Appeared in:
Allerton Conference on Control, Communication and Computing
October 1--3, 2003

Variational inference in graphical models: The view from the marginal polytope

Martin J .Wainwright*
wainwrig@eecs.berkeley.edu

Electrical Engineering & Computer Science*
UC Berkeley, CA 94720

Michael I. Jordan†
jordan@cs.berkeley.edu

Computer Science and Statistics†
UC Berkeley, CA, 94720

Abstract

Underlying a variety of techniques for approximate inference—among them mean field, sum-product, and cluster variational methods—is a classical variational principle from statistical physics, which involves a "free energy" optimization problem over the set of all distributions. Working within the framework of exponential families, we describe an alternative view, in which the optimization takes place over the (typically) much lower-dimensional space of mean parameters. The associated constraint set consists of all mean parameters that are globally realizable; for discrete random variables, we refer to this set as a *marginal polytope*. As opposed to the classical formulation, the representation given here clarifies that there are two distinct components to variational inference algorithms: (a) an approximation to the entropy function; and (b) an approximation to the marginal polytope. This viewpoint clarifies the essential ingredients of known variational methods, and also suggests novel relaxations. Taking the "zero-temperature limit" recovers a variational representation for MAP computation as a linear program (LP) over the marginal polytope. For trees, the max-product updates are a dual method for solving this LP, which provides a variational viewpoint that unifies the sum-product and max-product algorithms.

1 Introduction

Graphical models (e.g., Markov random fields, factor graphs) play a central role in a variety of fields, including error-correcting coding, statistical physics, machine learning, and statistical image processing. For a probability distribution defined by a graphical model, important problems include computing (approximate) marginal distributions, as well as maximum a posteriori (MAP) configurations. For graphs without cycles (i.e., trees), both of these *inference problems* can be solved efficiently by recursive algorithms of a dynamic programming nature (i.e., sum-product algorithm for marginal computation; max-product or min-sum algorithm for MAP computation). These techniques generalize to hypertrees via the junction tree algorithm [3], albeit with a cost exponential in the graph treewidth. Consequently, exact inference is intractable for general graphs, which motivates the use of approximate methods.

A variety of methods for *approximate inference*, including mean field algorithms, the sum-product algorithm and cluster variational methods [e.g., 6, 14], are based on approximations of a classical variational principle from statistical physics, which entails optimizing a "free energy" functional over the space of all distributions. This variational principle is intractable primarily because of its dimensionality, which is either of exponential size (for discrete random vectors on finite spaces) or infinite (in the continuous case). In this paper, we formulate an alternative variational principle, one based on a convex optimization problem over a (typically)

Work funded by NSF grant IIS-9988642, ARO MURI DAA19-02-1-0383, and a grant from Intel Corporation.

much lower dimensional space. Working within the framework of exponential families [1, 2], we show that the computation of the log partition function as well as relevant marginal probabilities can be reduced to the solution of a single optimization problem. This alternative formulation reveals that there are two distinct components to variational techniques for approximate inference. The first requirement is an approximation to the constraint set in this optimization problem, which corresponds to the set of expected sufficient statistics or mean parameters that are realizable under some global distribution. A second challenge is presented by the objective function itself, which lacks an explicit form in general. Overall, the perspective given here clarifies the essential ingredients that underlie various known methods (e.g., mean field, sum-product, cluster variational), and also suggests new techniques for approximate inference. Our presentation here will be brief; further details can be found in [12].

2 Background

We begin with necessary background on graphical models, the classical variational principle, and exponential families.

2.1 Graphical models

Let $G = (V, E)$ be an undirected graph. For each node $s \in V$, let x_s be a random variable taking values in some sample set $\mathcal{X}_s$. In general, this set may be continuous (e.g., $\mathcal{X}_s = \mathbb{R}$), or a discrete alphabet (e.g., $\mathcal{X}_s = \{0, 1, \ldots, m-1\}$). In this paper, our primary focus is the latter (discrete) case. The random vector $\mathbf{x} := \{x_s \mid s = 1, \ldots n\}$ takes values in the Cartesian product space $\mathcal{X}_1 \times \mathcal{X}_2 \times \cdots \times \mathcal{X}_n$, which we denote by $\mathcal{X}^n$.

A graph clique C is a fully-connected subset of V. For each clique, let $\psi_C : \mathcal{X}^n \to \mathbb{R}_+$ be a compatibility function that depends only on the subvector $x_C := \{x_s, s \in C\}$ of variables associated with C. With this notation, an undirected graphical model is a probability distribution that factorizes as

$$p(\mathbf{x}) \;=\; \frac{1}{Z_p} \prod_C \psi_C(x_C), \tag{1}$$

where the product is over all cliques of the graph, and $Z_p := \sum_{\mathbf{x}} \prod_c \psi_c(\mathbf{x})$ is the normalization constant or partition function.

For future reference, we briefly describe the *classical variational principle* for discrete random vectors, on which a variety of approximate inference techniques are based [e.g., 6, 14]. Let $\mathcal{Q}$ denote the set of all distributions (i.e., mass functions) on $\mathcal{X}^n$. For any $q \in \mathcal{Q}$, let $H(q) := -\sum_{\mathbf{x} \in \mathcal{X}^n} q(\mathbf{x}) \log q(\mathbf{x})$ denote the usual (Boltzmann-Shannon) entropy. Given any distribution p represented in the form (1), it is well-known that the associated log partition function $\log Z_p$ can be recovered as the solution of the following maximum entropy problem:

$$\log Z_p \;=\; \max_{q \in \mathcal{Q}} \big\{ \sum_{\mathbf{x}} q(\mathbf{x}) [\sum_c \log \psi_c(\mathbf{x})] + H(q) \big\}. \tag{2}$$

Moreover, the maximum is uniquely attained when $q = p$. It is straightforward to see that these assertions are equivalent to the fact $\min_{q \in \mathcal{Q}} D(q \parallel p) = 0$, attained for $q = p$, where $D(q \parallel p) := \sum_{\mathbf{x}} q(\mathbf{x}) \log[q(\mathbf{x})/p(\mathbf{x})]$ is the Kullback-Leibler divergence.

