let $x = (x_1, \ldots, x_n)$ be a configuration of all the variables, and $\mathsf{N}(i)$ be the neighbors of i. Primarily we focus on models with no 'hard' constraints, i.e. $p(x) > 0 \ \forall x$, though many of our results extend to this case. We may reparameterize the potential functions (Wainwright and Jordan, 2008) and define the energy E such that $p(x) = \frac{e^{-E(x)}}{Z}$ with

$$E = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} \frac{W_{ij}}{2} \left[x_i x_j + (1 - x_i)(1 - x_j) \right]. \tag{1}$$

This form allows edge coupling weights W_{ij} to be varied independently of the singleton potentials θ_i. If $W_{ij} > 0$ then an edge is *attractive*, if $W_{ij} < 0$ then it is *repulsive*. If all edges are attractive, then the model is attractive. We write μ_{ij} for pairwise marginals and, collecting together the θ_i and W_{ij} potential terms into a vector θ, with a slight abuse of notation, sometimes write (1) as $E = -\theta \cdot \mu$.

2.1 FREE ENERGY, VARIATIONAL APPROACH

Given any joint probability distribution $q(x)$ over all variables, the (Gibbs) free energy is defined as $\mathcal{F}_G(q) = \mathbb{E}_q(E) - S(q)$, where $S(q)$ is the (Shannon) entropy of the distribution.

It is easily shown (Wainwright and Jordan, 2008) that $-\log Z(\theta) = \min_q \mathcal{F}_G$, with the optimum when $q = p(\theta)$, the true distribution. This optimization is to be performed over all valid probability distributions, that is over the *marginal polytope*. However, this problem is intractable due to the difficulty of both computing the exact entropy S, and characterizing the polytope (Deza and Laurent, 2009).

2.2 BETHE APPROXIMATION

The standard approach of minimizing the *Bethe free energy* $\mathcal{F}$ makes two approximations:

1. The entropy S is approximated by the *Bethe entropy*

$$S_B(\mu) = \sum_{(i,j) \in \mathcal{E}} S_{ij}(\mu_{ij}) + \sum_{i \in \mathcal{V}} (1 - d_i) S_i(\mu_i), \tag{2}$$

   where S_{ij} is the entropy of μ_{ij}, S_i is the entropy of the singleton distribution of X_i and $d_i = |\mathsf{N}(i)|$ is the degree of i; and

2. The marginal polytope is relaxed to the *local polytope*, where we require only local (pairwise) consistency, that is we deal with a *pseudo-marginal* vector q, that may not be globally consistent, which consists of $\{q_i = q(X_i = 1) \ \forall i \in \mathcal{V}, \mu_{ij} = q(x_i, x_j) \ \forall (i,j) \in \mathcal{E}\}$ subject to the constraints $q_i = \sum_{j \in \mathsf{N}(i)} \mu_{ij}, q_j = \sum_{i \in \mathsf{N}(j)} \mu_{ij} \ \forall i, j \in \mathcal{V}$.

In general, the Bethe entropy S_B is not concave and hence, the Bethe free energy $\mathcal{F} = E - S_B$ is not convex.

The global optimum of the Bethe free energy $\mathcal{F} = \mathbb{E}_q(E) - S_B(q)$ is achieved by minimizing $\mathcal{F}$ over the local polytope, with the *Bethe partition function* Z_B defined such that the global minimum obtained equals $-\log Z_B$.

The local polytope constraints imply that, given q_i and q_j,

$$\mu_{ij} = \begin{pmatrix} 1 + \xi_{ij} - q_i - q_j & q_j - \xi_{ij} \\ q_i - \xi_{ij} & \xi_{ij} \end{pmatrix} \tag{3}$$

for some $\xi_{ij} \in [0, \min(q_i, q_j)]$, where $\mu_{ij}(a, b) = q(X_i = a, X_j = b)$.

As in (Welling and Teh, 2001), one can solve for the Bethe optimal ξ_{ij} explicitly in terms of q_i and q_j by minimizing $\mathcal{F}$, leading to

$$\xi^*_{ij}(q_i, q_j) = \frac{1}{2\alpha_{ij}} \left(Q_{ij} - \sqrt{Q_{ij}^2 - 4\alpha_{ij}(1 + \alpha_{ij}) q_i q_j} \right), \tag{4}$$

where $\alpha_{ij} = e^{W_{ij}} - 1, Q_{ij} = 1 + \alpha_{ij}(q_i + q_j)$.

Thus, we may consider the Bethe approximation as minimizing $\mathcal{F}$ over $q = (q_1, \ldots, q_n) \in [0,1]^n$. Further, the derivatives are given by

$$\frac{\partial \mathcal{F}}{\partial q_i} = -\phi_i + \log \left[\frac{(1-q_i)^{d_i - 1}}{q_i^{d_i - 1}} \prod_{j \in \mathsf{N}(i)} \frac{(q_i - \xi_{ij}^*)}{(1 + \xi_{ij}^* - q_i - q_j)} \right], \tag{5}$$

where $\phi_i = \theta_i - \frac{1}{2} \sum_{j \in \mathsf{N}(i)} W_{ij}$.

2.3 TREE-REWEIGHTED APPROXIMATION

Our primary focus in this paper is on the Bethe approximation but we shall find it helpful to compare results to another form of approximate inference. The *tree-reweighted* (TRW) approach may be regarded as a family of variational methods, where first one selects a point from the *spanning tree polytope*, that is the convex hull of all spanning trees of the model, represented as a weighting for each edge. Given this selection, the corresponding TRW entropy is the weighted combination of entropies on each of the possible trees. This is then combined with the energy and optimized over the local polytope, similarly to the Bethe approximation. Hence it provides an interesting contrast to the Bethe method, allowing us to focus on the difference in the entropy approximation. An important feature of TRW is that its entropy is concave and always upper bounds the true entropy (neither property is true in general for the Bethe entropy). Hence minimizing the TRW free energy is a convex problem and yields an upper bound on the true partition function. Sometimes we shall consider the optimal upper bound, i.e. the lowest upper bound achievable over all possible selections from the spanning tree polytope.

