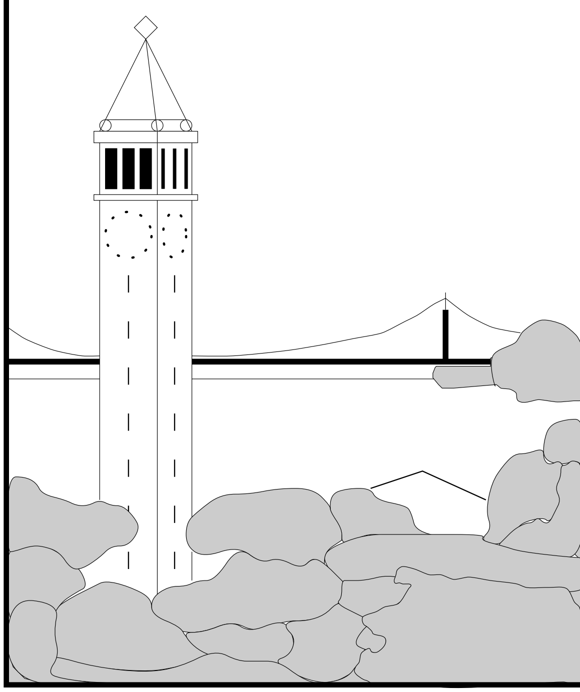
Semidefinite relaxations for approximate inference on graphs with cycles

Martin J. Wainwright and Michael I. Jordan



Report No. UCB/CSD-3-1226

January 2003

Computer Science Division (EECS)
University of California
Berkeley, California 94720

Semidefinite relaxations for approximate inference on graphs with cycles

Martin J. Wainwright,
Electrical Engineering & CS, UC Berkeley,
martinw@eecs.berkeley.edu

Michael I. Jordan
CS & Statistics Depts., UC Berkeley
jordan@cs.berkeley.edu

Abstract

We present a new method for calculating approximate marginals for probability distributions defined by graphs with cycles, based on a Gaussian entropy bound combined with a semidefinite outer bound on the marginal polytope. This combination leads to a log-determinant maximization problem that can be solved by efficient interior point methods [13]. As with the Bethe approximation and its generalizations [18], the optimizing arguments of this problem can be taken as approximations to the exact marginals. In contrast to Bethe/Kikuchi approaches, our variational problem is strictly convex and so has a unique global optimum. An additional desirable feature is that the value of the optimal solution is guaranteed to provide an upper bound on the log partition function. Such upper bounds are of interest in their own right (e.g., for parameter estimation, large deviations exponents, combinatorial enumeration). Finally, we show that taking the zero-temperature limit of our log-determinant relaxation recovers a class of well-known semidefinite relaxations for integer programming [e.g., 6].

Keywords: Graphical models; Markov random field; factor graph; approximate inference; sum-product algorithm; semidefinite constraints; determinant maximization; variational method; marginal polytope.

1 Introduction

Given a probability distribution defined by a graphical model (e.g., Markov random field, factor graph), one important problem is the computation of marginal distributions. Although highly efficient algorithms exist for trees, exact solutions are prohibitively complex for more general graphs of any substantial size. This difficulty motivates the use of algorithms for computing approximations to marginal distributions, a problem to which we refer as *approximate inference*. One widely-used algorithm is the belief propagation or sum-product algorithm [11]. As shown by Yedidia et al. [18], it can be interpreted as a method for attempting to solve a variational problem wherein the exact entropy is replaced by the Bethe approximation. Moreover, Yedidia et al. proposed extensions to the Bethe approximation based on clustering operations; such generalizations have been further explored in subsequent work by various researchers [e.g., 9, 10, 14].

An unattractive feature of the Bethe approach and its extensions is that with certain exceptions [e.g., 10, 9], the associated variational problems are typically not convex, thus leading to algorithmic complications, and also raising the possibility of multiple local optima. Secondly, in contrast to other variational methods (e.g., mean field [7]), the optimal values of Bethe-type variational problems fail to provide bounds on the log partition function. This function arises in various contexts, including approximate parameter estimation, large deviations, and combinatorial enumeration, so that such bounds are of interest in their own right.

In previous work [15], we derived a class of upper bounds on the log partition function via variational problems specified by "convexified" Bethe/Kikuchi entropy approximations. This paper introduces a new class of upper bounds based on solving a log-determinant maximization problem. Our derivation relies on a Gaussian upper bound on the discrete entropy of a suitably regularized random vector, and a semidefinite outer bound on the set of valid marginal distributions. The resulting variational problem has a unique optimum that can be found by efficient interior point methods [13]. As with the Bethe/Kikuchi approximations and sum-product algorithms, the optimizing arguments of this problem can be taken as approximations to the marginal distributions of the underlying graphical model. Moreover, taking the "zero-temperature" limit recovers a class of well-known semidefinite programming relaxations for integer programming problems [e.g., 6].

2 Problem set-up

We consider an undirected graph $G = (V, E)$ with $n = |V|$ nodes. Associated with each vertex $s \in V$ is a random variable x_s taking values in the discrete space $\mathcal{X} = \{0, 1, \ldots, m-1\}$. We let $\mathbf{x} = \{x_s \mid s \in V\}$ denote a random vector taking values in the Cartesian product space $\mathcal{X}^n$. Our analysis makes use of the following exponential representation of a graph-structured distribution $p(\mathbf{x})$. For some index set $\mathcal{I}$, we let $\boldsymbol{\phi} = \{\phi_\alpha \mid \alpha \in \mathcal{I}\}$ denote a collection of potential functions associated with the cliques of G, and let $\theta = \{\theta_\alpha \mid \alpha \in \mathcal{I}\}$ be a vector of parameters associated with these potential functions. The exponential family determined by $\boldsymbol{\phi}$ is the following collection:

$$p(\mathbf{x}; \theta) = \exp\Big\{ \sum_\alpha \theta_\alpha \phi_\alpha(\mathbf{x}) - \Phi(\theta) \Big\} \tag{1a}$$

$$\Phi(\theta) = \log \sum_{\mathbf{x} \in \mathcal{X}^n} \exp\Big\{ \sum_\alpha \theta_\alpha \phi_\alpha(\mathbf{x}) \Big\} \tag{1b}$$

Here $\Phi(\theta)$ is the *log partition function* that serves to normalize the distribution. In a *minimal* representation, the functions $\{\phi_\alpha\}$ are affinely independent, and $d = |\mathcal{I}|$ corresponds to the dimension of the family. For example, one minimal representation of a binary-valued random vector on a graph with pairwise cliques is the standard Ising model, in which $\boldsymbol{\phi} = \{x_s \mid s \in V\} \cup \{\, x_s x_t \mid (s,t) \in E\}$. Here the index set $\mathcal{I} = V \cup E$, and $d = n + |E|$. In order to incorporate higher order interactions, we simply add higher degree monomials (e.g., $x_s x_t x_u$ for a third order interaction) to the collection of potential functions. Similar representations exist for discrete processes on alphabets with $m > 2$ elements [e.g., 1].

