Log-Determinant Relaxation for Approximate Inference in Discrete Markov Random Fields

Martin J. Wainwright and Michael I. Jordan, *Fellow, IEEE*

***Abstract*—Graphical models are well suited to capture the complex and non-Gaussian statistical dependencies that arise in many real-world signals. A fundamental problem common to any signal processing application of a graphical model is that of computing approximate marginal probabilities over subsets of nodes. This paper proposes a novel method, applicable to discrete-valued Markov random fields (MRFs) on arbitrary graphs, for approximately solving this *marginalization problem*. The foundation of our method is a reformulation of the marginalization problem as the solution of a low-dimensional convex optimization problem over the marginal polytope. Exactly solving this problem for general graphs is intractable; for binary Markov random fields, we describe how to relax it by using a Gaussian bound on the discrete entropy and a semidefinite outer bound on the marginal polytope. This combination leads to a log-determinant maximization problem that can be solved efficiently by interior point methods, thereby providing approximations to the exact marginals. We show how a slightly weakened log-determinant relaxation can be solved even more efficiently by a dual reformulation. When applied to denoising problems in a coupled mixture-of-Gaussian model defined on a binary MRF with cycles, we find that the performance of this log-determinant relaxation is comparable or superior to the widely used sum-product algorithm over a range of experimental conditions.**

***Index Terms*—Belief propagation, denoising, Gaussian mixture, Markov random field, sum-product algorithm.**

I. Introduction

MANY classes of real-world signals, including speech [1], financial time series [2], and natural images [3], [4], exhibit complex and non-Gaussian statistical dependencies. In such settings, it is well known that classical approaches to denoising and detection, many of which are based on assumptions of independence or joint Gaussianity, may lead to markedly suboptimal performance. It is therefore of considerable interest to develop and explore signal processing methods that are capable of modeling and exploiting a broader class of dependency structures. One such class of methods is provided by *graphical models,* a formalism in which random variables are associated with the nodes of a graph and the edges of the graph represent statistical dependencies among these variables. Graphical models have proven useful in a wide range of signal processing problems; we refer the reader to the survey papers [5] and [6] for an overview.

In the graphical models most commonly encountered in signal processing applications, the underlying graph is a chain or a tree. For such cycle-free graphs, efficient recursive algorithms [5]–[7] are available for calculating various statistical quantities of interest (e.g., likelihoods and other marginal probabilities). The elegance and familiarity of these recursive algorithms should not, however, obscure the fact that chains and trees capture rather limited forms of statistical dependency, and there are numerous applications—among them image denoising [4] and sensor fusion [8]—that would be better served by the richer class of graphical models in which cycles are allowed in the underlying graph. Accordingly, the algorithmic treatment of such graphical models with cycles is the focus of this paper.

As a specific example of these issues, and as motivation for the experimental results that we present later in the paper, let us consider a graphical modeling approach to statistical signal processing in the wavelet domain. Crouse *et al.* [9] presented a statistical model for wavelets in which each wavelet coefficient is modeled as a finite mixture of Gaussians, typically with two mixture components indexed by a binary $\{0,1\}$-valued random variable. The wavelet coefficients are coupled together by introducing statistical dependencies among the binary variables underlying each local Gaussian mixture model; this setup is sufficiently flexible to capture the non-Gaussian dependencies present in signal classes such as natural images (e.g., [3] and [4]). Working within the graphical modeling framework, Crouse *et al.* investigated two types of dependency among the binary mixture component labels. The first class of model involves linking variables across space, separately for each scale [see Fig. 1(a)]. The second class of model involves linking components across scale according to a tree structure [see Fig. 1(b)]. The latter model is known as a *hidden Markov tree* ("hidden" because the states of the mixture component labels is not observed in the data), whereas each of the chains composing the former model is known as a *hidden Markov model*—perhaps better referred to as a *hidden Markov chain*.

