Bethe Bounds and Approximating the Global Optimum

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

Inference in general Markov random fields (MRFs) is NP-hard, though identifying the maximum a posteriori (MAP) configuration of pairwise MRFs with submodular cost functions is efficiently solvable using graph cuts. Marginal inference, however, even for this restricted class, is #P-hard. Restricting to binary pairwise models, we prove new formulations of derivatives of the Bethe free energy, provide bounds on the derivatives and bracket the locations of stationary points. Several results apply whether the model is associative or not. Applying these to discretized pseudo-marginals in the associative case, we present a polynomial time approximation scheme for global optimization of the Bethe free energy provided the maximum degree $\Delta = O(\log n)$, where n is the number of variables. Runtime is guaranteed $O(\epsilon^{-\frac{3}{2}} n^6 \Sigma^{\frac{3}{4}} \Omega^{\frac{3}{2}})$, where $\Sigma = O(\frac{\Delta}{n})$ is the fraction of possible edges present and Ω is a function of MRF parameters. We examine use of the algorithm in practice, demonstrating runtime that is typically much faster, and discuss several extensions.

1 Introduction

Markov random fields are fundamental tools in machine learning with broad application in areas including computer vision, speech recognition and computational biology. Two forms of inference are commonly employed: *maximum a posteriori* (MAP), where the most likely configuration is returned; and *marginal*, where the marginal probability distributions for each set of variables with a linking potential function are returned. In general, MAP inference is NP-hard [24] and marginal inference, even for pairwise models, is #P-hard [28, 3, 5].

An important class of MRFs, those with only unary and pairwise submodular cost functions, admits efficient MAP inference. This was first shown for binary models [9] and applied broadly in computer vision [2], where the graph cuts method is particularly effective [29]. Recent work extended the application of this approach to multi-label submodular energies of up to third order [21, 23]. Yet marginal inference, even for binary pairwise models, is intractable with few known exceptions. Belief propagation (BP) is efficient (and exact) for trees, and applying the same framework to general models, termed loopy belief propagation (LBP), is guaranteed to converge when the topology has one cycle [36].

However, while LBP has proved remarkably effective in some situations, it fails in others with no guarantee on convergence. A key result is that BP fixed points coincide with stationary points of the Bethe variational problem [37]. Stationary points, however, may not identify the global optimum of the the Bethe free energy. Subsequently, it was further shown that all stable BP fixed points are known to be local optima (rather than saddle points) of this problem, but not vice versa [11, 12]. Variational methods demonstrate that minimizing the Bethe free energy should deliver a good approximation to the true marginal distribution and recently [22] proved that for submodular MRFs, the Bethe optimum is an upper bound on the true free energy and thus yields a desirable lower bound on the partition function, for which methods such as [32] provide an upper bound.

Marginal inference is a crucial problem in probabilistic systems. A noteworthy example is the Quick Medical Reference (QMR) problem [27], a graphical model involving 600 diseases and 4000 possible findings. Therein, medical diagnostics are performed by computing the posterior marginal probability of each disease given a set of possible findings. The marginal

Appearing in Proceedings of the 16^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2013, Scottsdale, AZ, USA. Volume 31 of JMLR: W&CP 31. Copyright 2013 by the authors.

distribution over the presence of a disease must often be precisely estimated in order to determine the course of medical treatment. Thus, we seek the probability that a patient suffers from a condition, rather than the MAP estimate, which could be very different.

Marginal inference also arises during learning or parameter estimation in Markov random fields. For instance, computing the gradients of a partition function in a maximum likelihood estimation procedure is equivalent to marginal inference. In learning problems, the intractability of the marginal inference problem requires the exploration of marginal approximation schemes [7]. However, in the general case, *both* exact marginal inference and approximate marginal inference are NP-hard [3, 5].

For associative binary pairwise models, a FPRAS for the true partition function (not the Bethe approximation) was derived in [15], but the runtime is unwieldy at $O(\epsilon^{-2}m^3n^{11}\log n)$, which makes it impractical.

1.1 Contribution & Summary

We derive various properties of the Bethe free energy and apply them to discretized pseudo-marginals to prove a polynomial-time approximation scheme (PTAS) for the global minimum of the Bethe free energy for binary pairwise associative MRFs. We then go on to consider practical implementation.

The idea is that if we can find the optimal discretized point on a sufficiently fine mesh that covers all possible locations of an optimum point within a distance of δ, then we can bound the difference to the optimum by $\frac{1}{2}\Lambda\delta^2$ where Λ is the greatest directional second derivative. To our knowledge, we present the first rigorous bounds on Λ. One reason this is difficult is that derivatives tend to infinity as singleton marginals approach the boundary cases of 0 or 1. Hence we need to prove bounds on the location away from these edges.

We discuss preliminaries in section 2, then in section 3, derive various bounds, including on the location of any stationary point of the Bethe free energy. In section 4, we establish results for second and higher derivatives with a view to bounding Λ. Additional analysis yields the result (Theorem 8) that the discretized multi-label problem is submodular (see 2.1) on any mesh and hence the discretized optimum can be found efficiently using graph cuts [23].

In section 5, we use these earlier results to derive our main theoretical contribution, a deterministic PTAS for the global optimum of the Bethe free energy. The result may be summarized below, see sections 2 and 5.3 for notation and details:
For a binary associative pairwise MRF with n variables, maximum degree $\Delta = O(\log n)$, all singleton potentials bounded by T and pairwise potentials bounded by W s.t. all $\theta_i \in [-T, +T]$, and $\theta_{ij} = w_{ij}I$ with $0 \leq w_{ij} \leq W$, then within time $O(\epsilon^{-\frac{3}{2}}n^6\Sigma^{\frac{3}{4}}\Omega^{\frac{3}{2}})$, our algorithm is guaranteed to return a pseudo-marginal on the local polytope with Bethe free energy (equivalently log Bethe partition function) within ϵ of the global optimum. Σ reflects the density of edges and is $O(\frac{\Delta}{n})$; Ω is a measure of how extreme Λ, the curvature of the Bethe free energy, can be given the properties of the MRF, and is computed as $\Omega = \max(a, b)$ where $a = O(e^{W(1+\Delta)+2T})$ and $b = O(\Delta e^{W(1+\Delta/2)+T})$.

In section 6, we discuss practical implementation of the algorithm and present experimental results. We show how the analysis of section 3 may be extended to yield a fast algorithm that iteratively improves the earlier A_i, B_i bounds, often leading to greatly improved performance. We also discuss an existing alternative approach to improve the A_i, B_i bounds due to [19] which takes longer but can produce superior results.

Potential extensions are noted in the closing section 7, including applications to non-associative models, to models that are themselves multi-label and to models with higher order terms.

1.2 Structure of the overall algorithm

Input: Parameters $\{\theta_i, W_{ij}\}$ for an associative binary pairwise MRF, and a desired accuracy ϵ.

a) Compute bounds $\{A_i, B_i\}$ on the location of minima (see section 3 for the theoretical result, section 6 for improved performance in practice).

b) Compute $\Omega = \max(a, b)$ from Theorem 11 and bound Λ using equation (12).

c) Compute γ using $\Lambda n\gamma^2/2 \leq \epsilon$ (see start Section 5).

d) Generate a multi-label submodular MRF on $\prod_i [A_i, B_i]$ with mesh width γ.

e) Solve for the MAP solution using graph cuts [23].

1.3 Related work

A variety of heuristics have been proposed for marginal inference problems. Marginal inference in the QMR medical diagnostic problem has been explored with Markov Chain Monte Carlo (MCMC) [18, 26, 4] methods, variational methods [14], and search methods [6]. Many of these heuristics are restricted to certain classes of graphical model (such as QMR). Here we explore another approach to approximate marginal inference by minimizing the Bethe free energy.

The minimization of Bethe free energy is often approached using loopy Belief propagation (LBP). Several sets of sufficient conditions have been derived for convergence, such as [19]. In general, however, LBP may not converge, or may converge to a local optimum, which prevents its use as a PTAS for Bethe minimization [33]. This is still true for the restricted class of associative binary pairwise models [19]. An important contribution [35] showed that the Bethe free energy of a binary pairwise MRF may be considered as a function only of the singleton marginals, however this connection was provided without convergence results.

