Clamping Variables and Approximate Inference

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

It was recently proved using graph covers (Ruozzi, 2012) that the Bethe partition function is upper bounded by the true partition function for a binary pairwise model that is attractive. Here we provide a new, arguably simpler proof from first principles. We make use of the idea of clamping a variable to a particular value. For an attractive model, we show that summing over the Bethe partition functions for each sub-model obtained after clamping any variable can only raise (and hence improve) the approximation. In fact, we derive a stronger result that may have other useful implications. Repeatedly clamping until we obtain a model with no cycles, where the Bethe approximation is exact, yields the result. We also provide a related lower bound on a broad class of approximate partition functions of general pairwise multi-label models that depends only on the topology. We demonstrate that clamping a few wisely chosen variables can be of practical value by dramatically reducing approximation error.

1 Introduction

Marginal inference and estimating the partition function for undirected graphical models, also called Markov random fields (MRFs), are fundamental problems in machine learning. Exact solutions may be obtained via variable elimination or the junction tree method, but unless the treewidth is bounded, this can take exponential time (Pearl, 1988; Lauritzen and Spiegelhalter, 1988; Wainwright and Jordan, 2008). Hence, many approximate methods have been developed.

Of particular note is the Bethe approximation, which is widely used via the *loopy belief propagation* algorithm (LBP). Though this is typically fast and results are often accurate, in general it may converge only to a local optimum of the Bethe free energy, or may not converge at all (McEliece et al., 1998; Murphy et al., 1999). Another drawback is that, until recently, there were no guarantees on whether the returned approximation to the partition function was higher or lower than the true value. Both aspects are in contrast to methods such as the *tree-reweighted* approximation (TRW, Wainwright et al., 2005), which features a convex free energy and is guaranteed to return an upper bound on the true partition function. Nevertheless, empirically, LBP or convergent implementations of the Bethe approximation often outperform other methods (Meshi et al., 2009; Weller et al., 2014).

Using the method of graph covers (Vontobel, 2013), Ruozzi (2012) recently proved that the optimum Bethe partition function provides a lower bound for the true value, i.e. $Z_B \leq Z$, for discrete binary MRFs with submodular log potential cost functions of any arity. Here we provide an alternative proof for attractive binary pairwise models. Our proof does not rely on any methods of loop series (Sudderth et al., 2007) or graph covers, but rather builds on fundamental properties of the derivatives of the Bethe free energy. Our approach applies only to binary models (whereas Ruozzi, 2012 applies to any arity), but we obtain stronger results for this class, from which $Z_B \leq Z$ easily follows. We use the idea of *clamping* a variable and considering the approximate sub-partition functions over the remaining variables, as the clamped variable takes each of its possible values.

Notation and preliminaries are presented in §2. In §3, we derive a lower bound, not just for the standard Bethe partition function, but for a range of approximate partition functions over multi-label variables that may be defined from a variational perspective as an optimization problem, based only on the topology of the model. In §4, we consider the Bethe approximation for attractive binary pairwise models. We show that clamping any variable and summing the Bethe sub-partition functions over the remaining variables can only increase (hence improve) the approximation. Together with a similar argument to that used in §3, this proves that $Z_B \leq Z$ for this class of model. To derive the result, we analyze how the optimum of the Bethe free energy varies as the singleton marginal of one particular variable is fixed to different values in $[0, 1]$. Remarkably, we show that the negative of this optimum, less the singleton entropy of the variable, is a convex function of the singleton marginal. This may have further interesting implications. We present experiments in §5, demonstrating that clamping even a single variable selected using a simple heuristic can be very beneficial.

1.1 Related work

Branching or conditioning on a variable (or set of variables) and approximating over the remaining variables has a fruitful history in algorithms such as branch-and-cut (Padberg and Rinaldi, 1991; Mitchell, 2002), work on resolution versus search (Rish and Dechter, 2000) and various approaches of (Darwiche, 2009, Chapter 8). Cutset conditioning was discussed by Pearl (1988) and refined by Peot and Shachter (1991) as a method to render the remaining topology acyclic in preparation for belief propagation. Eaton and Ghahramani (2009) developed this further, introducing the *conditioned belief propagation* algorithm together with *back-belief-propagation* as a way to help identify which variables to clamp. Liu et al. (2012) discussed feedback message passing for inference in Gaussian (not discrete) models, deriving strong results for the particular class of attractive models. Choi and Darwiche (2008) examined methods to approximate the partition function by deleting edges.

2 Preliminaries

We consider a pairwise model with n variables $X_1, \dots, X_n$ and graph topology $(\mathcal{V}, \mathcal{E})$: $\mathcal{V}$ contains nodes $\{1, \dots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains an edge for each pairwise relationship. We sometimes consider multi-label models where each variable X_i takes values in $\{0, \dots, L_i - 1\}$, and sometimes restrict attention to binary models where $X_i \in \mathbb{B} = \{0, 1\}$ $\forall i$. Let $x = (x_1, \dots, x_n)$ be a configuration of all the variables, and $\mathcal{N}(i)$ be the neighbors of i. For all analysis of binary models, to be consistent with Welling and Teh (2001) and Weller and Jebara (2013), we assume a reparameterization such that $p(x) = \frac{e^{-E(x)}}{Z}$, where the energy of a configuration, $E = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} W_{ij} x_i x_j$, with singleton potentials θ_i and edge weights W_{ij}.

2.1 Clamping a variable and related definitions

We shall find it useful to examine sub-partition functions obtained by *clamping* one particular variable X_i, that is we consider the model on the $n-1$ variables $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_n$ obtained by setting X_i equal to one of its possible values.

Let $Z|_{X_i=a}$ be the sub-partition function on the model obtained by setting $X_i = a, a \in \{0, \dots, L_i - 1\}$. Observe that true partition functions and marginals are self-consistent in the following sense:

$$Z = \sum_{j=0}^{L_i - 1} Z|_{X_i=j} \; \forall i \in \mathcal{V}, \qquad p(X_i = a) = \frac{Z|_{X_i=a}}{\sum_{j=0}^{L_i-1} Z|_{X_i=j}}. \tag{1}$$

This is not true in general for approximate forms of inference,[1] but if the model has no cycles, then in many cases of interest, (1) does hold, motivating the following definition.

Definition 1. We say an approximation to the log-partition function Z_A is *ExactOnTrees* if it may be specified by the variational formula $-\log Z_A = \min_{q \in Q} F_A(q)$ where: (1) Q is some compact space that includes the marginal polytope; (2) F_A is a function of the (pseudo-)distribution q (typically a free energy approximation); and (3) For any model, whenever a subset of variables $\mathcal{V}' \subseteq \mathcal{V}$ is clamped to particular values $P = \{p_i \in \{0, \dots, L_i - 1\}, \; \forall X_i \in \mathcal{V}'\}$, i.e. $\forall X_i \in \mathcal{V}'$, we constrain

[1] For example, consider a single cycle with positive edge weights. This has $Z_B < Z$ (Weller et al., 2014), yet after clamping any variable, each resulting sub-model is a tree hence the Bethe approximation is exact.

$X_i = p_i$, which we write as $\mathcal{V}' \leftarrow P$, and the remaining induced graph on $\mathcal{V} \setminus \mathcal{V}'$ is acyclic, then the approximation is exact, i.e. $Z_A|_{\mathcal{V}' \leftarrow P} = Z|_{\mathcal{V}' \leftarrow P}$. Similarly, define an approximation to be in the broader class of *NotSmallerOnTrees* if it satisfies all of the above properties except that condition (3) is relaxed to $Z_A|_{\mathcal{V}' \leftarrow P} \geq Z|_{\mathcal{V}' \leftarrow P}$. Note that the Bethe approximation is ExactOnTrees, and approximations such as TRW are NotSmallerOnTrees, in both cases whether using the marginal polytope or any relaxation thereof, such as the cycle or local polytope (Weller et al., 2014).

We shall derive bounds on Z_A with the following idea: Obtain upper or lower bounds on the approximation achieved by clamping and summing over the approximate sub-partition functions; Repeat until an acyclic graph is reached, where the approximation is either exact or bounded. We introduce the following related concept from graph theory.

Definition 2. A *feedback vertex set* (FVS) of a graph is a set of vertices whose removal leaves a graph without cycles. Determining if there exists a feedback vertex set of a given size is a classical NP-hard problem (Karp, 1972). There is a significant literature on determining the minimum cardinality of an FVS of a graph G, which we write as $\nu(G)$. Further, if vertices are assigned non-negative weights, then a natural problem is to find an FVS with minimum weight, which we write as $\nu_w(G)$. An FVS with a factor 2 approximation to $\nu_w(G)$ may be found in time $O(|\mathcal{V}| + |\mathcal{E}| \log |\mathcal{E}|)$ (Bafna et al., 1999). For pairwise multi-label MRFs, we may create a weighted graph from the topology by assigning each node i a weight of $\log L_i$, and then compute the corresponding $\nu_w(G)$.

