A Lower Bound on the Partition Function of Attractive Graphical Models in the Continuous Case

Nicholas Ruozzi
University of Texas at Dallas

Abstract

Computing the partition function of an arbitrary graphical model is generally intractable. As a result, approximate inference techniques such as loopy belief propagation and expectation propagation are used to compute an approximation to the true partition function. However, due to general issues of intractability in the continuous case, our understanding of these approximations is relatively limited. In particular, a number of theoretical results known for these approximations in the discrete case are missing in the continuous case. In this work, we use graph covers to extend several such results from the discrete case to the continuous case. Specifically, we provide a graph cover based upper bound for continuous graphical models, and we use this characterization (along with a continuous analog of a discrete correlation-type inequality) to show that the Bethe partition function also provides a lower bound on the true partition function of attractive graphical models in the continuous case.

1 INTRODUCTION

Graphical models represent the factorization of a joint probability distribution over a hypergraph. The graph together with the factorization are then used to perform a variety of, either approximate or exact, inference tasks for prediction and learning (e.g., computing marginals and the partition function). Graphical models over discrete state spaces owe much of their popularity to simple approximate inference algorithms such as loopy belief propagation (BP) that are easy to implement (often as message-passing algorithms) and tend to provide reasonable approximations in practice. However, when the state space of the graphical model is continuous, loopy BP is much harder to apply: the algorithm requires computing potentially high dimensional integrals and even representing the messages that it passes becomes a non-trivial task.

A number of algorithms have been designed for approximate inference in continuous graphical models, many of which attempt to address one or more of the shortcomings of loopy BP: particle belief propagation (Ihler and McAllester, 2009), kernel belief propagation (Song et al., 2011), quantized stochastic belief propagation (Noorshams and Wainwright, 2013), expectation propagation (EP) (Minka, 2001), adaptive discretization (Isard et al., 2008), EPBP (Lienart et al., 2015), and many more. Irrespective of the approximate inference scheme that is used, we would like to understand the relationship between the true partition function and the approximate partition function generated by our scheme. This turns out to be surprisingly challenging, with the exception of simple methods/models (e.g., Gaussian graphical models, naïve mean field, etc.).

In this work, we take a few steps towards a better understanding of approximate variational inference in the continuous case. Here, we focus on BP and the closely related EP algorithms as they can both be viewed as algorithms to compute local optima of the Bethe free energy over a set of constraints. We provide two main theorems. The first demonstrates that the so-called Bethe partition function can be upper bounded via graph covers. A related result in the discrete case was previously demonstrated by Vontobel (2013), but additional effort is required to handle the continuous case. Second, we show that the Bethe partition function always lower bounds the true partition function for continuous, attractive graphical models, another result that was previously known only in the discrete setting (Ruozzi, 2012, 2013). Discrete, attractive graphical models have been used successfully in a variety computer vision applications, the theoretical

Preliminary work. Under review by AISTATS 2017. Do not distribute.

results presented here suggest that their continuous analogs could open up a new world of possible applications if efficient algorithms could be designed for continuous, attractive graphical models. We conclude with a discussion of the implications of the proposed theory.

2 PREREQUISITES

Let $f : \mathbb{R}^n \to \mathbb{R}_{\geq\epsilon}$ be a strictly positive function where $\mathbb{R}$ is the set of possible assignments of each variable and $\mathbb{R}_{\geq\epsilon}$ is the set of all real numbers larger than some $\epsilon > 0$. A function f factorizes with respect to a hypergraph $G = (V, \mathcal{A})$, if there exist potential functions $f_i : \mathbb{R} \to \mathbb{R}_{\geq\epsilon}$ for each $i \in V$ and $f_\alpha : \mathbb{R}^{|\alpha|} \to \mathbb{R}_{\geq\epsilon}$ for each $\alpha \in \mathcal{A}$ such that

$$f(x_1, \ldots, x_n) = \frac{1}{Z} \prod_{i \in V} f_i(x_i) \prod_{\alpha \in \mathcal{A}} f_\alpha(x_\alpha),$$

where the normalization constant, $Z \in \mathbb{R}_{\geq 0}$, ensures that f defines a probability distribution. The hypergraph G together with the potential functions $f_{i \in V}$ and $f_{\alpha \in \mathcal{A}}$ define a graphical model.

A typical inference task is to compute the marginals and/or the normalization constant, often called the partition function, of the graphical model.

$$Z = \int_{\mathbb{R}^n} f(x_1, \ldots, x_n) dx_1, \ldots, dx_n$$

The computation of the partition function is challenging for several reasons. First, there are significant numerical issues that need to be addressed when computing/approximating high-dimensional integrals. Second, computing the partition function of general discrete graphical models (which are a special case of this formulation) is a #P-hard problem. Third, the integral itself may not exist. In this paper, we will assume that the partition function is a positive number and that all potential functions are continuous almost everywhere and bounded over any finite interval (or product of intervals) of their domain[1].

2.1 The Bethe Free Energy

Because of the computational challenges associated with computing the exact partition function, approximate inference techniques are often employed in practice. One popular approach, given its relationship to loopy belief propagation, is to use the Bethe partition function as a surrogate for the true partition function.

[1]Here we work with Riemann integration. All of the results described herein can be extended to the more general case of product measures over $\mathbb{R}^n$.