A PTAS was recently proposed [25] for the location of a point whose derivative of the Bethe free energy has magnitude less than ϵ. However, this identifies only an approximately stationary point (which may not be even a local minimum) that could be arbitrarily far from the global optimum. That result applies for a general binary pairwise MRF subject to maximum degree $O(\log n)$. Here we primarily focus on associative models with the same degree restriction, but our deliverable not only satisfies the property in [25], but importantly, is also guaranteed to have Bethe free energy within ϵ of the optimum.

The PTAS in [25] may provide the global optimum when the fixed point is unique and recent work [34] has enumerated necessary and sufficient conditions for uniqueness. Nevertheless, aside from these restricted settings, there are no prior polynomial-time methods for finding or rigorously approximating the global minimum of the Bethe free energy. Earlier work considered discretizations of pseudo-marginals but presented incomplete results [16]. We go significantly further in deriving additional key results which together admit the PTAS and can also dramatically improve performance in practice. These include explicit forms and bounds on the second derivatives and on the locations of stationary points.

As discussed in sections 3, 5 and 6, our approach requires bounds on the location of minima of the Bethe free energy. We derive a new, fast method for this (BBP) but note that in practice, an existing approach [19] produces bounds that are no worse, and sometimes better, though it takes more time. In other contexts, bounds on the true marginals may be more useful. [17, 1] have derived such bounds, and [31, 13, 20] develop them in relation to pseudo-marginals from BP. Recent work has explored conditions under which the fixed points of the Bethe free energy may or may not correspond to the values of the true marginals [10].

2 Preliminaries & Notation

We focus on a binary pairwise MRF over n variables $X_1, \dots, X_n \in \mathbb{B} = \{0,1\}$ with topology $(\mathcal{V}, \mathcal{E})$ and generally follow the notation of [35]. We assume[1]

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

where the partition function $Z = \sum_x e^{-E(x)}$ is a normalizing constant. Let F be the Bethe free energy, so $F = E - S$ where S is the Bethe approximation to the true entropy, $S = \sum_{(i,j) \in \mathcal{E}} S_{ij} + \sum_{i \in \mathcal{V}} (1 - z_i) S_i$. S_{ij} is the entropy of a pseudo-marginal of (X_i, X_j) on the local polytope, S_i is the entropy of the singleton distribution and z_i is the degree of i, that is the number of variables to which X_i is adjacent. We assume the model is connected so all $z_i \geq 1$. For each node i define sum of positive and negative incident edge weights: $W_i = \sum_{j \in \mathsf{N}(i): W_{ij} > 0} W_{ij}$, $V_i = -\sum_{j \in \mathsf{N}(i): W_{ij} < 0} W_{ij}$ where $\mathsf{N}(i)$ indicates the neighbors of node i. For a pseudo-marginal distribution q, let $q_i = p(X_i = 1)$. Consistency and normalization constraints from the local polytope imply the pairwise marginal,

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

for some $\xi_{ij} \in [0, \min(q_i, q_j)]$, where $\mu_{ij}(a,b) = p(X_i = a, X_j = b)$. 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 *associative*; if negative then the edge is *repulsive*.[2] The MRF is associative if all edges are associative. As in [35], one can solve for ξ_{ij} explicitly in terms of q_i and q_j by minimizing F, leading to a quadratic equation with real roots,

$$\alpha_{ij} \xi_{ij}^2 - [1 + \alpha_{ij}(q_i + q_j)] \xi_{ij} + (1 + \alpha_{ij}) q_i q_j = 0. \quad (3)$$

For $\alpha_{ij} > 0$, $\xi_{ij}(q_i, q_j)$ is the lower root, for $\alpha_{ij} < 0$ it is the higher. Notice that when $\alpha_{ij} = 0$ (no edge relationship) this reduces as expected to $\xi_{ij} = p(X_i = 1, X_j = 1) = p(X_i = 1)p(X_j = 1) = q_i q_j$.

S_{ij} is the entropy of $\mu_{ij}(q_i, q_j)$. Hence

$$\begin{aligned} F(q) = \sum_{(i,j) \in \mathcal{E}} &-\big(W_{ij} \xi_{ij} + S_{ij}(q_i, q_j)\big) \\ &+ \sum_{i \in \mathcal{V}} \big(-\theta_i q_i + (z_i - 1) S_i(q_i) \big). \end{aligned} \quad (4)$$

Collecting the pairwise terms for one edge, define

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

[1] The energy E can always be thus reparameterized with finite θ_i and W_{ij} terms provided $p(x) > 0\ \forall x$. There are reasonable distributions where this does not hold, i.e. $\exists x : p(x) = 0$ but this can often be handled by assigning such configurations a sufficiently small positive probability ϵ.

[2] Our use of *associative* is equivalent to a submodular energy function. Other terms used are *attractive*, *regular* or *ferromagnetic*.

We are interested in *discretized pseudo-marginals* where for each q_i we restrict its possible values to a discrete set D_i of points in $[0, 1]$. Note we may often have $D_i \neq D_j$. Let $\mathcal{D} = \prod_{i \in V} D_i$.

In [35], the first partial derivative of the Bethe free energy is derived as

$$\frac{\partial F}{\partial q_i} = -\theta_i + \log Q_i \text{ , where} \tag{6}$$

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

Recall the sigmoid function $\sigma(x) = 1/(1 + \exp(-x))$ which will be used for Bethe bounds. We write A_i for the lower bound of q_i and B_i for the lower bound of $1 - q_i$ so $A_i \leq q_i \leq (1 - B_i)$. Define $\eta_i = \min(A_i, B_i)$.

2.1 Submodularity

In our context, 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

$$\forall x, y \in X_{ij},\ f(x \wedge y) + f(x \vee y) \leq f(x) + f(y) \tag{7}$$

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 associativity.

The key property for us is that if all the pairwise cost functions f_{ij} over $D_i \times D_j$ from (5) are submodular then the global discretized optimum may be found efficiently as a multi-label MAP inference problem using graph cuts [23].

3 Initial bounds

We use the technique of flipping variables, i.e. considering $Y_i = 1 - X_i$. Flipping a variable flips the parity of all its incident edges so associative ↔ repulsive. Flipping both ends of an edge leaves its parity unchanged.

3.1 Flipping all variables

Consider a new model with variables $\{Y_i = 1 - X_i, i = 1, \dots, n\}$ and the same edges. Instead of θ_is and W_{ij}s, let the new model have parameters ϕ_i and V_{ij}. We identify values such that the energies of all states are maintained up to a constant.[3]

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

Matching coefficients yields

$$V_{ij} = W_{ij},\ \phi_i = -\theta_i - \sum_{j \in \mathsf{N}(i)} W_{ij} = -\theta_i - W_i. \tag{8}$$

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

3.2 Flipping some variables

Sometimes we flip only a subset $\mathcal{R} \subseteq \mathcal{V}$ of the variables. This can be useful, for example, to make the model locally associative around a variable, which can always be achieved by flipping just those neighbors to which it has a repulsive edge. Let $Y_i = 1 - X_i$ if $i \in \mathcal{R}$, else $Y_i = X_i$ 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 3.1, solving for V_{ij} and ϕ_i such that energies are unchanged up to a constant,

$$\begin{aligned} V_{ij} &= \begin{cases} W_{ij} & (i,j) \in \mathcal{E}_0 \cup \mathcal{E}_2, \\ -W_{ij} & (i,j) \in \mathcal{E}_1 \end{cases} \\ \phi_i &= \begin{cases} \theta_i + \sum_{(i,j) \in \mathcal{E}_1} W_{ij} & i \in \mathcal{S}, \\ -\theta_i - \sum_{(i,j) \in \mathcal{E}_2} W_{ij} & i \in \mathcal{R}. \end{cases} \end{aligned} \tag{9}$$

Lemma 1. *Flipping any set of variables changes affected pseudo-marginal matrix entries' locations but not values. The Bethe free energy is unchanged up to a constant, hence the locations of stationary points are unaffected. Proof in Supplement.*

3.3 Bounds

We derive several results that are useful in bounding the Bethe free energy as well as the marginals.