2.1 Duality and marginal polytopes

It is well-known that Φ is convex in terms of θ, and strictly so for a minimal representation [1]. Accordingly, it is natural to consider its conjugate dual function, which is defined by the relation:

$$\Phi^*(\mu) = \sup_{\theta \in \mathbb{R}^d} \{\langle \mu,\, \theta \rangle - \Phi(\theta)\}. \tag{2}$$

Here the vector of *dual variables* μ is the same dimension as exponential parameter θ (i.e., $\mu \in \mathbb{R}^d$). It is straightforward to show that the partial derivatives of Φ with respect to θ correspond to cumulants of $\boldsymbol{\phi}(\mathbf{x})$; in particular, the first order derivatives are marginals:

$$\frac{\partial \Phi}{\partial \theta_\alpha}(\theta) = \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta) \phi_\alpha(\mathbf{x}) = \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})]. \tag{3}$$

In order to compute $\Phi^*(\widehat{\mu})$ for a given $\widehat{\mu}$, we take the derivative with respect to θ of the quantity within curly braces in equation (2). Setting this derivative to zero and making use of equation (3) yields defining conditions for the vector $\widehat{\theta}$ attaining the optimum in equation (2):

$$\widehat{\mu}_\alpha = \mathbb{E}_{\widehat{\theta}}[\phi_\alpha(\mathbf{x})] \qquad \forall\, \alpha \in \mathcal{I} \tag{4}$$

Since equation (4) involves taking an expectation, the dual variables μ have a natural interpretation as mean parameters. For example, given the standard minimal representation of the Ising model, the dual variables correspond to particular values of particular marginals (e.g., $\mathbb{E}_{\widehat{\theta}}[x_s] = p(x_s = 1; \widehat{\theta})$ when $\mathcal{X} = \{0, 1\}$.)

In order to calculate an explicit form for the conjugate dual Φ^*, we substitute the relation in equation (4) into the definition of Φ^*, thereby obtaining:

$$\Phi^*(\widehat{\mu}) \;=\; \langle \widehat{\mu},\, \widehat{\theta} \rangle - \Phi(\widehat{\theta}) \;=\; \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \widehat{\theta})\, \langle \widehat{\theta},\, \boldsymbol{\phi}(\mathbf{x}) \rangle - \Phi(\widehat{\theta}) \;=\; \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \widehat{\theta}) \log p(\mathbf{x}; \widehat{\theta}) \tag{5}$$

This relation establishes that the value of the conjugate dual $\Phi^*(\widehat{\mu})$ is given by the negative entropy of the distribution $p(\mathbf{x}; \widehat{\theta})$, where the pair $\widehat{\theta}$ and $\widehat{\mu}$ are dually coupled via equation (4). An additional consequence is that the dual parameters μ can be interpreted as *realizable* marginals; more precisely, they must belong to the set:

$$\mathrm{MARG}(G; \boldsymbol{\phi}) \;=\; \{\; \mu \in \mathbb{R}^d \;\Big|\; \sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}; \theta)\, \phi(\mathbf{x}) = \mu \;\;\text{for}\;\; \text{some} \;\; \theta \in \mathbb{R}^d \} \tag{6}$$

Note that this set is equivalent to the convex hull[1] of the finite collection of vectors $\{\boldsymbol{\phi}(\mathbf{x}) \mid \mathbf{x} \in \mathcal{X}^n\}$; consequently, the Minkowski-Weyl theorem [12] guarantees that it can be characterized by a finite number of linear inequality constraints. We refer to this set as the *marginal polytope* associated with the graph G and the potentials $\boldsymbol{\phi}$.

Since Φ is lower semi-continuous, taking the conjugate twice recovers the original function [12]; applying this fact to Φ^* and Φ, we obtain the following relation:

$$\Phi(\theta) \;=\; \max_{\mu \in \mathrm{MARG}(G; \boldsymbol{\phi})} \big\{ \langle \theta,\, \mu \rangle - \Phi^*(\mu) \big\} \tag{7}$$

Moreover, we are guaranteed that the optimum is attained uniquely at the exact marginals $\mu = \{\mu_\alpha\}$ of $p(\mathbf{x}; \theta)$. This variational formulation plays a central role in our development in the sequel.

2.2 Challenges with the variational formulation

There are two difficulties associated with the variational formulation (7). First of all, observe that the (negative) entropy Φ^*, as a function of *only* the local marginals, is implicitly defined; indeed, it is typically impossible to specify an explicit form for Φ^*. Key exceptions are trees and hypertrees, for which the entropy is well-known to decompose into a sum of local entropies defined by local marginals on the (hyper)edges [4]. Secondly, for a general graph with cycles, the marginal polytope $\mathrm{MARG}(G; \boldsymbol{\phi})$ is defined by a number of inequalities that grows rapidly in graph size [e.g., 5]. Trees and hypertrees again are important exceptions: in this case, the junction tree theorem [e.g., 4] provides a compact representation of the associated marginal polytopes.

The Bethe approach (and its generalizations) can be understood as consisting of two steps: (a) replacing the exact entropy $-\Phi^*$ with a tree (or hypertree) approximation; and (b) replacing the marginal polytope $\mathrm{MARG}(G; \boldsymbol{\phi})$ with constraint sets defined by tree (or hypertree) consistency conditions. However, since the (hyper)tree approximations used do not bound the exact entropy, the optimal values of Bethe-type variational problems do not provide a bound on the value of the log partition function $\Phi(\theta)$. Requirements for bounding Φ are both an *outer bound* on the marginal polytope, as well as an *upper bound* on the entropy $-\Phi^*$.

[1] Strictly speaking, the definition in equation (6) restricts the probability distribution specifying the convex combination to a member $p(\mathbf{x}; \theta)$ of the exponential family. However, it can be shown that (the closure of) $\mathrm{MARG}(G; \boldsymbol{\phi})$ thus defined is equivalent to the convex hull taken over all probability distributions.

3 Log-determinant relaxation

In this section, we state and prove a set of upper bounds based on the solution of a variational problem involving determinant maximization and semidefinite constraints. Although the ideas and methods described here are more generally applicable, for the sake of clarity in exposition we focus here on the case of a binary vector $\mathbf{x} \in \{-1, +1\}^n$ of "spins". It is also convenient to define all problems with respect to the complete graph K_n (i.e., fully connected). We use the standard (minimal) Ising representation for a binary problem, in terms of the potential functions $\boldsymbol{\phi} = \{x_s \mid s \in V\} \cup \{x_s x_t \mid (s,t)\}$. On the complete graph, there are $d = n + \binom{n}{2}$ such potential functions in total. Of course, any problem can be embedded into the complete graph by setting to zero a subset of the $\{\theta_{st}\}$ parameters. (In particular, for a graph $G = (V, E)$, we simply set $\theta_{st} = 0$ for all pairs $(s,t) \notin E$).

3.1 Outer bounds on the marginal polytope

We first focus on the the marginal polytope $\mathrm{MARG}(K_n)$ of valid dual variables $\{\mu_s, \mu_{st}\}$, as defined in equation (6). In this section, we describe a set of semidefinite and linear constraints that any valid dual vector $\mu \in \mathrm{MARG}(K_n)$ must satisfy.

3.1.1 Semidefinite outer bounds

Given an arbitrary vector $\mu \in \mathbb{R}^d$, consider the following $(n+1) \times (n+1)$ matrix:

$$M_1[\mu] \;\; \triangleq \;\; \begin{bmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_{n-1} & \mu_n \\ \mu_1 & 1 & \mu_{12} & \cdots & \cdots & \mu_{1n} \\ \mu_2 & \mu_{21} & 1 & \cdots & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n-1} & \vdots & \vdots & \vdots & \vdots & \mu_{n,(n-1)} \\ \mu_n & \mu_{n1} & \mu_{n2} & \cdots & \mu_{(n-1),n} & 1 \end{bmatrix} \tag{8}$$

The motivation underlying this definition is the following: suppose that the given dual vector μ actually belongs to $\mathrm{MARG}(K_n)$, in which case there exists some distribution $p(\mathbf{x};\theta)$ such that $\mu_s = \sum_{\mathbf{x}} p(\mathbf{x};\theta)\, x_s$ and $\mu_{st} = \sum_{\mathbf{x}} p(\mathbf{x};\theta)\, x_s x_t$. Thus, if $\mu \in \mathrm{MARG}(K_n)$, the matrix $M_1[\mu]$ can be interpreted as the matrix of second order moments for the vector $(1,\ \mathbf{x})$, as computed under $p(\mathbf{x};\theta)$. (Note in particular that the diagonal elements are all one, since $x_s^2 = 1$ when $x_s \in \{-1, +1\}$.) Since any such moment matrix must be positive semidefinite,[2] we have established the following:

Lemma 1 (Semidefinite outer bound). *The binary marginal polytope* $\mathrm{MARG}(K_n)$ *is contained within the semidefinite cone:*

$$\mathrm{SDEF}_1(K_n) \;\; = \;\; \left\{ \mu \in \mathbb{R}^d \;\middle|\; M_1[\mu] \succeq 0 \right\} \tag{9}$$

This semidefinite relaxation can be further strengthened by including higher order terms in the moment matrices, as described by Lasserre [e.g., 8]. For any integer $1 \le k \le n$, let $\mathcal{A}_k$ denote the set of all subsets of the vertex set with at most k elements (including the empty set). Associated with each such $\mathcal{A}_k$ is a random vector with $|\mathcal{A}_k| = \sum_{i=0}^{k} \binom{n}{i}$ elements, defined as

$$\mathbf{X}_{\mathcal{A}_k} \;\; \triangleq \;\; \left\{ \prod_{s \in S} x_s \mid S \in \mathcal{A}_k \right\}.$$

[2]To be explicit, letting $\widetilde{\mathbf{x}} = (1,\ \mathbf{x})$, then for any vector $a \in \mathbb{R}^{n+1}$, we have $a^T M_1[\mu] a = a^T \mathbb{E}[\widetilde{\mathbf{x}}\widetilde{\mathbf{x}}^T] a = \mathbb{E}[\|a^T \widetilde{\mathbf{x}}\|^2]$, which is certainly non-negative.

As a particular example of these definitions, we have $\mathcal{A}_1 = \{\emptyset, \{1\}, \ldots, \{n\}\}$, so that $\mathbf{X}_{\mathcal{A}_1}$ simply reduces[3] to $(1, \mathbf{x})$.

In analogy to $M_1[\mu]$, we then consider a $|\mathcal{A}_k| \times |\mathcal{A}_k|$ matrix $M_k[\mu]$ of second order moments associated with the vector $\mathbf{X}_{\mathcal{A}_k}$. In particular, the rows and columns of $M_k[\mu]$ are indexed by subsets of $\mathcal{A}_k$, and the (S,T) entry can be interpreted as a second order moment $\mu_{ST} = \mathbb{E}[(\prod_{s \in S} x_s)\, (\prod_{t \in T} x_t)]$. Note that for $k > 1$, the matrix $M_k[\mu]$ will involve not only the singleton and pairwise moments $\{\mu_s, \mu_{st}\}$ of $\mathbf{x}$, but also higher order moments (e.g., $\mu_{stu} = \mathbb{E}_\theta[x_s x_t x_u]$).

Example 1 (Higher-order semidefinite constraint). To provide a simple illustration, suppose that $n = 3$, so that

$$\mathbf{X}_{\mathcal{A}_2} = \{1,\ x_1,\ x_2,\ x_3,\ x_1x_2,\ x_1x_3,\ x_2x_3\}$$

In this case, the matrix $M_2[\mu]$ is 7×7, and takes the following form:

$$M_2[\mu] \quad = \quad \begin{bmatrix} 1 & \mu_1 & \mu_2 & \mu_3 & \mu_{12} & \mu_{13} & \mu_{23} \\ \mu_1 & 1 & \mu_{12} & \mu_{13} & \mu_2 & \mu_3 & \mu_{123} \\ \mu_2 & \mu_{12} & 1 & \mu_{23} & \mu_1 & \mu_{123} & \mu_3 \\ \mu_3 & \mu_{13} & \mu_{23} & 1 & \mu_{123} & \mu_1 & \mu_2 \\ \mu_{12} & \mu_2 & \mu_1 & \mu_{123} & 1 & \mu_{23} & \mu_{13} \\ \mu_{13} & \mu_3 & \mu_{123} & \mu_1 & \mu_{23} & 1 & \mu_{12} \\ \mu_{23} & \mu_{123} & \mu_3 & \mu_2 & \mu_{13} & \mu_{12} & 1 \end{bmatrix}$$

In calculating the form of $M_2[\mu] = \mathbb{E}_\theta[\mathbf{X}_{\mathcal{A}_2} \mathbf{X}_{\mathcal{A}_2}^T]$, we have frequently used the fact that $x_s^2 = 1$ whenever $x_s \in \{-1, +1\}$ in order to simplify the moment calculations. For example, in calculating the $(5,7)$ entry, we used the reduction $\mathbb{E}_\theta[(x_1x_2)\,(x_2x_3)] = \mathbb{E}_\theta[x_1x_3] = \mu_{13}$. $\square$

Now suppose that the moment vector $\mu = \{\mu_s, \mu_{st}\}$ belongs to $\mathrm{MARG}(K_n)$. In this case, there must exist a distribution $p(\mathbf{x}; \theta)$ such that $\mu_s = \mathbb{E}_\theta[x_s]$ and $\mu_{st} = \mathbb{E}_\theta[x_s x_t]$. Of course, we can also consider the higher order moments μ_{ST} for other pairs of subsets $S, T \in \mathcal{A}_k \backslash \mathcal{A}_1$, and use them (in conjunction with $\{\mu_s, \mu_{st}\}$) to form the matrix $M_k[\mu]$. This matrix must be positive semidefinite, since it is formed of moments.

With this intuition, we define for each $k \in \{1, \ldots, n\}$ a constraint set of the following form:

$$\mathrm{SDEF}_k(K_n) = \left\{ \{\mu_s, \mu_{st}\} \in \mathbb{R}^d \;\middle|\; \exists\, (\mu_{ST},\ S, T \in \mathcal{A}_k \backslash \mathcal{A}_1) \ \text{s.t}\ M_k[\mu] \succeq 0 \right\} \tag{10}$$

In words, $\mathrm{SDEF}_k(K_n)$ consists of those vectors $\{\mu_s, \mu_{st}\} \in \mathbb{R}^d$ for which there exists an extended sequence $\{\mu_{ST}, S, T \in \mathcal{A}_k \backslash\, \mathcal{A}_1\}$ of real numbers such that the matrix $M_k[\mu]$ is positive semidefinite. Note that when $k = 1$, the definition (10) agrees with our earlier specification of $\mathrm{SDEF}_1(K_n)$ in Lemma 1. Moreover, from the reasoning given above, we conclude that each $\mathrm{SDEF}_k(K_n)$ is an outer bound on the marginal polytope $\mathrm{MARG}(K_n)$. In addition, these outer bounds become tighter as k increases — viz.:

$$\mathrm{SDEF}_1(K_n) \ \supseteq\ \mathrm{SDEF}_2(K_n)\ \cdots\ \supseteq\ \mathrm{SDEF}_n(K_n) = \mathrm{MARG}(K_n). \tag{11}$$

The nested condition follows because $M_k[\mu]$ is a submatrix of the larger matrix $M_{k+1}[\mu]$ for all k.

[3]Our convention is that $\prod_{s \in \emptyset} x_s \equiv 1$.

3.1.2 Additional linear constraints

It is straightforward to augment these semidefinite constraints with additional linear constraints. In the case of a binary random vector, a large number of such constraints are known [5]. Here we focus in particular on two classes of constraints, referred to as rooted and unrooted triangle inequalities by Deza and Laurent [5], that are of especial relevance in the graphical model setting.

Pairwise edge constraints: Consider the mean parameters associated with each pair of random variables (x_s, x_t) — namely, μ_s, μ_t and μ_{st}. It is natural to require that this subset of mean parameters specify a valid pairwise marginal distribution over (x_s, x_t). Letting $\{a, b\}$ take values in $\{-1, +1\}^2$, consider the set of four linear constraints of the following form:

$$1 + a\,\mu_s + b\,\mu_t + ab\,\mu_{st} \quad \geq \quad 0. \tag{12}$$

As we show in Appendix A.1, these constraints are necessary and sufficient to guarantee the existence of a consistent pairwise marginal. Thus, there is an important connection to graphical models; in particular, by the junction tree theorem [4], this pairwise consistency guarantees that the constraints of equation (12) provide a complete description of the binary marginal polytope for any tree-structured graph. Moreover, for a general graph with cycles, they are equivalent to the tree-consistent constraint set used in the Bethe variational problem [18].