2.4 CYCLE POLYTOPE

We shall consider an additional relaxation of the marginal polytope termed the *cycle polytope*. This inherits all constraints of the local polytope, hence is at least as tight, and in addition enforces consistency around any cycle. A polyhedral approach characterizes this by requiring the following *cycle inequalities* to be satisfied (Barahona, 1993; Deza and Laurent, 2009; Sontag, 2010) for all cycles C and every subset of edges $F \subseteq C$ with $|F|$ odd:

$$\sum_{(i,j) \in F} (\mu_{ij}(0,0) + \mu_{ij}(1,1)) + \sum_{(i,j) \in C \setminus F} (\mu_{ij}(1,0) + \mu_{ij}(0,1)) \geq 1. \tag{6}$$

Each cycle inequality describes a facet of the marginal polytope (Barahona and Mahjoub, 1986). It is typically easier to optimize over the cycle polytope than the marginal polytope, and earlier work has shown that results are often similar (Sontag and Jaakkola, 2007).

2.5 SYMMETRIC AND HOMOGENEOUS MRFS

For analytic tractability, we shall often focus on particular forms of MRFs. We say a MRF is *homogeneous* if all singleton potentials are equal, all edge potentials are equal, and its graph has just one vertex and edge orbit.[1]

A MRF is *symmetric* if it has no singleton potentials, hence flipping all variables $0 \leftrightarrow 1$ leaves the energy unchanged, and the true marginals for each variable are $(\frac{1}{2}, \frac{1}{2})$. For symmetric, planar binary pairwise MRFs, it is known that the cycle polytope is equal to the marginal polytope (Barahona and Mahjoub, 1986). Using (4) and (5), it is easy to show the following result.

Lemma 1. *The Bethe free energy of any symmetric MRF has a stationary point at* $q_i = \frac{1}{2}\ \forall i$.

We remark that this is *not* always a minimum (see §5).

2.6 DERIVATIVES AND MARGINALS

It is known that the derivatives of $\log Z$ with respect to the potentials are the marginals, and that this also holds for any convex free energy, where pseudo-marginals replace marginals if a polytope other than the marginal is used (Wainwright, 2006). Using Danskin's theorem (Bertsekas, 1995), this can be generalized as follows.

Lemma 2. *Let* $\hat{F} = E - \hat{S}(\mu)$ *be any free energy approximation,* X *be a compact space, and* $\hat{A} = -\min_{\mu \in X} \hat{F}$ *be the corresponding approximation to* $\log Z$.
If the $\arg\min$ *is unique at pseudo-marginals* τ,
then $\frac{\partial \hat{A}}{\partial \theta_i} = \tau_i(1), \frac{\partial \hat{A}}{\partial W_{ij}} = \tau_{ij}(0,0) + \tau_{ij}(1,1)$.
If the $\arg\min$ *is not unique then let* $Q(\theta)$ *be the set of* $\arg\min$*s; the directional derivative of* $\hat{A}$ *in direction* $\theta \leftarrow \theta + y$ *is given by* $\nabla_y \hat{A} = \max_{\tau \in Q(\theta)} \tau \cdot y$.

In the next Section we begin to apply these results to analyze the locations and values of the minima of the Bethe free energy.

3 HOMOGENEOUS CYCLES

Since the Bethe approximation is exact for models with no cycles, it is instructive first to consider the case of one cycle on n variables, which we write as C_n. Earlier analysis considered the perspective of belief updates (Weiss, 2000; Aji, 2000). Here we examine the Bethe free energy, which in this context is convex (Pakzad and Anantharam, 2002) with a unique optimum.[2] We consider symmetric models, initially analyzing the homogeneous case.

[1] This means there is a graph isomorphism mapping any edge to any other, and the same for any vertex.

[2] This follows by considering (2) and observing that $S_{ij} - S_i$ (conditional entropy) is concave over the local consistency constraints, hence by appropriate counting, the total Bethe entropy is concave provided an MRF has at most one cycle.

With Lemma 1, we see that singleton marginals are $\frac{1}{2}$ across all approximation methods. For pairwise marginals, the following result holds due to convexity.

Lemma 3. *For any symmetric MRF and a free energy that is convex, the optimum occurs at uniform pseudo-marginals across all pairs of variables, either where the derivative is zero or at an extreme point of the range.*

The uniformity of the optimal edge pseudo-marginals, together with Lemma 1, shows that all are $\mu_{ij} = \begin{pmatrix} x & \frac{1}{2}-x \\ \frac{1}{2}-x & x \end{pmatrix}$ $\forall (i,j) \in \mathcal{E}$, where just x remains to be identified. The optimum x with zero derivative is always contained within the local polytope but we shall see that this is not always the case when we consider the cycle relaxation. Using (4), it is straightforward to derive the following result for the Bethe pairwise marginals.

Lemma 4. *For a symmetric homogeneous cycle, the Bethe optimum over the local polytope is at* $x = x_B(W) = \frac{1}{2}\sigma(W/2)$, *where we use standard sigmoid* $\sigma(y) := \frac{1}{1+e^{-y}}$. *Observe that* $x_B(-W) = 1/2 - x_B(W)$.

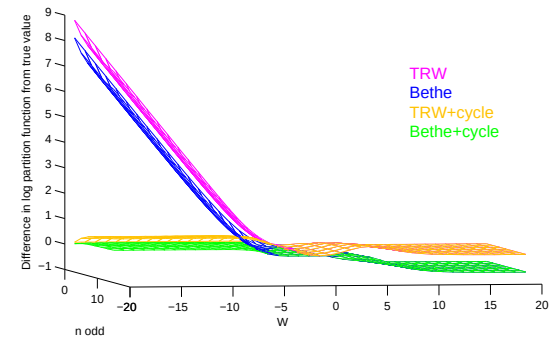
Further, we can derive the error of the Bethe pairwise marginals by using the loop series result given in Lemma 5 of §4, taking log, differentiating and using Lemma 2, to give the difference between true x and Bethe x_B as

$$x - x_B = \frac{1}{4} \frac{\operatorname{sech}^2 \frac{W}{4} \tanh^{n-1} \frac{W}{4}}{1 + \tanh^n \frac{W}{4}}. \tag{7}$$

