Approximating the Bethe partition function

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

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

Abstract

When belief propagation (BP) converges, it does so to a stationary point of the Bethe free energy $\mathcal{F}$, and is often strikingly accurate. However, it may converge only to a local optimum or may not converge at all. An algorithm was recently introduced for attractive binary pairwise MRFs which is guaranteed to return an ϵ-approximation to the global minimum of $\mathcal{F}$ in polynomial time provided the maximum degree $\Delta = O(\log n)$, where n is the number of variables. Here we significantly improve this algorithm and derive several results including a new approach based on analyzing first derivatives of $\mathcal{F}$, which leads to performance that is typically far superior and yields a fully polynomial-time approximation scheme (FPTAS) for attractive models without any degree restriction. Further, the method applies to general (non-attractive) models, though with no polynomial time guarantee in this case, leading to the important result that approximating $\log$ of the Bethe partition function, $\log Z_B = -\min \mathcal{F}$, for a general model to additive ϵ-accuracy may be reduced to a discrete MAP inference problem. We explore an application to predicting equipment failure on an urban power network and demonstrate that the Bethe approximation can perform well even when BP fails to converge.

1 INTRODUCTION

Undirected graphical models, also termed Markov random fields (MRFs), are flexible tools used in many areas including speech recognition, systems biology and computer vision. A set of variables and a score function is specified such that the probability of a configuration of variables is proportional to the value of the score function, which typically factorizes into sub-functions over subsets of variables in a way that defines a topology on the variables.

Three central problems are:

1. To evaluate the partition function Z, which is the sum of the score function over all possible settings, and hence is the normalization constant for the probability distribution.
2. Marginal inference, which is computing the probability distribution of a given subset of variables.
3. Maximum a posteriori (MAP) inference, which is the task of identifying a setting of all the variables which has maximum probability.

The first two problems are related (marginals are a ratio of two partition functions). Computing Z belongs to the class of counting problems #P (Valiant, 1979). Further, exact marginal inference is NP-hard (Cooper, 1990). The MAP problem is typically easier, yet is still NP-hard (Shimony, 1994), even to approximate (Abdelbar & Hedetniemi, 1998). Much work has focused on trying to find good approximate solutions, or restricted domains where exact solutions may be found efficiently. One popular method is to use a message-passing algorithm called belief propagation (Pearl, 1988), which returns an exact solution in linear time in n, the number of variables, if the topology of the model is a tree. If this method is applied to general topologies, termed loopy belief propagation (LBP), results are sometimes strikingly good (McEliece et al., 1998; Murphy et al., 1999), though in general it may not converge at all, and if it does, it may not be to a global optimum.

(Yedidia et al., 2001) showed a remarkable connection between LBP and an earlier approach from statistical physics (Bethe, 1935; Peierls & Born, 1936), in that any fixed point of LBP corresponds to a stationary point of a function of the system, termed the Bethe free energy $\mathcal{F}$. In fact, LBP can be seen as an iteration of the fixed point equations of the Bethe free energy. Variational approaches led to a better understanding of this relationship, showing that the negative of the global minimum of the Bethe free energy is the $\log$ of the Bethe partition function Z_B. Thus, Z_B should yield a good approximation to the true partition function Z, though this is not a formal result - there are cases where

Columbia University CUCS Technical Report.

it performs poorly, typically when there are many short cycles with strong edge interactions (Wainwright & Jordan, 2008, § 4.1). Even then, however, it can still be remarkably effective and in practice, LBP is widely used, often with excellent results. One motivation for our algorithm is to allow exploration of the limits for when Z_B performs well, even when LBP or other local optimization approaches fail, which has not previously been possible. We demonstrate this application in Experiments §6.

Another interesting example is the demonstration (Chandrasekaran et al., 2011) that the Bethe approximation is very useful to count independent sets of a graph. Further, it was shown that if the shortest cycle cover conjecture of Alon and Tarsi (Alon & Tarsi, 1985) is true, then the Bethe approximation is very good indeed for a random 3-regular graph.

Extensive analysis has focused on understanding conditions under which LBP is guaranteed to converge to the global optimum (Heskes, 2004; Mooij & Kappen, 2007; Watanabe, 2011), but outside these restricted settings, until recently, there were no polynomial time methods even to approximate Z_B. One major area of study is the important subclass of models which are *binary*, i.e. each variable takes one of just two possible values, and *pairwise*, i.e. all score sub-functions are evaluated over at most two variables. These play a key role in areas such as computer vision, both directly and as critical subroutines in solving more complex problems (Pletscher & Kohli, 2012). Further, it is possible to convert a general MRF into an equivalent binary pairwise model (Yedidia et al., 2001), though potentially with a much enlarged state space.

An algorithm was introduced in (Shin, 2012) guaranteed to return an approximately stationary point of $\mathcal{F}$ in polynomial time for such binary pairwise models, though with a bound on the maximum degree, $\Delta = O(\log n)$. (Weller & Jebara, 2013a) then used a discretizing approach to derive a polynomial-time approximation scheme (PTAS) for $\log Z_B$ for the significant subclass of *attractive*[1] binary pairwise models, also with $\Delta = O(\log n)$. Interestingly, (Ruozzi, 2012) recently proved that $Z_B \leq Z$ for attractive models. Similarly, for graphical models whose partition function is the permanent of a non-negative matrix, Z_B is recoverable via convex optimization and, here too, $Z_B \leq Z$ (Huang & Jebara, 2009; Vontobel, 2010; Watanabe & Chertkov, 2010; Gurvits, 2011). Otherwise, beyond trivial cases where the graph is acyclic, efficiently computing or approximating Z_B remains an active research topic.

[1] An *attractive* model has all pairwise relationships of the type that tend to pull adjacent variables toward the same value (see §2 for a more precise definition). Equivalent terms used are *associative*, *regular* or *ferromagnetic*.

1.1 Contribution and Summary

We obtain important new results for binary pairwise MRFs as described in the Abstract. We adopt ideas from (Weller & Jebara, 2013a) but go significantly further to derive much stronger results. The overall approach is to construct a *sufficient mesh* of discretized points in such a way that the optimum mesh point q^* is guaranteed to have $\mathcal{F}(q^*)$ within ϵ of the true optimum. The new, first derivative approach, generally results in a much coarser, yet still sufficient mesh, and also admits adaptive methods to focus points in regions where $\mathcal{F}$ may vary rapidly. Separately, we also refine the second derivative method of (Weller & Jebara, 2013a) to derive a method that performs well for very small ϵ. We then consider how best to solve the resulting discrete optimization problem, which may be framed as multi-label MAP inference, and for which many techniques are available, some of which are efficient for sub-classes of problem.[2]

In §2, we establish notation and present various preliminary results, then apply these in §3 to present our new approach for mesh construction based on analyzing first derivatives of $\mathcal{F}$. This leads to much improved performance (often by orders of magnitude), immediately admits general (non-attractive) models, and in the attractive setting yields a FPTAS for models with no restriction on topology.

In §4 we revisit the second derivative approach of (Weller & Jebara, 2013a). We show how this method can be refined and extended to yield better performance and also to admit non-attractive models, though for most cases of interest, unless ϵ is very small, the method of §3 will be superior.

In §5, we discuss the derived discrete optimization problem, which may be viewed as a multi-label MAP inference problem. In certain settings the problem is tractable, and in general we mention several features that can make it easier to find a satisfactory solution, or at least to bound its value. Experiments are described in §6 demonstrating practical application of the algorithm. Finally, we present conclusions in §7.

1.1.1 Structure of the overall algorithm

Input: Parameters $\{\theta_i, W_{ij}\}$ for a general binary pairwise MRF (convert format using the reparameterization of §2.1 if required), and a desired accuracy ϵ.

1. Preprocess by computing bounds $\{A_i, B_i\}$ on the locations of minima (see §2.4).
2. Construct a sufficient mesh using one of the methods in this paper. Indeed, all approaches are fast, so several may be used, then the most efficient mesh selected.

[2] Computing Z_B is at least PPAD or PLS-hard in general since it not only requires a fixed point but also the global minimizer (Shin, 2013; Daskalakis & Papadimitriou, 2011).

3. Attempt to solve the resulting multi-label MAP inference problem, see §5.
4. If unsuccessful, but a strongly persistent partial solution was obtained, then improved $\{A_i, B_i\}$ may be generated (see §5.2.1), repeat from 2.

At anytime, one may stop and compute bounds on $\mathcal{F}$, see §5.2.