Lemma 2. $\alpha_{ij} \geq 0 \Rightarrow \xi_{ij} \geq q_i q_j, \alpha_{ij} \leq 0 \Rightarrow \xi_{ij} \leq q_i q_j$

Proof. The quadratic equation (3) for ξ_{ij} may be rewritten $\xi_{ij} - q_i q_j = \alpha_{ij}(q_i - \xi_{ij})(q_j - \xi_{ij})$. Both terms in parentheses on the right are elements of the pseudo-marginal matrix μ so are constrained to be ≥ 0. □

This simple result is sufficient to bound the location of stationary points of the Bethe free energy away from the edges of 0 and 1, allowing us to construct our PTAS, though we improve the bounds in section 6.

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

Theorem 3. *If all edges incident to X_i are associative then at any stationary point of the Bethe free energy, $\sigma(\theta_i) \leq q_i \leq \sigma(\theta_i + W_i)$. Remark the same sandwich result holds for the true marginal p_i.*

Proof. We first prove the left inequality. Consider (6). Using $\alpha_{ij} > 0 \; \forall j \in \mathsf{N}(i)$ and Lemma 2 we have

$$\begin{aligned} Q_i &= \frac{\prod_{j \in \mathsf{N}(i)} (q_i - \xi_{ij})}{q_i^{z_i - 1}} \frac{(1-q_i)^{z_i-1}}{\prod_{j \in \mathsf{N}(i)} (1 + \xi_{ij} - q_i - q_j)} \\ &\leq \frac{\prod_{j \in \mathsf{N}(i)} q_i(1-q_j)}{q_i^{z_i-1}} \frac{(1-q_i)^{z_i-1}}{\prod_{j \in \mathsf{N}(i)} (1-q_i)(1-q_j)} \\ &= \frac{q_i}{1-q_i} \text{ which gives the result.} \end{aligned}$$

To obtain the right inequality, flip all variables as in section 3.1. Using the first inequality, (8) and Lemma 1 yields $1 - q_i \geq \sigma(-\theta_i - W_i) \Leftrightarrow q_i \leq \sigma(\theta_i + W_i)$ since $1 - \sigma(-x) = \sigma(x)$. To show the result for the true marginal, let $m_{i=a} = \sum_{x:x_i=a} \exp(\sum_{i \in V} \theta_i x_i + \sum_{(i,j) \in E} W_{ij} x_i x_j)$ then using (1), $p_i = \frac{m_{i=1}}{m_{i=1} + m_{i=0}}$. Since all $W_{ij} > 0$ the result follows. □

Using (9) we obtain a more powerful corollary.

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, $\sigma(\theta_i - V_i) \leq q_i \leq \sigma(\theta_i + W_i)$. The same result holds for the true marginal p_i.*

Proof. Using (9), flip all variables adjacent to X_i with a repulsive edge, i.e. set $\mathcal{R} = \{j \in \mathsf{N}(i) : W_{ij} < 0\}$. The resulting new model is fully associative around X_i so we may apply Theorem 3 to yield the result. □

Lemma 5. *For $q_i, q_j \in [0,1], 0 \leq q_i + q_j - 2q_iq_j \leq 1$. Proof in Supplement.*

Lemma 6 (Upper bound for ξ_{ij})**.** *If $\alpha_{ij} > 0$, then* $q_j - \xi_{ij} \geq \frac{q_j(1-q_i)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \geq \frac{q_j(1-q_i)}{1+\alpha_{ij}}$
$q_i - \xi_{ij} \geq \frac{q_i(1-q_j)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \geq \frac{q_i(1-q_j)}{1+\alpha_{ij}}$.
Also $\xi_{ij} \leq m(\alpha_{ij} + M)/(1 + \alpha_{ij}) \Rightarrow \xi_{ij} - q_iq_j \leq \frac{\alpha_{ij} m(1-M)}{1+\alpha_{ij}}$, where $m = \min(q_i, q_j)$ and $M = \max(q_i, q_j)$.

Proof. We prove the first inequality. The second follows by Lemma 5 and those for $q_i - \xi_{ij}$ follow by symmetry. The final inequality follows by combining the earlier ones. Let $\xi_{ij} = q_j + y$ and substitute into (3),

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

The function is a convex parabola which at $y = 0$ is at $q_j(q_i - 1) \leq 0$.[4] From Lemma 2 we know that the left root is at $\xi_{ij} \geq q_iq_j$ so we may take the derivative there, i.e. at $q_j + y = q_iq_j \Leftrightarrow y = q_j(q_i - 1)$ and by convexity establish a lower bound for $q_j - \xi_{ij}$. □

4 Higher derivatives & submodularity

We first derive a novel result for the second derivatives of an edge which will be crucial later for bounding the error of the discretized global optimum and also will allow us to show that the discretized multi-label problem is submodular.

4.1 Second derivatives for each edge

Theorem 7. *For any edge (i,j), for any α_{ij}, writing $f = f_{ij}$ and $\mu_{ab} = \mu_{ij}(a,b)$ from (2),*

$$\begin{aligned} \frac{\partial^2 f}{\partial q_i^2} &= \frac{1}{T_{ij}} q_j(1-q_j) \\ \frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} &= \frac{1}{T_{ij}} (\mu_{01}\mu_{10} - \mu_{00}\mu_{11}) \\ \frac{\partial^2 f}{\partial q_j^2} &= \frac{1}{T_{ij}} q_i(1-q_i) \end{aligned}$$

where $T_{ij} = q_iq_j(1-q_i)(1-q_j) - (\xi_{ij} - q_iq_j)^2 \geq 0$ with equality only if q_i or $q_j \in \{0,1\}$. Further $\mu_{01}\mu_{10} - \mu_{00}\mu_{11} = q_iq_j - \xi_{ij}$ and has the sign of $-\alpha_{ij}$.
Proof in Supplement.

Note that stronger edge interactions lead through higher $|\alpha_{ij}|$ to greater $(\xi_{ij} - q_iq_j)^2$ and hence larger second derivatives.

4.2 Submodularity

Theorem 8. *If a binary pairwise MRF is submodular on an edge (i,j), i.e. $\alpha_{ij} > 0$, then the multi-label discretized MRF for any discretization $\mathcal{D}$ is submodular for that edge. In particular, if the MRF is fully associative/submodular, i.e. $\alpha_{ij} > 0 \; \forall (i,j) \in \mathcal{E}$, then the multi-label discretized MRF is fully submodular for any discretization. Proof in Supplement.*

4.3 Second derivatives for singleton terms

Let $f_i(q_i)$ be the singleton terms from (4) for X_i. The only non-zero derivatives are with respect to q_i.

$$\begin{aligned} f_i(q_i) &= -\theta_i q_i + (z_i - 1) S_i(q_i) \\ \frac{\partial f_i}{\partial q_i} &= -\theta_i - (z_i - 1)[\log q_i - \log(1-q_i)] \\ \frac{\partial^2 f_i}{\partial q_i^2} &= -(z_i - 1) \frac{1}{q_i(1-q_i)} \leq 0 \text{ for a connected graph.} \end{aligned}$$

[4] This confirms neatly that we must take the left root else $y > 0 \Rightarrow \mu_{01} < 0$ (a contradiction).

$$\text{Hence, } -\frac{z_i-1}{\eta_i(1-\eta_i)} \le \frac{\partial^2 f}{\partial q_i^2} \le 0, \ \eta_i = \min(A_i, B_i). \tag{10}$$

5 Approximating the Global Optimum for an Associative Model

We now assemble earlier results to form the complete matrix H of second derivatives of the Bethe free energy F and use this to bound the error between the discretized optimum and the global Bethe optimum. In this section we assume the model is associative. Define the *Bethe box* to be the orthotope (or hyper-cuboid) given by $q_i \in [A_i, 1-B_i] \ \forall i \in \mathcal{V}$.

At the optimum (or any stationary point), all first derivatives are zero. If we choose our discretization mesh $\mathcal{D}$ to be sufficiently fine then we can be sure that some point in the mesh is within distance δ of a true optimum. In particular, if we choose each D_i so that in the q_i dimension every point in $[A_i, 1-B_i]$ is within distance γ of an optimum, then $\delta^2 \le n\gamma^2$.