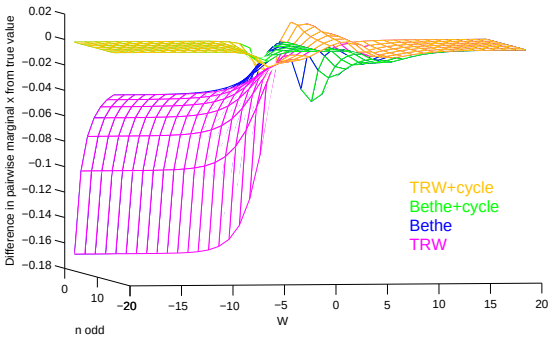
Remarks: Observe that at $W = 0$, $x - x_B = 0$; as $W \to \pm\infty$, $x - x_B \to 0$. For $W \neq 0$, $x - x_B$ is always > 0 unless n is even and $W < 0$, in which case it is negative. Differentiating (7) and solving for where x and x_B are most apart gives empirically $W \approx 2\log n + 0.9$ with corresponding max value of $x - x_B \approx \frac{1}{5n}$ for large n.

See Figure 1 for plots, where, for TRW, values were computed using optimal edge weights, as derived in the Appendix. Observe that at $W = 0$, all methods are exact. As W increases, the Bethe approximations to both $\log Z$ and the marginal x rise more slowly than the true values, though once W is high enough that x is large and cannot rise much further, then the Bethe x_B begins to catch up until they are both close to $\frac{1}{2}$ for large W. We remark that since the Bethe approximation is always a lower bound on the partition function for an attractive model (Ruozzi, 2012), and both the partition functions and marginals are equal at $W = 0$, we know from Lemma 2 that x_B must rise more slowly than x, as seen.

For $W > 0$, tightening the polytope makes no difference. The picture is different for negative W if n is odd, in which case we have a *frustrated cycle*, that is a cycle with an odd number of repulsive edges, which often causes difficulties with inference methods (Weller and Jebara, 2013b).



(a) Errors of $\log Z$ approximations



(b) Errors of pairwise marginal x

Figure 1: Homogeneous cycle C_n, n odd, edge weights W. By Lemma 2, the slope of the error of $\log Z$ wrt W is twice the error of x. For $W > 0$, local and cycle polytopes have the same values.

In this case, (6) is binding for $W < -2\log(n-1)$ and prevents the Bethe+cycle marginal x_{BC} from falling below $\frac{1}{2n}$. As $W \to -\infty$, the true x also does not fall below $\frac{1}{2n}$.[3] Thus, as $W \to -\infty$, the score (negative energy) and hence $\log Z \to -\infty$ for the true distribution. This also holds for Bethe or TRW on the cycle polytope, but on the local polytope, their energy and $\log Z \to 0$. Observe that for $W < 0$, Bethe generally outperforms TRW over both polytopes.

Tables 1 and 2 summarize results as $W \to \pm\infty$, again using optimal edge weights for TRW.

Model	$W \to -\infty$		$W \to \infty$	
	$\log Z'$	x	$\log \frac{Z'}{Z}$	x
Bethe	0	0	$-\log 2$	1/2
Bethe+cycle	0	0	$-\log 2$	1/2
TRW	$\log 2$	0	0	1/2
TRW+cycle	$\log 2$	0	0	1/2
True distribution	$\log 2$	0	0	1/2

Table 1: Analytic results for homogenous cycle C_n, n even. As $W \to \infty$, $\log Z'$ and $\log Z \to \infty$ so the difference is shown.

[3] To see this, note there are $2n$ configurations whose probabilities dominate as $W \to -\infty$: $01\ldots0$, its inverse flipping $0 \leftrightarrow 1$, and all n rotations; of these, just one has 00 and one has 11 for a specific edge.

Model	$W \to -\infty$		$W \to \infty$	
	$\log Z'$	x	$\log \frac{Z'}{Z}$	x
Bethe	0	0	$-\log 2$	1/2
Bethe+cycle	$-\infty$	$1/(2n)$	$-\log 2$	1/2
TRW	$\log 2$	0	0	1/2
TRW+cycle	$-\infty$	$1/(2n)$	0	1/2
True distribution	$-\infty$	$1/(2n)$	0	1/2

Table 2: Analytic results for homogeneous cycle C_n, n odd. As $W \to \infty, \log Z'$ and $\log Z \to \infty$ so the difference is shown.

4 NONHOMOGENEOUS CYCLES

The loop series method (Chertkov and Chernyak, 2006; Sudderth et al., 2007) provides a powerful tool to analyze the ratio of the true partition function to its Bethe approximation. In symmetric models with at most one cycle, by Lemma 3, we know that the unique Bethe optimum is at uniform marginals $q_i = \frac{1}{2}$. Using this and (4), and substituting into the loop series result yields the following.

Lemma 5. *For a symmetric MRF which includes exactly one cycle C_n, with edge weights $W_1, \ldots, W_n$, then $Z/Z_B = 1 + \prod_{i=1}^{n} \tanh \frac{W_i}{4}$.*

Remarks: In this setting, the ratio Z/Z_B is always ≤ 2 and ≈ 1 if even one cycle edge is weak, as might be expected since then the model is almost a tree. The ratio has no dependence on edges not in the cycle and those pairwise marginals will be exact. Further, since the Bethe entropy is concave, by Lemma 1, all singleton marginals are exact at $\frac{1}{2}$. Errors of pairwise pseudo-marginals on the cycle can be derived by using the expression for Z/Z_B from Lemma 5, taking $\log$ then differentiating and using Lemma 2.

Several principles are illustrated by considering 3 variables, A, B and C, connected in a triangle. Suppose AB and AC have strongly attractive edges with weight $W = 10$. We examine the effect of varying the weight of the third edge BC, see Figure 2.

