Clamping Improves TRW and Mean Field Approximations

Adrian Weller*
University of Cambridge

Justin Domke*
NICTA, The Australian National University

Abstract

We examine the effect of clamping variables for approximate inference in undirected graphical models with pairwise relationships and discrete variables. For any number of variable labels, we demonstrate that clamping and summing approximate sub-partition functions can lead only to a decrease in the partition function estimate for TRW, and an increase for the naive mean field method, in each case guaranteeing an improvement in the approximation and bound. We next focus on binary variables, add the Bethe approximation to consideration and examine ways to choose good variables to clamp, introducing new methods. We show the importance of identifying highly frustrated cycles, and of checking the singleton entropy of a variable. We explore the value of our methods by empirical analysis and draw lessons to guide practitioners.

1 INTRODUCTION

Undirected graphical models, also called Markov random fields (MRFs), are a powerful and compact way to represent dependencies between variables, and have become a central tool in machine learning. A key challenge is to estimate the normalizing partition function. For example, this may be used to compute the probability of evidence, and is often a critical component of learning a model. An exact solution may be obtained via the junction tree method but unless the treewidth is bounded, this can take exponential time (Lauritzen and Spiegelhalter, 1988). Hence, many approximate methods have been developed.

We focus on three popular approaches: the *tree-reweighted* approximation (TRW, Wainwright et al., 2005); the *naïve mean field* approximation (MF); and the *Bethe* approximation, often implemented via belief propagation (BP, Pearl, 1988; Yedidia et al., 2000). In each case, we shall examine the effect on the respective partition function estimate of *clamping* one or more variables to each possible setting then combining the approximate results obtained on the clamped sub-models. See §2 for all definitions. If all variables are clamped, then the exact solution is obtained but with time exponential in the number of variables. Intuitively, as more variables are clamped, one would hope for better results, but this is not always the case and demonstrating guarantees has been challenging.

Weller and Jebara (2014b) recently proved that for an *attractive* binary pairwise model (where it is known that the Bethe partition function yields a lower bound), the optimum Bethe partition function approximation can only increase (hence improve) for each variable clamped. They also provided an example of a non-attractive model (their Figure 5c) where clamping any variable leads to a *worse* approximation. Nevertheless, they introduced two heuristics for identifying a good variable to clamp, and for both attractive and mixed models, demonstrated empirically that approximation error can sometimes be significantly reduced by clamping even one variable.

We make the following contributions. For both TRW (which yields an upper bound) and MF (which provides a lower bound), we show that for pairwise models with any number of labels, and with any types of potentials, clamping can only improve the partition function estimate by decreasing and increasing the bounds respectively. Our proofs also yield insight into the approximate marginals returned. We next examine how to select a good *sequence* of variables to clamp. Although the methods of Weller and Jebara (2014b) can perform well for choosing one variable, we show that, for some models, their methods perform poorly, particularly for selecting multiple variables. We introduce methods that strip a model to its *core*, search for strongly *frustrated cycles*, and make use of approximate singleton entropy. We provide an empirical analysis of all approaches, including a comparison against the 'greedy' choice of the best variable to clamp in hindsight after an exhaustive exploration. We conclude with observations to help guide practitioners.

*Authors contributed equally.
Appearing in Proceedings of the 19^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2016, Cadiz, Spain. JMLR: W&CP volume 41.

1.1 Related Work

The technique of branching or conditioning on variables, and approximating over the remaining variables has been explored in algorithms such as branch-and-cut (Padberg and Rinaldi, 1991; Mitchell, 2002), work on resolution versus search (Rish and Dechter, 2000) and in (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 before using belief propagation. Eaton and Ghahramani (2009) developed this further, introducing *conditioned belief propagation*. Liu et al. (2012) explored feedback message passing for inference in Gaussian (not discrete) models, deriving strong results for attractive models. Bouchard and Zoeter (2009) discuss soft-binning to split configurations into subsets then apply the mean field approximation on each but without guarantees. Choi and Darwiche (2008) examined methods to approximate the partition function by deleting edges.

2 PRELIMINARIES AND NOTATION

We consider pairwise models with n variables $X_1, \ldots, X_n$ and graph topology $(\mathcal{V}, \mathcal{E})$: $\mathcal{V}$ contains nodes $\{1, \ldots, n\}$ where i corresponds to X_i, and $\mathcal{E} \subseteq \mathcal{V} \times \mathcal{V}$ contains an edge for each pairwise relationship. Sometimes we consider multi-label models where each $X_i \in \{0, \ldots, L_i - 1\}$, and sometimes we restrict attention to binary models where $X_i \in \mathbb{B} = \{0, 1\}\ \forall i$. Let $x = (x_1, \ldots, x_n)$ be a configuration of all variables, and $\mathcal{N}(i)$ be the neighbors of i.

We consider probability distribution $p(x) = e^{-E(x)}/Z(\theta)$, where $E(x)$ is the energy of configuration x. The *partition function* $Z(\theta)$ is a quantity of fundamental interest. It requires summing over all states to yield the normalizing constant $Z(\theta) = \sum_x \exp(-E(x))$ which ensures that $\sum_x p(x) = 1$. We denote the log-partition function, sometimes called the cumulant function, by $A(\theta) = \log Z(\theta)$.

For binary models, we assume a reparameterization such that $E(x) = -\sum_{i \in \mathcal{V}} \theta_i x_i - \sum_{(i,j) \in \mathcal{E}} \frac{W_{ij}}{2} [x_i x_j + (1 - x_i)(1 - x_j)]$, with singleton potentials θ_i and edge weights W_{ij}. If $W_{ij} \geq 0$ then the edge (i, j) is *attractive* (in which case, the edge tends to pull X_i and X_j toward the same value). If $W_{ij} < 0$ then the edge is *repulsive*. If all edges of a model are attractive, then the model is called attractive, else it is *mixed*.

For any (possibly non-binary) model, we write θ for the vector of all potentials, and μ for a vector of marginals, both using the standard overcomplete exponential family representation with $E(x) = -\theta \cdot \phi(x)$, where ϕ is the vector of sufficient statistics corresponding to the model (Wainwright and Jordan, 2008), so that $\mu = \mathbb{E}_\theta[\phi(X)]$.