The advantage of hidden Markov chains and hidden Markov trees is that they permit the use of fast recursive algorithms for computing marginal probabilities. The classical algorithm for chains is known as the "forward-backward" algorithm [10]. Crouse *et al.* [9] presented an analog of this algorithm for trees (see also [11] and [12]). The general algorithm for computing marginal probabilities on cycle-free graphs is known as the *sum-product algorithm* [13], also referred to as *belief propagation* [7].

Manuscript received April 19, 2005; revised June 28, 2005. This work was supported by the National Science Foundation under Grant IIS-9988642, ARO MURI DAA19-02-1-0383, and by a Grant from Intel Corporation. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Paul D. Fiore.

M. J. Wainwright is with the Department of Statistics and the Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720 USA (e-mail: wainwrig@eecs.berkeley.edu).

M. I. Jordan is with the Department of Electrical Engineering and Computer Science and the Department of Statistics, University of California, Berkeley, CA 94720 USA (e-mail: jordan@cs.berkeley.edu).

Digital Object Identifier 10.1109/TSP.2006.874409

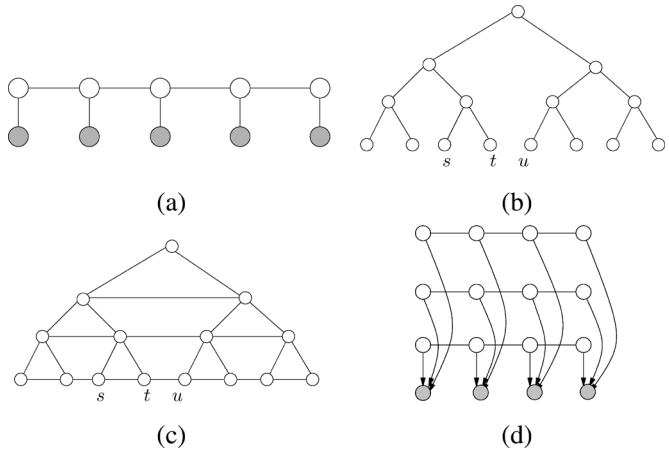

Fig. 1. Different types of graphical Markov models. (a) A chain-structured hidden Markov model: shaded nodes represent noisy observations of the hidden states represented by unshaded nodes. (b) A tree-structured model (showing only the hidden states for simplicity). (c) A marriage of the chain and tree models leads to a graphical model with cycles. (d) A factorial hidden Markov model, in which several Markov chains are coupled together; the arrows from the hidden state variables to the shaded node represent a conditional distribution that couples the Markov chains.

While hidden Markov chains and hidden Markov trees are associated with fast algorithms, they are not necessarily faithful models of the statistical dependencies among wavelet coefficients. In particular, tree-structured models are well known to introduce artifacts in the spatial domain [5], [14]. Consider for example the variables s, t, and u in Fig. 1(b). While the spatial separation of s and t is the same as the spatial separation between t and u, the vertices s and t are nearest neighbors in the tree whereas t and u are separated by a large distance in the tree. Although this type of boundary artifact is not present in the chain-structured model in Fig. 1(a), a chain-structured model fails to capture dependencies across scale. More desirable is the hybrid model shown in Fig. 1(c), in which both vertical and horizontal edges are present. This graphical model has cycles, however, and the computation of marginal probabilities no longer reduces to a straightforward recursion on a tree.

The need to couple multiple Markov chains arises not only in wavelet modeling [8], [15]. For example, many sensor fusion problems involve statistical dependencies among multiple time series, and one approach to treat such problems is to make use of a coupled set of hidden Markov chains, one for each time series. For example, Reyes *et al.* [8] described a model for separation of multiple speakers in which a hidden Markov chain is used for each speaker and the observable spectra are a function of the states of all of the chains [see Fig. 1(d)]. This graphical model has cycles and exact marginalization is feasible only for small numbers of chains.

