Chapter 11 in *New Directions in Statistical Signal Processing.*
Edited by S. Haykin, J. Principe, T. Sejnowski and J. McWhirter, MIT Press.
To appear in 2005. (This version: March 2005).

11 A Variational Principle for Graphical Models

Martin J. Wainwright and Michael I. Jordan
Department of Electrical Engineering and Computer Science
Department of Statistics
University of California, Berkeley
Berkeley, CA 94720
wainwrig@eecs.berkeley.edu *jordan@cs.berkeley.edu*

11.1 Introduction

Graphical models bring together graph theory and probability theory in a powerful formalism for multivariate statistical modeling. In statistical signal processing—as well as in related fields such as communication theory, control theory and bioinformatics—statistical models have long been formulated in terms of graphs, and algorithms for computing basic statistical quantities such as likelihoods and marginal probabilities have often been expressed in terms of recursions operating on these graphs. Examples include hidden Markov models, Markov random fields, the forward-backward algorithm and Kalman filtering [Rabiner and Juang (1993); Pearl (1988); Kailath et al. (2000)]. These ideas can be understood, unified and generalized within the formalism of graphical models. Indeed, graphical models provide a natural framework for formulating variations on these classical architectures, and for exploring entirely new families of statistical models.

The recursive algorithms cited above are all instances of a general recursive algorithm known as the *junction tree algorithm* [Lauritzen and Spiegelhalter, 1988]. The junction tree algorithm takes advantage of factorization properties of the joint probability distribution that are encoded by the pattern of missing edges in a graphical model. For suitably sparse graphs, the junction tree algorithm provides a systematic and practical solution to the general problem of computing likelihoods and other statistical quantities associated with a graphical model. Unfortunately, many graphical models of practical interest are not "suitably sparse," so that the junction tree algorithm no longer provides a viable computational solution to the problem of computing marginal probabilities and other expectations. One popular source of methods for attempting to cope with such cases is the *Markov chain Monte Carlo* (MCMC) framework, and indeed there is a significant literature on

the application of MCMC methods to graphical models [Besag and Green (1993); Gilks et al. (1996)]. However, MCMC methods can be overly slow for practical applications in fields such as signal processing, and there has been significant interest in developing faster approximation techniques.

The class of *variational methods* provides an alternative approach to computing approximate marginal probabilities and expectations in graphical models. Roughly speaking, a variational method is based on casting a quantity of interest (e.g., a likelihood) as the solution to an optimization problem, and then solving a perturbed version of this optimization problem. Examples of variational methods for computing approximate marginal probabilities and expectations include the "loopy" form of the *belief propagation* or *sum-product* algorithm [Yedidia et al., 2001; McEliece et al., 1998] as well as a variety of so-called *mean-field* algorithms [Jordan et al., 1999; Zhang, 1996].

Our principal goal in this chapter is to give a mathematically precise and computationally-oriented meaning to the term "variational" in the setting of graphical models—a meaning that reposes on basic concepts in the field of convex analysis [Rockafellar (1970)]. Compared to the somewhat loose definition of "variational" that is often encountered in the graphical models literature, our characterization has certain advantages, both in clarifying the relationships among existing algorithms, and in permitting fuller exploitation of the general tools of convex optimization in the design and analysis of new algorithms. Briefly, the core issues can be summarized as follows. In order to define an optimization problem, it is necessary to specify both a cost function to be optimized, and a constraint set over which the optimization takes place. Reflecting the origins of most existing variational methods in statistical physics, developers of variational methods generally express the function to be optimized as a "free energy", meaning a functional on probability distributions. The set to be optimized over is often left implicit, but it is generally taken to be the set of all probability distributions. A basic exercise in constrained optimization yields the "Boltzmann distribution" as the general form of the solution. While useful, this derivation has two shortcomings. First, the optimizing argument is a joint probability distribution, not a set of marginal probabilities or expectations. Thus, the derivation leaves us short of our goal of a variational representation for computing marginal probabilities. Second, the set of all probability distributions is a very large set, and formulating the optimization problem in terms of such a set provides little guidance in the design of computationally-efficient approximations.

Our approach addresses both of these issues. The key insight is to formulate the optimization problem not over the set of all probability distributions, but rather over a finite-dimensional set $\mathcal{M}$ of *realizable mean parameters*. This set is convex in general, and it is a polytope in the case of discrete random variables. There are several natural ways to approximate this convex set, and a broad range of extant algorithms turn out to involve particular choices of approximations. In particular, as we will show, the "loopy" form of the sum-product or belief propagation algorithm involves an *outer approximation* to $\mathcal{M}$, whereas the more classical mean-field algorithms, on the other hand, involve an *inner approximation*

to the set $\mathcal{M}$. The characterization of belief propagation as an optimization over an outer approximation of a certain convex set does not arise readily within the standard formulation of variational methods. Indeed, given an optimization over all possible probability distributions, it is difficult to see how to move "outside" of such a set. Similarly, while the standard formulation does provide some insight into the differences between belief propagation and mean-field methods (in that they optimize different "free energies"), the standard formulation does not involve the set $\mathcal{M}$, and hence does not reveal the fundamental difference in terms of outer versus inner approximations.

The core of the chapter is a variational characterization of the problem solved by the junction tree algorithm—that of computing exact marginal probabilities and expectations associated with subsets of nodes in a graphical model. These probabilities are obtained as the maximizing arguments of an optimization over the set $\mathcal{M}$. Perhaps surprisingly, this problem is a convex optimization problem for a broad class of graphical models. With this characterization in hand, we show how variational methods arise as "relaxations"—that is, simplified optimization problems that involve some approximation of the constraint set, the cost function or both. We show how a variety of standard variational methods, ranging from classical mean field to cluster variational methods, fit within this framework. We also discuss new methods that emerge from this framework, including a relaxation based on semidefinite constraints and a link between reweighted forms of the max-product algorithm and linear programming.

The remainder of the chapter is organized as follows. The first two sections are devoted to basics: Section 11.2 provides an overview of graphical models and Section 11.3 is devoted to a brief discussion of exponential families. In Section 11.4, we develop a general variational representation for computing marginal probabilities and expectations in exponential families. Section 11.5 illustrates how various exact methods can be understood from this perspective. The remainder of the chapter—Sections 11.6 through 11.8—is devoted to the exploration of various relaxations of this exact variational principle, which in turn yield various algorithms for computing approximations to marginal probabilities and other expectations.

11.2 Background

11.2.1 Graphical models

A graphical model consists of a collection of probability distributions that factorize according to the structure of an underlying graph. A graph $G = (V, E)$ is formed by a collection of vertices V, and a collection of edges E. An edge consists of a pair of vertices, and may either be directed or undirected. Associated with each vertex $s \in V$ is a random variable x_s taking values in some set $\mathcal{X}_s$, which may either be continuous (e.g., $\mathcal{X}_s = \mathbb{R}$) or discrete (e.g., $\mathcal{X}_s = \{0, 1, \ldots, m-1\}$). For any subset A of the vertex set V, we define $x_A := \{x_s \mid s \in A\}$.

Directed graphical models: In the directed case, each edge is directed from parent to child. We let $\pi(s)$ denote the set of all parents of given node $s \in V$. (If s has no parents, then the set $\pi(s)$ should be understood to be empty.) With this notation, a *directed graphical model* consists of a collection of probability distributions that factorize in the following way:

$$p(\mathbf{x}) = \prod_{s \in V} p(x_s \mid x_{\pi(s)}). \tag{11.1}$$

It can be verified that our use of notation is consistent, in that $p(x_s \mid x_{\pi(s)})$ is, in fact, the conditional distribution for the global distribution $p(\mathbf{x})$ thus defined.

Undirected graphical models: In the undirected case, the probability distribution factorizes according to functions defined on the *cliques* of the graph (i.e., fully-connected subsets of V). In particular, associated with each clique C is a *compatibility function* $\psi_C : \mathcal{X}^n \to \mathbb{R}_+$ that depends only on the subvector x_C. With this notation, an *undirected graphical model* (also known as a *Markov random field*) consists of a collection of distributions that factorize as

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{C} \psi_C(x_C), \tag{11.2}$$

where the product is taken over all cliques of the graph. The quantity Z is a constant chosen to ensure that the distribution is normalized. In contrast to the directed case (11.1), in general the compatibility functions ψ_C need not have any obvious or direct relation to local marginal distributions.

Families of probability distributions as defined as in (11.1) or (11.2) also have a characterization in terms of conditional independencies among subsets of random variables. We will not use this characterization in this chapter, but refer the interested reader to Lauritzen [1996] for a full treatment.

11.2.2 Inference problems and exact algorithms

Given a probability distribution $p(\cdot)$ defined by a graphical model, our focus will be solving one or more of the following *inference problems*:

(a) computing the likelihood.

(b) computing the marginal distribution $p(x_A)$ over a particular subset $A \subset V$ of nodes.

(c) computing the conditional distribution $p(x_A \mid x_B)$, for disjoint subsets A and B, where $A \cup B$ is in general a proper subset of V.

(d) computing a mode of the density (i.e., an element $\widehat{\mathbf{x}}$ in the set $\arg\max_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x})$).

Problem (a) is a special case of problem (b), because the likelihood is the marginal probability of the observed data. The computation of a conditional probability in (c) is similar in that it also requires marginalization steps, an initial one to obtain

the numerator $p(x_A, x_B)$, and a further step to obtain the denominator $p(x_B)$. In contrast, the problem of computing modes stated in (d) is fundamentally different, since it entails maximization rather than integration. Although problem (d) is not the main focus of this chapter, there are important connections between the problem of computing marginals and that of computing modes; these are discussed in Section 11.8.2.

To understand the challenges inherent in these inference problems, consider the case of a discrete random vector $\mathbf{x} \in \mathcal{X}^n$, where $\mathcal{X}_s = \{0, 1, \ldots, m-1\}$ for each vertex $s \in V$. A naive approach to computing a marginal at a single node—say $p(x_s)$—entails summing over all configurations of the form $\{\mathbf{x}' \mid x'_s = x_s\}$. Since this set has m^{n-1} elements, it is clear that a brute force approach will rapidly become intractable as n grows. Similarly, computing a mode entails solving an integer programming problem over an exponential number of configurations. For continuous random vectors, the problems are no easier[1] and typically harder, since they require computing a large number of integrals.

Both directed and undirected graphical models involve factorized expressions for joint probabilities, and it should come as no surprise that exact inference algorithms treat them in an essentially identical manner. Indeed, to permit a simple unified treatment of inference algorithms, it is convenient to convert directed models to undirected models and to work exclusively within the undirected formalism. Any directed graph can be converted, via a process known as moralization [Lauritzen and Spiegelhalter (1988)], to an undirected graph that—at least for the purposes of solving inference problems—is equivalent. Throughout the rest of the chapter, we assume that this transformation has been carried out.

11.2.2.1 Message-passing on trees

For graphs without cycles—also known as *trees*—these inference problems can be solved exactly by recursive "message-passing" algorithms of a dynamic programming nature, with a computational complexity that scales only linearly in the number of nodes. In particular, for the case of computing marginals, the dynamic programming solution takes the form of a general algorithm known as the *sum-product algorithm*, whereas for the problem of computing modes it takes the form of an analogous algorithm known as the *max-product algorithm*. Here we provide a brief description of these algorithms; further details can be found in various sources [Aji and McEliece (2000); Kschischang and Frey (1998); Lauritzen and Spiegelhalter (1988); Loeliger (2004)].

We begin by observing that the cliques of a tree-structured graph $T = (V, E(T))$ are simply the individual nodes and edges. As a consequence, any tree-structured

1. The Gaussian case is an important exception to this statement.

graphical model has the following factorization:

$$p(\mathbf{x}) = \frac{1}{Z} \prod_{s \in V} \psi_s(x_s) \prod_{(s,t) \in E(T)} \psi_{st}(x_s, x_t). \tag{11.3}$$

Here we describe how the sum-product algorithm computes the marginal distribution $\mu_s(x_s) := \sum_{\{\mathbf{x}' \mid x'_s = x_s\}} p(\mathbf{x})$ for every node of a tree-structured graph. We will focus on detail on the case of discrete random variables, with the understanding that the computations carry over (at least in principle) to the continuous case by replacing sums with integrals.

Sum-product algorithm: The essential principle underlying the sum-product algorithm on trees is divide and conquer: we solve a large problem by breaking it down into a sequence of simpler problems. The tree itself provides a natural way to break down the problem as follows. For an arbitrary $s \in V$, consider the set of its neighbors $\mathcal{N}(s) = \{u \in V | \ (s,u) \in E\}$. For each $u \in \mathcal{N}(s)$, let $T_u = (V_u, E_u)$ be the subgraph formed by the set of nodes (and edges joining them) that can be reached from u by paths that *do not* pass through node s. The key property of a tree is that each such subgraph T_u is again a tree, and T_u and T_v are disjoint for $u \neq v$. In this way, each vertex $u \in \mathcal{N}(s)$ can be viewed as the root of a subtree T_u, as illustrated in Figure 11.1(a). For each subtree T_t, we define $x_{V_t} := \{x_u \mid u \in V_t\}$. Now consider the collection of terms in equation (11.3) associated with vertices or edges in T_t: collecting all of these terms yields a subproblem $p(x_{V_t}; T_t)$ for this subtree.

Now the conditional independence properties of a tree allow the computation of the marginal at node μ_s to be broken down into a product of the form

$$\mu_s(x_s) \ \propto \ \psi_s(x_s) \prod_{t \in \mathcal{N}(s)} M^*_{ts}(x_s). \tag{11.4}$$

Each term $M^*_{ts}(x_s)$ in this product is the result of performing a partial summation for the subproblem $p(x_{V_t}; T_t)$ in the following way:

$$M^*_{ts}(x_s) \ = \sum_{\{x'_{T_t} \mid x'_s = x_s\}} \psi_{st}(x_s, x'_t) \ p(x'_{T_t}; T_t). \tag{11.5}$$

For fixed x_s, the subproblem defining $M^*_{ts}(x_s)$ is again a tree-structured summation, albeit involving a subtree T_t *smaller* than the original tree T. Therefore, it too can be broken down recursively in a similar fashion. In this way, the marginal at node s can be computed by a series of recursive updates.

Rather than applying the procedure described above to each node separately, the *sum-product algorithm* computes the marginals for all nodes simultaneously and in parallel. At each iteration, each node t passes a "message" to each of its neighbors $u \in \mathcal{N}(t)$. This message, which we denote by $M_{tu}(x_u)$, is a function of the possible states $x_u \in \mathcal{X}_u$ (i.e., a vector of length $|\mathcal{X}_u|$ for discrete random variables). On the full graph, there are a total of $2|E|$ messages, one for each direction of each edge. This full collection of messages is updated, typically in parallel, according to the

following recursion:

$$M_{ts}(x_s) \leftarrow \kappa \sum_{x'_t} \Big\{ \psi_{st}(x_s, x'_t)\, \psi_t(x'_t) \prod_{u \in \mathcal{N}(t)/s} M_{ut}(x'_t) \Big\}, \qquad (11.6)$$

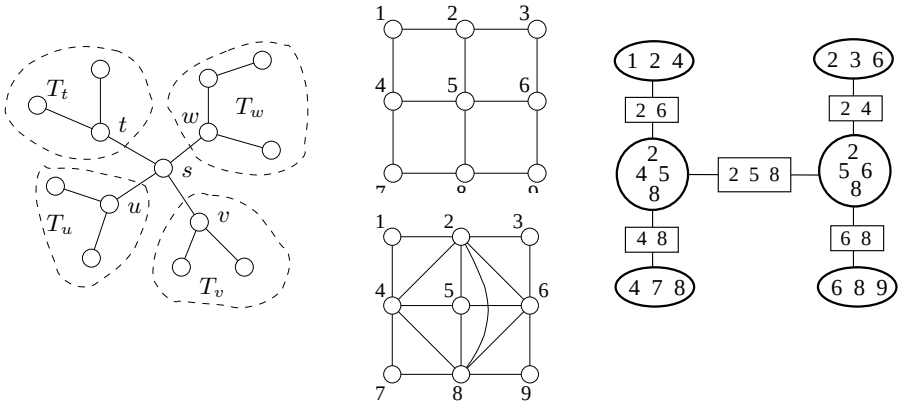
where $\kappa > 0$ is a normalization constant. It can be shown [Pearl (1988)] that for tree-structured graphs, iterates generated by the update (11.6) will converge to a unique fixed point $M^* = \{M^*_{st}, M^*_{ts},\ (s,t) \in E\}$ after a finite number of iterations. Moreover, component M^*_{ts} of this fixed point is precisely equal, up to a normalization constant, to the subproblem defined in equation (11.5), which justifies our abuse of notation post hoc. Since the fixed point M^* specifies the solution to all of the subproblems, the marginal μ_s at every node $s \in V$ can be computed easily via equation (11.4).