By considering KL divergence, standard variational methods (Wainwright and Jordan, 2008) show that $A(\theta) = \max_{\mu \in \mathbb{M}} \theta \cdot \mu + H(\mu)$, where $\mathbb{M}$, termed the *marginal polytope*, is the space of all marginal vectors μ that are consistent with a globally valid probability distribution over all configurations, and $H(\mu)$ is the entropy of the corresponding global distribution. We shall examine the following popular approximate inference methods, each of which may be defined as maximizing a negative free energy approximation over a space of marginals given by a particular polytope. We use a tilde above a symbol to indicate an approximate value, and show the method as a subscript.

Naive mean field MF $\tilde{A}_M(\theta) = \max_{\mu \in \mathbb{M}'} \theta \cdot \mu + H(\mu)$, where $\mathbb{M}'$ denotes the subspace of distributions where each variable is independent, i.e. distributions are restricted to the fully-factorized form $\prod_{i \in \mathcal{V}} \mu_i(x_i)$.

Bethe approximation $\tilde{A}_B(\theta) = \max_{\mu \in \mathbb{L}} \theta \cdot \mu + \tilde{H}_B(\mu)$, where $\mathbb{L}$ denotes the standard *local polytope* relaxation which enforces only pairwise consistency, i.e. it is required that $\mu_i(x_i) = \sum_{x_j} \mu_{ij}(x_i, x_j)\ \forall i \in \mathcal{V}, j \in \mathcal{N}(i)$, and $\tilde{H}_B(\mu)$ is the *Bethe entropy* approximation given by $\tilde{H}_B(\mu) = \sum_{i \in \mathcal{V}} H(\mu_i) - \sum_{(i,j) \in \mathcal{E}} I_{ij}(\mu_{ij})$, with the pairwise mutual information $I_{ij} = H(\mu_i) + H(\mu_j) - H(\mu_{ij}) \geq 0$.

Tree-reweighted approximation TRW $\tilde{A}_T(\theta) = \max_{\mu \in \mathbb{L}} \theta \cdot \mu + \tilde{H}_T(\mu)$, where the TRW entropy approximation (Wainwright et al., 2005) is specified by a convex sum of entropies of spanning trees, $\tilde{H}_T(\mu) = \sum_{\mathcal{T}} \rho_{\mathcal{T}} H(\mu_{\mathcal{T}})$, where, for any tree $\mathcal{T}$ and any $\mu \in \mathbb{L}$, $\mu_{\mathcal{T}}$ is the distribution that results from taking the tree-decomposition of marginals over $\mathcal{T}$. It is known that $H(\mu) \leq \tilde{H}_T(\mu)$ hence $A(\theta) \leq \tilde{A}_T(\theta)$. Further, it is easily shown that $\tilde{H}_T(\mu) = \sum_{i \in \mathcal{V}} H(\mu_i) - \sum_{(i,j) \in \mathcal{E}} c_{ij} I_{ij}(\mu_{ij})$ for edge counting numbers $\{c_{ij} \leq 1\}$. Thus, also $\tilde{H}_B(\mu) \leq \tilde{H}_T(\mu)$ and $\tilde{A}_B(\theta) \leq \tilde{A}_T(\theta)$.

Since $\mathbb{M}' \subseteq \mathbb{M}$, $\tilde{A}_M(\theta) \leq A(\theta)$. Also MF marginals are a subset of Bethe on which $\tilde{H}_B$ is exact, hence $\tilde{A}_M(\theta) \leq \tilde{A}_B(\theta)$. Collecting relationships, we have the following sandwich results (for any number of labels):

$$\tilde{A}_M(\theta) \leq \tilde{A}_B(\theta) \leq \tilde{A}_T(\theta), \text{ and } \tilde{A}_M(\theta) \leq A(\theta) \leq \tilde{A}_T(\theta). \tag{1}$$

For binary models with supermodular potentials (of any arity; in the case of pairwise models, supermodular potentials equate to an attractive model), Ruozzi (2012) proved that $\tilde{A}_B(\theta) \leq A(\theta)$, but in general, $\tilde{A}_B(\theta)$ can be above or below $A(\theta)$. $\tilde{A}_B(\theta)$ is often strikingly accurate, though there are settings where other methods are significantly better.

2.1 Clamping a Variable and Related Definitions

We are interested in sub-partition functions obtained by *clamping* some variable X_i, that is let $Z(x_i; \theta) =$

$Z(\theta)|_{X_i=x_i}$ be the sub-partition function on the model on $n-1$ variables $X_1,\dots,X_{i-1},X_{i+1},\dots,X_n$ obtained by setting $X_i = x_i$, with $x_i \in \{0,\dots,L_i-1\}$, with corresponding definitions for approximate $\tilde{Z}(x_i;\theta) = \tilde{Z}(\theta)|_{X_i=x_i}$. Observe that true values satisfy:

$$Z(\theta) = \sum_{x_i=0}^{L_i-1} Z(x_i;\theta) \text{ for any } X_i;\ p(X_i=x_i) = \frac{Z(x_i;\theta)}{Z(\theta)}.$$

These do not hold in general for approximate values but motivate the following definitions, that are achieved by clamping variable X_i and summing approximate sub-partition functions:

$$\tilde{Z}^{(i)}(\theta) = \sum_{x_i=0}^{L_i-1} \tilde{Z}(x_i;\theta); \quad \tilde{p}(x_i) = \frac{\tilde{Z}(x_i;\theta)}{\tilde{Z}^{(i)}(\theta)}. \qquad (2)$$

Correspondingly, we define $\tilde{A}^{(i)}(\theta) = \log \sum_{x_i=0}^{L_i-1} \exp \tilde{A}(x_i;\theta)$, where $\tilde{A}(x_i;\theta) = \log \tilde{Z}(x_i;\theta)$.