The Bethe partition function is defined via an optimization problem over the Bethe free energy.

$$\log F_{\mathrm{B}}(G, \tau) \triangleq U(G, \tau) - H(G, \tau)$$

where U is the energy,

$$U(G, \tau) = -\sum_{i \in V} \int_{\mathbb{R}} \tau_i(x_i) \log f_i(x_i) dx_i - \sum_{\alpha \in \mathcal{A}} \int_{\mathbb{R}^{|\alpha|}} \tau_\alpha(x_\alpha) \log f_\alpha(x_\alpha) dx_\alpha,$$

and H is an approximation of the differential entropy,

$$H(G, \tau) = -\sum_{i \in V} \int_{\mathbb{R}} \tau_i(x_i) \log \tau_i(x_i) dx_i - \sum_{\alpha \in \mathcal{A}} \int_{\mathbb{R}^{|\alpha|}} \tau_\alpha(x_\alpha) \log \frac{\tau_\alpha(x_\alpha)}{\prod_{k \in \alpha} \tau_k(x_k)} dx_\alpha.$$

The Bethe free energy is evaluated over a collection of so-called pseudomarginals that satisfy a set of local marginalization constraints, which define the local marginal polytope, $\mathcal{M}_L$.

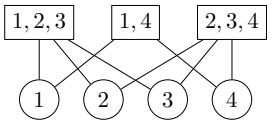
$$\mathcal{M}_L = \left\{ \tau : \begin{array}{l} \tau_i : \mathbb{R} \to [0,1], \tau_\alpha : \mathbb{R}^{|\alpha|} \to [0,1] \\ \text{are probability densities and} \\ \forall \alpha \in \mathcal{A}, i \in \alpha, x_i \in \mathcal{X}, \\ \int_{\mathbb{R}^{|\alpha|}} \tau_\alpha(x_\alpha) dx_{\alpha \setminus i} = \tau_i(x_i) \end{array} \right\}$$

The log-Bethe partition function is then determined by the minimum of $F_{\mathrm{B}}(G, \tau)$ over all $\tau \in \mathcal{M}_L$.

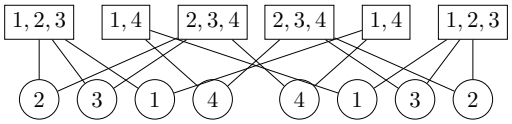
$$\log Z_{\mathrm{B}}(G) = -\min_{\tau \in \mathcal{M}_L} F_{\mathrm{B}}(G, \tau)$$

To estimate Z_{B} in the discrete case, one typically runs loopy belief propagation until a fixed point is reached. As fixed points of loopy belief propagation correspond to local optima of F_{B}, the attained fixed point must yield a lower bound on Z_{B} (Yedidia et al., 2005).

Similarly, the expectation propagation algorithm can be viewed as finding local optima of F_{B} over a weaker constraint set $\mathcal{M}_{\mathrm{EP}} \supseteq \mathcal{M}_L$, which replaces the marginalization constraint with one that only requires the approximate distributions described by τ to agree in expectation (Heskes and Zoeter, 2002). As a result, the corresponding approximation to the partition function must satisfy $Z_{\mathrm{EP}}(G) \geq Z_{\mathrm{B}}(G)$. However, in order to make the EP algorithm tractable in practice, the allowable pseudomarginals are often restricted to be from a "nice" family of distributions. In this case, the above inequality only holds if we similarly restrict the marginal polytope and compute the Bethe partition function over this restricted set.



(a) A hypergraph graph, G.



(b) One possible 2-cover of G.

Figure 1: An example of a graph cover of a factor graph. The nodes in the cover are labeled for the node that they copy in the base graph.

2.2 Graph covers

The primary theoretical tool that will enable the main results of this paper depends heavily on the notion of graph covers (sometimes called lifts of graphs).

Definition 2.1. *A graph H **covers** a graph $G = (V, E)$ if there exists a graph homomorphism $h : H \to G$ such that for all vertices $i \in G$ and all $j \in h^{-1}(i)$, h maps the neighborhood ∂j of j in H bijectively to the neighborhood ∂i of i in G.*

This definition can be extended to hypergraphs as well by using the factor graph formulation of a hypergraph. That is, each hypergraph G can be expressed as a standard graph as follows. Create a node in the factor graph representation for each vertex (called variable nodes) and each hyperedge (called factor nodes) of G. Each factor node is connected via an edge in the factor graph to the variable nodes on which the corresponding hyperedge depends. A hypergraph H is said to be an M-cover of G if every vertex and every hyperedge of G has exactly M copies in H. See Figure 1 for an example of this construction.

To any M-cover $H = (V^H, \mathcal{A}^H)$ of G given by the homomorphism h, we associate a collection of potentials as defined by the homomorhpism h: the potential at node $i \in V^H$ is equal to $f_{h(i)}$, the potential at node $h(i) \in G$, and for each $\beta \in \mathcal{A}^H$, we associate the potential $f_{h(\beta)}$. In this way, we can construct a function $f^H : \mathbb{R}^{M|V|} \to \mathbb{R}_{\geq \epsilon}$ such that f^H factorizes over H. The graphical model H is an M-cover of the graphical model G whenever H is an M-cover of G and f^H is chosen as described above. In the sequel, we will write $f^H(x^H) = f^H(x^1, \ldots, x^M)$ where x_i^m is the m^{th} copy of variable $i \in V$.

For discrete graphical models, there is a relationship between the Bethe partition function and the true partition function of each graph cover.

Theorem 2.2 (Theorem 27 of Vontobel (2013))**.** *For any graphical model over a finite state space,*

$$Z_{\mathrm{B}}(G) = \lim \sup_{M \to \infty} \sqrt[M]{\sum_{H \in \mathcal{C}^M(G)} Z(H)/|\mathcal{C}^M(G)|}$$

where $\mathcal{C}^M(G)$ is the set of all M-covers of G.

The proof of this theorem relies on a counting argument: assignments on graph covers are mapped to pseudomarginals in the local marginal polytope. Under the mapping, every assignment on some M-cover that maps to the same collection of pseudomarginals must have the same energy. Then, to estimate the quantity under the M^{th}-root in the statement of the theorem, it suffices only to count how many assignments over all M-covers can map down to each specific collection of pseudomarginals in the local marginal polytope.

