Bethe and Related Pairwise Entropy Approximations

Adrian Weller
Department of Engineering
University of Cambridge
aw665@cam.ac.uk

Abstract

For undirected graphical models, belief propagation often performs remarkably well for approximate marginal inference, and may be viewed as a heuristic to minimize the Bethe free energy. Focusing on binary pairwise models, we demonstrate that several recent results on the Bethe approximation may be generalized to a broad family of related pairwise free energy approximations with arbitrary counting numbers. We explore approximation error and shed light on the empirical success of the Bethe approximation.

1 INTRODUCTION

Undirected graphical models, also called Markov random fields (MRFs), have become a central tool in machine learning, providing a powerful and compact way to describe relationships between variables. Fundamental problems are to compute the normalizing partition function, and to solve for the marginal distribution of a subset of variables (marginal inference). Both tasks are computationally intractable (Cooper, 1990), prompting great interest in approximate algorithms that perform well. One popular approach is *belief propagation* (BP, Pearl, 1988). When the underlying model topology is acyclic, this returns exact values in linear time. If the method is applied to models with cycles, termed *loopy belief propagation* (LBP), results are often strikingly good but not always, and it may not converge at all (McEliece et al., 1998).

Yedidia et al. (2001) demonstrated that fixed points of LBP correspond to stationary points of the *Bethe free energy* $\mathcal{F}_B$ (Bethe, 1935), see §2 for definitions. Further, Heskes (2002) showed that stable fixed points correspond to local minima of the Bethe free energy. In this paper, we summarize recent results on the Bethe approximation (Welling and Teh, 2001; Weller and Jebara, 2013, 2014a,b; Weller et al., 2014), and in each case consider how the result may be generalized by considering the broad class of pairwise entropy approximations specified by arbitrary *counting numbers*, which includes the Bethe and *tree-reweighted* approximations (TRW, Wainwright et al., 2005) as special cases. We discuss consequences and related applications, including in §5 minimizing the approximate free energy, which Weller and Jebara (2014a) recently showed, for the specific case of the Bethe approximation on attractive models, can be approximated to any ϵ-accuracy with a *fully polynomial-time approximation scheme* (FPTAS).

In §6, we compare this family of entropy approximations to the *true* entropy, and consider how differences interact with the other form of approximation typically employed: the marginal polytope, which enforces global variable consistency, is relaxed to the local polytope, which enforces only local (pairwise) consistency. We also provide fresh insights on balanced and frustrated cycles by considering the loop series approach of Sudderth et al. (2007).

1.1 RELATED WORK

Related work is discussed throughout the text but here we clarify the context and contributions of our results up to §5 that build to show how to approximate the global optimum of the approximate free energy to arbitrary accuracy for general counting numbers.

Context. All for attractive binary pairwise models: The problem of identifying a most probable configuration (MAP inference) is solvable in polynomial-time via graph cuts (Greig et al., 1989); this generalizes to multi-label pairwise models with submodular cost functions (Schlesinger and Flach, 2006). However, aside from restricted cases (e.g. low treewidth or the *fully polynomial-time randomized approximation scheme* (FPRAS) of Jerrum and Sinclair (1993) for uniform external field), there is no way to estimate the partition function Z accurately in polynomial-time. LBP is a heuristic to find the Bethe partition function by minimizing the Bethe free energy, with $\log Z_B = -\min \mathcal{F}_B$, and for these mod-

els we know that Z_B is a lower bound and usually a good estimate of Z (Sudderth et al., 2007; Ruozzi, 2012; Weller and Jebara, 2014b), but LBP may find only a local optimum or not converge at all. Various methods (e.g. CCCP, Yuille, 2002) were introduced which converge but only to a local minimum of $\mathcal{F}_B$ with no time guarantee. Shin (2012) introduced the first polynomial-time method but this returns an approximately stationary point of the Bethe $\mathcal{F}_B$ (i.e. a point where |derivative of $\mathcal{F}_B| < \epsilon$, which is useful for loop series methods, but this point may have $\mathcal{F}_B$ value far from the global optimum; attractive not required) subject to a sparsity condition that max degree is $O(\log n)$. Weller and Jebara (2013) derived a PTAS for the global optimum of $\mathcal{F}_B$ with the same sparsity condition. Weller and Jebara (2014a) improved this, providing the first FPTAS for $\log Z_B$ for an attractive model with any topology. These applied only for the Bethe approximation.

Contributions. Here we broaden analysis significantly to consider any counting numbers, relying on our new Theorems 2, 5, 6 and 7, and Lemmas 3 and 4. All these extend previous results that applied only to the Bethe approximation. It is somewhat remarkable that it emerges that an attractive model admits a FPTAS for $\log Z_A$ for any counting numbers. This is significant theoretically and will allow the benefits of non-convex free energy approximations to be explored further in future work. Theorems 2, 5 and 6 importantly apply to general (non-attractive models), as does Algorithm 1, allowing $\log Z_A$ with any counting numbers to be computed to arbitrary accuracy, though with no polynomial-time guarantee if not attractive - still this will be useful to learn insights from small models and to benchmark accuracy of faster methods.

2 PRELIMINARIES

We adopt notation consistent with (Welling and Teh, 2001; Weller and Jebara, 2013, 2014a,b). Consider a binary pairwise model with n variables $X_1, \ldots, X_n \in \mathbb{B} = \{0, 1\}$ and graph topology $(\mathcal{V}, \mathcal{E})$ with $m = |\mathcal{E}|$ edges; that is $\mathcal{V}$ contains nodes $\{1, \ldots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq V \times V$ contains an edge for each pairwise score relationship. Let $\mathcal{N}(i)$ be the neighbors of i. Let $x = (x_1, \ldots, x_n)$ be one particular configuration, and define its *energy* $E(x)$ via the relationships

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

where the partition function $Z = \sum_x e^{-E(x)}$ is the normalizing constant, and $\{\theta_i, W_{ij}\}$ specify the potentials of the model.[1] If $W_{ij} \geq 0$, the edge (i, j) is *attractive* (tending to pull its variables toward the same value); if $W_{ij} < 0$ then it is *repulsive* (tending to push apart its variables to different values). A model is attractive iff all its edges are attractive.

2.1 VARIATIONAL INFERENCE AND COUNTING NUMBERS

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

Since this minimization is often computationally intractable, two pairwise approximations are typically made:

1. The marginal polytope is relaxed to the *local polytope* $\mathbb{L}$, where only *local* consistency is required - that is we deal with a *pseudomarginal* vector q, which in our context may be considered $\{q_i = q(X_i = 1)\ \forall i \in \mathcal{V}, \mu_{ij}(x_i, x_j) = q(x_i, x_j)\ \ \forall (i,j) \in \mathcal{E}\}$, subject to constraints $q_i = \sum_{x_j \in \mathbb{B}} \mu_{ij}(1, x_j), q_j = \sum_{x_i \in \mathbb{B}} \mu_{ij}(x_i, 1)\ \forall (i,j) \in \mathcal{E}$.

