The Bethe Partition Function of Log-supermodular Graphical Models

Nicholas Ruozzi
Communication Theory Laboratory
EPFL
Lausanne, Switzerland
nicholas.ruozzi@epfl.ch

Abstract

Sudderth, Wainwright, and Willsky conjectured that the Bethe approximation corresponding to any fixed point of the belief propagation algorithm over an attractive, pairwise binary graphical model provides a lower bound on the true partition function. In this work, we resolve this conjecture in the affirmative by demonstrating that, for any graphical model with binary variables whose potential functions (not necessarily pairwise) are all log-supermodular, the Bethe partition function always lower bounds the true partition function. The proof of this result follows from a new variant of the "four functions" theorem that may be of independent interest.

1 Introduction

Graphical models have proven to be a useful tool for performing approximate inference in a wide variety of application areas including computer vision, combinatorial optimization, statistical physics, and wireless networking. Computing the partition function of a given graphical model, a typical inference problem, is an NP-hard problem in general. Because of this, the inference problem is often replaced by a variational approximation that is, hopefully, easier to solve. The Bethe approximation, one such standard approximation, is of great interest both because of its practical performance and because of its relationship to the belief propagation (BP) algorithm: stationary points of the Bethe free energy function correspond to fixed points of belief propagation [1]. However, the Bethe partition function is only an approximation to the true partition function and need not provide an upper or lower bound.

In certain special cases, the Bethe approximation is conjectured to provide a lower bound on the true partition function. One such example is the class of attractive pairwise graphical models: models in which the interaction between any two neighboring variables places a greater weight on assignments in which the two variables agree. Many applications in computer vision and statistical physics can be expressed as attractive pairwise graphical models (e.g., the ferromagnetic Ising model). Sudderth, Wainwright, and Willsky [2] used a loop series expansion of Chertkov and Chernyak [3, 4] in order to study the fixed points of BP over attractive graphical models. They provided conditions on the fixed points of BP under which the stationary points of the Bethe free energy function corresponding to these fixed points are a lower bound on the true partition function. Empirically, they observed that, even when their conditions were not satisfied, the Bethe partition function appeared to lower bound the true partition function, and they conjectured that this is always the case for attractive pairwise binary graphical models.

Recent work on the relationship between the Bethe partition function and the graph covers of a given graphical model has suggested a new approach to resolving this conjecture. Vontobel [5] demonstrated that the Bethe partition function can be precisely characterized by the average of the true partition functions corresponding to graph covers of the base graphical model. The primary contribution of the present work is to show that, for graphical models with log-supermodular potentials, the partition function associated with any graph cover of the base graph, appropriately normalized, must lower bound the true partition function. As pairwise binary graphical models are log-supermodular if and only if they are attractive, combining our result with the observations of [5] resolves the conjecture of [2].

The key element in our proof, and the second contribution of this work, is a new variant of the "four functions" theorem that is specific to log-supermodular functions. We state and prove this variant in Section 3.1, and in Section 4.1, we use it to resolve the conjecture. As a final contribution, we demonstrate that our variant of the "four functions" theorem has applications beyond log-supermodular functions: as an example, we use it to show that the Bethe partition function can also provide a lower bound on the number of independent sets in a bipartite graph.

2 Undirected Graphical Models

Let $f : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ be a non-negative function. We say that f factors with respect to a hypergraph $G = (V, \mathcal{A})$ where $\mathcal{A} \subseteq 2^V$, if there exist potential functions $\phi_i : \{0,1\} \to \mathbb{R}_{\geq 0}$ for each $i \in V$ and $\psi_\alpha : \{0,1\}^{|\alpha|} \to \mathbb{R}_{\geq 0}$ for each $\alpha \in \mathcal{A}$ such that

$$f(x) = \prod_{i \in V} \phi_i(x_i) \prod_{\alpha \in \mathcal{A}} \psi_\alpha(x_\alpha)$$

where x_α is the subvector of the vector x indexed by the set α.