Considering the variational perspective above, note that $\tilde{A}(x_i;\theta) = \max_{\mu\in\mathbb{P}(x_i)} \tilde{\mathcal{F}}(\mu)$, i.e. the constrained optimization where $\mathbb{P}$ is the standard space over which the method optimizes ($\mathbb{P}=\mathbb{M}'$ for MF, $\mathbb{P}=\mathbb{L}$ for Bethe and TRW), $\mathbb{P}(x_i)$ is the sub-space constrained to $\mu(X_i=x_i)=1$, and $\tilde{\mathcal{F}}(\mu)$ is the respective negative free energy approximation being maximized.

For all our approximate methods, if *all* variables are clamped, leading to a sum over all possible configurations, then the exact partition function will be obtained but in exponential time. Weller and Jebara (2014b) showed for the Bethe approximation that: for an *attractive* binary pairwise model and any variable X_i, clamping helps in that $\tilde{A}_B(\theta) \le \tilde{A}_B^{(i)}(\theta) \le A(\theta)$; for a *mixed* model, however, clamping a variable *can lead to a worse* approximation, yet empirically it was shown often to help significantly.

We next show that for TRW and MF, clamping and summing always reduces error and improves bounds for any model type.

3 NEW RESULTS FOR THE TREE-REWEIGHTED APPROXIMATION TRW

Theorem 1. *Using definitions from §2, for any model and any variable $X_i \in \{0,\dots,L_i-1\}$,*

$$\tilde{A}_T^{(i)}(\theta) = \sum_{x_i=0}^{L_i-1} \tilde{p}_T(x_i) \max_{\mu^{x_i}\in\mathbb{L}(x_i)} \left(\theta\cdot\mu^{x_i} + \tilde{H}_T(\mu^{x_i})\right) + H(\tilde{p}_T) \qquad [\tilde{p}_T \text{ is defined in } (2)]$$
$$= \max_{\nu\in\mathbb{L}^{(i)}} \left(\theta\cdot\nu + \tilde{H}_T(\nu)\right),$$

where $\mathbb{L}^{(i)}$ denotes all convex combinations of the polytopes $\mathbb{L}(X_i=0),\dots,\mathbb{L}(X_i=L_i-1)$, i.e. $\mu\in\mathbb{L}^{(i)}$ if any only if $\mu = \sum_{x_i=0}^{L_i} r(x_i)\mu^{x_i}$ where r is a distribution and $\mu^{x_i}\in\mathbb{L}(X_i=x_i)$.

Proof. Since log-sum-exp is the convex conjugate of the negative entropy, for any a, $\log\sum_i \exp a_i = \sup_{w\in\Delta}\sum_i a_i w_i + H(w)$, where Δ is the probability simplex. Moreover, the maximizing w^* is $w_i^* = e^{a_i}/\sum_j e^{a_j}$. Applying this to $\tilde{A}_T^{(i)}$,

$$\tilde{A}_T^{(i)}(\theta) = \log\sum_{x_i}\exp\tilde{A}_T(x_i;\theta)$$
$$= \max_{w\in\Delta}\sum_{x_i} w(x_i)\left(\max_{\mu^{x_i}\in\mathbb{L}(x_i)} \theta\cdot\mu^{x_i} + \tilde{H}_T(\mu^{x_i})\right) + H(w)$$
$$= \max_{w\in\Delta}\max_{\{\mu^{x_i}\in\mathbb{L}(x_i)\}}\sum_{x_i} w(x_i)\left(\theta\cdot\mu^{x_i} + \tilde{H}_T(\mu^{x_i})\right) + H(w),$$

where Δ denotes the probability simplex for labels of X_i, and μ^{x_i} are pseudomarginal vectors that result from clamping $X_i = x_i$. The final equality above follows because a sum of maximizations over independent sets of variables is equivalent to the joint maximization of all variables over a sum of the objectives. The above form for w^* gives the first equality of the theorem (since $w^* = \tilde{p}_T$ from (2)).

Next we shall consider $\sum_{x_i} w(x_i)\tilde{H}_T(\mu^{x_i})$. For any fixed w and $\{\mu^{x_i}\}$, let $\nu = \sum_{x_i} w(x_i)\mu^{x_i}$ and observe that for any tree $\mathcal{T}$, $\nu_\mathcal{T}(x) = \sum_{x_i} w(x_i)\mu_\mathcal{T}^{x_i}(x_{-i})$. We shall also need that $\sum_{x_i} w(x_i)H(\mu_\mathcal{T}^{x_i}) = H(\nu_\mathcal{T}) - H(w)$. This follows, essentially, from $H_{\nu_\mathcal{T}}[X_{-i}|X_i] = H_{\nu_\mathcal{T}}[X] - H_{\nu_\mathcal{T}}[X_i]$ (Cover and Thomas, 1991). Now given the TRW distribution ρ over spanning trees of the model,

$$\sum_{x_i} w(x_i)\tilde{H}_T(\mu^{x_i}) = \sum_{x_i} w(x_i)\sum_\mathcal{T}\rho_\mathcal{T} H(\mu_\mathcal{T}^{x_i})$$
$$= \sum_\mathcal{T}\rho_\mathcal{T}\left(H(\nu_\mathcal{T}) - H(w)\right)$$
$$= \tilde{H}_T(\nu) - H(w), \quad \text{and hence,}$$

$\tilde{A}_T^{(i)}(\theta) = \max_{w\in\Delta}\max_{\{\mu^{x_i}\in\mathbb{L}(x_i)\}}\left(\theta\cdot\nu + \tilde{H}_T(\nu)\right)$. This is equivalent to the stated result, using the dependence of ν on w and μ^{x_i}. □

Observe that the result of clamping a single variable with TRW is precisely to tighten the local polytope from $\mathbb{L}$ to $\mathbb{L}^{(i)}$. As an immediate corollary, $\tilde{A}_T^{(i)}(\theta) \le \tilde{A}_T(\theta)$. Further, $\tilde{A}_T^{(i)}(\theta) = \log\sum_{x_i}\tilde{Z}_T(x_i;\theta) \ge \log\sum_{x_i} Z(x_i;\theta) = \log Z(\theta) = A(\theta)$. Hence, we have shown the following.