Although introducing cycles into a graphical model leads to a more expressive class of probability distributions, it also raises the fundamental computational challenge of computing marginal probabilities in the presence of cycles. In principle, any Markov model on a graph with cycles can be converted, by a procedure of clustering nodes and augmenting the associated states, into the so-called *junction tree* form [16], to which exact recursive algorithms akin to the sum-product algorithm can be applied. However, the overall algorithm of the junction tree approach has computational cost exponential in the size of the augmented state space, a quantity which is unacceptably large in many applications. Thus, a key problem to be addressed—if graphical models with cycles are to be applied fruitfully to signal processing problems—is the development of efficient methods for computing approximate marginal distributions.

The sum-product algorithm involves a simple message-passing protocol in which each node in the graph computes outgoing messages based on transformations (sums of products) of the messages arriving from its neighbors. When there are cycles in the graph, the algorithm can still be implemented, but it is no longer guaranteed to converge, and the answers obtained (assuming convergence occurs) must be viewed as approximations to the underlying marginal probabilities. Despite these serious problems, this "loopy" form of the sum-product algorithm is widely used in practice and indeed it is the state-of-the-art approach to various signal processing problems involving graphical models with cycles [6], [13], [17]. Interestingly, the "loopy" sum-product algorithm can be characterized in terms of optimization: in particular, Yedidia *et al.* [18] showed its fixed points correspond to stationary points of the "Bethe free energy." This important result not only provides an analysis tool, but also motivates seeking alternative algorithms via other optimization-based formulations.

The framework that we pursue in this paper begins by formulating the problem of exact marginalization as an optimization problem; our approximate marginalization algorithm is based on solving a relaxed version of this exact formulation. More precisely, we show in Section III how the general problem of computing marginal probabilities in graphical models for discrete random variables (and for a more general class of models known as "exponential family models") can be formulated as a *convex* optimization problem—the problem involves the maximization of a certain concave cost function over a convex set. Both of these mathematical objects—the cost function and the constraint set—can be complex, however, and the optimization problem is intractable in general. The "Bethe free energy" approach involves approximating the cost function and relaxing to a simpler constraint set. However, as the Bethe free energy is often nonconvex, the loopy sum-product algorithm may have multiple fixed points, and may converge to a nonglobal optimum.

Rather than replacing a convex problem with a nonconvex one, it would seem desirable to maintain convexity in any relaxation. This is the contribution of the current paper: we propose a convex relaxation of the general problem of computing marginal probabilities on graphs with cycles. At the foundation of our method is a conjugate dual relation [19] that holds for any graphical model. The natural constraint set arising from this duality is the *marginal polytope* of all globally realizable marginal vectors. Our convex relaxation involves a semidefinite outer bound on the marginal polytope together with an upper bound on the conjugate dual function. The resulting problem is strictly convex, and its unique optimum can be found by efficient interior point methods [20]. We illustrate our relaxation in the context of denoising using the coupled mixture-of-Gaussians model of Crouse *et al.* [19] for graphs with cycles. As we will

show, the performance of our method is either comparable or superior to the sum-product algorithm over a wide range of experimental conditions [i.e., coupling strengths, signal-to-noise ratio (SNR)]. A significant fact is that the improvement in performance over sum-product is particularly large for more strongly coupled problems, the regime in which accounting for statistical dependency is most relevant.

The remainder of this paper is organized as follows. Section II is devoted to background on graphical Markov models, their use in modeling coupled mixture of Gaussians (MOGs), and the role of marginalization in signal processing applications. In Section III, we show how the problem of computing marginal probabilities can be reformulated as a low-dimensional optimization problem. In Section IV, we develop a convex relaxation of this optimization problem. Experimental results from applying this relaxation as a technique for performing denoising in the coupled mixture-of-Gaussians model are described in Section V.