We will express the hypergraph G as a bipartite graph that consists of a variable node for each $i \in V$, a factor node for each $\alpha \in \mathcal{A}$, and an edge joining the factor node corresponding to α to the variable node representing i if $i \in \alpha$. This is typically referred to as the *factor graph* representation of G.

Definition 2.1. *A function $f : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ is **log-supermodular** if for all $x, y \in \{0,1\}^n$*

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

*where $(x \wedge y)_i = \min\{x_i, y_i\}$ and $(x \vee y)_i = \max\{x_i, y_i\}$. Similarly, a function $f : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ is **log-submodular** if for all $x, y \in \{0,1\}^n$*

$$f(x)f(y) \geq f(x \wedge y) f(x \vee y).$$

Definition 2.2. *A factorization of a function $f : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ over $G = (V, \mathcal{A})$ is log-supermodular if for all $\alpha \in \mathcal{A}$, $\psi_\alpha(x_\alpha)$ is log-supermodular.*

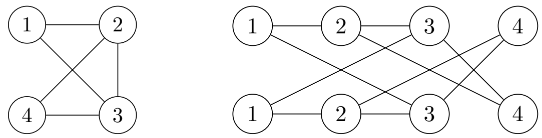
Every function that admits a log-supermodular factorization is necessarily log-supermodular, products of log-supermodular functions are easily seen to be log-supermodular, but the converse may not be true outside of special cases. If $|\alpha| \leq 2$ for each $\alpha \in \mathcal{A}$, then we call the factorization pairwise. For any pairwise factorization, f is log-supermodular if and only if ψ_{ij} is log-supermodular for each i and j.

Pairwise graphical models such that $\psi_\alpha(x_\alpha)$ is log-supermodular for all $\alpha \in \mathcal{A}$ are referred to as *attractive* graphical models. A generalization of attractive interactions to the non-pairwise case is presented in [2]: for all $\alpha \in \mathcal{A}$, ψ_α, when appropriately normalized, has non-negative central moments. However, the relationship between this generalization and log-supermodularity remains unclear.

2.1 Graph Covers

Graph covers have played an important role in our understanding of graphical models [5, 6]. Roughly, if a graph H covers a graph G, then H looks locally the same as G.

Definition 2.3. *A graph H **covers** a graph $G = (V, E)$ if there exists a graph homomorphism $h : H \to G$ such that for all vertices $v \in G$ and all $w \in h^{-1}(v)$, h maps the neighborhood ∂w of w in H bijectively to the neighborhood ∂v of v in G. If $h(w) = v$, then we say that $w \in H$ is a copy of $v \in G$. Further, H is a k-cover of G if every vertex of G has exactly k copies in H.*



(a) A graph, G.

(b) One possible cover of G.

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

For an example of a graph cover, see Figure 1.

For the factor graph corresponding to $G = (V, \mathcal{A})$, each k-cover consists of a variable node for each of the $k|V|$ variables, a factor node for each of the $k|\mathcal{A}|$ factors, and an edge joining each copy of $\alpha \in \mathcal{A}$ to a distinct copy of each $i \in \alpha$. To any k-cover $H = (V_H, \mathcal{A}_H)$ of G given by the homomorphism h, we can associate a collection of potentials: the potential at node $i \in V_H$ is equal to $\phi_{h(i)}$, the potential at node $h(i) \in G$, and for each $\alpha \in \mathcal{A}_H$, we associate the potential $\psi_{h(\alpha)}$. In this way, we can construct a function $f^H : \{0,1\}^{kn} \to \mathbb{R}_{\geq 0}$ such that f^H factorizes over H.

Notice that if f^G admits a log-supermodular factorization over G and H is a k-cover of G, then f^H admits a log-supermodular factorization over H.