3 Lower Bound on Approximate Partition Functions

We obtain a lower bound on any approximation that is NotSmallerOnTrees by observing that $Z_A \geq Z_A|_{X_n=j} \; \forall j$ from the definition (the sub-partition functions optimize over a subset).

Theorem 3. *If a pairwise MRF has topology with an FVS of size n and corresponding values $L_1, \dots, L_n$, then for any approximation that is NotSmallerOnTrees, $Z_A \geq \frac{Z}{\prod_{i=1}^n L_i}$.*

Proof. We proceed by induction on n. The base case $n = 0$ holds by the assumption that Z_A is NotSmallerOnTrees. Now assume the result holds for $n - 1$ and consider a MRF which requires n vertices to be deleted to become acyclic. Clamp variable X_n at each of its L_n values to create the approximation $Z_A^{(n)} := \sum_{j=0}^{L_n-1} Z_A|_{X_n=j}$. By the definition of NotSmallerOnTrees, $Z_A \geq Z_A|_{X_n=j} \; \forall j$; and by the inductive hypothesis, $Z_A|_{X_n=j} \geq \frac{Z|_{X_n=j}}{\prod_{i=1}^{n-1} L_i}$.
Hence, $L_n Z_A \geq Z_A^{(n)} = \sum_{j=0}^{L_n-1} Z_A|_{X_n=j} \geq \frac{1}{\prod_{i=1}^{n-1} L_i} \sum_{j=0}^{L_n-1} Z|_{X_n=j} = \frac{Z}{\prod_{i=1}^{n-1} L_i}$. □

By considering an FVS with minimum $\prod_{i=1}^n L_i$, Theorem 3 is equivalent to the following result.

Theorem 4. *For any approximation that is NotSmallerOnTrees, $Z_A \geq Ze^{-\nu_w}$.*

This bound applies to general multi-label models with any pairwise and singleton potentials (no need for attractive). The bound is trivial for a tree, but already for a binary model with one cycle we obtain that $Z_B \geq Z/2$ for any potentials, even over the marginal polytope. The bound is tight, at least for uniform $L_i = L \; \forall i$.[2] The bound depends only on the vertices that must be deleted to yield a graph with no cycles, not on the number of cycles (which clearly upper bounds $\nu(G)$). For binary models, exact inference takes time $\Theta((|\mathcal{V}| - |\nu(G)|)2^{\nu(G)})$.

4 Attractive Binary Pairwise Models

In this Section, we restrict attention to the standard Bethe approximation. We shall use results derived in (Welling and Teh, 2001) and (Weller and Jebara, 2013), and adopt similar notation. The Bethe partition function, Z_B, is defined as in Definition 1, where Q is set as the *local polytope* relaxation and F_A is the Bethe free energy, given by $\mathcal{F}(q) = \mathbb{E}_q(E) - S_B(q)$, where E is the energy

[2] For example, in the binary case: consider a sub-MRF on a cycle with no singleton potentials and uniform, very high edge weights. This can be shown to have $Z_B \approx Z/2$ (Weller et al., 2014). Now connect ν of these together in a chain using very weak edges (this construction is due to N. Ruozzi).

and S_B is the Bethe pairwise entropy approximation (see Wainwright and Jordan, 2008 for details). We consider attractive binary pairwise models and apply similar clamping ideas to those used in §3. In §4.1 we show that clamping can never decrease the approximate Bethe partition function, then use this result in §4.2 to prove that $Z_B \leq Z$ for this class of model. In deriving the clamping result of §4.1, in Theorem 7 we show an interesting, stronger result on how the optimum Bethe free energy changes as the singleton marginal q_i is varied over $[0, 1]$.

4.1 Clamping a variable can only increase the Bethe partition function

Let Z_B be the Bethe partition function for the original model. Clamp variable X_i and form the new approximation $Z_B^{(i)} = \sum_{j=0}^{1} Z_B|_{X_i=j}$. In this Section, we shall prove the following Theorem.

Theorem 5. *For an attractive binary pairwise model and any variable X_i, $Z_B^{(i)} \geq Z_B$.*

We first introduce notation and derive preliminary results, which build to Theorem 7, our strongest result, from which Theorem 5 easily follows. Let $q = (q_1, \ldots, q_n)$ be a location in n-dimensional pseudomarginal space, i.e. q_i is the singleton pseudomarginal $q(X_i = 1)$ in the local polytope. Let $\mathcal{F}(q)$ be the Bethe free energy computed at q using Bethe optimum pairwise pseudomarginals given by the formula for $q(X_i = 1, X_j = 1) = \xi_{ij}(q_i, q_j, W_{ij})$ in (Welling and Teh, 2001), i.e. for an attractive model, for edge (i, j), ξ_{ij} is the lower root of

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

where $\alpha_{ij} = e^{W_{ij}} - 1$, and $W_{ij} > 0$ is the strength (associativity) of the log-potential edge weight.

Let $\mathcal{G}(q) = -\mathcal{F}(q)$. Note that $\log Z_B = \max_{q \in [0,1]^n} \mathcal{G}(q)$. For any $x \in [0, 1]$, consider the optimum constrained by holding $q_i = x$ fixed, i.e. let $\log Z_{Bi}(x) = \max_{q \in [0,1]^n : q_i = x} \mathcal{G}(q)$. Let $r^*(x) = (r_1^*(x), \ldots, r_{i-1}^*(x), r_{i+1}^*(x), \ldots, r_n^*(x))$ with corresponding pairwise terms $\{\xi_{ij}^*\}$, be an $\arg\max$ for where this optimum occurs. Observe that $\log Z_{Bi}(0) = \log Z_B|_{X_i=0}, \log Z_{Bi}(1) = \log Z_B|_{X_i=1}$ and $\log Z_B = \log Z_{Bi}(q_i^*) = \max_{q \in [0,1]^n} \mathcal{G}(q)$, where q_i^* is a location of X_i at which the global optimum is achieved.

To prove Theorem 5, we need a sufficiently good upper bound on $\log Z_{Bi}(q_i^*)$ compared to $\log Z_{Bi}(0)$ and $\log Z_{Bi}(1)$. First we demonstrate what such a bound could be, then prove that this holds. Let $S_i(x) = -x \log x - (1 - x) \log(1 - x)$ be the standard singleton entropy.

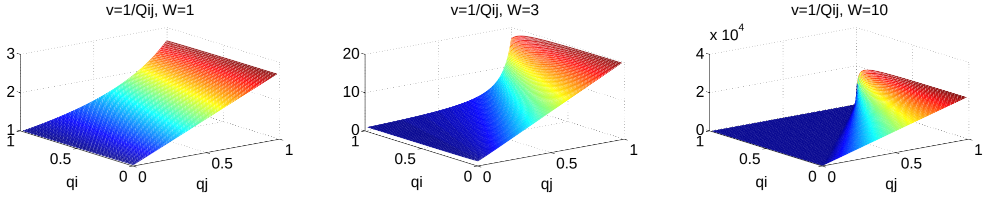
Lemma 6 (Demonstrating what would be a sufficiently good upper bound on $\log Z_B$)**.** *If $\exists x \in [0, 1]$ such that $\log Z_B \leq x \log Z_{Bi}(1) + (1 - x) \log Z_{Bi}(0) + S_i(x)$, then:*
(i) $Z_{Bi}(0) + Z_{Bi}(1) - Z_B \geq e^m f_c(x)$ where $f_c(x) = 1 + e^c - e^{xc+S_i(x)}$,
$m = \min(\log Z_{Bi}(0), \log Z_{Bi}(1))$ and $c = |\log Z_{Bi}(1) - \log Z_{Bi}(0)|$; and
(ii) $\forall x \in [0, 1]$, $f_c(x) \geq 0$ with equality iff $x = \sigma(c) = 1/(1 + \exp(-c))$, the sigmoid function.

Proof. (i) This follows easily from the assumption. (ii) This is easily checked by differentiating. It is also given in (Koller and Friedman, 2009, Proposition 11.8). □

See Figure 6 in the Supplement for example plots of the function $f_c(x)$. Lemma 6 motivates us to consider if perhaps $\log Z_{Bi}(x)$ might be upper bounded by $x \log Z_{Bi}(1) + (1-x) \log Z_{Bi}(0) + S_i(x)$, i.e. the linear interpolation between $\log Z_{Bi}(0)$ and $\log Z_{Bi}(1)$, plus the singleton entropy term $S_i(x)$. It is easily seen that this would be true if $r^*(q_i)$ were constant. In fact, we shall show that $r^*(q_i)$ varies in a particular way which yields the following, stronger result, which, together with Lemma 6, will prove Theorem 5.