1.2 Related work

Methods such as CCCP (Yuille, 2002) or UPS (Teh & Welling, 2002) are guaranteed to converge to a local minimum of the Bethe free energy, but this may be far from the global optimum. In earlier work, a fully polynomial-time randomized approximation scheme (FPRAS) for the true partition function was derived (Jerrum & Sinclair, 1993), but only when singleton potentials are uniform (i.e. a uniform external field) and the resulting runtime is high at $O(\epsilon^{-2}m^3n^{11}\log n)$. It was recently shown (Heinemann & Globerson, 2011) that models exist such that the true marginal probability cannot possibly be the location of a minimum of the Bethe free energy. Our work demonstrates an interesting connection between MAP inference techniques (NP-hard) and estimating the partition function Z (#P-hard). Recently (Hazan & Jaakkola, 2012) showed a different connection by using MAP inference on randomly perturbed models to approximate and bound Z.

2 NOTATION & PRELIMINARIES

Our notation is similar to (Weller & Jebara, 2013a) and (Welling & Teh, 2001). We focus on a binary pairwise model with n variables $X_1, \ldots, X_n \in \mathbb{B} = \{0, 1\}$ and graph topology $(\mathcal{V}, \mathcal{E})$ with $m = |\mathcal{E}|$; that is $\mathcal{V}$ contains nodes $\{1, \ldots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq V \times V$ contains an edge for each pairwise score relationship. Let $\mathsf{N}(i)$ be the neighbors of i. Let $x = (x_1, \ldots, x_n)$ be one particular configuration, and introduce the notion of *energy* $E(x)$ through[3]

$$p(x) = \frac{e^{-E(x)}}{Z},\ E = -\sum_{i\in\mathcal{V}} \theta_i x_i - \sum_{(i,j)\in\mathcal{E}} W_{ij}x_i x_j, \quad (1)$$

where the partition function $Z = \sum_x e^{-E(x)}$ is the normalizing constant.

Given any joint probability distribution $p(X_1, \ldots, X_n)$ over all variables, the (Gibbs) free energy is defined as $\mathcal{F}_G(p) = \mathbb{E}_p(E) - S(p)$, where $S(p)$ is the (Shannon) entropy of the distribution. Using variational methods, a remarkable result is easily shown (Wainwright & Jordan, 2008): minimizing $\mathcal{F}_G$ over the set of all globally valid distributions (termed the *marginal polytope*) yields a value of $-\log Z$, exactly at the true marginal distribution, given in (1).

Minimizing $\mathcal{F}_G$ is, however, computationally intractable, hence the approach of minimizing the Bethe free energy $\mathcal{F}$ makes two approximations: (i) the marginal polytope is relaxed to the *local polytope*, where we require only *local* consistency, that is we deal with a *pseudo-marginal* distribution q, which in our context may be considered $\{q_i = q(X_i = 1)\ \forall i \in \mathcal{V}, \mu_{ij} = q(x_i, x_j)\ \forall (i,j) \in \mathcal{E}\}$ subject to $q_i = \sum_j \mu_{ij}\ \forall i \in \mathcal{V}, j \in \mathsf{N}(i)$; and (ii) the entropy S is approximated by the Bethe entropy $S_B = \sum_{(i,j)\in\mathcal{E}} S_{ij} + \sum_{i\in\mathcal{V}}(1-d_i)S_i$, where S_{ij} is the entropy of μ_{ij}, S_i is the entropy of the singleton distribution and $d_i = |\mathsf{N}(i)|$ is the degree of i. We assume the model is connected so $d_i \geq 1\ \forall i$ (else each component may be analyzed independently), and take $x \log x = 0$ for $x = 0$. Hence, the global optimum of the Bethe free energy,

$$\begin{aligned}\mathcal{F}(q) &= \mathbb{E}_q(E) - S_B(q) \quad (2)\\ &= \sum_{(i,j)\in\mathcal{E}} -\big(W_{ij}\xi_{ij} + S_{ij}(q_i, q_j)\big)\\ &\quad + \sum_{i\in\mathcal{V}} \big(-\theta_i q_i + (z_i - 1)S_i(q_i)\big),\end{aligned}$$

is achieved by minimizing $\mathcal{F}$ over the local polytope, with Z_B defined s.t. the result obtained equals $-\log Z_B$. See (Wainwright & Jordan, 2008) for details.

Considering the local polytope, given q_i and q_j, we must have

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

for some $\xi_{ij} \in [0, \min(q_i, q_j)]$, where $\mu_{ij}(a,b) = q(X_i = a, X_j = b)$. Let $\alpha_{ij} = e^{W_{ij}} - 1$. $\alpha_{ij} = 0 \Leftrightarrow W_{ij} = 0$ may be assumed not to occur else the edge (i,j) may be deleted. α_{ij} has the same sign as W_{ij}, if positive then the edge (i,j) is *attractive*; if negative then the edge is *repulsive*. The MRF is attractive if all edges are attractive. As in (Welling & Teh, 2001), one can solve for ξ_{ij} explicitly in terms of q_i and q_j by minimizing $\mathcal{F}$, leading to a quadratic equation with real roots,

$$\alpha_{ij}\xi_{ij}^2 - [1 + \alpha_{ij}(q_i + q_j)]\xi_{ij} + (1 + \alpha_{ij})q_i q_j = 0. \quad (4)$$

For $\alpha_{ij} > 0$, $\xi_{ij}(q_i, q_j)$ is the lower root, for $\alpha_{ij} < 0$ it is the higher. Collecting the pairwise terms of $\mathcal{F}$ from (2) for one edge, define

$$f_{ij}(q_i, q_j) = -W_{ij}\xi_{ij}(q_i, q_j) - S_{ij}(q_i, q_j). \quad (5)$$

Thus we may consider the minimization of $\mathcal{F}$ over $q = (q_1, \ldots, q_n) \in [0,1]^n$.

[3] The probability or score function can always be reparameterized in this way, with finite θ_i and W_{ij} terms provided $p(x) > 0\ \forall x$, which is a requirement for our approach. There are reasonable distributions where this does not hold, i.e. distributions where $\exists x : p(x) = 0$, but this can often be handled by assigning such configurations a sufficiently small positive probability ϵ.

We are interested in *discretized pseudo-marginals* where for each q_i, we restrict its possible values to a discrete mesh $\mathcal{M}_i$ of points in $[0,1]$, which may be spaced unevenly. We allow $\mathcal{M}_i \neq \mathcal{M}_j$. Write $\mathcal{M}$ for the entire mesh. Let $N_i = |\mathcal{M}_i|$ and define $N = \sum_{i\in\mathcal{V}} N_i$ and $\Pi = \prod_{i\in\mathcal{V}} N_i$, the sum and product respectively of the number of mesh points in each dimension. Let $\hat{q}$ be the location of a global optimum of $\mathcal{F}$. We say that a mesh construction $\mathcal{M}(\epsilon)$ is *sufficient* if, given $\epsilon > 0$, it can be guaranteed that $\exists$ a mesh point $q^* \in \prod_{i\in\mathcal{V}} \mathcal{M}_i$ s.t. $\mathcal{F}(q^*) - \mathcal{F}(\hat{q}) \leq \epsilon$.

We shall make use of the standard sigmoid function, $\sigma(x) = 1/(1+\exp(-x))$ for various bounds.

2.1 Input model specification

Throughout this paper, we assume the reparameterization in (1) for all analysis, but a different specification is more natural for input models avoiding bias. We assume an input model is given with singleton terms θ_i as in (1), but with pairwise energy terms instead given by $-\frac{W_{ij}}{2}x_i x_j - \frac{W_{ij}}{2}(1-x_i)(1-x_j)$. With this format, varying W_{ij} simply alters the degree of push/pull between i and j, without also changing the probability that each variable will be 0 or 1, as is the case with the format of (1). We assume maximum possible values W and T are known with $|\theta_i| \leq T\ \forall i \in \mathcal{V}$, and $|W_{ij}| \leq W\ \forall (i,j) \in \mathcal{E}$. The required transformation to convert from input model to the format of (1), simply takes $\theta_i \leftarrow \theta_i - \sum_{j\in\mathsf{N}(i)} W_{ij}/2$, leaving W_{ij} unaffected.

2.2 Submodularity

In our context, a pairwise multi-label function on a set of ordered labels $X_{ij} = \{1,\ldots,K_i\} \times \{1,\ldots,K_j\}$ is *submodular* iff $\forall x, y \in X_{ij},\ f(x \wedge y) + f(x \vee y) \leq f(x) + f(y)$, where for $x = (x_1, x_2)$ and $y = (y_1, y_2)$, $(x \wedge y) = (\min(x_1, y_1), \min(x_2, y_2))$ and $(x \vee y) = (\max(x_1, y_1), \max(x_2, y_2))$. For binary variables, submodular energy is equivalent to being attractive.

The key property for us is that if all pairwise cost functions f_{ij} over $\mathcal{M}_i \times \mathcal{M}_j$ from (5) are submodular, then the global discretized optimum may be found efficiently using graph cuts (Schlesinger & Flach, 2006).