Using a first order Taylor expansion of F around a true optimum, with the remainder expressed in terms of the second derivative, the error of our discretized optimum versus the true Bethe optimum $\le \frac{1}{2}\Lambda\delta^2$, where Λ is the largest eigenvalue of H evaluated at some intermediate point, which we shall bound. Observe that any Bethe optimum must lie in the Bethe box, hence we need only bound the largest eigenvalue of H anywhere inside it.[5]

Note that our error is one-sided since our discretized optimum can never be better than the true optimum. This may facilitate further analysis to find a better approximation by using points in the neighborhood to estimate the likely error.

5.1 Complete matrix of second derivatives

Theorem 7 and (10) provide all the terms.

Lemma 9. *All entries on the main diagonal of H are strictly positive, all others are ≤ 0.*

Proof. Apply Theorem 7. If $(i,j) \in \mathcal{E}$ then $H_{ij} = (q_i q_j - \xi_{ij})/T_{ij} \le 0$. If $(i,j) \notin \mathcal{E}$, $i \ne j$ then $H_{ij} = 0$. On the main diagonal,

$$H_{ii} = -\frac{z_i-1}{q_i(1-q_i)} + \sum_{j\in\mathsf{N}(i)} \frac{q_j(1-q_j)}{T_{ij}} \tag{11}$$

$$\ge \frac{1-z_i}{q_i(1-q_i)} + \sum_{j\in\mathsf{N}(i)} \frac{q_j(1-q_j)}{q_iq_j(1-q_i)(1-q_j)} = \frac{1}{q_i(1-q_i)}. \ \square$$

5.2 Max eigenvalue & complexity bound

We have shown that H is a real symmetric matrix with strictly positive main diagonal and all other entries ≤ 0. To further bound the entries we derive a lower bound for T_{ij} at any point in the Bethe box. Define $K_{ij} = \eta_i\eta_j(1-\eta_i)(1-\eta_j)\frac{2\alpha_{ij}+1}{(\alpha_{ij}+1)^2}$. All terms are known from the data prior to the discrete optimization.

Lemma 10. *At any point in the Bethe box, $T_{ij} \ge K_{ij}$.*

Proof. Using Theorem 7 and Lemma 6,

$$\begin{aligned} T_{ij} &\ge q_iq_j(1-q_i)(1-q_j) - \Big(\frac{\alpha_{ij} m(1-M)}{1+\alpha_{ij}}\Big)^2 \\ &\ge q_iq_j(1-q_i)(1-q_j)\Big[1 - \Big(\frac{\alpha_{ij}}{1+\alpha_{ij}}\Big)^2\Big]. \qquad \square \end{aligned}$$

Theorem 11. *At any point in the Bethe box, each entry H_{ij} satisfies $-a \le H_{ij} \le b$ where*

$$\begin{aligned} a &= \frac{1}{4} \max_{(i,j)\in\mathcal{E}} \frac{\alpha_{ij}}{\alpha_{ij}+1}\frac{1}{K_{ij}} \\ &= \max_{(i,j)\in\mathcal{E}} \frac{\alpha_{ij}(\alpha_{ij}+1)}{4(2\alpha_{ij}+1)\eta_i\eta_j(1-\eta_i)(1-\eta_j)}, \\ b &= \max_{i\in\mathcal{V}} \frac{1}{\eta_i(1-\eta_i)}\Big(1 - z_i + \sum_{j\in\mathsf{N}(i)} \frac{(\alpha_{ij}+1)^2}{2\alpha_{ij}+1}\Big). \end{aligned}$$

Proof. For any edge $(i,j) \in \mathcal{E}$,

$$-H_{ij} = \frac{\xi_{ij} - q_iq_j}{T_{ij}} \le \frac{m(1-M)\alpha_{ij}}{1+\alpha_{ij}}\frac{1}{K_{ij}} \le \frac{1}{4}\frac{\alpha_{ij}}{1+\alpha_{ij}}\frac{1}{K_{ij}}.$$

Using (11) and the expression from the proof of Lemma 10,

$$\begin{aligned} H_{ii} &\le \frac{1-z_i}{\eta_i(1-\eta_i)} + \sum_{j\in\mathsf{N}(i)} \frac{1}{q_i(1-q_i)\Big[1 - \Big(\frac{\alpha_{ij}}{1+\alpha_{ij}}\Big)^2\Big]} \\ &\le \frac{1}{\eta_i(1-\eta_i)}\Big(1 - z_i + \sum_{j\in\mathsf{N}(i)} \frac{(\alpha_{ij}+1)^2}{2\alpha_{ij}+1}\Big). \qquad \square \end{aligned}$$

Since $\alpha_{ij}+1 < 2\alpha_{ij}+1$ we have the corollary that $H_{ii} < \frac{1+\sum_{j\in\mathsf{N}(i)}\alpha_{ij}}{\eta_i(1-\eta_i)}$. We remark that at any minimum of the Bethe free energy, all eigenvalues are ≥ 0 so at these locations the maximum eigenvalue $\le$ Tr $H <$ $\sum_{i\in\mathcal{V}} \frac{1}{\eta_i(1-\eta_i)} + \sum_{(i,j)\in\mathcal{E}} \alpha_{ij}\Big(\frac{1}{\eta_i(1-\eta_i)} + \frac{1}{\eta_j(1-\eta_j)}\Big)$.

To bound the largest eigenvalue anywhere in the Bethe box, we may use recent results such as Corollary 2 in [38], though we suspect that the particular properties of H given in Lemma 9 may admit more precise bounds. Here we use an elementary bound relating to edge sparsity or maximum degree as in [25]. Let Σ be the proportion of non-zero entries in H so the number

[5] This value can also be used to find an approximately stationary point [25] if required by considering the Taylor expansion of F' around a stationary point.

of non-zero entries is $n^2\Sigma \le n + n\Delta \Rightarrow \Sigma \le \frac{\Delta+1}{n}$. Let $\Omega = \max(a,b)$ from Theorem 11, then we have

$$\Lambda \le \sqrt{\mathrm{tr}(H^T H)} \le \sqrt{\Sigma n^2 \Omega^2} = n\Omega\sqrt{\Sigma}. \qquad (12)$$

Returning to our objective at the start of this section 5, note that by using N_i points in D_i, we can ensure $\gamma \le (1 - B_i - A_i)/(N_i + 1)$. Using worst case Bethe bounds ($A_i = B_i = 0$), we achieve maximum γ distance in each dimension with $\frac{1}{\gamma}$ points for each variable, so the total number of nodes in the max-flow graph we need to solve the multi-label graph cuts problem is $N \le \frac{n}{\gamma}$. We require $n\gamma^2 \le \frac{2\epsilon}{\Lambda}$ hence $N^2 \ge \frac{n^3\Lambda}{2\epsilon}$. Using (12) it is sufficient if $N^2 \ge \frac{n^4\Omega\sqrt{\Sigma}}{2\epsilon}$. Graph cuts is a max-flow algorithm for which there are push-relabel methods guaranteed to run in time $O(N^3)$ [8]. Hence our algorithm has worst case runtime of $O(\epsilon^{-\frac{3}{2}} n^6 \Sigma^{\frac{3}{4}} \Omega^{\frac{3}{2}})$. However, in practice, runtime for this class of max-flow problem using algorithms such as Boykov-Kolmogorov [2] can approach $O(N)$ for much faster performance.

5.3 Model specification

Note Ω above may depend on n. For our analysis throughout this paper, we assumed the reparameterization in (1) but a natural specification to assume for input models avoiding bias is to provide maximum possible values W and T with

$$\theta_{ij} = \begin{pmatrix} W_{ij}/2 & 0 \\ 0 & W_{ij}/2 \end{pmatrix} \text{ s.t. } 0 < W_{ij} \le W \; \forall (i,j) \in \mathcal{E}$$

$$|\theta_i| \le T \; \forall i \in \mathcal{V}.$$

The required reparameterization for edge (i,j) takes $\theta_i \leftarrow \theta_i - W_{ij}/2$, hence reparameterizing all edges takes $\theta_i \leftarrow \theta_i - \sum_{j \in \mathsf{N}(i)} W_{ij}/2$. A sufficient condition for $\frac{1}{\eta_i(1-\eta_i)}$ to have a polynomial upper bound is that the maximum degree $\Delta := \max_{i \in \mathcal{V}} z_i = O(\log n)$, the same degree restriction as in [25]. In this case, $\frac{1}{\eta_i(1-\eta_i)} = O(e^{T+\Delta W/2})$.