Theorem 7. *Let $A_i(q_i) = \log Z_{Bi}(q_i) - S_i(q_i)$. For an attractive binary pairwise model, $A_i(q_i)$ is a convex function.*

Proof. We outline the main points of the proof. Observe that $A_i(x) = \max_{q \in [0,1]^n : q_i = x} \mathcal{G}(q) - S_i(x)$, where $\mathcal{G}(q) = -\mathcal{F}(q)$. Note that there may be multiple $\arg\max$ locations $r^*(x)$. As shown in (Weller and Jebara, 2013), $\mathcal{F}$ is at least thrice differentiable in $(0, 1)^n$ and all stationary points lie in the interior $(0, 1)^n$. Given our conditions, the 'envelope theorem' of (Milgrom, 1999, Theorem



(a) W=1 (b) W=3 (c) W=10

Figure 1: 3d plots of $v_{ij} = Q_{ij}^{-1}$, using $\xi_{ij}(q_i, q_j, W)$ from (Welling and Teh, 2001).

1) applies, showing that A_i is continuous in $[0,1]$ with right derivative[3]

$$A'_{i+}(x) = \max_{r*(q_i=x)} \frac{\partial}{\partial x}\left[\mathcal{G}(q_i = x, r^*(x)) - S_i(x)\right] = \max_{r*(q_i=x)} \frac{\partial}{\partial x}\left[\mathcal{G}(q_i = x, r^*(x))\right] - \frac{dS_i(x)}{dx}. \tag{3}$$

We shall show that this is non-decreasing, which is sufficient to show the convexity result of Theorem 7. To evaluate the right hand side of (3), we use the derivative shown by Welling and Teh (2001):

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

where $\log Q_i = \log \dfrac{(1-q_i)^{d_i-1}}{q_i^{d_i-1}} \dfrac{\prod_{j\in\mathcal{N}(i)}(q_i - \xi_{ij})}{\prod_{j\in\mathcal{N}(i)}(1+\xi_{ij} - q_i - q_j)}$ (as in Weller and Jebara, 2013)

$$= \log \frac{q_i}{1-q_i} + \log \prod_{j\in\mathcal{N}(i)} Q_{ij}, \text{ here defining } Q_{ij} = \left(\frac{q_i - \xi_{ij}}{1+\xi_{ij} - q_i - q_j}\right)\left(\frac{1-q_i}{q_i}\right).$$

A key observation is that the $\log \frac{q_i}{1-q_i}$ term is exactly $-\frac{dS_i(q_i)}{dq_i}$, and thus cancels the $-\frac{dS_i(x)}{dx}$ term at the end of (3). Hence, $A'_{i+}(q_i) = \max_{r*(q_i)}\left[-\sum_{j\in\mathcal{N}(i)} \log Q_{ij}(q_i, r_j^*, \xi_{ij}^*)\right]$. [4]

It remains to show that this expression is non-decreasing with q_i. We shall show something stronger, that at every $\arg\max r^*(q_i)$, and for all $j \in \mathcal{N}(i)$, $-\log Q_{ij}$ is non-decreasing $\Leftrightarrow v_{ij} = Q_{ij}^{-1}$ is non-decreasing. The result then follows since the $\max$ of non-decreasing functions is non-decreasing.

See Figure 1 for example plots of the v_{ij} function, and observe that v_{ij} appears to decrease with q_i (which is unhelpful here) while it increases with q_j. Now, in an attractive model, the Bethe free energy is *submodular*, i.e. $\frac{\partial^2 \mathcal{F}}{\partial q_i \partial q_j} \le 0$ (Weller and Jebara, 2013), hence as q_i increases, $r_j^*(q_i)$ can only increase (Topkis, 1978). For our purpose, we must show that $\frac{dr_j^*}{dq_i}$ is sufficiently large such that $\frac{dv_{ij}}{dq_i} \ge 0$. This forms the remainder of the proof.

At any particular $\arg\max r^*(q_i)$, writing $v = v_{ij}[q_i, r_j^*(q_i), \xi_{ij}^*(q_i, r_j^*(q_i))]$, we have

$$\begin{aligned} \frac{dv}{dq_i} &= \frac{\partial v}{\partial q_i} + \frac{\partial v}{\partial \xi_{ij}}\frac{d\xi_{ij}^*}{dq_i} + \frac{\partial v}{\partial q_j}\frac{dr_j^*}{dq_i} \\ &= \frac{\partial v}{\partial q_i} + \frac{\partial v}{\partial \xi_{ij}}\frac{\partial \xi_{ij}^*}{\partial q_i} + \frac{dr_j^*}{dq_i}\left(\frac{\partial v}{\partial \xi_{ij}}\frac{\partial \xi_{ij}^*}{\partial q_j} + \frac{\partial v}{\partial q_j}\right). \end{aligned} \tag{4}$$

From (Weller and Jebara, 2013), $\frac{\partial \xi_{ij}}{\partial q_i} = \frac{\alpha_{ij}(q_j - \xi_{ij}) + q_j}{1 + \alpha_{ij}(q_i - \xi_{ij} + q_j - \xi_{ij})}$ and similarly, $\frac{\partial \xi_{ij}}{\partial q_j} = \frac{\alpha_{ij}(q_i - \xi_{ij}) + q_i}{1 + \alpha_{ij}(q_j - \xi_{ij} + q_i - \xi_{ij})}$, where $\alpha_{ij} = e^{W_{ij}} - 1$. The other partial derivatives are easily derived: $\frac{\partial v}{\partial q_i} = \frac{q_i(q_j - 1)(1-q_i) + (1 + \xi_{ij} - q_i - q_j)(q_i - \xi_{ij})}{(1-q_i)^2 (q_i - \xi_{ij})^2}$, $\frac{\partial v}{\partial \xi_{ij}} = \frac{q_i(1-q_j)}{(1-q_i)(q_i - \xi_{ij})^2}$, and $\frac{\partial v}{\partial q_j} = \frac{-q_i}{(1-q_i)(q_i - \xi_{ij})}$.

The only remaining term needed for (4) is $\frac{dr_j^*}{dq_i}$. The following results are proved in the Appendix, subject to a technical requirement that at an $\arg\max$, the reduced Hessian $H_{\setminus i}$, i.e. the matrix of

[3] This result is similar to Danskin's theorem (Bertsekas, 1995). Intuitively, for multiple $\arg\max$ locations, each may increase at a different rate, so here we must take the $\max$ of the derivatives over all the $\arg\max$.

[4] We remark that Q_{ij} is the ratio $\left(\frac{p(X_i=1, X_j=0)}{p(X_i=0, X_j=0)}\right) \Big/ \left(\frac{p(X_i=1)}{p(X_i=0)}\right) = \frac{p(X_j=0|X_i=1)}{p(X_j=0|X_i=0)}$.

second partial derivatives of $\mathcal{F}$ after removing the ith row and column, must be non-singular in order to have an invertible locally linear function. Call this required property $\mathcal{P}$. By nature, each $H_{\setminus i}$ is positive semi-definite. If needed, a small perturbation argument allows us to assume that no eigenvalue is 0, then in the limit as the perturbation tends to 0, Theorem 7 holds since the limit of convex functions is convex. Let $[n] = \{1, \dots, n\}$ and G be the topology of the MRF.

Theorem 8. *For any $k \in [n] \setminus i$, let C_k be the connected component of $G \setminus i$ that contains X_k. If $C_k + i$ is a tree, then $\frac{dr_k^*}{dq_i} = \prod_{(s \to t) \in P(i \rightsquigarrow k)} \frac{\xi_{st}^* - r_s^* r_t^*}{r_s^*(1-r_s^*)}$,where $P(i \rightsquigarrow k)$ is the unique path from i to k in $C_k + i$, and for notational convenience, define $r_i^* = q_i$. Proof in Appendix (subject to $\mathcal{P}$).*

In fact, this result applies for any combination of attractive and repulsive edges. The result is remarkable, yet also intuitive. In the numerator, $\xi_{st} - q_s q_t = \mathrm{Cov}_q(X_s, X_t)$, increasing with W_{ij} and equal to 0 at $W_{ij} = 0$ (Weller and Jebara, 2013), and in the denominator, $q_s(1 - q_s) = \mathrm{Var}_q(X_s)$, hence the ratio is exactly what is called in finance the beta of X_t with respect to X_s.[5]

In particular, Theorem 8 shows that for any $j \in \mathcal{N}(i)$ whose component is a tree, $\frac{dr_j^*}{dq_i} = \frac{\xi_{ij}^* - q_i r_j^*}{q_i(1-q_i)}$. The next result shows that in an attractive model, additional edges can only reinforce this sensitivity.