2.2 Bethe Approximations

For a function $f : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ that factorizes over $G = (V, \mathcal{A})$, we are interested computing the partition function $Z(G) = \sum_x f(x)$. In general, this is an NP-hard problem, but in practice, algorithms, such as belief propagation, based on variational approximations produce reasonable estimates in many settings. One such variational approximation, the Bethe approximation at temperature $T = 1$, is defined as follows:

$$\begin{aligned}\log Z_{\mathrm{B}}(G, \tau) = &\sum_{i \in V} \sum_{x_i} \tau_i(x_i) \log \phi_i(x_i) + \sum_{\alpha \in \mathcal{A}} \sum_{x_\alpha} \tau_\alpha(x_\alpha) \log \psi_\alpha(x_\alpha) \\ &- \sum_{i \in V} \sum_{x_i} \tau_i(x_i) \log \tau_i(x_i) - \sum_{\alpha \in \mathcal{A}} \sum_{x_\alpha} \tau_\alpha(x_\alpha) \log \frac{\tau_\alpha(x_\alpha)}{\prod_{i \in \alpha} \tau_i(x_i)}\end{aligned}$$

for τ in the local marginal polytope,

$$\mathcal{T} \triangleq \{\tau \geq 0 \mid \forall \alpha \in \mathcal{A}, i \in \alpha, \sum_{x_{\alpha \setminus i}} \tau_\alpha(x_\alpha) = \tau_i(x_i) \text{ and } \forall i \in V, \sum_{x_i} \tau_i(x_i) = 1\}.$$

The fixed points of the belief propagation algorithm correspond to stationary points of $\log Z_{\mathrm{B}}(G, \tau)$ over $\mathcal{T}$, the set of pseudomarginals [1], and the Bethe partition function is defined to be the maximum value achieved by this approximation over $\mathcal{T}$:

$$Z_{\mathrm{B}}(G) = \max_{\tau \in \mathcal{T}} Z_{\mathrm{B}}(G, \tau).$$

For a fixed factor graph G, we are interested in the relationship between the true partition function, $Z(G)$, and the Bethe approximation corresponding to G, $Z_{\mathrm{B}}(G)$. While, in general, $Z_{\mathrm{B}}(G)$ can be either an upper or a lower bound on the true partition function, in this work, we address the following conjecture of [2]:

Conjecture 2.4. *If $f : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ admits a pairwise, log-supermodular factorization over $G = (V, \mathcal{A})$, then $Z_{\mathrm{B}}(G) \leq Z(G)$.*

We resolve this conjecture in the affirmative, and show that it continues to hold for a larger class of log-supermodular functions. Our results are based, primarily, on two observations: a variant of the "four functions" theorem [7] and the following, recent theorem of Vontobel [5]:

Theorem 2.5.

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

where $\mathcal{C}^k(G)$ is the set of all k-covers of G. [1]

Proof. See Theorem 27 of [5]. $\square$

Theorem 2.5 suggests that a reasonable strategy for proving that $Z_{\mathrm{B}}(G) \le Z(G)$ would be to show that $Z(H) \le Z(G)^k$ for any k-cover H of G. This is the strategy that we adopt in the remainder of this work.

3 The "Four Functions" Theorem and Related Results

The "four functions" theorem [7] is a general result concerning nonnegative functions over distributive lattices. Many correlation inequalities from statistical physics, such as the FKG inequality, can be seen as special cases of this theorem [8].

Theorem 3.1 ("Four Functions" Theorem)**.** *Let $f_1, f_2, f_3, f_4 : \{0,1\}^n \to \mathbb{R}_{\ge 0}$ be nonnegative real-valued functions. If for all $x, y \in \{0,1\}^n$,*

$$f_1(x)f_2(y) \le f_3(x \wedge y)f_4(x \vee y),$$

then

$$\Big[\sum_{x\in\{0,1\}^n} f_1(x)\Big]\Big[\sum_{x\in\{0,1\}^n} f_2(x)\Big] \le \Big[\sum_{x\in\{0,1\}^n} f_3(x)\Big]\Big[\sum_{x\in\{0,1\}^n} f_4(x)\Big].$$

The following lemma is a direct consequence of the four functions theorem:

Lemma 3.2. *If $f : \{0,1\}^n \to \mathbb{R}_{\ge 0}$ is log-supermodular, then every marginal of f is also log-supermodular.*