It was recently proved (Ruozzi, 2012) that $Z_B \leq Z$ for attractive models. A natural conjecture is that the Bethe optimum pseudo-marginal in the local polytope must lie inside the marginal polytope. However, our example, when BC is weakly attractive, proves this conjecture to be false. As a consequence, tightening the local polytope to the marginal polytope for the Bethe free energy in this case worsens the approximation of the log-partition function (though it improves the marginals), see Figure 2 near 0 BC edge weight. For this model, the two aspects of the Bethe approximation to $\log Z$ act in opposing directions - the result is more accurate with both than with either one alone. For intuition, note that via the path $B-A-C$, in the globally consistent probability distribution, B and C are overwhelmingly likely to take the same value. Given that singleton marginals are $\frac{1}{2}$, the Bethe approximation, however, decomposes into a separate optimization for each edge, which for the weak edge BC, yields that B and C are almost independent, leading to a conflict with the true marginal. This causes the Bethe optimum over the local polytope to lie outside the marginal polytope. The same conclusion may be drawn rigorously by considering the cycle inequality (6), taking the edge set $F = \{BC\}$ and observing that the terms are approximately $\frac{1}{4} + \frac{1}{4} + 2(0 + 0) \approx \frac{1}{2} < 1$. Recall that here the cycle and marginal polytopes are the same (see §2.5). The same phenomenon can also be shown to occur for the TRW approximation with uniform edge appearance probabilities.

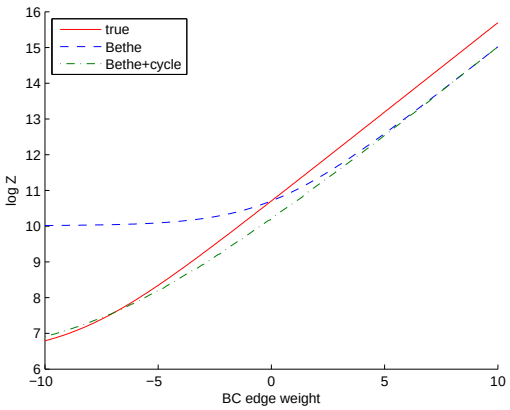


Figure 2: Log partition function and approximations for ABC triangle, see §4. Edge weights for AB and AC are 10 (strongly attractive) while BC is varied as shown. Near 0: Bethe is a better approximation to $\log Z$ but Bethe+cycle has better derivative, hence better marginals by Lemma 2; since Bethe+cycle is below Bethe in this region, its optimum does not lie in the local polytope.

Notice in Figure 2 that as the BC edge strength rises above 0, the Bethe marginals (given by the derivative) improve while the $\log Z$ approximation deteriorates. We remark that the exactness of the Bethe approximation on a tree can be very fragile in the sense that adding a very weak edge between variables to complete a cycle may expose that pairwise marginal as being (perhaps highly) inaccurate.

5 GENERAL HOMOGENEOUS GRAPHS

We discuss how the Bethe entropy approximation leads to a 'phase shift' in behavior for graphs with more than one cycle when W is above a positive threshold.

The true entropy is always maximized at $q_i = \frac{1}{2}$ for all variables. This also holds for the TRW approximation. However, in densely connected attractive models, the Bethe approximation pulls singleton marginals towards 0 or 1. This behavior has been discussed previously (Heskes, 2004; Mooij and Kappen, 2005) and described in terms of algorithmic stability (Wainwright and Jordan, 2008, §7.4), or heuristically as a result of LBP over-counting information when going around cycles (Ihler, 2007), but here we

explain it as a consequence of the Bethe entropy approximation.

We focus on symmetric homogeneous models which are *d-regular*, i.e. each node has the same degree d. One example is the complete graph on n variables, K_n. For this model, $d = n - 1$. The following result is proved in the Appendix, using properties of the Hessian from (Weller and Jebara, 2013a).

Lemma 6. *Consider a symmetric homogeneous MRF on n vertices with $d-$regular topology and edge weights W. $q = (\frac{1}{2}, \ldots, \frac{1}{2})$ is a stationary point of the Bethe free energy but for W above a critical value, this is not a minimum. Specifically, let H be the Hessian of the Bethe free energy at q, x_B be the value from Lemma 4 and $\mathbf{1}$ be the vector of length n with 1 in each dimension; then $\mathbf{1}^T H \mathbf{1} = n[d - 4x_B(d-1)]/x_B < 0$ if $x_B > \frac{1}{4}\frac{d}{d-1} \Leftrightarrow W > 2 \log \frac{d}{d-2}$.*

To help understand this result, consider (2) for the Bethe entropy S_B, and recall that $\sum_i d_i = 2m$ (m is the number of edges, handshake lemma), hence $S_B = mS_{ij} - (2m - n)S_i$. For large W, all the probability mass for each edge is pulled onto the main diagonal, thus $S_{ij} \approx S_i$. For $m > n$, which interestingly is exactly the case of more than one cycle, in order to achieve the optimum S_B, each entropy term $\to 0$ by tending to pairwise marginal $\begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}$ or symmetrically $\begin{pmatrix} 0 & 0 \\ 0 & 1 \end{pmatrix}$. See the second row of Figure 3 for an illustration of how the Bethe entropy surface changes dramatically as W rises, even sometimes going negative, and the top row to see how the Bethe free energy surfaces changes rapidly as W moves through the critical threshold.

Reinforcing this pull of singleton marginals away from $\frac{1}{2}$ is the shape of the energy surface, when optimized for free energy subject to given singleton marginals. In the Bethe approximation, this is achieved by computing ξ_{ij} terms according to (4), as illustrated in the bottom row of Figure 3, but for any reasonable entropy term (including TRW), always $\xi_{ij} < \min(q_i, q_j)$, hence the energy is lower towards the extreme values 0 or 1.

Remarks: (i) This effect is specifically due to the Bethe entropy approximation, and is not affected by tightening the polytope relaxation, as we shall see in §6. (ii) To help appreciate the consequences of Lemma 6, observe that $\log \frac{d}{d-2}$ is positive, monotonically decreasing to 0 as d increases. Thus, for larger, more densely connected topologies, the threshold for this effect is at lower positive edge weights. Above the threshold, $q_i = \frac{1}{2}$ is no longer a minimum but becomes a saddle point.[4] (iii) This explains the observation made after (Heinemann and Globerson, 2011,

[4]The Hessian at $q_i = \frac{1}{2}$ is neither positive nor negative definite. Moving away from the valley where all q_i are equal, the Bethe free energy rises quickly.