Theorem 1 (Submodularity for any discretization of an attractive model, (Weller & Jebara, 2013a) Theorem 8, (Korc et al., 2012))**.** *If a binary pairwise MRF is submodular on an edge (i,j), i.e. $W_{ij} > 0$, then the multi-label discretized MRF for any mesh $\mathcal{M}$ is submodular for that edge. In particular, if the MRF is fully attractive, i.e. $W_{ij} > 0\ \forall (i,j) \in \mathcal{E}$, then the multi-label discretized MRF is fully submodular for any discretization. Proof in (Weller & Jebara, 2013a) .*

2.3 Flipping variables

As in (Weller & Jebara, 2013a) , we use the techniques below for flipping variables, i.e. we can consider a new model with variables $\{X'_i\}$, where $X'_i = 1 - X_i$ for some selection of i. Flipping a variable flips the parity of all its incident edges so attractive $\leftrightarrow$ repulsive. Flipping both ends of an edge leaves its parity unchanged.

2.3.1 Flipping all variables

Consider a new model with variables $\{X'_i = 1 - X_i, i = 1,\ldots,n\}$ and the same edges. Instead of θ_i and W_{ij} parameters, let those of the new model be θ'_i and W'_{ij}. Identify values such that the energies of all states are maintained up to a constant[4]:

$$\begin{aligned} E &= -\sum_{i\in\mathcal{V}} \theta_i X_i - \sum_{(i,j)\in\mathcal{E}} W_{ij} X_i X_j \\ &= const - \sum_{i\in\mathcal{V}} \theta'_i (1 - X_i) - \sum_{(i,j)\in\mathcal{E}} W'_{ij}(1-X_i)(1-X_j). \end{aligned}$$

Matching coefficients gives

$$W'_{ij} = W_{ij},\ \theta'_i = -\theta_i - \sum_{j\in\mathsf{N}(i)} W_{ij}. \tag{6}$$

If the original model was attractive, so too is the new.

2.3.2 Flipping some variables

Sometimes it is helpful to flip only a subset $\mathcal{R} \subseteq \mathcal{V}$ of the variables. This can be useful, for example, to make the model locally attractive around a variable, which can always be achieved by flipping just those neighbors to which it has a repulsive edge. Let $X'_i = 1 - X_i$ if $i \in \mathcal{R}$, else $X'_i = X_i$ for $i \in \mathcal{S}$, where $\mathcal{S} = \mathcal{V} \setminus \mathcal{R}$. Let $\mathcal{E}_t = \{$edges with exactly t ends in $\mathcal{R}\}$ for $t = 0, 1, 2$.

As in 2.3.1, solving for W'_{ij} and θ'_i such that energies are unchanged up to a constant,

$$W'_{ij} = \begin{cases} W_{ij} & (i,j) \in \mathcal{E}_0 \cup \mathcal{E}_2, \\ -W_{ij} & (i,j) \in \mathcal{E}_1 \end{cases}$$

$$\theta'_i = \begin{cases} \theta_i + \sum_{(i,j)\in\mathcal{E}_1} W_{ij} & i \in \mathcal{S}, \\ -\theta_i - \sum_{(i,j)\in\mathcal{E}_2} W_{ij} & i \in \mathcal{R}. \end{cases} \tag{7}$$

Lemma 2. *Flipping variables changes affected pseudo-marginal matrix entries' locations but not values. $\mathcal{F}$ is unchanged up to a constant, hence the locations of stationary points are unaffected. (Proof in (Weller & Jebara, 2013a))*

[4] Any constant difference will be absorbed into the partition function and leave probabilities unchanged.

2.4 Preliminary bounds

We use the following results from (Weller & Jebara, 2013a).

Lemma 3 ((Weller & Jebara, 2013a) Lemma 2)**.** $\alpha_{ij} \geq 0 \Rightarrow \xi_{ij} \geq q_i q_j, \alpha_{ij} \leq 0 \Rightarrow \xi_{ij} \leq q_i q_j$

Theorem 4 ((Weller & Jebara, 2013a) Theorem 4)**.** *For general edge types (associative or repulsive), let* $W_i = \sum_{j \in \mathsf{N}(i):W_{ij}>0} W_{ij}$, $V_i = -\sum_{j \in \mathsf{N}(i):W_{ij}<0} W_{ij}$. *At any stationary point of the Bethe free energy,* $\sigma(\theta_i - V_i) \leq q_i \leq \sigma(\theta_i + W_i)$.

For the efficiency of our overall approach, it is very desirable to tighten the bounds on locations of minima of $\mathcal{F}$ since this both reduces the search space and allows a lower density of discretizing points in our mesh. This may be achieved efficiently by running either of the following two algorithms: Bethe bound propagation (BBP) from (Weller & Jebara, 2013a), or using the approach from (Mooij & Kappen, 2007) which we term MK. Either method can achieve striking results quickly, though MK is our preferred method[5] - it considers cavity fields around each variable and determines the range of possible beliefs after iterating LBP, starting from any initial values; since any minimum of $\mathcal{F}$ corresponds to a fixed point of LBP (Yedidia et al., 2001), this bounds all minima.

Let the lower bounds obtained for q_i and $1-q_i$ respectively be A_i and B_i so that $A_i \leq q_i \leq 1 - B_i$, and let the *Bethe box* be the orthotope given by $\prod_{i \in \mathcal{V}}[A_i, 1 - B_i]$. Define $\eta_i = \min(A_i, B_i)$, i.e. the closest that q_i can come to the extreme values of 0 or 1.

Lemma 5 (Upper bound for ξ_{ij} for an attractive edge, (Weller & Jebara, 2013a) Lemma 6)**.** *If* $\alpha_{ij} > 0$, *then* $\xi_{ij} - q_i q_j \leq \frac{\alpha_{ij} m(1-M)}{1+\alpha_{ij}}$, *where* $m = \min(q_i, q_j)$ *and* $M = \max(q_i, q_j)$.

2.5 Derivatives of $\mathcal{F}$

In (Welling & Teh, 2001), first partial derivatives of the Bethe free energy are derived as

$$\frac{\partial \mathcal{F}}{\partial q_i} = -\theta_i + \log Q_i, \tag{8}$$

$$\text{where } Q_i = \frac{(1-q_i)^{d_i-1}}{q_i^{d_i-1}} \frac{\prod_{j \in \mathsf{N}(i)}(q_i - \xi_{ij})}{\prod_{j \in \mathsf{N}(i)}(1 + \xi_{ij} - q_i - q_j)}.$$

Theorem 6 (Second derivatives for each edge, (Weller & Jebara, 2013a) Theorem 7)**.** *For any edge* (i,j), *for any* α_{ij},

$$\frac{\partial^2 f_{ij}}{\partial q_i^2} = \frac{1}{T_{ij}} q_j(1-q_j), \quad \frac{\partial^2 f_{ij}}{\partial q_j^2} = \frac{1}{T_{ij}} q_i(1-q_i)$$

$$\frac{\partial^2 f_{ij}}{\partial q_i \partial q_j} = \frac{\partial^2 f_{ij}}{\partial q_j \partial q_i} = \frac{1}{T_{ij}}(q_i q_j - \xi_{ij}),$$

$$\text{where } T_{ij} = q_i q_j (1-q_i)(1-q_j) - (\xi_{ij} - q_i q_j)^2 \tag{9}$$

$$\geq 0 \text{ with equality iff } q_i \text{ or } q_j \in \{0, 1\}.$$

Incorporating all singleton terms gives the following result.

Theorem 7 (All terms of the Hessian, see (Weller & Jebara, 2013a) §4.3 and Lemma 9)**.** *Let* H *be the Hessian of* $\mathcal{F}$ *for a binary pairwise model, i.e.* $H_{ij} = \frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j}$, *and* d_i *be the degree of variable* X_i, *then*

$$H_{ii} = -\frac{d_i - 1}{q_i(1-q_i)} + \sum_{j \in \mathsf{N}(i)} \frac{q_j(1-q_j)}{T_{ij}} \geq \frac{1}{q_i(1-q_i)},$$

$$H_{ij} = \begin{cases} \frac{q_i q_j - \xi_{ij}}{T_{ij}} & (i,j) \in \mathcal{E} \\ 0 & (i,j) \notin \mathcal{E}, i \neq j. \end{cases}$$

3 NEW APPROACH

We develop a new approach to constructing a sufficient mesh $\mathcal{M}$ by analyzing bounds on the first derivatives of $\mathcal{F}$. This yields several attractive features:

- For attractive models, we obtain a FPTAS with worst case runtime $O(\epsilon^{-3} n^3 m^3 W^3)$ and no restriction on topology, as was required in (Weller & Jebara, 2013a).