2.2 Exponential families

In this section, we introduce the background on exponential families [e.g., 1, 2] necessary for subsequent development. An exponential family consists of a particular class of densities taken with respect to a base measure $\boldsymbol{\nu}$, which is defined as follows. Given some arbitrary function[1] $h : \mathcal{X}^n \rightarrow \mathbb{R}_+$, we endow $\mathcal{X}^n$ with the measure $\boldsymbol{\nu}$ defined via $d\boldsymbol{\nu} = h(\mathbf{x})\, d\mathbf{x}$, where component dx_s in the product $d\mathbf{x} = \prod_{s=1}^n dx_s$ is (a suitably restricted version of) counting measure for a discrete space, or Lebesgue measure in the continuous case.

Now let $\boldsymbol{\phi} = \{\phi_\alpha \mid \alpha \in \mathcal{I}\}$ be a collection of functions $\phi_\alpha : \mathcal{X}^n \rightarrow \mathbb{R}$, known either as *potentials* or *sufficient statistics*. Let the index set $\mathcal{I}$ have $d := |\mathcal{I}|$ elements, so that $\boldsymbol{\phi}$ itself can be viewed as a vector-valued mapping from $\mathcal{X}^n$ to $\mathbb{R}^d$. Associated with $\boldsymbol{\phi}$ is a vector $\theta = \{\, \theta_\alpha \mid \alpha \in \mathcal{I} \,\}$ of *exponential* parameters. For each fixed $\mathbf{x} \in \mathcal{X}^n$, let $\langle \theta,\, \boldsymbol{\phi}(\mathbf{x}) \rangle$ denote the (Euclidean) inner product in $\mathbb{R}^d$ of the two vectors θ and $\boldsymbol{\phi}(\mathbf{x})$. With this notation, the *exponential family* associated with $\boldsymbol{\phi}$ consists of the following parameterized collection of density functions (taken with respect to $d\boldsymbol{\nu}$):

$$p(\mathbf{x}; \theta) \;\; = \;\; \exp\big\{ \langle \theta,\, \boldsymbol{\phi}(\mathbf{x}) \rangle \;-\; A(\theta) \big\}. \tag{3}$$

The quantity A, known as the *log partition function*, is defined by the integral:

$$A(\theta) \;\; = \;\; \log \int_{\mathcal{X}^n} \exp \langle \theta,\, \boldsymbol{\phi}(\mathbf{x}) \rangle \, \boldsymbol{\nu}(d\mathbf{x}). \tag{4}$$

The exponential parameters θ of interest belong to the set $\Theta := \{\theta \in \mathbb{R}^d \mid A(\theta) < \infty\}$. We restrict our attention to the case in which the set Θ is open, corresponding to a *regular* exponential family [2]. Note that the exponential family associated with any random vector taking only a finite number of configurations is always regular, since $\Theta = \mathbb{R}^d$.

We summarize some well-known but important properties of A in the following:

Lemma 1. *(a) The function A is convex and lower semi-continuous on $\mathbb{R}^d$.*

(b) Moreover, A is differentiable on Θ, with derivatives corresponding to cumulants—in particular:

$$\frac{\partial A}{\partial \theta_\alpha}(\theta) \;\; = \;\; \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})] \; := \; \int_{\mathcal{X}^n} \phi_\alpha(\mathbf{x}) p(\mathbf{x}; \theta) \boldsymbol{\nu}(d\mathbf{x}). \tag{5}$$

An exponential family is *minimal* if there is no affine combination of the potential functions $\boldsymbol{\phi} = \{\phi_\alpha, \alpha \in \mathcal{I}\}$ that is equal to a constant $\boldsymbol{\nu}$-a.e.. Otherwise, we say that the exponential family is *overcomplete*. In a minimal representation, there is a one-to-one correspondence between exponential parameters θ and distributions $p(\mathbf{x}; \theta)$. Moreover, the log partition function is strictly convex on Θ.

Many graphical models can be formulated as exponential families; in particular, all graphical models on discrete variables can be represented as exponential family distributions. We illustrate with some simple examples:

Example 1. The *Ising model* from statistical physics constitutes a classical example of an exponential family. In this model, the base measure is the counting measure restricted to $\{0, 1\}^n$. For a given undirected graph $G = (V, E)$, consider the collection of potential functions $\boldsymbol{\phi}(\mathbf{x}) = \{x_s,\; s \in V\} \cup \{x_s x_t \mid (s, t) \in E\}$ associated with nodes and edges of the graph. The Ising model consists of the collection of densities $p(\mathbf{x}; \theta) = \exp\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t - \Phi(\theta)\}$, taken with respect to counting measure on $\{0, 1\}^n$.

[1] Here $\mathbb{R}_+ = \{y \in \mathbb{R} \mid y \geq 0\}$.