The four functions theorem can be extended to more than four functions, by generalizing $\wedge$ and $\vee$. For any collection of vectors $x^1, \ldots, x^k \in \mathbb{R}^n$, let $z^i(x^1, \ldots, x^k)$ be the vector whose j^{th} component is the i^{th} largest element of $x^1_j, \ldots, x^k_j$ for each $j \in \{1, \ldots, n\}$. As an example, for vectors $x^1, \ldots, x^k \in \{0,1\}^n$, $z^i(x^1, \ldots, x^k)_j = \{\sum_{a=1}^k x^a_j \ge i\}$ where $\{\cdot \ge \cdot\}$ is one if the inequality is satisfied and zero otherwise. The "four functions" theorem is then a special case of the more general "2k functions" theorem [9, 10, 11]:

Theorem 3.3 ("2k Functions" Theorem)**.** *Let $f_1, \ldots, f_k : \{0,1\}^n \to \mathbb{R}_{\ge 0}$ and $g_1, \ldots, g_k : \{0,1\}^n \to \mathbb{R}_{\ge 0}$ be nonnegative real-valued functions. If for all $x^1, \ldots, x^k \in \{0,1\}^n$,*

$$\prod_{i=1}^k g_i(x^i) \le \prod_{i=1}^k f_i(z^i(x^1, \ldots, x^k)), \tag{1}$$

then

$$\prod_{i=1}^k \Big[\sum_{x\in\{0,1\}^n} g_i(x)\Big] \le \prod_{i=1}^k \Big[\sum_{x\in\{0,1\}^n} f_i(x)\Big].$$

3.1 A Variant of the "Four Functions" Theorem

A natural generalization of Theorem 3.3 would be to replace the product of functions on the left-hand side of Equation 1 with an arbitrary function over $x^1, \ldots, x^k$: we will show that we can replace this product with an arbitrary log-supermodular function while preserving the conclusion of the theorem. The key property of log-supermodular functions that makes this possible is the following lemma:

[1] The proof of the theorem is demonstrated for "normal" factor graphs, but it easily extends to the factor graphs described above by replacing variable nodes with equality constraints.

Lemma 3.4. *If $g : \{0,1\}^n \to \mathbb{R}_{\geq 0}$ is log-supermodular, then for any integer $k \geq 1$ and $x^1, \ldots, x^k \in \{0,1\}^n$, $\prod_{i=1}^k g(x^i) \leq \prod_{i=1}^k g(z^i(x^1, \ldots, x^k))$.*

Proof. This follows directly from the log-supermodularity of g. □

The proof of our variant of the "$2k$ functions theorem" uses the properties of weak majorizations:

Definition 3.5. *A vector $x \in \mathbb{R}^n$ is **weakly majorized** by a vector $y \in \mathbb{R}^n$, denoted $x \prec_w y$, if $\sum_{i=1}^t z^i(x_1, \ldots, x_n) \leq \sum_{i=1}^t z^i(y_1, \ldots, y_n)$ for all $t \in \{1, \ldots, n\}$.*

For the purposes of this paper, we will only need the following result concerning weak majorizations:

Theorem 3.6. *For $x, y \in \mathbb{R}^n$, $x \prec_w y$ if and only if $\sum_{i=1}^n g(x_i) \leq \sum_{i=1}^n g(y_i)$ for all continuous, increasing, and convex functions $g : \mathbb{R} \to \mathbb{R}$.*

Proof. See 3.C.1.b and 4.B.2 of [12]. □

We now state and prove our variant of the $2k$ functions theorem in two pieces. First, we consider the case where $n = 1$:

Lemma 3.7. *Let $f_1, \ldots, f_k : \{0,1\} \to \mathbb{R}_{\geq 0}$ and $g : \{0,1\}^k \to \mathbb{R}_{\geq 0}$ be nonnegative real-valued functions such that g is log-supermodular. If for all $x^1, \ldots, x^k \in \{0,1\}$,*

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

then

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

Proof. For each $c \in \{0, \ldots, k\}$, define $X^c = \{(x^1, \ldots, x^k) : x^1 + \ldots + x^k = c\}$. Let $G^c \in \mathbb{R}^{\binom{k}{c}}$ be the vector obtained from by evaluating g at each element of X^c, and define F^c similarly for $f(x^1, \ldots, x^k) \triangleq \prod_{i=1}^k f_i(x^i)$.

