Approximating the Bethe Partition Function

Adrian Weller
Department of Computer Science
Columbia University
New York NY 10027
adrian@cs.columbia.edu

Tony Jebara
Department of Computer Science
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 by Weller and Jebara 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 extend their approach and derive a new method based on analyzing first derivatives of $\mathcal{F}$, which leads to much better performance and, for attractive models, yields a fully polynomial-time approximation scheme (FPTAS) without any degree restriction. Further, our methods apply to general (non-attractive) models, though with no polynomial time guarantee in this case, demonstrating 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. This allows the merits of the global Bethe optimum to be tested.

1 INTRODUCTION

Undirected graphical models, also termed Markov random fields (MRFs), are central tools in machine learning. 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 factorizes into subfunctions 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; and (3) Maximum a posteriori (MAP) inference, which is the task of identifying a setting of all the variables which has maximum probability.

All these are NP-hard, and (1) and (2) are closely related (marginals are a ratio of two partition functions). Variational methods show that the partition function may be obtained by minimizing the free energy over the marginal polytope, and that if instead the Bethe free energy (Bethe, 1935) is minimized over the local polytope, this should yield a good approximation[1]. Although this is not a formal result, and there are cases where it performs poorly - typically when there are many short cycles with strong edge interactions (Wainwright and Jordan, 2008, § 4.1), still, the approach has proved very popular and often strikingly accurate. Belief propagation is often used to perform this minimization (Pearl, 1988; Yedidia et al., 2001). Performance is often excellent (McEliece et al., 1998; Murphy et al., 1999), but when applied to models with cycles, termed loopy belief propagation (LBP), convergence is not guaranteed in general, even to a local minimum. Some conjectured that when LBP behaves poorly, it is likely that the Bethe approximation, as given by the global minimum, also performs poorly, but it has not previously been possible to test this.

Approaches such as gradient descent (Welling and Teh, 2001), double-loop methods (Yuille, 2002) or Frank-Wolfe (Belanger et al., 2013) will converge but only to a local minimum, and with no runtime guarantee. Recently, two methods with polynomial runtime were given for the important subclass of binary pairwise models: one returns an approximately stationary point (Shin, 2012), though its value may be far even from a local minimum; the other returns an ϵ-approximate global optimum value (Weller and Jebara, 2013a) but only for the restricted case of attractive models (where pairwise relationships tend to pull connected variables to the same value)[1]. Both these methods restrict the topology to have maximum degree $O(\log n)$,

[1] All terms are defined in §2.

where n is the number of variables.

1.1 CONTRIBUTION AND SUMMARY

We obtain results for binary pairwise MRFs by expanding on ideas from Weller and Jebara (2013a). The 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. Our first derivative method typically results in a mesh that is much coarser (by many orders of magnitude, see §6.1), yet still sufficient, and admits adaptive methods to focus points in regions where $\mathcal{F}$ may vary rapidly. This leads to a FPTAS for attractive models with no restriction on topology. In addition, we refine and extend the second derivative approach of Weller and Jebara (2013a) to derive a method that performs well for very small ϵ. With our new methods, both approaches apply to general binary pairwise models (not necessarily attractive) to reduce the problem of finding an ϵ-approximate global optimum to solving a derived discrete optimization problem, which may be framed as multi-label MAP inference, where a rich family of methods already exists.

There are several motivations for this work:

- To our knowledge, we present the first way to solve for the global Bethe optimum (within ϵ accuracy) of a general binary pairwise MRF. Runtime is practical for small real-world problems.
- This now allows the accuracy of the global Bethe optimum to be tested.
- For attractive models, we obtain a fully polynomial time approximation scheme for any topology, thus answering an open theoretical question.

In §2, we establish notation and present preliminary results, then apply these in §3 to derive our new approach for mesh construction based on analyzing first derivatives of $\mathcal{F}$. In §4 we revisit the second derivative approach of Weller and 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 is much superior.

In §5, we discuss the resulting discrete optimization problem. In certain settings this 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. Conclusions are presented in §7.

For a sketch of the overall approach, see Algorithm, 1.