Regarding Theorem 11, now $a = O(e^{W(1+\Delta)+2T})$ and $b = O(\Delta e^{W(1+\Delta/2)+T})$ with $\Omega = \max(a,b)$ and $\Sigma = O(\Delta/n)$ yielding the polynomial result.

6 Practical considerations

6.1 Improving A_i, B_i bounds

In practice, the runtime of our approach is dramatically improved if we obtain better bounds $[A_i, 1 - B_i]$ on the location of optima since then: (i) the search space to discretize is directly reduced, and (ii) the upper bound on Λ is decreased through lower Ω, thus a less fine mesh is required for a given level of accuracy. Extending the analysis of Section 3 leads to a novel algorithm to improve these bounds iteratively, which we term Bethe bound propagation (BBP), see Supplement for derivation and comments. The algorithm is shown below, where we suggest using THRESH= 0.002, MAXITER= 20. BBP runs very rapidly in time $O(|\mathcal{E}|)$ and can be used on general binary pairwise models (no need for associativity), sometimes leading to impressive results without further work.

Algorithm 1 BBP for a general binary pairwise model

```
{Initialize}
for all $i \in \mathcal{V}$ do
    $W_i = \sum_{j \in \mathsf{N}(i): W_{ij} > 0} W_{ij}$,
    $V_i = -\sum_{j \in \mathsf{N}(i): W_{ij} < 0} W_{ij}$,
    $A_i = \sigma(\theta_i - V_i)$, $B_i = 1 - \sigma(\theta_i + W_i)$[6]
end for
for all $(i,j) \in \mathcal{E}$ do
    $\alpha_{ij} = \exp(|W_{ij}|) - 1$
end for
{Main loop}
repeat
    for all $i \in \mathcal{V}$ do
        $L_i = 1$, $U_i = 1$ {Initialize for this pass}
        for all $j \in \mathsf{N}(i)$ do
            if $W_{ij} > 0$ then
                {Associative edge}
                $L_i * = 1 + \frac{\alpha_{ij} A_j}{1 + \alpha_{ij}(1-B_i)(1-A_j)}$
                $U_i * = 1 + \frac{\alpha_{ij} B_j}{1 + \alpha_{ij}(1-A_i)(1-B_j)}$
            else
                {Repulsive edge}
                $L_i * = 1 + \frac{\alpha_{ij} B_j}{1 + \alpha_{ij}(1-B_i)(1-B_j)}$
                $U_i * = 1 + \frac{\alpha_{ij} A_j}{1 + \alpha_{ij}(1-A_i)(1-A_j)}$
            end if
        end for
        $A_i = 1/(1 + \exp(-\theta_i + V_i)/L_i)$
        $B_i = 1/(1 + \exp(\theta_i + W_i)/U_i)$
    end for
until All $A_i$,$B_i$ changed by < THRESH or run MAXITER times
```

A different approach, which we term MK, was derived in [19], based on considering the set of possible beliefs after iterating LBP, starting from any initial values. Since any minimum of the Bethe free energy corresponds to a fixed point of LBP [36], this method may be used as an alternative to BBP. MK considers cavity fields around each variable, which requires more time (often by orders of magnitude), but the bounds obtained are no worse, and sometimes significantly better. In difficult cases, the additional time required to run MK is more than compensated by faster runtime for the later, graph cuts part of the overall algorithm, and hence was our preferred experimental method.

6.2 Experiments

See section 1.2 for the overall algorithm, and Figure 1 for results. Theoretical bounds have no units and are shown scaled to fit the axes. For all experiments, $\epsilon = 0.01$ was used; once the discretized multilabel MAP problem was formed, it was solved via the Schlesinger-Flach construction [23] to reduce to a binary max-flow problem and then using the Boykov-Kolmogorov algorithm [2]. Typically, inference methods are more challenged as the number of variables, n, increases, or as the number and strength of edge interactions increase relative to single variable potentials.

Models are specified using the notation of section 5.3.

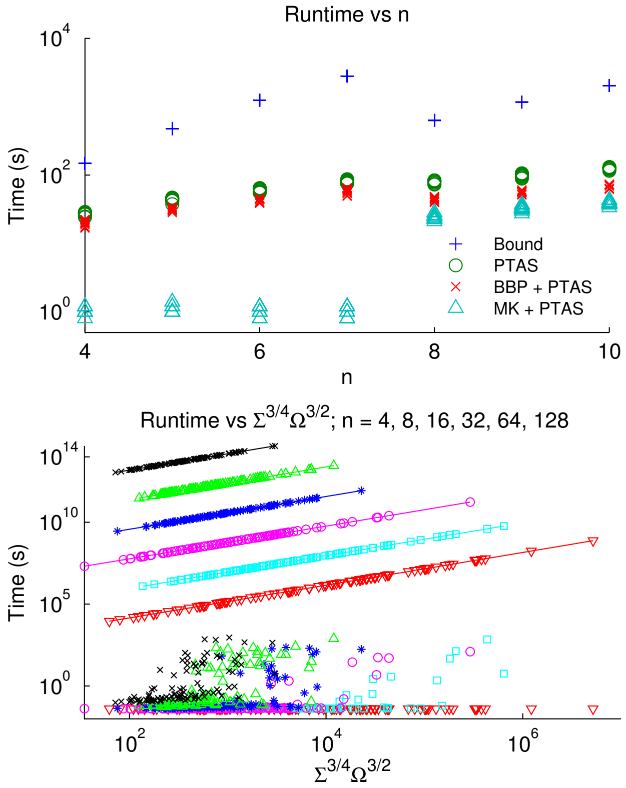

Figure 1: Experiment results, see text for details.

The top graph shows the effect on runtime of varying n while using different approaches to compute the $\{A_i, B_i\}$ bounds. The slowest runs use the PTAS bounds from Theorem 3. We also show the improvement in overall performance obtained by using BBP or MK to improve these initial $\{A_i, B_i\}$ bounds. To isolate these effects, for each n, 12 random connected regular graphs of uniform degree d_n, with $T = 0$ and constant edge weights W_n were used. We set $d_n = 2\lfloor \ln n \rfloor$ and fixed $d_n W_n = 6$.

The lower graph shows results using MK for random connected Erdős-Rényi graphs with edge probability $\frac{\log_2 n}{n-1}$, hence expected degree $d_n = \log_2 n$. $T = 0$ and each $W_{ij} \sim \text{Uniform}[0, W_n]$ where $d_n W_n = 8$. For each n, 100 instances were generated and the runtime is shown against the instance-specific terms of the theoretical bound, given by $\Sigma^{3/4}\Omega^{3/2}$. The empirical worst-case runtimes follow the shape of the theoretical bound shown higher on the graph, though performance is often much faster.

7 Conclusion & Extensions

To our knowledge, we have proved the first PTAS for the global optimum of the Bethe free energy of an associative binary pairwise MRF[7]. In doing so, we derived a range of results, including several for general edges and models (associative or not), which may prove useful in their own right, including our results on second derivatives and Bethe bound propagation. The approach is useful in practice, especially when combined with BBP or MK for initial bounds.

Although the algorithm is only weakly polynomial, we are not sure if more is possible. If input parameters are unrestricted, then potentially α values could be infinite, corresponding to distributions with zero probability for some states (which may be reasonable), which will lead to infinite derivatives as some pseudomarginal entries will be driven to 0.

[30] has shown that graph cuts is in a strong sense equivalent to max-product belief propagation with careful scheduling and damping. Together with our result this shows an interesting link between max-product and sum-product techniques. One direction to explore is how sum-product belief propagation fares using a scheme similar to [30].

Our approach immediately also applies to approximating optimum mean field marginals. In addition, it may readily extend to allow approximate marginal inference for multi-label and third order submodular MRFs, both of which can be mapped to equivalent associative binary pairwise MRFs [23, 21].