- Our sufficient mesh is typically dramatically coarser than the earlier method of (Weller & Jebara, 2013a), leading to a much simpler subsequent MAP problem unless ϵ is very small. Here, the sum of the number of discretizing points in each dimension, $N = O\left(\frac{nmW}{\epsilon}\right)$. For comparison, the earlier method, even after our improvements in §4, forms a mesh with $N = O\left(\epsilon^{-1/2} n^{7/4} \Delta^{3/4} \exp\left[\frac{1}{2}(W(1+\Delta/2)+T)\right]\right)$. As an example, for the model in the experiments of §6, our new approach with the adaptive minsum method (see §3.1.2), yields a mesh with N that is 8 orders of magnitude smaller than the earlier method.

- Our approach immediately handles a general model with both attractive and repulsive edges. Hence approximating $\log Z_B$ may be reduced to a discrete multi-label MAP inference problem. This is valuable due to the availability of many MAP techniques. We discuss this in §5, where we consider when the MAP problem is tractable and examine approaches which may be tried in general.

First assume we have a model which is fully attractive around variable X_i, i.e. $W_{ij} > 0 \ \forall j \in \mathsf{N}(i)$. From (8) and Lemma 3, we obtain

$$\frac{\partial \mathcal{F}}{\partial q_i} = -\theta_i + \log Q_i \leq -\theta_i + \log \frac{q_i}{1-q_i}. \tag{10}$$

[5] Both BBP and MK are anytime methods that converge quickly, and can be implemented such that each iteration runs in $O(m)$ time. MK takes a little longer but can yield tighter bounds.

Flip all variables (see §2.3.1). Write $'$ for the parameters of the new flipped model, which is also fully attractive, then using (6) and (10),

$$\frac{\partial \mathcal{F}'}{\partial q_i'} \leq -\theta_i' + \log \frac{q_i'}{1-q_i'}$$
$$\Leftrightarrow -\theta_i - W_i + \log \frac{q_i}{1-q_i} \leq \frac{\partial \mathcal{F}}{\partial q_i}.$$

Combining this with (10) yields the sandwich result

$$-\theta_i - W_i + \log \frac{q_i}{1-q_i} \leq \frac{\partial \mathcal{F}}{\partial q_i} \leq -\theta_i + \log \frac{q_i}{1-q_i}.$$

Now generalize to consider the case that i has some neighbors $\mathcal{R}$ to which it is adjacent by repulsive edges. In this case, flip those nodes $\mathcal{R}$ (see §2.3.2) to yield a model, which we denote by $''$, which is fully attractive around i, hence we may apply the above result. By (7) we have $\theta_i'' = \theta_i - V_i$, and using $W_i'' = W_i + V_i$, we obtain that for a general model,

$$-\theta_i - W_i + \log \frac{q_i}{1-q_i} \leq \frac{\partial \mathcal{F}}{\partial q_i} \leq -\theta_i + V_i + \log \frac{q_i}{1-q_i}. \tag{11}$$

This bounds each first derivative $\frac{\partial \mathcal{F}}{\partial q_i}$ within a range of width $V_i + W_i = \sum_{j \in \mathsf{N}(i)} |W_{ij}|$, which will be sufficient for the main theoretical result to come in (15). We take the opportunity, however, to narrow this range, thereby improving the result in practice, by using just one step of the belief propagation algorithm (BBP) of (Weller & Jebara, 2013a).

Following the derivation of BBP in the Supplement of (Weller & Jebara, 2013a), where better bounds are derived on the q_i location of stationary points by taking account of $[A_j, 1-B_j]$ bounds on neighbors $j \in \mathsf{N}(i)$, we may refine the result of (11) to yield

$$f_i^L(q_i) \leq \frac{\partial \mathcal{F}}{\partial q_i} \leq f_i^U(q_i), \text{ where}$$
$$f_i^L(q_i) = -\theta_i - W_i + \log U_i + \log \frac{q_i}{1-q_i}$$
$$f_i^U(q_i) = -\theta_i + V_i - \log L_i + \log \frac{q_i}{1-q_i}. \tag{12}$$

L_i, U_i are each > 1 with $\log L_i + \log U_i \leq V_i + W_i$. They are computed as $L_i = \prod_{j \in \mathsf{N}(i)} L_{ij}$, $U_i = \prod_{j \in \mathsf{N}(i)} U_{ij}$, with $L_{ij} = \begin{cases} 1 + \frac{\alpha_{ij} A_j}{1+\alpha_{ij}(1-B_i)(1-A_j)} & \text{if } W_{ij} > 0 \\ 1 + \frac{\alpha_{ij} B_j}{1+\alpha_{ij}(1-B_i)(1-B_j)} & \text{if } W_{ij} < 0 \end{cases}$,
$U_{ij} = \begin{cases} 1 + \frac{\alpha_{ij} B_j}{1+\alpha_{ij}(1-A_i)(1-B_j)} & \text{if } W_{ij} > 0 \\ 1 + \frac{\alpha_{ij} A_j}{1+\alpha_{ij}(1-A_i)(1-A_j)} & \text{if } W_{ij} < 0 \end{cases}$.

See Figure 1 for an example. We make the following observations:

- The upper bound is equal to the lower bound plus the constant $D_i = V_i + W_i - \log L_i - \log U_i \geq 0$.

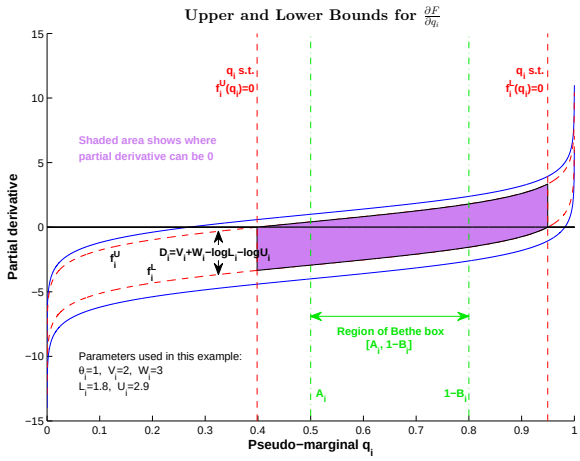

Figure 1: Upper and Lower Bounds for $\frac{\partial \mathcal{F}}{\partial q_i}$. Solid blue curves show worst case bounds (11) as functions of q_i, and are different by a constant $V_i + W_i = \sum_{j \in \mathsf{N}(i)} |W_{ij}|$. Dashed red curves show the upper $f_i^U(q_i)$ and lower $f_i^L(q_i)$ bounds (12) after being lowered by $\log L_i$ and raised by $\log U_i$ respectively, which incorporate the information from the bounds of neighboring variables. All bounding curves are strictly monotonic. The Bethe box region for q_i must lie within the shaded region demarcated by vertical red dashed lines, but we may have better bounds available, e.g. from MK, as shown by A_i and $1 - B_i$.

- The bound curves are monotonically increasing with q_i, ranging from $-\infty$ to $+\infty$ as q_i ranges from 0 to 1.

- A necessary condition to be within the Bethe box is that the upper bound is ≥ 0 and the lower bound is ≤ 0. Hence, anywhere within the Bethe box, we must have bounded derivative, $|\frac{\partial \mathcal{F}}{\partial q_i}| \leq D_i$. BBP generates $\{[A_i, 1-B_i]\}$ bounds by iteratively updating with L_i, U_i terms. In general, however, we may have better bounds from any other method, such as MK, which lead to higher L_i and U_i parameters and lower D_i.

$\mathcal{F}$ is continuous on $[0,1]^n$ and differentiable everywhere in $(0,1)^n$ with partial derivatives satisfying (12). $f_i^L(q_i)$ and $f_i^U(q_i)$ are continuous and integrable. Indeed, using the notation $\left[\phi(x)\right]_{x=a}^{x=b} = \phi(b) - \phi(a)$,

$$\int_a^b C + \log \frac{q_i}{1-q_i} dq_i = \Big[C q_i + q_i \log q_i + (1-q_i) \log(1-q_i) \Big]_{q_i=a}^{q_i=b} \tag{13}$$

for $0 \leq a \leq b \leq 1$, which relates to the binary entropy function $H(p) = -p \log p - (1-p) \log(1-p)$, recall the definition of $\mathcal{F}$. We remark that although $\frac{\partial \mathcal{F}}{\partial q_i}$ tends to $-\infty$ or $+\infty$ as q_i tends to 0 or 1, the integral converges (taking $0 \log 0 = 0$).