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

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

for some $\xi_{ij} \in [\max(0, q_i + q_j - 1), \min(q_i, q_j)]$.

Thus we may adopt a minimal representation with pseudomarginals specified by $\{q_i\ \ \forall i \in \mathcal{V}\}$ singleton and $\{\xi_{ij}\ \forall (i,j) \in \mathcal{E}\}$ pairwise terms.

2. The entropy S is replaced by an approximation S_A that incorporates singleton and pairwise entropy terms via *counting numbers* $\{c_i\ \forall i \in \mathcal{V}, \rho_{ij}\ \forall (i,j) \in \mathcal{E}\}$:

$$S_A(q) = \sum_{i \in \mathcal{V}} c_i S_i - \sum_{(i,j) \in \mathcal{E}} \rho_{ij} I_{ij}. \quad (3)$$

Here $S_i(q_i)$ is the entropy of the singleton distribution of X_i, and $I_{ij}(\mu_{ij})$ is the mutual information of edge (i, j) given by $I_{ij} = S_i + S_j - S_{ij}$, where $S_{ij}(\mu_{ij})$ is the entropy of the pairwise distribution μ_{ij}. Note that always $I_{ij} \geq 0$.[2]

In this paper, we shall consider the approximate partition function Z_A obtained by minimizing the corresponding approximate free energy $\mathcal{F}_A$, defined as follows,

$$-\log Z_A = \min_{q \in \mathbb{L}} \mathcal{F}_A(q), \quad \mathcal{F}_A(q) = \mathbb{E}_q(E) - S_A(q). \quad (4)$$

We shall also be interested in the approximate marginals given by the $\arg\min$ of (4).

[1] It is easily shown (Wainwright and Jordan, 2008) that any binary pairwise model may be reparameterized to the form in (1). Eaton and Ghahramani (2013) showed that any discrete model may be arbitrarily well approximated by a binary pairwise model, though the state space may be large.

[2] Some instead define $S_A = \sum_{i \in \mathcal{V}} c'_i S_i + \sum_{(i,j) \in \mathcal{E}} c'_{ij} S_{ij}$, which is equivalent via $c'_{ij} = \rho_{ij}, c'_i = c_i - \sum_{j \in \mathcal{N}(i)} \rho_{ij}$.

2.2 CHOICE OF COUNTING NUMBERS

In the standard Bethe entropy approximation S_B, all counting numbers c_i and ρ_{ij} are set to 1. This often performs very well, yet leads to a non-convex approximate free energy $\mathcal{F}_B$ that can be hard to optimize.

Another choice yields the well-known *tree-reweighted* approximation (TRW, Wainwright et al., 2005) S_T. Here again all $c_i = 1$ but now the edge weights ρ_{ij} are selected from the *spanning tree polytope*, resulting in all $\rho_{ij} \leq 1$. Since $I_{ij} \geq 0$, this immediately implies that $S_T \geq S_B$, and hence $Z_T \geq Z_B$. It is also known that TRW values are bounded by true values in that $S_T \geq S$, hence $Z_T \geq Z$ (whereas for many counting numbers, S_A may be above or below S, similarly Z_A may be above or below Z; indeed, in some cases including Bethe, S_A may even be negative). We note also that S_T is concave leading to the corresponding free energy approximation $\mathcal{F}_T$ being convex, allowing easier optimization.

Other choices of counting numbers yield a rich family of approximations, which has been studied previously. Yedidia et al. (2005) discuss counting numbers for the broader concept of *regions* which may contain any number of variables (in particular more than two). This naturally relates to *generalized belief propagation* (GBP) and associated *Kikuchi free energy approximations*. Pakzad and Anantharam (2005) and Heskes (2006) derived sufficient conditions for such free energy approximations to be convex. In this paper, we consider only pairwise counting numbers. In this context, Meshi et al. (2009) explored a wide range of pairwise counting numbers to try to find a convex free energy approximation with performance competitive to Bethe. For a subrange of models, they observed that this was possible yet still overall, Bethe performed very well. This is one of the motivations for this work, to understand better why Bethe performs so well.

Following Yedidia et al. (2005) and Meshi et al. (2009), we say that an approximation is *variable valid* if $c_i = 1 \; \forall i \in \mathcal{V}$, and is *edge valid* if $\rho_{ij} = 1 \; \forall (i,j) \in \mathcal{E}$. Their earlier work showed that variable valid approximations typically perform well compared to others, and we shall focus more attention on these models, though many of our results apply more generally to arbitrary counting numbers. Note that if all variables are independent, then variable validity is required to return the true entropy. If variables are connected in a tree, then edge validity is necessary to be exact. Bethe is unique in always being both variable and edge valid.

On a related theme, Weller et al. (2014) teased apart the two aspects of the Bethe approximation, i.e. the polytope and entropy as described in §2.1. Their results indicate that even if the optimization of (4) is performed over the marginal polytope, still the Bethe entropy approximation typically performs better than TRW. We consider polytope effects in §6.2.

2.3 SUBMODULARITY

A (set) function $f : 2^X \to \mathbb{R}$ is *submodular* if $\forall S, T \subseteq X, f(S \cap T) + f(S \cup T) \leq f(S) + f(T)$. For finite X, this is equivalent to diminishing returns, i.e. $\forall S \subseteq T, x \in X \setminus T, \; f(T \cup \{x\}) - f(T) \leq f(S \cup \{x\}) - f(S)$.

Submodular functions have been studied extensively (Edmonds, 1970; Lovász, 1983; Bach, 2013). In some ways, they are a discrete analogue of convex functions and can be minimized efficiently. The concept can be generalized to consider any *lattice*, i.e. a partially ordered set $(L, \preceq)$ such that $\forall x, y \in L, \; \exists$ a greatest lowest bound (glb or *meet*) $x \wedge y \in L$ and a least upper bound (lub or *join*) $x \vee y \in L$. A (lattice) function $f : L \to \mathbb{R}$ is *submodular* if $\forall x, y \in L, f(x \wedge y) + f(x \vee y) \leq f(x) + f(y)$.