2.3 Log-supermodularity

A nonnegative, real-valued function, $g : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$, is log-supermodular (equivalently, multivariate totally positive of order two (Karlin and Rinott, 1980)) if

$$g(x)g(y) \leq g(x \wedge y)g(x \vee y)$$

for all $x, y \in \mathbb{R}^n$, where $x \vee y$ is the componentwise maximum of the vectors x and y and $x \wedge y$ is their componentwise minimum. A strictly positive twice continuously differentiable function $g : \mathbb{R}^n \to \mathbb{R}_{>0}$ is log-supermodular if and only if $\frac{\partial^2 \log g}{\partial x_i \partial x_j} \geq 0$ for all $i \neq j \in V$.

A graphical model is said to be log-supermodular decomposable if the objective can be factorized as a product of log-supermodular potentials. Such graphical models are sometimes said to be "attractive": the potential functions encourage agreement between the variables on which they depend. Discrete log-supermodular decomposable graphical models, in particular the ferromagnetic Ising model, have been very popular in computer vision applications, so much so that fast MAP inference algorithms have been developed for this special case (Kolmogorov and Zabih, 2002).

For log-supermodular decomposable graphical models over discrete state spaces, it was conjectured and then shown that $Z_{\mathrm{B}}(G) \leq Z(G)$ (Sudderth et al., 2007; Ruozzi, 2012): the proof of this result makes use of Theorem 2.2 and a correlation inequality for log-supermodular functions (Ruozzi, 2012).

3 THE CONTINUOUS CASE

In the remainder of this work, we explain how to partially extend Theorem 2.2 to continuous state spaces

by viewing the continuous case as a limit of discrete partition function computations. This combined with a correlation inequality will allow us to demonstrate that $Z_{\mathrm{B}}(G) \leq Z(G)$ for continuous log-supermodular decomposable graphical models. Expectation propagation may also yield a lower bound for such attractive models, though the techniques presented here are not sufficient to establish such a result.

3.1 Graph Covers and Z_{B}

We begin by proving an upper bound analogous to Theorem 2.2 in the continuous case.

Theorem 3.1. *For any continuous graphical model whose potential functions are bounded from below by some $\epsilon > 0$,*

$$Z_{\mathrm{B}}(G) \leq \limsup_{k \to \infty} \sqrt[M]{\sum_{H \in \mathcal{C}^M(G)} Z(H)/|\mathcal{C}^M(G)|}$$

where $\mathcal{C}^M(G)$ is the set of all M-covers of G.

Note that the above upper bound is all that is required for the proof of the main result. We conjecture that, as in the discrete case, equality should hold in the continuous case.

A complete proof of this theorem is described in Appendix A, but we describe the basic approach here. At a high level, we consider the special case of probability distributions with bounded support $[-t, t]^n$ for some large $t > 0$. We can carve up this domain into equal sized buckets of volume $1/2^{sn}$ for some $s \in \mathbb{Z}_{>0}$ by partitioning $[-t, t]$ into intervals of size $1/2^s$. The argument then considers only those pseudomarginal distributions in $\mathcal{M}_L$ that are constant over these partitions. As any Riemann-integrable distribution can be arbitrarily well approximated by distributions of this form, the result will follow by taking the limit as $s \to \infty$.

The key observation is that the Bethe free energy restricted to pseudomarginals that are constant over each partition can *almost* be expressed as the Bethe free energy of a discrete graphical model (there is a somewhat technical issue as differential entropy is not a limit of discrete entropy). We then apply Theorem 2.2 to this discrete graphical model and argue that the upper bound holds in the limit. This requires a somewhat more nuanced approach than the proof described by Vontobel (2013), but the basic approach remains the same.

3.2 Correlation Inequalities

We would like to use Theorem 3.1 to extend the lower bound results of Ruozzi (2012) for log-supermodular decomposable graphical models to the continuous case. Again, we accomplish this by proving the continuous analogs of the theorems used for the proof in the discrete case.

In particular, we prove a correlation inequality for continuous log-supermodular functions. For any collection of vectors $x^1, \ldots, x^M \in \mathbb{R}^n$, let $z^i(x^1, \ldots, x^M)$ be the vector whose j^{th} component is the i^{th} largest element of $x^1_j, \ldots, x^M_j$ for each $j \in \{1, \ldots, n\}$. We have the following theorem.

Theorem 3.2. *Let $f_1, \ldots, f_M : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and $g : \mathbb{R}^{Mn} \to \mathbb{R}_{\geq 0}$ be nonnegative real-valued functions such that g is log-supermodular. If for all $x^1, \ldots, x^M \in \mathbb{R}^n$,*

$$g(x^1, \ldots, x^M) \leq \prod_{i=1}^M f_i(z^i(x^1, \ldots, x^M)), \text{ then}$$

$$\int_{\mathbb{R}^{Mn}} g(x^1, \ldots, x^M) dx^1 \ldots dx^M \leq \prod_{i=1}^M \Big[\int_{\mathbb{R}^n} f_i(x) dx \Big].$$

The proof of this theorem can be found in Appendix B. It relies on extending a discrete version of the above inequality to the continuous case (Ruozzi, 2012). The argument is relatively straightforward: approximate the integrals as sums over a finite distributive lattice, apply the known theorem for the discrete case, and then take a limit as the discretization becomes finer and finer.

3.3 Putting It All Together

Finally, we are ready to state and prove that the Bethe partition function provides a lower bound on the true partition function of log-supermodular decomposable graphical models.

Theorem 3.3. *If $f^G : \mathbb{R}^n \to \mathbb{R}_{\geq \epsilon}$ is log-supermodular decomposable over $G = (V, \mathcal{A})$, then for any M-cover, H, of G, $Z(H) \leq Z(G)^M$.*

Proof. We emulate the proof of Ruozzi (2012). Let H be an M-cover of G. As described previously, each variable and hyperedge of G must appear M times in the cover. We denote the assignment of the i^{th} copy of each variable in G by the vector x^i.