Theorem 2. *For any discrete model (any number of labels, any types of potentials) and any variable X_i, $A(\theta) \le \tilde{A}_T^{(i)}(\theta) \le \tilde{A}_T(\theta)$.*

Note that all the analysis above assumes that the distribution ρ over trees $\mathcal{T}$ used by TRW is constant. However, when a variable X_i is clamped, its edges may be removed from the graph, which can only further decrease the bound. To see this, construct a new distribution ρ' over subgraphs $\{\mathcal{U}\}$ as follows: for each original tree $\mathcal{T}$, let $\mathcal{R}(\mathcal{T})$ be $\mathcal{T}$ less its edge(s) to X_i, and let $\rho'_{\mathcal{U}} = \sum_{\mathcal{T}:\mathcal{U}=\mathcal{R}(\mathcal{T})} \rho_{\mathcal{T}}$. In the clamped model, if $\mathcal{U} = \mathcal{R}(\mathcal{T})$, then $H(\mu^{x_i}_{\mathcal{U}}) = H(\mu^{x_i}_{\mathcal{T}})\ \forall \mu^{x_i} \in \mathbb{L}(x_i)$. If $\mathcal{U}$ is disconnected, new edges (not incident to X_i) may be added to it to make a tree, and this can only reduce $H(\mu^{x_i}_{\mathcal{U}})$ (Wainwright and Jordan, 2008). In addition, tree weights may be reoptimized to reduce the bound still further.

4 NEW RESULTS FOR THE NAIVE MEAN FIELD APPROXIMATION MF

In this Section, we consider the mean field negative free energy approximation,

$$\tilde{\mathcal{F}}_M(\mu) = \sum_{i\in\mathcal{V}}\sum_{x_i}\theta_i(x_i)\mu_i(x_i) + \sum_{(i,j)\in\mathcal{E}}\sum_{x_i,x_j}\theta_{ij}(x_i,x_j)\mu_i(x_i)\mu_j(x_j) + \sum_{i\in\mathcal{V}} H(\mu_i), \tag{3}$$

where $\mu \in \mathbb{M}'$ is a fully-factorized distribution. Given θ, let $\hat{\mu}$ be any local maximizer of $\tilde{\mathcal{F}}_M(\mu)$ over $\mathbb{M}'$. Define $\hat{\mu}^{x_i}(X) = \hat{\mu}(X_{-i})\mathbb{I}[X_i = x_i]$, where $\mathbb{I}$ is the standard indicator function. We first show a key Lemma, from which the main result in Theorem 4 easily follows.

Lemma 3. $\tilde{\mathcal{F}}_M(\hat{\mu}) = \log\sum_{x_i}\exp\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i})$.

Proof. Write that $\log\sum_{x_i}\exp\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i}) = \sum_{x_i} w(x_i)\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i}) + H(w)$, where $w(x_i) = \exp\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i})/\sum_{x'_i}\exp\tilde{\mathcal{F}}_M(\hat{\mu}^{x'_i})$. Hence, $w(x_i) \propto \exp\left[\theta_i(x_i) + \sum_j\sum_{x_j}\theta_{ij}(x_i,x_j)\hat{\mu}_j(x_j)\right]$. This is precisely the mean-field update if $\tilde{\mathcal{F}}_M$ is maximized with respect to μ_i while holding all other marginals fixed (Wainwright and Jordan, 2008, §5.3). Hence, $w(x_i) = \hat{\mu}_i(x_i)$, since $\hat{\mu}$ is a local maximizer, which is unique when $\{\hat{\mu}_j, j \neq i\}$ are fixed. It thus follows that

$$\log\sum_{x_i}\exp\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i}) = \sum_{x_i}\hat{\mu}_i(x_i)\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i}) + H(\hat{\mu}_i).$$

Now consider (3) and observe that $\sum_{x_i}\hat{\mu}_i(x_i)\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i}) = \tilde{\mathcal{F}}_M(\hat{\mu}) - H(\hat{\mu}_i)$, hence the result follows. □

Theorem 4. *For any model (any number of labels) and any variable X_i, $\tilde{A}_M(\theta) \leq \tilde{A}^{(i)}_M(\theta) \leq A(\theta)$.*

Proof. Using Lemma 3, we have

$$\tilde{A}_M(\theta) = \tilde{\mathcal{F}}_M(\hat{\mu}) = \log\sum_{x_i}\exp\tilde{\mathcal{F}}_M(\hat{\mu}^{x_i}) \leq \log\sum_{x_i}\exp\max_{\mu\in\mathbb{M}'(x_i)}\tilde{\mathcal{F}}_M(\mu) = \tilde{A}^{(i)}_M(\theta),$$

where $\mathbb{M}'(x_i)$ is the constrained sub-space defined in §2.1.

For the other inequality, $\tilde{A}^{(i)}_M(\theta) = \log\sum_{x_i}\tilde{Z}_M(x_i;\theta) \leq \log\sum_{x_i} Z(x_i;\theta) = \log Z(\theta) = A(\theta)$. □

In practice, there may be locally optimum solutions for the clamped problems that show worse performance than the parent. However, the analysis above shows that this concern is guaranteed to be avoided if the clamped optimizations are initialized at the solution of the parent problem.

5 METHODS TO SELECT WHICH VARIABLES TO CLAMP

Henceforth, we focus on binary pairwise models. As shown above, clamping a variable and summing over approximate sub-partition functions will always reduce error and improve bounds for both MF and TRW. Further, Weller and Jebara (2014b) proved that for the Bethe approximation, this is also true for attractive models, and empirically it is often helpful for mixed models.

This leads to the question of how to choose which variable, or sequence of variables, to clamp. Weller and Jebara (2014b) introduced two selection heuristics, motivated by trying to break *strong* cycles (we say a subgraph is strong if all its edges have weights with high absolute value; since Bethe and TRW are exact on trees, this goal is reasonable), and demonstrated that these heuristics were effective in several contexts. We first describe these earlier approaches.

maxW is a simple $O(|\mathcal{E}|)$ method which picks a variable X_i with $\max_{i\in\mathcal{V}}\sum_{j\in\mathcal{N}(i)}|W_{ij}|$. 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* was introduced as a more complex approach to attempt to avoid this problem (we introduce a simpler method in §5.1) by considering the convergent series of powers of a modified $|W|$ matrix, which approximates a weighted count around all cycles. It was shown that Mpower outperforms maxW in some cases, though for many examples, their performance was similar.