Max-product algorithm: Suppose that the summation in the update (11.6) is replaced by a maximization. The resulting *max-product* algorithm solves the problem of finding a mode of a tree-structured distribution $p(\mathbf{x})$. In this sense, it represents a generalization of the Viterbi algorithm [Forney (1973)] from chains to arbitrary tree-structured graphs. More specifically, the max-product updates will converge to another unique fixed point M^*—distinct, of course, from the sum-product fixed point. This fixed point can be used to compute the *max-marginal* $\nu_s(x_s) := \max_{\{\mathbf{x}' \mid x'_s = x_s\}} p(\mathbf{x}')$ at each node of the graph, in an analogous way to the computation of ordinary sum-marginals. Given these max-marginals, it is straightforward to compute a mode $\widehat{\mathbf{x}} \in \arg\max_{\mathbf{x}} p(\mathbf{x})$ of the distribution [Dawid (1992); Wainwright et al. (2004)]. More generally, updates of this form apply to arbitrary *commutative semirings* on tree-structured graphs [Dawid (1992); Aji and McEliece (2000)]. The pairs "sum-product" and "max-product" are two particular examples of such an algebraic structure.

11.2.2.2 Junction tree representation

We have seen that inference problems on trees can be solved exactly by recursive message-passing algorithms. Given a graph with cycles, a natural idea is to cluster its nodes so as to form a *clique tree*—that is, an acyclic graph whose nodes are formed by cliques of G. Having done so, it is tempting to simply apply a standard algorithm for inference on trees. However, the clique tree must satisfy an additional restriction so as to ensure consistency of these computations. In particular, since a given vertex $s \in V$ may appear in multiple cliques (say C_1 and C_2), what is required is a mechanism for enforcing consistency among the different appearances of the random variable x_s.

In order to enforce consistency, it turns out to be necessary to restrict attention to those clique trees that satisfy a particular graph-theoretic property. In particular, we say that a clique tree satisfies the *running intersection property* if for any two clique nodes C_1 and C_2, all nodes on the unique path joining them contain the



(a) (b) (c)

Figure 11.1 (a): Decomposition of a tree, rooted at node s, into subtrees. Each neighbor (e.g., u) of node s is the root of a subtree (e.g., T_u). Subtrees T_u and T_v, for $t \neq u$, are disconnected when node s is removed from the graph. (b), (c) Illustration of junction tree construction. Top panel in (b) shows original graph: a 3×3 grid. Bottom panel in (b) shows triangulated version of original graph. Note the two 4-cliques in the middle. (c) Corresponding junction tree for triangulated graph in (b), with maximal cliques depicted within ellipses. The rectangles are separator sets; these are intersections of neighboring cliques.

intersection $C_1 \cap C_2$. Any clique tree with this property is known as a *junction tree*.

For what type of graphs can one build junction trees? An important result in graph theory asserts that a graph G has a junction tree if and only if it is *triangulated*.[2] This result underlies the *junction tree algorithm* [Lauritzen and Spiegelhalter (1988)] for exact inference on arbitrary graphs, which consists of the following three steps:

1. Given a graph with cycles G, triangulate it by adding edges as necessary.
2. Form a junction tree associated with the triangulated graph.
3. Run a tree inference algorithm on the junction tree.

We illustrate these basic steps with an example.

Example 11.1

Consider the 3×3 grid shown in the top panel of Figure 11.1(b). The first step is to form a triangulated version, as shown in the bottom panel of Figure 11.1(b). Note that the graph would *not* be triangulated if the additional edge joining nodes 2 and 8 were not present. Without this edge, the 4-cycle $(2 - 4 - 8 - 6 - 2)$ would lack

2. A graph is triangulated means that every cycle of length four or longer has a chord.

a chord. Panel (c) shows a junction tree associated with this triangulated graph, in which circles represent maximal cliques (i.e., fully-connected subsets of nodes that cannot be augmented with an additional node and remain fully-connected), and boxes represent *separator sets* (intersections of cliques adjacent in the junction tree). ♢

An important by-product of the junction tree construction is an alternative representation of the probability distribution defined by a graphical model. Let $\mathcal{C}$ denote the set of all maximal cliques in the triangulated graph, and define $\mathcal{S}$ as the set of all separator sets in the junction tree. For each separator set $S \in \mathcal{S}$, let $d(S)$ denote the number of maximal cliques to which it is adjacent. The junction tree framework guarantees that the distribution $p(\cdot)$ factorizes in the form

$$p(\mathbf{x}) = \frac{\prod_{C \in \mathcal{C}} \mu_C(x_C)}{\prod_{S \in \mathcal{S}} [\mu_S(x_S)]^{d(S)-1}}, \tag{11.7}$$

where μ_C and μ_S are the marginal distributions over the cliques and separator sets respectively. Observe that unlike the representation of equation (11.2), the decomposition of equation (11.7) is directly in terms of marginal distributions, and does not require a normalization constant (i.e., $Z = 1$).

Example 11.2 Markov chain

Consider the Markov chain $p(x_1, x_2, x_3) = p(x_1)\, p(x_2 \,|\, x_1)\, p(x_3 \,|\, x_2)$. The cliques in a graphical model representation are $\{1, 2\}$ and $\{2, 3\}$, with separator $\{2\}$. Clearly the distribution cannot be written as the product of marginals involving only the cliques. However, if we include the separator, it can be factorized in terms of its marginals—viz. $p(x_1, x_2, x_3) = \frac{p(x_1, x_2) p(x_2, x_3)}{p(x_2)}$. ♢

To anticipate the development in the sequel, it is helpful to consider the following "inverse" perspective on the junction tree representation. Suppose that we are given a set of functions $\tau_C(x_C)$ and $\tau_S(x_S)$ associated with the cliques and separator sets in the junction tree. What conditions are necessary to ensure that these functions are valid marginals for some distribution? Suppose that the functions $\{\tau_S, \tau_C\}$ are *locally consistent* in the following sense:

$$\sum_{x_S} \tau_S(x_S) = 1 \qquad \text{normalization} \tag{11.8a}$$

$$\sum_{\{\mathbf{x}'_C \,|\, \mathbf{x}'_S = x_S\}} \tau_C(x'_C) = \tau_S(x_S) \qquad \text{marginalization} \tag{11.8b}$$

The essence of the junction tree theory described above is that such local consistency is both necessary and sufficient to ensure that these functions are valid marginals for some distribution.

Finally, turning to the computational complexity of the junction tree algorithm, the computational cost grows exponentially in the size of the maximal clique in the junction tree. The size of the maximal clique over all possible triangulations of a graph defines an important graph-theoretic quantity known as the *treewidth*

of the graph. Thus, the complexity of the junction tree algorithm is exponential in the treewidth. For certain classes of graphs, including chains and trees, the treewidth is small and the junction tree algorithm provides an effective solution to inference problems. Such families include many well-known graphical model architectures, and the junction tree algorithm subsumes many classical recursive algorithms, including the forward-backward algorithms for hidden Markov models [Rabiner and Juang (1993)], the Kalman filtering-smoothing algorithms for state-space models [Kailath et al. (2000)], and the pruning and peeling algorithms from computational genetics [Felsenstein (1981)]. On the other hand, there are many graphical models (e.g., grids) for which the treewidth is infeasibly large. Coping with such models requires leaving behind the junction tree framework, and turning to approximate inference algorithms.

11.2.3 Message-passing algorithms for approximate inference

In the remainder of the chapter, we present a general variational principle for graphical models that can be used to derive a class of techniques known as *variational inference algorithms*. To motivate our later development, we pause to give a high-level description of two variational inference algorithms, with the goal of highlighting their simple and intuitive nature.

The first variational algorithm that we consider is a so-called "loopy" form of the sum-product algorithm (also referred to as the *belief propagation* algorithm). Recall that the sum-product algorithm is designed as an exact method for trees; from a purely algorithmic point of view, however, there is nothing to prevent one from running the procedure on a graph with cycles. More specifically, the message updates (11.6) can be applied at a given node while ignoring the presence of cycles—essentially pretending that any given node is embedded in a tree. Intuitively, such an algorithm might be expected to work well if the graph is suitably "tree-like," such that the effect of messages propagating around cycles is appropriately diminished. This algorithm is in fact widely used in various applications that involve signal processing, including image processing, computer vision, computational biology, and error-control coding.

A second variational algorithm is the so-called *naive mean field* algorithm. For concreteness, we describe it in application to a very special type of graphical model, known as the Ising model. The Ising model is a Markov random field involving a binary random vector $\mathbf{x} \in \{0,1\}^n$, in which pairs of adjacent nodes are coupled with a weight θ_{st}, and each node has an observation weight θ_s. (See Examples 11.4 and 11.11 for a more detailed description of this model.) To motivate the mean field updates, we consider the Gibbs sampler for this model, in which the basic update step is to choose a node $s \in V$ randomly, and then to update the state of the associated random variable according to the conditional probability with neighboring states fixed. More precisely, denoting by $\mathcal{N}(s)$ the neighbors of a node $s \in V$, and letting $x^{(p)}_{\mathcal{N}(s)}$ denote the state of the neighbors of s at iteration p, the

Gibbs update for x_s takes the following form:

$$x_s^{(p+1)} = \begin{cases} 1 & \text{if } u \leq \{1 + \exp[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} x_t^{(p)})]\}^{-1} \\ 0 & \text{otherwise} \end{cases}, \tag{11.9}$$

where u is a sample from a uniform distribution $\mathcal{U}(0,1)$. It is well-known that this procedure generates a sequence of configurations that converge (in a stochastic sense) to a sample from the Ising model distribution.

In a dense graph, such that the cardinality of $\mathcal{N}(s)$ is large, we might attempt to invoke a law of large numbers or some other concentration result for $\sum_{t \in \mathcal{N}(s)} \theta_{st} x_t^{(p)}$. To the extent that such sums are concentrated, it might make sense to replace sample values with expectations, which motivates the following averaged version of equation (11.9):

$$\mu_s \leftarrow \left\{ 1 + \exp\big[-(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t) \big] \right\}^{-1}, \tag{11.10}$$

in which μ_s denotes an estimate of the marginal probability $p(x_s = 1)$. Thus, rather than flipping the random variable x_s with a probability that depends on the state of its neighbors, we update a parameter μ_s using a deterministic function of the corresponding parameters $\{\mu_t \mid t \in \mathcal{N}(s)\}$ at its neighbors. Equation (11.10) defines the naive mean field algorithm for the Ising model, which can be viewed as a message-passing algorithm on the graph.

At first sight, message-passing algorithms of this nature might seem rather mysterious, and do raise some questions. Do the updates have fixed points? Do the updates converge? What is the relation between the fixed points and the exact quantities? The goal of the remainder of this chapter is to shed some light on such issues. Ultimately, we will see that a broad class of message-passing algorithms, including the mean field updates, the sum-product and max-product algorithms, as well as various extensions of these methods can all be understood as solving either exact or approximate versions of a certain variational principle for graphical models.

11.3 Graphical models in exponential form

We begin by describing how many graphical models can be viewed as particular types of exponential families. Further background can be found in the books by Efron [1978] and Brown [1986]. This exponential family representation is the foundation of our later development of the variational principle.

11.3.1 Maximum entropy

One way in which to motivate exponential family representations of graphical models is through the principle of maximum entropy. The set-up for this principle

is as follows: given a collection of functions $\phi_\alpha : \mathcal{X}^n \to \mathbb{R}$, suppose that we have observed their expected values—that is, we have

$$\mathbb{E}[\phi_\alpha(\mathbf{x})] = \mu_\alpha \qquad \text{for all } \alpha \in \mathcal{I}, \tag{11.11}$$

where $\mu = \{\mu_\alpha \mid \alpha \in \mathcal{I}\}$ is a real vector, $\mathcal{I}$ is an index set, and $d := |\mathcal{I}|$ is the length of the vectors μ and $\boldsymbol{\phi} := \{\phi_\alpha \mid \alpha \in \mathcal{I}\}$.

Our goal is use the observations to infer a full probability distribution. Let $\mathcal{P}$ denote the set of all probability distributions p over the random vector $\mathbf{x}$. Since there are (in general) many distributions $p \in \mathcal{P}$ that are consistent with the observations (11.11), we need a principled method for choosing among them. The principle of maximum entropy is to choose the distribution p_{ME} such that its *entropy*, defined as $H(p) := -\sum_{\mathbf{x} \in \mathcal{X}^n} p(\mathbf{x}) \log p(\mathbf{x})$, is maximized. More formally, the maximum entropy solution p_{ME} is given by the following constrained optimization problem:

$$p_{ME} \;:=\; \arg\max_{p \in \mathcal{P}} H(p) \qquad \text{subject to constraints (11.11).} \tag{11.12}$$

One interpretation of this principle is as choosing the distribution with maximal uncertainty while remaining faithful to the data.

Presuming that problem (11.12) is feasible, it is straightforward to show using a Lagrangian formulation that its optimal solution takes the form

$$p(\mathbf{x}; \theta) \;\propto\; \exp\Big\{ \sum_{\alpha \in \mathcal{I}} \theta_\theta \phi_\alpha(\mathbf{x}) \Big\}, \tag{11.13}$$

which corresponds to a distribution in exponential form. Note that the exponential decomposition (11.13) is analogous to the product decomposition (11.2) considered earlier.

In the language of exponential families, the vector $\theta \in \mathbb{R}^d$ is known as the *canonical parameter*, and the collection of functions $\boldsymbol{\phi} = \{\phi_\alpha \mid \alpha \in \mathcal{I}\}$ are known as *sufficient statistics*. In the context of our current presentation, each canonical parameter θ_α has a very concrete interpretation as the Lagrange multiplier associated with the constraint $\mathbb{E}[\phi_\alpha(\mathbf{x})] = \mu_\alpha$.

11.3.2 Exponential families

We now define exponential families in more generality. Any exponential family consists of a particular class of densities taken with respect to a fixed based measure $\boldsymbol{\nu}$. The base measure is typically counting measure (as in our discrete example above), or Lebesgue measure (e.g., for Gaussian families). Throughout this chapter, we use $\langle a, b \rangle$ to denote the ordinary Euclidean inner product between two vectors a and b of the same dimension. Thus, for each fixed $\mathbf{x} \in \mathcal{X}^n$, the quantity $\langle \theta, \boldsymbol{\phi}(\mathbf{x}) \rangle$ is the Euclidean inner product in $\mathbb{R}^d$ of the two vectors $\theta \in \mathbb{R}^d$ and $\boldsymbol{\phi}(\mathbf{x}) = \{\phi_\alpha(\mathbf{x}) \mid \alpha \in \mathcal{I}\}$.

With this notation, the *exponential family* associated with $\boldsymbol{\phi}$ consists of the

following parameterized collection of density functions:

$$p(\mathbf{x};\theta) \;=\; \exp\big\{\langle\theta,\, \boldsymbol{\phi}(\mathbf{x})\rangle \;-\; A(\theta)\big\}. \tag{11.14}$$

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

$$A(\theta) \;=\; \log \int_{\mathcal{X}^n} \exp\langle\theta,\, \boldsymbol{\phi}(\mathbf{x})\rangle\, \boldsymbol{\nu}(d\mathbf{x}). \tag{11.15}$$

Presuming that the integral is finite, this definition ensures that $p(\mathbf{x};\theta)$ is properly normalized (i.e., $\int_{\mathcal{X}^n} p(\mathbf{x};\theta)\boldsymbol{\nu}(d\mathbf{x}) = 1$). With the set of potentials $\boldsymbol{\phi}$ fixed, each parameter vector θ indexes a particular member $p(\mathbf{x};\theta)$ of the family. The canonical parameters θ of interest belong to the set

$$\Theta \;:=\; \{\theta \in \mathbb{R}^d \mid A(\theta) < \infty\}. \tag{11.16}$$

Throughout this chapter, we deal exclusively with *regular* exponential families, for which the set Θ is assumed to be open.

We summarize for future reference some well-known properties of A:

Lemma 11.1
The cumulant generating function A is convex in terms of θ. Moreover, it is infinitely differentiable on Θ, and its derivatives correspond to cumulants.

As an important special case, the first derivatives of A take the form

$$\frac{\partial A}{\partial \theta_\alpha} \;=\; \int_{\mathcal{X}^n} \phi_\alpha(\mathbf{x}) p(\mathbf{x};\theta)\boldsymbol{\nu}(d\mathbf{x}) \;=\; \mathbb{E}_\theta[\phi_\alpha(\mathbf{x})], \tag{11.17}$$

and define a vector $\mu := \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})]$ of *mean parameters* associated with the exponential family. There are important relations between the canonical and mean parameters, and many inference problems can be formulated in terms of the mean parameters. These correspondences and other properties of the cumulant generating function are fundamental to our development of a variational principle for solving inference problems.

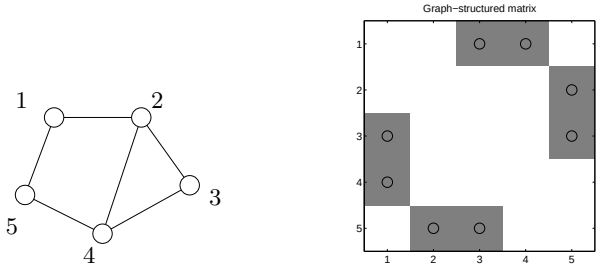
11.3.3 Illustrative examples

In order to illustrate these definitions, we now discuss some particular classes of graphical models that commonly arise in signal and image processing problems, and how they can be represented in exponential form. In particular, we will see that graphical structure is reflected in the choice of sufficient statistics, or equivalently in terms of constraints on the canonical parameter vector.

We begin with an important case—the Gaussian Markov random field— which is widely used for modeling various types of imagery and spatial data [Luettgen et al. (1994); Szeliski (1990)].