II. Background and Problem Setup

In this section, we begin by providing some background on graphical Markov models; we refer the reader to [16] and [21] and survey papers [5] and [6] for further details. We then describe the coupled mixture-of-Gaussian model and its use in modeling and noisy prediction.

A. Graphical Markov Models

There exist various but essentially equivalent formalisms for describing graphical models, depending on the type of graph used. In this paper, we make use of an undirected graph $G = (V, E)$, where $V = \{1, \ldots, n\}$ is the vertex set and E is a set of edges joining pairs of vertices. We say that a set $S \subset V$ is a *vertex cutset* in G if removing S from V breaks the graph into two or more disconnected components—say, A and B. To each node $s \in V$, we associate a random variable X_s taking values in some configuration space $\mathcal{X}_s$. For any subset $A \subseteq V$, we define $X_A := \{X_s | s \in A\}$ with configuration space $\mathcal{X}_A$. The link between the random vector $X \equiv X_V$ and the graphical structure arises from Markov properties that are imposed by the graph. In particular, the random vector X is a *Markov random field* with respect to the graph G if, for all subsets A and B that are separated by some vertex cutset S, the random variables X_A and X_B are conditionally independent given X_S. Fig. 1(a) provides a simple example of a graphical Markov model, in which the random variables $X_1, \ldots, X_n$ of a Markov chain are associated with the nodes of a chain. In this chain, each vertex is a cutset; this property implies the familiar conditional independence properties of a Markov chain, in which the past and future are conditionally independent given the present.

An alternative specification of a Markov random field (MRF) is in terms of a particular factorization of the distribution that respects the structure of the graph. In this paper, we focus on pairwise MRFs, in which the factorization is specified by terms associated with nodes and edges of the underlying graph. It is convenient to specify the factorization as an additive decomposition in the exponential domain, as we now describe for a discrete random vector X (i.e., for which $\mathcal{X}_s = \{0, 1, \ldots, m\}$. For each $s \in V$ and $j \in \mathcal{X}_s$, let us define an indicator function $\mathbb{I}_j[x_s]$ (equal to one if $x_s = j$ and zero otherwise). We then consider the singleton functions θ_s at each node, and coupling functions θ_{st} on each edge (s, t) that are weighted combinations of these indicator functions

$$\begin{aligned} \theta_s(x_s) &= \sum_{j \in \mathcal{X}_s} \theta_{s;j} \mathbb{I}_j[x_s] \\ \theta_{st}(x_s, x_t) &= \sum_{(j,k)} \theta_{st;jk} \mathbb{I}_j[x_s] \mathbb{I}_k[x_t]. \end{aligned} \tag{1}$$

The collection of functions $\phi(x) := \{\mathbb{I}_j[x_s]\} \cup \{\mathbb{I}_j[x_s]\mathbb{I}_k[x_t]\}$ are known as the *sufficient statistics,* and the vector θ with elements $\{\theta_{s;j}\} \cup \{\theta_{st;jk}\}$ is the (canonical) *parameter vector*. As we have described it, the vector θ is d'-dimensional, where $d' = |V|m + |E|m^2$. In fact, since the indicator functions are linearly dependent (e.g., $\sum_j \mathbb{I}_j[x_s] = 1$ for all x_s), it is possible to describe the same model family using only a total of $d = |V|(m-1) + |E|(m-1)^2$ parameters.

For a pairwise MRF, the distribution of the random vector X, denoted by $p(x; \theta) \equiv p(X = x; \theta)$, decomposes as

$$p(x; \theta) = \exp\left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) - A(\theta) \right\} \tag{2a}$$

where the quantity

$$\begin{aligned} A(\theta) := \log \sum_{x \in \mathcal{X}^n} \exp & \\ \times \left\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \right\} & \end{aligned} \tag{2b}$$

is the *log partition function* that serves to ensure that the distribution is normalized properly. Note that A is a function from $\mathbb{R}^d$ to $\mathbb{R}$; from the definition (2b), it can be seen that A is both convex and continuous.