For a pairwise function f over binary variables, f is submodular iff $f(0,0)+f(1,1) \leq f(0,1)+f(1,0)$. It is easily shown that the energy (or cost) of an edge (i,j) is submodular iff it is attractive, i.e. iff $W_{ij} \geq 0$. Further, the set of vectors in $\mathbb{R}^n$ with $x \preceq y$ if $x_i \leq y_i$ for all components i, is a lattice. Here $x \wedge y$ has ith component of $\min(x_i, y_i)$ and $x \vee y$ has ith component of $\max(x_i, y_i)$.

2.4 FLIPPING VARIABLES

The method of *flipping* (sometimes called *switching*) binary variables will be useful for our analysis in §3.3. Given a model on variables $\{X_i\}$, consider a new model on $\{X'_i\}$ where we flip a subset $\mathcal{R}$ of the variables, i.e. $X'_i = 1 - X_i$ for variables $i \in \mathcal{R} \subseteq \mathcal{V}$, and $X'_i = X_i$ for $i \in \mathcal{V} \setminus \mathcal{R}$. We identify new model parameters $\{\theta'_i, W'_{ij}\}$ as in (Weller and Jebara, 2013, §3) in order to preserve energies of all states up to a constant, hence the probability distribution over states is unchanged. If all variables are flipped (i.e. $\mathcal{R} = \mathcal{V}$), new parameters are given by

$$W'_{ij} = W_{ij}, \; \theta'_i = -\theta_i - \sum_{j \in \mathcal{N}(i)} W_{ij}. \tag{5}$$

If the original model was attractive, so too is the new model. In general, if a subset $\mathcal{R} \subseteq \mathcal{V}$ is flipped, let $\mathcal{E}_t = \{\text{edges with exactly } t \text{ ends in } \mathcal{R}\}$ for $t = 0, 1, 2$, then we obtain

$$W'_{ij} = \begin{cases} W_{ij} & (i,j) \in \mathcal{E}_0 \cup \mathcal{E}_2, \\ -W_{ij} & (i,j) \in \mathcal{E}_1, \end{cases}$$
$$\theta'_i = \begin{cases} \theta_i + \sum_{(i,j) \in \mathcal{E}_1} W_{ij} & i \in \mathcal{V} \setminus \mathcal{R}, \\ -\theta_i - \sum_{(i,j) \in \mathcal{E}_2} W_{ij} & i \in \mathcal{R}. \end{cases} \tag{6}$$

The proof of the following result for general counting numbers follows the argument used by Weller and Jebara (2013) for the specific case of the Bethe approximation.

Lemma 1. *Flipping variables changes affected pseudo-marginal matrix entries' locations but not values. For*

any counting numbers, $\mathcal{F}_A$ is unchanged up to a constant, hence the locations of stationary points are unaffected.

2.5 ATTRACTIVE AND BALANCED MODELS

A model is *attractive* iff all its edges are attractive, i.e. iff $W_{ij} \geq 0 \ \forall (i,j) \in \mathcal{E}$. As suggested by §2.3, attractive models have desirable properties, e.g. a MAP assignment may be found in polynomial time (Greig et al., 1989), and as shown in §5, we can construct a FPTAS for Z_A for any counting numbers. We remark that, as observed by Harary (1953), a general model (which may contain repulsive edges) can be mapped to an attractive model by flipping a subset of variables iff the initial model is *balanced*, that is iff it contains no *frustrated* cycles, i.e. a cycle with an odd number of repulsive edges. Hence, results that apply to attractive models may readily be extended to the wider class of balanced models.

3 FIRST DERIVATIVES OF $\mathcal{F}_A$

Combining (4) with (1), (2) and (3), yields

$$\mathcal{F}_A(q) = -\sum_{i \in \mathcal{V}} \theta_i q_i - \sum_{(i,j) \in \mathcal{E}} W_{ij} \xi_{ij} - \sum_{i \in \mathcal{V}} c_i S_i + \sum_{(i,j) \in \mathcal{E}} \rho_{ij} (S_i + S_j - S_{ij}). \quad (7)$$

3.1 OPTIMUM PAIRWISE PSEUDOMARGINALS

Differentiating (7) with respect to ξ_{ij}, we obtain

$$\frac{\partial \mathcal{F}_A}{\partial \xi_{ij}} = -W_{ij} - \rho_{ij} \frac{\partial S_{ij}}{\partial \xi_{ij}} = -W_{ij} + \rho_{ij} \log \left[\frac{\xi_{ij}(1 + \xi_{ij} - q_i - q_j)}{(q_i - \xi_{ij})(q_j - \xi_{ij})} \right].$$

Note that this is independent of the singleton counting numbers $\{c_i\}$. Welling and Teh (2001) considered the specific case of the Bethe approximation, where $\rho_{ij} = 1$. Solving the general case for $\frac{\partial \mathcal{F}_A}{\partial \xi_{ij}} = 0$ leads to a quadratic equation,

$$\alpha_{ij} \xi_{ij}^2 - [1 + \alpha_{ij}(q_i + q_j)] \xi_{ij} + (1 + \alpha_{ij}) q_i q_j = 0, \quad (8)$$

where we define $\alpha_{ij} = e^{W_{ij}/\rho_{ij}} - 1$. Observe that here W_{ij}/ρ_{ij} plays the 'edge count modified' role typically performed by W_{ij} in the standard Bethe approximation. It is easily shown that (8) has just one feasible solution (Welling and Teh, 2001; Weller and Jebara, 2013), as given in the following result.

Theorem 2. *For general counting numbers, given singleton pseudomarginals, optimum pairwise terms (which minimize the approximate free energy) are given by*

$$\xi_{ij}^*(q_i, q_j) = \frac{1}{2\alpha_{ij}} \left(x_{ij} - \sqrt{x_{ij}^2 - 4\alpha_{ij}(1 + \alpha_{ij}) q_i q_j} \right),$$

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

Henceforth we shall often consider $\mathcal{F}_A$ as a function of just the singleton pseudomarginals $\{q_i\}$, with all pairwise ξ_{ij} terms being implicitly specified by their optimum values as given by Theorem 2.

As noted by Weller and Jebara (2013), (8) may be rewritten as $\xi_{ij} - q_i q_j = \alpha_{ij}(q_i - \xi_{ij})(q_j - \xi_{ij})$. The terms in parentheses are elements of the pairwise marginal (2), constrained to be ≥ 0. By its definition, α_{ij} takes the same sign as W_{ij}/ρ_{ij}, hence the following result holds.

Lemma 3. $\frac{W_{ij}}{\rho_{ij}} \geq 0 \Rightarrow \xi_{ij} \geq q_i q_j, \frac{W_{ij}}{\rho_{ij}} \leq 0 \Rightarrow \xi_{ij} \leq q_i q_j$.