Algorithm 1 Mesh method to return ϵ-approximate global optimum $\log Z_B$ for a general binary pairwise model

```
Input: $\epsilon$, model parameters (convert using §2.1 if required)
Output: estimate of global optimum $\log Z_B$ guaranteed to be in range $[\log Z_B - \epsilon, \log Z_B]$, together with corresponding pseudo-marginal as $\arg$ for the discrete optimum
1: Preprocess by computing bounds on the locations of minima, see §2.4.
2: Construct a sufficient mesh using one of the methods in this paper, see §3 & 4. All approaches are fast, so several may be used and the most efficient mesh selected.
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 location bounds may be generated (see §5.2.1), repeat from 2.
At anytime, one may stop and compute bounds on $\log Z_B$, see §5.2.
```

1.2 RELATED WORK

Jerrum and Sinclair (1993) derived a fully polynomial-time randomized approximation scheme (FPRAS) for the true partition function, but only when singleton potentials are uniform (i.e. a uniform external field), and the runtime is high at $O(\epsilon^{-2}m^3n^{11}\log n)$. Heinemann and Globerson (2011) have shown that models exist such that the true marginal probability cannot possibly be the location of a minimum of the Bethe free energy. Approaches have been developed to solve related convex problems but results are typically less good (Meshi et al., 2009). Our work demonstrates an interesting connection between MAP inference techniques (NP-hard) and estimating the partition function Z (#P-hard). A different connection was shown by using MAP inference on randomly perturbed models to approximate and bound Z (Hazan and Jaakkola, 2012).

2 NOTATION & PRELIMINARIES

Our notation is similar to Weller and Jebara (2013a) and Welling and Teh (2001). We focus on a binary pairwise model with n variables $X_1, \dots, X_n \in \mathbb{B} = \{0, 1\}$ and graph topology $(\mathcal{V}, \mathcal{E})$ with $m = |\mathcal{E}|$; that is $\mathcal{V}$ contains nodes $\{1, \dots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq V \times V$ contains an edge for each pairwise score relationship. Let $\mathcal{N}(i)$ be the neighbors of i. Let $x = (x_1, \dots, x_n)$ be one particular configuration, and introduce the notion of *energy* $E(x)$ through[2]

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

[2] 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.

where the partition function $Z = \sum_x e^{-E(x)}$ is the normalizing constant, and $\{\theta_i, W_{ij}\}$ are parameters of the model.

Given any joint probability distribution $p(X_1, \dots, X_n)$ 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 and Jordan, 2008): minimizing $\mathcal{F}_G$ over the set of all globally valid distributions (termed the *marginal polytope*) yields a value of $-\log Z$ at the true marginal distribution, given in (1).

This minimization 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* vector 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 $q_i = \sum_{j \in \mathcal{N}(i)} \mu_{ij}, q_j = \sum_{i \in \mathcal{N}(j)} \mu_{ij}\ \forall i,j \in \mathcal{V}$; 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 $d_i = |\mathcal{N}(i)|$ is the degree of i. The local polytope constraints imply that, given q_i and q_j,

$$\mu_{ij} = \begin{pmatrix} 1 + \xi_{ij} - q_i - q_j & q_j - \xi_{ij} \\ q_i - \xi_{ij} & \xi_{ij} \end{pmatrix} \tag{2}$$

for some $\xi_{ij} \in [0, \min(q_i, q_j)]$, where $\mu_{ij}(a,b) = q(X_i = a, X_j = b)$. Hence, the global optimum of the Bethe free energy,

$$\begin{aligned} \mathcal{F}(q) &= \mathbb{E}_q(E) - S_B(q) \qquad (3) \\ &= \sum_{(i,j)\in\mathcal{E}} -\big(W_{ij}\xi_{ij} + S_{ij}(q_i, q_j)\big) \\ &\quad + \sum_{i\in\mathcal{V}} \big(-\theta_i q_i + (d_i - 1) S_i(q_i)\big), \end{aligned}$$

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 and Jordan, 2008) for details. 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 shown by Welling and 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 with real roots,

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

For $\alpha_{ij} > 0$, $\xi_{ij}(q_i, q_j)$ is the lower root, for $\alpha_{ij} < 0$ it is the higher. Thus we may consider the minimization of $\mathcal{F}$ over $q = (q_1, \dots, q_n) \in [0,1]^n$. Collecting the pairwise terms of $\mathcal{F}$ from (3) for one edge, define

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

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 $[0,1]$. The points may be spaced unevenly and we may have $\mathcal{M}_i \neq \mathcal{M}_j$. 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. Write $\mathcal{M}$ for the entire mesh. 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. $\mathcal{F}(q^*) - \mathcal{F}(\hat{q}) \le \epsilon$. The resulting discrete optimization problem may be framed as MAP inference in a multi-label MRF, where variable i takes values in $\mathcal{M}_i$, with the same topology (see §5).