Hence if $\hat{q} = (\hat{q}_1, \ldots, \hat{q}_n)$ is the location of a global minimum, then for any $q = (q_1, \ldots, q_n)$ in the Bethe box,

$$\mathcal{F}(q) - \mathcal{F}(\hat{q}) \leq \sum_{i:\hat{q}_i \leq q_i} \int_{\hat{q}_i}^{q_i} f_i^U(q_i) dq_i + \sum_{i:q_i < \hat{q}_i} \int_{q_i}^{\hat{q}_i} -f_i^L(q_i) dq_i. \tag{14}$$

To construct a sufficient mesh, a simple initial bound relies on $|\frac{\partial \mathcal{F}}{\partial q_i}| \leq D_i$. If mesh points $\mathcal{M}_i$ are chosen s.t. in dimension i there must be a point q^* within γ_i of a global minimum (which can be achieved using a mesh width in each dimension of $2\gamma_i$), then by setting $\gamma_i = \frac{\epsilon}{nD_i}$, we obtain $\mathcal{F}(q^*) - \mathcal{F}(\hat{q}) \leq \sum_i D_i \frac{\epsilon}{nD_i} = \epsilon$. It is easily seen that $N_i \leq 1 + \lceil \frac{1}{2\gamma_i} \rceil$, hence the total number of mesh points, $N = \sum_{i \in \mathcal{V}} N_i$, satisfies

$$\begin{aligned} N &\leq 2n + \frac{n}{2\epsilon} \sum_i D_i \leq 2n + \frac{n}{\epsilon} \sum_{(i,j) \in \mathcal{E}} |W_{ij}| \\ &= O\left(\frac{n}{\epsilon} \sum_{(i,j) \in \mathcal{E}} |W_{ij}| \right) = O\left(\frac{nmW}{\epsilon} \right), \end{aligned} \tag{15}$$

since $D_i \leq V_i + W_i = \sum_{j \in \mathsf{N}(i)} |W_{ij}|$. Here $W = \max_{(i,j) \in \mathcal{E}} |W_{ij}|$ and $m = |\mathcal{E}|$ is the number of edges.

If the initial model is fully attractive, then by Theorem 1 we obtain a submodular multi-label MAP problem which is solvable using graph cuts with worst case runtime $O(N^3) = O(\epsilon^{-3} n^3 m^3 W^3)$ (Schlesinger & Flach, 2006; Greig et al., 1989; Goldberg & Tarjan, 1988).

Note from the first expression in (15) that if we have information on individual edge weights then we have a better bound using $\sum_{(i,j) \in \mathcal{E}} |W_{ij}|$ rather than just mW.

For comparison, the earlier second derivative approach of (Weller & Jebara, 2013a) has runtime $O(\epsilon^{-\frac{3}{2}} n^6 \Sigma^{\frac{3}{4}} \Omega^{\frac{3}{2}})$, where, even using the improved method in §4 here, $\Omega = O(\Delta e^{W(1+\Delta/2)+T})$. Unless ϵ is very small, the new first derivative approach is typically dramatically more efficient and more useful in practice. Further, it naturally handles both attractive and repulsive edge weights in the same way.

3.1 Refinements, adaptive methods

Since the resulting multi-label MAP inference problem is NP-hard in general (Shimony, 1994), it is helpful to minimize its size. As noted above, setting $\gamma_i = \frac{\epsilon}{nD_i}$, which we term the *simple method*, yields a sufficient mesh, where $|\frac{\partial \mathcal{F}}{\partial q_i}| \leq D_i = V_i + W_i - \log L_i - \log U_i$. However, since the bounding curves are monotonic with $f_i^U \geq 0$ and $f_i^L \leq 0$, a better bound for the magnitude of the derivative is often available by setting $D_i = \max\{f_i^U(1 - B_i), -f_i^L(A_i)\}$.

3.1.1 The minsum method

We define $N_i =$ the number of mesh points in dimension i, with sum $N = \sum_{i \in \mathcal{V}} N_i$ and product $\Pi = \prod_{i \in \mathcal{V}} N_i$.

For a fully attractive model, the resulting MAP problem may be solved in time $O(N^3)$ by graph cuts (Theorem 1, (Schlesinger & Flach, 2006; Greig et al., 1989; Goldberg & Tarjan, 1988)), so it is sensible to minimize N. In other cases, however, it is less clear what to minimize. For example, a brute force search over all points would take time $\Theta(\Pi)$.

Define the spread of possible values in dimension i as $S_i = 1 - B_i - A_i$ and note $N_i = 1 + \lceil \frac{S_i}{2\gamma_i} \rceil$ is required to cover the whole range. To minimize N while ensuring the mesh is sufficient, consider the Lagrangian $\mathcal{L} = \sum_{i \in \mathcal{V}} \frac{S_i}{2\gamma_i} - \lambda(\epsilon - \sum_{i \in V} \gamma_i D_i)$, where D_i is set as in the simple method (§3.1). Optimizing gives

$$\gamma_i = \frac{\epsilon}{\sum_{j \in \mathcal{V}} \sqrt{S_j D_j}} \sqrt{\frac{S_i}{D_i}}, \text{ with } N \leq 2n + \frac{1}{2\epsilon} \left(\sum_{i \in V} \sqrt{S_i D_i} \right)^2 \tag{16}$$

which we term the *minsum method*. Note $D_i \leq d_i W$ where d_i is the degree of X_i, hence $\left(\sum_{i \in V} \sqrt{S_i D_i}\right)^2 \leq W \left(\sum_{i \in V} \sqrt{d_i}\right)^2$. By Cauchy-Schwartz and the handshake lemma, $\left(\sum_{i \in V} \sqrt{d_i}\right)^2 \leq n \sum_{i \in \mathcal{V}} d_i = 2mn$, with equality iff the d_i are constant, i.e. the graph is regular.

If instead Π is minimized, rather than N, a similar argument shows that the simple method (§3.1) is optimal.

3.1.2 Adaptive methods

The previous methods rely on one bound D_i for $|\frac{\partial \mathcal{F}}{\partial q_i}|$ over the whole range $[A_i, 1 - B_i]$. However, we may increase efficiency by using local bounds to vary the mesh width across the range. A bound on the maximum magnitude of the derivative over any sub-range may be found by checking just $-f_i^L$ at the lower end and f_i^U at the upper end.

This may be improved by using the exact integral as in (14). First, constant proportions $k_i > 0$ should be chosen with $\sum_i k_i = 1$. Next, the first (lowest) mesh point $\gamma_1^i \in \mathcal{M}_i$ should be set s.t. $\int_{A_i}^{\gamma_1^i} f_i^U(q_i) dq_i = k_i \epsilon$. This will ensure that γ_1^i covers all points to its left in the sense that $\mathcal{F}[q_i = \gamma_1^i] - \mathcal{F}[q_i \in [A_i, \gamma_1^i]] \leq k_i \epsilon$ where all other variables $q_j, j \neq i$, are held constant at any values within the Bethe box. γ_1^i also covers all points to its right up to what we term its *reach*, i.e. the point r_1^i s.t. $\int_{\gamma_1^i}^{r_1^i} -f_i^L(q_i) dq_i = k_i \epsilon$. Next, γ_2^i is chosen as before, using r_1^i as the left extreme rather than A_i, and so on, until the final mesh point is computed with reach $\geq 1 - B_i$. This yields an optimal mesh for the choice of $\{k_i\}$.

If $k_i = \frac{1}{n}$, we achieve an optimized *adaptive simple* method. If $k_i = \frac{\sqrt{S_i D_i}}{\sum_{j \in \mathcal{V}} \sqrt{S_j D_j}}$, we achieve an *adaptive minsum* method. For many problems, this adaptive minsum method will be the most efficient.

Integrals are easily computed using (13). To our knowl-

edge, computing optimal points $\{\gamma_s^i\}$ is not possible analytically, but each may be found with high accuracy in just a few iterations using a search method, hence total time to compute the mesh is $O(N)$, which is negligible compared to solving the subsequent MAP problem.

4 REVISITING THE SECOND DERIVATIVE APPROACH

We review the second derivative approach used in (Weller & Jebara, 2013a) (see §5 there). As here, the possible location of a global minimum $\hat{q}$ was first bounded in the Bethe box given by $\prod_{i\in\mathcal{V}}[A_i, 1-B_i]$. Next an upper bound Λ was derived on the maximum possible eigenvalue of the Hessian H of $\mathcal{F}$ anywhere within the Bethe box, where it was required that all edges be attractive. Then a mesh of constant width in every dimension was introduced s.t. the nearest mesh point q^* to $\hat{q}$ was at most γ away in each dimension. Hence the ℓ_2 distance δ satisfies $\delta^2 \leq n\gamma^2$ and by Taylor's theorem, $F(q^*) \leq F(\hat{q}) + \frac{1}{2}\Lambda\delta^2$. Λ was computed by bounding the maximum magnitude of any element of H. Considering Theorem 7, this involves separate analysis of diagonal H_{ii} terms, which are positive and were bounded above by the term b; and edge H_{ij} terms, which are negative for attractive edges, whose magnitude was bounded above by a. Then Ω was set as $\max(a,b)$, and Σ as the proportion of non-zero entries in H. Finally, $\Lambda \leq \sqrt{\text{tr}(H^T H)} \leq \sqrt{\Sigma n^2 \Omega^2} = n\Omega\sqrt{\Sigma}$.