We remark that, given singleton marginals $\{q_i\}$, a lower edge counting number $|\rho_{ij}|$ implies a more extreme pairwise marginal term in the sense of greater $|\xi_{ij} - q_i q_j|$. This is true, for example, of TRW compared to Bethe.

3.2 FIRST DERIVATIVES WRT q_i, ASSUMING OPTIMUM PAIRWISE PSEUDOMARGINALS

We follow the approach of Welling and Teh (2001), noting that at the optimum pairwise pseudomarginals, $\frac{\partial \mathcal{F}_A}{\partial \xi_{ij}} = 0$ for all edges, hence, holding q_j fixed $\forall j \neq i$,

$$\begin{aligned} \left. \frac{d\mathcal{F}_A}{dq_i} \right|_{\{q_j\}} &= \left. \frac{\partial \mathcal{F}_A}{\partial q_i} \right|_{\{q_j, \xi_{ij}\}} + \sum_{j \in \mathcal{N}(i)} \frac{\partial \mathcal{F}_A}{\partial \xi_{ij}} \frac{\partial \xi_{ij}}{\partial q_i} \\ &= -\theta_i - c_i \frac{\partial S_i}{\partial q_i} + \sum_{j \in \mathcal{N}(i)} \rho_{ij} \frac{\partial}{\partial q_i}(S_i - S_{ij}) \\ &= -\theta_i + c_i \log \frac{q_i}{1 - q_i} \\ &\quad + \sum_{j \in \mathcal{N}(i)} \rho_{ij} \left(-\log \frac{q_i}{1 - q_i} + \log \frac{q_i - \xi_{ij}}{1 + \xi_{ij} - q_i - q_j} \right) \\ &= -\theta_i + c_i \log \frac{q_i}{1 - q_i} + \sum_{j \in \mathcal{N}(i)} \rho_{ij} \log Q_{ij}, \end{aligned} \quad (9)$$

where as in (Weller and Jebara, 2014b), we define[3]

$$Q_{ij} = \left(\frac{q_i - \xi_{ij}}{1 + \xi_{ij} - q_i - q_j} \right) \left(\frac{1 - q_i}{q_i} \right). \quad (10)$$

Considering (10) and Lemma 3 yields the following.

Lemma 4. *If edge (i,j) is attractive, i.e. $W_{ij} \geq 0$, then $\rho_{ij} \log Q_{ij} \leq 0$.*

Gradient descent methods may be used to try to minimize $\mathcal{F}_A$ but note these may find only a local optimum.

[3] Note $Q_{ij} = \frac{\partial}{\partial q_i}(S_i - S_{ij}) = \frac{p(X_j=0|X_i=1)}{p(X_j=0|X_i=0)}$ by (2).

3.3 BOUNDS ON FIRST DERIVATIVES WRT q_i

We generalize the approach of Weller and Jebara (2014a) to bound the range of first derivatives (9) for free energy approximations with arbitrary counting numbers. An important application is the construction of an ϵ-*sufficient mesh* to estimate $\log Z_A$, see §5.

Initially assume a model that is locally attractive around X_i, i.e. $W_{ij} \geq 0 \; \forall j \in \mathcal{N}(i)$. From (9) and Lemma 4, we obtain $\frac{\partial \mathcal{F}_A}{\partial q_i} \leq -\theta_i + c_i \log \frac{q_i}{1-q_i}$.

Now flip all variables, see §2.4, to consider a model with $\{X'_i = 1 - X_i \; \forall i \in \mathcal{V}\}$, keeping the same counting numbers. We obtain $W'_{ij} = W_{ij}$ and can apply the result above to yield

$$\begin{aligned} \frac{\partial \mathcal{F}_A}{\partial q'_i} &\leq -\theta'_i + c_i \log \frac{q'_i}{1-q'_i} \\ \Leftrightarrow -\frac{\partial \mathcal{F}_A}{\partial q_i} &\leq \theta_i + W_i^+ - c_i \log \frac{q_i}{1-q_i} \quad \text{(see §2.4)}, \end{aligned}$$

where we define $W_i^+ = \sum_{j \in \mathcal{N}(i): W_{ij} \geq 0} W_{ij}$. Combine this with the earlier result to yield a sandwich inequality,

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

Now generalize to consider the case that X_i has some neighbors $X_j \in \mathcal{R}$ to which it is adjacent by repulsive edges, i.e. where $W_{ij} < 0$. First flip just the variables in $\mathcal{R}$, see §2.4, and then apply the above sandwich result to yield the following Theorem, where we define the nonnegative value $W_i^- = \sum_{j \in \mathcal{N}(i): W_{ij} \leq 0} -W_{ij}$.

Theorem 5. *For arbitrary counting numbers, assuming optimum pairwise pseudomarginals, first derivatives of $\mathcal{F}_A$ are sandwiched in the range*

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

Note that both upper and lower bounds are monotonic in q_i (increasing with q_i if $c_i > 0$, else nonincreasing), ranging from $-\infty$ to ∞, separated by the constant value $W_i^- + W_i^+ = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$. See Figure 1 for an example.

4 SECOND DERIVATIVES OF $\mathcal{F}_A$

We extend the analysis of Weller and Jebara (2013) to derive all terms of the Hessian H for free energy approximations $\mathcal{F}_A$ with arbitrary counting numbers.

Theorem 6 ($H_{ij} = \frac{\partial^2 \mathcal{F}_A}{\partial q_i \partial q_j}$ second derivatives of $\mathcal{F}_A(q_1, \ldots, q_n)$ at optimum pairwise marginals ξ_{ij})**.**

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

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

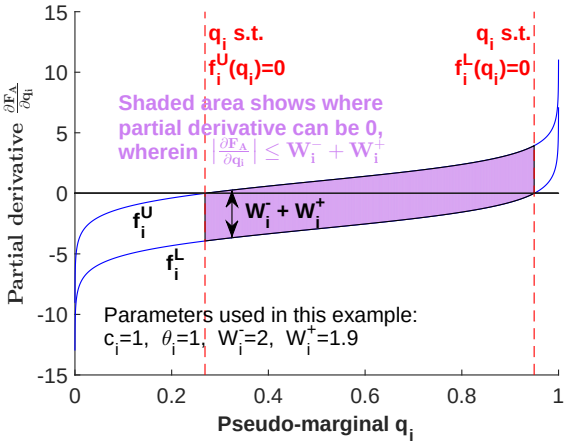


Figure 1: An example of upper and lower bounds for $\frac{\partial \mathcal{F}_A}{\partial q_i}$. Blue curves show monotonic upper $f_i^U(q_i)$ and lower $f_i^L(q_i)$ bound curves from Theorem 5, separated by constant $W_i^- + W_i^+$. In preprocessing, the search space is shrunk to within the dashed red lines, within which $|\frac{\partial \mathcal{F}_A}{\partial q_i}| \leq W_i^- + W_i^+ = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$.

where ξ_{ij} takes its optimum value from Theorem 2, and $T_{ij} = q_i q_j (1-q_i)(1-q_j) - (\xi_{ij} - q_i q_j)^2 \geq 0$, with equality iff q_i or $q_j \in \{0, 1\}$. Proof in Appendix.