2.1 INPUT MODEL SPECIFICATION

To be consistent with Welling and Teh (2001) and Weller and Jebara (2013a), for all theoretical analysis in this paper, we assume the reparameterization in (1). However, when an input model is specified, in order to avoid bias, we use singleton terms θ_i as in (1), but instead use pairwise energy terms given by $-\frac{W_{ij}}{2} x_i x_j - \frac{W_{ij}}{2}(1-x_i)(1-x_j)$. With this form, varying W_{ij} simply alters the degree of association between i and j. We assume maximum possible values W and T are known with $|\theta_i| \le T\ \forall i \in \mathcal{V}$, and $|W_{ij}| \le W\ \forall (i,j) \in \mathcal{E}$. The required transformation to convert from input model to the format of (1), simply takes $\theta_i \leftarrow \theta_i - \sum_{j\in\mathcal{N}(i)} W_{ij}/2$, leaving W_{ij} unaffected.

2.2 SUBMODULARITY

If all pairwise cost functions f_{ij} over $\mathcal{M}_i \times \mathcal{M}_j$ from (5) are submodular[3], then the global discretized optimum may be found efficiently using graph cuts (Schlesinger and Flach, 2006). We require the following earlier result.

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

2.3 FLIPPING VARIABLES

A useful technique for our analysis is to consider a model where some variables are flipped, i.e. given a model on $\{X_i\}$, consider a new model on $\{X'_i\}$ where $X'_i = 1 - X_i$ for some $i \in \mathcal{V}$. New model parameters $\{\theta'_i, W'_{ij}\}$ may be identified as in (Weller and Jebara, 2013a, §3) to preserve

[3] Here a pairwise multi-label function on a set of ordered labels $X_{ij} = \{1, \dots, K_i\} \times \{1, \dots, K_j\}$ is *submodular* iff $\forall x, y \in X_{ij}$, $f(x \wedge y) + f(x \vee y) \le f(x) + f(y)$, 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 this is equivalent to the edge potential being attractive.

energies of all states up to a constant. If all variables are flipped, new parameters are given by

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

If the original model was attractive, so too is the new. If only a subset $\mathcal{R} \subseteq \mathcal{V}$ is flipped, let $X'_i = 1 - X_i$ if $i \in \mathcal{R}$, else $X'_i = X_i$ for $i \in \mathcal{S}$, where $\mathcal{S} = \mathcal{V} \backslash \mathcal{R}$. Let $\mathcal{E}_t = \{$edges with exactly t ends in $\mathcal{R}\}$ for $t = 0, 1, 2$. Then we obtain

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

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

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 and Jebara (2013a)*

2.4 PRELIMINARY BOUNDS

We use the following earlier results.

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

Lemma 4 (Upper bound for ξ_{ij} for an attractive edge, Weller and 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 $M = \max(q_i, q_j)$.*

Theorem 5 (Weller and Jebara (2013a) Theorem 4)**.** *For general edge types (associative or repulsive), let $W_i = \sum_{j \in \mathcal{N}(i): W_{ij} > 0} W_{ij}$, $V_i = -\sum_{j \in \mathcal{N}(i): W_{ij} < 0} W_{ij}$. At any stationary point of the Bethe free energy, $\sigma(\theta_i - V_i) \leq q_i \leq \sigma(\theta_i + W_i)$, where $\sigma(x) = 1/(1 + \exp(-x))$ (sigmoid).*

For the efficiency of our overall approach, it is very desirable to tighten these bounds on locations of minima of $\mathcal{F}$ since this both reduces the search space and allows a lower density of discretizing points in the mesh. For our theoretical results, we do not assume this can be done but in practice, it can be attempted efficiently by running either of the following two algorithms: Bethe bound propagation (BBP) from (Weller and Jebara, 2013a, §6), or using the approach from Mooij and Kappen (2007) which we term MK. Either method can achieve striking results quickly, though MK is our preferred method[4] - this 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.