Example 11.3 Gaussian MRF

Consider a graph $G = (V, E)$, such as that illustrated in Figure 11.2(a), and suppose that each vertex $s \in V$ has an associated Gaussian random variable x_s. Any such scalar Gaussian is a (two-dimensional) exponential family specified by sufficient statistics x_s and x_s^2. Turning to the Gaussian random vector $\mathbf{x} := \{x_s \mid s \in V\}$, it has an exponential family representation in terms of the sufficient statistics $\{x_s, x_s^2 \mid s \in V\} \cup \{x_s x_t \mid (s,t) \in E\}$, with associated canonical parameters $\{\theta_s,\ \theta_{ss} \mid s \in V\} \cup \{\theta_{st} \mid (s,t) \in E\}$. Here the additional cross-terms $x_s x_t$ allow for possible correlation between components x_s and x_t of the Gaussian random vector. Note that there are a total of $d = 2n + |E|$ sufficient statistics.



(a) (b)

Figure 11.2 (a) A simple Gaussian model based on a graph G with 5 vertices. (b) The adjacency matrix of the graph G in (a), which specifies the sparsity pattern of the matrix $Z(\theta)$.

The sufficient statistics and parameters can be represented compactly as $(n+1) \times (n+1)$ symmetric matrices:

$$\mathbf{X} = \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix} \qquad U(\theta) := \begin{bmatrix} 0 & \theta_1 & \theta_2 & \dots & \theta_n \\ \theta_1 & \theta_{11} & \theta_{12} & \dots & \theta_{1n} \\ \theta_2 & \theta_{21} & \theta_{22} & \dots & \theta_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \theta_n & \theta_{n1} & \theta_{n2} & \dots & \theta_{nn} \end{bmatrix} \tag{11.18}$$

We use $Z(\theta)$ to denote the lower $n \times n$ block of $U(\theta)$; it is known as the *precision matrix*. We say that $\mathbf{x}$ forms a Gaussian Markov random field if its probability density function decomposes according to the graph $G = (V, E)$. In terms of our canonical parameterization, this condition translates to the requirement that $\theta_{st} = 0$ whenever $(s,t) \notin E$. Alternatively stated, the precision matrix $Z(\theta)$ must have the same zero-pattern as the adjacency matrix of the graph, as illustrated in Figure 11.2(b).

For any two symmetric matrices C and D, it is convenient to define the inner product $\langle C,\ D \rangle := \text{trace}(C\,D)$. Using this notation leads to a particularly compact

representation of a Gaussian MRF:

$$p(\mathbf{x};\theta) \;=\; \exp\big\{\langle U(\theta),\; \mathbf{X}\rangle - A(\theta)\big\}, \tag{11.19}$$

where $A(\theta) := \log \int_{\mathbb{R}^n} \exp\big[\langle U(\theta),\; \mathbf{X}\rangle\big] d\mathbf{x}$ is the log cumulant generating function. The integral defining $A(\theta)$ is finite only if the $n \times n$ precision matrix $Z(\theta)$ is negative definite, so that the domain of A has the form $\Theta = \{\theta \in \mathbb{R}^d \mid Z(\theta) \prec 0\}$.

Note that the mean parameters in the Gaussian model have a clear interpretation. The singleton elements $\mu_s = \mathbb{E}_\theta[x_s]$ are simply the Gaussian mean, whereas the elements $\mu_{ss} = \mathbb{E}_\theta[x_s^2]$ and $\mu_{st} = \mathbb{E}_\theta[x_s x_t]$ are second-order moments. $\diamondsuit$

Markov random fields involving *discrete* random variables also arise in many applications, including image processing, bioinformatics, and error-control coding [Geman and Geman (1984); Kschischang et al. (2001); Loeliger (2004); Durbin et al. (1998)]. As with the Gaussian case, this class of Markov random fields also has a natural exponential representation.

Example 11.4 Multinomial MRF

Suppose that each x_s is a multinomial random variable, taking values in the space $\mathcal{X}_s = \{0, 1, \ldots, m_s - 1\}$. In order to represent a Markov random field over the vector $\mathbf{x} = \big\{x_s \mid s \in V\big\}$ in exponential form, we now introduce a particular set of sufficient statistics that will be useful in the sequel. For each $j \in \mathcal{X}_s$, let $\mathbb{I}_j(x_s)$ be an indicator function for the event $\{x_s = j\}$. Similarly, for each pair $(j,k) \in \mathcal{X}_s \times \mathcal{X}_t$, let $\mathbb{I}_{jk}(x_s, x_t)$ be an indicator for the event $\{(x_s, x_t) = (j,k)\}$. These building blocks yield the following set of sufficient statistics:

$$\big\{\mathbb{I}_j(x_s) \mid s \in V,\; j \in \mathcal{X}_s\big\} \cup \big\{\mathbb{I}_j(x_s)\mathbb{I}_k(x_t) \mid (s,t) \in E,\; (j,k) \in \mathcal{X}_s \times \mathcal{X}_t\big\}. \tag{11.20}$$

The corresponding canonical parameter θ has elements of the form

$$\theta = \big\{\theta_{s;j} \mid s \in V,\; j \in \mathcal{X}_s\big\} \cup \big\{\theta_{st;jk} \mid (s,t) \in E,\; (j,k) \in \mathcal{X}_s \times \mathcal{X}_t\big\}. \tag{11.21}$$

It is convenient to combine the canonical parameters and indicator functions using the shorthand notation $\theta_s(x_s) := \sum_{j \in \mathcal{X}_s} \theta_{s;j}\mathbb{I}_j(x_s)$; the quantity $\theta_{st}(x_s, x_t)$ can be defined similarly.

With this notation, a multinomial MRF with pairwise interactions can be written in exponential form as

$$p(\mathbf{x};\theta) \;=\; \exp\Big\{\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) - A(\theta)\Big\}, \tag{11.22}$$

where the cumulant generating function is given by the summation

$$A(\theta) := \log \sum_{\mathbf{x} \in \mathcal{X}^n} \exp\Big\{\sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t)\Big\}.$$

In signal processing applications of these models, the random vector $\mathbf{x}$ is often viewed as hidden or partially observed (for instance, corresponding to the correct segmentation of an image). Thus, it is frequently the case that the functions θ_s

are determined by noisy observations, whereas the terms θ_{st} control the coupling between variables x_s and x_t that are adjacent on the graph (e.g., reflecting spatial continuity assumptions). See Figure 11.3(a) for an illustration of such a multinomial MRF defined on a two-dimensional lattice, which is a widely-used model in statistical image processing [Geman and Geman (1984)]. In the special case that $\mathcal{X}_s = \{0, 1\}$ for all $s \in V$, the family (11.22) is known as the *Ising model*.

Note that the mean parameters associated with this model correspond to particular marginal probabilities. For instance, the mean parameters associated with vertex s have the form $\mu_{s;j} = \mathbb{E}_\theta[\mathbb{I}_j(x_s)] = p(x_s = j; \theta)$, and the mean parameters μ_{st} associated with edge (s, t) have an analogous interpretation as pairwise marginal values.

$\diamondsuit$

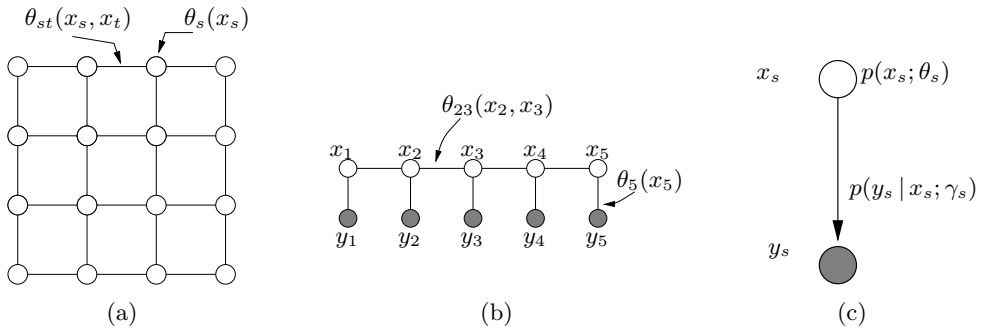


Figure 11.3 (a) A multinomial MRF on a 2-D lattice model. (b) A hidden Markov model (HMM) is a special case of a multinomial MRF for a chain-structured graph. (c) The graphical representation of a scalar Gaussian mixture model: the multinomial x_s indexes components in the mixture, and y_s is conditionally Gaussian (with exponential parameters γ_s) given the mixture component x_s.

Example 11.5 Hidden Markov model

A very important special case of the multinomial MRF is the hidden Markov model (HMM), which is a chain-structured graphical model widely used for the modeling of time series and other one-dimensional signals. It is conventional in the HMM literature to refer to the multinomial random variables $\mathbf{x} = \{x_s \mid s \in V\}$ as "state variables." As illustrated in Figure 11.3(b), the edge set E defines a chain linking the state variables. The parameters $\theta_{st}(x_s, x_t)$ define the *state transition matrix*; if this transition matrix is the same for all pairs s and t, then we have a *homogeneous* Markov chain. Associated with each multinomial state variable x_s is a noisy *observation* y_s, defined by the conditional probability distribution $p(y_s|x_s)$. If we condition on the observed value of y_s, this conditional probability is simply a function of x_s, which we denote by $\theta_s(x_s)$. Given these definitions, equation (11.22) describes the conditional probability distribution $p(\mathbf{x} \mid \mathbf{y})$ for the

HMM. In Figure 11.3(b), this conditioning is captured by shading the corresponding nodes in the graph. Note that the cumulant generating function $A(\theta)$ is, in fact, equal to the log likelihood of the observed data. ♢

Graphical models are not limited to cases in which the random variables at each node belong to the same exponential family. More generally, we can consider heterogeneous combinations of exponential family members. A very natural example, which combines the two previous types of graphical model, is that of a Gaussian mixture model. Such mixture models are widely used in modeling various classes of data, including natural images, speech signals, and financial time series data; see the book [Titterington et al. (1986)] for further background.

Example 11.6 Mixture model

As shown in Figure 11.3(c), a *scalar* mixture model has a very simple graphical interpretation. In particular, let x_s be a multinomial variable, taking values in $\mathcal{X}_s = \{0, 1, 2, \ldots, m_s - 1\}$, specified in exponential parameter form with a function $\theta_s(x_s)$. The role of x_s is to specify the choice of mixture component in the mixture model, so that our mixture model has m_s components in total. We now let y_s be conditionally Gaussian given x_s, so that the conditional distribution $p(y_s \,|\, x_s; \gamma_s)$ can be written in exponential family form with canonical parameters γ_s that are a function of x_s. Overall, the pair (x_s, y_s) form a very simple graphical model in exponential form, as shown in Figure 11.3(c).

The pair (x_s, y_s) serves a basic block for building more sophisticated graphical models. For example, one model is based on assuming that the mixture vector $\mathbf{x}$ is a multinomial MRF defined on an underlying graph $G = (V, E)$, whereas the components of $\mathbf{y}$ are conditionally independent given the mixture vector $\mathbf{x}$. These assumptions lead to an exponential family $p(\mathbf{y}, \mathbf{x}; \theta, \gamma)$ of the form:

$$\prod_{s \in V} p(y_s \,|\, x_s; \gamma_s) \; \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t)] \Big\}. \tag{11.23}$$

For tree-structured graphs, Crouse et al. [1998] have applied this type of mixture model to applications in wavelet-based signal processing. ♢

This type of mixture model is a particular example of a broad class of graphical models that involve heterogeneous combinations of exponential family members (e.g., hierarchical Bayesian models).

11.4 An exact variational principle for inference

With this set-up, we can now re-phrase inference problems in the language of exponential families. In particular, this chapter focuses primarily on the following two problems:

(a) computing the cumulant generating function $A(\theta)$

(b) computing the vector of mean parameters $\mu := \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})]$

In Section 11.8.2 we discuss a closely related problem—namely, that of computing a mode of the distribution $p(\mathbf{x};\theta)$.

The problem of computing the cumulant generating function arises in a variety of signal processing problems, including likelihood ratio tests (for classification and detection problems) and parameter estimation. The computation of mean parameters is also fundamental, and takes different forms depending on the underlying graphical model. For instance, it corresponds to computing means and covariances in the Gaussian case, whereas for a multinomial MRF it corresponds to computing marginal distributions.

The goal of this section is to show how both of these inference problems can be represented *variationally*—as the solution of an optimization problem. The variational principle that we develop, though related to the classical "free energy" approach of statistical physics [Yedidia et al. (2001)], also has important differences. The classical principle yields a variational formulation for the cumulant generating function (or log partition function) in terms of optimizing over the space of all distributions. In our approach, on the other hand, the optimization is not defined over all distributions—a very high or infinite-dimensional space—but rather over the much lower-dimensional space of mean parameters. As an important consequence, solving this variational principle yields not only the cumulant generating function but also the full set of mean parameters $\mu = \{\mu_\alpha \mid \alpha \in \mathcal{I}\}$.

11.4.1 Conjugate duality

The cornerstone of our variational principle is the notion of *conjugate duality*. In this section, we provide a brief introduction to this concept, and refer the interested reader to the standard texts [Rockafellar (1970); Hiriart-Urruty and Lemaréchal (1993)] for further details. As is standard in convex analysis, we consider *extended* real-valued functions, meaning that they take values in the extended real line $\mathbb{R}_* := \mathbb{R} \cup \{+\infty\}$. Associated with any convex function $f : \mathbb{R}^d \to \mathbb{R}_*$ is a conjugate dual function $f^* : \mathbb{R}^d \to \mathbb{R}_*$, which is defined as follows:

$$f^*(y) := \sup_{x \in \mathbb{R}^d} \{\langle y, x\rangle - f(x)\}. \tag{11.24}$$

This definition illustrates the concept of a *variational definition*: the function value $f^*(y)$ is specified as the solution of an optimization problem parameterized by the vector $y \in \mathbb{R}^d$.

As illustrated in Figure 11.4, the value $f^*(y)$ has a natural geometric interpretation as the (negative) intercept of the hyperplane with normal $(y, -1)$ that supports the epigraph of f. In particular, consider the family of hyperplanes of the form $\langle y, x\rangle - c$, where y is a fixed normal direction and $c \in \mathbb{R}$ is the intercept to be adjusted. Our goal is to find the smallest c such that the resulting hyperplane supports the epigraph of f. Note that the hyperplane $\langle y, x\rangle - c$ lies below the epigraph of f if and only if the inequality $\langle y, x\rangle - c \leq f(x)$ holds for all $x \in \mathbb{R}^d$.

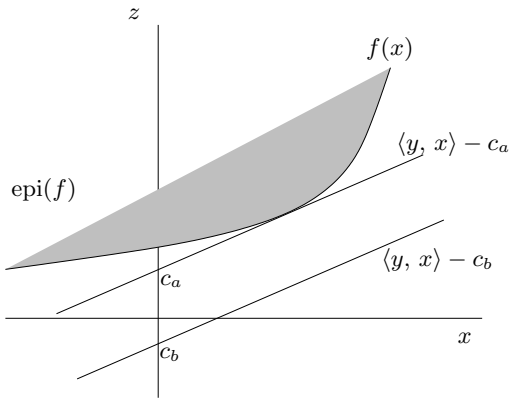


Figure 11.4 Interpretation of conjugate duality in terms of supporting hyperplanes to the epigraph of f, defined as $\text{epi}(f) := \{(x, y) \in \mathbb{R}^d \times \mathbb{R} \mid f(x) \leq y\}$. The dual function is obtained by translating the family of hyperplane with normal y and intercept $-c$ until it just supports the epigraph of f (the shaded region).

Moreover, it can be seen that the smallest c for which this inequality is valid is given by $c^* = \sup_{x \in \mathbb{R}^d} \{\langle y, x \rangle - f(x)\}$, which is precisely the value of the dual function. As illustrated in Figure 11.4, the geometric interpretation is that of moving the hyperplane (by adjusting the intercept c) until it is just tangent to the epigraph of f.

For convex functions meeting certain technical conditions, taking the dual *twice* recovers the original function. In analytical terms, this fact means that we can generate a variational representation for convex f in terms of its dual function as follows:

$$f(x) = \sup_{y \in \mathbb{R}^d} \{\langle x, y \rangle - f^*(y)\}. \tag{11.25}$$

Our goal in the next few section is to apply conjugacy to the cumulant generating function A associated with an exponential family, as defined in equation (11.15). More specifically, its dual function takes the form

$$A^*(\mu) := \sup_{\theta \in \Theta} \{\langle \theta, \mu \rangle - A(\theta)\}, \tag{11.26}$$

where we have used the fact that, by definition, the function value $A(\theta)$ is finite only if $\theta \in \Theta$. Here $\mu \in \mathbb{R}^d$ is a vector of so-called dual variables of the same dimension as θ. Our choice of notation—using μ for the dual variables—is deliberately suggestive: as we will see momentarily, these dual variables turn out to be precisely the mean parameters defined in equation (11.17).

Example 11.7

To illustrate the computation of a dual function, consider a scalar Bernoulli random variable $x \in \{0, 1\}$, whose distribution can be written in the exponential family

form as $p(x;\theta) = \exp\{\theta x - A(\theta)\}$. The cumulant generating function is given by $A(\theta) = \log[1 + \exp(\theta)]$, and there is a single dual variable $\mu = \mathbb{E}_\theta[x]$. Thus, the variational problem (11.26) defining A^* takes the form:

$$A^*(\mu) \;=\; \sup_{\theta \in \mathbb{R}} \big\{\theta\mu - \log[1 + \exp(\theta)]\big\}. \tag{11.27}$$

If $\mu \in (0,1)$, then taking derivatives shows that the supremum is attained at the unique $\theta \in \mathbb{R}$ satisfying the well-known logistic relation $\theta = \log[\mu/(1-\mu)]$. Substituting this logistic relation into equation (11.27) yields that for $\mu \in (0,1)$, we have $A^*(\mu) = \mu \log \mu + (1-\mu)\log(1-\mu)$. By taking limits $\mu \to 1-$ and $\mu \to 0+$, it can be seen that this expression is valid for μ in the closed interval $[0,1]$.