Note that both maxW and Mpower rely exclusively on the absolute value of edge weights $|W_{ij}|$, while ignoring their signs, and also ignoring singleton potentials. In the remainder of this Section, we demonstrate how these earlier methods can perform poorly in certain circumstances, and introduce new approaches. Details of all selection methods are provided in the Appendix.

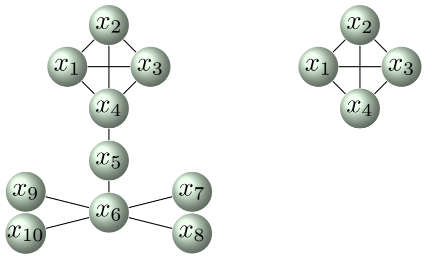


Figure 1: On the left, a model with the 'lamp' topology from Weller and Jebara (2014b). On the right, the *core* of the same model. This is obtained by iteratively removing variables with degree 1. If maxW is applied to the original model, it often chooses to clamp X_6 (which has highest degree) whereas any of $X_1, \ldots, X_4$ would be better, and will be selected if the model is first stripped to its core. See §5.1.

5.1 Stripping to the Core

Following Sudderth et al. (2007), we define the *core* of a graph (model), to be what remains after iteratively pruning nodes (variables) with degree 1. Equivalently, it is the subgraph (submodel) induced on the nodes (variables) which either belong to some cycle, or lie on a path between cycles. An example is shown in Figure 1. For pairwise entropy approximations, we should expect it will be better to strip a model to its core before applying any method to select a variable (then clamping it in the original model). In many cases that were previously challenging for maxW, this quick pre-processing step enables maxW to perform as well as the more expensive Mpower method.

5.2 Balanced Models, Frustrated Cycles and Strong Cycles

A *frustrated cycle* is a cycle with an odd number of repulsive edges (see §2 for definitions). These cause difficulties for many methods of inference. A *balanced* model (or submodel) is one that does not contain any frustrated cycles. It is easily shown (Harary, 1953; Weller, 2015) that a model may be mapped into an equivalent attractive model by flipping an appropriately chosen subset of variables iff it is balanced (in which case such a set can be identified in linear time). Hence, results for attractive models readily extend to the broader class of balanced models.

Pairwise approximations such as Bethe and TRW are exact on models without any cycles. Further, it is known that for Bethe and TRW, frustrated cycles can lead to more trouble than balanced cycles (Weller, 2015, §6.3). This is illustrated in the top row of Figure 2, which shows approximation error for symmetric models (i.e. no singleton potentials) with uniform edge weights. As edge weights rise, both MF and Bethe underestimate with error that is bounded and tends to $\log 2$,[1] while TRW tends to the correct solution. For strong negative weights, however, which lead to frustrated cycles, Bethe and TRW show rapidly increasing error without bound. Note that the model on the C_4 cycle is balanced even with negative edge weights (since there are an even number of edges), with symmetric error either side of 0 edge weights.[2] The observation that Bethe and TRW can perform arbitrarily badly for strong frustrated cycles, whereas the error for MF is bounded (recall the MF solution lies in $\mathbb{M}$), explains our experimental results where MF outperforms Bethe, see §6.1.

This motivates trying to identify strong frustrated cycles. Both the maxW and Mpower earlier methods of Weller and Jebara (2014b) consider only $|W_{ij}|$, hence are unable to differentiate between balanced and frustrated cycles. To find strong frustrated cycles is NP-hard but we introduce heuristics that build on a recent algorithm by Sontag et al. (2012), which was used in a cutting plane approach to tighten the local polytope for MAP inference. We combine ideas from their algorithm with cycle scores based on the loop series method (Chertkov and Chernyak, 2006; Sudderth et al., 2007; Weller et al., 2014) and present two new heuristics for identifying a good variable to clamp: *frustCycles*, which seeks to identify a variable lying on strong frustrated cycles, which if clamped, would remove those cycles; and *strongCycles*, which attempts also to take into due consideration the (lower) value of removing strong balanced cycles. Details are provided in the Appendix.

5.3 Using Singleton Entropies

All previous clamping selection methods examine only edge weights. However, if a variable already has very low singleton entropy (typically this would be due to a strong singleton potential: this might be present in the original model, or could have arisen as a result of earlier clamping rounds), then it has effectively already been held fixed to one value, and there is little to be gained from clamping it. On the other hand, if a variable has high entropy and is strongly connected to many others, without a frustrated cycle, then clamping it can effectively lead to a cascade where many other variables will also be 'effectively clamped', yielding a significant improvement in approximation error. Afterward, there is little residual value in actually clamping those other variables.

This effect is illustrated by comparing the rows of Figure 2. Observe that even in the K_5 fully connected model, when no frustrated cycle is present (i.e. when edge weights are positive), just one clamping is sufficient to obtain almost zero error. A further illustration is provided by considering

[1] For high positive weights: with MF, all singleton marginals are pulled toward 0 or 1; with Bethe, for K_4 and K_5 the same happens, for cycles all edge marginals approach $(1/2\ 0; 0\ 1/2)$; in all cases, $\tilde{H}(\mu') \to 0$, whereas the true distribution has two dominating states (all 0s or all 1s), hence $H(\mu) \to \log 2$.

[2] With $W < 0$, the models with fully connected topologies K_n for $n > 3$ contain both balanced and frustrated cycles but the number and strength of the frustrated cycles dominate.