2.5 DERIVATIVES OF $\mathcal{F}$

Welling and Teh (2001) derived first partial derivatives of the Bethe free energy as

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

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

Weller and Jebara (2013a) derived all second partial derivatives.

Theorem 6 (All terms of the Hessian, see Weller and 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*

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

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

$$\begin{aligned} \textit{where } T_{ij} &= q_i q_j (1-q_i)(1-q_j) - (\xi_{ij} - q_i q_j)^2 \\ &\geq 0 \textit{ with equality iff } q_i \textit{ or } q_j \in \{0, 1\}. \end{aligned} \tag{9}$$

3 NEW APPROACH: GRADMESH

We develop a new approach to constructing a sufficient mesh $\mathcal{M}$ by analyzing bounds on the first derivatives of $\mathcal{F}$. To help distinguish between methods, we call the new first derivative approach *gradMesh*, and the earlier, second derivative approach *curvMesh*. The new gradMesh approach 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, unlike earlier work (Weller and Jebara, 2013a) which required max degree $\Delta = O(\log n)$.
- Our sufficient mesh is typically dramatically coarser than the earlier method of Weller and Jebara (2013a) unless ϵ is very small, leading to a much smaller subsequent MAP problem. 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 $N = O\left(\epsilon^{-1/2} n^{7/4} \Delta^{3/4} \exp\left[\frac{1}{2}(W(1 + \Delta/2) + T)\right]\right)$. See §6.1 for examples.

[4] 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.

- The 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, see §5.

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

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

Flip all variables (see §2.3). Write $'$ for the parameters of the new flipped model, which is also fully attractive, then using (6) and (10),

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

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) 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 $V_i + W_i = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$, which is sufficient for the main theoretical result, see (15). We take the opportunity, however, to describe a method which sometimes significantly narrows this range, thereby improving the result in practice.

Using one $O(m)$ iteration of the belief propagation algorithm (BBP) derived in (Weller and Jebara, 2013a, Supplement), allows us to refine the bounds for variable X_i of (11) based on the $[A_j, 1-B_j]$ location bounds on its neighbors $j \in \mathcal{N}(i)$, to show

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

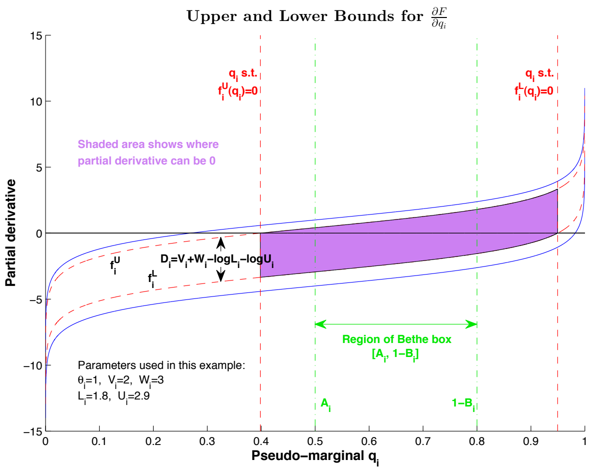

Figure 1: Upper and Lower Bounds for $\frac{\partial \mathcal{F}}{\partial q_i}$. Solid blue curves show worst case bounds (11) as functions of q_i, and are different by a constant $V_i + W_i = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$. Dashed red curves show the upper $f_i^U(q_i)$ and lower $f_i^L(q_i)$ 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 $1 - B_i$.

L_i, U_i are each > 1 with $\log L_i + \log U_i \leq V_i + W_i$. They are computed as $L_i = \prod_{j \in \mathcal{N}(i)} L_{ij}$, $U_i = \prod_{j \in \mathcal{N}(i)} U_{ij}$, 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 $D_i = V_i + W_i - \log L_i - \log U_i \geq 0$.