4.1 Improved bound for an attractive model

We improve the upper bound for Λ by improving the a bound for attractive edges to derive $\tilde{a}$, a better upper bound on $-H_{ij}$. Essentially, a more careful analysis allows a potentially small term in the numerator and denominator to be canceled before bounding. Writing $\bar{\eta} = \min_{i\in\mathcal{V}} \eta_i(1-\eta_i)$, i.e. the closest that any dimension can come to 0 or 1, the result is that

$$-H_{ij} \leq \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right) \Big/ \bar{\eta}\left(1-\left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2\right) \quad (17)$$
$$= O(e^{W(1+\Delta/2)+T}).$$

Thus, $\tilde{a} = O(e^{W(1+\Delta/2)+T})$ which compares favorably to the earlier bound in (Weller & Jebara, 2013a) , where $a = O(e^{W(1+\Delta)+2T})$. Recall $b = O(\Delta e^{W(1+\Delta/2)+T})$ and $\Omega = \max(a,b)$, so using the new $\tilde{a}$ bound, now $\Omega = O(\Delta e^{W(1+\Delta/2)+T})$. Details and derivation are in the supplement.

4.2 Extending the second derivative approach to a general (non-attractive) model

Using flipping arguments from §2.3, we are able to extend the method of (Weller & Jebara, 2013a) to apply to general models. Interestingly, the theoretical bounds derived for $\Omega = \max(a,b)$ take exactly the same form as for the purely attractive case, except that now $-W \leq W_{ij} \leq W$, whereas previously it was required that $0 \leq W_{ij} \leq W$. Since it is a second derivative approach, the mesh size (measured by N, the total number of points summed over the dimensions) grows as $O(\epsilon^{-1/2})$ rather than as $O(\epsilon^{-1})$ in the new first derivative approach. In practice, however, particularly for harder cases where n and W are above small values, unless ϵ is very small, the method of §3 is much more efficient. Details and derivations are in the supplement.

5 RESULTING MULTI-LABEL MAP

After computing a sufficient mesh, it remains to solve the multi-label MAP inference problem on a MRF with the same topology as the initial model, where each q_i takes values in $\mathcal{M}_i$. In general, this is NP-hard (Shimony, 1994).

5.1 Tractable cases

If it happens that all cost functions are submodular (as is always the case if the initial model is fully attractive by Theorem 1), then as already noted, it may be solved efficiently using graph cut methods, which rely on solving a max flow/min cut problem on a related graph, with worst case runtime $O(N^3)$ (Schlesinger & Flach, 2006; Greig et al., 1989; Goldberg & Tarjan, 1988). Using the Boykov-Kolmogorov algorithm (Boykov & Kolmogorov, 2004), performance is typically much faster, sometimes approaching $O(N)$. This submodular setting is the only known class of problem which is solvable for any topology.

Alternatively, the topological restriction of bounded tree-width allows tractable inference (Pearl, 1988). Further, under mild assumptions, this was shown to be the only restriction which will allow efficient inference for any cost functions (Chandrasekaran et al., 2008). We note that if the problem has bounded tree-width, then so too does the original binary pairwise model, hence exact inference (to yield the true marginals or the true partition function Z) on the original model is tractable, making our approximation result less interesting for this class. In contrast, although MAP inference is tractable for any attractive binary pairwise model, marginal inference and computing Z are not (Jerrum & Sinclair, 1993).

A recent approach reducing MAP inference to identifying a maximum weight stable set in a derived weighted graph ((Jebara, 2013), (Weller & Jebara, 2013b)) shows promise, allowing efficient inference if the derived graph is perfect. Further, testing if this graph is perfect can be performed in polynomial time ((Jebara, 2013), (Chudnovsky et al., 2005)).

5.2 All other cases

Many different methods are available, see (Kappes et al., 2013) for a recent survey. Some, such as dual approaches, may provide a helpful bound even if the optimum is not found. Indeed, a LP relaxation will run in polynomial time and return an upper bound on $\log Z_B$ that may be useful. A lower bound may be found from any discrete point, and this may be improved using local search methods. Note also that BBP bounds $q_i \in [A_i, 1 - B_i]$ apply *for all* the Bethe box, but for a particular value of q_i say, then the BBP approach provides tighter bounds on each of its neighbors $j \in \mathsf{N}(i)$, which may be helpful for pruning the solution space.

5.2.1 Persistent partial optimization approaches

MQPBO (Kohli et al., 2008) and Kovtun's method (Kovtun, 2003) are examples of this class. Both consider LP-relaxations and run in polynomial time. In our context, the output consists of ranges (which in the best case could be one point) of settings for some subset of the variables. If any such ranges are returned, the strong persistence property ensures that *any* MAP solution satisfies the ranges. Hence, these may be used to update $\{A_i, B_i\}$ bounds (padding the discretized range to the full continuous range covered by the end points if needed), compute a new, smaller, sufficient mesh and repeat until no improvement is obtained.

6 EXPERIMENTS

As a first step toward applying our algorithm to explore the usefulness of the global optimum of the Bethe approximation, here we consider one setting where LBP fails to converge, yet still we achieve reasonable results.

We aim to predict transformer failures in a power network (Rudin et al., 2012). Since the real data is sensitive, our experiments use synthetic data. Let $X_i \in \{0, 1\}$ indicate if transformer i has failed or not. Each transformer has a probability of failure on its own which is represented by a singleton potential θ_i. However, when connected in a network, a transformer can propagate its failure to nearby nodes (as in viral contagion) since the edges in the network form associative dependencies. We assume that homogeneous attractive pairwise potentials couple all transformers that are connected by an edge, i.e. $W_{ij} = W \; \forall (i, j) \in \mathcal{E}$. The network topology creates a Markov random field specifying the distribution $p(X_1, \ldots, X_n)$. Our goal is to compute the marginal probability of failure of each transformer within the network (not simply in isolation as in (Rudin et al., 2012)). Since recovering $p(X_i)$ is hard, we estimate Bethe pseudo-marginals $q_i = q(X_i = 1)$ through our algorithm, which emerge as the $\arg\min$ when optimizing the Bethe free energy.

A simulated sub-network of 55 connected transformers with average degree 2 was generated using a random preferential attachment model. Typical settings of $\theta_i = -2$ and $W = 4$ were specified (using the input model specification of §2.1). We attempted to run BP using the libDAI package (Mooij, 2010) but were unable to achieve convergence, even with multiple initial values, using various sequential or parallel settings and with damping. However, running our algorithm with $\epsilon = 1$ achieved reasonable results as shown in Table 1, where true values were obtained with the junction tree algorithm.

$\epsilon = 1$ PTAS for $\log Z_B$	Error vs true value
Mean ℓ_1 error of single marginals	0.003
Log-partition function	0.26

Table 1: Results on simulated power network

General folklore has suggested that the Bethe approximation is poor when BP fails to converge, thus this initial result suggests further work, which is now feasible using our algorithm.

7 DISCUSSION & FUTURE WORK

To our knowledge, we have derived the first ϵ-approximation algorithm for $\log Z_B$ for a general binary pairwise model. The approach is useful in practice, and much more efficient than the previous method of (Weller & Jebara, 2013a), though can take a long time to run for large, densely connected problems or when coupling is high. From experiments run, we note that the ϵ bounds appear to be close to tight since we have found models where the optimum returned when run with $\epsilon = 1$ is more than 0.5 different to that for $\epsilon = 0.1$. When applied to attractive models, we guarantee a FPTAS with no degree restriction.

Future work includes further improving the efficiency of the mesh, considering how it should be selected to simplify the subsequent discrete optimization problem, and exploring applications. Interesting avenues include using it as a subroutine in a dual decomposition approach to optimize over a tighter relaxation of the marginal polytope, and it provides the opportunity to examine rigorously the performance of other Bethe approaches that typically run more quickly, such as LBP or CCCP (Yuille, 2002), against the true Bethe global optimum.

Acknowledgments

We are grateful to Kui Tang for help with coding, and to David Sontag, Kui Tang, Nicholas Ruozzi and Tomaž Slivnik for helpful discussions. This material is based upon work supported by the National Science Foundation under Grant No. 1117631.