Example 2. A *parity check code* defined by a collection $\{f_c\}$ of parity checks can be viewed as an exponential family in the following way. Associated with each bit $s = 1, \ldots, n$ is the potential function $\phi_s(x_s) = x_s$. The base measure takes the form $\boldsymbol{\nu}(d\mathbf{x}) = h(\mathbf{x})d\mathbf{x}$, where $d\mathbf{x}$ is counting measure restricted to $\{0,1\}^n$ and $h(\mathbf{x}) = \prod_{c\in\mathcal{C}} f_c(\mathbf{x})$. Letting θ_s denote the likelihood associated with bit s, the overall exponential family takes the form $p(\mathbf{x};\theta) = \exp\left\{\sum_{s=1}^{n}\theta_s x_s - A(\theta)\right\} h(\mathbf{x})$.

3 Mean parameters and duality

We now turn to consideration of mean parameters, with particular emphasis on the set of realizable mean parameters, which plays a key role in the variational principle to be formulated.

3.1 Mean parameters and marginal polytopes

Given a potential function vector $\boldsymbol{\phi} : \mathcal{X}^n \to \mathbb{R}^d$, it is of interest to consider the set of realizable *expected sufficient statistics* or *mean parameters*, meaning vectors $\mu \in \mathbb{R}^d$ that arise as expectations of $\boldsymbol{\phi}$ under an arbitrary distribution that is absolutely continuous with respect to $\boldsymbol{\nu}$. With this motivation, we define the following set:

$$\mathcal{M} \;:=\; \left\{ \mu \in \mathbb{R}^d \;\middle|\; \exists\, p(\cdot) \text{ such that } \int \boldsymbol{\phi}(\mathbf{x})\; p(\mathbf{x})\boldsymbol{\nu}(d\mathbf{x}) = \mu \right\}. \tag{6}$$

Of particular interest in this paper is the case of a discrete random vector, for which the set $\mathcal{M}$ is simply the convex hull of a finite number of vectors—namely, $\{\boldsymbol{\phi}(\mathbf{x}),\ \mathbf{x} \in \mathcal{X}^n\}$. Therefore, by the Minkowski-Weyl theorem [8], the set has an equivalent representation of the form

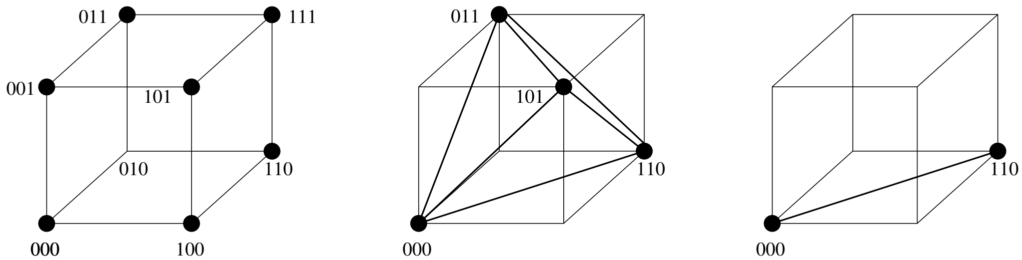
$$\mathcal{M} \;=\; \{\, \mu \in \mathbb{R}^d \mid \langle a_j,\, \mu\rangle \leq b_j \quad \forall\, j \in \mathcal{J}\}. \tag{7}$$

Since any convex set can be represented as the intersection of half-spaces containing it [8], the crucial part of this representation is that the index set $\mathcal{J}$ in equation (7) must be finite. For discrete random vectors, we use $\mathrm{MARG}(G;\boldsymbol{\phi})$ (or the shorter notation $\mathrm{MARG}(G)$) to denote the *marginal polytope* defined by the potential functions $\boldsymbol{\phi}$ associated with the G. The extreme points of $\mathrm{MARG}(G)$ are of the form $\mu_{\mathbf{e}} := \boldsymbol{\phi}(\mathbf{e})$, and realized by the distributions $\delta_{\mathbf{e}}(\mathbf{x})$, which are equal to one if $\mathbf{x} = \mathbf{e}$ and 0 otherwise, where $\mathbf{e} \in \mathcal{X}^n$ is a given configuration.

To illustrate the notions of mean parameters and marginal polytopes, we continue with the examples of the previous section.

Example 3. The mean parameters associated with the Ising model of Example 1 are the single node marginal probabilities $\mu_s = \mathbb{E}_\theta[x_s] = p(x_s = 1;\theta)$ for each $s \in V$, and the joint marginal probability $\mu_{st} = \mathbb{E}_\theta[x_s x_t] = p(x_s = 1, x_t = 1;\theta)$ for each $(s,t) \in E$. The marginal polytope consists of all vectors $\mu = \{\mu_s, s \in V\} \cup \{\mu_{st}, (s,t) \in E\}$ of marginal probabilities that are globally realizable by some distribution. This particular case of a marginal polytope has been extensively studied in the combinatorial optimization literature [e.g., 4].

Example 4. We now consider the marginal polytopes associated with some simple binary parity check codes. Building on Example 2, a code of length $n = 3$ can be represented as an exponential family in terms of the potential functions $\boldsymbol{\phi}(\mathbf{x}) = \{x_1, x_2, x_3\}$. For a given binary code, the marginal polytope corresponds to the convex hull of all possible codewords, which we refer to as a codeword polytope [5]. Three different examples are shown in Figure 1.