Our strategy will be to show that $\log G^c \prec_w \log F^c$ for each c or, equivalently, that $\prod_{t=1}^T z^t(G_1^c, \ldots, G^c_{\binom{k}{c}}) \leq \prod_{t=1}^T z^t(F_1^c, \ldots, F^c_{\binom{k}{c}})$ for all $c \in \{0, \ldots, k\}$ and $T \leq \binom{k}{c}$. Then, by Theorem 3.6 and the fact that 2^x is convex and increasing, we will have

$$\sum_{(x^1, \ldots, x^k) \in X^c} g(x^1, \ldots, x^k) = \sum_{t=1}^{\binom{k}{c}} 2^{\log G_t^c} \leq \sum_{t=1}^{\binom{k}{c}} 2^{\log F_t^c} = \sum_{(x^1, \ldots, x^k) \in X^c} \prod_{i=1}^k f_i(x^i)$$

for all c. As the X^c are disjoint, this will complete the proof. We note that, by continuity arguments, this analysis holds even when some values of g and f are equal to zero.

Now, fix $c \in \{0, \ldots, k\}$ and $T \in \{1, \ldots, \binom{k}{c}\}$. Suppose $v^1, \ldots, v^T \in X^c$ are T distinct vectors. By Lemma 3.4, we must have

$$\prod_{t=1}^T g(v^t) \leq \prod_{t=1}^T g(z^t(v^1, \ldots, v^T)) \leq \prod_{t=1}^T f(w^t)$$

where $w_j^t = z^j(z^t(v^1, \ldots, v^T)_1, \ldots, z^t(v^1, \ldots, v^T)_k)$ for each $j \in \{1, \ldots, k\}$. Given any such $v^1, \ldots, v^T \in X^c$, we will show how to construct distinct vectors $\overline{v}^1, \ldots, \overline{v}^T \in X^c$ such that $\prod_{t=1}^T f(w^t) \leq \prod_{t=1}^T f(\overline{v}^t)$. Consequently, we will have

$$\prod_{t=1}^T g(v^t) \leq \prod_{t=1}^T f(\overline{v}^t) \leq \prod_{t=1}^T z^t(F_1^c, \ldots, F^c_{\binom{k}{c}}).$$

As our construction will work for any choice of distinct vectors $v^1, \ldots, v^T \in X^c$, it will work, in particular, for the T distinct vectors in X^c that maximize $\prod_{t=1}^T g(v^t)$, and the lemma will then follow as a consequence of our previous arguments.

We now describe how to construct the vectors $\overline{v}^1, \ldots, \overline{v}^T$ from the vectors $v^1, \ldots, v^T$. Let $A \in \mathbb{R}^{k\times T}$ be the matrix whose i^{th} column is given by the vector v^i. Construct $\overline{A} \in \mathbb{R}^{k\times T}$ from A by swapping the rows of A so that for each $i < j \in \{1, \ldots, k\}, \sum_p \overline{A}_{ip} \geq \sum_p \overline{A}_{jp}$. Intuitively, the first row of $\overline{A}$ corresponds to the row of A with the most nonzero elements, the second row of $\overline{A}$ corresponds to the row of A with the second largest number of nonzero elements, and so on. Let $\overline{v}^1, \ldots, \overline{v}^T$ be the columns of $\overline{A}$. Notice that $\overline{v}^1, \ldots, \overline{v}^T$ are distinct vectors in X^c and that, by construction, $z^j(z^t(\overline{v}^1, \ldots, \overline{v}^T)_1, \ldots, z^t(\overline{v}^1, \ldots, \overline{v}^T)_k) = z^t(\overline{v}^1, \ldots, \overline{v}^T)_j$ for each $j \in \{1, \ldots, k\}$ and $t \in \{1, \ldots, T\}$. Therefore, we must have