Theorem 9. *In an attractive model with edge (i, j), $\frac{dr_j^*(q_i)}{dq_i} \geq \frac{\xi_{ij}^* - q_i r_j^*}{q_i(1-q_i)}$. Proof in Appendix (subject to $\mathcal{P}$).*

Now collecting all terms, substituting into (4), and using (2), after some algebra yields that $\frac{dv}{dq_i} \geq 0$, as required to prove Theorem 7. This now also proves Theorem 5. □

4.2 The Bethe partition function lower bounds the true partition function

Theorem 5, together with an argument similar to the proof of Theorem 3, easily yields a new proof that $Z_B \leq Z$ for an attractive binary pairwise model.

Theorem 10 (first proved by Ruozzi, 2012)**.** *For an attractive binary pairwise model, $Z_B \leq Z$.*

Proof. We shall use induction on n to show that the following statement holds for all n:
If a MRF may be rendered acyclic by deleting n vertices $v_1, \dots, v_n$, then $Z_B \leq Z$.
The base case $n = 0$ holds since the Bethe approximation is ExactOnTrees. Now assume the result holds for $n-1$ and consider a MRF which requires n vertices to be deleted to become acyclic. Clamp variable X_n and consider $Z_B^{(n)} = \sum_{j=0}^{1} Z_B|_{X_n=j}$. By Theorem 5, $Z_B \leq Z_B^{(n)}$; and by the inductive hypothesis, $Z_B|_{X_n=j} \leq Z|_{X_n=j}\ \forall j$. Hence, $Z_B \leq \sum_{j=0}^{1} Z_B|_{X_n=j} \leq \sum_{j=0}^{1} Z|_{X_n=j} = Z$. □

5 Experiments

For an approximation which is ExactOnTrees, it is natural to try clamping a few variables to remove cycles from the topology. Here we run experiments on binary pairwise models to explore the potential benefit of clamping even just one variable, though the procedure can be repeated. For exact inference, we used the junction tree algorithm. For approximate inference, we used Frank-Wolfe (FW) (Frank and Wolfe, 1956): At each iteration, a tangent hyperplane to the approximate free energy is computed at the current point, then a move is made to the best computed point along the line to the vertex of the local polytope with the optimum score on the hyperplane. This proceeds monotonically, even on a non-convex surface, hence will converge (since it is bounded), though it may be only to a local optimum and runtime is not guaranteed. This method typically produces good solutions in reasonable time compared to other approaches (Belanger et al., 2013; Weller et al., 2014) and allows direct comparison to earlier results (Meshi et al., 2009; Weller et al., 2014). To further facilitate comparison, in this Section we use the same unbiased reparameterization used by Weller et al. (2014), with $E = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} \frac{W_{ij}}{2} \left[x_i x_j + (1 - x_i)(1 - x_j) \right]$.

[5] Sudderth et al. (2007) defined a different, symmetric $\beta_{st} = \frac{\xi_{st} - q_s q_t}{q_s(1-q_s) q_t(1-q_t)}$ for analyzing loop series. In our context, we suggest that the ratio defined above may be a better Bethe beta.

Test models were constructed as follows: For n variables, singleton potentials were drawn $\theta_i \sim U[-T_{max}, T_{max}]$; edge weights were drawn $W_{ij} \sim U[0, W_{max}]$ for attractive models, or $W_{ij} \sim U[-W_{max}, W_{max}]$ for general models. For models with random edges, we constructed Erdős-Renyi random graphs (rejecting disconnected samples), where each edge has independent probability p of being present. To observe the effect of increasing n while maintaining approximately the same average degree, we examined $n = 10, p = 0.5$ and $n = 50, p = 0.1$. We also examined models on a complete graph topology with 10 variables for comparison with TRW in (Weller et al., 2014). 100 models were generated for each set of parameters with varying T_{max} and W_{max} values.

Results are displayed in Figures 2 to 4 showing average absolute error of $\log Z_B$ vs $\log Z$ and average ℓ_1 error of singleton marginals. The legend indicates the different methods used: *Original* is FW on the initial model; then various methods were used to select the variable to clamp, before running FW on the 2 resulting submodels and combining those results. *avg Clamp* for $\log Z$ means average over all possible clampings, whereas *all Clamp* for marginals computes each singleton marginal as the estimated $\hat{p_i} = Z_B|_{X_i=1}/(Z_B|_{X_i=0} + Z_B|_{X_i=1})$. *best Clamp* uses the variable which with hindsight gave the best improvement in $\log Z$ estimate, thereby showing the best possible result for $\log Z$. Similarly, *worst Clamp* picks the variable which showed worst performance. Where one variable is clamped, the respective marginals are computed thus: for the clamped variable X_i, use $\hat{p_i}$ as before; for all others, take the weighted average over the estimated Bethe pseudomarginals on each sub-model using weights $1 - \hat{p_i}$ and $\hat{p_i}$ for sub-models with $X_i = 0$ and $X_i = 1$ respectively.

maxW and Mpower are heuristics to try to pick a good variable in advance. Ideally, we would like to break heavy cycles, but searching for these is NP-hard. maxW is a simple $O(|\mathcal{E}|)$ method which picks a variable X_i with $\max_{i \in \mathcal{V}} \sum_{j \in \mathcal{N}(i)} |W_{ij}|$, and can be seen to perform well (Liu et al., 2012 proposed the same maxW approach for inference in Gaussian models). One way in which maxW can make a poor selection is to choose a variable at the centre of a large star configuration but far from any cycle. Mpower attempts to avoid this by considering the convergent series of powers of a modified W matrix, but on the examples shown, this did not perform significantly better. See §8.1 in the Appendix for more details on Mpower and further experimental results.

FW provides no runtime guarantee when optimizing over a non-convex surface such as the Bethe free energy, but across all parameters, the average combined runtimes on the two clamped sub-models was the same order of magnitude as that for the original model, see Figure 5.

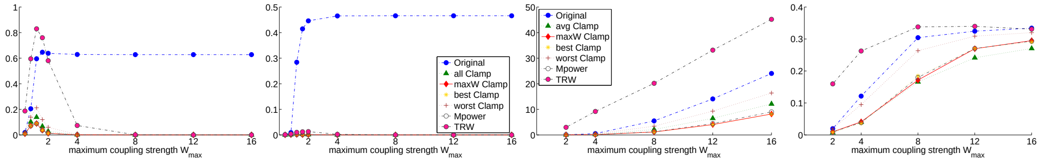
6 Discussion

The results of §4 immediately also apply to any binary pairwise model where a subset of variables may be flipped to yield an attractive model, i.e. where the topology has no frustrated cycle (Weller et al., 2014), and also to any model that may be reduced to an attractive binary pairwise model (Schlesinger and Flach, 2006; Zivny et al., 2009). For this class, together with the lower bound of §3, we have sandwiched the range of Z_B (equivalently, given Z_B, we have sandwiched the range of the true partition function Z) and bounded its error; further, clamping any variable, solving for optimum $\log Z_B$ on sub-models and summing is guaranteed to be more accurate than solving on the original model. In some cases, it may also be faster; indeed, some algorithms such as LBP may fail on the original model but perform well on clamped sub-models.

Methods presented may prove useful for analyzing general (non-attractive) models, or for other applications. As one example, it is known that the Bethe free energy is convex for a MRF whose topology has at most one cycle (Pakzad and Anantharam, 2002). In analyzing the Hessian of the Bethe free energy, we are able to leverage this to show the following result, which may be useful for optimization (proof in Appendix; this result was conjectured by N. Ruozzi).

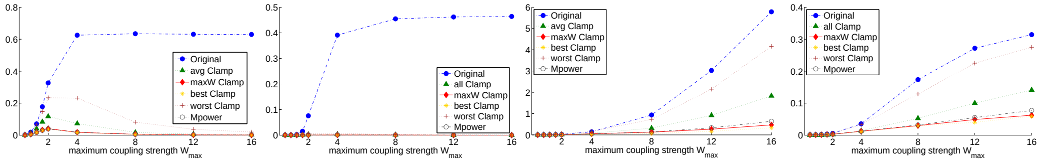
Lemma 11. *In a binary pairwise MRF (attractive or repulsive edges, any topology), for any subset of variables $S \subseteq \mathcal{V}$ whose induced topology contains at most one cycle, the Bethe free energy (using optimum pairwise marginals) over S, holding variables $\mathcal{V} \setminus S$ at fixed singleton marginals, is convex.*

In §5, clamping appears to be very helpful, especially for attractive models with low singleton potentials where results are excellent (overcoming TRW's advantage in this context), but also for general models, particularly with the simple maxW selection heuristic. We can observe some decline in benefit as n grows but this is not surprising when clamping just a single variable. Note, however,



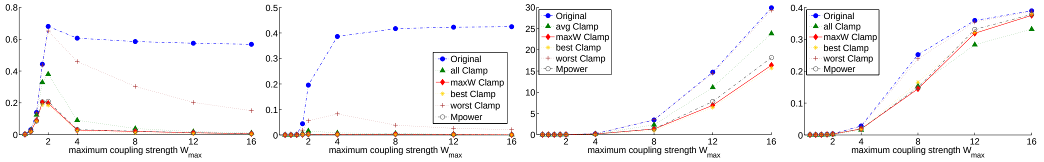
(a) attractive $\log Z$, $T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z$, $T_{max} = 2$ (d) general margs, $T_{max} = 2$

Figure 2: Average errors vs true, **complete graph on $n = 10$. TRW in pink**. Consistent legend throughout.