(a) Uncoded (b) One parity check (c) Two parity checks

Figure 1. Marginal polytopes associated with some simple parity check codes of length $n = 3$. (a) With no parity checks, the codeword polytope is simply the unit cube $[0, 1]^3$. (b) Introducing a single parity check eliminates the odd parity vertices, resulting in a tetrahedron. (c) When a second parity check (say constraining $x_1 = x_2$) is added, then the codeword polytope is simply a line, lying entirely within the lower-dimensional box $[0, 1]^2$.

3.2 Legendre mapping

Given an arbitrary member of the exponential family defined by $\boldsymbol{\phi}$, we can define a mapping $\Lambda : \Theta \to \mathcal{M}$ as follows:

$$\Lambda(\theta) \;\; := \;\; \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})] \;=\; \int_{\mathcal{X}^n} \boldsymbol{\phi}(\mathbf{x})\, p(\mathbf{x}; \theta)\, \boldsymbol{\nu}(d\mathbf{x}). \tag{8}$$

The mapping Λ associates to each $\theta \in \Theta$ a vector of mean parameters $\mu := \Lambda(\theta)$ belonging to the set $\mathcal{M}$. The following two issues are of interest: (1) determining when Λ to one-to-one and hence invertible on its image, and (2) characterizing the image of Θ under the mapping Λ. We begin with a result [2] that addresses the first question:

Proposition 1. *The mapping Λ is one-to-one if and only if the exponential representation is minimal.*

For an overcomplete representation, the inverse image $\Lambda^{-1}(\mu) := \{\theta \in \Theta \mid \Lambda(\theta) = \mu\}$, rather than being a singleton (as it would be for an invertible mapping), is a (non-trivial) affine subset of Θ. In general, although there is no longer a bijection between Θ and $\Lambda(\Theta)$ in an overcomplete representation, there is still a bijection between each element of $\Lambda(\Theta)$ and an affine subset of Θ (see [12]).

We now turn to the second question regarding the range of Λ. The relative interior (denoted ri) of a convex set is its interior taken relative to its affine hull; it coincides with the usual interior for a full-dimensional convex set. A key fact is that any convex set has a non-empty relative interior [8]. With this notion, we have the following result (see [2, 12] for a proof):

Theorem 1. *The mean parameter mapping Λ is onto the (relative) interior of $\mathcal{M}$.*

Typically, the exponential family $\{p(\mathbf{x}; \theta) \mid \theta \in \Theta\}$ describes only a strict subset of all possible densities, whereas the definition (6) of $\mathcal{M}$ allows the density $p(\cdot)$ to be arbitrary. The significance of Theorem 1 lies in the fact that any $\mu \in \mathrm{ri}\,\mathcal{M}$ can be realized by an exponential family member. For a minimal exponential family, Proposition 1 guarantees that there is a *unique* exponential parameter $\theta(\mu)$ such that $\Lambda(\theta(\mu)) = \mu$. However, whenever the exponential family describes a strict subset of all densities (as in most cases of interest), then there exists at least some other density $p(\cdot)$—albeit not a member of the exponential family—that also realizes μ (i.e., for which $\int \boldsymbol{\phi}(\mathbf{x}) p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x}) = \mu$). The distinguishing feature of $p(\mathbf{x}; \theta(\mu))$ lies in its characterization in terms of maximum entropy, as we now discuss.

3.3 Fenchel-Legendre conjugate

We now turn to consideration of the Fenchel-Legendre conjugate [8] of the log partition function A. In particular, this conjugate dual function, which we denote by A^*, is defined by the following optimization problem:

$$A^*(\mu) \quad := \quad \sup_{\theta \in \Theta} \{\langle \mu, \theta \rangle - A(\theta)\}. \tag{9}$$

Our re-use of the notation $\mu \in \mathbb{R}^d$ is deliberate, in that these dual variables turn out to have a natural interpretation as mean parameters.

The (Boltzmann-Shannon) entropy of the exponential family member $p(\mathbf{x};\theta)$ with respect to $\boldsymbol{\nu}$ is defined as $H(p(\mathbf{x};\theta)) = -\int_{\mathcal{X}^n} p(\mathbf{x};\theta) \log [p(\mathbf{x};\theta)]\, \boldsymbol{\nu}(d\mathbf{x})$. One assertion of the following theorem is that when $\mu \in \operatorname{ri} \mathcal{M}$, then the value of the dual function $A^*(\mu)$ is precisely the negative entropy of $p(\mathbf{x};\theta(\mu))$, where $\theta(\mu)$ is an element of the inverse image $\Lambda^{-1}(\mu)$. Of equal importance is the fact that the value of the dual function is $+\infty$ for $\mu \notin \operatorname{cl} \mathcal{M}$. More formally, we state and prove the following:

Theorem 2.
(a) For any $\mu \in \operatorname{ri} \mathcal{M}$, let $\theta(\mu)$ denote a member of $\Lambda^{-1}(\mu)$. The Fenchel-Legendre dual of A has the following form:

$$A^*(\mu) \quad = \quad \begin{cases} -H(p(\mathbf{x};\theta(\mu))) & \text{if } \mu \in \operatorname{ri} \mathcal{M} \\ +\infty & \text{if } \mu \notin \operatorname{cl} \mathcal{M}. \end{cases} \tag{10}$$

For any boundary point $\mu \in \operatorname{bd} \mathcal{M} := \operatorname{cl} \mathcal{M} \backslash \operatorname{ri} \mathcal{M}$, we have $A^*(\mu) = \lim_{n \to +\infty} [-H(p(\mathbf{x};\theta(\mu^n)))]$, taken over a sequence $\{\mu^n\} \subset \operatorname{ri} \mathcal{M}$ converging to μ.