$$\prod_{t=1}^T g(\overline{v}^t) \leq \prod_{t=1}^T g(z^t(\overline{v}^1, \ldots, \overline{v}^T)) \leq \prod_{t=1}^T f(z^t(\overline{v}^1, \ldots, \overline{v}^T)) = \prod_{t=1}^T f(\overline{v}^t)$$

where the equality follows from the definition of f as a product of the f_i. In addition, the vector $z^t(v^1, \ldots, v^T)$ is simply a permuted version of the vector $z^t(\overline{v}^1, \ldots, \overline{v}^T)$ which means that their j^{th} largest elements must agree:

$$\begin{aligned} w_j^t &= z^j(z^t(v^1, \ldots, v^T)_1, \ldots, z^t(v^1, \ldots, v^T)_k) \\ &= z^j(z^t(\overline{v}^1, \ldots, \overline{v}^T)_1, \ldots, z^t(\overline{v}^1, \ldots, \overline{v}^T)_k) \\ &= z^t(\overline{v}^1, \ldots, \overline{v}^T)_j. \end{aligned}$$

Therefore,

$$\prod_{t=1}^T g(v^t) \leq \prod_{t=1}^T f(w^t) = \prod_{t=1}^T f(z^t(\overline{v}^1, \ldots, \overline{v}^T)) = \prod_{t=1}^T f(\overline{v}^t)$$

and the lemma follows as a consequence . □

Remark. In the case that $n = 1$ and $k \geq 1$, this lemma is a more general result than the $2k$ functions theorem: if $g(x^1, \ldots, x^k) = \prod_i g_i(x^i)$ for $g_1, \ldots, g_k : \{0, 1\} \to \mathbb{R}_{\geq 0}$, then g is log-supermodular.

As in the proof of the 2k functions theorem, the general theorem for $n \geq 1$ follows by induction on n.This inductive proof closely follows the inductive argument in the proof of the "four functions" theorem described in [8] with the added observation that marginals of log-supermodular functions continue to be log-supermodular.

Theorem 3.8. *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 $x^1, \ldots, x^k \in \{0, 1\}^n$,*

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

then

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

Proof. We will prove the result for general k and n by induction on n. The base case of $n = 1$ follows from Lemma 3.7. Now, for $n \geq 2$, suppose that the result holds for $k \geq 1$ and $n - 1$, and 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.

Define $f' : \{0, 1\}^{n-1} \to \mathbb{R}_{\geq 0}$ and $g' : \{0, 1\}^{k(n-1)} \to \mathbb{R}_{\geq 0}$ as

$$\begin{aligned} f_i'(y) &= f_i(y, 0) + f_i(y, 1) \\ g'(y^1, \ldots, y^k) &= \sum_{s^1,\ldots,s^k\in\{0,1\}} g(y^1, s^1, \ldots, y^k, s^k) \end{aligned}$$

Notice that g' is log-supermodular because it is the marginal of a log-supermodular function (see Lemma 3.2). If we can show that

$$g'(y^1, \dots, y^k) \le \prod_{i=1}^{k} f'_i(z^i(y^1, \dots, y^k))$$

for all $y^1, \dots, y^k \in \{0,1\}^{n-1}$, then the result will follow by induction on n. To show this, fix $\overline{y}^1, \dots, \overline{y}^k \in \{0,1\}^{n-1}$ and define $\overline{f} : \{0,1\} \to \mathbb{R}_{\ge 0}$ and $\overline{g} : \{0,1\}^k \to \mathbb{R}_{\ge 0}$ as

$$\overline{f}_i(s) = f_i(z^i(\overline{y}^1, \dots, \overline{y}^k), s)$$
$$\overline{g}(s^1, \dots, s^k) = g(\overline{y}^1, s^1, \dots, \overline{y}^k, s^k).$$