- 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 $\big[\phi(x)\big]_{x=a}^{x=b} = \phi(b) - \phi(a)$,

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

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 $\hat{q} = (\hat{q}_1, \dots, \hat{q}_n)$ is the location of a global minimum, then for any $q = (q_1, \dots, q_n)$ 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 $D_i \leq V_i + W_i = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$. 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 $O(N^3) = O(\epsilon^{-3}n^3m^3W^3)$ (Schlesinger and Flach, 2006; Greig et al., 1989; Goldberg and 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 and 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 (which is not submodular in general) is NP-hard (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 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 and Flach, 2006; Greig et al., 1989; Goldberg and 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

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

which we term the *minsum method*. Note $D_i \leq d_iW$ where d_i is the degree of X_i, hence $\left(\sum_{i\in V}\sqrt{S_iD_i}\right)^2 \leq W\left(\sum_{i\in V}\sqrt{d_i}\right)^2$. 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 or smallest 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 knowledge, 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: CURVMESH

We shall review and then refine the second derivative approach used in (Weller and Jebara, 2013a, §5), which we call *curvMesh*. Its 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 gradMesh approach. In practice, however, unless ϵ is very small, gradMesh is much more efficient (see Figure 2).

As in this paper, 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 6, 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

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

Thus, $\tilde{a} = O(e^{W(1+\Delta/2)+T})$ which compares favorably to the earlier bound in Weller and 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 derivations are in the supplement.

4.2 EXTENDING TO A GENERAL MODEL

Using flipping arguments from §2.3, we are able to extend the method of Weller and Jebara (2013a) to apply to general (non-attractive) models. Interestingly, the 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 $0 \leq W_{ij} \leq W$. 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 and Flach, 2006; Greig et al., 1989; Goldberg and Tarjan, 1988). Using the algorithm of Boykov and 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 using the junction tree algorithm, 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 and Sinclair, 1993).

A recent approach reducing MAP inference to identifying a maximum weight stable set in a derived weighted graph (Jebara, 2014; Weller and 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, 2014; Chudnovsky et al., 2005).

5.2 INTRACTABLE MAP 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 that the Bethe box bounds on each $q_i \in [A_i, 1 - B_i]$ are worst case, irrespective of other variables. However, given a particular value for one or more $q_j, j \in \mathcal{N}(i)$, either BBP (Weller and Jebara, 2013a, §6) or MK (Mooij and Kappen, 2007) can produce better bounds on q_i, which may be helpful for pruning the solution space.

5.2.1 Persistent partial optimization approaches

The multi-label implementation of quadratic pseudo-Boolean optimization (Kohli et al., 2008, MQPBO), and the method of 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

6.1 COMPARISON TO EARLIER WORK

We compared the new mesh construction methods from this paper with the earlier approach by Weller and Jebara (2013a), see Figure 2. We considered two values of ϵ: 1 (medium resolution) and 0.1 (fine resolution). For each value, we generated random MRFs on n variables, all pairwise connected, where $\theta_i \sim U[-2, 2]$ and $W_{ij} \sim U[-W, W]$, using the input convention of §2.1.[5] We show results first for fixed $W = 5$ as n is varied from 3 to 20, then for fixed $n = 10$ as W is varied from 1 to 10, generating 10 random models for each value. Of the various first derivative gradMesh methods, only minsum is shown since the others would not be sufficiently distinguishable on these plots.[6]