References

Abdelbar, A., & Hedetniemi, S. (1998). Approximating MAPs for belief networks is NP-hard and other theorems. *Artificial Intelligence, 102*, 21–38.

Alon, N., & Tarsi, M. (1985). Covering multigraphs by simple circuits. *SIAM Journal on Algebraic Discrete Methods, 6*, 345–350.

Bethe, H. (1935). Statistical theory of superlattices. *Proc. R. Soc. Lond. A, 150*, 552–575.

Boykov, Y., & Kolmogorov, V. (2004). An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell., 26*, 1124–1137.

Chandrasekaran, V., Chertkov, M., Gamarnik, D., Shah, D., & Shin, J. (2011). Counting independent sets using the Bethe approximation. *SIAM J. Discrete Math., 25*, 1012–1034.

Chandrasekaran, V., Srebro, N., & Harsha, P. (2008). Complexity of inference in graphical models. *UAI* (pp. 70–78). AUAI Press.

Chudnovsky, M., Cornuéjols, G., Liu, X., Seymour, P., & Vuskovic, K. (2005). Recognizing Berge graphs. *Combinatorica, 25*, 143–186.

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

Daskalakis, C., & Papadimitriou, C. (2011). Continuous local search. *Proceedings of ACM-SIAM Symposium on Discrete Algorithms (SODA)* (pp. 790–804).

Goldberg, A., & Tarjan, R. (1988). A new approach to the maximum flow problem. *Journal of the ACM, 35*, 921–940.

Greig, D., Porteous, B., & Seheult, A. (1989). Exact maximum a posteriori estimation for binary images. *J. Royal Statistical Soc., Series B, 51*, 271–279.

Gurvits, L. (2011). Unleashing the power of Schrijver's permanental inequality with the help of the Bethe approximation. *Elec. Coll. Comp. Compl.*

Hazan, T., & Jaakkola, T. (2012). On the partition function and random maximum a-posteriori perturbations. *ICML*.

Heinemann, U., & Globerson, A. (2011). What cannot be learned with Bethe approximations. *UAI* (pp. 319–326).

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

Huang, B., & Jebara, T. (2009). *Approximating the permanent with belief propagation* (Technical Report).

Jebara, T. (2013). *Tractability: Practical approaches to hard problems*, chapter Perfect graphs and graphical modeling. Cambridge Press.

Jerrum, M., & Sinclair, A. (1993). Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput., 22*, 1087–1116.

Kappes, J., Andres, B., Hamprecht, F., Schnörr, C., Nowozin, S., Batra, D., Kim, S., Kausler, B., Lellmann, J., Komodakis, N., & Rother, C. (2013). A comparative study of modern inference techniques for discrete energy minimization problems. *CVPR*.

Kohli, P., Shekhovtsov, A., Rother, C., Kolmogorov, V., & Torr, P. (2008). On partial optimality in multi-label MRFs. *ICML* (pp. 480–487). ACM.

Korc, F., Kolmogorov, V., & Lampert, C. (2012). *Approximating marginals using discrete energy minimization* (Technical Report). IST Austria.

Kovtun, I. (2003). Partial optimal labeling search for a NP-hard subclass of (max, +) problems. *DAGM-Symposium* (pp. 402–409). Springer.

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

Mooij, J. (2010). libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research, 11*, 2169–2173.

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

Murphy, K., Weiss, Y., & Jordan, M. (1999). Loopy belief propagation for approximate inference: An empirical study. *Proceedings of the Fifteenth conference on Uncertainty in artificial intelligence* (pp. 467–475).

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: Networks of plausible inference*. Morgan Kaufmann.

Peierls, R., & Born, M. (1936). On Ising's model of ferromagnetism. *Proc. Camb. Phil. Soc., 32*, 477.

Pletscher, P., & Kohli, P. (2012). Learning low-order models for enforcing high-order statistics. *Artificial Intelligence and Statistics*.

Rudin, C., Waltz, D., Anderson, R., Boulanger, A., Salleb-Aouissi, A., Chow, M., Dutta, H., Gross, P., Huang, B., & Ierome, S. (2012). Machine learning for the New York City power grid. *IEEE Trans. Pattern Anal. Mach. Intell.*, *34*, 328–345.

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

Schlesinger, D., & Flach, B. (2006). *Transforming an arbitrary minsum problem into a binary one* (Technical Report). Dresden University of Technology.

Shimony, S. (1994). Finding MAPs for belief networks is NP-hard. *Aritifical Intelligence*, *68*, 399–410.

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

Shin, J. (2013). The complexity of approximating a Bethe equilibrium. *CoRR, abs/1109.1724*.

Teh, Y., & Welling, M. (2002). The unified propagation and scaling algorithm. *Advances in Neural Information Processing Systems*.

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

Vontobel, P. (2010). The Bethe permanent of a non-negative matrix. *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on* (pp. 341–346).

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

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

Watanabe, Y., & Chertkov, M. (2010). Belief propagation and loop calculus for the permanent of a non-negative matrix. *Journal of Physics A: Mathematical and Theoretical*, *43*, 242002.

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

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

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

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

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

APPENDIX: SUPPLEMENTARY MATERIAL FOR APPROXIMATING THE BETHE PARTITION FUNCTION

Here we provide further details and proofs of several of the results in the main paper, using the original numbering.

4 REVISITING THE SECOND DERIVATIVE APPROACH

4.1 Improved bound for an attractive model

In this section, we improve the upper bound for Λ by improving the a bound for attractive edges to derive $\tilde{a}$, an improved upper bound on $-H_{ij}$. Essentially, a more careful analysis allows a potentially small term in the numerator and denominator to be canceled before bounding.

Using Theorem 7, equation (9) and Lemma 5,

$$\begin{aligned} -H_{ij} &= (\xi_{ij} - q_i q_j)\frac{1}{T_{ij}} \\ &\leq \frac{m(1-M)\alpha_{ij}}{1+\alpha_{ij}} \frac{1}{m(1-M)\left[(1-m)M - m(1-M)\left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2\right]} \\ &= \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right) \frac{1}{(1-m)M - m(1-M)\left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2} \end{aligned} \tag{18}$$

where $m = \min(q_i, q_j), M = \max(q_i, q_j)$. Now we use the following result.

Lemma 8. *For any $k \in (0,1)$, let $y = \min_{q_i \in [A_i, 1-B_i], q_j \in [A_j, 1-Bj]} (1-m)M - m(1-M)k$, then*

$$y = \begin{cases} B_i A_j - (1-B_i)(1-A_j)k & \text{if } (1-B_i) \leq A_j \quad \textit{i range} \leq \textit{j range} \\ (1-k)\min\{A_j(1-A_j), B_i(1-B_i)\} & \text{if } A_i \leq A_j \leq 1-B_i \leq 1-B_j \quad \textit{ranges overlap, i lower} \\ (1-k)\min\{A_j(1-A_j), B_j(1-B_j)\} & \text{if } A_i \leq A_j \leq 1-B_j \leq 1-B_i \quad \textit{j range} \subseteq \textit{i range} \\ (1-k)\min\{A_i(1-A_i), B_i(1-B_i)\} & \text{if } A_j \leq A_i \leq 1-B_i \leq 1-B_j \quad \textit{i range} \subseteq \textit{j range} \\ (1-k)\min\{A_i(1-A_i), B_j(1-B_j)\} & \text{if } A_j \leq A_i \leq 1-B_j \leq 1-B_i \quad \textit{ranges overlap, j lower} \\ B_j A_i - (1-B_j)(1-A_i)k & \text{if } (1-B_j) \leq A_i \quad \textit{j range} \leq \textit{i range.} \end{cases}$$

Proof. The minimum is achieved by minimizing the larger and maximizing the smaller of q_i and q_j. The result follows for cases where their ranges are disjoint. If ranges overlap, then the minimum is achieved at some $q_i = q_j$ in the overlap, with value $q_i(1-q_i)(1-k)$, which is concave and minimized at an extreme of the overlap range. □

Lemma 8 is useful in practice, and should be used to compute $\tilde{a} = \max_{(i,j)\in\mathcal{E}}$ of the bound above. To analyze the theoretical worst case, it is straightforward to see the corollary that $y \geq (1-k)\bar{\eta}$, where $\bar{\eta} = \min_{i\in\mathcal{V}} \eta_i(1-\eta_i)$. This bound can be met, for example, if all ranges coincide. Hence, from (18), and with the reasoning for $\frac{1}{\bar{\eta}}$ from (Weller & Jebara, 2013a) §5.3, where it is shown that $\frac{1}{\eta_i(1-\eta_i)} = O(e^{T+\Delta W/2})$, and using $\alpha_{ij} = e^{W_{ij}} - 1$, we obtain

$$-H_{ij} \leq \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right) \Big/ \bar{\eta}\left(1 - \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2\right) = O(e^{W(1+\Delta/2)+T}). \tag{19}$$

Thus, $\tilde{a} = O(e^{W(1+\Delta/2)+T})$ which compares favorably to the earlier bound in (Weller & Jebara, 2013a) , where $a = O(e^{W(1+\Delta)+2T})$. Recall $b = O(\Delta e^{W(1+\Delta/2)+T})$ and $\Omega = \max(a,b)$, so using the new $\tilde{a}$ bound, now $\Omega = O(\Delta e^{W(1+\Delta/2)+T})$.