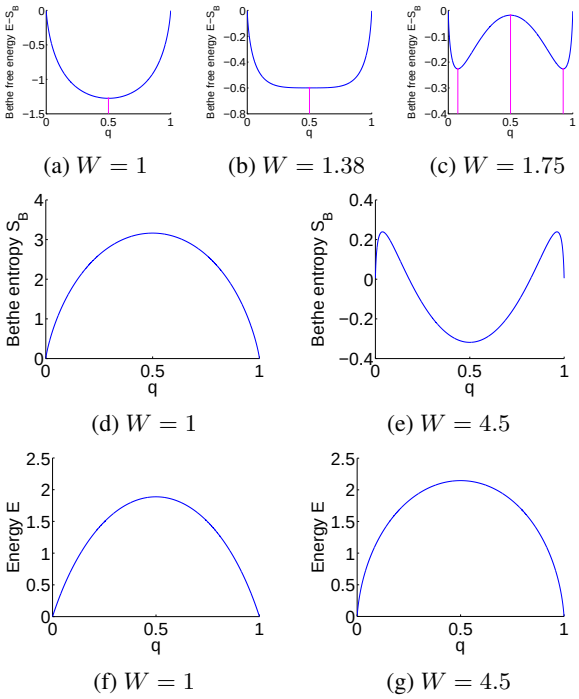


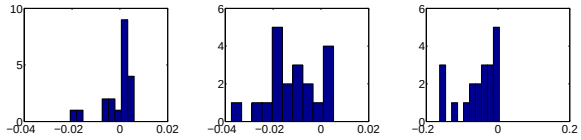
Figure 3: Bethe free energy $E - S_B$ with stationary points highlighted (top), then entropy S_B (middle) and energy E (bottom) vs $q_i = q \ \forall i$ for symmetric homogeneous complete graph K_5. **All quantities are evaluated at the optimum over pairwise marginals**, i.e. $\{\xi_{ij}\}$ are computed as in (4). These figures are described in Lemma 6 and the text thereafter. $W \approx 1.38$ is the critical threshold, above which Bethe singleton marginals are rapidly pulled toward 0 or 1. $W = 4.5$ is sufficiently high that the Bethe entropy becomes negative at $q = \frac{1}{2}$ (middle row).

Lemma 3), where it is pointed out that for an attractive model as $n \to \infty$, if $n/m \to 0$, a marginal distribution (other than the extreme of all 0 or all 1) is unlearnable by the Bethe approximation (because the effect we have described pushes all singleton marginals to 0 or 1). (iv) As W rises, although the Bethe singleton marginals can be poor, the Bethe partition function does not perform badly: For a symmetric model, as $W \to \infty$, there are 2 dominating MAP states (all 0 or all 1) with equal probability. The true marginals are at $q_i = \frac{1}{2}$ which picks up the benefit of $\log 2$ entropy, whereas the Bethe approximation converges to one or other of the MAP states with 0 entropy, hence has $\log 2$ error.

To see why a similar effect does not occur as $W \to -\infty$, note that for $W < 0$ around a frustrated cycle, the minimum energy solution on the local polytope is at $q_i = \frac{1}{2}$. Indeed, this can pull singleton Bethe marginals *toward* $\frac{1}{2}$ in this case. See §5.1 in the Appendix for further analysis.

6 EXPERIMENTS

We are interested in the empirical performance of the optimum Bethe marginals and partition function, as the re-



(a) local (b) cycle (c) marginal

Figure 4: Histogram of differences observed in optimum returned Bethe free energy, FW-mesh primal, over the 20 models in the validation set (mesh using $\epsilon = 0.1$, less than ϵ is insignificant). Negative numbers indicate FW outperformed mesh.

laxation of the marginal polytope is tightened. Many methods have been developed to attempt the optimization over the local polytope, primarily addressing its non-convexity, though none is guaranteed to return the global optimum. Recently, an algorithm was derived to return an ϵ-approximation to the optimum $\log Z_B$ based on constructing a discretized mesh of pseudo-marginals (Weller and Jebara, 2013a, 2014). One method for optimizing over tighter relaxations is to use this algorithm as an inner solver in an iterative dual decomposition approach with subgradient updates (Sontag, 2010; Sontag et al., 2011), where it can be shown that, when minimizing the Bethe free energy, the dual returned less ϵ lower bounds $-\log Z_B$ over the tighter polytope. This would be our preferred approach, but for the models on which we would like to run experiments, the runtime is prohibitive.

Hence we explored two other methods: (i) We replaced the inner solver with a faster, convergent double-loop method, the HAK-BETHE option in libDAI (Heskes et al., 2003; Mooij, 2010), though this is guaranteed only to return a local optimum at each iteration, hence we have no guarantee on the quality of the final result; (ii) We applied the Frank-Wolfe algorithm (FW) (Frank and Wolfe, 1956; Jaggi, 2013; Belanger et al., 2013). At each iteration, a tangent hyperplane is computed at the current point, then a move is made to the best computed point along the line to the vertex (of the appropriate polytope) with the optimum score on the hyperplane. This proceeds monotonically, even on a non-convex surface such as the Bethe free energy, hence will converge (since it is bounded), though runtime is guaranteed only for a convex surface as in TRW.

FW can be applied directly to optimize over marginal, cycle or local polytopes, and performed much better than HAK: runtime was orders of magnitude faster, and the energy found was in line with HAK.[5] To further justify using FW, which may only reach a local optimum, on our main test cases, we compared its performance on a small validation set against the benchmark of dual decomposition using the guaranteed ϵ-approximate mesh method (Weller and Jebara, 2014) as an inner solver.

[5]The average difference between energies found was < 0.1.