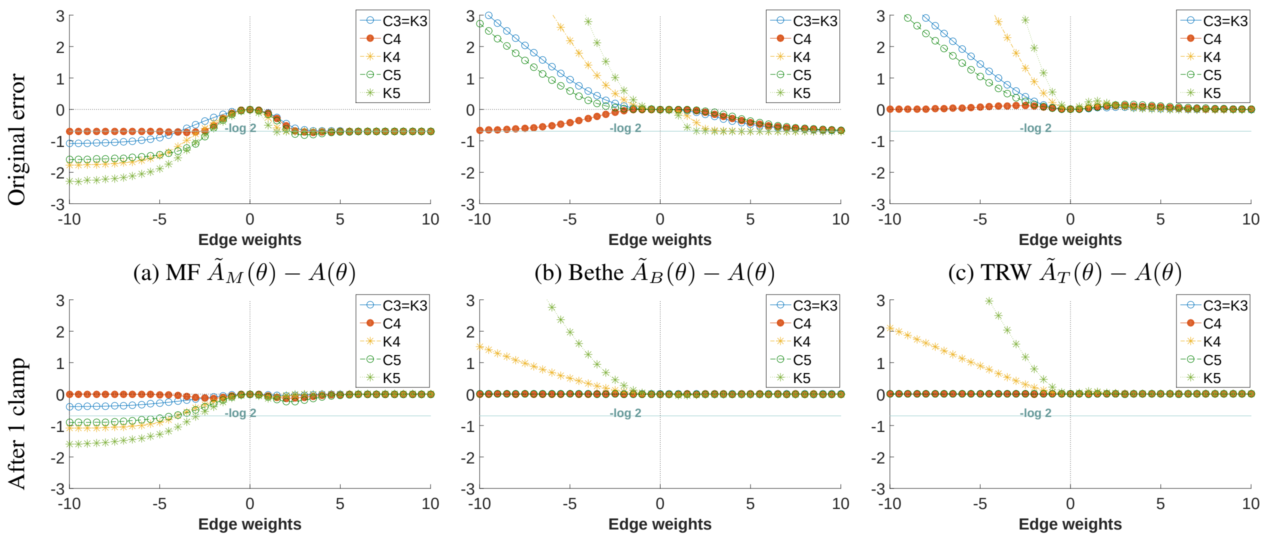


Figure 2: Top: Approximate log-partition function minus true value, i.e. $\tilde{A}(\theta) - A(\theta)$, for symmetric models (no singleton potentials) on 3, 4 and 5 variables with cycle C_n and complete graph K_n topologies, with uniform edge weights that are varied, for (a) MF, (b) Bethe, and (c) TRW. Bottom: Error $\tilde{A}^{(i)}(\theta) - A(\theta)$ for the same models and methods after clamping (any) one variable and summing approximate sub-partition functions. Observe, before clamping: Even strong positive weights lead to underestimates with bounded error; while strong negative weights with frustrated cycles lead to unbounded overestimates for Bethe and TRW. After clamping: All methods are significantly improved; if there are no frustrated cycles remaining, all methods are almost exact. See discussion in §5.2.

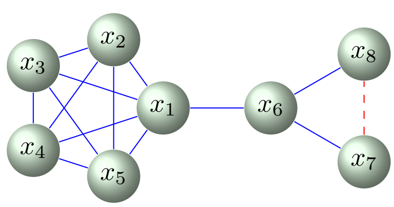


Figure 3: Example of a model where the earlier maxW and Mpower heuristics of Weller and Jebara (2014b) perform poorly to select variables to clamp but our new methods perform well. Solid blue (dashed red) edges are strongly attractive (repulsive). maxW and Mpower both repeatedly select variables in $X_1, \ldots, X_5$: the first clamp is good but less so than picking from the frustrated cycle X_6, X_7, X_8; then repeat clampings in $X_1, \ldots, X_5$ reap little benefit. It is best to pick X_6 then X_1.

the model in Figure 3. Recognizing this effect, ideally we would compute singleton entropies by exact inference, but that would clearly be too costly, hence, we use approximate inference. Specifically, we introduce TRE versions of each earlier method: for each variable, multiply its respective earlier heuristic clamp score by its TRw Entropy (we want both high) and choose the best. We use the TRW approximate entropy for two reasons (Weller et al., 2014): (i) TRW singleton marginals typically have similarly good accuracy to Bethe, while often being easier to estimate (since the TRW free energy is convex); and (ii) we are particularly interested in cases where edge potentials are high around a variable, and in this setting, Bethe marginals can be poor, being pulled toward 0 or 1 even if the true marginal is close to 1/2. TRE versions of all heuristics perform well for multiple clampings on models such as the one in Figure 3.

6 EXPERIMENTS

We tested all approaches on 100 runs each of various randomly generated binary pairwise models, exploring up to 5 clampings. We used all topologies shown in Figure 4, with the following parameters: all used random singleton potentials $\theta_i \sim U[-2, 2]$; attractive models had $W_{ij} \sim U[0, 2]$ (typically a difficult lower intermediate value for Bethe) or $W_{ij} \sim U[0, 6]$; mixed models had $W_{ij} \sim U[-6, 6]$ or $W_{ij} \sim U[-12, 12]$. Exact values were computed using the junction tree algorithm. All inference methods were implemented using the standard open source libDAI library (Mooij, 2010). In addition, we performed experiments on complete graphs with fewer variables but a similar number of edges, see Figure 5, and on random 4-regular graphs. Figure 5(d) shows typical timings vs. performance. Full details and results are provided in the Appendix.

We implemented all variable selection methods, specifically: maxW, Mpower, frustCycles and strongCycles, always first stripping to the core. We also tried the original maxW without stripping (as a comparison, which was previously shown to perform well). In addition, we used TRE versions of all these, for a total of 10 heuristics. We also implemented a greedy search over all possible clampings up to 3, to see how this would perform compared to our heuristics, and we implemented pseudo-greedy, which tried only the 10 heurstics in our basket and picked the best performer. This best performer was determined for MF (TRW) by the highest (lowest) solution. Similarly for Bethe for attractive models, best performer meant the variable which led to the