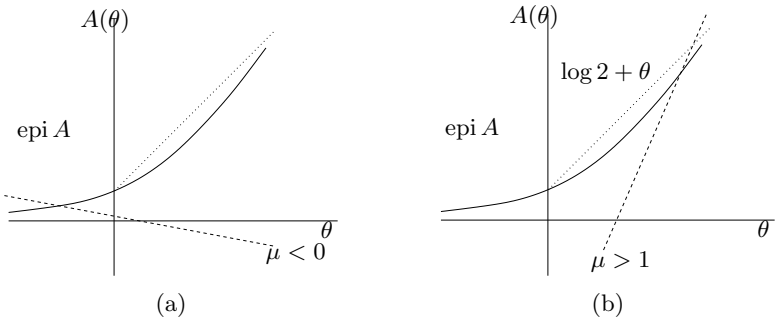


Figure 11.5 Behavior of the supremum defining $A^*(\mu)$ for (a) $\mu < 0$ and (b) $\mu > 1$. The value of the dual function corresponds to the negative intercept of the supporting hyperplane to $\operatorname{epi} A$ with slope μ.

Figure 11.5 illustrates the behavior of the supremum (11.27) for $\mu \notin [0,1]$. From our geometric interpretation of the value $A^*(\mu)$ in terms of supporting hyperplanes, the dual value is $+\infty$ if no supporting hyperplane can be found. In this particular case, the log partition function $A(\theta) = \log[1 + \exp(\theta)]$ is bounded below by the line $\theta = 0$. Therefore, as illustrated in Figure 11.5(a), any slope $\mu < 0$ cannot support $\operatorname{epi} A$, which implies that $A^*(\mu) = +\infty$. A similar picture holds for the case $\mu > 1$, as shown in Figure 11.5(b). Consequently, the dual function is equal to $+\infty$ for $\mu \notin [0,1]$. ♢

As the preceding example illustrates, there are two aspects to characterizing the dual function A^*:

(a) determining its domain (i.e., the set on which it takes a finite value)

(b) specifying its precise functional form on the domain.

In Example 11.7, the domain of A^* is simply the closed interval $[0,1]$, and its functional form on its domain is that of the binary entropy function. In the following two sections, we consider each of these aspects in more detail for general graphical

models in exponential form.

11.4.2 Sets of realizable mean parameters

For a given $\mu \in \mathbb{R}^d$, consider the optimization problem on the right-hand side of equation (11.26): since the cost function is differentiable, a first step in the solution is to take the derivative with respect to θ and set it equal to zero. Doing so yields the zero-gradient condition:

$$\mu \;=\; \nabla A(\theta) \;=\; \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})], \tag{11.28}$$

where the second equality follows from the standard properties of A given in Lemma 11.1.

We now need to determine the set of $\mu \in \mathbb{R}^d$ for which equation (11.28) has a solution. Observe that any $\mu \in \mathbb{R}^d$ satisfying this equation has a natural interpretation as a *globally realizable mean parameter*—i.e., a vector that can be realized by taking expectations of the sufficient statistic vector $\boldsymbol{\phi}$. This observation motivates defining the following set

$$\mathcal{M} \;:=\; \Big\{ \mu \in \mathbb{R}^d \;\Big|\; \exists \;\; p(\cdot) \;\; \text{such that} \;\; \int \boldsymbol{\phi}(\mathbf{x}) p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x}) = \mu \Big\}, \tag{11.29}$$

which corresponds to all realizable mean parameters associated with the set of sufficient statistics $\boldsymbol{\phi}$.

Example 11.8 Gaussian mean parameters
The Gaussian MRF, first introduced in Example 11.3, provides a simple illustration of the set $\mathcal{M}$. Given the sufficient statistics that define a Gaussian, the associated mean parameters are either first-order moments (e.g., $\mu_s = \mathbb{E}[x_s]$), or second-order moments (e.g., $\mu_{ss} = \mathbb{E}[x_s^2]$ and $\mu_{st} = \mathbb{E}[x_s x_t]$). This full collection of mean parameters can be compactly represented in matrix form:

$$W(\mu) \;:=\; \mathbb{E}_\theta \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix} \;=\; \begin{bmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_n \\ \mu_1 & \mu_{11} & \mu_{12} & \cdots & \mu_{1n} \\ \mu_2 & \mu_{21} & \mu_{22} & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_n & \mu_{n1} & \mu_{n2} & \cdots & \mu_{nn} \end{bmatrix} \tag{11.30}$$

The Schur product lemma [Horn and Johnson (1985)] implies that $\det W(\mu) = \det \operatorname{cov}(\mathbf{x})$, so that a mean parameter vector $\mu = \big\{\mu_s \mid s \in V\big\} \cup \big\{\mu_{st} \mid (s,t) \in E\big\}$ is globally realizable if and only if the matrix $W(\mu)$ is strictly positive definite. Thus, the set $\mathcal{M}$ is straightforward to characterize in the Gaussian case. $\diamondsuit$

Example 11.9 Marginal polytopes
We now consider the case of a multinomial MRF, first introduced in Example 11.4. With the choice of sufficient statistics (11.20), the associated mean parameters are

simply local marginal probabilities—viz.:

$$\mu_{s;j} := p(x_s = j; \theta) \ \ \forall\, s \in V, \quad \mu_{st;jk} := p((x_s, x_t) = (j,k); \theta) \ \ \forall\, (s,t) \in E \quad (11.31)$$

In analogy to our earlier definition of $\theta_s(x_s)$, we define functional versions of the mean parameters as follows:

$$\mu_s(x_s) := \sum_{j \in \mathcal{X}_s} \mu_{s;j} \mathbb{I}_j(x_s), \quad \mu_{st}(x_s, x_t) := \sum_{(j,k) \in \mathcal{X}_s \times \mathcal{X}_t} \mu_{st;jk} \mathbb{I}_{jk}(x_s, x_t). \quad (11.32)$$

With this notation, the set $\mathcal{M}$ consists of all singleton marginals μ_s (as s ranges over V) and pairwise marginals μ_{st} (for edges (s,t) in the edge set E) that can be realized by a distribution with support on $\mathcal{X}^n$. Since the space $\mathcal{X}^n$ has a finite

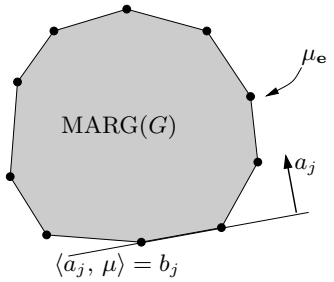


Figure 11.6 Geometrical illustration of a marginal polytope. Each vertex corresponds to the mean parameter $\mu_{\mathbf{e}} := \boldsymbol{\phi}(\mathbf{e})$ realized by the distribution $\delta_{\mathbf{e}}(\mathbf{x})$ that puts all of its mass on the configuration $\mathbf{e} \in \mathcal{X}^n$. The faces of the marginal polytope are specified by hyperplane constraints $\langle a_j, \mu \rangle \leq b_j$.

number of elements, the set $\mathcal{M}$ is formed by taking the convex hull of a finite number of vectors. As a consequence, it must be a *polytope*, meaning that it can be described by a finite number of linear inequality constraints. In this discrete case, we refer to $\mathcal{M}$ as a *marginal polytope*, denoted by MARG(G); see Figure 11.6 for an idealized illustration.

As discussed in Section 11.5.2, it is straightforward to specify a set of necessary conditions, expressed in terms of local constraints, that any element of MARG(G) must satisfy. However—and in sharp contrast to the Gaussian case—characterizing the marginal polytope exactly for a general graph is intractable, as it must require an exponential number of linear inequality constraints. Indeed, if it were possible to characterize MARG(G) with polynomial-sized set of constraints, then this would imply the polynomial-time solvability of various NP-complete problems (see Section (11.8.2) for further discussion of this point). $\diamondsuit$

11.4.3 Entropy in terms of mean parameters

We now turn to the second aspect of the characterization of the conjugate dual function A^*—that of specifying its precise functional form on its domain $\mathcal{M}$. As might be expected from our discussion of maximum entropy in Section 11.3.1, the form of the dual function A^* turns out to be closely related to entropy. Accordingly, we begin by defining the entropy in a bit more generality: Given a density function p taken with respect to base measure $\boldsymbol{\nu}$, its entropy is given by

$$H(p) = -\int_{\mathcal{X}^n} p(\mathbf{x}) \log [p(\mathbf{x})] \, \boldsymbol{\nu}(d\mathbf{x}) \;=\; -\mathbb{E}_p[\log p(\mathbf{x})]. \tag{11.33}$$

With this set-up, now suppose that μ belongs to the interior of $\mathcal{M}$. Under this assumption, it can be shown [Brown (1986); Wainwright and Jordan (2003a)] that there exists an canonical parameter $\theta(\mu) \in \Theta$ such that

$$\mathbb{E}_{\theta(\mu)}[\boldsymbol{\phi}(\mathbf{x})] = \mu. \tag{11.34}$$

Substituting this relation into the definition (11.26) of the dual function yields

$$A^*(\mu) \;=\; \langle \mu, \theta(\mu) \rangle - A(\theta(\mu)) \;=\; \mathbb{E}_{\theta(\mu)}\big[\log p(\mathbf{x}; \theta(\mu))\big],$$

which we recognize as the negative entropy $-H(p(\mathbf{x}; \theta(\mu)))$, where μ and $\theta(\mu)$ are dually coupled via equation (11.34).

Summarizing our development thus far, we have established that the dual function A^* has the following form:

$$A^*(\mu) = \begin{cases} -H(p(\mathbf{x}; \theta(\mu))) & \text{if } \mu \text{ belongs to the interior of } \mathcal{M} \\ +\infty & \text{if } \mu \text{ is outside the closure of } \mathcal{M}. \end{cases} \tag{11.35}$$

An alternative way to interpret this dual function A^* is by returning to the maximum entropy problem originally considered in Section 11.3.1. More specifically, suppose that we consider the optimal value of the maximum entropy problem (11.12), considered parametrically as a function of the constraints μ. Essentially, what we have established that the parametric form of this optimal value function is the dual function—that is:

$$A^*(\mu) \;=\; \max_{p \in \mathcal{P}} H(p) \quad \text{such that } \mathbb{E}_p[\phi_\alpha(\mathbf{x})] = \mu_\alpha \text{ for all } \alpha \in \mathcal{I}. \tag{11.36}$$

In this context, the property that $A^*(\mu) = +\infty$ for a constraint vector μ outside of $\mathcal{M}$ has a concrete interpretation: it corresponds to *infeasibility* of the maximum entropy problem (11.12).

11.4.3.1 Exact variational principle

Given the form (11.35) of the dual function, we can now use the conjugate dual relation (11.25) to express A in terms of an optimization problem involving its dual

function and the mean parameters:

$$A(\theta) \;=\; \sup_{\mu \in \mathcal{M}} \big\{ \langle \theta, \mu \rangle - A^*(\mu) \big\}. \tag{11.37}$$

Note that the optimization is restricted to the set $\mathcal{M}$ of globally realizable mean parameters, since the dual function A^* is infinite outside of this set. Thus, we have expressed the cumulant generating function as the solution of an optimization problem that is convex (since it entails maximizing a concave function over the convex set $\mathcal{M}$), and low-dimensional (since it is expressed in terms of the mean parameters $\mu \in \mathbb{R}^d$).

In addition to representing the value $A(\theta)$ of the cumulant generating function, the variational principle (11.35) also has another important property. More specifically, the nature of our dual construction ensures that the optimum is always attained at the vector of mean parameters $\mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})]$. Consequently, solving this optimization problem yields both the value of the cumulant generating function *as well as* the full set of mean parameters. In this way, the variational principle (11.37) based on exponential families differs fundamentally from the classical free energy principle from statistical physics.

11.5 Exact inference in variational form

In order to illustrate the general variational principle (11.37), it is worthwhile considering important cases in which it can be solved exactly. Accordingly, this section treats in some detail the case of a Gaussian MRF on an arbitrary graph— for which we re-derive the normal equations— as well as the case of a multinomial MRF on a tree, for which we sketch out a derivation of the sum-product algorithm from a variational perspective. In addition to providing a novel perspective on exact methods, the variational principle (11.37) also underlies a variety of methods for approximate inference, as we will see in Section 11.6.

11.5.1 Exact inference in Gaussian MRFs

We begin by considering the case of a Gaussian Markov random field (MRF) on an arbitrary graph, as discussed in Examples 11.3 and 11.8. In particular, we showed in the latter example that the set $\mathcal{M}_{Gauss}$ of realizable Gaussian mean parameters μ is determined by a positive definiteness constraint on the matrix $W(\mu)$ of mean parameters defined in equation (11.30).

We now consider the form of the dual function $A^*(\mu)$. It is well-known [Cover and Thomas (1991)] that the entropy of a multivariate Gaussian random vector can be written as

$$H(p) \;=\; \frac{1}{2} \log \det \operatorname{cov}(\mathbf{x}) + \frac{n}{2} \log 2\pi e,$$

where $\mathrm{cov}(\mathbf{x})$ is the $n \times n$ covariance matrix of $\mathbf{x}$. By recalling the definition (11.30) of $W(\mu)$ and applying the Schur complement formula [Horn and Johnson (1985)], we see that $\det \mathrm{cov}(\mathbf{x}) = \det W(\mu)$, which implies that the dual function for a Gaussian can be written in the form

$$A^*_{Gauss}(\mu) = -\frac{1}{2} \log \det W(\mu) - \frac{n}{2} \log 2\pi e, \tag{11.38}$$

valid for all $\mu \in \mathcal{M}_{Gauss}$. (To understand the negative signs, recall from equation (11.35) that A^* is equal to the negative entropy for $\mu \in \mathcal{M}_{Gauss}$.) Combining this exact expression for A^*_{Gauss} with our characterization of $\mathcal{M}_{Gauss}$ leads to

$$A_{Gauss}(\theta) = \sup_{W(\mu) \succ 0,\ W_{11}(\mu)=1} \big\{ \langle U(\theta),\ W(\mu) \rangle + \frac{1}{2} \log \det W(\mu) + \frac{n}{2} \log 2\pi e \big\}, \tag{11.39}$$

which corresponds to the variational principle (11.37) specialized to the Gaussian case.

We now show how solving the optimization problem (11.39) leads to the *normal equations* for Gaussian inference. In order to do so, it is convenient to introduce the following notation for different blocks of the matrices $W(\mu)$ and $U(\theta)$:

$$W(\mu) \ = \ \begin{bmatrix} 1 & z^T(\mu) \\ z(\mu) & Z(\mu) \end{bmatrix}, \qquad\qquad U(\theta) \ = \ \begin{bmatrix} 0 & z^T(\theta) \\ z(\theta) & Z(\theta) \end{bmatrix}. \tag{11.40}$$

In this definition, the submatrices $Z(\mu)$ and $Z(\theta)$ are $n \times n$, whereas $z(\mu)$ and $z(\theta)$ are $n \times 1$ vectors.

Now if $W(\mu) \succ 0$ were the only constraint in problem (11.39), then, using the fact that $\nabla \log \det W = W^{-1}$ for any symmetric positive matrix W, the optimal solution to problem (11.39) would simply be $W(\mu) = -2[U(\theta)]^{-1}$. Accordingly, if we enforce the constraint $[W(\mu)]_{11} = 1$ using a Lagrange multiplier λ, then it follows from the Karush-Kuhn-Tucker conditions [Bertsekas (1995)] that the optimal solution will assume the form $W(\mu) = -2[U(\theta) + \lambda^* E_{11}]^{-1}$, where λ^* is the optimal setting of the Lagrange multiplier and E_{11} is an $(n+1) \times (n+1)$ matrix with a one in the upper left hand corner, and zero in all other entries. Using the standard formula for the inverse of a block-partitioned matrix [Horn and Johnson (1985)], it is straightforward to verify that the blocks in the optimal $W(\mu)$ are related to the blocks of $U(\theta)$ by the relations:

$$Z(\mu) - z(\mu) z^T(\mu) = -2[Z(\theta)]^{-1} \tag{11.41a}$$

$$z(\mu) = -[Z(\theta)]^{-1}\ z(\theta) \tag{11.41b}$$

(The multiplier λ^* turns out not to be involved in these particular blocks.) In order to interpret these relations, it is helpful to return to the definition of $U(\theta)$ given in equation (11.18), and the Gaussian density of equation (11.19). In this way, we see that the first part of equation (11.41) corresponds to the fact that the covariance matrix is the inverse of the precision matrix, whereas the second part corresponds to the normal equations for the mean $z(\mu)$ of a Gaussian. Thus, as a special case of

the general variational principle (11.37), we have re-derived the familiar equations for Gaussian inference.