Triplet constraints: Of course, it is natural to extend local consistency to triplets $\{x_s, x_t, x_u\}$ (and even more generally, to higher order subsets). For the triplet case, consider the following set of constraints (and permutations thereof) among the pairwise mean parameters $\{\mu_{st}, \mu_{su}, \mu_{tu}\}$:

$$\mu_{st} + \mu_{su} + \mu_{tu} \quad \geq \quad -1 \tag{13a}$$

$$\mu_{st} - \mu_{su} - \mu_{tu} \quad \geq \quad -1 \tag{13b}$$

In Appendix A.2, we prove that these constraints, in conjunction with the pairwise constraints (12), are necessary and sufficient to ensure that the collection of mean parameters $\{\mu_s, \mu_t, \mu_u, \mu_{st}, \mu_{su}, \mu_{tu}\}$ uniquely determine a valid marginal over the triplet (x_s, x_t, x_u). Once again, by applying the junction tree theorem [4], we conclude that the constraints (12) and (13) provide a complete characterization of the binary marginal polytope for hypertrees of width two. It is worthwhile observing that this set of constraints is equivalent to those that are implicitly enforced by any Kikuchi approximation with clusters of size three (when applied to a binary problem).

3.2 Gaussian entropy bound

We now turn to the task of upper bounding the entropy. Our starting point is the familiar interpretation of the Gaussian as the maximum entropy distribution subject to covariance constraints [see 3]:

Lemma 2. *The (differential) entropy $h(\widetilde{\mathbf{x}})$ of any continuous random vector $\widetilde{\mathbf{x}}$ is upper bounded by the entropy $\frac{1}{2}\log\det\operatorname{cov}(\widetilde{\mathbf{x}}) + \frac{n}{2}\log(2\pi e)$ of a Gaussian with matched covariance.*

Of interest to us is the discrete entropy of a discrete-valued random vector $\mathbf{x} \in \{-1, +1\}^n$, whereas the Gaussian bound of Lemma 2 applies to the differential entropy of a continuous-valued random vector. Therefore, we need to convert our discrete vector to the continuous space. In order to do so, we define a new continuous random vector via $\widetilde{\mathbf{x}} = \frac{1}{2}\mathbf{x} + \mathbf{u}$, where $\mathbf{u}$ is a random vector

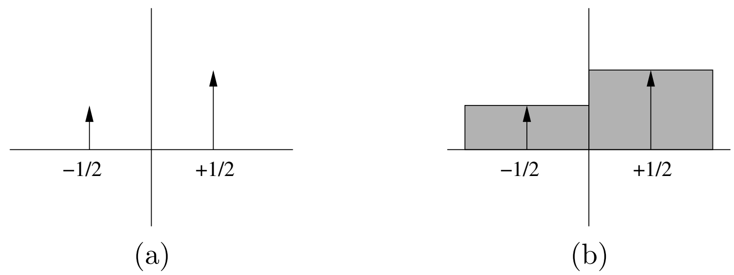


Figure 1. Illustration of the smoothing procedure. (a) Original probability mass function with impulses at $\{-\frac{1}{2}, +\frac{1}{2}\}$. (b) Transformed version, where the impulses are spread out with a uniform random variable on $[-\frac{1}{2}, \frac{1}{2}]$. By construction, the (differential) entropy of the continuous random variable in (b) is equivalent to the discrete entropy of the original one in (a).

independent of $\mathbf{x}$, with each element independently and identically distributed[4] as $u_s \sim \mathcal{U}[-\frac{1}{2}, \frac{1}{2}]$. This construction is illustrated for the scalar case in Figure 1. The motivation for rescaling $\mathbf{x}$ by $\frac{1}{2}$ is to pack the boxes as tightly together as possible.

Lemma 3. *We have $h(\widetilde{\mathbf{x}}) = H(\mathbf{x})$, where h and H denote the differential and discrete entropies of $\widetilde{\mathbf{x}}$ and $\mathbf{x}$ respectively.*

Proof. Letting $\mathcal{D} = \{\widetilde{\mathbf{x}} \in \mathbb{R}^n \mid p(\widetilde{\mathbf{x}}) > 0 \}$, then we have $h(\widetilde{\mathbf{x}}) = -\int_{\mathcal{D}} p(\widetilde{\mathbf{x}}) \log p(\widetilde{\mathbf{x}}) d\widetilde{\mathbf{x}}$. By construction, $\mathcal{D}$ can be decomposed into a disjoint union of hyperboxes $\cup_{\mathbf{e}} B(\mathbf{e})$ of unit volume, one for each configuration $\mathbf{e} \in \{-\frac{1}{2}, +\frac{1}{2}\}^n$. Accordingly, we write the differential entropy as $h(\widetilde{\mathbf{x}}) = -\sum_{\mathbf{e}} \int_{B(\mathbf{e})} p(\widetilde{\mathbf{x}}) \log p(\widetilde{\mathbf{x}}) d\widetilde{\mathbf{x}}$. Note that the quantity $p(\widetilde{\mathbf{x}}) \log p(\widetilde{\mathbf{x}})$ is equal to the constant $P(\mathbf{e}) \log P(\mathbf{e})$ over each hyperbox, where $P(\mathbf{e})$ is the probability of the discrete configuration $\mathbf{e} \in \{-\frac{1}{2}, \frac{1}{2}\}^n$. Accordingly, we can write the differential entropy as $h(\widetilde{\mathbf{x}}) = -\sum_{\mathbf{e}} P(\mathbf{e}) \log P(\mathbf{e}) \operatorname{vol}(B(\mathbf{e}))$, which is seen to be equal to $H(\mathbf{x})$ as claimed. $\square$

3.3 Log-determinant relaxation

Equipped with these building blocks, we are now ready to state and prove a log-determinant relaxation for the log partition function.

Theorem 1. *Let $\mathbf{x}$ be a random vector taking values in $\{-1, +1\}^n$, and let $\mathrm{OUT}(K_n)$ be any convex outer bound on $\mathrm{MARG}(K_n)$ that is contained within $\mathrm{SDEF}_1(K_n)$. Then the log partition function $\Phi(\theta)$ is upper bounded by the solution of the following variational problem:*

$$\Phi(\theta) \quad \leq \quad \max_{\mu \in \mathrm{OUT}(K_n)} \left\{ \langle \theta, \, \mu \rangle + \frac{1}{2} \log \det \left[M_1(\mu) + \frac{1}{3} \operatorname{blkdiag}[0, I_n] \right] \right\} + \frac{n}{2} \log(\frac{\pi e}{2}) \tag{14}$$

where $\operatorname{blkdiag}[0, I_n]$ *is a* $(n+1) \times (n+1)$ *block-diagonal matrix.*

Remarks: The inclusion $\mathrm{OUT}(K_n) \subseteq \mathrm{SDEF}_1(K_n)$ guarantees that the matrix $M_1(\mu)$ (and hence $M_1(\mu) + \frac{1}{3} \operatorname{blkdiag}[0, I_n]$) will always be positive semidefinite. Importantly, the optimization problem in equation (14) is a determinant maximization problem, for which efficient interior point methods have been developed [e.g., 13].