These second derivatives may be combined with the earlier gradients (9) for more efficient local minimization of $\mathcal{F}_A$.

4.1 SUBMODULARITY OF $\mathcal{F}_A$

Considering the expression for H_{ij} from Theorem 6 together with Lemma 3, observe that provided $\rho_{ij} \neq 0$ and $q_i, q_j \notin \{0, 1\}$, $W_{ij} \geq 0 \Leftrightarrow \frac{\partial^2 \mathcal{F}_A}{\partial q_i \partial q_j} \leq 0$ (whatever the sign of ρ_{ij}). Since third derivatives exist and are finite in this range, this yields the following result.

Theorem 7. *For any counting numbers with $\rho_{ij} \neq 0 \; \forall (i,j) \in \mathcal{E}$, and any discretization, an attractive model yields a submodular discrete optimization problem to estimate* $\log Z_A$*. Proof in Appendix.*

This means that considering $\mathcal{F}_A(q_1, \ldots, q_n)$ with pairwise marginals given by Theorem 2, for any discrete mesh $\mathcal{M} = \prod_{i=1}^n M_i$, where M_i is a finite set of points for q_i in $[0, 1]$, and for any counting numbers, then the discrete optimization to find the point in $\mathcal{M}$ with lowest $\mathcal{F}_A$ is submodular for any attractive model (hence can be solved efficiently).

5 OPTIMIZING THE APPROXIMATE FREE ENERGY $\mathcal{F}_A$

True marginal inference is NP-hard (Cooper, 1990), even to approximate (Dagum and Luby, 1993). However, Weller and Jebara (2014a) derived an algorithm to approximate the Bethe log-partition function, $\log Z_B$, to within any

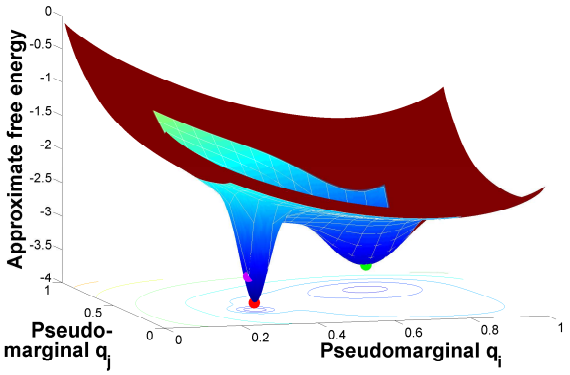


Figure 2: Stylized example for optimizing the approximate free energy over two variables. The search space is first shrunk to exclude the outer red region, then the inner blue region is discretized using an *ϵ-sufficient mesh*. The red dot indicates the (continuous) global minimum. On the mesh: the purple dot has the closest location, guaranteed to have value within ϵ, while the green dot is the lowest point, hence is the discretized optimum returned.

ϵ by constructing an *ϵ-sufficient mesh* $\mathcal{M}(\epsilon)$, i.e. a discrete mesh over the space of singleton marginals $[0,1]^n$ such that the mesh point q^* with $\min_{q \in \mathcal{M}(\epsilon)} \mathcal{F}_B(q)$ is guaranteed to have $\mathcal{F}_B(q^*)$ within ϵ of the global optimum of $-\log Z_B$. In the case of an attractive model, the discrete optimization problem was shown to be submodular, leading to a FPTAS for $\log Z_B$. Using Theorems 5 and 7, we extend their approach to obtain similar results for any counting numbers.

The overall mesh method is outlined in Algorithm 1 and illustrated in Figure 2. Note that we need search only over the space of singleton marginals $[0,1]^n$, since pairwise terms may be computed with Theorem 2. First the search space is shrunk using the bounds of Theorem 5, since we need check only where $\frac{\partial \mathcal{F}_A}{\partial q_i}$ can be 0. Within this range, $\left|\frac{\partial \mathcal{F}_A}{\partial q_i}\right| \leq W_i^- + W_i^+ = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$, see Figure 1. Next, discrete mesh points for each variable's singleton marginal q_i may be selected in its range such that the step size δ_i satisfies $\delta_i \max \left|\frac{\partial \mathcal{F}_A}{\partial q_i}\right| \approx \frac{\epsilon}{n}$. This ensures that, wherever the global minimum is within the space, $\mathcal{F}_A$ cannot rise by more than $n\frac{\epsilon}{n} = \epsilon$ at the closest mesh point. This leads to a number of mesh points in dimension i of $N_i = O(\frac{1}{\delta_i}) = O(\frac{n}{\epsilon} \sum_{j \in \mathcal{N}(i)} |W_{ij}|)$. If an upper bound W on edge strengths is known such that $|W_{ij}| \leq W \ \forall (i,j) \in \mathcal{E}$, then the sum of mesh points in each dimension, $N = \sum_{i \in \mathcal{V}} N_i = O(\frac{nmW}{\epsilon})$, where $m = |\mathcal{E}|$.

If the model is attractive, we obtain a FPTAS since by Theorem 7, the resulting submodular multilabel optimization problem may be solved in time $O(N^3) = O\left(\left(\frac{nmW}{\epsilon}\right)^3\right)$ using earlier graph cut results (Schlesinger and Flach, 2006; Greig et al., 1989; Goldberg and Tarjan, 1988). If the model is balanced, then a subset of variables may be efficiently identified such that flipping them yields an attractive model (see §2.4), hence the FPTAS extends to balanced models. If the model is not balanced, there is an extensive range of methods available, see (Koller and Friedman, 2009, §13) or (Kappes et al., 2013) for recent surveys.

Algorithm 1 Mesh method to return ϵ-approximate global optimum $\log Z_A$ for any counting numbers.

Input: ϵ, model parameters $\{\theta_i, W_{ij}\}$ and counting numbers $\{c_i, \rho_{ij}\}$
Output: Estimate of global optimum $\log Z_A$ guaranteed in $[\log Z_A - \epsilon, \log Z_A]$, with corresponding pseudomarginals as arg for the discrete optimum

1: For each X_i: Compute upper and lower bound curves for $\frac{\partial \mathcal{F}_A}{\partial q_i}$ from Theorem 5, use these to shrink the search space to a range wherein $\left|\frac{\partial \mathcal{F}_A}{\partial q_i}\right| \leq W_i^- + W_i^+ = \sum_{j \in \mathcal{N}(i)} |W_{ij}|$, see Figure 1.
2: Construct an ϵ-sufficient mesh as described in §5.
3: Solve the resulting discrete optimization problem (efficient by Theorem 7 if the model is attractive), see §5.