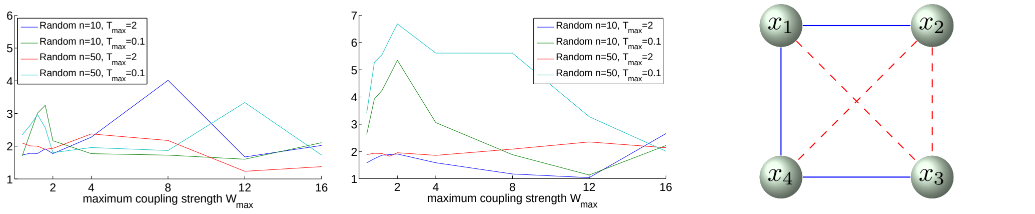
(a) attractive $\log Z$, $T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z$, $T_{max} = 2$ (d) general margs, $T_{max} = 2$

Figure 3: Average errors vs true, **random graph on $n = 10, p = 0.5$**. Consistent legend throughout.



(a) attractive $\log Z$, $T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z$, $T_{max} = 2$ (d) general margs, $T_{max} = 2$

Figure 4: Average errors vs true, **random graph on $n = 50, p = 0.1$**. Consistent legend throughout.



(a) attractive random graphs (b) general random graphs (c) Blue (dashed red) edges are attractive (repulsive) with edge weight $+2$ (-2). No singleton potentials.

Figure 5: Left: Average ratio of combined sub-model runtimes to original runtime (using maxW, other choices are similar). Right: Example model where *clamping any variable worsens* the Bethe approximation to $\log Z$.

that non-attractive models exist such that clamping and summing over *any variable* can lead to a *worse* Bethe approximation of $\log Z$, see Figure 5c for a simple example on four variables.

It will be interesting to explore the extent to which our results may be generalized beyond binary pairwise models. Further, it is tempting to speculate that similar results may be found for other approximations. For example, some methods that upper bound the partition function, such as TRW, might always yield a lower (hence better) approximation when a variable is clamped.

Acknowledgments. We thank Nicholas Ruozzi for careful reading, and Nicholas, David Sontag, Aryeh Kontorovich and Tomaž Slivnik for helpful discussion and comments. This work was supported in part by NSF grants IIS-1117631 and CCF-1302269.

References

V. Bafna, P. Berman, and T. Fujito. A 2-approximation algorithm for the undirected feedback vertex set problem. *SIAM Journal on Discrete Mathematics*, 12(3):289–9, 1999.

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

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

A. Choi and A. Darwiche. Approximating the partition function by deleting and then correcting for model edges. In *Uncertainty in Artificial Intelligence (UAI)*, 2008.

A. Darwiche. *Modeling and Reasoning with Bayesian Networks*. Cambridge University Press, 2009.

F. Eaton and Z. Ghahramani. Choosing a variable to clamp: Approximate inference using conditioned belief propagation. In *Artificial Intelligence and Statistics*, 2009.

K. Fan. Topological proofs for certain theorems on matrices with non-negative elements. *Monatshefte fr Mathematik*, 62:219–237, 1958.

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.

R. Karp. *Complexity of Computer Computations*, chapter Reducibility Among Combinatorial Problems, pages 85–103. New York: Plenum., 1972.

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

S. Lauritzen and D. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society series B*, 50:157–224, 1988.

Y. Liu, V. Chandrasekaran, A. Anandkumar, and A. Willsky. Feedback message passing for inference in Gaussian graphical models. *IEEE Transactions on Signal Processing*, 60(8):4135–4150, 2012.

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*, 2009.

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

J. Mitchell. Branch-and-cut algorithms for combinatorial optimization problems. *Handbook of Applied Optimization*, pages 65–77, 2002.

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

M. Padberg and G. Rinaldi. A branch-and-cut algorithm for the resolution of large-scale symmetric traveling salesman problems. *SIAM review*, 33(1):60–100, 1991.

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

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

M. Peot and R. Shachter. Fusion and propagation with multiple observations in belief networks. *Artificial Intelligence*, 48(3):299–318, 1991.

I. Rish and R. Dechter. Resolution versus search: Two strategies for SAT. *Journal of Automated Reasoning*, 24 (1-2):225–275, 2000.

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.

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

D. Topkis. Minimizing a submodular function on a lattice. *Operations Research*, 26(2):305–321, 1978.

P. Vontobel. Counting in graph covers: A combinatorial characterization of the Bethe entropy function. *Information Theory, IEEE Transactions on*, 59(9):6018–6048, Sept 2013. ISSN 0018-9448.

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 *AISTATS*, 2013.

A. Weller and T. Jebara. Approximating the Bethe partition function. In *UAI*, 2014.

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.

S. Zivny, D. Cohen, and P. Jeavons. The expressive power of binary submodular functions. *Discrete Applied Mathematics*, 157(15):3347–3358, 2009.

APPENDIX: SUPPLEMENTARY MATERIAL FOR *CLAMPING VARIABLES AND APPROXIMATE INFERENCE*

In this Appendix, we provide:

- Figure 6 showing examples of the $f_c(x)$ function introduced in Lemma 6;
- In Section 7, theoretical results on the Hessian leading to proofs of Theorem 8 and (a stronger version of) Theorem 9 from §4.1, and Lemma 11 from §6; and
- In Section 8, additional illustrative experimental results with details on the Mpower selection heuristic.

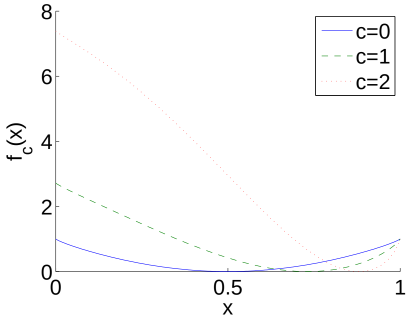


Figure 6: Plots of upper bound $f_c(x)$ against x for various values of c

7 The Hessian and Proofs of Earlier Results

In this Section, we first discuss properties of the Hessian in §7.1, then use these in §7.2 to prove Theorems 8 and 9, and Lemma 11. Define the *interior* to be all points $q \in (0,1)^n$. Recall that $r^*(x) = (r_1^*(q_i), \dots, r_{i-1}^*(q_i), r_{i+1}^*(q_i), \dots, r_n^*(q_i))$ with corresponding pairwise terms $\{\xi_{ij}^*\}$, is an $\arg\max$ of $\mathcal{G}(q) = -\mathcal{F}(q)$ where q_i is held fixed at a particular value. For notational convenience, define $r_i^* = q_i$.

7.1 Properties of the Hessian

From (Weller and Jebara, 2013), we have all terms of the Hessian matrix $H_{jk} = \frac{\partial^2 \mathcal{F}}{\partial q_j \partial q_k}$:

$$H_{jk} = \begin{cases} \frac{q_j q_k - \xi_{jk}}{T_{jk}} & \text{if } (j,k) \in \mathcal{E} \\ 0 & \text{if } (j,k) \notin \mathcal{E} \end{cases}, \quad H_{jj} = -\frac{d_j - 1}{q_j(1-q_j)} + \sum_{k \in \mathcal{N}(j)} \frac{q_k(1-q_k)}{T_{jk}}, \tag{5}$$

where $d_j = |\mathcal{N}(j)|$ is the degree of j, and $T_{jk} = q_j q_k (1-q_j)(1-q_k) - (\xi_{jk} - q_j q_k)^2 \geq 0$, with equality only at an edge (i.e. q_j or $q_k \in \{0,1\}$). For an attractive edge (j,k), in the interior, as shown in (Weller and Jebara, 2013, Lemma 14 in Supplement), $\xi_{jk} - q_j q_k > 0$ and hence $H_{jk} < 0$.