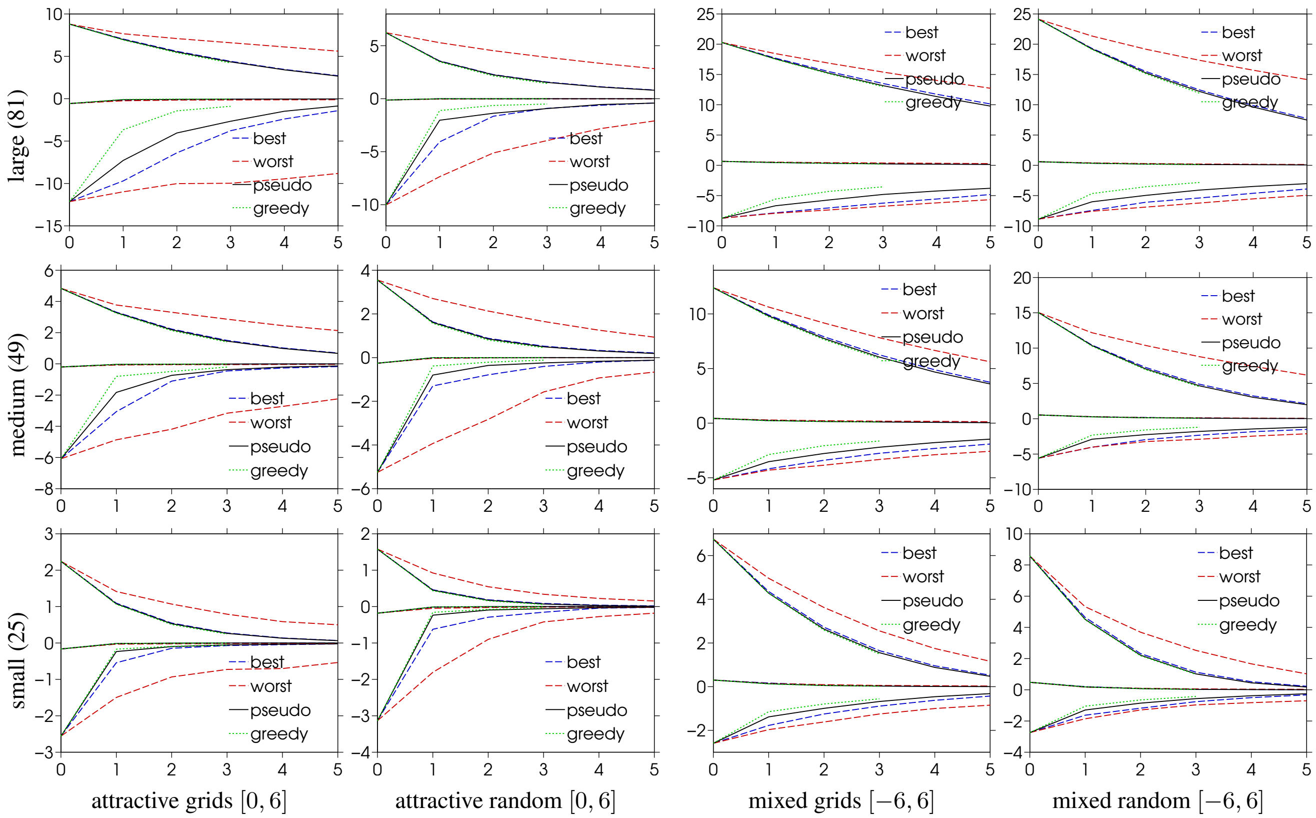
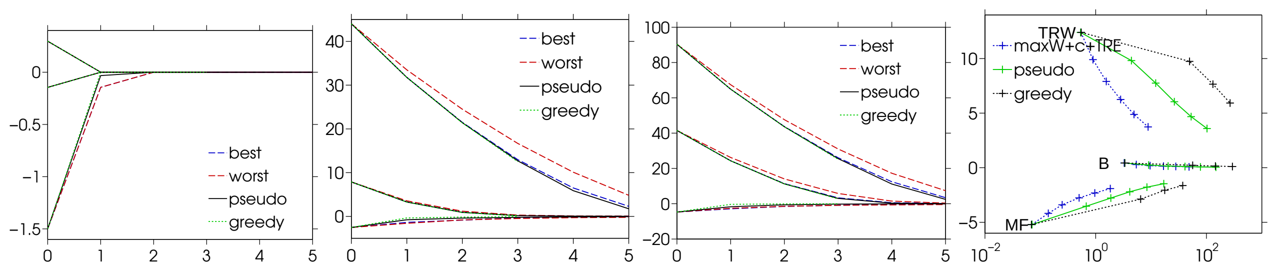


Figure 4: Average error plots $\tilde{A}(\theta) - A(\theta)$ for TRW, Bethe and MF over 100 runs. Each plot shows a particular model type and size, arranged by model type (horizontally) and model size (vertically). Within each plot, to aid comparison, we show: top curves for TRW, middle curves for Bethe, and bottom curves for MF, as justified by equation (1); in each case, for 0 to 5 clampings. Grids are toroidal 5x5, 7x7 and 9x9. Random models are Erdös-Renyi with the same number of variables, edge probability s.t. avg degree is 4 to match grids. All models shown have $\theta_i \sim U[-2, 2]$, with: attractive $W_{ij} \sim [0, 6]$, or mixed $W_{ij} \sim [-6, 6]$. Error shown for Bethe on mixed models is $|\tilde{A}(\theta) - A(\theta)|$. *best* and *worst* curves indicate the best and worst of our 10 selection heuristics, run from the start up to that clamp point.



(a) attractive K_{15}, $[0, 6]$ (b) mixed K_{15}, $[-6, 6]$ (c) mixed K_{15}, $[-12, 12]$ (d) time (log axis) vs. error 7x7 grid, $W_{ij} \sim U[-6, 6]$

Figure 5: Left three plots (a)-(c) show average error $\tilde{A}(\theta) - A(\theta)$ for TRW, Bethe and MF over 100 runs; each plot shows a model on a complete graph topology on 15 variables using $\theta_i \sim U[-2, 2]$ and edge weights drawn uniformly from the respective range shown. Within each plot, to aid comparison, we show: top curves for TRW, middle curves for Bethe, and bottom curves for MF, as justified by equation (1); in each case, for 0 to 5 clampings. Error shown for Bethe on mixed models is $|\tilde{A}(\theta) - A(\theta)|$. *best* and *worst* curves indicate the best and worst of our 10 selection heuristics, run from the start up to that clamp point.