Proof of Theorem 1:
The proof is based on the variational representation of Φ given in equation (7). Examining this

[4]The notation $\mathcal{U}[a, b]$ denotes the uniform distribution on the interval $[a, b]$.

representation, we see that an upper bound on Φ can be obtained via an upper bound on the entropy $-\Phi^*$. For any $\mu \in \text{MARG}(K_n)$, let $\mathbf{x}$ be a random vector with these marginals. Consider the continuous-valued random vector $\widetilde{\mathbf{x}} = \frac{1}{2}\mathbf{x}+\mathbf{u}$. From Lemma 3, we have $H(\mathbf{x}) = h(\widetilde{\mathbf{x}})$; combining this equality with Lemma 2, we obtain the upper bound $H(\mathbf{x}) \leq \frac{1}{2} \log\det\text{cov}(\widetilde{\mathbf{x}}) + \frac{n}{2}\log(2\pi e)$. Since $\mathbf{x}$ and $\mathbf{u}$ are independent, we can write $\text{cov}(\widetilde{\mathbf{x}}) = \frac{1}{4}\text{cov}(\mathbf{x}) + \frac{1}{12}I_n$, where we have used the fact that $\text{cov}(\mathbf{u}) = \frac{1}{12}I_n$ for an independent uniform random vector $\mathbf{u}$ on $[-1/2,\, 1/2]$. Next we use the Schur complement formula [13] to express the log determinant as follows:

$$\begin{aligned}\log\det\text{cov}(\widetilde{\mathbf{x}}) &= \log\det\Big\{\frac{1}{4}\big[\,\text{cov}(\mathbf{x}) + \frac{1}{3}I_n\big]\Big\} \\ &= \log\det\big\{M_1[\mu] + \frac{1}{3}\,\text{blkdiag}[0, I_n]\big\} + n\log\frac{1}{4}\end{aligned} \tag{15}$$

Combining equation (15) with the Gaussian upper bound leads to the following expression:

$$H(\mathbf{x}) \;=\; -\Phi^*(\mu) \;\;\leq\;\; \frac{1}{2}\log\det\big(M_1[\mu] + \frac{1}{3}\,\text{blkdiag}[0, I_n]\big) + \frac{n}{2}\log(\frac{\pi e}{2})$$

Substituting this upper bound into the variational representation of equation (7) yields

$$\begin{aligned}\Phi(\theta) &\leq \max_{\mu\in\text{MARG}(K_n)}\Big\{\langle\theta,\,\mu\rangle + \frac{1}{2}\log\det\big[M_1(\mu) + \frac{1}{3}\,\text{blkdiag}[0, I_n]\big] + \frac{n}{2}\log(\frac{\pi e}{2})\Big\} \\ &\leq \max_{\mu\in\text{OUT}(K_n)}\Big\{\langle\theta,\,\mu\rangle + \frac{1}{2}\log\det\big[M_1(\mu) + \frac{1}{3}\,\text{blkdiag}[0, I_n]\big]\Big\} + \frac{n}{2}\log(\frac{\pi e}{2})\end{aligned}$$

where the second inequality follows because $\text{OUT}(K_n)$ is an outer bound on the marginal polytope by assumption. □

4 Experimental results

At least two aspects of Theorem 1 are of interest for applications. First of all, bounds on the log partition function are useful in various contexts (e.g., bounds on marginals, parameter estimation, combinatorial enumeration). The second aspect, and the one on which we focus here, is that the maximizing arguments $\widehat{\mu} \in \text{OUT}(K_n)$ of equation (14) can be taken as approximations to the exact marginals of the distribution $p(\mathbf{x};\theta)$. So as to test the performance of the log-determinant relaxation as an inference method, we performed extensive experiments on the complete graph (fully connected), as well as the 2-D nearest-neighbor lattice model. So as to enable comparison to the exact answer, we show here results for relatively small problems with 16 nodes.

Random problems: For any given trial, we specified the distribution $p(\mathbf{x};\theta)$ by a random choice choice of the exponential parameter vector θ in the following way. Let $\mathcal{U}[a,b]$ denote the uniform distribution on the interval $[a,b]$. For every trial shown here, we set each single node parameter $\theta_s \sim \mathcal{U}[-d_{\text{obs}}, +d_{\text{obs}}]$ independently for each node, where $d_{\text{obs}} = 0.25$. In every trial, we set the parameter θ_{st} for each edge in an IID manner, where the underlying distribution depended on the experimental condition. For a given coupling strength $d_{\text{coup}} > 0$, we investigated three possible types of coupling: (a) for *repulsive (or anti-ferromagnetic)* interactions, we set $\theta_{st} \sim \mathcal{U}[-2d_{\text{coup}}, 0]$; (b) for *mixed (or paramagnetic)* interactions, we set $\theta_{st} = \mathcal{U}[-d_{\text{coup}}, +d_{\text{coup}}]$; (c) for *attractive (or ferromagnetic)* interactions, we set $\theta_{st} = \mathcal{U}[0, 2d_{\text{coup}}]$.

Methodological specifics: Given a problem $p(\mathbf{x};\theta)$, we performed the following computations: (a) the exact marginal probability $p(x_s = 1;\theta)$ at each node; and (b) approximate marginals computed from the Bethe approximation with the sum-product algorithm [17], or (c) log-determinant approximate marginals from Theorem 1 using the outer bound $\mathrm{OUT}(K_n)$ given by the first semidefinite relaxation $\mathrm{SDEF}_1(K_n)$ in conjunction with the pairwise linear constraints in equation (12). We computed the exact marginal values either by exhaustive summation (complete graph), or by the junction tree algorithm (lattices). We used the standard parallel message-passing form of the sum-product algorithm with a damping factor[5] $\gamma = 0.05$. The log-determinant problem of Theorem 1 was solved using the SDPSOL program [16] with a MATLAB interface. For each graph (fully connected or grid), we examined a total of 6 conditions: 2 different potential strengths (weak or strong) for each of the 3 types of coupling (attractive, mixed, and repulsive). To assess the error in the approximation, we used the following L_1-based measure

$$\frac{1}{n}\sum_{s=1}^{n} |p(x_s = 1;\theta) - \widehat{\mu}_s|, \tag{16}$$

where $\widehat{\mu}_s$ was the approximate marginal computed either by SP or by LD.

Results: Table 1 shows quantitative results[6] for 100 trials performed in each of the 12 experimental conditions. The potential strength is given as the pair $(d_{\text{obs}}, d_{\text{coup}})$; note that $d_{\text{obs}} = 0.25$ in all trials. For each method, we show the sample mean plus or minus one standard deviation, the sample median, and the range (min, max) of the errors. Overall, the performance of LD is better than that of SP in terms of mean or median error. The performance of SP is slightly better in the regime of weak coupling and relatively strong observations (θ_s values); see the entries marked with * in the table. In the remaining cases, the LD method outperforms SP, and often with a large margin (particularly for examples with strong coupling). The two methods also differ substantially in the ranges of the approximation error. The SP method exhibits some instability, with the error for certain problems being larger than 0.5; for the same problems, the LD error ranges are much smaller, with a worst case maximum error over all trials and conditions of 0.13. In addition, the behavior of SP can change dramatically between the weakly coupled and strongly coupled conditions. For instance, in the attractive condition for K_{16}, the mean error changes by a factor of roughly 20, even though the coupling strength was only doubled between the weak and strong condition (0.06 and 0.12 respectively). In contrast, the error for LD remains nearly constant between the weak and strong conditions.

Figures 2 and 3 show a representative set of trials from the 6 conditions for the fully connected K_{16} and the grid, respectively. Each figure contains six panels, corresponding to the 6 experimental conditions. The top plot in each panel shows the SP approximations versus the exact marginals $p(x_s = 1;\theta)$, whereas the bottom plot compares the LD approximation to the exact answer. For the weakly coupled cases in K_{16} (top row of Figure 2), both methods perform well, and the results for these problem instances are essentially comparable. For the strongly coupled instances (bottom row), the performance of SP degrades, particularly for the repulsive and attractive cases (shown in (d) and (f) respectively). For the attractive case in (f), note that the SP method gives approximate marginals that are all very near zero, despite the fact that the true marginals are roughly equal to 0.6. This phenomenon is fairly typical: for attractive problems with sufficient coupling, SP typically

[5] More precisely, we updated messages in the log domain as $\gamma \log M_{st} + (1-\gamma)\log M_{st}$.