For each $\alpha \in \mathcal{A}$, let y^i_α denote the assignment to the i^{th} copy of α by the elements of $x^1, \ldots, x^M$. By repeated application of the definition of log-supermodularity, we have

$$\begin{aligned} \prod_{i=1}^M \psi_\alpha(y^i_\alpha) &\leq \prod_{i=1}^M \psi_\alpha(z^i(y^1_\alpha, \ldots, y^M_\alpha)) \\ &= \prod_{i=1}^M \psi_\alpha(z^i(x^1_\alpha, \ldots, x^M_\alpha)) \end{aligned}$$

$$= \prod_{i=1}^{M} \psi_\alpha(z^i(x^1, \ldots, x^M)_\alpha).$$

From this, we can conclude that $f^H(x^1, \ldots, x^M) \leq \prod_{i=1}^{k} f^G(z^i(x^1, \ldots, x^M))$. Now, by Theorem 3.2,

$$\begin{aligned} Z(H) &= \int_{\mathbb{R}^{Mn}} f^H(x^1, \ldots, x^M) dx^1, \ldots dx^M \\ &\leq \prod_{i=1}^{M} \Big[\int_{\mathbb{R}^n} f^G(x) dx \Big] \\ &= Z(G)^M \end{aligned}$$

as claimed. □

The desired lower bound is then a simple corollary of this theorem and Theorem 3.1.

Corollary 3.4. *Under the conditions of Theorem 3.3,* $Z_{\mathrm{B}}(G) \leq Z(G)$.

4 DISCUSSION

A few general remarks about the above theorems in the continuous case are in order. First, even if the partition function $Z(G)$ exists and is finite, it does not mean that $Z_{\mathrm{B}}(G)$ is necessarily finite. In particular, there could exist an M-cover H of G such that $Z(H)$ is not finite. As a result, for any M', there exists an $M'' > M'$ and an M''-cover H'' such that $Z(H'')$ is not finite. Because all of the potentials are nonnegative, as we consider larger and larger supports $[-t, t]^n$, $Z_{\mathrm{B}}(G)$ could tend towards infinity.

This does indeed happen in practice. The canonical example is given by pairwise Gaussian graphical models, i.e., graphical models of the form

$$\begin{aligned} p(x) &\propto \exp\left(-1/2 x^T A x + b^T x\right) \\ &= \prod_{i \in V} \exp\left(-\frac{1}{2} A_{ii} x_i^2 + b_i x_i\right) \prod_{(i,j) \in E} \exp\left(-A_{ij} x_i x_j\right) \end{aligned}$$

for some symmetric positive definite matrix $A \in \mathbb{R}^{n \times n}$ (the inverse covariance matrix) and some vector $b \in \mathbb{R}^n$. It has been shown that for such models that when the matrix A is not walk-summable (Malioutov et al., 2006), then there exists a 2-cover whose covariance matrix is not positive semidefinite (Ruozzi and Tatikonda, 2013). Such models are not normalizable, i.e., they do not have finite partition functions. Separately, Cseke and Heskes (2011) showed that the Bethe free energy is unbounded in this case. If equality could be established in Theorem 3.1, then combined with the results of Ruozzi and Tatikonda (2013); Ruozzi et al. (2009) this would immediately yield an alternative proof.

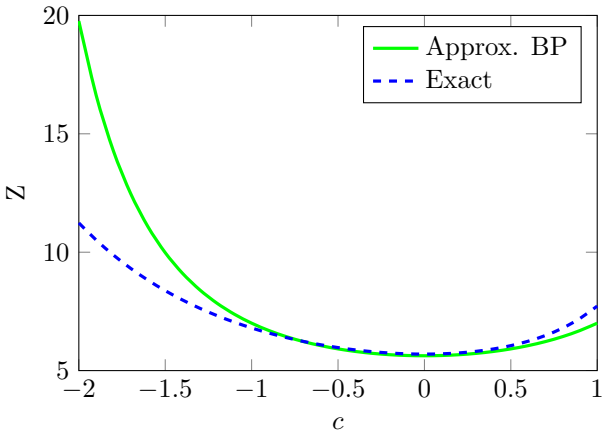


Figure 2: Locally optimal Gaussian beliefs in the Bethe free energy versus the exact partition function for the model $f(x_1, x_2, x_3) \propto \exp\left(-|x_1|^3 - |x_2|^3 - |x_3|^3 + cx_1x_2 + cx_1x_3 + cx_2x_3\right)$ for various values of c. The model is log-supermodular for $c \geq 0$. As $|c|$ increases the model becomes increasingly multimodal, and the Bethe free energy over only Gaussian distributions provides a poorer and poorer estimate of the true partition function.

More generally, as the above example illustrates, the Bethe partition function of log-concave probability distributions is not necessarily well-behaved.

Second, if the partition function exists and is finite for a log-supermodular decomposable graphical model, then the Bethe partition function exists and is finite. This makes log-supermodular decomposable functions a nice family to work with in the continuous setting as the Bethe partition function always provides a meaningful lower bound on the true partition function for these models. Further, the approximate MAP inference problem, obtained by dropping the entropy terms from the Bethe approximation, is exact for log-supermodular models (Ruozzi, 2015). Previous work has shown that EP behaves well for strongly log-concave potential functions whose third through sixth derivatives are bounded (Dehaene and Barthelmé, 2015), but these restrictions are quite severe compared to log-supermodularity.