It is worthwhile noting that the derivation did not exploit any particular features of the graph structure. The Gaussian case is remarkable in this regard, in that both the dual function A^* and the set $\mathcal{M}$ of realizable mean parameters can be characterized simply for an arbitrary graph. However, many methods for solving the normal equations (11.41) as efficiently as possible, including Kalman filtering on trees [Willsky (2002)], make heavy use of the underlying graphical structure.

11.5.2 Exact inference on trees

We now turn to the case of tree-structured Markov random fields, focusing for concreteness on the multinomial case, first introduced in Example 11.4 and treated in more depth in Example 11.9. Recall from the latter example that for a multinomial MRF, the set $\mathcal{M}$ of realizable mean parameters corresponds to a marginal polytope, which we denote by $\mathrm{MARG}(G)$.

There is an obvious set of local constraints that any member of $\mathrm{MARG}(G)$ must satisfy. For instance, given their interpretation as local marginal distributions, the vectors μ_s and μ_{st} must of course be non-negative. In addition, they must satisfy normalization conditions (i.e., $\sum_{x_s} \mu_s(x_s) = 1$), and the pairwise marginalization conditions (i.e., $\sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s)$). Accordingly, we define for any graph G the following constraint set:

$$\mathrm{LOCAL}(G) \;:=\; \{\, \mu \geq 0 \mid \sum_{x_s} \mu_s(x_s) = 1, \;\; \sum_{x_t} \mu_{st}(x_s, x_t) = \mu_s(x_s) \;\forall (s,t) \in E \}. \tag{11.42}$$

Since any set of singleton and pairwise marginals (regardless of the underlying graph structure) must satisfy these local consistency constraints, we are guaranteed that $\mathrm{MARG}(G) \subseteq \mathrm{LOCAL}(G)$ for *any* graph G. This fact plays a significant role in our later discussion in Section 11.7 of the Bethe variational principle and sum-product on graphs with cycles. Of most importance to the current development is the following consequence of the junction tree theorem (see Section 11.2.2.2): when the graph G is tree-structured, then $\mathrm{LOCAL}(T) = \mathrm{MARG}(T)$. Thus, the marginal polytope $\mathrm{MARG}(T)$ for trees has a very simple description (11.42).

The second component of the exact variational principle (11.37) is the dual function A^*. Here the junction tree framework is useful again: in particular specializing representation (11.7) to a tree yields the following factorization

$$p(\mathbf{x}; \mu) \;=\; \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)} \tag{11.43}$$

for a tree-structured distribution in terms of its mean parameters μ_s and μ_{st}.

From this decomposition, it is straightforward to compute the entropy *purely* as a function of the mean parameters by taking the logarithm, expectations and

simplifying. Doing so yields the expression

$$-A^*(\mu) \;=\; \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st}) \tag{11.44}$$

where the singleton entropy H_s and mutual information I_{st} are given by

$$H_s(\mu_s) := -\sum_{x_s} \mu_s(x_s)\log\mu_s(x_s), \qquad I_{st}(\mu_{st}) := \sum_{x_s,x_t} \mu_{st}(x_s,x_t)\log\frac{\mu_{st}(x_s,x_t)}{\mu_s(x_s)\mu_t(x_t)},$$

respectively. Putting the pieces together, the general variational principle (11.37) takes the following particular form:

$$A(\theta) \;=\; \max_{\mu\in \mathrm{LOCAL}(T)} \Big\{ \langle\theta,\, \mu\rangle + \sum_{s\in V} H_s(\mu_s) - \sum_{(s,t)\in E} I_{st}(\mu_{st}) \Big\}. \tag{11.45}$$

There is an important link between this variational principle for multinomial MRFs on trees, and the sum-product updates (11.6). In particular, the sum-product updates can be derived as an iterative algorithm for solving a Lagrangian dual formulation of the problem (11.45). This will be clarified in our discussion of the Bethe variational principle in Section 11.7.

11.6 Approximate inference in variational form

Thus far, we have seen how well-known methods for exact inference—specifically, the computation of means and covariances in the Gaussian case and the computation of local marginal distributions by the sum-product algorithm for tree-structured problems—can be re-derived from the general variational principle (11.37). It is worthwhile isolating the properties that permit an exact solution of the variational principle. First, for both of the preceding cases, it is possible to characterize the set $\mathcal{M}$ of globally realizable mean parameters in a straightforward manner. Second, the entropy can be expressed as a closed-form function of the mean parameters μ, so that the dual function $A^*(\mu)$ has an explicit form.

Neither of these two properties hold for a general graphical model in exponential form. As a consequence, there are significant challenges associated with exploiting the variational representation. More precisely, in contrast to the simple cases discussed thus far, many graphical models of interest have the following properties:

(a) the constraint set $\mathcal{M}$ of realizable mean parameters is extremely difficult to characterize in an explicit manner.

(b) the negative entropy function A^* is defined indirectly—in a variational manner—so that it too typically lacks an explicit form.

These difficulties motivate the use of approximations to $\mathcal{M}$ and A^*. Indeed, a broad class of methods for approximate inference—ranging from mean field theory to cluster variational methods—are based on this strategy. Accordingly, the remainder

of the chapter is devoted to discussion of approximate methods based on relaxations of the exact variational principle.

11.6.1 Mean field theory

We begin our discussion of approximate algorithms with mean field methods, a set of algorithms with roots in statistical physics [Chandler (1987)]. Working from the variational principle (11.37), we show that mean field methods can be understood as solving an approximation thereof, with the essential restriction that the optimization is limited to a subset of distributions for which the dual function A^* is relatively easy to characterize. Throughout this section, we will refer to a distribution with this property as a *tractable* distribution.

11.6.1.1 Tractable families

Let H represent a subgraph of G over which it feasible to perform exact calculations (e.g., a graph with small treewidth); we refer to any such H as a *tractable subgraph*. In an exponential formulation, the set of all distributions that respect the structure of H can be represented by a linear subspace of canonical parameters. More specifically, letting $\mathcal{I}(H)$ denote the subset of indices associated with cliques in H, the set of canonical parameters corresponding to distributions structured according to H is given by:

$$\mathcal{E}(H) := \{\theta \in \Theta \mid \theta_\alpha = 0 \quad \forall \ \alpha \in \mathcal{I} \backslash \mathcal{I}(H)\}. \tag{11.46}$$

We consider some examples to illustrate:

Example 11.10 Tractable subgraphs
The simplest instance of a tractable subgraph is the completely disconnected graph $H_0 = (V, \emptyset)$ (see Figure 11.7(b)). Permissible parameters belong to the subspace $\mathcal{E}(H_0) := \{\theta \in \Theta \mid \theta_{st} = 0 \ \forall \ (s,t) \in E\}$, where θ_{st} refers to the collection of canonical parameters associated with edge (s,t). The associated distributions are of the product form $p(\mathbf{x};\theta) = \prod_{s \in V} p(x_s;\theta_s)$ where θ_s refers to the collection of canonical parameters associated with vertex s.

To obtain a more structured approximation, one could choose a spanning tree $T = (V, E(T))$, as illustrated in Figure 11.7(c). In this case, we are free to choose the canonical parameters corresponding to vertices and edges in T, but we must set to zero any canonical parameters corresponding to edges not in the tree. Accordingly, the subspace of tree-structured distributions is given by $\mathcal{E}(T) = \{\theta \mid \theta_{st} = 0 \ \forall \ (s,t) \notin E(T)\}$. $\diamondsuit$

For a given subgraph H, consider the set of all possible mean parameters that are realizable by tractable distributions:

$$\mathcal{M}_{tract}(G;H) \ := \ \{\mu \in \mathbb{R}^d \mid \mu = \mathbb{E}_\theta[\boldsymbol{\phi}(\mathbf{x})] \text{ for some } \theta \in \mathcal{E}(H)\}. \tag{11.47}$$

The notation $\mathcal{M}_{tract}(G;H)$ indicates that mean parameters in this set arise from taking expectations of sufficient statistics associated with the graph G, but that they must be realizable by a tractable distribution—i.e., one that respects the structure of H. See Example 11.11 for an explicit illustration of this set when the tractable subgraph H is the fully disconnected graph. Since any μ that arises from a tractable distribution is certainly a valid mean parameter, the inclusion $\mathcal{M}_{tract}(G;H) \subseteq \mathcal{M}(G)$ always holds. In this sense, $\mathcal{M}_{tract}$ is an *inner approximation* to the set $\mathcal{M}$ of realizable mean parameters.

11.6.1.2 Optimization and lower bounds

We now have the necessary ingredients to develop the mean field approach to approximate inference. Let $p(\mathbf{x};\theta)$ denote the *target distribution* that we are interested in approximating. The basis of the mean field method is the following fact: any valid mean parameter specifies a lower bound on the cumulant generating function. Indeed, as an immediate consequence of the variational principle (11.37), we have:

$$A(\theta) \;\geq\; \langle \theta,\, \mu \rangle - A^*(\mu). \tag{11.48}$$

for any $\mu \in \mathcal{M}$. This inequality can also be established by applying Jensen's inequality [Jordan et al. (1999)].

Since the dual function A^* typically lacks an explicit form, it is not possible, at least in general, to compute the lower bound (11.48). The mean field approach circumvents this difficulty by restricting the choice of μ to the tractable subset $\mathcal{M}_{tract}(G;H)$, for which the dual function has an explicit form A^*_H. As long as μ belongs to $\mathcal{M}_{tract}(G;H)$, then the lower bound (11.48) will be computable.

Of course, for a non-trivial class of tractable distributions, there are many such bounds. The goal of the mean field method is the natural one: find the best approximation μ^{MF}, as measured in terms of the tightness of the bound. This optimal approximation is specified as the solution of the optimization problem

$$\sup_{\mu \in \mathcal{M}_{tract}(G;H)} \big\{ \langle \mu,\, \theta \rangle - A^*_H(\mu) \big\}, \tag{11.49}$$

which is a relaxation of the exact variational principle (11.37). The optimal value specifies a lower bound on $A(\theta)$, and it is (by definition) the best one that can be obtained by using a distribution from the tractable class.

An important alternative interpretation of the mean field approach is in terms of minimizing the Kullback-Leibler (KL) divergence between the approximating (tractable) distribution and the target distribution. Given two densities p and q, the KL divergence is given by

$$D(p \,\|\, q) \;=\; \int_{\mathcal{X}^n} \log \frac{p(\mathbf{x})}{q(\mathbf{x})} p(\mathbf{x}) \boldsymbol{\nu}(d\mathbf{x}). \tag{11.50}$$

To see the link to our derivation of mean field, consider for a given mean parameter $\mu \in \mathcal{M}_{tract}(G;H)$, the difference between the log partition function $A(\theta)$ and the

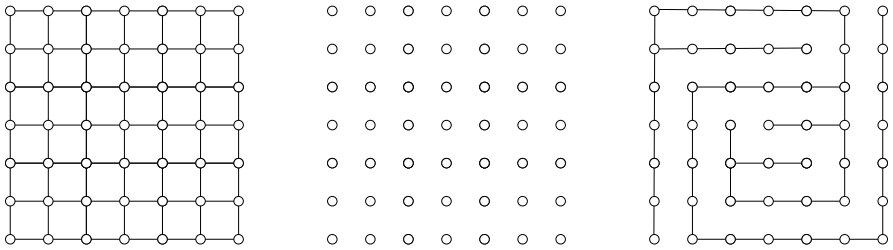


Figure 11.7 Graphical illustration of the mean field approximation. (a) Original graph is a 7×7 grid. (b) Fully disconnected graph, corresponding to a naive mean field approximation. (c) A more structured approximation based on a spanning tree.

quantity $\langle \mu, \theta \rangle - A_H^*(\mu)$:

$$D(\mu \parallel \theta) \;=\; A(\theta) + A_H^*(\mu) - \langle \mu, \theta \rangle.$$

A bit of algebra shows that this difference is equal to the KL divergence (11.50) with $q = p(\mathbf{x}; \theta)$ and $p = p(\mathbf{x}; \mu)$ (i.e., the exponential family member with mean parameter μ). Therefore, solving the mean field variational problem (11.49) is equivalent to minimizing the KL divergence subject to the constraint that μ belongs to tractable set of mean parameters, or equivalently that p is a tractable distribution.

11.6.2 Naive mean field updates

The *naive mean field* approach corresponds to choosing a fully factorized or product distribution in order to approximate the original distribution. The naive mean field updates are a particular set of recursions for finding a stationary point of the resulting optimization problem.

Example 11.11

As an illustration, we derive the naive mean field updates for the *Ising model*, which is a special case of the multinomial MRF defined in Example 11.4. It involves binary variables, so that $\mathcal{X}_s = \{0, 1\}$ for all vertices $s \in V$. Moreover, the canonical parameters are of the form $\theta_s(x_s) = \theta_s x_s$ and $\theta_{st}(x_s, x_t) = \theta_{st} x_s x_t$ for real numbers θ_s and θ_{st}. Consequently, the exponential representation of the Ising model has the form

$$p(\mathbf{x}; \theta) \;\propto\; \exp\Big\{ \sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t \Big\}.$$

Letting H_0 denote the fully disconnected graph (i.e., without any edges), the tractable set $\mathcal{M}_{tract}(G; H_0)$ consists of all mean parameters $\{\mu_s, \mu_{st}\}$ that arise

from a product distribution. Explicitly, in this binary case, we have

$$\mathcal{M}_{tract}(G; H_0) \;:=\; \{(\mu_s, \mu_{st}) \mid 0 \;\leq\; \mu_s \;\leq 1, \;\; \mu_{st} = \mu_s \; \mu_t \; \}.$$

Moreover, the negative entropy of a product distribution over binary random variables decomposes into the sum $A^*_{H_0}(\mu) = \sum_{s \in V} \big[\mu_s \log \mu_s + (1-\mu_s) \log(1-\mu_s)\big]$. Accordingly, the associated naive mean field problem takes the form

$$\max_{\mu \in \mathcal{M}_{tract}(G; H_0)} \big\{ \langle \mu, \, \theta \rangle - A^*_{H_0}(\mu) \big\}.$$

In this particular case, it is convenient to eliminate μ_{st} by replacing it by the product $\mu_s \mu_t$. Doing so leads to a reduced form of the problem:

$$\max_{\{\mu_s\} \in [0,1]^n} \Big\{ \sum_{s \in V} \theta_s \mu_s + \sum_{(s,t) \in E} \theta_{st} \mu_s \mu_t - \sum_{s \in V} \big[\mu_s \log \mu_s + (1 - \mu_s) \log(1 - \mu_s)\big] \Big\}. \tag{11.51}$$

Let F denote the function of μ within curly braces in equation (11.51). It can be seen that the function F is strictly concave in a given fixed coordinate μ_s when all the other coordinates are held fixed. Moreover, it is straightforward to show that the maximum over μ_s with $\mu_t, t \neq s$ fixed is attained in the interior $(0, 1)$, and can be found by taking the gradient and setting it equal to zero. Doing so yields the following update for μ_s:

$$\mu_s \;\leftarrow\; \sigma\big(\theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st} \mu_t\big), \tag{11.52}$$

where $\sigma(z) := [1 + \exp(-z)]^{-1}$ is the logistic function. Applying equation (11.52) iteratively to each node in succession amounts to performing coordinate ascent in the objective function for the mean field variational problem (11.51). Thus, we have derived the update equation presented earlier in equation (11.10). ♢

Similarly, it is straightforward to apply the naive mean field approximation to other types of graphical models, as we illustrate for a multivariate Gaussian.

Example 11.12 Gaussian mean field
The mean parameters for a multivariate Gaussian are of the form $\mu_s = \mathbb{E}[x_s]$, $\mu_{ss} = \mathbb{E}[x_s^2]$ and $\mu_{st} = \mathbb{E}[x_s x_t]$ for $s \neq t$. Using only Gaussians in product form, the set of tractable mean parameters takes the form

$$\mathcal{M}_{tract}(G; H_0) \;=\; \{\mu \in \mathbb{R}^d \mid \mu_{st} = \mu_s \mu_t \; \forall s \neq t, \;\; \mu_{ss} - \mu_s^2 > 0 \; \}.$$

As with naive mean field on the Ising model, the constraints $\mu_{st} = \mu_s \mu_t$ for $s \neq t$ can be imposed directly, thereby leaving only the inequality $\mu_{ss} - \mu_s^2 > 0$ for each node. The negative entropy of a Gaussian in product form can be written as $A^*_{Gauss}(\mu) = -\sum_{s=1}^n \frac{1}{2} \log(\mu_{ss} - \mu_s^2) - \frac{n}{2} \log 2\pi e$. Combining A^*_{Gauss} with the

constraints leads to the naive MF problem for a multivariate Gaussian:

$$\sup_{\{(\mu_s,\mu_{ss}) \mid \mu_{ss}-\mu_s^2>0\}} \big\{ \langle U(\theta),\ W(\mu) \rangle + \sum_{s=1}^{n} \frac{1}{2} \log(\mu_{ss} - \mu_s^2) + \frac{n}{2} \log 2\pi e \big\},$$

where the matrices $U(\theta)$ and $W(\mu)$ are defined in equation (11.40). Here it should be understood that any terms $\mu_{st}, s \neq t$ contained in $W(\mu)$ are replaced with the product $\mu_s \mu_t$.