(b) In terms of this dual, the log partition function has the following variational representation:

$$A(\theta) \quad = \quad \sup_{\mu \in \mathcal{M}} \big\{\langle \theta, \mu \rangle - A^*(\mu)\big\}. \tag{11}$$

Proof. (a) Case (i) $\mu \in \operatorname{ri} \mathcal{M}$: In this case, Theorem 1 guarantees that the inverse image $\Lambda^{-1}(\mu)$ is non-empty. Any point in this inverse image attains the supremum in equation (9). In a minimal representation, there is only one optimizing point, whereas there is an affine subset for an overcomplete representation. Nonetheless, for any $\theta(\mu) \in \Lambda^{-1}(\mu)$, the value of the optimum is $A^*(\mu) = \langle \theta(\mu), \mu \rangle - A(\theta(\mu))$. We conclude by observing that

$$-H(p(\mathbf{x};\theta(\mu))) \quad = \quad \mathbb{E}_\theta[\langle \theta(\mu), \boldsymbol{\phi}(\mathbf{x}) \rangle - A(\theta(\mu))] \quad = \quad \langle \theta(\mu), \mu \rangle - A(\theta(\mu)).$$

Case (ii) $\mu \notin \operatorname{cl} \mathcal{M}$: Let $\operatorname{dom} A^* = \{\mu \in \mathbb{R}^d \mid A^*(\mu) < +\infty\}$ denote the effective domain of A^*. With this notation, we must prove that $\operatorname{cl} \mathcal{M} \supseteq \operatorname{dom} A^*$. From Lemma 1, the function A is lower semi-continuous; moreover, it can be shown [2, 12] that A is also essentially smooth [8]. From Theorem 1, we have $\nabla A(\Theta) = \operatorname{ri} \mathcal{M}$. By Corollary 26.4.1 of Rockafellar [8], these conditions guarantee that $\operatorname{ri} \operatorname{dom} A^* \subseteq \operatorname{ri} \mathcal{M} \subseteq \operatorname{dom} A^*$. Since both $\mathcal{M}$ and $\operatorname{dom} A^*$ are convex sets, taking closures in the these inclusions yields that $\operatorname{cl} \operatorname{dom} A^* = \operatorname{cl} \operatorname{ri} \mathcal{M} = \operatorname{cl} \mathcal{M}$, where the second equality follows by the convexity of $\mathcal{M}$. Therefore, by definition of the effective domain, $A^*(\mu) = +\infty$ for any $\mu \notin \operatorname{cl} \mathcal{M}$.

Case (iii) $\mu \in \operatorname{cl} \mathcal{M} \backslash \operatorname{ri} \mathcal{M}$: Since A^* is defined as a conjugate function, it is lower semi-continuous. Therefore, the value of $A^*(\mu)$ for any boundary point $\mu \in \operatorname{cl} \mathcal{M} \backslash \operatorname{ri} \mathcal{M}$ is determined by the limit over a sequence approaching μ from inside $\operatorname{ri} \mathcal{M}$, as claimed.

(b) From Lemma 1(a), A is lower semi-continuous, which ensures that $(A^*)^* = A$ so that we

can write $A(\theta) = \sup_{\mu \in \mathrm{cl}\,\mathrm{dom}\,A^*}\{\langle \theta,\, \mu\rangle - A^*(\mu)\}$. Part (a) shows that $\mathrm{cl}\,\mathrm{dom}\,A^* = \mathrm{cl}\,\mathcal{M}$, so that equation (11) follows. Whether the supremum is taken over $\mathcal{M}$ or over $\mathrm{cl}\,\mathcal{M}$ is inconsequential.
$\square$

4 Variational methods for inference

For the purposes of approximate inference, the key part of Theorem 2 is equation (11), which is a variational representation in two distinct senses. First, as with the classical variational principle (2), it specifies the log partition function as the solution of an optimization problem; the crucial difference is that the optimization takes place over the lower-dimensional space of mean parameters, as opposed to the space of all distributions. Second, in contrast to the classical principle, a single optimization problem provides a variational procedure for computing all mean parameters, as stated formally in the following:

Proposition 2. *For all $\theta \in \Theta$, the supremum in equation* (11) *is attained uniquely at the vector $\mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})]$. This statement holds in both minimal and overcomplete representations.*

As a consequence of this result, an additional by-product of solving problem (11), apart from computing the log partition function, is the set of the mean parameters $\mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})]$ associated with $p(\mathbf{x};\theta)$. One might think that the problem of computing mean parameters is now solved, since it has been "reduced" to a convex optimization problem. Indeed, for simple scalar examples (e.g., Bernoulli, Poisson, Gaussian), the variational problem (11) is easily solved. For general multivariate exponential families, in contrast, there are two primary challenges associated with the variational representation (11). Here we limit our discussion of these issues to the case of discrete random variables.

Nature of marginal polytopes: Recall from equation (7) that any marginal polytope can be characterized by a finite number of halfplane constraints. The difficulty is that for general graphs, this number can grow very quickly with increasing problem size. As might be expected, the nature of a marginal polytope $\mathrm{MARG}(G)$ depends critically on the structure of the underlying graph. As a concrete example, the marginal polytope for the Ising model on the complete graph with $n = 7$ nodes is known to have more than 2×10^8 facets [4]. In sharp contrast, for trees (and more generally, for hypertrees), the junction tree theorem [3] guarantees that the number of facets grows only linearly in the number of nodes.