Third, log-supermodular decomposable models need not be unimodal: there are log-supermodular models that are log-concave, log-convex, and even multimodal. Figure 2 illustrates the behavior of the Bethe partition function for a log-supermodular model whose corresponding probability distribution is multimodal. We note that it is not particularly difficult to construct multimodal log-supermodular decomposable graphical models with a finite partition function. In particular, one can take any log-supermodular decomposable

graphical model that is possibly unbounded on $\mathbb{R}^n$ and restrict it to a box in $\mathbb{R}^n$ on which it is bounded. As every box in $\mathbb{R}^n$ is a sublattice of $\mathbb{R}^n$, the resulting graphical model is log-supermodular decomposable over the box.

Finally, recall that expectation propagation further relaxes the Bethe free energy, i.e., $Z_{\mathrm{B}} \leq Z_{\mathrm{EP}}$. This means that, in general, if $Z_{\mathrm{B}} \geq Z$, then Z_{B} is necessarily a better approximation to the true partition function than Z_{EP}. However, for log-supermodular models, if Z_{EP} is less than Z, then it must necessarily outperform the Bethe free energy approximation.

In summary, we have provided an upper bound on the Bethe partition function of continuous graphical models. We used this upper bound to prove that $Z \geq Z_{\mathrm{B}}$ for continuous log-supermodular graphical models, which matches a result for discrete log-supermodular graphical models. To apply these theoretical results to real-world log-supermodular decomposable models would require fast/practical algorithms. We leave this as the subject of future work.

ACKNOWLEDGMENTS

This work was supported in part by National Science Foundation grant III-1527312.

A PROOF OF THEOREM 3.1

For simplicity, we work with joint distributions with bounded support over $[-t,t]^{|V|}$ for some positive integer t and pseudomarginals that are supported on subsets of this space. Specifically, $\tau_i : [-t,t] \to \mathbb{R}_{\geq \epsilon}$ for each $i \in V$ and $\tau_\alpha : [-t,t]^{|\alpha|} \to \mathbb{R}_{\geq \epsilon}$ for each $\alpha \in \mathcal{A}$.

We can partition each $[-t,t]$ interval into boxes of size $\Delta_s = \frac{1}{2^s}$ for some positive integer s. Let $\mathcal{M}_L^s \subseteq \mathcal{M}_L$ denote the set of all pseudomarginals that are piecewise constant over each of the boxes of size Δ_s. Any Riemann integrable probability distribution over $[-t,t]^{|V|}$ can be arbitrarily well approximated as $s \to \infty$, therefore any collection of pseudomarginals $\tau \in \mathcal{M}_L$ can be arbitrarily well approximated by a sequence of $\tau^s \in \mathcal{M}_L^s$ for $s \in \mathbb{N}$ (i.e., $\tau^s \to \tau$ as $s \to \infty$). We will consider maximizing the Bethe free energy over $\mathcal{M}_L^s$.

Let $\mathcal{Y}_s \triangleq \{1, \ldots, 2t/\Delta_s\}$, and for any $k \in \mathcal{Y}_s$, let $I_k^s \subseteq [-t,t]$ denote the k^{th} partition of size Δ_s. We extend this notation to vectors by forming the product, $I_y^s \triangleq \prod_{i \in V} I_{y_i}^s$ for each $y \in \mathcal{Y}_s^{|V|}$.

For any $\tau^s \in \mathcal{M}_L^s$, we can define a collection of discretized pseudomarginals μ^s for each $y \in \mathcal{Y}_s^{|V|}$.

$$\begin{aligned}\mu_i^s(y_i^s) &\triangleq \int_{\mathcal{I}_{y_i}^s} \tau_i^s(x_i) dx_i \\ &= \Delta_s \tau_i^s(x_i), \text{ for any } x_i \in \mathcal{I}_{y_i}^s \\ \mu_\alpha^s(y_\alpha^s) &\triangleq \Delta_s^{|\alpha|} \tau_\alpha^s(x_\alpha), \text{ for any } x_\alpha \in \mathcal{I}_{y_\alpha}^s\end{aligned}$$

With these definitions, the Bethe free energy of $\tau^s \in \mathcal{M}_L^s$ is given by

$$F_{\mathrm{B}}(G, \tau^s) \geq F_{\mathrm{B}}^s(G, \mu^s) \triangleq U^s(G, \mu^s) - H^s(G, \mu^s)$$

where

$$\begin{aligned}U^s(G,\mu^s) = &- \sum_{i \in V} \sum_{y_i} \mu_i^s(y_i) \inf_{x_i \in I_{y_i}^s} \log f_i(x_i) \\ &- \sum_{\alpha \in \mathcal{A}} \sum_{y_\alpha} \mu_\alpha^s(y_\alpha) \inf_{x_\alpha \in I_{y_\alpha}^s} \log f_\alpha(x_\alpha), \\ H^s(G,\mu^s) = &- \sum_{i \in V} \left[-\log \Delta_s + \sum_{y_i} \mu_i^s(y_i) \log \mu_i^s(y_i) \right] \\ &- \sum_{\alpha \in \mathcal{A}} \sum_{y_\alpha} \mu_\alpha^s(y_\alpha) \log \frac{\mu_\alpha^s(y_\alpha)}{\prod_{i \in \alpha} \mu_i^s(y_i)}.\end{aligned}$$

This is a "Δ_s-corrected" version of the Bethe free energy of the discrete pseudomarginals μ corresponding to a discrete graphical model over the hypergraph G with potential functions

$$\widehat{f}_i^s(y_i) \triangleq \inf_{x_i \in I_{y_i}^s} f_i(x_i)$$

$$\widehat{f}_\alpha^s(y_\alpha) \triangleq \inf_{x_\alpha \in I_{y_\alpha}^s} \log f_\alpha(x_\alpha)$$

and joint distribution

$$\widehat{f}_s(y_1, \ldots, y_{|V|}) \propto \prod_{i \in V} \widehat{f}_i^s(y_i) \prod_{\alpha \in \mathcal{A}} \widehat{f}_\alpha^s(y_\alpha).$$

This, in turn, provides a lower bound on the original joint distribution with a distribution that is constant over each of the partitions of $[-t,t]^{|V|}$. Note that the approximations are nondecreasing in s.