Taking derivatives with respect to μ_{ss} and μ_s and re-arranging yields the stationary conditions $\frac{1}{2(\mu_{ss}-\mu_s^2)} = -\theta_{ss}$ and $\frac{\mu_s}{2(\mu_{ss}-\mu_s^2)} = \theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st}\mu_t$. Since $\theta_{ss} < 0$, we can combine both equations into the update $\mu_s \leftarrow -\frac{1}{\theta_{ss}} \big\{ \theta_s + \sum_{t \in \mathcal{N}(s)} \theta_{st}\mu_t \big\}$. In fact, the resulting algorithm is equivalent to the Gauss-Jacobi method for solving the normal equations, and so is guaranteed to converge under suitable conditions [Demmel (1997)], in which case the algorithm computes the correct mean vector $[\mu_1 \ldots \mu_n]$. ♢

11.6.3 Structured mean field and other extensions

Of course, the essential principles underlying the mean field approach are not limited to fully factorized distributions. More generally, one can consider classes of tractable distributions that incorporate additional structure. This *structured mean field approach* was first proposed by Saul and Jordan [1996], and further developed by various researchers. In this section, we discuss only particular example in order to illustrate the basic idea, and refer the interested reader elsewhere [Wiegerinck (2000); Wainwright and Jordan (2003a)] for further details.

Example 11.13 Structured MF for factorial HMMs
The factorial hidden Markov model, as described in Ghahramani and Jordan [1997], has the form shown in Figure 11.8(a). It consists of a set of M Markov chains ($M = 3$ in this diagram), which share at each time a common observation (shaded nodes). Such models are useful, for example, in modeling the joint dependencies between speech and video signals over time.

Although the separate chains are independent a priori, the common observation induces an effective coupling between all nodes at each time (a coupling which is captured by the moralization process mentioned earlier). Thus, an equivalent model is shown in panel (b), where the dotted ellipses represent the induced coupling of each observation.

A natural choice of approximating distribution in this case is based on the subgraph H consisting of the decoupled set of M chains, as illustrated in panel (c). The decoupled nature of the approximation yields valuable savings on the computational side. In particular, it can be shown [Saul and Jordan (1996); Wainwright and Jordan (2003a)] that all intermediate quantities necessary for implementing the structured mean field updates can be calculated by applying the forward-backward algorithm (i.e., the sum-product updates as an exact method) to each chain separately. ♢

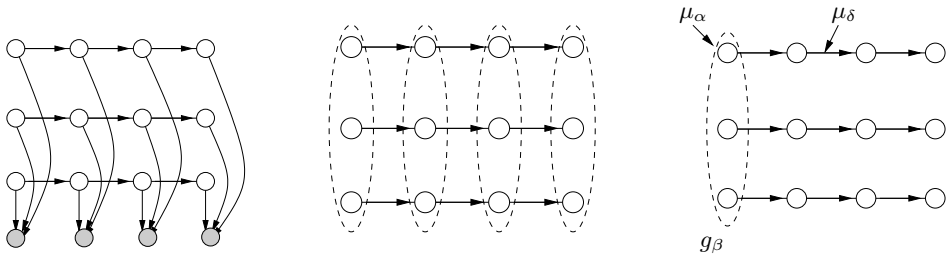


Figure 11.8 Structured mean field approximation for a factorial HMM. (a) Original model consists of a set of hidden Markov models (defined on chains), coupled at each time by a common observation. (b) An equivalent model, where the ellipses represent interactions among all nodes at a fixed time, induced by the common observation. (c) Approximating distribution formed by a product of chain-structured models. Here μ_α and μ_δ are the sets of mean parameters associated with the indicated vertex and edge respectively.

In addition to structured mean field, there are various other extensions to naive mean field, which we mention only in passing here. A large class of techniques, including linear response theory and the TAP method [Plefka (1982); Kappen and Rodriguez (1998); Opper and Saad (2001)], seek to improve the mean field approximation by introducing higher-order correction terms. Although the lower bound on the log partition function is not usually preserved by these higher-order methods, Leisinck and Kappen [2001] demonstrated how to generate tighter lower bounds based on higher-order expansions.

11.6.4 Geometric view of mean field

An important fact about the mean field approach is that the variational problem (11.49) may be non-convex, so that there may be local minima, and the mean field updates can have multiple solutions.

One way to understand this non-convexity is in terms of the set of tractable mean parameters: under fairly mild conditions, it can be shown [Wainwright and Jordan (2003a)] that the set $\mathcal{M}_{tract}(G; H)$ is non-convex. Figure 11.9 provides a geometric illustration for the case of a multinomial MRF, for which the set $\mathcal{M}$ is a marginal polytope.

A practical consequence of this non-convexity is that the mean field updates are often sensitive to the initial conditions. Moreover, the mean field method can exhibit "spontaneous symmetry breaking," wherein the mean field approximation is asymmetric even though the original problem is perfectly symmetric; see Jaakkola [2001] for an illustration of this phenomenon. Despite this non-convexity, the mean field approximation becomes exact for certain types of models as the number of nodes n grows to infinity [Baxter, 1982].

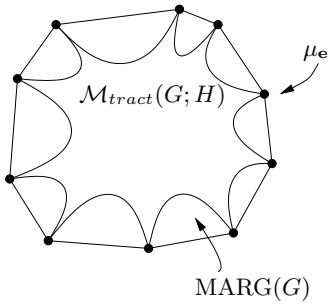


Figure 11.9 The set $\mathcal{M}_{tract}(G;H)$ of mean parameters that arise from tractable distributions is a non-convex inner bound on $\mathcal{M}(G)$. Illustrated here is the multinomial case where $\mathcal{M}(G) \equiv \text{MARG}(G)$ is a polytope. The circles correspond to mean parameters that arise from delta distributions with all their mass on a single configuration , and belong to both $\mathcal{M}(G)$ and $\mathcal{M}_{tract}(G;H)$.

11.6.5 Parameter estimation and variational EM

Mean field methods also play an important role in the problem of parameter estimation, in which the goal is to estimate model parameters on the basis of partial observations. The expectation-maximization (EM) algorithm [Dempster et al. (1977)] provides a general approach to maximum likelihood parameter estimation in the case in which some subset of variables are observed whereas others are unobserved. Although the EM algorithm is often presented as an alternation between an expectation step (E step) and a maximization step (M step), it is also possible to take a variational perspective on EM, and view both steps as maximization steps [Csiszár and Tusnády (1984); Neal and Hinton (1999)]. More concretely, in the exponential family setting, the E step reduces to the computation of expected sufficient statistics—i.e., mean parameters. As we have seen, the variational framework provides a general class of methods for computing approximations of mean parameters. This observation suggests a general class of *variational EM algorithms*, in which the approximation provided by a variational inference algorithm is substituted for the mean parameters in the E step. In general, as a consequence of making such a substitution, one loses the guarantees that are associated with the EM algorithm. In the specific case of mean field algorithms, however, a convergence guarantee is retained: in particular, the algorithm will converge to a stationary point of a lower bound for the likelihood function [Wainwright and Jordan (2003a)].

11.7 Bethe entropy approximation and sum-product algorithm

In this section, we turn to another important message-passing algorithm for approximate inference, known either as *belief propagation*, or the *sum-product algorithm*. In

Section 11.5.2, we described the use of the sum-product algorithm for trees, in which context it is guaranteed to converge and perform exact inference. When the same message-passing updates are applied to graphs with cycles, in contrast, there are no such guarantees; nonetheless, this "loopy" form of the sum-product algorithm is widely used to compute approximate marginals in various signal processing applications, including phase unwrapping [Frey et al. (2001)], low-level vision [Freeman et al. (2000)], and channel decoding [Richardson and Urbanke (2001)].

The main idea of this section is the connection between the sum-product updates and the Bethe variational principle. The presentation given here differs from the original work of Yedidia et al. [2001], in that we formulate the problem purely in terms of mean parameters and marginal polytopes. This perspective highlights a key point: mean field and sum-product, though similar as message-passing algorithms, are fundamentally different at the variational level. In particular, whereas the essence of mean field is to *restrict* optimization to a limited class of distributions for which the negative entropy and mean parameters can be characterized *exactly*, the the sum-product algorithm, in contrast, is based on *enlarging* the constraint set and *approximating* the entropy function.

The standard Bethe approximation applies to an undirected graphical model with potential functions involving at most pairs of variables, which we refer to as a *pairwise Markov random field*. In principle, by selectively introducing auxiliary variables, any undirected graphical model can be converted into an equivalent pairwise form to which the Bethe approximation can be applied; see Freeman and Weiss [2000] for a detailed description of this procedure. Moreover, although the Bethe approximation can be developed more generally, we also limit our discussion to a multinomial MRF, as discussed earlier in Examples 11.4 and 11.9. We also make use of the local marginal functions $\mu_s(x_s)$ and $\mu_{st}(x_s, x_t)$, as defined in equation (11.32). As discussed in Example 11.9, the set $\mathcal{M}$ associated with a multinomial MRF is the marginal polytope $\mathrm{MARG}(G)$.

Recall that there are two components to the general variational principle (11.37): the set of realizable mean parameters (given by a marginal polytope in this case), and the dual function A^*. Developing an approximation to the general principle requires approximations to both of these components, which we discuss in turn in the following sections.

11.7.1 Bethe entropy approximation

From equation (11.35), recall that dual function A^* corresponds to the maximum entropy distribution consistent with a given set of mean parameters; as such, it typically lacks a closed form expression. An important exception to this general rule is the case of a tree-structured distribution: as discussed in Section 11.5.2, the function A^* for a tree-structured distribution has a closed-form expression that is straightforward to compute; see, in particular, equation (11.44).

Of course, the entropy of a distribution defined by a graph with cycles will not, in general, decompose additively like that of a tree. Nonetheless, one can imagine

using the decomposition in equation (11.44) as an approximation to the entropy. Doing so yields an expression known as the *Bethe approximation* to the entropy on a graph with cycles:

$$H_{Bethe}(\mu) \; := \; \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E} I_{st}(\mu_{st}). \tag{11.53}$$

To be clear, the quantity $H_{Bethe}(\mu)$ is an approximation to the negative dual function $-A^*(\mu)$. Moreover, our development in Section 11.5.2 shows that this approximation is exact when the graph is tree-structured.

An alternative form of the Bethe entropy approximation can be derived by writing mutual information in terms of entropies as $I_{st}(\mu_{st}) = H_s(\mu_s) + H_t(\mu_t) - H_{st}(\mu_{st})$. In particular, expanding the mutual information terms in this way, and then collecting all the single node entropy terms yields $H_{Bethe}(\mu) = \sum_{s \in V}(1 - d_s)H_s(\mu_s) + \sum_{(s,t) \in E} H_{st}(\mu_{st})$, where d_s denotes the number of neighbors of node s. This representation is the form of the Bethe entropy introduced by Yedidia et al. [2001]; however, the form given in equation (11.53) turns out to be more convenient for our purposes.

11.7.2 Tree-based outer bound

Note that the Bethe entropy approximation H_{Bethe} is certainly well-defined for any $\mu \in \text{MARG}(G)$. However, as discussed earlier, characterizing this polytope of realizable marginals is a very challenging problem. Accordingly, a natural approach is to specify a subset of necessary constraints, which leads to an outer bound on $\text{MARG}(G)$. Let $\tau_s(x_s)$ and $\tau_{st}(x_s, x_t)$ be a set of candidate marginal distributions. In Section 11.5.2, we considered the following constraint set:

$$\text{LOCAL}(G) \; = \; \{ \, \tau \geq 0 \mid \sum_{x_s} \tau_s(x_s) = 1, \;\; \sum_{x_s} \tau_{st}(x_s, x_t) = \tau_t(x_t) \, \}. \tag{11.54}$$

Although $\text{LOCAL}(G)$ is an exact description of the marginal polytope for a tree-structured graph, it is only an outer bound for graphs with cycles. (We demonstrate this fact more concretely in Example 11.14.) For this reason, our change in notation—i.e., from μ to τ—is quite deliberate, with the goal of emphasizing that members τ of $\text{LOCAL}(G)$ need not be realizable. We refer to members of $\text{LOCAL}(G)$ as *pseudomarginals* (these are sometimes referred to as "beliefs").

Example 11.14 Pseudomarginals

We illustrate using a binary random vector on the simplest possible graph for which $\text{LOCAL}(G)$ is not an exact description of $\text{MARG}(G)$—namely, a single cycle with three nodes. Consider candidate marginal distributions $\{\tau_s, \tau_{st}\}$ of the form

$$\tau_s := \begin{bmatrix} 0.5 & 0.5 \end{bmatrix}, \qquad \tau_{st} := \begin{bmatrix} \beta_{st} & 0.5 - \beta_{st} \\ 0.5 - \beta_{st} & \beta_{st} \end{bmatrix}, \tag{11.55}$$

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 LOCAL(G) for any choice of $\beta_{st} \in [0, 0.5]$.

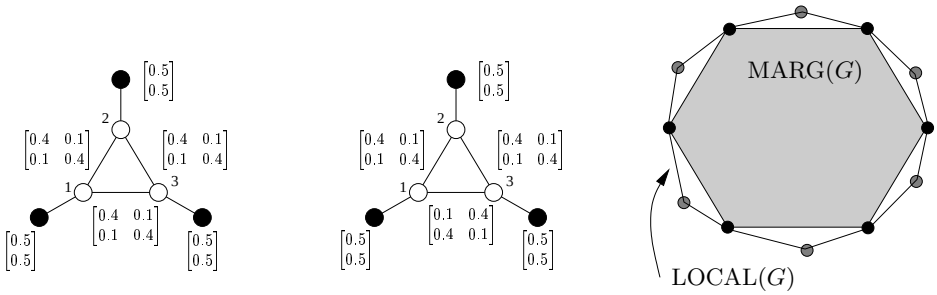


Figure 11.10 (a), (b): Illustration of the marginal polytope for a single cycle graph on three nodes. Setting $\beta_{st} = 0.4$ for all three edges gives a globally consistent set of marginals. (b) With β_{13} perturbed to 0.1, the marginals (though locally consistent) are no longer globally so. (c) For a more general graph, an idealized illustration of the tree-based constraint set LOCAL(G) as an outer bound on the marginal polytope MARG(G).

First, consider the setting $\beta_{st} = 0.4$ for all edges (s, t), as illustrated in panel (a). It is not difficult to show that the resulting marginals thus defined are realizable; in fact, they can be obtained from the distribution that places probability 0.35 on each of the configurations [0 0 0] and [1 1 1], and probability 0.05 on each of the remaining six configurations. Now suppose that we perturb one of the pairwise marginals—say τ_{13}—by setting $\beta_{13} = 0.1$. The resulting problem is illustrated in panel (b). Observe that there are now strong (positive) dependencies between the pairs of variables (x_1, x_2) and (x_2, x_3): both pairs are quite likely to agree (with probability 0.8). In contrast, the pair (x_1, x_3) can only share the same value relatively infrequently (with probability 0.2). This arrangement should provoke some doubt. Indeed, it can be shown that $\tau \notin$ MARG(G) by attempting but *failing* to construct a distribution that realizes τ, or alternatively and much more directly using the idea of semidefinite constraints (see Example 11.15). ♢

More generally, Figure 11.10(c) provides an idealized illustration of the constraint set LOCAL(G), and its relation to the exact marginal polytope MARG(G). Observe that the set LOCAL(G) is another polytope that is a *convex outer approximation* to MARG(G). It is worthwhile contrasting with the *non-convex inner approximation* used by a mean field approximation, as illustrated in Figure 11.9.

11.7.3 Bethe variational problem and sum-product

Note that the Bethe entropy is also well-defined for any pseudomarginal in LOCAL(G). Therefore, it is valid to consider a constrained optimization problem over the set LOCAL(G) in which the cost function involves the Bethe entropy approximation H_{Bethe}. Indeed, doing so leads to the so-called *Bethe variational problem*:

$$\max_{\tau \in \mathrm{LOCAL}(G)} \big\{ \langle \theta, \tau \rangle + \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} I_{st}(\tau_{st}) \big\}. \tag{11.56}$$

Although ostensibly similar to a (structured) mean field approach, the Bethe variational problem (BVP) is fundamentally different in a number of ways. First, as discussed in Section 11.6.1, a mean field method is based on an exact representation of the entropy, albeit over a limited class of distributions. In contrast, with the exception of tree-structured graphs, the Bethe entropy is a bona fide *approximation* to the entropy. For instance, it is not difficult to see that it can be negative, which of course can never happen for an exact entropy. Second, the mean field approach entails optimizing over an *inner bound* on the marginal polytope, which ensures that any mean field solution is always globally consistent with respect to at least one distribution, and that it yields a lower bound on the log partition function. In contrast, since LOCAL(G) is a strict outer bound on the set of realizable marginals MARG(G), the optimizing pseudomarginals τ^* of the BVP may not be globally consistent with any distribution.

11.7.4 Solving the Bethe variational problem

Having formulated the Bethe variational problem, we now consider iterative methods for solving it. Observe that the set LOCAL(G) is a polytope defined by $\mathcal{O}(n + |E|)$ constraints. A natural approach to solving the BVP, then, is to attach Lagrange multipliers to these constraints, and find stationary points of the Lagrangian. A remarkable fact, established by Yedidia et al. [2001], is that sum-product updates (11.6) can be re-derived as a method for trying to find such Lagrangian stationary points.