6.1 IMPLEMENTATION AND VALIDATION

To validate FW for the Bethe approximations on each polytope, we compared log partition functions and pairwise marginals across 20 MRFs, each on a complete graph with 5 variables. Each edge potential was drawn $W_{ij} \sim [-8, 8]$ and each singleton potential $\theta_i \sim [-2, 2]$. To handle the tighter polytope relaxations using the mesh method, we used a dual decomposition approach as follows. For the cycle polytope, one Lagrangian variable was introduced for each cycle constraint (6) with projected subgradient descent updates. For the marginal polytope, rather than imposing each facet constraint, which would quickly become unmanageable[6], instead a lift-and-project method was employed (Sontag, 2010). These algorithms may be of independent interest and are provided in the Supplement.

For all mesh runs, we used $\epsilon = 0.1$. Note that strong duality is not guaranteed for Bethe since the objective is non-convex, hence we are guaranteed only an upper bound on $\log Z_B$; yet we were able to monitor the duality gap by using rounded primals and observed that the realized gaps were typically within ϵ, see Figure 6.

For FW, we always initialized at the uniform distribution, i.e. $\mu_{ij} = \begin{pmatrix} \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} \end{pmatrix} \; \forall (i,j) \in \mathcal{E}$, note this is always within the marginal polytope. At each iteration, to determine how far to go along the line to the optimum vertex, we used Matlab's fminbnd function. This induces a minimum move of 10^{-6} along the line to the optimum vertex, which was helpful in escaping from local minima. When we tried allowing zero step size, performance became worse. Our stopping criterion was to run for $10,000$ iterations (which did not take long) or until the objective value changed by $< 10^{-6}$, at which point we output the best value found so far, and the corresponding pseudo-marginals.

Results on the validation set are shown in Figure 4, indicating that FW performed well compared to mesh + dual decomposition (the best standard we have for the Bethe optimum). Note, however, that good performance on $\log Z_B$ estimation does not necessarily imply that the Bethe optimal marginals were being returned for either method. There may be several local optima where the Bethe free energy has value close to the global optimum, and methods may return different locations. This is a feature of the non-convex surface which should be borne in mind when considering later results, hence we should not be surprised that in the validation set, although $17/20$ of the runs had ℓ_1 error in singleton marginals under 0.05, there were 3 runs with larger differences, in one case as high as 0.7 (not shown).[7]

[6]The number of facets of the marginal polytope grows extremely rapidly (Deza and Laurent, 2009).

[7]Recall the example from §5, where a symmetric homogeneous MRF with complete graph K_n topology and high edge

Given this performance, we used FW for all Bethe optimizations on the test cases. FW was also used for all TRW runs, where edge appearance probabilities were obtained using the matrix-tree theorem with weights proportional to each edge's coupling strength $|W_{ij}|$, as was used in (Sontag and Jaakkola, 2007).

6.2 TEST SETS

Models with 10 variables connected in a complete graph were drawn with random potentials. This allows comparison to earlier work such as (Sontag and Jaakkola, 2007) and (Meshi et al., 2009, Appendix). In addition to examining error in $\log$ partition functions and singleton marginals as was done in earlier work, given our theoretical observations in §3-5, we also explored the error in pairwise marginals. To do this, we report the ℓ_1 error in the estimated probability that a pair of variables is equal, averaged over all edges, i.e. we report average ℓ_1 error of $\mu_{ij}(0,0) + \mu_{ij}(1,1)$. We used FW to minimize the Bethe and TRW free energies over each of the local, cycle and marginal polytopes. For each maximum coupling value used, 100 models were generated and results averaged for plotting. Given the theoretical observations of §3-5, we are interested in behavior both for attractive and general (non-attractive) models.

For general models, potentials were drawn for single variables $\theta_i \sim U[-2,2]$ and edges $W_{ij} \sim U[-y,y]$ where y was varied to observe the impact of coupling strength.[8] Results are shown in Figure 5. Tightening the relaxation of the polytope from local to cycle or marginal, dramatically improves both Bethe and TRW approximations on all measures, with little difference between the cycle or marginal polytopes. This confirms observations in (Sontag and Jaakkola, 2007).

The relative performance of Bethe compared to TRW depends on the criteria used. Looking at the error of singleton marginals, Bethe is better than TRW for low coupling strengths, but for high coupling strengths the methods perform equally well on the local polytope, whereas on the cycle or marginal polytopes, TRW outperforms Bethe (though Bethe is still competitive). Thus, tightening the relaxation of the local polytope at high coupling does not lead to Bethe being superior on all measures. However, in terms of partition function and pairwise marginals, which are important in many applications, Bethe does consistently outperform TRW in all settings, and over all polytopes.

For attractive models, in order to explore our observations in §5, much lower singleton potentials were used. We drew $\theta_i \sim U[-0.1, 0.1]$ and $W_{ij} \sim U[0,y]$ where y is varied. This is consistent with parameters used by Meshi et al. (2009). Results are shown in Figure 7. When coupling is high, the Bethe entropy approximation pushes singleton marginals away from $\frac{1}{2}$. This effect quickly becomes strong above a threshold. Hence, when singleton potentials are very low, i.e. true marginals are close to $\frac{1}{2}$, the Bethe approximation will perform poorly irrespective of polytope, as observed in our attractive experiments. We note, however, that this effect rarely causes singleton marginals to cross over to the other side of $\frac{1}{2}$. Further, as discussed in §5, the partition function approximation is not observed to deviate by more than $\log 2$ on average.

7 CONCLUSIONS

We have used analytic and empirical methods to explore the two aspects of the Bethe approximation: the polytope relaxation and the entropy approximation. We found Frank-Wolfe to be an effective method for optimization, and note that for the cycle polytope, the runtime of each iteration scales polynomially with the number of variables (see §6.1.3 in the Appendix for further details).

For general models with both attractive and repulsive edges, tightening the relaxation of the polytope from local to cycle or marginal, dramatically improves both Bethe and TRW approximations on all measures, with little difference between the cycle or marginal polytopes. For singleton marginals, except when coupling is low, there does not appear to be a significant advantage to solving the non-convex Bethe free energy formulation compared to convex variational approaches such as TRW. However, for log-partition function estimation, Bethe does provide significant benefits. Empirically, in both attractive and mixed models, Bethe pairwise marginals appear consistently better than TRW.