Denote the local marginalization constraints in the discrete case as

$$\begin{aligned}\mathcal{T}_L^s \triangleq \Bigg\{ &\mu^s \geq 0 : \sum_{y_i} \mu_i^s(y_i) = 1, \forall i \in V, \text{ and} \\ &\sum_{y_{\alpha \setminus \{i\}}} \mu_\alpha^s(y_\alpha) = \mu_i^s(y_i), \forall \alpha \in \mathcal{A}, i \in \alpha, y_i \in \mathcal{Y}_s \Bigg\}\end{aligned}$$

with $Z_{\mathrm{B}}^s = \exp\left(-\inf_{\mu \in \mathcal{T}_L^s} F_{\mathrm{B}}^s(\mu)\right)$. Additionally, we will write $Z^s(G) \triangleq Z(G; \widehat{f}_{i \in V}^s, \widehat{f}_{\alpha \in \mathcal{A}}^s)$ to denote the partition function of the discrete model.

We have that the limit of the discrete partition functions is equal to the true partition function and that the limit of the discrete Bethe approximations is equal to the exact Bethe approximation.

$$\lim_{s\to\infty} Z_s(G)\Delta_s^{|V|} = Z(G)$$
$$\lim_{s\to\infty} Z_{\mathrm{B}}^s(G) = \sup_{s>0} Z_{\mathrm{B}}^s(G) = Z_{\mathrm{B}}(G) \tag{1}$$

The roadmap for the remainder of the proof is as follows. We have shown how to represent exact partition function computations in the continuous case as a limit of the partition function of discrete graphical models. To evaluate the average partition function over all M-covers as required by the theorem, we will map assignments on M-covers to discrete pseudomarginals in $\mathcal{T}_L^s$. As any assignment that maps to the same pseudomarginal must have the same energy (i.e., evaluate to the same number when plugged into the joint distribution of the appropriate M-cover), we will only need to count how many assignments map to each pseudomarginal. We can then combine all of the above observations to prove the desired result.

Definition A.1. *Consider the following mapping, φ_M^s, from assignments/configurations on covers to pseudomarginals in $\mathcal{T}_L^s$:*

$$\varphi_M^s : \{(H, y^H) | H \in \mathcal{C}^M(G), y^H \in \mathcal{Y}_s^{M|V|}\} \to \mathcal{T}_L^s$$
$$: (H, y^H) \mapsto \mu$$

Let $h(\cdot)$ be the covering map from H to G. The components of μ are given by

$$\mu_\alpha(y_\alpha) = \sum_{\beta\in\mathcal{A}(H):h(\beta)=\alpha} \frac{1_{y_\beta^H = x_\alpha}}{M}, \forall\alpha\in\mathcal{A}(G), y_\alpha \in \mathcal{Y}_s^{|\alpha|}$$
$$\mu_i(y_i) = \sum_{j\in V(H):h(j)=i} \frac{1_{y_j^H = y_i}}{M}, \forall i \in V(G), y_i \in \mathcal{Y}_s.$$

Each $\mu_i(y_i)$ corresponds to the number of times that each of the M copies of the vertex $i \in G$ in the M-cover, H, are equal to y_i in the assignment y^H, divided by M. Note that, by construction, $\exp(-M\cdot U(\mu)) = \widehat{f}_s^H(y^H)$.

Definition A.2. *The set of all pseudomarginals realizable by some configuration on some M-cover, denoted $\tilde{\mathcal{T}}_s^M \subseteq \mathcal{T}_L^s$, is the image of the pseudo-marginal mapping.*

$$\tilde{\mathcal{T}}_s^M \triangleq image(\varphi_M^s)$$

The size of the set of all pseudo-marginal lift realizable vectors, $|\tilde{\mathcal{T}}_s^M|$, grows polynomially with M for a fixed s.

$$|\tilde{\mathcal{T}}_s^M| \le (M+1)^{(|V||\mathcal{Y}_s| + \sum_{\alpha\in\mathcal{A}} |\mathcal{Y}_s|^{|\alpha|})} \tag{2}$$

This can be seen by observing that for some vertex i, μ_i has $|\mathcal{Y}_s|$ possible assignments, each of which can take one of the $M+1$ values in the set $\{\frac{0}{M}, \frac{1}{M}, ..., \frac{M}{M}\}$. Similar reasoning applies to each factor $\alpha \in \mathcal{A}$.

Our goal is to count the average number of assignments over all M-covers that map down via φ_M^s to a specific $\mu \in \tilde{\mathcal{T}}_s^M$. We will denote this quantity by

$$\bar{C}_M^s(\mu) \triangleq \frac{|(\varphi_M^s)^{-1}(\mu)|}{|\mathcal{C}^M(G)|}.$$

Lemma A.3. *For each $\mu \in \tilde{\mathcal{T}}_s^M$,*

$$\bar{C}_M^s(\mu) = \prod_{i\in V} \binom{M}{M\cdot\mu_i}^{|1-\partial i|} \prod_{\alpha\in\mathcal{A}} \binom{M}{M\cdot\mu_\alpha}$$
$$= \exp(M\cdot H_{\mathrm{B}}(\mu) + o(M))$$

where

$$\binom{M}{M\cdot\mu_i} \triangleq \frac{M!}{\prod_{y_i\in\mathcal{Y}_s}(M\mu_i(y_i))!}$$
$$\binom{M}{M\cdot\mu_\alpha} \triangleq \frac{M!}{\prod_{y_\alpha\in\mathcal{Y}_s^{|\alpha|}}(M\mu_\alpha(y_\alpha))!}.$$