Acknowledgments

We are grateful to the anonymous reviewers, Joris Mooij and Nicholas Ruozzi for helpful comments, and to Kui Tang for running experiments. This material is based upon work supported by the National Science Foundation under Grant No. 1117631.

[7] Although the theoretical result guaranteeing polynomial runtime requires $\Delta = O(\log n)$, in practice, for any (dense) associative binary pairwise model, BBP or MK can first be run quickly and then in many cases, the improved $\{A_i, B_i\}$ bounds and resulting Ω will ensure that the overall algorithm will run efficiently.

References

[1] B. Bidyuk and R. Dechter. An anytime scheme for bounding posterior beliefs. In *AAAI*, pages 1095–1100, 2006.

[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.

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

[4] P. Dagum and E. Horvitz. A Bayesian analysis of simulation algorithms for inference in belief networks. *Networks*, 23:499–516, 1993.

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

[6] R. Dechter. Mini-buckets: A general scheme of generating approximation in automated reasoning. In *Proceedings of the Fifteenth International Joint Conference on Artificial Intelligence*, 1997.

[7] V. Ganapathi, D. Vickrey, J. Duchi, and D. Koller. Constrained approximate maximum entropy learning of Markov random fields. In *Uncertainty in Artificial Intelligence*, 2008.

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

[9] 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.

[10] U. Heinemann and A. Globerson. What cannot be learned with Bethe approximations. In F. G. Cozman and A. Pfeffer, editors, *UAI*, pages 319–326. AUAI Press, 2011.

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

[12] T. Heskes. Convexity arguments for efficient minimization of the Bethe and Kikuchi free energies. *Journal of Artificial Intelligence Research*, 26:153–190, 2006.

[13] A. T. Ihler. Accuracy bounds for belief propagation. In *Proceedings of the Twenty-Third Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-07)*, pages 183–190, Corvallis, Oregon, 2007. AUAI Press.

[14] T. Jaakkola and M. Jordan. Variational probabilistic inference and the QMR-DT network. *Journal of Artificial Intelligence Research*, 10:291–322, 1999.

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

[16] F. Korc, V. Kolmogorov, and C. Lampert. Approximating marginals using discrete energy minimization. Technical report, IST Austria, 2012.

[17] M. A. R. Leisink and H. J. Kappen. Bound propagation. *J. Artif. Intell. Res. (JAIR)*, 19:139–154, 2003.

[18] D. MacKay. *Learning in graphical models*, chapter Introduction to Monte Carlo methods. MIT Press, 1998.

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

[20] J. M. Mooij and H. J. Kappen. Bounds on marginal probability distributions. In *Neural Information Processing Systems*, pages 1105–1112, 2008.

[21] S. Ramalingam, P. Kohli, K. Alahari, and P. Torr. Exact inference in multi-label CRFs with higher order cliques. In *Computer Vision and Pattern Recognition*, 2008.

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

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

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

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

[26] M. Shwe and G. Cooper. An empirical analysis of likelihood-weighting simulation on a large, multiply connected medicial belief network. *Computers and Biomedical Research*, 24:453–475, 1991.

[27] M. Shwe, B. Middleton, D. Heckerman, M. Henrion, E. Horvitz, H. Lehmann, and G. Cooper. Probabilistic diagnosis using a reformulation of the INTERNIST-1/QMR knowledge base: Part-i. *Methods of Information in Medicine*, 30:241–255, 1991.

[28] A. Sinclair and M. Jerrum. Approximate counting, uniform generation and rapidly mixing Markov chains. *Information and Computation*, 82(1):93–133, 1989.

[29] R. Szeliski, R. Zabih, D. Scharstein, O. Veksler, V. Kolmogorov, A. Agarwala, M. Tappen, and C. Rotheret. A comparative study of energy minimization methods for Markov random fields. In *ECCV*, 2006.

[30] D. Tarlow, I. Givoni, R. Zemel, and B. Frey. Graph cuts is a max-product algorithm. In F. Gagliardi Cozman and A. Pfeffer, editors, *UAI*, pages 671–680. AUAI Press, 2011.

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

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

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

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

[35] M. Welling and Y.W. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Uncertainty in Artificial Intelligence*, 2001.

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

[37] J. Yedidia, W. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *IEEE Trans. Information Theory*, pages 2282–2312, 2005.

[38] X. Zhan. Extremal eigenvalues of real symmetric matrices with entries in an interval. *SIAM J. Matrix Analysis Applications*, 27(3):851–860, 2005.

8 APPENDIX - SUPPLEMENTARY MATERIAL

Here we provide proofs of several of the results in the main paper, using the original numbering. To establish these, we also derive additional preliminary results where required, starting with Lemma number 12. We hope this numbering aids clarity rather than confusion.

Section 3

Lemma 1. Flipping any set of variables changes affected pseudo-marginal matrix entries' locations but not values. The Bethe free energy is unchanged up to a constant, hence the locations of stationary points are unaffected.

Proof. By construction, energies are the same up to a constant. The singleton entropies are symmetric functions of q_i and $1-q_i$ so are unaffected. The impact on pseudo-marginal matrix entries follows directly from definitions. Thus Bethe entropy is unaffected. □

Lemma 5. For $q_i, q_j \in [0,1], 0 \leq q_i + q_j - 2q_iq_j \leq 1$.

Proof. Let $f = q_i+q_j-2q_iq_j$. To show the left inequality, consider $m = \min(q_i, q_j)$ and $M = \max(q_i, q_j)$, then $f \geq 2m(1-M) \geq 0$. For the right inequality observe $1-f = (1-q_i)(1-q_j) + q_iq_j \geq 0$. □

Lemma 12. *Unless q_i or $q_j \in \{0,1\}$, all entries of the pseudo-marginal μ_{ij} are strictly > 0, whether (i,j) is associative or repulsive.*[8]

Proof. First assume $\alpha_{ij} > 0$. Considering (2) and using Lemmas 2 and 6, we have that element-wise

$$\mu_{ij} \geq \begin{pmatrix} (1-q_i)(1-q_j) & q_j(1-q_i)/(1+\alpha_{ij}) \\ q_i(1-q_j)/(1+\alpha_{ij}) & q_iq_j \end{pmatrix} \tag{13}$$

which proves the result for this case. If $\alpha_{ij} < 0$ then flip either q_i or q_j. As in the proof of Lemma 1, pseudo-marginal entries change position but not value. □

Section 4

Theorem 7. For any edge (i,j), for any α_{ij}, writing $f = f_{ij}$ and $\mu_{ab} = \mu_{ij}(a,b)$ from (2),

$$\frac{\partial^2 f}{\partial q_i^2} = \frac{1}{T_{ij}} q_j(1-q_j)$$
$$\frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} = \frac{1}{T_{ij}}(\mu_{01}\mu_{10} - \mu_{00}\mu_{11})$$
$$\frac{\partial^2 f}{\partial q_j^2} = \frac{1}{T_{ij}} q_i(1-q_i)$$

where $T_{ij} = q_iq_j(1-q_i)(1-q_j) - (\xi_{ij} - q_iq_j)^2 \geq 0$ with equality only for q_i or $q_j \in \{0,1\}$. Further $\mu_{01}\mu_{10} - \mu_{00}\mu_{11} = q_iq_j - \xi_{ij}$ and has the sign of $-\alpha_{ij}$.

Proof. We begin with the same approach as [16] but extend the analysis and derive stronger results.

For notational convenience add a third pseudo-dimension restricted to the value 1. Let $\mathbf{y} = (y_1, y_2, y_3)$ be the vector with components $y_1 = x_i$, $y_2 = x_j$ and $y_3 = 1$ where $x_i, x_j \in \mathbb{B}$. Define $\pi(\mathbf{y}) = \mu_{ij}(x_i, x_j)$, and $\phi(\mathbf{y}) = W_{ij}$ if $\mathbf{y} = (1,1,1)$ or $\phi(\mathbf{y}) = 0$ otherwise. Let $\mathbf{r} = (q_i, q_j, 1)$. Define function h used in entropy calculations as $h(z) = -z \log z$.