A bit more formally, for each $x_s \in \mathcal{X}_s$, let $\lambda_{st}(x_s)$ be a Lagrange multiplier associated with the constraint $C_{ts}(x_s) = 0$, where $C_{ts}(x_s) := \tau_s(x_s) - \sum_{x_t} \tau_{st}(x_s, x_t)$. Our approach is to consider the following partial Lagrangian corresponding to the Bethe variational problem (11.56):

$$\mathcal{L}(\tau; \lambda) \;:=\; \langle \theta, \tau \rangle + H_{Bethe}(\tau) + \sum_{(s,t) \in E} \big[\sum_{x_s} \lambda_{ts}(x_s) C_{ts}(x_s) + \sum_{x_t} \lambda_{st}(x_t) C_{st}(x_t) \big].$$

The key insight of Yedidia et al. [2001] is that any fixed point of the sum-product updates specifies a pair (τ^*, λ^*) such that:

$$\nabla_\tau \mathcal{L}(\tau^*; \lambda^*) \;=\; 0, \qquad\qquad \nabla_\lambda \mathcal{L}(\tau^*; \lambda^*) \;=\; 0 \tag{11.57}$$

In particular, the Lagrange multipliers can be used to specify messages of the form $M_{ts}(x_s) = \exp(\lambda_{ts}(x_s))$. After taking derivatives of the Lagrangian and equating them to zero, some algebra then yields the familiar message-update rule:

$$M_{ts}(x_s) \;=\; \kappa \sum_{x_t} \exp\left\{\theta_{st}(x_s, x_t) + \theta_t(x_t)\right\} \prod_{u \in \mathcal{N}(t)\setminus s} M_{ut}(x_t). \tag{11.58}$$

We refer the reader to Yedidia et al. [2001] or Wainwright and Jordan [2003a] for further details of this derivation. By construction, any fixed point M^* of these updates specifies a pair (τ^*, λ^*) that satisfies the stationary[3] conditions (11.57).

This variational formulation of the sum-product updates—namely, as an algorithm for solving a constrained optimization problem—has a number of important consequences. First of all, it can be used to guarantee the existence of sum-product fixed points. Observe that the cost function in the Bethe variational problem (11.56) is continuous and bounded above, and the constraint set LOCAL(G) is non-empty and compact; therefore, at least some (possibly local) maximum is attained. Moreover, since the constraints are linear, there will always be a set of Lagrange multipliers associated with any local maximum [Bertsekas (1995)]. For any optimum in the relative interior of LOCAL(G), these Lagrange multipliers can be used to construct a fixed point of the sum-product updates.

For graphs with cycles, this Lagrangian formulation provides no guarantees on the convergence of the sum-product updates; indeed, whether or not the algorithm converges depends both on the potential strengths and the topology of the graph. Several researchers [Yuille (2002); Welling and Teh (2001); Heskes et al. (2003)] have proposed alternatives to sum-product that are guaranteed to converge, albeit at the price of increased computational cost. It should also be noted that with the exception of trees and other special cases [Pakzad and Anantharam (2002); McEliece and Yildirim (2002)], the BVP is usually a non-convex problem, in that H_{Bethe} fails to be concave. As a consequence, there may be multiple local optima to the BVP, and there are no guarantees that sum-product (or other iterative algorithms) will find a global optimum.

As illustrated in Figure 11.10(c), the constraint set LOCAL(G) of the Bethe variational problem is a strict outer bound on the marginal polytope MARG(G). Since the exact marginals of $p(\mathbf{x};\theta)$ must always lie in the marginal polytope, a natural question is whether solutions to the Bethe variational problem ever fall into the region LOCAL(G)\ MARG(G). There turns out be a straightforward answer to this question, stemming from an alternative reparameterization-based characterization of sum-product fixed points [Wainwright et al. (2003b)]. One consequence of this characterization is that for any vector τ of pseudomarginals in the interior of LOCAL(G), it is possible to specify a distribution for which τ is a sum-product fixed point. As a particular example, it is possible to construct a

3. Some care is required in dealing with the boundary conditions $\tau_s(x_s) \geq 0$ and $\tau_{st}(x_s, x_t) \geq 0$; see Yedidia et al. for further discussion.

distribution $p(\mathbf{x};\theta)$ such that the pseudomarginal τ discussed in Example 11.14 is a fixed point of the sum-product updates.

11.7.5 Extensions based on clustering and hypertrees

From our development in the previous section, it is clear that there are two *distinct* components to the Bethe variational principle: (a) the entropy approximation H_{Bethe}, and (b) the approximation LOCAL(G) to the set of realizable marginal parameters. In principle, the BVP could be strengthened by improving either one, or both, of these components. One natural generalization of the BVP, first proposed by Yedidia et al. [2002] and further explored by various researchers [Heskes et al. (2003); McEliece and Yildirim (2002); Minka (2001)], is based on working with clusters of variables. The approximations in the Bethe approach are based on trees, which are special cases of junction trees based on cliques of size two. A natural strategy, then, is to strengthen the approximations by exploiting more complex junction trees, also known as hypertrees. Our description of this procedure is very brief, but further details can be found in various sources [Yedidia et al. (2002); Wainwright and Jordan (2003a)].

Recall that the essential ingredients in Bethe variational principle are local (pseudo)marginal distributions on nodes and edges (i.e., pairs of nodes). These distributions, subject to edge-wise marginalization conditions, are used to specify the Bethe entropy approximation. One way to improve the Bethe approach, which is based on pairs of nodes, is to build entropy approximations and impose marginalization constraints on *larger* clusters of nodes. To illustrate, suppose that the original graph is simply the 3×3 grid shown in Figure 11.11(a). A particular

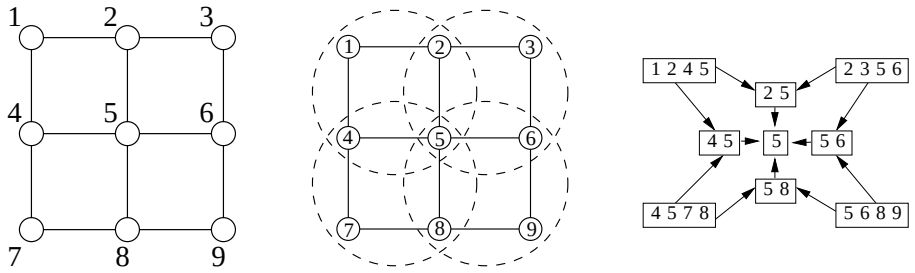


Figure 11.11 (a) Ordinary 3×3 grid. (b) Clustering of the vertices into groups of 4, known as Kikuchi 4-plaque clustering. (c) Poset diagram of the clusters as well as their overlaps. Pseudomarginals on these subsets must satisfy certain local consistency conditions, and are used to define a higher-order entropy approximation.

grouping of the nodes, which is known as Kikuchi 4-plaque clustering in statistical physics [Yedidia et al. (2002)], is illustrated in panel (b). This operation creates

four new "supernodes" or clusters, each consisting of four nodes from the original graph. These clusters, as well as their overlaps—which turn out to be critical to track for certain technical reasons [Yedidia et al. (2002)])—are illustrated in panel (c).

Given a clustering of this type, we now consider a set of marginal distributions τ_h, where h ranges over the clusters. As with the singleton τ_s and pairwise τ_{st} that define the Bethe approximation, we require that these higher-order cluster marginals are suitably normalized (i.e.,$\sum_{x'_h} \tau_h(x'_h) = 1$), and are consistent with one another whenever they overlap. More precisely, for any pair $g \subseteq h$, the following *marginalization* condition $\sum_{\{x'_h \mid x'_g = x_g\}} \tau_h(x'_h) = \tau_g(x_g)$ must hold. Imposing these normalization and marginalization conditions leads to a higher-order analog of the constraint LOCAL(G) previously defined in equation (11.54).

In analogy to the Bethe entropy approximation, we can also consider a hypertree-based approximation to the entropy. There are certain technical aspects to specifying such entropy approximations, in that it turns out to be critical to ensure that the local entropies are weighted with certain "over-counting" numbers [Yedidia et al. (2002); Wainwright and Jordan (2003a)]. Without going into these details here, the outcome is another relaxed variational principle, which can be understood as a higher-level analog of the Bethe variational principle.

11.8 From the exact principle to new approximations

The preceding sections have illustrated how a variety of known methods—both exact and approximate—can be understood in an unified manner on the basis of the general variational principle (11.37). In this final section, we turn to a brief discussion of several new approximate methods that also emerge from this same variational principle. Given space constraints, our discussion in this chapter is necessarily brief, but we refer to reader to the papers [Wainwright and Jordan (2003a,b); Wainwright et al. (2002, 2003a)] for further details.

11.8.1 Exploiting semidefinite constraints for approximate inference

As discussed in Section 11.6, one key component in any relaxation of the exact variational principle is an approximation of the set $\mathcal{M}$ of realizable mean parameters. Recall that for graphical models that involve discrete random variables, we refer to this set as a *marginal polytope*. Since any polytope is specified by a finite collection of halfspace constraints (see Figure 11.6), one very natural way in which to generate an outer approximation is by including only a *subset* of these halfspace constraints. Indeed, as we have seen in Section 11.7, it is precisely this route that the Bethe approximation and its clustering-based extensions follow.

However, such *polyhedral relaxations* are not the only way in which to generate outer approximations to marginal polytopes. Recognizing that elements of the marginal polytope are essentially *moments* leads very naturally to the idea of a

semidefinite relaxation. Indeed, the use of semidefinite constraints for characterizing moments has a very rich history, both with classical work [Karlin and Studden (1966)] on scalar random variables, and more recent work [Lasserre (2001); Parrilo (2003)] on the multivariate case.

11.8.1.1 Semidefinite outer bounds on marginal polytopes

We use the case of a multinomial MRF defined by a graph $G = (V, E)$, as discussed in Example 11.4, in order to illustrate the use of semidefinite constraints. Although the basic idea is quite generally applicable [Wainwright and Jordan (2003a)]), herein we restrict ourselves to binary variables (i.e., $\mathcal{X}_s = \{0, 1\}$) so as to simplify the exposition. Recall that the sufficient statistics in a binary MRF take the form of certain indicator functions, as defined in equation (11.20). In fact, this representation is overcomplete (in that there are linear dependencies among the indicator functions); in the binary case, it suffices to consider only the sufficient statistics $x_s = \mathbb{I}_1(x_s)$ and $x_s x_t = \mathbb{I}_{11}(x_s, x_t)$. Our goal, then, is to characterize the set of all first- and second-order moments, defined by $\mu_s = \mathbb{E}[x_s]$ and $\mu_{st} = \mathbb{E}[x_s x_t]$ respectively, that arise from taking expectations with a distribution with its support restricted to $\{0, 1\}^n$. Rather than focusing on just the pairs μ_{st} for edges $(s, t) \in E$, it is convenient to consider the full collection of pairwise moments $\{\mu_{st} \,|\, s, t \in V\}$.

Suppose that we are given a vector $\mu \in \mathbb{R}^d$ (where $d = n + \binom{n}{2}$), and wish to assess whether or not it is a globally realizable moment vector (i.e., whether there exists some distribution $p(\mathbf{x})$ such that $\mu_s = \sum_{\mathbf{x}} p(\mathbf{x})\, x_s$ and $\mu_{st} = \sum_{\mathbf{x}} p(\mathbf{x})\, x_s x_t$). In order to derive a *necessary* condition, we suppose that such a distribution p exists, and then consider the following $(n+1) \times (n+1)$ moment matrix:

$$\mathbb{E}_p\left\{ \begin{bmatrix} 1 \\ \mathbf{x} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{x} \end{bmatrix} \right\} = \begin{bmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_{n-1} & \mu_n \\ \mu_1 & \mu_1 & \mu_{12} & \cdots & \cdots & \mu_{1n} \\ \mu_2 & \mu_{21} & \mu_2 & \cdots & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n-1} & \vdots & \vdots & \vdots & \vdots & \mu_{n,(n-1)} \\ \mu_n & \mu_{n1} & \mu_{n2} & \cdots & \mu_{(n-1),n} & \mu_n \end{bmatrix}, \quad (11.59)$$

which we denote by $M_1[\mu]$. Note that in calculating the form of this moment matrix, we have made use of the relation $\mu_s = \mu_{ss}$, which holds because $x_s = x_s^2$ for any binary-valued quantity.

We now observe that any such moment matrix is necessarily positive semidefinite, which we denote by $M_1[\mu] \succeq 0$. (This positive semidefiniteness can be verified as follows: letting $\mathbf{y} := (1,\ \mathbf{x})$, then for any vector $a \in \mathbb{R}^{n+1}$, we have $a^T M_1[\mu] a = a^T \mathbb{E}[\mathbf{y}\mathbf{y}^T] a = \mathbb{E}[\|a^T \mathbf{y}\|^2]$, which is certainly non-negative). Therefore, we conclude that the semidefinite constraint set $\text{SDEF}_1 := \{\mu \in \mathbb{R}^d \,|\, M_1[\mu] \succeq 0\}$ is an outer bound on the exact marginal polytope.

Example 11.15
To illustrate the use of the outer bound SDEF_1, recall the pseudomarginal vector τ that we constructed in Example 11.14 for the single cycle on three nodes. In terms of our reduced representation (involving only expectations of the singletons x_s and pairwise functions $x_s x_t$), this pseudomarginal can be written as follows:

$$\tau_s \;=\; 0.5 \;\; \text{for } s = 1, 2, 3, \qquad \tau_{12} \;=\; \tau_{23} \;=\; 0.4, \qquad \tau_{13} \;=\; 0.1.$$

Suppose that we now construct the matrix M_1 for this trial set of mean parameters; it takes the following form:

$$M_1[\tau] \;=\; \begin{bmatrix} 1 & 0.5 & 0.5 & 0.5 \\ 0.5 & 0.5 & 0.4 & 0.1 \\ 0.5 & 0.4 & 0.5 & 0.4 \\ 0.5 & 0.1 & 0.4 & 0.5 \end{bmatrix}.$$

A simple calculation shows that it is not positive definite, so that $\tau \notin \text{SDEF}_1$. Since SDEF_1 is an outer bound on the marginal polytope, this reasoning shows—in a very quick and direct manner— that τ is not a globally valid moment vector.

In fact, the semidefinite constraint set SDEF_1 can be viewed as the first in a sequence of progressively tighter relaxations on the marginal polytope.

11.8.1.2 Log-determinant relaxation

We now show how to use such semidefinite constraints in approximate inference. Our approach is based on combining the first-order semidefinite outer bound SDEF_1 with Gaussian-based entropy approximation. The end result is a log-determinant problem that represents another relaxation of the exact variational principle [Wainwright and Jordan (2003b)]. In contrast to the Bethe/Kikuchi approaches, this relaxation is convex (and hence has a unique optimum), and moreover provides an upper bound on the cumulant generating function.

Our starting point is the familiar interpretation of the Gaussian as the maximum entropy distribution subject to covariance constraints [Cover and Thomas (1991)]. In particular, given a continuous random vector $\widetilde{\mathbf{x}}$, its differential entropy $h(\widetilde{\mathbf{x}})$ is always upper bounded by the entropy of a Gaussian with matched covariance, or in analytical terms

$$h(\widetilde{\mathbf{x}}) \;\leq\; \frac{1}{2} \log \det \text{cov}(\widetilde{\mathbf{x}}) + \frac{n}{2} \log(2\pi e), \tag{11.60}$$

where $\text{cov}(\widetilde{\mathbf{x}})$ is the covariance matrix of $\widetilde{\mathbf{x}}$. The upper bound (11.60) is not directly applicable to a random vector taking values in a discrete space (since differential entropy in this case diverges to minus infinity). However, a straightforward discretization argument shows that for any discrete random vector $\mathbf{x} \in \{0, 1\}^n$, its (ordinary) discrete entropy can be upper bounded in terms of the matrix $M_1[\mu]$ of

mean parameters as

$$H(\mathbf{x}) = -A^*(\mu) \;\leq\; \frac{1}{2}\log\det\Big\{M_1[\mu] + \frac{1}{12}\,\mathrm{blkdiag}[0, I_n]\Big\} + \frac{n}{2}\log(2\pi e). \quad (11.61)$$

where $\mathrm{blkdiag}[0, I_n]$ is a $(n+1)\times(n+1)$ block-diagonal matrix with a 1×1 zero block, and an $n \times n$ identity block.

Finally, putting all the pieces together leads to the following result [Wainwright and Jordan (2003b)]: the cumulant generating function $A(\theta)$ is upper bounded by the solution of the following *log-determinant optimization problem*:

$$A(\theta) \;\leq\; \max_{\tau\in \mathrm{SDEF}_1}\left\{\langle\theta,\, \mu\rangle + \frac{1}{2}\log\det\big[M_1(\tau) + \frac{1}{12}\,\mathrm{blkdiag}[0, I_n]\big]\right\} + \frac{n}{2}\log(2\pi e). \quad (11.62)$$

Note that the constraint $\tau \in \mathrm{SDEF}_1$ ensures that $M_1(\tau) \succeq 0$, and hence a fortiori that $M_1(\tau) + \frac{1}{12}\,\mathrm{blkdiag}[0, I_n]$ is positive definite. Moreover, an important fact is that the optimization problem in equation (11.62) is a determinant maximization problem, for which efficient interior point methods have been developed [Vandenberghe et al. (1998)].