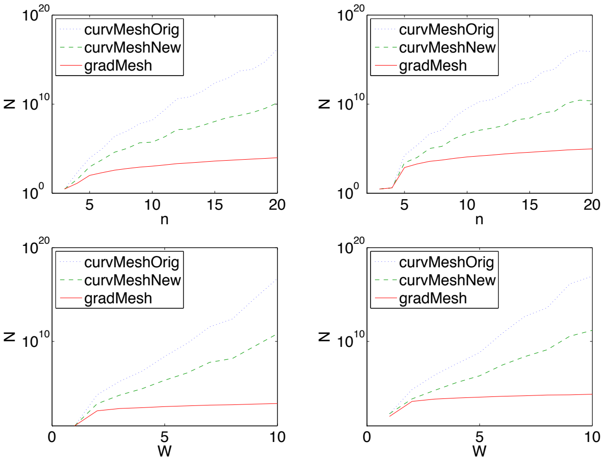

Figure 2: Variation in N = sum of number of mesh points in each dimension, **log scale**, as: (top) n = number of variables is changed, keeping $W = 5$ fixed; (bottom) W = maximum coupling strength is changed, keeping $n = 10$ fixed. On the left, $\epsilon = 1$ (medium resolution); on the right, $\epsilon = 0.1$ (fine resolution). In each case the topology is a complete graph, edge weights are chosen $W_{ij} \sim U[-W, W]$ and $\theta_i \sim U[-2, 2]$. Average over 10 random models for each value. *curvMeshOrig* is the original method of Weller and Jebara (2013a); *curvMeshNew* is our refinement, see §4; *gradMesh* is our new first derivative minsum method, see §3. For more details, see text of §6.1.

Note that N is shown on a $\log$ axis, thus we observe that the new methods dramatically outperform that of Weller and Jebara (2013a) by many orders of magnitude for most cases of interest, even for small ϵ. Further, recall that $N = \sum_i N_i$ is the sum of the number of mesh points in each dimension. The runtime of the overall algorithm is certainly $\Omega(N)$, even for attractive models[7], and for general models is typically a significantly higher power, thus further demonstrating the benefit of the new methods.

[5] The original method of Weller and Jebara (2013a) could only handle attractive models but we augment it as in §4.2. Plots for attractive models, where $W_{ij} \sim U[0, W]$ are very similar to those shown.

[6] In practice, the adaptive methods typically produce a mesh with about half the number of points in each dimension.

[7] In our experiments on attractive models, the Boykov-Kolmogorov algorithm typically runs in time $O(N^{1.5})$ to $O(N^{2.5})$.

6.2 POWER NETWORK

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 single simulated sub-network of 55 connected transformers was generated using a random preferential attachment model, resulting in average degree 2 (see Figure 3 in the Appendix). 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 gradMesh adaptive minsum 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 from true value
Mean ℓ_1 error of single marginals	0.003
Log-partition function	0.26

Table 1: Results on simulated power network

It has been suggested that the Bethe approximation is poor when BP fails to converge (Mooij and Kappen, 2005). Our new method will allow this to be explored rigorously in future work. The initial result above is a promising first step and justifies further investigation.

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. Our approaches are useful in practice, and much more efficient than the earlier method of Weller and Jebara (2013a). From experiments run, we note that the ϵ bounds for the adaptive minsum first derivative *gradMesh* approach 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.

As described in §6.2, Bethe pseudo-marginals may be recovered from our approach by taking the q^* that is returned as the $\arg\min$ of $\mathcal{F}$ over the discrete mesh. However, although $\mathcal{F}(q^*)$ is guaranteed within ϵ of the optimum, there is no guarantee that q^* will necessarily be close to a true Bethe optimum pseudo-marginal. For example, the surface could be very flat over a wide region, or the true optimum might be $\frac{\epsilon}{2}$ better at a location far from q^*. We sketch out how our approach may be used to bound the location of a global optimum pseudo-marginal, though note that there is no runtime guarantee. First pick an initial ϵ_1 and run the main algorithm to find q_1^*. Now use any method to solve for the second best discretized mesh point q_2^*. If it happens that $\mathcal{F}(q_2^*) \geq \mathcal{F}(q_1^*) + \epsilon_1$ then, by the nature of the mesh construction, there must be a global minimum within the orthotope given by the neighboring mesh points of q_1^* in each dimension[8] and we terminate. On the other hand, if $\mathcal{F}(q_2^*) < \mathcal{F}(q_1^*)$ then we reduce ϵ, for example to $\frac{\epsilon_1}{2}$ and repeat until we're successful.

Future work includes further reducing the size of the mesh, considering how it should be selected to simplify the subsequent discrete optimization problem, and exploring applications. Importantly, we now have the opportunity to examine rigorously the performance of the global Bethe optimum. In addition, this will provide a benchmark against which to compare other (non-global) Bethe approaches that typically run more quickly, such as LBP or CCCP (Yuille, 2002). Another interesting avenue is to use our algorithm as a subroutine in a dual decomposition approach to optimize over a tighter relaxation of the marginal polytope.