[6] In each case, we performed 100 trials, but discarded those trials for which SP failed to converge in the analysis. Of course, this procedure is favorable to sum-product, since its mean error invariably increased with the inclusion of non-convergent cases, whereas the LD error remained the same or decreased.

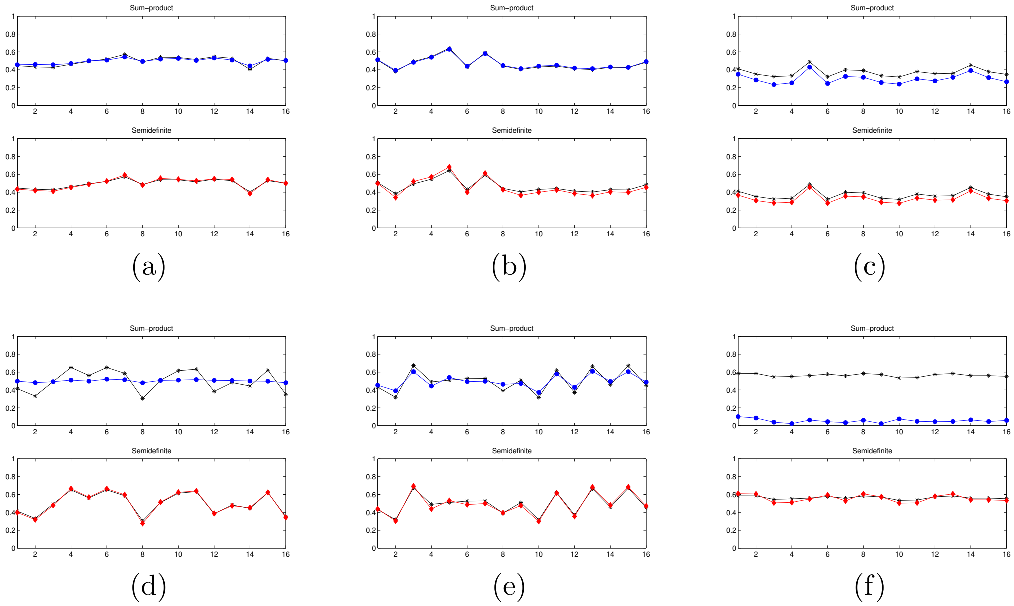


Figure 2. Specific instances of the sum-product (SP) algorithm versus the log-determinant (LD) relaxation on the complete graph K_{16}. Each panel shows the exact marginal probability $p(x_s = 1; \theta)$ at node s, versus the SP approximations (top plot) and LD approximations (bottom plot). Panels in the top and bottom rows correspond to weaker and stronger coupling, respectively. (See Table 1 for precise numbers). The panels in the left, middle, and right columns correspond to the repulsive, mixed, and attractive conditions respectively.

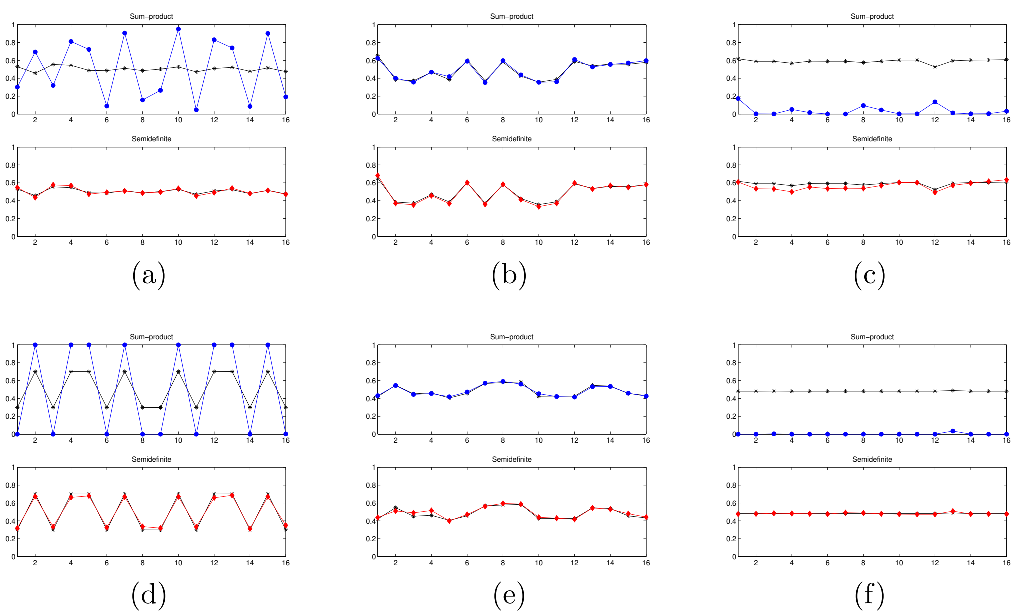


Figure 3. Specific instances of the sum-product (SP) algorithm versus the log-determinant (LD) relaxation on the four-nearest neighbor grid on 16 nodes. Each panel shows the exact marginal probability $p(x_s = 1; \theta)$ at node s, versus the SP approximations (top plot) and LD approximations (bottom plot). The panel layout is as in Figure 2.

Problem type			Method					
			Sum-product			Log-determinant		
Graph	Coupling	Strength	Mean ± std	Median	Range	Mean ± std	Median	Range
Full	Repulsive	$(0.25, 0.25)$	0.037 ± 0.015	0.035	$[0.01, 0.10]$	0.020 ± 0.005	0.020	$[0.01, 0.03]$
	Repulsive	$(0.25, 0.50)$	0.071 ± 0.032	0.066	$[0.03, 0.20]$	0.018 ± 0.005	0.017	$[0.01, 0.04]$
	Mixed*	$(0.25, 0.25)$	0.004 ± 0.005	0.003	$[0.00, 0.04]$	0.020 ± 0.005	0.019	$[0.01, 0.03]$
	Mixed	$(0.25, 0.50)$	0.055 ± 0.060	0.035	$[0.01, 0.31]$	0.021 ± 0.010	0.010	$[0.01, 0.06]$
	Attractive*	$(0.25, 0.06)$	0.024 ± 0.016	0.021	$[0.00, 0.08]$	0.027 ± 0.015	0.026	$[0.01, 0.06]$
	Attractive	$(0.25, 0.12)$	0.435 ± 0.196	0.422	$[0.08, 0.86]$	0.033 ± 0.019	0.023	$[0.01, 0.09]$
Grid	Repulsive	$(0.25, 1.0)$	0.294 ± 0.124	0.285	$[0.04, 0.59]$	0.047 ± 0.028	0.041	$[0.01, 0.12]$
	Repulsive	$(0.25, 2.0)$	0.342 ± 0.167	0.342	$[0.04, 0.78]$	0.041 ± 0.030	0.033	$[0.00, 0.12]$
	Mixed*	$(0.25, 1.0)$	0.014 ± 0.024	0.008	$[0.00, 0.20]$	0.016 ± 0.004	0.016	$[0.01, 0.02]$
	Mixed	$(0.25, 2.0)$	0.095 ± 0.111	0.053	$[0.01, 0.54]$	0.038 ± 0.024	0.032	$[0.01, 0.11]$
	Attractive	$(0.25, 1.0)$	0.440 ± 0.200	0.404	$[0.06, 0.90]$	0.047 ± 0.030	0.037	$[0.01, 0.13]$
	Attractive	$(0.25, 2.0)$	0.520 ± 0.226	0.550	$[0.06, 0.94]$	0.042 ± 0.031	0.031	$[0.00, 0.12]$

Table 1. Statistics of the L_1-approximation error for the sum-product (SP) and log-determinant (LD) methods. Shown are results for the fully connected graph K_{16}, as well as the 4-nearest neighbor grid with 16 nodes, with varying coupling and a range of potential strengths $(d_{\text{obs}}, d_{\text{coup}})$. (See text for definitions of coupling and the potential strengths.) Each experiment is nominally based on 100 trials, excluding those trials for which SP failed to converge.

outputs approximations that are either all close to zero or to one, which can be very far from the exact answer. The LD method, in contrast, remains quite accurate. Similar patterns are observed for the grid[7] in Figure 3. For the grid, both the repulsive and the ferromagnetic cases again cause a great deal of difficulty for SP, while the performance of LD remains good on these same problem instances.