Now write

$$H_{jj} = \frac{1}{q_j(1-q_j)} + \sum_{k \in \mathcal{N}(j)} \left(\frac{q_k(1-q_k)}{T_{jk}} - \frac{1}{q_j(1-q_j)} \right). \tag{6}$$

Consider the term in large parentheses for some $k \in \mathcal{N}(j)$. First observe that the term is ≥ 0, strictly > 0 in the interior, whether the edge is attractive or repulsive. Since $H_{jj} > 0$, on the surface $\frac{\partial \mathcal{F}}{\partial q_j}\Big|_{r^*} = 0$, we have

$$\frac{\partial r_j^*}{\partial r_k^*} = -\frac{H_{jk}}{H_{jj}}\Big|_{r^*}, \tag{7}$$

which also holds for $k = i$ where we define $r_i^* = q_i$.

Further, we may incorporate the term for k to obtain

$$H_{jj} \geq \frac{1}{q_j(1-q_j)} + \frac{q_k(1-q_k)}{T_{jk}} - \frac{1}{q_j(1-q_j)} = \frac{q_k(1-q_k)}{T_{jk}},$$

with equality iff j has no neighbor other than k (again allowing $k = i$), in which case,

$$\frac{\partial r_j^*}{\partial r_k^*} = \frac{\xi_{jk}^* - r_j^* r_k^*}{r_k^*(1-r_k^*)}. \tag{8}$$

We also show the following results, though the remainder of this Section §7.1 is not used until later when we prove Theorem 9 in §7.2.1.

Considering the term in large parentheses from (6), using the definition of T_{jk}, we may write

$$\left(\frac{q_k(1-q_k)}{T_{jk}} - \frac{1}{q_j(1-q_j)} \right) = \left(\frac{\xi_{jk} - q_j q_k}{T_{jk}} \right) \left(\frac{\xi_{jk} - q_j q_k}{q_j(1-q_j)} \right) = -H_{jk}\beta_{j \to k}, \tag{9}$$

where we define $\beta_{j\to k} = \frac{\xi_{jk}-q_jq_k}{q_j(1-q_j)}$, which as mentioned in the main paper after Theorem 8, is equal to $\frac{\mathrm{Cov}_q(X_j,X_k)}{\mathrm{Var}_q(X_j)}$, called in finance the beta of X_k with respect to X_j. This is clearly positive for an attractive edge. We next show that the range of $\beta_{j\to k}$ is bounded, as would be expected for beta.

Lemma 12. *In the interior, for an edge* (j,k)*: if attractive,* $0 < \beta_{j\to k} \leq \frac{\alpha_{jk}}{\alpha_{jk}+1} = 1 - e^{-W_{jk}} < 1$*; if repulsive,* $-1 < e^{W_{jk}} - 1 = \alpha_{jk} \leq \beta_{j\to k} < 0$*. In either case,* $|\beta_{j\to k}| = \left| \frac{\xi_{jk}-q_jq_k}{q_j(1-q_j)} \right| \leq 1 - e^{-|W_{jk}|} < 1$.

Proof. This follows from (Weller and Jebara, 2013, Lemma 6) and the corresponding flipped result (Weller and Jebara, 2014, Lemma 10 in Supplement; consider each of the 2 cases for p_{jk} therein). □

Define $\beta^*_{j\to k} = \beta_{j\to k}\big|_{r^*}$. Regarding (8), note that $\beta^*_{j\to k} \geq \frac{\partial r_k^*}{\partial r_j^*}$ with equality iff $\mathcal{N}(k) = \{j\}$. This notation will become clear when we use it in §7.2.1 to prove Theorem 9.

7.2 Derivation of earlier results

Using the results of §7.1, we first provide a general Theorem from which Lemma 11 follows as an immediate corollary.

Theorem 13. *For any binary pairwise MRF where the Bethe free energy is convex, adding further variables to the model and holding them at fixed singleton marginal values (optimum pairwise marginals are computed using the formula of Welling and Teh, 2001), leaves the Bethe free energy over the original variables convex.*

Proof. The Bethe free energy is convex ⇔ the Hessian is everywhere positive semi-definite. When new variables are added to the system, considering (5) and (6), the only effect on the sub-Hessian restricted to the original variables is potentially to increase the diagonal terms H_{jj} for any original variable j which is adjacent to a new variable. By Weyl's inequality, this can only increase the minimum eigenvalue of the sub-Hessian, and the result follows. □

Since the Bethe free energy is convex for any model whose entire topology contains at most one cycle (Pakzad and Anantharam, 2002), Lemma 11 follows.

We next turn to Theorem 8, then use this to prove a stronger version of Theorem 9. Keep in mind that, as shown in (Weller and Jebara, 2013), each stationary point lies in an open region in the interior $q \in (0,1)^n$. Further, as discussed in §4.1, we assume that at any $\arg\max$ point $r^*(q_i)$, the reduced Hessian $H_{\setminus i}$ is non-singular. Hence, writing $\nabla_{n-1}\mathcal{F}\big|_{q_i}$ for the $(n-1)$-vector of partial derivatives $\frac{\partial \mathcal{F}(q)}{\partial q_j}\Big|_{q_i}$ $\forall j \neq i$, there is an open region around any $(q_i, r^*(q_i))$ where the function $\nabla_{n-1}\mathcal{F}\big|_{q_i} = 0$ may be well approximated by an invertible linear function, allowing us to solve (as in the implicit function theorem) for the total derivatives $\frac{dr_j^*}{dq_i}$ as the unique solutions to the linear system $\frac{dr_j^*}{dq_i} = \frac{\partial r_j^*}{\partial q_i} + \sum_{k\notin\{i,j\}} \frac{\partial r_j^*}{\partial r_k^*} \frac{dr_k^*}{dq_i}\ \forall j \neq i$, where here $\frac{\partial r_j^*}{\partial r_k^*}$ always means on the surface $\nabla_{n-1}\mathcal{F}\big|_{q_i} = 0$. In addition, since $H_{\setminus i}$ is real, symmetric, positive definite, with all main diagonal ≥ 0 and all off-diagonal ≤ 0, it is an M-matrix (indeed a Stieltjes matrix), which we shall use in §7.2.1. We assume these points for the rest of this Section.

Notation: Let $D_j = \frac{dr_j^*}{dq_i}$, and $\partial_{jk} = \frac{\partial r_j^*}{\partial r_k^*}$, so $D_j = \sum_{k\notin\{i,j\}} \partial_{jk} D_k + \partial_{ji}\ \forall j \neq i$. For notational convenience, define $r_i^* = q_i$ and take $D_i = 1$. Let $[n] = \{1, \dots, n\}$ and $[n] \setminus i = \{1, \dots, n\} \setminus \{i\}$. Note that $\partial_{jk} = \frac{\partial r_j^*}{\partial r_k^*} \leq \beta^*_{k\to j}$ (equality iff j has no neighbor other than k), as defined above. We shall write Hessian terms such as H_{jk} to mean $H_{jk}\big|_{r^*}$ where this is implied by the context.

We first need the following Lemma.

Lemma 14. *Consider a MRF with n variables, where then one more variable X_{n+1} is added with singleton marginal r^*_{n+1}, adjacent to exactly one of the original n variables, say X_a with $a \in [n]$ (note we allow $a = i$), then: $D_1, \dots, D_n$ are unaffected, and $D_{n+1} = \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{r^*_a(1-r^*_a)} D_a$.*

Proof. We have the linear system $D_j = \sum_{k\notin\{i,j\}} \partial_{jk} D_k + \partial_{ji}\ \forall j \in [n] \setminus i$. When X_{n+1} is added, this yields a new equation for D_{n+1}, which as shown in (8), is $D_{n+1} = \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{r^*_a(1-r^*_a)} D_a$, and the only other equation that changes is the one for D_a, where we write ∂'_{ak} and ∂'_{ai} for the new coefficients. Hence, it is sufficient to show that the earlier solutions for $D_1, \dots, D_n$ satisfy the new equation for D_a, i.e. if $D_a = \sum_{k\in[n+1]\setminus\{i,a\}} \partial'_{ak} D_k + \partial'_{ai}$.