Various refinements to improve efficiency are discussed by Weller and Jebara (2014a) for the Bethe case. All those techniques may also be applied here, and can help significantly in practice, though they do not improve the theoretical worst case.

Other approaches to attempt to minimize the Bethe free energy have been developed (Welling and Teh, 2001; Yuille, 2002; Heskes et al., 2003; Shin, 2012), and some generalize to other counting numbers, including the message passing methods of Hazan and Shashua (2008) (guaranteed to converge for a convex free energy), Wiegerinck and Heskes (2003) or Meshi et al. (2009), but unless $\mathcal{F}_A$ is convex, none guarantees a solution close to the global optimum.

6 UNDERSTANDING APPROXIMATION ERROR

We examine how the entropy approximation S_A may lead to error in the marginals, then consider other factors affecting error in the estimate of the partition function.

6.1 EFFECT OF APPROXIMATE ENTROPY ON MARGINALS

It has previously been observed that in cyclic graphs, there are situations where the Bethe entropy tends to pull approximate singleton marginals toward extreme values near 0 or 1, and that this tends to occur as a 'phase transition' in behavior when edge weights rise above some threshold (Heskes, 2004; Mooij and Kappen, 2005).[4] One perspec-

[4]Note that we describe a transition in the accuracy of approximate singleton marginals. A quite different symmetry-breaking effect is the 'ferromagnetic-paramagnetic' transition that relates

tive on this is algorithmic stability (Wainwright and Jordan, 2008, §7.4). A different heuristic interpretation is that it occurs as a result of LBP overcounting information when going around cycles (Ihler, 2007). Here we extend the explanatory approach of Weller et al. (2014) by considering the entropy approximation and examining the effect of different counting numbers.

To illustrate the principles, we analyze a simple model with n vertices connected such that each vertex has exactly d neighbors (such models are called *d-regular*), with all edge potentials symmetric of weight W and no singleton potentials (we call these models *symmetric* and *homogeneous*). Using (9), it is easily shown that, for any counting numbers, there is a stationary point of $\mathcal{F}_A$ at a location with $q_i = \frac{1}{2}\ \forall i \in \mathcal{V}$, which by symmetry clearly also give the true singleton marginals. However, for certain counting numbers, including the Bethe parameters, when W is above a critical threshold, this stationary point is no longer a minimum, and new minima emerge that pull singleton marginals away to extreme values. The following result considers an approximation with uniform counting numbers (i.e. all $c_i = c, \rho_{ij} = \rho$), and demonstrates conditions for when $q_i = \frac{1}{2}\ \forall i \in \mathcal{V}$ is not a minimum, by explicitly providing a direction showing that the Hessian H is not positive semidefinite.

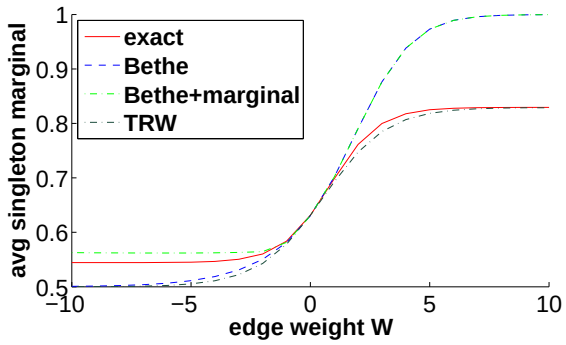
Lemma 8. *For a symmetric homogeneous d-regular model on n vertices, let H be the Hessian of the approximate free energy at $q_i = \frac{1}{2}\ \forall i \in \mathcal{V}$, using uniform counting numbers $c_i = c\ \forall i \in \mathcal{V}, \rho_{ij} = \rho\ \forall (i,j) \in \mathcal{E}$, then $\mathbf{1}^T H \mathbf{1} = n\left[4(c - d\rho) + \frac{d}{\rho\xi}\right]$, where $\xi = \frac{1}{2}\sigma\left(\frac{W}{2\rho}\right)$ is the uniform optimum edge marginal term, and $\sigma(u) = \frac{1}{1+e^{-u}}$ is the standard sigmoid function. Proof in Appendix.*

Hence, $q_i = \frac{1}{2}\ \forall i$ is not a minimum if $\omega = 4(c - d\rho) + \frac{d}{\rho\xi} < 0$. First, note that for the Bethe approximation $c = \rho = 1$, and this condition reduces to $\xi > \frac{1}{4}\frac{d}{d-1} \Leftrightarrow W > 2\log\frac{d}{d-2}$. Indeed, when W rises above this critical threshold, singleton marginals will move away from $\frac{1}{2}$ (Weller et al., 2014).

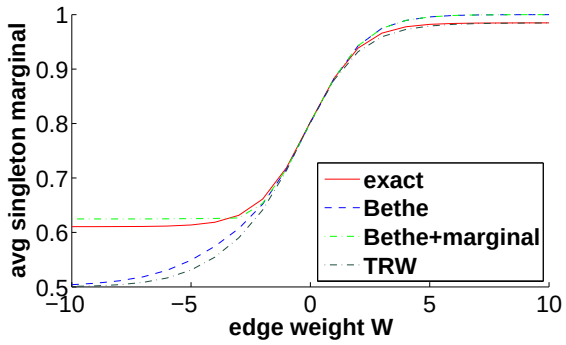
In general, higher singleton counting numbers c and lower edge counting numbers ρ raise ω, making it harder to satisfy the condition. The effect of the density of connectivity d is less clear, and depends on the other parameters. For example, consider the TRW approximation with $c = 1$ and uniform edge weights $\rho = \frac{2(n-1)}{nd} < 1$, declining with d, which are optimum in this setting (Weller et al., 2014, Lemma 7), then ω is positive and increases rapidly with d (whereas Bethe suffers in this regard by keeping $\rho = 1$ fixed).

To understand this behavior, recall the definition of S_A in (3). As singleton counting numbers c_i rise, we add more S_i which are concave, thereby increasing convexity of $\mathcal{F}_A$ around $\frac{1}{2}$ and making it more likely to be a minimum. On the other hand, increasing edge terms ρ_{ij} leads to more mutual information I_{ij} being subtracted, thereby increasing concavity of $\mathcal{F}_A$ around $\frac{1}{2}$ and potentially pushing marginals away from $\frac{1}{2}$. This perspective helps to understand why a convex free energy approximation leads to algorithmic stability (Wainwright and Jordan, 2008, §7.4).

to the true global distribution of states (mostly aligned or not).