Acknowledgements

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 work was supported in part by NSF grants IIS-1117631 and CCF-1302269.

References

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.

H. Bethe. Statistical theory of superlattices. *Proc. R. Soc. Lond. A*, 150(871):552–575, 1935.

[8] In fact, the optimum must be within a tighter orthotope based on the *reach* down and up, in each dimension, of q_1^*, see 3.1.2.

Y. Boykov and V. Kolmogorov. An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(9):1124–1137, 2004.

V. Chandrasekaran, N. Srebro, and P. Harsha. Complexity of inference in graphical models. In D. McAllester and P. Myllymäki, editors, *UAI*, pages 70–78. AUAI Press, 2008. ISBN 0-9749039-4-9.

M. Chudnovsky, G. Cornuéjols, X. Liu, P. Seymour, and K. Vuskovic. Recognizing Berge graphs. *Combinatorica*, 25 (2):143–186, 2005.

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

D. Greig, B. Porteous, and A. Seheult. Exact maximum a posteriori estimation for binary images. *J. Royal Statistical Soc., Series B*, 51(2):271–279, 1989.

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

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

T. Jebara. *Tractability: Practical Approaches to Hard Problems*, chapter Perfect graphs and graphical modeling. Cambridge Press, 2014.

M. Jerrum and A. Sinclair. Polynomial-time approximation algorithms for the Ising model. *SIAM J. Comput.*, 22(5):1087–1116, 1993.

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

P. Kohli, A. Shekhovtsov, C. Rother, V. Kolmogorov, and P. Torr. On partial optimality in multi-label MRFs. In W. Cohen, A. McCallum, and S. Roweis, editors, *ICML*, volume 307 of *ACM International Conference Proceeding Series*, pages 480–487. ACM, 2008. ISBN 978-1-60558-205-4.

F. Korč, V. Kolmogorov, and C. Lampert. Approximating marginals using discrete energy minimization. Technical report, IST Austria, 2012.

I. Kovtun. Partial optimal labeling search for a NP-hard subclass of (max, +) problems. In B. Michaelis and G. Krell, editors, *DAGM-Symposium*, volume 2781 of *Lecture Notes in Computer Science*, pages 402–409. Springer, 2003. ISBN 3-540-40861-4.

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. Sufficient conditions for convergence of loopy belief propagation. In *Proceedings of the 21st Annual Conference on Uncertainty in Artificial Intelligence (UAI-05)*, pages 396–403. AUAI Press, 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.

J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Morgan Kaufmann, 1988.

C. Rudin, D. Waltz, R. Anderson, A. Boulanger, A. Salleb-Aouissi, M. Chow, H. Dutta, P. Gross, B. Huang, and S. Ierome. Machine learning for the New York City power grid. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34 (2):328–345, February 2012. ISSN 0162-8828. doi: 10.1109/TPAMI.2011.108.

D. Schlesinger and B. Flach. Transforming an arbitrary minsum problem into a binary one. Technical report, Dresden University of Technology, 2006.

S. Shimony. Finding MAPs for belief networks is NP-hard. *Artifical Intelligence*, 68(2):399–410, 1994.

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

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

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.

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

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

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

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 (the maximum eigenvalue of the Hessian H) 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 6, equation (9) and Lemma 4,

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

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

Lemma 7. *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*

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

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 7 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 $y \ge (1-k)\bar{\eta}$, 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 and 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

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

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

Here we extend the analysis of Weller and 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 6 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 $0 \le W_{ij} \le W$, now we assume $|W_{ij}| \le W$.

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 7 using $k = -\alpha_{ij}$ and $A'_j = B_j, B'_j = A_j$. The theoretical bound for an attractive edge from (19) becomes $H_{ij} \le \frac{-\alpha_{ij}}{\bar{\eta}(1-\alpha_{ij}^2)}$. As we should expect from the attractive case, the following result holds.