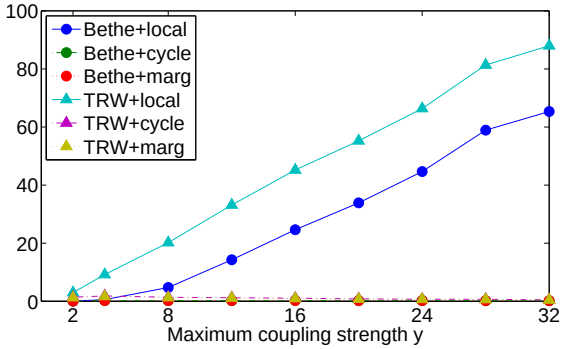
In our experiments with attractive models, the polytope approximation appears to makes little difference. However, we have shown theoretically that in some cases it can cause a significant effect. In particular, our discussion of non-homogeneous attractive cycles in §4 shows that even in the attractive setting, tightening the polytope can affect the Bethe approximation - improving marginals but worsening the partition function. It is possible that to observe this phenomenon empirically, one needs a different distribution over models.

Acknowledgements

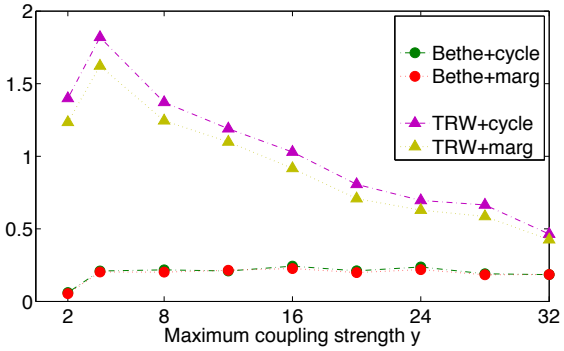
We thank U. Heinemann and A. Globerson for helpful correspondence. Work by A.W., K. T. and T.J. was supported in part by NSF grants IIS-1117631 and CCF-1302269. Work by D.S. was supported in part by the DARPA PPAML program under AFRL contract no. FA8750-14-C-0005.

weights was shown to have 2 locations at the global minimum, with average ℓ_1 distance between them approaching 1.

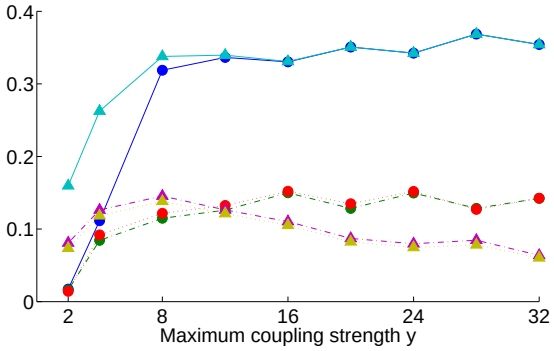
[8] These settings were chosen to facilitate comparison with the results of (Sontag and Jaakkola, 2007), though in that paper, variables take values in $\{-1,1\}$ so the equivalent singleton potential ranges coincide. To compare couplings, our y values should be divided by 4.



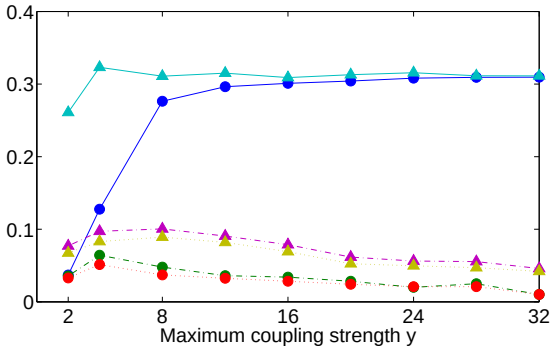
(a) log partition error



(b) log partition error, **local polytope removed**



(c) Singleton marginals, average ℓ_1 error



(d) Pairwise marginals, average ℓ_1 error

Figure 5: Results for general models showing error vs true values. $\theta_i \sim \mathbf{U}[-\mathbf{2}, \mathbf{2}]$. **The legend is consistent across plots.** These may be compared to plots in (Sontag and Jaakkola, 2007).

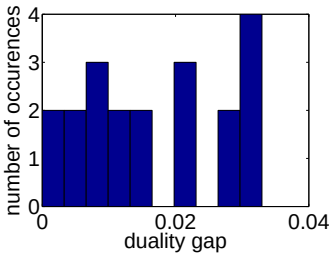
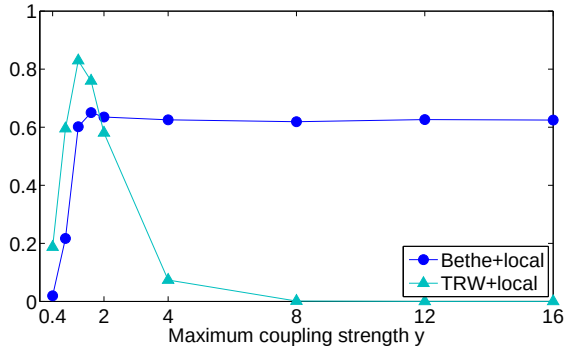
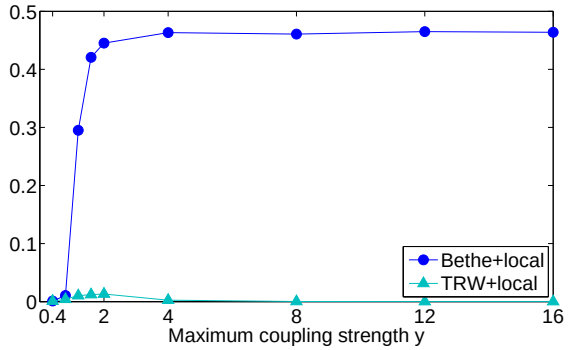