(a) $T_{max} = 1$



(b) $T_{max} = 3$

Figure 3: Average over 20 runs of singleton marginal vs. uniform symmetric edge weight W for: exact inference, Bethe approximation, Bethe+marginal polytope, and TRW (all $\rho_{ij} = 2/3$). Triangle topology with random singleton potentials $\theta_i \sim [0, T_{max}]$. For $W > 0$: Bethe and Bethe+marginal overlap, exact and TRW almost overlap. For $W < 0$ (frustrated cycle): Bethe and TRW almost overlap, as do exact and Bethe+marginal.

The severity of this problem for estimating singleton marginals is high when true marginals are near $\frac{1}{2}$, which typically occurs for small singleton potentials, but it is less problematic when true marginals are themselves near 0 or 1. The effect is illustrated in Figure 3. Note how, for positive W, the Bethe marginals are pulled toward 1 whereas TRW is almost exactly correct. The effect for $W < 0$ is dominated instead by a polytope effect, which we discuss in the next Section.

We remark that although the entropy approximation may have a dramatic effect on the accuracy of singleton marginals, particularly for low singleton potentials (where true marginals are near $\frac{1}{2}$), the effect on estimating pairwise marginals and the partition function is less clear. In-

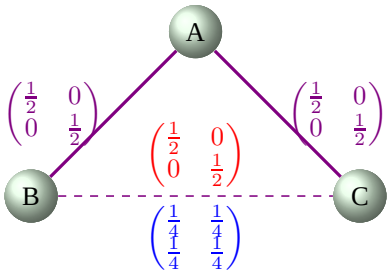


Figure 4: Illustration of the polytope effect on edge marginals. A-B and A-C are strongly coupled, B-C is very weakly coupled with all edges symmetric and attractive, and no singleton potentials. Edge marginals are shown. For B-C, above the edge (red) is the optimum in the marginal polytope (global consistency), below the edge (blue) is the optimum for the local polytope. See §6.2.

deed, Bethe typically outperforms TRW on these measures (Weller et al., 2014).

6.2 EFFECT OF LOCAL POLYTOPE

We revisit and expand on an example from Weller et al. (2014) to show that the impact of each of the two aspects (i.e. polytope and entropy, see §2.1) of an approximation to the partition function can pull in opposite directions. Hence, improving just the entropy approximation could lead to a *worse* approximation.

Consider the model in Figure 4, where 3 variables are connected in a triangle. Two edges are strongly attractive, and the third is very weakly attractive. The strong edge $A-B$ ensures that A and B take the same value, similarly for $B-C$. Hence, in the globally consistent marginal polytope, B and C must take the same value. The global states 000 and 111 each have probability of almost $\frac{1}{2}$, and the pairwise marginals are shown along the edges of Figure 4. Since the model is almost a tree, we know that $Z_B \approx Z$. We shall examine how this arises by starting with exact inference, then switch to use the Bethe entropy approximation on the marginal polytope, and then relax the constraint set to the local polytope. We shall ignore the energy terms since they are equal here for true or approximate inference.

As noted, there are 2 states that dominate the global probability distribution, hence true $S \approx \log 2$. Computing the Bethe entropy on the marginal polytope, we obtain $S_B \approx 3\log 2 - 3\log 2 = 0$, which is too low by $\log 2$. However, when the polytope is relaxed, a better optimum is found by maximizing the edge entropy of $B-C$ as shown under the edge in Figure 4. Since only local consistency is required, there is no longer any need for B to be equal to C and we gain the difference in edge entropy of $2\log 2 - \log 2 = \log 2$, thus exactly offsetting the deficit due to Bethe entropy on the marginal polytope.

This example demonstrates that focusing exclusively on the entropy approximation, without also considering the polytope approximation, may lead to difficulties. We highlight another aspect of the polytope approximation, in that it introduces half-integral vertices (Wainwright and Jordan, 2008). In a balanced cycle (even number of repulsive edges), this is of little consequence since the optimum energy (MAP solution) is always at an integral vertex, but in a frustrated cycle (odd number of repulsive edges, see §2.5), the energy can cause singleton marginals to be pulled towards $\frac{1}{2}$.[5] Hence, although the Bethe entropy pulls these marginals away from $\frac{1}{2}$ on balanced cycles, the polytope effect pushes the other way on frustrated cycles, which in some cases may provide a helpful 'balance'. Since many optimization techniques (including message passing methods) exploit the efficiencies possible with the local polytope approximation, it may in fact be desirable overall to have an entropy approximation such as Bethe, for this offsetting effect. See Figure 3 in the region $W < 0$ for an illustration, where the Bethe+marginal optimization was performed using the Frank-Wolfe algorithm (Frank and Wolfe, 1956).

6.3 BOUNDS ON Z_A

While the TRW approximation has $Z_T \geq Z$ by construction, until recently there were no guarantees on the performance of the Bethe approximation, though it typically yields very good results. Sudderth et al. (2007) proved that $Z_B \leq Z$ for a range of attractive binary pairwise models, and conjectured that this bound holds for all attractive models. This was proved true by Ruozzi (2012) using the method of graph covers, and then also by Weller and Jebara (2014b) by combining the idea of clamping variables with analyzing properties of the derivatives of $\mathcal{F}_B$.

In this Section, we use the loop series method (Sudderth et al., 2007; Chertkov and Chernyak, 2006) to show that for certain other models, we can prove that $Z_B \geq Z$. For such models, this immediately implies that the Bethe approximation is better for estimating Z than any approximation with $c_i = 1 \ \forall i \in \mathcal{V}$ (variable valid) and $\rho_{ij} \leq 1 \ \forall (i,j) \in \mathcal{E}$ (from the definition of S_A, see §2.1-2.2). In particular, for these models, $Z \leq Z_B \leq Z_T$.

Sudderth et al. (2007) showed that $Z/Z_B = 1 +$ a series of terms, one term for each *generalized loop*, which is a subgraph such that no vertex has degree 1, and demonstrated that each of the terms in the series is ≥ 0 for certain models, and hence $Z_B \leq Z$ for these cases. See Appendix for background on this approach. In particular, if there is exactly one cycle in the model, then there is only one term in the series and if the cycle is attractive, then this term is positive. We note that this immediately generalizes to a cycle that is balanced (see §2.5 for definitions).