Just as the Bethe variational principle (11.56) is a tree-based approximation, the log-determinant relaxation (11.62) is a Gaussian-based approximation. In particular, it is worthwhile comparing the structure of the log-determinant relaxation (11.62) to the exact variational principle for a multivariate Gaussian, as described in Section 11.5.1. In contrast to the Bethe variational principle, in which all of the constraints defining the relaxation are local, this new principle (11.62) imposes some quite *global* constraints on the mean parameters. Empirically, these global constraints are important for strongly coupled problems, in which the performance log-determinant relaxation appears is much more robust then the sum-product algorithm [Wainwright and Jordan (2003b)]. In summary, starting from the exact variational principle (11.37), we have derived a new relaxation, whose properties are rather different than the Bethe and Kikuchi variational principles.

11.8.2 Relaxations for computing modes

Recall from our introductory comments in Section 11.2.2 that, in addition to the problem of computing expectations and likelihoods, it is also frequently of interest to compute the mode of a distribution. This section is devoted to a brief discussion of mode computation, and more concretely how the exact variational principle (11.37), as well as relaxations thereof, again turns out to play an important role.

11.8.2.1 Zero-temperature limits

In order to understand the role of the exact variational principle (11.37) in computing modes, consider a multinomial MRF of the form $p(\mathbf{x};\theta)$, as discussed in Example 11.4. Of interest to us is the 1-parameter family of distributions $\{p(\mathbf{x};\beta\theta)\,|\,\beta > 0\}$, where β is the real number to be varied. At one extreme, if

$\beta = 0$, then there is no coupling, and the distribution is simply uniform over all possible configurations. The other extreme, as $\beta \to +\infty$, is more interesting; in this limit, the distribution concentrates all of its mass on the configuration (or subset of configurations) that are modes of the distribution. Taking this limit $\beta \to +\infty$ is known as "zero-temperature" limit, since the parameter β is typically viewed as inverse temperature in statistical physics. This argument suggests that there should be a link between computing modes and the limiting behavior of the marginalization problem as $\beta \to +\infty$.

In order to develop this idea a bit more formally, we begin by observing that the exact variational principle (11.37) holds for the distribution $p(\mathbf{x}; \beta\theta)$ for any value of $\beta \geq 0$. It can be shown [Wainwright and Jordan (2003a)] that if we actually take a suitably scaled limit of this exact variational principle as $\beta \to +\infty$, then we recover the following variational principle for computing modes:

$$\max_{\mathbf{x} \in \mathcal{X}^n} \langle \theta, \boldsymbol{\phi}(\mathbf{x}) \rangle \;=\; \max_{\mu \in \mathrm{MARG}(G)} \langle \theta, \mu \rangle \tag{11.63}$$

Since the log probability $\log p(\mathbf{x}; \theta)$ is equal to $\langle \theta, \boldsymbol{\phi}(\mathbf{x}) \rangle$ (up to an additive constant), the left-hand side is simply the problem of computing the mode of the distribution $p(\mathbf{x}; \theta)$. On the right-hand side, we simply have a linear program, since the constraint set $\mathrm{MARG}(G)$ is a polytope, and the cost function $\langle \theta, \mu \rangle$ is linear in μ (with θ fixed). This equivalence means that, at least in principle, we can compute a mode of the distribution by solving a linear program (LP) over the marginal polytope. The geometric interpretation is also clear: as illustrated in Figure 11.6, vertices of the marginal polytope are in one-to-one correspondence with configurations $\mathbf{x}$. Since any LP achieves its optimum at a vertex [Bertsimas and Tsitsikilis (1997)], solving the LP is equivalent to finding the mode.

11.8.2.2 Linear programming and tree-reweighted max-product

Of course, the LP-based reformulation (11.63) is not practically useful for precisely the same reasons as before—it is extremely challenging to characterize the marginal polytope $\mathrm{MARG}(G)$ for a general graph. Many computationally intractable optimization problems (e.g., MAX-CUT) can be reformulated as an LP over the marginal polytope, as in equation (11.63), which underscores the inherent complexity of characterizing marginal polytopes. Nonetheless, this variational formulation motivates the idea of forming relaxations using outer bounds on the marginal polytope. For various classes of problems in combinatorial optimization, both linear programming and semidefinite relaxations of this flavor have been studied extensively.

Here we briefly describe an LP relaxation that is very natural given our development of the Bethe variational principle in Section 11.7. In particular, we consider using the local constraint set $\mathrm{LOCAL}(G)$, as defined in equation (11.54), as an outer bound of the marginal polytope $\mathrm{MARG}(G)$. Doing so leads to the following

LP relaxation for the problem of computing the mode of a multinomial MRF:

$$\max_{\mathbf{x}\in\mathcal{X}^n} \langle \theta, \boldsymbol{\phi}(\mathbf{x}) \rangle \;=\; \max_{\mu\in \mathrm{MARG}(G)} \langle \theta, \mu \rangle \;\leq\; \max_{\tau\in \mathrm{LOCAL}(G)} \langle \theta, \mu \rangle. \tag{11.64}$$

Since the relaxed constraint set LOCAL(G)—like the original set MARG(G)—is a polytope, the relaxation on the right-hand side of equation (11.64) is a linear program. Consequently, the optimum of the relaxed problem must be attained at a vertex (possibly more than one) of the polytope LOCAL(G).

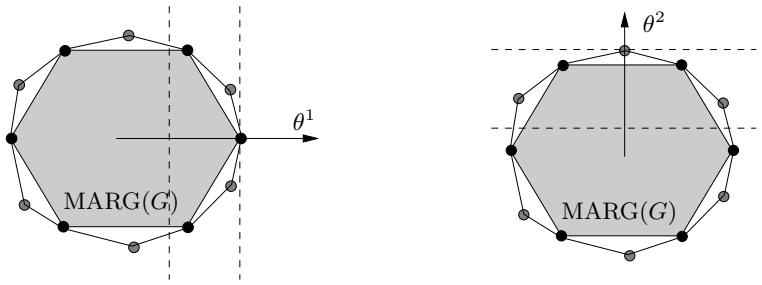


Figure 11.12 The constraint set LOCAL(G) is an outer bound on the exact marginal polytope. Its vertex set includes all the vertices of MARG(G), which are in one-to-one correspondence with optimal solutions of the integer program. It also includes additional fractional vertices, which are *not* vertices of MARG(G).

We say that a vertex of LOCAL(G) is *integral* if all of its components are zero or one, and *fractional* otherwise. The distinction between fractional and integral vertices is crucial, because it determines whether or not the LP relaxation (11.64) specified by LOCAL(G) is tight. In particular, there are only two possible outcomes to solving the relaxation:

(a) the optimum is attained at a vertex of MARG(G), in which case the upper bound in equation (11.64) is tight, and a mode can be obtained.

(b) the optimum is attained only at one or more fractional vertices of LOCAL(G), which lie strictly outside MARG(G). In this case, the upper bound of equation (11.64) is loose, and the relaxation does not output the optimal configuration.

Figure 11.12 illustrates both of these possibilities. The vector θ^1 corresponds to case (a), in which the optimum is attained at a vertex of MARG(G). The vector θ^2 represents a less fortunate setting, in which the optimum is attained only at a fractional vertex of LOCAL(G). In simple cases, one can explicitly demonstrate a fractional vertex of the polytope LOCAL(G).

Given the link between the sum-product algorithm and the Bethe variational principle, it would be natural to conjecture that the max-product algorithm can be derived as an algorithm for solving the LP relaxation (11.64). For trees (in which case the LP (11.64) is exact), this conjecture is true: more precisely, it can

be shown [Wainwright et al. (2003a)] that the max-product algorithm (or the Viterbi algorithm) is an iterative method for solving the dual of the LP (11.64). However, this statement is false for graphs with cycles, since it is straightforward to construct problems (on graphs with cycles) for which the max-product algorithm will output a non-optimal configuration. Consequently, the max-product algorithm does not specify solutions to the dual problem, since any LP relaxation will either output a configuration with a guarantee of correctness, or a fractional vertex. However, Wainwright et al. [2003a] derive a tree-reweighted analog of the max-product algorithm, which does have provable connections to dual optimal solutions of the tree-based relaxation (11.64).

11.9 Conclusion

A fundamental problem that arises in applications of graphical models—whether in signal processing, machine learning, bioinformatics, communication theory, or other fields—is that of computing likelihoods, marginal probabilities, and other expectations. We have presented a variational characterization of the problem of computing likelihoods and expectations in general exponential-family graphical models. Our characterization focuses attention both on the constraint set and the objective function. In particular, for exponential-family graphical models, the constraint set $\mathcal{M}$ is a convex subset in a finite-dimensional space, consisting of all realizable mean parameters. The objective function is the sum of a linear function and an entropy function. The latter is a concave function, and thus the overall problem—that of maximizing the objective function over $\mathcal{M}$—is a convex problem. In this chapter, we discussed how the junction tree algorithm and other exact inference algorithms can be understood as particular methods for solving this convex optimization problem. In addition, we showed that a variety of approximate inference algorithms—including loopy belief propagation, general cluster variational methods and mean-field methods—can be understood as methods for solving particular relaxations of the general variational principle. More concretely, we saw that belief propagation involves an outer approximation of $\mathcal{M}$ whereas mean field methods involve an inner approximation of $\mathcal{M}$. In addition, this variational principle suggests a number of new inference algorithms, as we briefly discussed.

It is worth noting certain limitations inherent to the variational framework as presented in this chapter. In particular, we have not discussed curved exponential families, but instead limited our treatment to regular families. Curved exponential families are useful in the context of directed graphical models, and further research is required to develop a general variational treatment of such models. Similarly, we have dealt exclusively with exponential family models, and not treated nonparametric models. One approach to exploiting variational ideas for nonparametric models is through exponential family approximations of nonparametric distributions; for example, Blei and Jordan [2004] have presented inference methods for Dirichlet process mixtures that are based on the variational framework presented here.

References

S.M. Aji and R.J. McEliece. The generalized distributive law. *IEEE Trans. Info. Theory*, 46:325–343, March 2000.

R. J. Baxter, editor. *Exactly Solved Models in Statistical Mechanics*. Academic Press, New York, 1982.

D.P. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

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

J. Besag and P. J. Green. Spatial statistics and Bayesian computation. *J. R. Stat. Soc. B*, 55(1):25–37, 1993.

D. M. Blei and M. I. Jordan. Variational methods for the Dirichlet process. In *International Conference on Machine Learning*, New York, NY, 2004. ACM Press.

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

D. Chandler. *Introduction to modern statistical mechanics*. Oxford University Press, Oxford, 1987.

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

M.S. Crouse, R.D. Nowak, and R.G. Baraniuk. Wavelet-based statistical signal processing using hidden Markov models. *IEEE Trans. on Signal Processing*, 46: 886–902, April 1998.

I. Csiszár and G. Tusnády. Information geometry and alternating minimization procedures. In E. J. Dudewisc et al., editor, *Recent results in estimation theory and related topics*. Unknown, 1984.

A. P. Dawid. Applications of a general propagation algorithm for probabilistic expert systems. *Statistics and Computing*, 2:25–36, 1992.

J.W. Demmel. *Applied numerical linear algebra*. SIAM, Philadelphia, 1997.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm (with discussion). *J. Royal Stat. Soc. B*, 39:1–38, 1977.

R. Durbin, S. Eddy, A. Krogh, and G. Mitchison, editors. *Biological Sequence*

Analysis. Cambridge University Press, Cambridge, 1998.

B. Efron. The geometry of exponential families. *Annals of Statistics*, 6:362–376, 1978.

J. Felsenstein. Evolutionary trees from DNA sequences: a maximum likelihood approach. *Journal of Molecular Evolution*, 17:368–376, 1981.

G. D. Forney, Jr. The Viterbi algorithm. *Proc. IEEE*, 61:268–277, March 1973.

W. T. Freeman, E. C. Pasztor, and O. T. Carmichael. Learning low-level vision. *Intl. J. Computer Vision*, 40(1):25–47, 2000.

B. Frey, R. Koetter, and N. Petrovic. Very loopy belief propagation for unwrapping phase images. In *NIPS 14*. MIT Press, 2001.

S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Pat. Anal. Mach. Intell.*, 6:721–741, 1984.

Z. Ghahramani and M. Jordan. Factorial hidden Markov models. *Machine Learning*, 29:245–273, 1997.

W. Gilks, S. Richardson, and D. Spiegelhalter, editors. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, New York, NY, 1996.

T. Heskes, K. Albers, and B. Kappen. Approximate inference and constrained optimization. In *Uncertainty in Artificial Intelligence*, volume 13, page to appear, 2003.

J. Hiriart-Urruty and C. Lemaréchal. *Convex analysis and minimization algorithms*, volume 1. Springer-Verlag, New York, 1993.

R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1985.

T. S. Jaakkola. Tutorial on variational approximation methods. In M. Opper and D. Saad, editors, *Advanced mean field methods: Theory and practice*, pages 129–160. MIT Press, 2001.

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.

T. Kailath, A. H. Sayed, and B. Hassibi. *Linear Estimation*. Prentice Hall, New Jersey, 2000.

H. Kappen and P. Rodriguez. Efficient learning in Boltzmann machines using linear response theory. *Neural Computation*, 10:1137–1156, 1998.

S. Karlin and W. Studden. *Tchebycheff systems, with applications in analysis and statistics*. Interscience Publishers, New York, NY, 1966.

F. Kschischang and B. Frey. Iterative decoding of compound codes by probability propagation in graphical models. *IEEE Sel. Areas Comm.*, 16(2):219–230, February 1998.

F.R. Kschischang, B.J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. Info. Theory*, 47:498–519, February 2001.

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

S. L. Lauritzen. *Graphical Models*. Oxford University Press, Oxford, 1996.

S. L. Lauritzen and D. J. Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems (with discussion). *Journal of the Royal Statistical Society B*, 50:155–224, January 1988.

M.A.R. Leisink and H.J. Kappen. A tighter bound for graphical models. In *NIPS 13*, pages 266–272. MIT Press, 2001.

H. A. Loeliger. An introduction to factor graphs. *IEEE Signal Processing Magazine*, 21:28–41, 2004.

M. Luettgen, W. Karl, and A. Willsky. Efficient multiscale regularization with application to optical flow. *IEEE Trans. Image Processing*, 3(1):41–64, Jan. 1994.

R. J. McEliece and M. Yildirim. Belief propagation on partially ordered sets. In D. Gilliam and J. Rosenthal, editors, *Mathematical Theory of Systems and Networks*. Institute for Mathematics and its Applications, 2002.

R.J. McEliece, D.J.C. McKay, and J.F. Cheng. Turbo decoding as an instance of Pearl's belief propagation algorithm. *IEEE Jour. Sel. Comm.*, 16(2):140–152, February 1998.

T. P. Minka. *A family of algorithms for approximate Bayesian inference*. PhD thesis, MIT, January 2001.

R. Neal and G. E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*. MIT Press, Cambridge, MA, 1999.

M. Opper and D. Saad. Adaptive TAP equations. In M. Opper and D. Saad, editors, *Advanced mean field methods: Theory and practice*, pages 85–98. MIT Press, 2001.

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

P. Parrilo. Semidefinite programming relaxations for semialgebraic problems. *Mathematical Programming, Ser. B*, 96:293–320, 2003.

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

T. Plefka. Convergence condition of the TAP equation for the infinite-ranged Ising model. *Journal of Physics A*, 15(6):1971–1978, 1982.

L. R. Rabiner and B. H. Juang. *Fundamentals of speech recognition*. Prentice Hall, Englewood Cliffs, N.J., 1993.

T. Richardson and R. Urbanke. The capacity of low-density parity check codes under message-passing decoding. *IEEE Trans. Info. Theory*, 47:599–618, February 2001.

G. Rockafellar. *Convex Analysis*. Princeton University Press, Princeton, 1970.

L. K. Saul and M.I. Jordan. Exploiting tractable substructures in intractable networks. In *NIPS 8*, pages 486–492. MIT Press, 1996.

R. Szeliski. Bayesian modeling of uncertainty in low-level vision. *International Journal of Computer Vision*, 5(3):271–301, 1990.

D. M. Titterington, A.F.M. Smith, and U.E. Makov, editors. *Statistical analysis of finite mixture distributions*. Wiley, New York, 1986.

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.

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. MAP estimation via agreement on (hyper)trees: Message-passing and linear programming approaches. In *Proc. Allerton Conference on Communication, Control and Computing*, October 2002.

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 2003a.

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, 2003b.

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree consistency and bounds on the max-product algorithm and its generalizations. *Statistics and Computing*, 14:143–166, April 2004.

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

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 2003b.

Y. Weiss and W. T. Freeman. Correctness of belief propagation in Gaussian graphical models of arbitrary topology. In *NIPS 12*, pages 673–679. MIT Press, 2000.

M. Welling and Y. Teh. Belief optimization: A stable alternative to loopy belief propagation. In *Uncertainty in Artificial Intelligence*, July 2001.

W. Wiegerinck. Variational approximations between mean field theory and the junction tree algorithm. In *UAI 2000*, San Francisco, CA, 2000. Morgan Kaufmann Publishers.

A. S. Willsky. Multiresolution Markov models for signal and image processing. *Proceedings of the IEEE*, 90(8):1396–1458, 2002.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Generalized belief propagation. In

NIPS 13, pages 689–695. MIT Press, 2001.

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.

A. Yuille. CCCP algorithms to minimize the Bethe and Kikuchi free energies: Convergent alternatives to belief propagation. *Neural Computation*, 14:1691–1722, 2002.

J. Zhang. The application of the Gibbs-Bogoliubov-Feynman inequality in mean-field calculations for Markov random-fields. *IEEE Trans. on Image Processing*, 5(7):1208–1214, July 1996.