Lemma 8. *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 and Jebara (2013a) may similarly be combined with flipping to provide a worse upper bound for H_{ij} when (i,j) is repulsive. See Weller and 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 and Jebara (2013a) Theorem 11, the equivalent bound is $H_{ij} \le \frac{-\alpha_{ij}}{4K_{ij}}$ 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 9 (Lower bound for ξ_{ij} for a repulsive edge, analogue of Lemma 4)**.** *For any repulsive edge* (i,j)*,*
$q_iq_j - \xi_{ij} \le -\alpha_{ij}p_{ij}$ *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 (2). 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 4 to the new model, $\xi'_{ij} - q'_iq'_j \le \frac{\alpha'_{ij}}{1+\alpha'_{ij}} m'(1-M')$. Substituting in $\xi'_{ij} = q_i - \xi_{ij}$ from (21) and using (20), we have $(q_i - \xi_{ij}) - q_i(1-q_j) \le -\alpha_{ij}m'(1-M')$. Since $m' = \min\{q_i, 1-q_j\}$ and $M' = \max\{q_i, 1-q_j\}$, noting $q_i \le 1-q_j \Leftrightarrow q_i + q_j \le 1 \Leftrightarrow q_iq_j \le (1-q_i)(1-q_j)$, the result follows. □

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

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

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

4.2.2 Diagonal terms

Consider the H_{ii} terms from Theorem 6, which is true for a general model. If all neighbors of X_i are adjacent via attractive edges, then, as in Weller and Jebara (2013a) Theorem 11, $H_{ii} \leq \frac{1}{\eta_i(1-\eta_i)}\left(1 - d_i + \sum_{j \in \mathcal{N}(i)} \frac{1}{1-\left(\frac{\alpha_{ij}}{1+\alpha_{ij}}\right)^2}\right)$.

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,

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

Similarly to the analysis in §4.2.1, using Lemma 8 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 $W = \max |W_{ij}|$.

6.2 POWER NETWORK

Here we show the simulated sub-network used in the experiment.

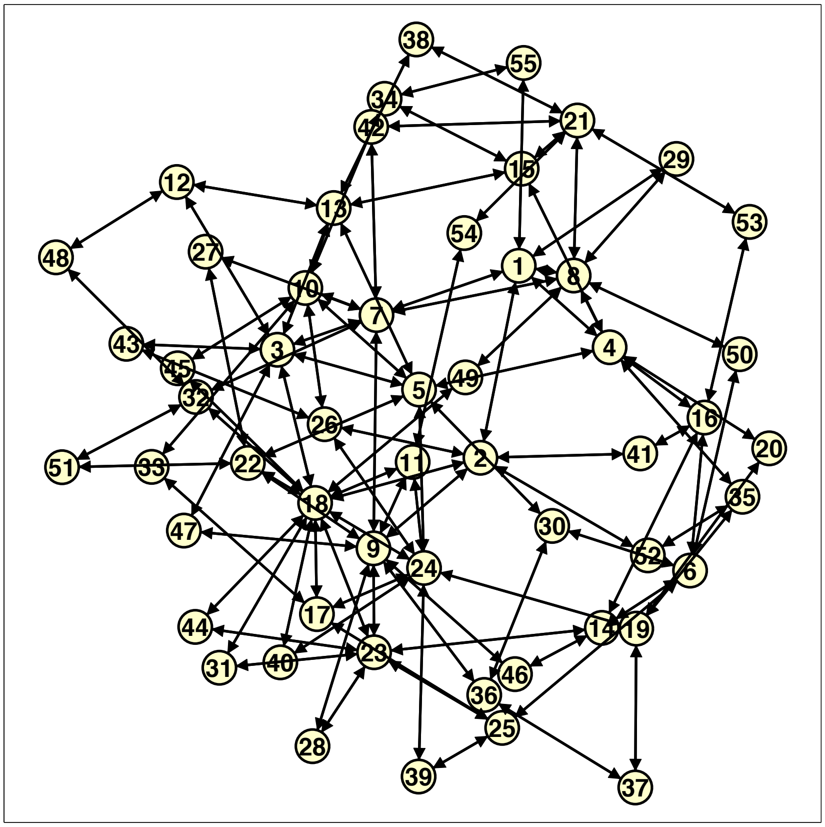

Figure 3: Sub-network used for the experiment described in the main text §6.2.