With regards to computational complexity, the interior point method [13] for solving LD has a guarantee of polynomial-time complexity to solve the problem to within order ϵ. In contrast, there are no such guarantees associated with the SP algorithm; indeed, as we saw, it may even fail to converge. In practice, however, we found that the SP algorithm was typically more efficient than the LD method, particularly on the easier problems (for which it converges rapidly).

5 Connection to integer programming

The results of the previous section demonstrate that the performance of the LD method remains robust over a wide range of coupling types and strengths. Of particular interest is that (unlike SP) the performance degrades gracefully as the interaction strength is increased. The goal of this section is to understand the behavior for strong coupling in a more formal manner.

For a fixed parameter vector $\theta \in \mathbb{R}^d$, we consider the 1-parameter family of distributions $\{ p(\mathbf{x}; \theta/t) \mid t > 0 \}$. Here the parameter t plays the role of "temperature". For instance, a large

[7] Here the performance of SP appears worse overall than for the fully connected K_{16}, an anomaly due to the fact that we tested the grids with couplings that were stronger (in a relative sense) than those for K_{16}. Our primary reason was that SP failed to converge very frequently for strong coupling on K_{16}.

choice of t weakens the coupling, so that the distribution is close to uniform. Of interest to us is the opposite extreme: namely, the so-called *zero-temperature limit* $t \to 0^+$, in which the distribution concentrates all its mass on the most likely configurations. It turns out that actually taking the zero temperature limit of the log-determinant relaxation in Theorem 1 leads to well-known semidefinite relaxations for quadratic binary integer programming problems.

Proposition 1 (Integer programming). *The limiting form of the log-determinant relaxation of Theorem 1 as $t \to 0^+$ is the following semidefinite relaxation for a quadratic binary integer program:*

$$\max_{\mathbf{x} \in \{-1,+1\}^n} \big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t)} \theta_{st} x_s x_t \big\} \quad \leq \quad \max_{\mu \in \mathrm{OUT}(K_n)} \big\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t)} \theta_{st} \mu_{st} \big\}. \tag{17}$$

Proof. With reference to the LHS of equation (14), we have the well-known relation:

$$\lim_{t \to 0^+} t \, \Phi(\theta/t) \quad = \quad \max_{\mathbf{x} \in \{-1,+1\}^n} \Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t)} \theta_{st} x_s x_t \Big\}.$$

Now considering the same limit for the RHS of equation (14), it can be seen that it reduces to evaluating $\lim_{t \to 0^+} t \, \widetilde{\Phi}(\theta/t)$, where the function $\widetilde{\Phi}$ is defined for all $\theta \in \mathbb{R}^d$ as:

$$\widetilde{\Phi}(\theta) \quad = \quad \max_{\mu \in \mathrm{OUT}(K_n)} \Big\{ \langle \theta, \, \mu \rangle + \frac{1}{2} \log\det \big[M_1(\mu) + \frac{1}{3} \, \mathrm{blkdiag}[0, I_n] \big] \Big\}.$$

By definition, $\widetilde{\Phi}$ is conjugate to the proper convex function of μ defined as

$$\widetilde{\Phi}^*(\mu) = \begin{cases} -\frac{1}{2} \log\det \big[M_1(\mu) + \frac{1}{3} \, \mathrm{blkdiag}[0, I_n] \big] & \text{if} \quad \mu \in \mathrm{OUT}(K_n) \\ +\infty & \text{otherwise.} \end{cases}$$

Therefore, $\widetilde{\Phi}$ is proper and lower semi-continuous (closed), so that the quantity $\lim_{t \to 0^+} t \, \widetilde{\Phi}(\theta/t)$ is equivalent to the *recession function* of $\widetilde{\Phi}$ (see Theorem 8.5 in Rockafellar [12]). For a proper and closed convex function, the recession function of $\widetilde{\Phi}$ is given by the support function of $\mathrm{dom}\, \widetilde{\Phi}^*$ (Theorem 13.3. in Rockafellar [12]). In analytical terms, we have

$$\lim_{t \to 0^+} t \, \widetilde{\Phi}(\frac{\theta}{t}) \quad = \quad \max_{\mu \in \mathrm{OUT}(K_n)} \langle \theta, \, \mu \rangle \quad = \quad \max_{\mu \in \mathrm{OUT}(K_n)} \big\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t)} \theta_{st} \mu_{st} \big\},$$

which completes the proof. □

Note that the integer program on the LHS of equation (17) corresponds to the problem of finding the most likely configuration (i.e., maximizing $\log p(\mathbf{x}; \theta)$.) Semidefinite relaxations, such as that on the RHS of equation (17), are widely-used in combinatorial optimization; notably, Goemans and Williamson [6] have provided sharp worst-case guarantees for its performance on the MAX-CUT problem[8] in the case $\mathrm{OUT}(K_n) = \mathrm{SDEF}_1(K_n)$.

[8]The MAX-CUT problem is a special case of the integer program on the LHS of equation (17), for which $\theta_s = 0$ for all $s \in V$ and $\theta_{st} \leq 0$ for all pairs (s,t).

6 Discussion

This paper demonstrated the utility of semidefinite techniques for computing approximate marginals in graphs with cycles. In particular, we developed a method based on the combination of a Gaussian entropy bound with semidefinite constraints on the marginal polytope. The resultant log-determinant maximization problem can be solved by efficient interior point methods [13]. In experimental trials, we found that the log-determinant method was either comparable to or outperformed the sum-product algorithm, and by a substantial margin for certain problem classes.

Although this paper focused exclusively on the binary problem, the methods described here can be extended to the general multinomial case. It also remains to develop a deeper understanding of the interaction between the two choices involved in these approximations (i.e., the entropy bound, and the outer bound on the marginal polytope), as well as how to tailor approximations to particular graph structures. It is certainly possible to combine semidefinite constraints with entropy approximations (preferably convex) other than the Gaussian bound used in this paper. For instance, it would be interesting to investigate the behavior of "convexified" Bethe/Kikuchi entropy approximations [15] in conjunction with semidefinite constraints. Finally, we showed how the zero-temperature limit of the log-determinant variational problem coincides with well-known semidefinite relaxations for integer programming. One open question is whether techniques for bounding the performance of such semidefinite relaxations [e.g., 6] can be adapted to the finite temperature case.

A Linear constraints on the marginal polytope

In this appendix, we clarify the connection between the linear constraints defined in Section 3.1.2, local consistency, and graphical models.