We can easily check that $\overline{g}(s^1, \dots, s^k)$ is log-supermodular and that $\overline{g}(s^1, \dots, s^k) \le \prod_{i=1}^{k} \overline{f}_i(z^i(s^1, \dots, s^k))$ for all $s^1, \dots, s^k \in \{0,1\}$. Hence, by Lemma 3.7,

$$g'(\overline{y}^1, \dots, \overline{y}^k) = \sum_{s^1, \dots, s^k} \overline{g}(s^1, \dots, s^k) \le \prod_{i=1}^{k} \sum_{s \in \{0,1\}} \overline{f}_i(s) = \prod_{i=1}^{k} f'_i(z^i(\overline{y}^1, \dots, \overline{y}^k)),$$

which completes the proof of the theorem. □

4 Graph Covers and the Partition Function

In this section, we show how to apply Theorem 3.8 in order to resolve Conjecture 2.4. In addition, we show that the theorem can be applied more generally to yield similar results for a class of functions that can be converted into log-supermodular functions by a change of variables.

4.1 Log-supermodularity and Graph Covers

The following theorem follows easily from Theorem 3.8:

Theorem 4.1. *If $f^G : \{0,1\}^n \to \mathbb{R}_{\ge 0}$ admits a log-supermodular factorization over $G = (V, \mathcal{A})$, then for any k-cover, H, of G, $Z(H) \le Z(G)^k$.*

Proof. Let H be a k-cover of G. Divide the vertices of H into k sets $S_1, \dots, S_k$ such that each set contains exactly one copy of each vertex $i \in V$. Let the assignments to the variables in the set S_i be denoted 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 $x^1, \dots, x^k$. By Lemma 3.4,

$$\prod_{i=1}^{k} \psi_\alpha(y^i_\alpha) \le \prod_{i=1}^{k} \psi_\alpha(z^i(y^1_\alpha, \dots, y^k_\alpha)) = \prod_{i=1}^{k} \psi_\alpha(z^i(x^1_\alpha, \dots, x^k_\alpha)) = \prod_{i=1}^{k} \psi_\alpha(z^i(x^1, \dots, x^k)_\alpha).$$

From this, we can conclude that $f^H(x^1, \dots, x^k) \le \prod_{i=1}^{k} f^G(z^i(x^1, \dots, x^k))$. Now, by Theorem 3.8,

$$Z(H) = \sum_{x^1, \dots, x^k} f^H(x^1, \dots, x^k) \le \prod_{i=1}^{k} \Big[\sum_{x^i} f^G(x^i) \Big] = Z(G)^k.$$

□

This theorem settles the conjecture of [2] for any log-supermodular function that admits a pairwise binary factorization, and the conjecture continues to hold for any graphical model that admits a log-supermodular factorization.

Corollary 4.2. *If $f : \{0,1\}^n \to \mathbb{R}_{\ge 0}$ admits a log-supermodular factorization over $G = (V, \mathcal{A})$, then $Z_{\mathrm{B}}(G) \le Z(G)$.*

Proof. This follows directly from Theorem 4.1 and Theorem 2.5. □

As the value of the Bethe approximation at any of the fixed points of BP is always a lower bound on $Z_{\mathrm{B}}(G)$, the conclusion of the corollary holds for any fixed point of the BP algorithm as well.

4.2 Beyond Log-supermodularity

While Theorem 4.1 is a statement only about log-supermodular functions, we can use it to infer similar results even when the function under consideration is not log-supermodular. As an example of such an application, we consider the problem of counting the number of independent sets in a given graph, $G = (V, E)$. An independent set, $I \subseteq V$, in G is a subset of the vertices such that no two adjacent vertices are in I. We define the following function:

$$I^G(x_1, \ldots, x_{|V|}) = \prod_{(i,j)\in E} (1 - x_i x_j)$$

which is equal to one if the nonzero x_i's define an independent set and zero otherwise. As every potential function depends on at most two variables, I^G factorizes over the graph $G = (V, E)$. Notice that f^G is log-submodular, not log-supermodular.

In this section, we will focus on bipartite graphs: $G = (V, E)$ is *bipartite* if we can partition the vertex set into two sets $A \subseteq V$ and $B = V \setminus A$ such that A and B are independent sets. Examples of bipartite graphs include single cycles, trees, and grid graphs. We will denote bipartite graphs as $G = (A, B, E)$.