Nature of dual function: From equation (9), the dual function A^* is itself defined in a variational manner, and hence typically lacks an explicit form. More specifically, it is defined implicitly via the composition of two functions: (i) first compute an exponential parameter $\theta(\mu)$ in the inverse image $\Lambda^{-1}(\mu)$; and then (ii) compute the negative entropy of the distribution $p(\mathbf{x};\theta(\mu))$. In general, computing the inverse map Λ^{-1} is as difficult as performing inference (which corresponds to computing the forward map); moreover, even if it were possible to compute an element $\theta(\mu) \in \Lambda^{-1}(\mu)$, computing the usual entropy would be intractable. Again, tree and hypertree-structured distributions are important exceptions to this rule, for which A^* has an explicit form.

As we now discuss, various methods for approximate inference can be differentiated in terms of how they circumvent the challenges imposed by the nature of marginal polytopes and the dual function.

4.1 Mean field approach

Within the current framework, the mean field approach to approximate inference can be understood in the following way. For any $\mu \in \mathrm{MARG}(G)$, the variational representation (11) immediately yields the lower bound $A(\theta) \geq \langle \mu, \, \theta \rangle - A^*(\mu)$. Unfortunately, for an arbitrary μ, this lower bound cannot be evaluated, given the implicit nature of the dual function A^*. It can, however, be evaluated for those subsets of $\mathrm{MARG}(G)$ that arise from "tractable" graphical models, including fully-factorized (i.e., product) distributions as well as tree-structured distributions, for which the dual function A^* has an explicit form in terms of mean parameters. For a given tractable subgraph H of G, let $\mathcal{E}(H)$ denote the set of exponential parameters that respect its structure (i.e., non-zero entries are permitted only for entries associated with the cliques of H, which defines an e-flat manifold [1].). Consider the set of mean parameters realized by such tractable distributions

$$\mathrm{TRACT}(G; H) \quad = \quad \{\mu \in \mathbb{R}^d \, | \mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})] \quad \text{for some } \theta \in \mathcal{E}(H)\}. \tag{12}$$

Mean field seeks to solve the optimization problem $\max_{\mu \in \mathrm{TRACT}(G;H)}\{\langle \theta, \, \mu \rangle - A^*(\mu)\}$, which amounts to finding the tightest lower bound, subject to the constraint that μ can be realized by a tractable distribution. By definition, the set $\mathrm{TRACT}(G; H)$ is an inner approximation to $\mathrm{MARG}(G)$, and it is a non-convex set under mild restrictions [see 12]. Figure 2(a) provides an idealized illustration of $\mathrm{TRACT}(G; H)$, and its relation to the exact marginal polytope.

4.2 Bethe approximation and sum-product:

As established by Yedidia et al. [14], the sum-product algorithm is based on the so-called Bethe variational problem, which can be understood as follows within the current framework. For a tree-structured problem, the dual function A^* has an explicit representation as a sum of single node entropy and edgewise mutual information terms, defined as follows:

$$H_s(\mu_s) \quad := \quad -\sum_{x_s} \mu_s(x_s) \log \mu_s(x_s), \tag{13a}$$

$$I_{st}(\mu_{st}) \quad := \quad \sum_{x_s, x_t} \mu_{st}(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \quad = \quad H_s(\mu_s) + H_t(\mu_t) - H_{st}(\mu_{st}). \tag{13b}$$

In general, the entropy of a distribution defined by a graph with cycles will not decompose in this additive manner; however, making this assumption leads to the *Bethe approximation* $H_{Bethe}(\mu) := \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st})$ of the entropy on a graph with cycles. (Although this form of the Bethe approximation differs from that used by Yedidia et al. [14], the two forms are equivalent on the constraint set $\mathrm{LOCAL}(G)$, defined below.)

Although an exact characterization of the marginal polytope $\mathrm{MARG}(G)$ is intractable, it is possible to provide a subset of *necessary* constraints. Specifically, it is clear that any set $\tau_s(x_s)$ and $\tau_{st}(x_s, x_t)$ of candidate marginal distributions must belong to the polytope

$$\mathrm{LOCAL}(G) \quad = \quad \{\, \tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \;\; \sum_{x_s} \tau_{st}(x_s, x_t) = \tau_t(x_t) \,\}.$$

For a tree T, the junction tree theorem [3] ensures that $\mathrm{LOCAL}(T)$ is an exact description of the marginal polytope; for a graph with cycles, in contrast, it is a convex outer bound, as illustrated in Figure 2(b). By construction, fixed points of the sum-product algorithm (i.e., stationary points of the Bethe variational problem [14]) belong to $\mathrm{LOCAL}(G)$. Since the exact