Here we apply similar analysis (Sudderth et al., 2007, §3-4, or see Appendix), and observe that if there is exactly one

[5] This can lead the Bethe optimum of a strongly frustrated cycle to occur at a location where $S_B < 0$.

cycle and it is frustrated, then the term is negative, thus proving that for such models, $Z_B \geq Z$.

Interestingly, Weller and Jebara (2014b) have shown that for the case of a model with one balanced cycle, $\frac{1}{2}Z \leq Z_B \leq Z$, so although Z_B is lower than Z, it cannot be by much even for very strong edge weights; whereas for a single frustrated cycle, there is no limit to how large Z_B/Z can rise. This suggests that for a general model, the accuracy of Z_B will depend on the blend of balanced and frustrated cycles, where in a sense frustrated cycles cause greater trouble than balanced cycles, though to understand how the effects combine in a model with multiple cycles will require further analysis. Since Z_B performs well even for attractive models (Sudderth et al., 2007), this indicates that, for estimating the partition function, practitioners should use approximations with $\rho_{ij} < 1$ (such as TRW) with caution.

The loop series method extends to models with more than one cycle but the analysis becomes more complicated. Again using the approach of Sudderth et al. (2007), we can conclude more generally that $Z_B \geq Z$ for any model such that every generalized loop contains an odd number of repulsive edges (this is a sort of generalized frustrated cycle), and the Bethe optimum marginals for every variable that has an odd degree ≥ 3 in any generalized loop, are either all $\leq \frac{1}{2}$ or all $\geq \frac{1}{2}$ (see Appendix).

6.4 DERIVATIVES WRT COUNTING NUMBERS

We are interested in exploring which counting numbers lead to accurate inference as measured by errors in the estimates of the partition function and marginals. Considering (7) and using the envelope theorem (Milgrom, 1999, Theorem 1), we have right derivatives:

$$\frac{\partial \log Z_A}{\partial c_i}^+ = \max_{q \in X} S_i(q_i),$$
$$\frac{\partial \log Z_A}{\partial \rho_{ij}}^+ = \max_{q \in X} \left[S_{ij}(\mu_{ij}) - S_i(q_i) - S_j(q_j) \right], \quad (11)$$

where X is the set of all $\arg\min \mathcal{F}_A$.[6] The left derivatives correspondingly take the $\min$ rather than the $\max$ of the same expressions. If the minimum of $\mathcal{F}_A$ is unique, as is the case for any convex $\mathcal{F}_A$, then the right and left derivatives are equal.

For tractable models, where the exact partition function Z may be computed, this will allow exploration over the range of counting numbers that yield accurate partition functions. It will be interesting to investigate robustness of the quality of the partition function estimate to changes in model potentials, and accuracy of marginals, though this is outside the scope of the current work.

Others have investigated ways to optimize counting numbers. Wiegerinck and Heskes (2003) proposed a method using linear response theory. They also discussed alpha-divergence measures, an idea developed further by Minka (2005), who fascinatingly frames (fractional) BP and (power) EP under a general framework of iterative minimization of alpha-divergence, yielding insight into which measures may be expected to perform well for different objectives, though concluding that this is difficult to predict.

7 CONCLUSION

We have shown how recent results for the Bethe approximation may be extended to handle the broad range of pairwise approximations using any counting numbers. Our analysis builds on earlier work (Welling and Teh, 2001; Yedidia et al., 2005; Meshi et al., 2009; Sudderth et al., 2007; Weller and Jebara, 2013, 2014a), providing new insights and deepening our understanding of how best to perform inference in practice. This is important given the popularity of LBP and TRW approximations. Further, it provides a valuable toolbox for further exploration.

Areas for future investigation include trying to understand better how to predict which approach will work well for a given model, and analyzing the performance of message passing algorithms with different counting numbers (where our ϵ-accurate approach provides a valuable benchmark).

Acknowledgements

The author thanks Ofer Meshi for fruitful discussions and for sharing code, and the anonymous reviewers for helpful comments and suggestions.

References

F. Bach. Learning with submodular functions: A convex optimization perspective. *Foundations and Trends in Machine Learning*, 6(2-3):145–373, 2013.

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

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

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

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

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

F. Eaton and Z. Ghahramani. Model reductions for inference: Generality of pairwise, binary, and planar factor graphs. *Neural Computation*, 25(5):1213–1260, 2013.

[6] This generalizes an earlier result for convex free energies (Meshi et al., 2009, Prop 5.2), which itself generalized a result of Wainwright et al. (2005). The envelope theorem is similar to Danskin's theorem (Bertsekas, 1995). Recall $\log Z_A = -\min \mathcal{F}_A$. Intuitively, for multiple $\arg\min$ locations, each may vary at a different rate, thus for the right derivative, we must take the $\max$ of the derivative over all the locations.

J. Edmonds. Submodular functions, matroids, and certain polyhedra. *Edited by G. Goos, J. Hartmanis, and J. van Leeuwen*, page 11, 1970.

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

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.

F. Harary. On the notion of balance of a signed graph. *Michigan Mathematical Journal*, 2:143–146, 1953.

T. Hazan and A. Shashua. Convergent message-passing algorithms for inference over general graphs with convex free energies. In *UAI*, 2008.

T. Heskes. Stable fixed points of loopy belief propagation are minima of the Bethe free energy. In *Neural Information Processing Systems*, 2002.

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

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

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

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

M. 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.

D. Koller and N. Friedman. *Probabilistic Graphical Models - Principles and Techniques*. MIT Press, 2009.

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

L. Lovász. Submodular functions and convexity. In A. Bachem, M. Grötschel, and B. Korte, editors, *Mathematical Programming – The State of the Art*, pages 235–257, Berlin, 1983. Springer-Verlag.

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.

P. Milgrom. The envelope theorems. *Department of Economics, Standford University, Mimeo*, 1999. URL `http://www-siepr.stanford.edu/workp/swp99016.pdf`.

T. Minka. Divergence measures and message passing. *Technical Report MSR-TR-2005-173*, 2005.

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

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

P. Pakzad and V. Anantharam. Estimation and marginalization using Kikuchi approximation methods. *Neural Computation*, 17(8):1836–1873, 2005.

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

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

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

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

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

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

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

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

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

A. Weller and T. Jebara. Clamping variables and approximate inference. In *Neural Information Processing Systems (NIPS)*, 2014b.

A. Weller, K. Tang, D. Sontag, and T. Jebara. Understanding the Bethe approximation: When and how can it go wrong? In *Uncertainty in Artificial Intelligence (UAI)*, 2014.

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

W. Wiegerinck and T. Heskes. Fractional belief propagation. In S. Becker, S. Thrun, and K. Obermayer, editors, *Advances in Neural Information Processing Systems 15*, pages 438–445. MIT Press, 2003.

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

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

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