For any bipartite graph $G = (A, B, E)$, I^G can be converted into a log-supermodular graphical model by a simple change of variables. Define $y_a = x_a$ for all $a \in A$ and $y_b = 1 - x_b$ for all $b \in B$. We then have

$$\begin{aligned} I^G(x_1, \ldots, x_{|V|}) &= \prod_{(i,j)\in E} (1 - x_i x_j) \\ &= \prod_{(a,b)\in E, a\in A, b\in B} (1 - y_a(1 - y_b)) \\ &\triangleq \overline{I}^G(y_1, \ldots, y_{|V|}). \end{aligned}$$

$\overline{I}^G$ admits a log-supermodular factorization over G and $\sum_y \overline{I}^G(y) = \sum_x I^G(x)$. Similarly, for any graph cover H of G, we have $\sum_y \overline{I}^H(y) = \sum_x I^H(x)$. Consequently, by Theorem 4.1, we can conclude that $Z(G) \geq Z_{\mathrm{B}}(G)$. Similar observations can be used to show that the Bethe partition function provides a lower bound on the true partition function for other problems that factor over pairwise bipartite graphical models (e.g., the antiferromagnetic Ising model on a grid).

5 Conclusions

While the results presented above were discussed in the case that the temperature parameter, T, was equal to one, they easily extend to any $T \geq 0$ (as exponentiation preserves log-supermodularity in this case). Hence, all of the bounds discussed above can be extended to the problem of maximizing a log-supermodular function. In particular, the inequality in Theorem 4.1 shows that the maximum value of the objective function on any graph cover is achieved by a lift of a maximizing assignment on the base graph.

This work also suggests a number of directions for future research. Related work on the Bethe approximation for permanents suggests that conjectures similar to those discussed above can be made for other classes of functions [13, 14]. While, like the "four functions" theorem, many of the above results can be extended to general distributive lattices, understanding when similar results may hold for non-binary problems may be of interest for graphical models that arise in certain application areas such as computer vision.

Acknowledgments

The author would like to thank Pascal Vontobel and Nicolas Macris for useful discussions and suggestions during the preparation of this work. This work was supported by EC grant FP7-265496, "STAMINA".

References

[1] 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.

[2] 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.

[3] M. Chertkov and V. Y. Chernyak. Loop series for discrete statistical models on graphs. *J. Stat. Mech.*, 2006.

[4] V. Gómez, J. M. Mooij, and H. J. Kappen. Truncating the loop series expansion for BP. *Journal of Machine Learning Research (JMLR)*, 2007.

[5] P. O. Vontobel. Counting in graph covers: A combinatorial characterization of the Bethe entropy function. *CoRR*, abs/1012.0065, 2010.

[6] P. O. Vontobel and R. Koetter. Graph-cover decoding and finite-length analysis of message-passing iterative decoding of LDPC codes. *CoRR*, abs/cs/0512078, 2005.

[7] R. Ahlswede and D. E. Daykin. An inequality for the weights of two families of sets, their unions and intersections. *Probability Theory and Related Fields*, 43:183–185, 1978.

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

[9] R. Aharoni and U. Keich. A generalization of the Ahlswede-Daykin inequality. *Discrete Mathematics*, 152(13):1 – 12, 1996.

[10] Y. Rinott and M. Saks. Correlation inequalities and a conjecture for permanents. *Combinatorica*, 13:269–277, 1993.

[11] Y. Rinott and M. Saks. On FKG-type and permanental inequalities. *Lecture Notes-Monograph Series*, 22:pp. 332–342, 1992.

[12] A. W. Marshall and I. Olkin. *Inequalities: Theory of Majorization and its Applications*. Academic Press, New York, 1979.

[13] P. O. Vontobel. The Bethe permanent of a non-negative matrix. In *Communication, Control, and Computing (Allerton), 2010 48th Annual Allerton Conference on*, pages 341 –346, Oct. 2010.

[14] L. Gurvits. Unleashing the power of Schrijver's permanental inequality with the help of the Bethe approximation. *ArXiv e-prints*, June 2011.