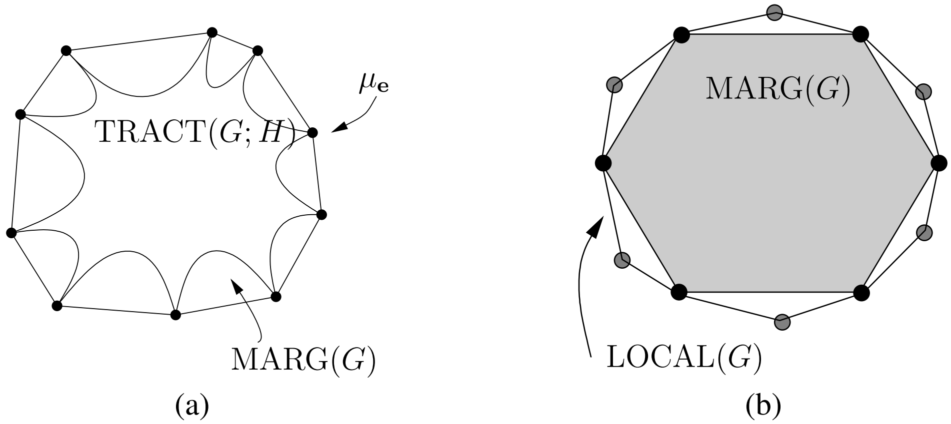


Figure 2. (a) Mean field theory uses a *non-convex inner approximation* $\mathrm{TRACT}(G; H)$ to the marginal polytope $\mathrm{MARG}(G)$. The circles correspond to mean parameters that arise from delta distributions, and belong to both $\mathrm{MARG}(G)$ and $\mathrm{TRACT}(G; H)$. (b) The Bethe approximation uses a tree-based *convex outer approximation* $\mathrm{LOCAL}(G)$ to the marginal polytope. We refer to points in $\mathrm{LOCAL}(G)$ as *pseudomarginals* since they are not necessarily the marginals of some probability distribution.

marginals belong to $\mathrm{MARG}(G)$, it is natural to ask whether sum-product fixed points ever fall in the region $\mathrm{LOCAL}(G) \setminus \mathrm{MARG}(G)$. In fact, it can be shown [11] that the sum-product algorithm, viewed as a multi-function from $\mathbb{R}^d$ to $\mathrm{LOCAL}(G)$, is onto the relative interior of $\mathrm{LOCAL}(G)$. As a consequence, for any $\tau \in \mathrm{ri}\,\mathrm{LOCAL}(G)$, there exists a problem $p(\mathbf{x}; \theta)$ for which τ is a sum-product fixed point, as illustrated in the following.

Example 5 (Globally inconsistent fixed point). We illustrate using a binary random vector on the simplest possible graph for which sum-product is not exact—namely, a single cycle graph $G = (V, E)$ with three nodes. Consider candidate marginal distributions $\{\tau_s, \tau_{st}\}$ of the form

$$\tau_s := \begin{bmatrix} 0.5 & 0.5 \end{bmatrix} \ \forall s \in V, \qquad \tau_{st} := \begin{bmatrix} \beta_{st} & 0.5 - \beta_{st} \\ 0.5 - \beta_{st} & \beta_{st} \end{bmatrix} \ \forall (s,t) \in E, \tag{14}$$

where $\beta_{st} \in [0, 0.5]$ is a parameter to be specified independently for each edge (s, t). It is straightforward to verify that $\{\tau_s, \tau_{st}\}$ belong to $\mathrm{LOCAL}(G)$ for any choice of $\beta_{st} \in [0, 0.5]$.

Suppose that we choose $\beta_{12} = \beta_{23} = 0.4$, but then set $\beta_{13} = 0.1$. Observe that this choice induces strong (positive) dependencies between the pairs of variables (x_1, x_2) and (x_2, x_3), whereas the pair (x_1, x_3) can only share the same value with low probability. The pseudomarginal τ thus constructed is not a member of $\mathrm{MARG}(G)$, as can be shown by contradiction, or more directly using semidefinite constraints, as discussed in the following section. Finally, to construct a problem $p(\mathbf{x}; \theta)$ for which τ is a fixed point of the sum-product algorithm, we define $\theta_s(x_s) = \log \tau_s(x_s)$ and $\theta_{st}(x_s, x_t) = \log \tau_{st}(x_s, x_t)/[\tau_s(x_s)\tau_t(x_t)]$. It can be verified that $\tau \in \mathrm{LOCAL}(G) \setminus \mathrm{MARG}(G)$ is a fixed point of the sum-product algorithm, when applied to $p(\mathbf{x}; \theta)$ with the messages M_{st} initialized to all ones. $\diamondsuit$

4.3 Semidefinite constraints and log-determinant relaxation

Since the sum-product algorithm can lead to globally inconsistent marginals, a natural approach is to add constraints to $\mathrm{LOCAL}(G)$, thereby obtaining a tighter outer bound on $\mathrm{LOCAL}(G)$. Indeed, one component of the Kikuchi and other cluster variational methods [14] is such a tighter outer bound, based on additional linear constraints. In recent work [13], we have demonstrated an alternative approach to approximate inference, based on semidefinite constraints. For concreteness, we focus on the Ising model of Example 1 defined on the complete graph K_n. Given an arbitrary vector $\mu \in \mathbb{R}^d$, consider the symmetric $(n+1) \times (n+1)$ matrix with entries defined as follows:

$$M_1[\mu]_{\alpha\beta} \;\; := \;\; \begin{cases} 1 & \text{if } \alpha = \beta = 1 \\ \mu_s & \text{if } (\alpha, \beta) = (s+1, 1) \text{ or } (\alpha, \beta) = (s+1, s+1) \\ \mu_{st} & \text{if } \alpha = s+1 \text{ and } \beta = t+1 \end{cases} \tag{15}$$

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 $\mu_{st} = \sum_{\mathbf{x}} p(\mathbf{x}; \theta) x_s x_t$. Consequently, the matrix $M_1[\mu]$ can be interpreted as the second order moment matrix for $(1,\; \mathbf{x})$, as computed under $p(\mathbf{x}; \theta)$. Since any such moment matrix must be positive semidefinite, the binary marginal polytope $\mathrm{MARG}(K_n)$ must be contained within the set $\mathrm{SDEF}_1(K_n) := \{\; \mu \in \mathbb{R}^d \;\mid\; M_1[\mu] \;\succeq\; 0 \;\}$. This semidefinite constraint can be further strengthened by including higher order terms in the moment matrices [7]. Combining semidefinite constraints with a Gaussian-based upper bound on the entropy leads to the following convex relaxation for inference [13]:

Theorem 3. *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 is upper bounded as follows:*

$$A(\theta) \;\; \leq \;\; \max_{\mu \in \mathrm{OUT}(K_n)} \left\{ \langle \theta,\, \mu \rangle + \frac{1}{2} \log\det \left[M_1(\mu) + \frac{1}{12}\, \mathrm{blkdiag}[0, I_n] \right] \right\} + \frac{n}{2} \log(2\pi e) \tag{16}$$

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

Importantly, the optimization problem in equation (16) is a (strictly concave) determinant maximization problem, for which efficient interior point methods have been developed [9]. For certain problem classes, this relaxation outperforms the sum-product algorithm by a considerable margin [13].