Consider (5) but instead of solving for ξ_{ij} explicitly, express f as an optimization problem, minimizing free energy subject to local consistency and normalization constraints in order to use techniques from convex optimization. We have $f(q_i, q_j) = g(\mathbf{r})$ where

$$g(\mathbf{r}) = \min_{\pi} \sum_{\mathbf{y}} \big(-\phi(\mathbf{y})\pi(\mathbf{y}) - h(\pi(\mathbf{y}))\big)$$
$$\text{s.t.} \sum_{\mathbf{y}:y_k=1} \pi(\mathbf{y}) = r_k \;\; k = 1,2,3. \tag{14}$$

The Lagrangian can be written as

$$L_{\mathbf{r}}(\pi, \boldsymbol{\lambda}) = \sum_{\mathbf{y}} [(-\phi(\mathbf{y}) - \langle \mathbf{y}, \boldsymbol{\lambda} \rangle)\pi(\mathbf{y}) - h(\pi(\mathbf{y}))] + \langle \mathbf{r}, \boldsymbol{\lambda} \rangle$$

and its derivative is

$$\frac{\partial L_{\mathbf{r}}(\pi, \boldsymbol{\lambda})}{\partial \pi} = -\phi(\mathbf{y}) - \langle \mathbf{y}, \boldsymbol{\lambda} \rangle + 1 + \log \pi$$

which yields a minimum at

$$\pi_{\boldsymbol{\lambda}}(\mathbf{y}) = \exp(\phi(\mathbf{y}) + \langle \mathbf{y}, \boldsymbol{\lambda} \rangle - 1). \tag{15}$$

Since the minimization problem in (14) is convex and satisfies the weak Slater's condition (the constraints are affine), strong duality applies and $g(\mathbf{r}) = \max_{\boldsymbol{\lambda}} G(\mathbf{r}, \boldsymbol{\lambda}) = G(\mathbf{r}, \boldsymbol{\lambda}^*(\mathbf{r}))$ where the dual is simply

$$G(\mathbf{r}, \boldsymbol{\lambda}) = \min_{\pi} L_{\mathbf{r}}(\pi, \boldsymbol{\lambda}) = -\sum_{\mathbf{y}} \pi_{\boldsymbol{\lambda}}(\mathbf{y}) + \langle \mathbf{r}, \boldsymbol{\lambda} \rangle. \tag{16}$$

[8] Here we assume α_{ij} is finite, see footnote 1.

Let $D_k(\mathbf{r},\boldsymbol{\lambda}) = \frac{\partial G(\mathbf{r},\boldsymbol{\lambda})}{\partial \lambda_k}$ then $D_k(\mathbf{r},\boldsymbol{\lambda}^*) = 0,\ k = 1,2,3.$

Hence $\frac{\partial g}{\partial r_k} = \frac{\partial G}{\partial r_k} = \lambda_k$ using (16). Focusing on our goal of obtaining second derivatives, we consider $\frac{\partial^2 g}{\partial r_l \partial r_k} = \frac{\partial \lambda_k}{\partial r_l}$ which we shall express in terms of $C_{kl} := \frac{\partial^2 G}{\partial \lambda_l \partial \lambda_k} = \frac{\partial D_k}{\partial \lambda_l}$.

Differentiating $D_k(\mathbf{r},\boldsymbol{\lambda}^*) = 0$ with respect to r_l,

$$0 = \frac{\partial D_k(\mathbf{r},\boldsymbol{\lambda}^*)}{\partial r_l} = \frac{\partial D_k}{\partial r_l} + \sum_{p=1}^{3} \frac{\partial D_k}{\partial \lambda_p}\frac{\partial \lambda_p}{\partial r_l} \quad k,l = 1,2,3.$$

Considering (16), $\frac{\partial D_k}{\partial r_l} = \frac{\partial^2 G}{\partial r_l \partial \lambda_k} = \delta_{kl}$ hence $0 = \delta_{kl} + \sum_p C_{kp} \frac{\partial^2 g}{\partial r_l \partial r_p}$. Thus $\frac{\partial^2 g}{\partial r_l \partial r_k} = -[C^{-1}]_{kl}$. Using its definition and (16), we have

$$\begin{aligned} C_{kl} = \frac{\partial^2 G}{\partial \lambda_l \partial \lambda_k} &= \frac{\partial}{\partial \lambda_l}\Big(-\sum_{\mathbf{y}} y_k \pi_{\boldsymbol{\lambda}}(\mathbf{y}) + r_k \Big) \\ &= -\sum_{\mathbf{y}} y_k y_l \pi_{\boldsymbol{\lambda}}(\mathbf{y}) = -\sum_{\mathbf{y}: y_k = y_l = 1} \pi_{\boldsymbol{\lambda}}(\mathbf{y}). \end{aligned}$$

Earlier work [16] stopped here, recognizing that $\det C \leq 0$. We more precisely characterize this matrix

$$C = -\begin{pmatrix} \mu_{10}+\mu_{11} & \mu_{11} & \mu_{10}+\mu_{11} \\ \mu_{11} & \mu_{01}+\mu_{11} & \mu_{01}+\mu_{11} \\ \mu_{10}+\mu_{11} & \mu_{01}+\mu_{11} & 1 \end{pmatrix} \qquad (17)$$

Recall constraints $\mu_{00}+\mu_{01}+\mu_{10}+\mu_{11} = 1$, $\mu_{01}+\mu_{11} = q_j$, $\mu_{10}+\mu_{11} = q_i$. Note C is symmetric.

Applying the result above and using Cramer's rule,

$$\frac{\partial^2 f}{\partial q_i^2} = \frac{\partial^2 g}{\partial r_1^2} = -\frac{1}{\det C}(\mu_{01}+\mu_{11})(\mu_{00}+\mu_{10}) = \frac{q_j(1-q_j)}{-\det C}$$

$$\frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} = \frac{\partial^2 g}{\partial r_1 \partial r_2} = \frac{(\mu_{01}\mu_{10} - \mu_{00}\mu_{11})}{-\det C}$$

$$\frac{\partial^2 f}{\partial q_j^2} = \frac{\partial^2 g}{\partial r_2^2} = -\frac{1}{\det C}(\mu_{10}+\mu_{11})(\mu_{00}+\mu_{01}) = \frac{q_i(1-q_i)}{-\det C}.$$

Using (17) and simplifying, we obtain $-\det C = \mu_{00}\mu_{10}\mu_{11} + \mu_{10}\mu_{11}\mu_{01} + \mu_{11}\mu_{10}\mu_{00} + \mu_{01}\mu_{00}\mu_{10}$. By Lemma 12 this is strictly > 0 unless q_i or $q_j \in \{0,1\}$. Substituting in terms from (2) and simplifying establishes $-\det C = T_{ij}$ from the statement of the theorem, and $\mu_{01}\mu_{10} - \mu_{00}\mu_{11} = q_i q_j - \xi_{ij}$. The sign follows from Lemma 2 or observing from (15) that $\frac{\mu_{00}\mu_{11}}{\mu_{01}\mu_{10}} = e^{W_{ij}} = \alpha_{ij} + 1$. □

Lemma 13 (Finite 3rd derivatives)**.** *For any edge (i,j) with $\alpha_{ij} > 0$, if $q_i, q_j \in (0,1)$ then all third derivatives exist and are finite.*

Proof. Using Theorem 7 noting $T_{ij} > 0$ strictly and considering (2), it is sufficient to show $\frac{\partial \xi_{ij}}{\partial q_k}$ is finite. We may assume $k \in \{i,j\}$ else the derivative is 0 and by symmetry need only check $\frac{\partial \xi_{ij}}{\partial q_i}$. Differentiating (3),

$$\frac{\partial \xi_{ij}}{\partial q_i} = \frac{\alpha_{ij}(q_j - \xi_{ij}) + q_j}{1 + \alpha_{ij}(q_i - \xi_{ij} + q_j - \xi_{ij})},$$

clearly finite for $\alpha_{ij} > 0$ since recalling (2), $q_i - \xi_{ij}$ and $q_j - \xi_{ij}$ are elements of the pseudo-marginal and hence are non-negative (or use Lemma 6). □

Theorem 8. If a binary pairwise MRF is submodular on an edge (i,j), i.e. $\alpha_{ij} > 0$, then the multi-label discretized MRF for any discretization $\mathcal{D}$ is submodular for that edge. In particular, if the MRF is fully associative/submodular, i.e. $\alpha_{ij} > 0\ \forall (i,j) \in \mathcal{E}$, then the multi-label discretized MRF is fully submodular for any discretization.