4.2 Extending the second derivative approach to a general (non-attractive) model

Here we extend the analysis of (Weller & Jebara, 2013a) by considering repulsive edges to show that for a general binary pairwise model, we can still calculate useful bounds (which turn out to be very similar to the earlier bounds for attractive models) for a sufficient mesh width.

Our main tool for dealing with a repulsive edge is to flip the variable at one end (see §2.3) to yield an attractive edge, then we can apply earlier results. We denote the flipped model parameters with a $'$. For example, if just variable X_j is flipped, then $q'_j = q(X'_j = 1) = q(1 - X_j = 1) = 1 - q_j$. Since $\alpha_{ij} = e^{W_{ij}} - 1$ and here $W'_{ij} = -W_{ij}$, the following relationship holds if one end of an edge is flipped,

$$\frac{\alpha'_{ij}}{1 + \alpha'_{ij}} = \frac{e^{-W_{ij}} - 1}{e^{-W_{ij}}} = 1 - e^{W_{ij}} = -\alpha_{ij}. \tag{20}$$

Note that, for an attractive edge, $\frac{\alpha'_{ij}}{1+\alpha'_{ij}} \in (0, 1)$, as is $-\alpha_{ij}$ for a repulsive edge. Recall that when we flip some set of variables, by construction $\mathcal{F}' = \mathcal{F} + constant$ (see §2.3).

The Hessian terms from Theorem 7 still apply. Our goal is to bound the magnitude of each entry H_{ij} for a general binary pairwise model, then the earlier analysis will provide the result. Whereas for a fully attractive model, we assumed a maximum edge weight W with $0 \leq W_{ij} \leq W$, now we assume $|W_{ij}| \leq W$.

4.2.1 Edge terms

First consider H_{ij} for an edge $(i, j) \in \mathcal{E}$. If the edge is attractive, then the earlier analysis holds (it makes no difference if other edges are attractive or repulsive). If it is repulsive, then $H_{ij} > 0$. Consider a model where just X_j is flipped. $H_{ij} = \frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j} = -\frac{\partial^2 \mathcal{F}'}{\partial q'_i \partial q'_j} = -H'_{ij}$. Hence using (18) and (20), in practice an upper bound may be computed from Lemma 8 using $k = -\alpha_{ij}$ and $A'_j = B_j, B'_j = A_j$. The theoretical bound for an attractive edge from (19) becomes $H_{ij} \leq \frac{-\alpha_{ij}}{\bar{\eta}(1-\alpha_{ij}^2)}$. As we should expect from the attractive case, the following result holds.

Lemma 9. *For a repulsive edge,* $\frac{1}{1-\alpha_{ij}^2} = O(e^{-W_{ij}})$.

Proof. Let $u = -W_{ij}$, then $\alpha_{ij} = e^{-u} - 1$ and $\frac{1}{1-\alpha_{ij}^2} = \frac{1}{(1-\alpha_{ij})(1+\alpha_{ij})} = \frac{1}{e^{-u}(2-e^{-u})} = O(e^u)$. □

Hence, noting that we may flip any neighbors j of i which are adjacent via repulsive edges to obtain $\frac{1}{\eta_i(1-\eta_i)} = O(e^{T+\Delta W/2})$ as before, where now $W = \max_{(i,j)\in\mathcal{E}} |W_{ij}|$, we see that for our new second derivative method, just as in the fully attractive case, $\tilde{a} = O(e^{W(1+\Delta/2)+T})$.

For comparison interest, we also show how the earlier, worse bound for an attractive edge given in (Weller & Jebara, 2013a) may similarly be combined with flipping to provide a worse upper bound for H_{ij} when (i, j) is repulsive. See (Weller & Jebara, 2013a) §5.2: considering the proof of Lemma 10 and using (20) from this paper, we see that for a repulsive edge, the K_{ij} minimum bound for T_{ij} becomes $K_{ij} = \eta_i\eta_j(1 - \eta_i)(1 - \eta_j)(1 - \alpha_{ij}^2)$; then from (Weller & Jebara, 2013a) Theorem 11, the equivalent bound is $H_{ij} \leq \frac{-\alpha_{ij}}{4K_{ij}}$ which gives $a = O(e^{W(1+\Delta)+2T})$ as it was for the fully attractive case.

We provide a further interesting result, deriving a lower bound for ξ_{ij} for a repulsive edge.

Lemma 10 (Lower bound for ξ_{ij} for a repulsive edge, analogue of Lemma 5)**.** *For any repulsive edge* (i, j), $q_iq_j - \xi_{ij} \leq -\alpha_{ij}p_{ij}$ *where* $p_{ij} = \min\{q_iq_j, (1 - q_i)(1 - q_j)\}$.

Proof. Consider a model where just variable X_j is flipped, and let all new quantities be designated by the symbol $'$. Consider the joint pseudo-marginal (3). In the new model the columns are switched since $\mu'_{ij}(a, b) = q(X'_i = a, X'_j = b) = q(X_i = a, X_j = 1 - b) = \mu_{ij}(a, 1 - b)$, hence

$$\mu'_{ij} = \begin{pmatrix} 1 + \xi'_{ij} - q'_i - q'_j & q'_j - \xi'_{ij} \\ q'_i - \xi'_{ij} & \xi'_{ij} \end{pmatrix} = \begin{pmatrix} q_j - \xi_{ij} & 1 + \xi_{ij} - q_i - q_j \\ \xi_{ij} & q_i - \xi_{ij} \end{pmatrix}. \tag{21}$$

Applying Lemma 5 to the new model, $\xi'_{ij} - q'_iq'_j \leq \frac{\alpha'_{ij}}{1+\alpha'_{ij}} m'(1 - M')$. Substituting in $\xi'_{ij} = q_i - \xi_{ij}$ from (21) and using (20), we have $(q_i - \xi_{ij}) - q_i(1 - q_j) \leq -\alpha_{ij}m'(1 - M')$. Since $m' = \min\{q_i, 1 - q_j\}$ and $M' = \max\{q_i, 1 - q_j\}$, noting $q_i \leq 1 - q_j \Leftrightarrow q_i + q_j \leq 1 \Leftrightarrow q_iq_j \leq (1 - q_i)(1 - q_j)$, the result follows. □

Hence for a repulsive edge (i,j), using (9), we have

$$T_{ij} = q_i q_j (1-q_i)(1-q_j) - (\xi_{ij} - q_i q_j)^2 \geq p_{ij} P_{ij} - \alpha_{ij}^2 p_{ij}^2,$$

where $P_{ij} = \max\{q_i q_j, (1-q_i)(1-q_j)\}$.

4.2.2 Diagonal terms

Consider the H_{ii} terms from Theorem 7, which is true for a general model. If all neighbors of X_i are adjacent via attractive edges, then, as in (Weller & Jebara, 2013a) Theorem 11, $H_{ii} \leq \frac{1}{\eta_i(1-\eta_i)} \left(1 - d_i + \sum_{j \in \mathsf{N}(i)} \frac{1}{1 - \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2} \right)$.

If any neighbors are connected to X_i by a repulsive edge, then consider a new model where those neighbors are flipped, so now all edges incident to X_i are attractive, and designate the new model parameters with a $'$. As before, observe $\mathcal{F} = \mathcal{F}' + constant$, hence $H_{ii} = \frac{\partial^2 \mathcal{F}}{\partial q_i^2} = \frac{\partial^2 \mathcal{F}'}{\partial q_i'^2} = H'_{ii}$. Using (20) we obtain that for a general model,

$$H_{ii} \leq \frac{1}{\eta_i(1-\eta_i)} \left(1 - d_i + \sum_{j \in \mathsf{N}(i): W_{ij} > 0} \frac{1}{1 - \left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2} + \sum_{j \in \mathsf{N}(i): W_{ij} < 0} \frac{1}{1 - \alpha_{ij}^2} \right). \tag{22}$$

Similarly to the analysis in §4.2.1, using Lemma 9 gives that for a general model, $b = \max_{i \in \mathcal{V}} H_{ii} = O(\Delta e^{W(1+\Delta/2)+T})$, just as for a fully attractive model, where now $W = \max |W_{ij}|$.