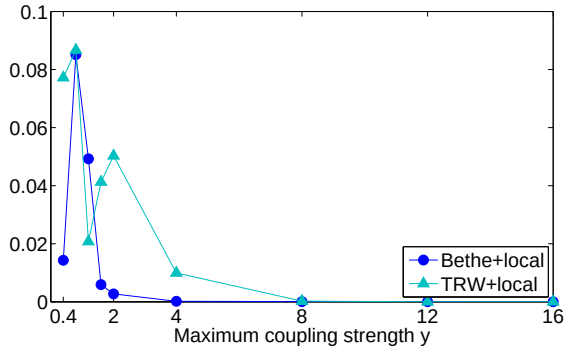
Figure 6: Duality gaps observed on the validation set using mesh approach + dual decomposition over 20 models, cycle polytope, $\epsilon = 0.1$. See text in §6.1



(a) log partition error



(b) Singleton marginals, average ℓ_1 error



(c) Pairwise marginals, average ℓ_1 error. **Note small scale.**

Figure 7: Results for attractive models showing error vs true values. $\theta_i \sim \mathbf{U}[-\mathbf{0.1}, \mathbf{0.1}]$. Only local polytope shown, **results for other polytopes are almost identical**.

References

S. Aji. *Graphical models and iterative decoding*. PhD thesis, California Institute of Technology, 2000.

F. Barahona. On cuts and matchings in planar graphs. *Math. Program.*, 60:53–68, 1993.

F. Barahona and A. Mahjoub. On the cut polytope. *Mathematical Programming*, 36(2):157–173, 1986. ISSN 0025-5610. doi: 10.1007/BF02592023.

D. Belanger, D. Sheldon, and A. McCallum. Marginal inference in MRFs using Frank-Wolfe. In *NIPS Workshop on Greedy Optimization, Frank-Wolfe and Friends*, December 2013.

D. Bertsekas. *Nonlinear Programming*. Athena Scientific, 1995.

S. Boyd and A. Mutapcic. Subgradient Methods, notes for EE364b, Jan 2007. `http://www.stanford.edu/class/ee364b/notes/subgrad_method_notes.pdf`, 2007.

M. Chertkov and M. Chernyak. Loop series for discrete statistical models on graphs. *J. Stat. Mech.*, 2006.

G. Cooper. The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42:393–405, 1990.

P. Dagum and M. Luby. Approximate probabilistic reasoning in Bayesian belief networks is NP-hard. *Artificial Intelligence*, 60:141–153, 1993.

M. Deza and M. Laurent. *Geometry of Cuts and Metrics*. Springer Publishing Company, Incorporated, 1st edition, 2009. ISBN 3642042945, 9783642042942.

M. Frank and P. Wolfe. An algorithm for quadratic programming. *Naval Research Logistics Quarterly*, 3(1-2):95–110, 1956. ISSN 1931-9193. doi: 10.1002/nav.3800030109.

U. Heinemann and A. Globerson. What cannot be learned with Bethe approximations. In *UAI*, pages 319–326, 2011.

T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Neural Information Processing Systems*, 2003.

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

T. Heskes, K. Albers, and B. Kappen. Approximate inference and constrained optimization. In *UAI*, pages 313–320, 2003.

A. Ihler. Accuracy bounds for belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*, 2007.

M. Jaggi. Revisiting Frank-Wolfe: Projection-free sparse convex optimization. In *ICML (1)*, pages 427–435, 2013.

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

O. Meshi, A. Jaimovich, A. Globerson, and N. Friedman. Convexifying the Bethe free energy. In *UAI*, pages 402–410, 2009.

J. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, August 2010. URL `http://www.jmlr.org/papers/volume11/mooij10a/mooij10a.pdf`.

J. Mooij and H. Kappen. On the properties of the Bethe approximation and loopy belief propagation on binary networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2005.

J. Mooij and H. Kappen. Sufficient conditions for convergence of the sum-product algorithm. *IEEE Transactions on Information Theory*, 53(12):4422–4437, December 2007.

K. Murphy, Y. Weiss, and M. Jordan. Loopy belief propagation for approximate inference: An empirical study. In *Uncertainty in Artificial Intelligence (UAI)*, 1999.

P. Pakzad and V. Anantharam. Belief propagation and statistical physics. In *Princeton University*, 2002.

N. Ruozzi. The Bethe partition function of log-supermodular graphical models. In *Neural Information Processing Systems*, 2012.

J. Shin. Complexity of Bethe approximation. In *Artificial Intelligence and Statistics*, 2012.

D. Sontag. *Approximate Inference in Graphical Models using LP Relaxations*. PhD thesis, Massachusetts Institute of Technology, Department of Electrical Engineering and Computer Science, 2010.

D. Sontag and T. Jaakkola. New outer bounds on the marginal polytope. In *NIPS*, 2007.

D. Sontag, A. Globerson, and T. Jaakkola. Introduction to dual decomposition for inference. In S. Sra, S. Nowozin, and S. Wright, editors, *Optimization for Machine Learning*. MIT Press, 2011.

E. Sudderth, M. Wainwright, and A. Willsky. Loop series and Bethe variational bounds in attractive graphical models. In *NIPS*, 2007.

L. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8(2):189–201, 1979.

M. Wainwright. Estimating the "wrong" graphical model: Benefits in the computation-limited setting. *Journal of Machine Learning Research*, 7:1829–1859, 2006.

M. Wainwright and M. Jordan. Graphical models, exponential families and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

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

Y. Watanabe. Uniqueness of belief propagation on signed graphs. In *Neural Information Processing Systems*, 2011.

Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12(1):1–41, 2000.

A. Weller and T. Jebara. Bethe bounds and approximating the global optimum. In *Artificial Intelligence and Statistics*, 2013a.

A. Weller and T. Jebara. On MAP inference by MWSS on perfect graphs. In *Uncertainty in Artificial Intelligence (UAI)*, 2013b.

A. Weller and T. Jebara. Approximating the Bethe partition function. In *Uncertainty in Atrificial Intelligence (UAI)*, 2014.

M. Welling and Y. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*, 2001.

T. Werner. Primal view on belief propagation. In *Uncertainty in Artificial Intelligence (UAI)*, 2010.

J. Yedidia, W. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *International Joint Conference on Artificial Intelligence, Distinguished Lecture Track*, 2001.