Proof. Vontobel (2013) proves a similar result (see Lemma 29 and Theorem 30) for factor graphs in normal form (into which any graphical model can easily be converted). We state the result here for the general case and to emphasize that terms in the $o(M)$ term do not depend on s. As the argument is nearly identical to (Vontobel, 2013), we omit it due to space constraints. □

We now have all of the tools necessary for the proof of the theorem. Define

$$Z_{\mathrm{B},M}^s(G) \triangleq \sqrt[M]{\frac{\sum_{H\in\mathcal{C}^M(G)} Z_s(H)\Delta_s^{|V|M}}{|\mathcal{C}^M(G)|}}.$$

Using the above counting arguments, we can relate $Z_{\mathrm{B},M}^s$ to the Bethe free energy of pseudomarginals $\mu \in \mathcal{T}_L^s$.

$$Z_{\mathrm{B},M}^s(G)^M$$
$$= \sum_{H\in\mathcal{C}^M(G)} \sum_{y^H\in\mathcal{Y}_s^{M|V|}} \frac{\widehat{f}_s^H(y^H)}{|\mathcal{C}^M(G)|}\Delta_s^{|V|M}$$
$$= \sum_{\mu\in\tilde{\mathcal{T}}_s^M} \sum_{H\in\mathcal{C}^M(G)} \sum_{y^H:\varphi_M^s(H,y^H)=\mu} \frac{\widehat{f}_s^H(y^H)}{|\mathcal{C}^M(G)|}\Delta_s^{|V|M}$$
$$= \sum_{\mu\in\tilde{\mathcal{T}}_s^M} \sum_{H\in\mathcal{C}^M(G)} \sum_{y^H:\varphi_M^s(H,y^H)=\mu} \frac{\exp(-M\cdot U^s(\mu))}{|\mathcal{C}^M(G)|}\Delta_s^{|V|M}$$
$$= \sum_{\mu\in\tilde{\mathcal{T}}_s^M} \exp(-M\cdot U^s(\mu))\cdot\bar{C}_M^s(\mu)\Delta_s^{|V|M}$$

$$= \sum_{\mu \in \tilde{\mathcal{T}}_s^M} \exp\left(-M \cdot F_{\mathrm{B}}^s(\mu) + o(M)\right)$$

To complete the proof, we need to argue that, after taking the appropriate limits, this quantity upper bounds $Z_{\mathrm{B}}(G)$. Fix an $s' > 0$.

$$\begin{aligned}
&\limsup_{M\to\infty} \lim_{s\to\infty} Z_{B,M}^s \\
&\overset{(a)}{\geq} \limsup_{M\to\infty} \sqrt[M]{\sum_{\mu \in \tilde{\mathcal{T}}_{s'}^M} \exp\left(-M \cdot F_{\mathrm{B}}^{s'}(\mu) + o(M)\right)} \\
&\overset{(b)}{=} \limsup_{M\to\infty} \sqrt[M]{\max_{\mu \in \tilde{\mathcal{T}}_{s'}^M} \exp(-M \cdot F_{\mathrm{B}}^{s'}(\mu) + o(M))} \\
&= \limsup_{M\to\infty} \max_{\mu \in \tilde{\mathcal{T}}_{s'}^M} \exp(-F_{\mathrm{B}}^{s'}(\mu) + o(1)) \\
&= \sup_{\mu \in \tilde{\mathcal{T}}_L^{s'}} \exp(-F_{\mathrm{B}}^{s'}(\mu)) \\
&= \exp(-\inf_{\mu \in \tilde{\mathcal{T}}_L^{s'}} F_{\mathrm{B}}(\mu)) \\
&= Z_{\mathrm{B}}^{s'}
\end{aligned}$$

Step (a) follows from the observation that $\lim_{s\to\infty} = \sup_{s\to\infty}$, see (1). Step (b) follows from the observation that, for a fixed s, the number of terms in the sum is growing polynomially in M, see (2).

As the lower bound holds for any s', we must have

$$\begin{aligned}
\limsup_{M\to\infty} \lim_{s\to\infty} Z_{B,M}^s &\geq \sup_s Z_{\mathrm{B}}^s \\
&= \lim_s Z_{\mathrm{B}}^s \\
&= Z_{\mathrm{B}}
\end{aligned}$$

as desired.

B PROOF OF THEOREM 3.2

Again, we will assume that all of the functions involved are bounded over every finite interval, continuous almost everywhere, and Riemann integrable. We will extend the following result to the continuous case.

Theorem B.1 (Theorem 3.8 of (Ruozzi, 2012))**.** *Let $f_1, \ldots, f_k : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ and $g : \{0,1\}^{kn} \to \mathbb{R}_{\geq 0}$ be nonnegative real-valued functions such that g is log-supermodular. If for all $x^1, \ldots, x^k \in \{0,1\}^n$,*

$$g(x^1, \ldots, x^k) \leq \prod_{i=1}^k f_i(z^i(x^1, \ldots, x^k)), \text{ then}$$

$$\sum_{x_1,\ldots,x_k \in \{0,1\}^n} g(x^1, \ldots, x^k) \leq \prod_{i=1}^k \Big[\sum_{x \in \{0,1\}^n} f_i(x) \Big].$$

Every finite distributive lattice can be embedded as a sublattice of $\{0,1\}^n$ (Alon and Spencer, 2000). As this embedding preserves log-supermodularity, Theorem B.1 applies over any finite distributive lattice, not just $\{0,1\}^n$.