As shown in (c), when a mixed model is highly connected with strong edge weights, *MF can be much more accurate than Bethe*. We believe this is because Bethe (and TRW) can return arbitrarily high error for strong frustrated cycles, see §5.2.

Right (d) shows a typical plot of runtime (secs) vs. error for our sparse models, here showing average results for a 7x7 grid with mixed edge weights $W_{ij} \sim U[-6, 6]$. 'B' indicates Bethe. Time plots for models with dense edges, such as those on a complete graph, look quite different, with MF performing significantly better.

Full details and results are provided in the Appendix.

greatest increase. For pseudo-greedy for Bethe on mixed models, where there is no way to be sure which is best, we tried various options, settling on picking the variable that gave the best improvement (i.e. the biggest fall) in TRW.

6.1 Discussion

Looking across all results (Figures 4 and 5, with more details in the Appendix), we make the following observations.

Bethe typically dominates for accuracy as an inference method, as has been previously observed. However, as mixed models become more densely interconnected with strong edges, MF becomes competitive, and can even be far superior to Bethe, e.g. see Figure 5(c). We believe this is because Bethe (and TRW) can return arbitrarily high error for strong frustrated cycles, see §5.2. Figure 18 in the Appendix shows a histogram of Bethe errors on mixed models.

Clamping improves accuracy significantly, particularly when models have many strong edge weights. The improvement is greater for random models than for those with fixed degree; this is likely because some high degree variables will be present which will be good to clamp. Our heuristics perform well. Pseudo-greedy (which takes the best of our set of heuristics at each stage) performs almost identically to true greedy (which tries all possible clampings), except for MF. There, we believe that part of the effect is due to the highly non-convex optimization, so that true greedy effectively gets the benefit of many random initializations. There is clear value to probing with our portfolio of heuristics and picking the best, since no one method dominated. The best single performer was maxW after being augmented with our core and TRE updates (these augmentations were particularly helpful after the first clamping), which was best only about half the time (Figure 19 in the Appendix). See the Appendix for more details, including error plots zoomed in around the Bethe results.

Runtime varies significantly, see Figure 5(d) for a typical example. Considering clamping approaches, greedy takes the longest time though yields little benefit over pseudo-greedy. maxW, augmented with our core and TRE updates, is fast and yields the best time-adjusted results. See Appendix for all timings, and note that sometimes, clamping makes the subsequent optimization problems easier to solve, hence the total time with clamping is occasionally *lower* than without, while also being significantly more accurate. All branches over multiple clampings can be parallelized, as clearly can (pseudo-)greedy approaches. As an inference method, MF runs the fastest but this could be influenced by our implementation, with all timings sensitive to parameters. In order to get TRW to converge, we used damping which significantly slows it down, though there may be faster convergent methods. Further, edge weights could be optimized which is not implemented in libDAI. For Bethe, we used the HAK double-loop algorithm (Heskes et al., 2003), which was needed to ensure convergence.

Additional discussion on greedily selecting which next variable to clamp is provided in §10 of the Appendix.

7 CONCLUSION

TRW and MF are important and widely-used methods of approximate inference, yielding useful upper and lower bounds on the true partition function. For both methods, we have derived guarantees on the beneficial effect of clamping any variable and summing approximate sub-partition functions, for any model type (attractive or not) and any number of discrete labels. Such guarantees have been difficult to obtain, and do not apply in general for Bethe, with the only prior result to our knowledge being that of Weller and Jebara (2014b) for Bethe, only for the restricted case of attractive binary pairwise models. By clamping TRW or MF, this leads directly to useful improved upper or lower bounds on the true partition function, and also helpfully on the optimum Bethe approximation. Further, our derivation provides a surprising interpretation in terms of a tightening of the local polytope relaxation ($\mathbb{L}^{(i)}$ in Theorem 1).

Earlier approaches to selecting a variable to clamp can perform poorly in some settings. We examined when this is likely to occur, and introduced new methods based on first stripping to the core, looking for heavy (frustrated) cycles, and using singleton entropies. These new methods empirically yielded significant benefits.

Based on an experimental comparison across the different inference approaches and clamping selection methods, including examining the accuracy improvement vs. time tradeoff, we are able to suggest the following practical recommendations:

- As has been previously observed, typically Bethe is the best approach, provided convergence difficulties do not arise. However, perhaps surprisingly, for densely connected mixed models with strong edges, MF can be much more accurate (e.g. Figure 5(c)).
- Clamping can be very helpful, more so for denser models with stronger edge weights, a setting where inference on the original model is hard.
- For variable selection, if speed is critical, use just the updated maxW heuristic (augmented with core and TRE). Otherwise, use our basket of approaches and pick the pseudo-greedy best option. For Bethe on mixed models, use TRW to guide pseudo-greedy selection.
- In many cases, it will be helpful to run MF and TRW in order to obtain guaranteed bounds on the true partition function. If a Bethe method is used, bounds can also be useful to check if a poor local optimum was returned (below the MF value).

Acknowledgements

We thank the anonymous referees for helpful comments. NICTA is funded by the Australian Government through the Department of Communications and the Australian Research Council through the ICT Centre of Excellence Program.

References

G. Bouchard and O. Zoeter. Split variational inference. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 57–64. ACM, 2009.

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

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

T. Cover and J. Thomas. *Elements of Information Theory*. John Wiley & Sons, 1991.

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.

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

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

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.

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

J. Mooij. libDAI: A free and open source C++ library for discrete approximate inference in graphical models. *Journal of Machine Learning Research*, 11:2169–2173, August 2010. URL `http://www.jmlr.org/papers/volume11/mooij10a/mooij10a.pdf`.

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.

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. Sontag, D. Choe, and Y. Li. Efficiently searching for frustrated cycles in MAP inference. In *UAI*, 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. Bethe and related pairwise entropy approximations. In *Uncertainty in Artificial Intelligence (UAI)*, 2015.

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.

J. Yedidia, W. Freeman, and Y. Weiss. Generalized belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, 2000.