Observe from (7) that $\partial'_{ak} = \partial_{ak} H_{aa}/H'_{aa}\ \ \forall k \in [n]$, where H'_{aa} incorporates the new X_{n+1} variable. Hence,

$$\begin{aligned}
\sum_{k\in[n+1]\setminus\{i,a\}} \partial'_{ak} D_k + \partial'_{ai} &= \frac{H_{aa}}{H'_{aa}} \left(\sum_{k\notin\{i,j\}} \partial_{ak} D_k + \partial_{ai} \right) + \partial'_{a,n+1} D_{n+1} \\
&= \frac{H_{aa}}{H'_{aa}} D_a + \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{T_{a,n+1} H'_{aa}} \frac{\xi^*_{a,n+1} - r^*_a r^*_{n+1}}{r^*_a(1-r^*_a)} D_a \quad \text{by (7), (5) and just above} \\
&= \frac{D_a}{H'_{aa}} \left[H_{aa} + \frac{\left(\xi^*_{a,n+1} - r^*_a r^*_{n+1}\right)^2}{T_{a,n+1} r^*_a(1-r^*_a)} \right] \\
&= \frac{D_a}{H'_{aa}} \left[H_{aa} + \left(\frac{r^*_{n+1}(1-r^*_{n+1})}{T_{a,n+1}} - \frac{1}{r^*_a(1-r^*_a)} \right) \right] \quad (\text{definition of } T_{a,n+1}) \\
&= \frac{D_a}{H'_{aa}} \left[H_{aa} + (H'_{aa} - H_{aa}) \right] = D_a
\end{aligned}$$

□

Theorem 8 may now be proved by induction on $|C_k|$. The base case $|C_k| = 1$ follows from (8). The inductive step follows from Lemma 14 by considering a leaf.

7.2.1 Proof of (stronger version of) Theorem 9:

As above, we have the linear system given by the following equations:

$$D_j = \sum_{k\notin\{i,j\}} \partial_{jk} D_k + \partial_{ji} \quad \forall j \neq i \qquad\qquad \Leftrightarrow -\partial_{ji} = \sum_{k\neq i} [\partial_{jk} - \delta_{jk}] D_k \tag{10}$$

with $\partial_{jk} = \dfrac{\partial r_j^*}{\partial r_k*} = -\dfrac{H_{jk}}{H_{jj}}\ \ k \notin \{i,j\},\ \ \partial_{jj} := 0, \qquad \partial_{ji} = \dfrac{\partial r_j^*}{\partial q_i} = -\dfrac{H_{ji}}{H_{jj}},\ \ \delta_{jk} = \begin{cases} 1 & j = k \\ 0 & j \neq k \end{cases}.$

Hence we may rewrite (10), multiplying by $-H_{jj}$, to give the equivalent system

$$\sum_{k\neq i} H_{jk} D_k = -H_{ji} \quad \forall j \neq i \tag{11}$$

Note equation (11) makes intuitive sense: for each variable X_j, we have $\mathcal{F}_j = 0$ at a stationary point, then taking the total derivative with respect to q_i gives $H_{ji} + \sum_{k \neq i} H_{jk} D_k = 0$.

By Theorem 8, we have the complete solution vector $D_k \ \forall k \neq i$ provided the topology is acyclic. In this setting, we rewrite the result of Theorem 8 using the β^* notation from above: $D_k = \prod_{(s \to t) \in P(i \rightsquigarrow k)} \beta^*_{s \to t}$, where here $P(i \rightsquigarrow k)$ is the *unique* path from i to k.

For a general graph, there may be many paths from i to k. Let $\Pi(i \rightsquigarrow k)$ be the set of all such directed paths. For any r^*, for any particular path $P(i \rightsquigarrow k) \in \Pi(i \rightsquigarrow k)$, define its *weight* to be $W[P(i \rightsquigarrow k)] = \prod_{(s \to t) \in P(i \rightsquigarrow k)} \beta^*_{s \to t}$. We shall prove the following result:

$$D_k \geq \max_{P(i \rightsquigarrow k) \in \Pi(i \rightsquigarrow k)} W[P(i \rightsquigarrow k)]. \qquad (12)$$

Note this is clearly stronger than Theorem 9 since $\forall j \in \mathcal{N}(i)$, the path going directly $i \to j$ is one member of $\Pi(i \rightsquigarrow j)$, though in general there may be many others.

For any particular r^*, let G' be the weighted directed graph formed from the topology of the MRF by replacing each undirected edge $s - t$ by two directed edges: $s \to t$ with weight $\beta^*_{s \to t}$ and $t \to s$ with weight $\beta^*_{t \to s}$. Note that in an attractive model, all $\beta^*_{s \to t} \in (0, 1)$, see Lemma 12.

It is a simple application of Dijkstra's algorithm to construct from G' a tree of all maximum weight directed paths from i to each vertex $j \neq i$, which we call $\mathcal{T}$.[6] (For our purpose we just need to know that such a tree $\mathcal{T}$ exists.)

We want to solve (11), which we write as $H_{\setminus i} D = -H_i$, where we want to solve for D, which is the vector of $D_k \ \forall k \neq i$, and H_i is the ith column of H without its ith element. Let $H^{\mathcal{T}}_{\setminus i}$ be the reduced Hessian for the model on $\mathcal{T}$ (which is missing some edges), and $H^{\mathcal{T}}_i$ be the ith column of the Hessian for the model on $\mathcal{T}$ without its ith element. In the sub-model with only the edges of $\mathcal{T}$, by construction and Theorem 8, $D^{\mathcal{T}}_k = \max_{P(i \rightsquigarrow k) \in \Pi(i \rightsquigarrow k)} W[P(i \rightsquigarrow k)]$. Hence, it is sufficient to show that adding the extra edges from $\mathcal{T}$ to G cannot decrease any D_k. This forms the remainder of the proof, where we shall require the following nonsingular M-matrix property of $H_{\setminus i}$: its inverse is elementwise non-negative (Fan, 1958, Theorem 5').

Let $\Delta = H_{\setminus i} - H^{\mathcal{T}}_{\setminus i}$ (this accounts for edges in $E(G) \setminus E(\mathcal{T})$ not incident to i), $\eta = H_i - H^{\mathcal{T}}_i$ (this accounts for edges in $E(G) \setminus E(\mathcal{T})$ incident to i) and $\delta = D - D^{\mathcal{T}}$. We must show that $\delta \geq 0$ elementwise. We have $H^{\mathcal{T}}_{\setminus i} D^{\mathcal{T}} = -H^{\mathcal{T}}_i$ and $H_{\setminus i} D = -H_i$, hence $H^{\mathcal{T}}_{\setminus i} D^{\mathcal{T}} - \eta = -H^{\mathcal{T}}_i - \eta = -H_i = H_{\setminus i} D = (H^{\mathcal{T}}_{\setminus i} + \Delta)(D^{\mathcal{T}} + \delta)$, hence $-\eta = (H^{\mathcal{T}}_{\setminus i} + \Delta)\delta + \Delta D^{\mathcal{T}} \Leftrightarrow \delta = (H_{\setminus i})^{-1}(-\eta - \Delta D^{\mathcal{T}})$. Thus, it is sufficient to show that the $(n-1)$ vector $-\eta - \Delta D^{\mathcal{T}}$ is elementwise non-negative.

Recall (5) and (9). $-\eta - \Delta D^{\mathcal{T}}$ may be written as the sum of $-\eta_e - \Delta_e D^{\mathcal{T}}$, with one η_e and Δ_e for each edge $e = (s, t)$ in $E(G) \setminus E(\mathcal{T})$. For each such edge e, we have 2 cases:

Case 1, $i \notin \{s, t\}$: $\eta_e = 0$; Δ_e has only 4 non-zero elements, at locations $(s, s), (s, t), (t, s), (t, t)$. Showing only these elements,

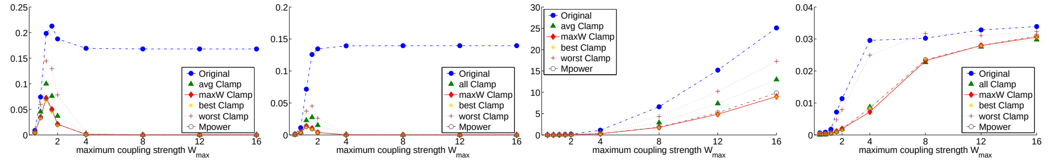
$$\Delta_e = \begin{array}{c c} & \begin{array}{cc} s & t \end{array} \\ \begin{array}{c} s \\ t \end{array} & \begin{pmatrix} -H_{st}\beta^*_{s \to t} & H_{st} \\ H_{st} & -H_{st}\beta^*_{t \to s} \end{pmatrix} \end{array} = -H_{st} \begin{array}{c c} & \begin{array}{cc} s & t \end{array} \\ \begin{array}{c} s \\ t \end{array} & \begin{pmatrix} \beta^*_{s \to t} & -1 \\ -1 & \beta^*_{t \to s} \end{pmatrix} \end{array}, \text{where } -H_{st} > 0 \text{ for an attractive edge.}$$

Hence, $-\eta_e - \Delta_e D^{\mathcal{T}}$ is 0 everwhere except element s which is $-H_{st}(D^{\mathcal{T}}_t - D^{\mathcal{T}}_s \beta^*_{s \to t})$, and element t which is $-H_{st}(D^{\mathcal{T}}_s - D^{\mathcal{T}}_t \beta^*_{t \to s})$. Observe that both expressions are ≥ 0 by construction of $\mathcal{T}$ (for example, considering the first bracketed term, observe that $D^{\mathcal{T}}_t$ is the maximum weight of a path from i to t, whereas $D^{\mathcal{T}}_s \beta^*_{s \to t}$ is the weight of a path to t going through s).