Proof. For any edge (i,j), let f be the pairwise function f_{ij} from (5) and note the submodularity requirement (7). Let $x = (x_1,x_2)$, $y = (y_1,y_2)$ be any points in $[0,1]^2$. Define $s(x,y) = (s_1,s_2) = (\min(x_1,y_1), \min(x_2,y_2))$, and $t(x,y) = (t_1,t_2) = (\max(x_1,y_1), \max(x_2,y_2))$. Let $g(x,y) = f(s_1,s_2) + f(t_1,t_2) - f(s_1,t_2) - f(s_2,t_1)$, call this the submodularity of the rectangle defined by x,y. We must show $g(x,y) \leq 0$. Note f is continuous in $[0,1]^2$ hence so also is g. We shall show that $\forall (x,y) \in (0,1)^2$, $g(x,y) < 0$ then the result follows by continuity.

Assume $x,y \in (0,1)^2$. Consider derivatives of f in the compact set $R = [s_1,t_1] \times [s_2,t_2]$. Using (6) and Lemma 12, first derivatives exist and are bounded. By Theorem 7 and Lemma 13 the same holds for second and third derivatives. Further, Theorem 7 and Lemma 14 show that $\frac{\partial^2 f}{\partial q_i \partial q_j} = \frac{\partial^2 f}{\partial q_j \partial q_i} < 0$.

If a rectangle is sliced fully along each dimension so as to be subdivided into sub-rectangles then summing the submodularities of all the sub-rectangles, internal terms cancel and we obtain the submodularity of the original rectangle.

Hence there exists an ϵ such that if we subdivide the rectangle defined by x,y into sufficiently small sub-rectangles with sides $< \epsilon$ and apply Taylor's theorem up to second order with the remainder expressed in terms of the third derivative evaluated in the interval, then the second order terms dominate and the submodularity of each small sub-rectangle < 0. Summing over all sub-rectangles provides the result. □

Section 6

In order to derive our approach of Bethe bound propagation (BBP), we extend the analysis of bounds on ξ_{ij} from Section 3.

Lemma 14 (Better lower bound for ξ_{ij})**.** *If $\alpha_{ij} > 0$, then $\xi_{ij} \geq q_i q_j + \alpha_{ij} q_i q_j (1-q_i)(1-q_j)/[1+\alpha_{ij}(q_i + q_j - 2q_i q_j)]$, equality only possible at an edge, i.e. one or both of $q_i, q_j \in \{0,1\}$.*

Proof. Write $\xi_{ij} = q_i q_j + y$ and substitute into (3),

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

We have a convex parabola which at $y=0$ is above the abscissa (unless q_i or $q_j \in \{0,1\}$) and has negative gradient by Lemma 5. Hence all roots are at $y \geq 0$ and given convexity we can bound below using the tangent at $y=0$ which yields the result. □

Bethe bound propagation (BBP)

We have already derived bounds on stationary points in Theorems 3 and 4. Here we show for variables with only associative edges how we can iteratively improve these bounds, sometimes with striking results. Note that a fully associative model is not required, and as in section 3.2, *any* model may be selectively flipped to yield local associativity around a particular node.

We first assume all $\alpha_{ij} \geq 0$ and adopt the approach of Theorem 3, now using the better bound from Lemma 14 to obtain

$$\begin{aligned} q_i - \xi_{ij} &\leq q_i - q_i q_j - \frac{\alpha_{ij} q_i q_j (1-q_i)(1-q_j)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \\ &= q_i(1-q_j)\Big[1 - \frac{\alpha_{ij} q_j (1-q_i)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)}\Big], \end{aligned}$$

$$\begin{aligned} &1 + \xi_{ij} - q_i - q_j \geq \\ &1 + q_i q_j - q_i - q_j + \frac{\alpha_{ij} q_i q_j (1-q_i)(1-q_j)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)} \\ &= (1-q_i)(1-q_j)\Big[1 + \frac{\alpha_{ij} q_i q_j}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)}\Big]. \end{aligned}$$

Hence $Q_i \leq \frac{q_i}{1-q_i} \prod_{j \in \mathsf{N}(i)} R_{ij}^{-1}$ where

$$R_{ij} = \frac{1 + \frac{\alpha_{ij} q_i q_j}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)}}{1 - \frac{\alpha_{ij} q_j (1-q_i)}{1+\alpha_{ij}(q_i+q_j-2q_iq_j)}} = 1 + \frac{\alpha_{ij} q_j}{1+\alpha_{ij} q_i (1-q_j)},$$

monotonically increasing with q_j and decreasing with q_i. Hence

$$e^{W_{ij}} = 1+\alpha_{ij} \geq R_{ij} \geq L_{ij} := 1 + \frac{\alpha_{ij} A_j}{1+\alpha_{ij}(1-B_i)(1-A_j)} \quad (18)$$

Using Theorem 3, we initialize $A_i = \sigma(\theta_i)$ and $B_i = 1 - \sigma(\theta_i + W_i)$.

Using (6), at any stationary point we must have

$$q_i \geq 1/[1+\exp(-\theta_i)/L_i]$$

where $L_i = \prod_{j \in \mathsf{N}(i)} L_{ij}$. Intuitively, in an associative model, if variable i has neighbors j which are likely to be 1 (i.e. high A_j) then this pulls up the probability that i will be 1 (i.e. raises A_i).

Flipping all variables,

$$1 - q_i \geq 1/[1+\exp(\theta_i + W_i)/U_i]$$

where $U_i = \prod_{j \in \mathsf{N}(i)} U_{ij}$ with

$$e^{-W_{ij}} \geq U_{ij} := 1 + \frac{\alpha_{ij} B_j}{1+\alpha_{ij}(1-A_i)(1-B_j)}.$$

It is also possible to write this as

$$\sigma(\theta_i + \log L_i) \leq q_i \leq \sigma(\theta_i + W_i - \log U_i).$$

This establishes a message passing type of algorithm for iteratively improving the bounds $\{A_i, B_i\}$. Repeat until convergence:

$$\begin{aligned} \text{new } A_i &\leftarrow (1+\exp(-\theta_i)/L_i)^{-1} \\ \text{new } B_i &\leftarrow (1+\exp(\theta_i+W_i)/U_i)^{-1} \\ &\text{recompute } L_i, U_i \text{ using new } A_i, B_i. \end{aligned}$$

Lemma 15. *At every iteration, all of A_i, B_i, L_{ij}, U_{ij} monotonically increase.*

Proof. All of the dependencies are monotonically increasing on all inputs. The first iteration yields an increase since each $L_{ij}, U_{ij} > 1$. □

Since $A_i + B_i \leq 1$, each is bounded above and we achieve monotonic convergence. Combining this with the main global optimization approach can dramatically reduce the range of values that need be considered, leading to significant time savings. Convergence is rapid even for large, densely connected graphs. Each iteration takes $O(|\mathcal{E}|)$ time; a good heuristic is to run for up to 20 iterations, terminating early if all parameters improve by less than a threshold value. This adds negligible time to the global optimization.

This procedure alone can produce impressive results. For example, running on a 100-node graph with independent random edge probability 0.04 (hence average degree 4), each W_{ij} and θ_i drawn randomly from Uniform $[0,1]$ and then adjusting $\theta_i \leftarrow \theta_i - \sum_{j \in \mathsf{N}(i)} W_{ij}/2$ in order to be unbiased, convergence takes about 11 iterations yielding final average bracket width of 0.05 after starting with average bracket width of 0.40. Greater connectivity, higher edge strengths and smaller individual node potentials make the problem more challenging and may widen the returned final brackets significantly.

BBP for general models

A repulsive edge (i, j) may always be flipped to associative by flipping variable j, which flips its Bethe bounds $A_j \leftrightarrow B_j$. Using Theorem 4 we can extend the analysis above to run BBP on any model, see Algorithm 1 in section 6. Performance in terms of convergence speed and final bracket width is similar for associative and non-associative models.