4.4 Zero temperature limit and MAP computation

Finally, we consider the link between the variational representation of Theorem 2(b), and the problem of MAP estimation for a discrete random vector (i.e., computing an element $\mathbf{x}^*$ of the set $\arg\max_{\mathbf{x}} p(\mathbf{x}; \theta)$, or equivalently of $\arg\max_{\mathbf{x}} \langle \theta,\, \boldsymbol{\phi}(\mathbf{x}) \rangle$). To provide intuition for the result to follow, consider the family of scaled distributions $\{p(\mathbf{x}; \beta\theta) \;\mid\; \beta > 0\}$. For each $\beta > 0$, equation (11) yields $A(\beta\theta) = \max_{\mu \in \mathrm{MARG}(G)} \left\{ \langle \beta\theta,\, \mu \rangle - A^*(\mu) \right\}$. As $\beta \to +\infty$, the distribution $p(\mathbf{x}; \beta\theta)$ should place increasing amounts of mass on configurations $\mathbf{x}^*$ in the set $\arg\max_{\mathbf{x}} \langle \theta,\, \boldsymbol{\phi}(\mathbf{x}) \rangle$. This type of rescaling, which equivalent to the "zero-temperature limit" of statistical physics, suggests that the limiting behavior of the variational representation should have a close connection to the problem of MAP estimation. More formally, we have the following result (see [12] for a proof):

Theorem 4. *The problem of MAP computation has the following alternative representations:*

$$\max_{\mathbf{x} \in \mathcal{X}^n} \langle \theta,\, \boldsymbol{\phi}(\mathbf{x}) \rangle \;\; \stackrel{(a)}{=} \;\; \lim_{\beta \to +\infty} \frac{A(\beta\theta)}{\beta} \;\; \stackrel{(b)}{=} \;\; \max_{\mu \in \mathrm{MARG}(G)} \langle \theta,\, \mu \rangle \tag{17}$$

The final expression shows that MAP estimation is equivalent to a linear program (LP) over the marginal polytope $\mathrm{MARG}(G)$, which forms the basis of LP relaxations for integer programming [e.g., 4, 5]. Given our derivation from a zero-temperature limit, it is natural to conjecture a link between this LP and the max-product algorithm. In fact, for trees (for which $\mathrm{MARG}(T) = \mathrm{LOCAL}(T)$), it can be shown [10, 12] that the max-product algorithm is a dual method for solving the LP $\max_{\mu \in \mathrm{LOCAL}(T)} \langle \theta,\, \mu \rangle$. For a graph with cycles, the analogous statement *fails* to hold, since fixed points of max-product can specify an incorrect (i.e., non-MAP optimal) configuration. However, a tree-reweighted variant of the max-product algorithm [10] has similar properties for graphs with cycles.

5 Discussion

Working within the framework of exponential families, we described an alternative view of variational inference. In contrast to the classical principle, the formulation given here is entirely in terms of mean parameters, which leads a low-dimensional convex optimization problem. This perspective clarifies two distinct ingredients that underlie variational methods: approximations to marginal polytopes, and approximations to the entropy (as a function only of mean parameters). This representation also suggests novel convex relaxations for inference, and shows how to obtain upper bounds on the log partition function. Among other open questions, it remains to explore the relative roles of the approximations to the marginal polytope and entropy functions in controlling the accuracy of variational methods for approximate inference.

References

[1] S. Amari and H. Nagaoka. *Methods of information geometry*. AMS, Providence, RI, 2000.

[2] L. Brown. *Fundamentals of statistical exponential families*. Institute of Mathematical Statistics, Hayward, CA, 1986.

[3] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter. *Probabilistic networks and expert systems*. Statistics for Engineering and Information Science. Springer-Verlag, 1999.

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

[5] J. Feldman, D. R. Karger, and M. J. Wainwright. Using linear programming to decode LDPC codes. In *Conference on Information Science and Systems*, March 2003.

[6] M. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In *Learning in graphical models*, pages 105–161. MIT Press, 1999.

[7] J. B. Lasserre. An explicit equivalent positive semidefinite program for nonlinear 0-1 programs. *SIAM Journal on Optimization*, 12:756–769, 2001.

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

[9] 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.

[10] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Exact MAP estimates via agreement on (hyper)trees: Linear programming and message-passing approaches. Technical report, UC Berkeley, UCB/CSD-3-1269, August 2003.

[11] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Trans. Info. Theory*, 49(5):1120–1146, 2003.

[12] M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. Technical report, UC Berkeley, Department of Statistics, No. 649, 2003.

[13] M. J. Wainwright and M. I. Jordan. Semidefinite relaxations for approximate inference on graphs with cycles. Technical report, UC Berkeley, UCB/CSD-3-1226, January 2003.

[14] 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.