Case 2, $i \in \{s, t\}$: WLOG suppose the edge is (i, s). $-\eta_e$ is zero everywhere except element s which is $-H_{is}$ (positive). Δ_e has just one non-zero element at (s, s) which is $-H_{is}\beta^*_{s \to i}$. Hence, $-\eta_e - \Delta_e D^{\mathcal{T}}$ is 0 everwhere except element s which is $-H_{is}(1 - D^{\mathcal{T}}_s \beta^*_{s \to i}) > 0$ by Lemma 12.

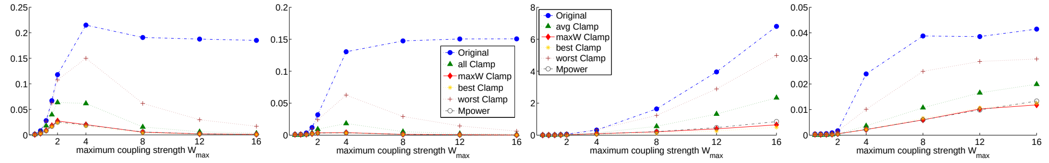
This completes the proof.

[6] We want the max of the prod of edge weights $\Leftrightarrow$ max of the log of the prod of edge weights $\Leftrightarrow$ max of the sum of the log of edge weights (all negative) $\Leftrightarrow$ min of the sum of - log of the edge weights (all positive); so really we construct the usual shortest directed paths tree using - log of the edge weights, which are all positive.



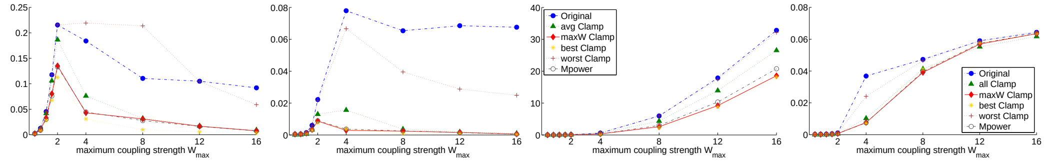
(a) attractive $\log Z$, $T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z$, $T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure 7: Average errors vs true, **complete graph on $n = 10$**. Consistent legend throughout.



(a) attractive $\log Z$, $T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z$, $T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure 8: Average errors vs true, **random graph on $n = 10, p = 0.5$**. Consistent legend throughout.



(a) attractive $\log Z$, $T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z$, $T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure 9: Average errors vs true, **random graph on $n = 50, p = 0.1$**. Consistent legend throughout.

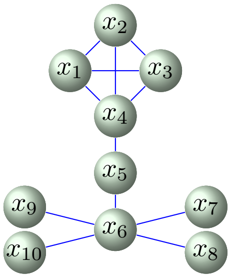


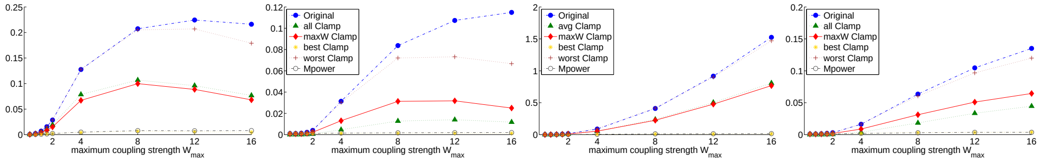
Figure 10: 'Lamp' topology. maxW is likely to choose x_6 since it has the highest degree, but x_4 is typically a better choice since it lies on cycles. Mpower can recognize this and make a better choice.

8 Additional Experiments

All of the experiments reported in §5 were also run at other settings. In particular, the earlier results show the poor performance of the standard Bethe approximation in estimating singleton marginals for attractive models with low singleton potentials, and indicate how clamping repairs this. Here, in Figures 7-9, we show results for the same topologies using the higher singleton potentials $T_{max} = 2$ for attractive models, and also show results with low singleton potentials $T_{max} = 0.1$ for general (non-attractive) models.

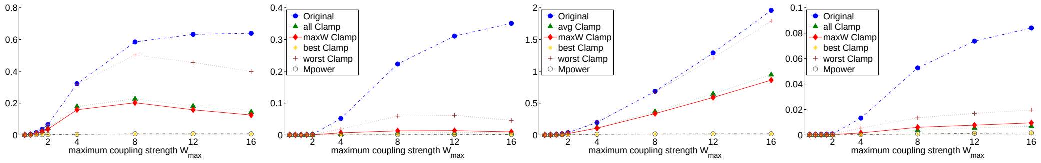
Note that in some examples of attractive models, when the 'worst clamp' variable was clamped, the resulting Bethe approximation to $\log Z$ appears to worsen (see Figure 9a), which seems to conflict with Theorem 5. The explanation is that in these examples, Frank-Wolfe is failing to find the global Bethe optimum, as was confirmed by spot checking.

Next we show results for a particular fixed topology we call a 'lamp', see Figure 10, which illustrates how maxW can sometimes select a poor variable to clamp. We explain the Mpower selection heuristic and demonstrate that it performs much better on this topology.



(a) attractive $\log Z, T_{max} = 2$ (b) attractive margs, $T_{max} = 2$ (c) general $\log Z, T_{max} = 2$ (d) general margs, $T_{max} = 2$

Figure 11: Average errors vs true, **'lamp' topology $\boldsymbol{T_{max} = 2}$**. Consistent legend throughout. Mpower performs well, significantly better than maxW.



(a) attractive $\log Z, T_{max} = 0.1$ (b) attractive margs, $T_{max} = 0.1$ (c) general $\log Z, T_{max} = 0.1$ (d) general margs, $T_{max} = 0.1$

Figure 12: Average errors vs true, **'lamp' topology $\boldsymbol{T_{max} = 0.1}$**. Consistent legend throughout. Mpower performs well, significantly better than maxW for $\log Z$.

8.1 Mpower heuristic

We would like an efficient way to select a variable to clamp which lies on many heavy simple cycles. One problem is how to define heavy. Even with a good definition, it is still NP-hard to search over all simple cycles. The idea for Mpower is as follows: assign each edge (i, j) a weight based on $|W_{ij}|$ and create a matrix M of these weights. If M is raised to the kth power, then the ith diagonal element in M^k is the sum over all paths of length k from i to i of the product of the edge weights along the path. Ideally, we might consider the sum $\sum_{k=1}^{\infty} M^k$ and use the diagonal elements to rank the vertices, choosing the one with highest total score. Recalling (12), it is sensible to assign edge weights M_{ij} based on possible $\beta^*_{i \to j}$ values. Given Lemma 12, a first idea is to use $1 - e^{-|W_{ij}|}$.

However, we'd like to be sure that the matrix series $\sum_{k=1}^{\infty} M^k$ is convergent, allowing it to be computed as $(I - M)^{-1} - I$ (since we shall be interested only in ranking the diagonal terms, in fact there is no need to subtract I at the end). Thus, we need the spectral radius $\rho(M) < 1$. A sufficient condition is that all row sums are < 1. Since each term $1 - e^{-|W_{ij}|} < 1$ and there at most $n - 1$ such elements in any row, our first heuristic was to set $M_{ij} = \frac{1}{n-1}(1 - e^{-|W_{ij}|})$. We then made two adjustments.

First, note that the series $\sum_{k=1}^{\infty} M^k$ overcounts all cycles, though at an exponentially decaying rate. It is hard to repair this. However, it also includes relatively high value terms coming from paths from i to any neighbor j and straight back again, along with all powers of these. We should like to discard all of these, hence from each ith diagonal term of $(I - M)^{-1}$, we subtract $s_i/(1 - s_i)$, where s_i is the ith diagonal term of M^2. This is very similar to the final version we used, and gives only very marginally worse results on the examples we considered.

For our final version, we observe that $1 - e^{-|W_{ij}|}$ decays rapidly, and $\approx \tanh \frac{|W_{ij}|}{2}$. Given the form of the loop series expansion for a single cycle, which contains $\tanh \frac{W_{ij}}{4}$ terms (Weller et al., 2014, Lemma 5), we tried instead using $M_{ij} = \frac{1}{n-1} \tanh \frac{|W_{ij}|}{4}$, and it is for this heuristic that results are shown in Figures 11 (for $T_{max} = 2$) and 12 (for $T_{max} = 0.1$). Observe that for this topology, Mpower performs close to optimally (almost the same results as for best Clamp), significantly outperforming maxW in most settings. Note, however, that in the experiments on random graphs reported in §5, Mpower did not outperform the simpler maxW heuristic. In future work, we hope to improve the selection methods.