As in the statement of Theorem 3.2, let $f_1, \ldots, f_M : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ and $g : \mathbb{R}^{Mn} \to \mathbb{R}_{\geq 0}$ be nonnegative real-valued functions such that g is log-supermodular and for all $x^1, \ldots, x^M \in \mathbb{R}^n$,

$$g(x^1, \ldots, x^M) \leq \prod_{i=1}^M f_i(z^i(x^1, \ldots, x^M)).$$

As in the proof of Theorem 3.1, we will work with joint distributions with bounded support over $[-t,t]^n$ for some positive integer t and pseudomarginals that are supported on subsets of this space. Specifically, $f_i : [-t,t]^n \to \mathbb{R}_{\geq 0}$ for each $i \in V$ and $g : [-t,t]^{Mn} \to \mathbb{R}_{\geq 0}$. We can partition each $[-t,t]$ interval into boxes of size $\Delta_s = \frac{1}{2^s}$ for some positive integer s. Let $\mathcal{Y}_s \triangleq \{1, \ldots, 2t/\Delta_s\}$, and for any $k \in \mathcal{Y}_s$, let $I_k^s \subseteq [-t,t]$ denote the k^{th} partition of size Δ_s. Finally, let $\mathcal{X}_s$ denote the set of left endpoints of the partitions.

We will construct discrete approximations of the integrals by defining

$$\forall x \in \mathcal{X}_s^n, \widehat{f}_i^s(x) \triangleq f_i(x)\Delta_s^n \text{ and}$$

$$\forall x^1, \ldots, x^M \in \mathcal{X}_s^n, \widehat{g}^s(x^1, \ldots, x^M) \triangleq g(x^1, \ldots, x^M)\Delta_s^{Mn}.$$

We have that

$$\lim_{s\to\infty} \sum_{x \in \mathcal{X}_s^n} \widehat{f}_i^s(x) = \int_{x \in [-t,t]^n} f_i(x)dx \text{ and}$$

$$\lim_{s\to\infty} \sum_{x^1,\ldots,x^M \in \mathcal{X}_s^n} \widehat{g}^s(x^1, \ldots, x^M) = \int_{x^1,\ldots,x^M \in [-t,t]^n} g(x)dx.$$

Moreover, $\widehat{g}^s$ is log-supermodular, and, for all $x^1, \ldots, x^M \in \mathcal{X}_s$,

$$\widehat{g}^s(x^1, \ldots, x^M) \leq \prod_{i=1}^M \widehat{f}_i^s(z^i(x^1, \ldots, x^M)).$$

Now, by Theorem B.1,

$$\sum_{x^1,\ldots,x^M \in \mathcal{X}_s^n} \widehat{g}^s(x^1, \ldots, x^M) \quad \leq \prod_{i=1}^M \left[\sum_{x \in \mathcal{X}_s^n} \widehat{f}_i^s(x) \right].$$

Taking limits on both sides of the inequality gives the desired result.

References

N. Alon and J.H. Spencer. *The probabilistic method*. Wiley-Interscience series in discrete mathematics and optimization. Wiley, 2000. ISBN 9780471370468.

B. Cseke and T. Heskes. Properties of Bethe free energies and message passing in Gaussian models. *Journal of Artificial Intelligence Research*, 41:1–24, 2011.

G. P. Dehaene and S. Barthelmé. Bounding errors of expectation-propagation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 244–252. 2015.

T. Heskes and O. Zoeter. Expectation propagation for approximate inference in dynamic Bayesian networks. In *Proceedings of the Eighteenth conference on Uncertainty in Artificial Intelligence (UAI)*, pages 216–223, 2002.

A. Ihler and D. McAllester. Particle belief propagation. In *Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, FClearwater Beach, Florida USA, Apr. 2009.

M. Isard, J. MacCormick, and K. Achan. Continuously-adaptive discretization for message-passing algorithms. In *Advances in Neural Information Processing Systems (NIPS)*, pages 737–744, Vancouver, B.C., Dec. 2008.

S. Karlin and Y. Rinott. Classes of orderings of measures and related correlation inequalities. i. multivariate totally positive distributions. *Journal of Multivariate Analysis*, 10(4):467 – 498, 1980.

V. Kolmogorov and R. Zabih. What energy functions can be minimized via graph cuts? In *Computer VisionECCV 2002*, pages 65–81. Springer, 2002.

T. Lienart, Y. W. Teh, and A. Doucet. Expectation particle belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3609–3617, 2015.

D. M. Malioutov, J. K. Johnson, and A. S. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *Journal of Machine Learning Research*, 7:2031–2064, 2006.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369, 2001.

N. Noorshams and M. J. Wainwright. Belief propagation for continuous state spaces: Stochastic message-passing with quantitative guarantees. *J. Mach. Learn. Res. (JMLR)*, 14(1):2799–2835, Jan. 2013.

N. Ruozzi. The Bethe partition function of log-supermodular graphical models. In *Neural Information Processing Systems (NIPS)*, Lake Tahoe, NV, Dec. 2012.

N. Ruozzi. Beyond log-supermodularity: Lower bounds and the bethe partition function. In *Proceedings of the Twenty-Ninth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI-13)*, pages 546–555, Corvallis, Oregon, 2013. AUAI Press.

N. Ruozzi. Approximate MAP inference in continuous MRFs. In *Advances in Neural Information Processing Systems (NIPS)*, Dec. 2015.

N. Ruozzi and S. Tatikonda. Message-passing algorithms for quadratic minimization. *Journal of Machine Learning Research*, 14:2287–2314, 2013.

N. Ruozzi, J. Thaler, and S. Tatikonda. Graph covers and quadratic minimization. In *Communication, Control, and Computing, 2009. Allerton 2009. 47th Annual Allerton Conference on*, pages 1590–1596. IEEE, 2009.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, Fort Lauderdale, FL, USA, Apr. 2011.

E. B. Sudderth, M. J. Wainwright, and A. S. Willsky. Loop series and Bethe variational bounds in attractive graphical models. In *Neural Information Processing Systems (NIPS)*, Vancouver, BC, Canada, Dec. 2007.

P. O. Vontobel. Counting in graph covers: A combinatorial characterization of the Bethe entropy function. *Information Theory, IEEE Transactions on*, Jan. 2013.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282 – 2312, July 2005.