A.1 Pairwise constraints

We first establish that the constraints (12) on $\{\mu_s, \mu_t, \mu_{st}\}$ are necessary and sufficient to guarantee the existence of a consistent pairwise marginal for (x_s, x_t). In order to do so, it is convenient to work with an alternative set of random variables $y_s = \frac{1}{2}(1 + x_s)$ that take values in $\{0, 1\}$. We use λ_s and λ_{st} to denote the associated set of mean parameters $\mathbb{E}[y_s]$ and $\mathbb{E}[y_s y_t]$, which are linked to μ by the linear relations

$$\lambda_s = \frac{1}{2}(1 + \mu_s) \tag{18a}$$

$$\lambda_{st} = \frac{1}{4}(1 + \mu_s + \mu_t + \mu_{st}) \tag{18b}$$

The $\{0, 1\}$ representation is convenient because the mean parameters λ correspond directly to particular marginal probabilities: specifically, we have $\lambda_s = p(x_s = 1)$ and $\lambda_{st} = p(x_s = 1, x_t = 1)$. Consequently, it is straightforward to see that $\{\lambda_s, \lambda_t, \lambda_{st}\}$ specify a pairwise marginal as follows:

$$p(x_s, x_t) = \begin{bmatrix} (1 + \lambda_{st} - \lambda_s - \lambda_t) & (\lambda_t - \lambda_{st}) \\ (\lambda_s - \lambda_{st}) & \lambda_{st} \end{bmatrix} \tag{19}$$

Note the sum of these four entries in this 2×2 matrix is equal to one for all λ, as it must for a joint marginal. Moreover, as a marginal probability, each entry must lie in the interval $[0, 1]$. Given the sum condition, it is necessary and sufficient to impose non-negativity for each entry:

$$\lambda_{st} \geq 0 \tag{20a}$$
$$\lambda_t - \lambda_{st} \geq 0 \tag{20b}$$
$$\lambda_s - \lambda_{st} \geq 0 \tag{20c}$$
$$1 + \lambda_{st} - \lambda_s - \lambda_t \geq 0 \tag{20d}$$

By making use of the relation (18) between λ and μ, elementary calculations show that the constraints (20) are equivalent to the four constraints specified by equation (12).

A.2 Triplet constraints

We now show how the triplet constraints (13) are necessary and sufficient to specify a valid marginal over the triplet (x_s, x_t, x_u). Once again, it is convenient to work in terms of the variables $y_s \in \{0, 1\}$. Considering the triplet entails introducing the additional mean parameter:

$$\lambda_{stu} = \mathbb{E}[y_s y_t y_u] = p(y_s = 1, y_t = 1, y_u = 1) \tag{21}$$

The full collection of seven mean parameters $\{\lambda_s, \lambda_t, \lambda_u, \lambda_{st}, \lambda_{su}, \lambda_{tu}, \lambda_{stu}\}$ suffices to specify the eight entries of the triplet marginal, where the final degree of freedom is associated with the sum constraint. More specifically, we have the following relations:

$$p(y_s = 1, y_t = 1, y_u = 1) = \lambda_{stu} \tag{22a}$$
$$p(y_s = 0, y_t = 0, y_u = 0) = 1 - \lambda_s - \lambda_t - \lambda_u + \lambda_{st} + \lambda_{su} + \lambda_{tu} - \lambda_{stu} \tag{22b}$$
$$p(y_s = 1, y_t = 1, y_u = 0) = \lambda_{st} - \lambda_{stu} \tag{22c}$$
$$p(y_s = 1, y_t = 0, y_u = 0) = \lambda_s - \lambda_{st} - \lambda_{su} + \lambda_{stu} \tag{22d}$$

Equation (22a) follows by definition, whereas equation (22b) follows by applying the inclusion-exclusion principle. Equation (22d) is derived most easily by drawing a Venn diagram. Of course, we have three copies of each of the last two equations, corresponding to the three possible positions of the zero (equation (22c)) or the one (equation (22d)).

Once again, it is clear that the sum constraint is satisfied. (Specifically, for all choices of λ, the sum of equations (22a) through (22d) is one, remembering that we have three copies of the last two equations.) Therefore, it is necessary and sufficient to force each marginal to be non-negative, which yields a set of inequality constraints for the mean parameters λ. In order to derive the triplet constraints (13), we need to project the polytope down to lower dimension by eliminating λ_{stu} from the description. In order to do so, we use Fourier-Motzkin elimination [see 2], as applied to the inequalities:

$$\lambda_{stu} \geq 0 \tag{23a}$$
$$\lambda_{stu} \geq -\lambda_s + \lambda_{st} + \lambda_{su} \tag{23b}$$
$$\lambda_{stu} \leq 1 - \lambda_s - \lambda_t - \lambda_u + \lambda_{st} + \lambda_{su} + \lambda_{tu} \tag{23c}$$
$$\lambda_{stu} \leq \lambda_{st}, \lambda_{su}, \lambda_{tu} \tag{23d}$$

Combining the ($\leq$) constraints with the ($\geq$) constraints in pairs, as specified by the Fourier-Motzkin procedure, yields the following inequalities:

$$
\begin{array}{rcll}
\lambda_{st}, \lambda_{su}, \lambda_{tu} & \geq & 0 & \text{(a) and (d)} \\
1 + \lambda_{tu} - \lambda_t - \lambda_u & \geq & 0 & \text{(b) and (c)} \\
\lambda_s - \lambda_{su} & \geq & 0 & \text{(b) and (d) with } \lambda_{st} \\
\lambda_s + \lambda_{tu} - \lambda_{su} - \lambda_{st} & \geq & 0 & \text{(b) and (d) with } \lambda_{tu} \\
1 - \lambda_s - \lambda_t - \lambda_u + \lambda_{st} + \lambda_{su} + \lambda_{tu} & \geq & 0 & \text{(a) and (c)}
\end{array}
$$

The first three sets of inequalities should be familiar; from the discussion in Appendix A.1, these constraints (and permutations thereof) guarantee validity of the three sets of pairwise marginals. (The last two inequalities, in contrast, cannot be derived by such pairwise considerations.) Finally, by using the relation (18) between λ and μ, it can be seen that the last two inequalities are equivalent to the constraints in equation (13).

Acknowledgements

Thanks to Constantine Caramanis and Laurent El Ghaoui for helpful discussions. Work funded by NSF grant IIS-9988642, ONR-MURI grant N00014-00-1-0637, and a grant from Intel Corporation.

References

[1] S. Amari. Information geometry on a hierarchy of probability distributions. *IEEE Trans. on Information Theory*, 47(5):1701–1711, 2001.

[2] D. Bertsimas and J. Tsitsikilis. *Introduction to linear optimization.* Athena Scientific, Belmont, MA, 1997.

[3] T.M. Cover and J.A. Thomas. *Elements of Information Theory.* John Wiley and Sons, New York, 1991.

[4] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probablistic Networks and Expert Systems.* Statistics for Engineering and Information Science. Springer-Verlag, 1999.

[5] M. Deza and M. Laurent. *Geometry of cuts and metric embeddings.* Springer-Verlag, New York, 1997.

[6] M. X. Goemans and D. P. Williamson. Improved approximation algorithms for maximum cut and satisfiability problems using semidefinite programming. *Int. Journal of the ACM*, 42:1115–1145, 1995.

[7] M. Jordan. *Learning in graphical models.* MIT Press, Cambridge, MA, 1999.

[8] J. B. Lasserre. Global optimization with polynomials and the problem of moments. *SIAM Journal on Optimization*, 11(3):796–817, 2001.

[9] R. J. McEliece and M. Yildirim. Belief propagation on partially ordered sets. In *Mathematical Theory of Systems and Networks*, August 2002.

[10] P. Pakzad and V. Anantharam. Iterative algorithms and free energy minimization. In *CISS*, March 2002.

[11] J. Pearl. *Probabilistic reasoning in intelligent systems.* Morgan Kaufman, San Mateo, 1988.

[12] G. Rockafellar. *Convex Analysis.* Princeton University Press, Princeton, 1970.

[13] L. Vandenberghe, S. Boyd, and S. Wu. Determinant maximization with linear matrix inequality constraints. *SIAM Journal on Matrix Analysis and Applications*, 19:499–533, 1998.

[14] M. J. Wainwright. *Stochastic processes on graphs with cycles: geometric and variational approaches.* PhD thesis, MIT, January 2002.

[15] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. A new class of upper bounds on the log partition function. In *Proc. Uncertainty in Artificial Intelligence*, volume 18, pages 536–543, August 2002.

[16] S.P. Wu and S. Boyd. SDPSOL: A parser/solver for semidefinite programs with matrix structure. In *Recent advances in LMI methods for control*, pages 79–91. SIAM, 2000.

[17] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In *NIPS 13*, pages 689–695. MIT Press, 2001.

[18] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. Technical Report TR2001-22, Mitsubishi Electric Research Labs, January 2002.