B. Coupled Mixture-of-Gaussians and Denoising

We now use the Markov random field to define a more general graphical model involving mixture variables. Let us view the random variable $X_s \in \{0, 1, \ldots, m-1\}$ as indexing a Gaussian mixture with m components. More concretely, we define a random variable Z_s whose conditional distribution is given by

$$\begin{aligned} & p\left(Z_s = z_s | X_s = j; \nu_s, \sigma_s^2\right) \\ & \quad = \frac{1}{\sqrt{2\pi\sigma_{s;j}^2}} \exp\left\{ -\frac{(z_s - \nu_{s;j})^2}{2\sigma_{s;j}^2} \right\}, \\ & \quad \text{for } j = 0, 1, \ldots, m-1 \end{aligned} \tag{3}$$

where $\sigma^2 := \{\sigma_{s;j}^2 | j \in \mathcal{X}_s\}$ and $\nu_s := \{\nu_{s;j} | j \in \mathcal{X}_s\}$ are m-vectors specifying the Gaussian variances and means respectively. Summing the joint distribution over the values of X_s

yields a Gaussian mixture distribution for Z_s. See Fig. 2(a) for a simple graphical representation of this model.

The *coupled mixture-of-Gaussians* model is a joint probability model over (X, Z), defined by the distribution

$$p(X = x, Z = z; \theta, \sigma^2, \nu) = p(x; \theta) \prod_{s \in V} p\left(z_s | x_s; \nu_s, \sigma_s^2\right). \tag{4}$$

The wavelet signal processing framework of Crouse *et al.* [9] involves a model of the form (4), in which the underlying graph is a tree and each variable Z_s is a mixture of $m = 2$ Gaussian components. Our main focus in this paper is the generalization of this model when the underlying graph has cycles, such as the lattice shown in Fig. 2(b).

One important application of the model (4) is to signal denoising. The problem setup is as follows: given a vector Y of *noisy observations* of the Gaussian mixture vector Z, we would like to use the noisy observations to form an optimal prediction of Z. When Z_s is no longer directly observed [as it was in Fig. 2(a)], we have the new local model illustrated in Fig. 2(c), in which the third (shaded) node represents the noisy observation variable Y_s. One common observation model, and the one that we consider in this paper, takes the form

$$Y = \alpha Z + \sqrt{1 - \alpha^2} W \tag{5}$$

where $W \sim \mathcal{N}(0, I)$ is a Gaussian random noise vector and $\alpha \in [0, 1]$ controls the SNR. For continuous random variables, it is common to assess prediction performance using the mean-squared error (MSE), in which case it is well known that the optimal predictor of Z_s, given observations $Y = y$, is the conditional mean $\widehat{z}_s(y) := \mathbb{E}[Z_s | Y = y]$. For an observation model of the form (5), it is straightforward to derive that the conditional mean takes the form

$$\widehat{z}_s(y) = \sum_{j \in \mathcal{X}_s} p(X_s = j | y; \theta) \times \left\{ \nu_{s;j} + \frac{\alpha \sigma_{s;j}^2}{\alpha^2 \sigma_{s;j}^2 + (1 - \alpha^2)} [y_s - \nu_{s;j}] \right\}. \tag{6}$$

Note that $\widehat{z}_s$ is a combination of linear least squares estimators (LLSEs), in which the LLSE for Gaussian component j is weighted by the marginal probability $p(X_s = j | y; \theta)$. Thus, the main challenge associated with performing optimal prediction is the computation of these marginal probabilities. For our development to follow, it is convenient to observe that since y is an observed quantity (and hence fixed), the computation of the conditional marginal distribution $p(X_s = j | y; \theta)$ can be reformulated as the computation of an ordinary marginal distribution $p(X_s = j; \widetilde{\theta})$, where $\widetilde{\theta}$ is a modified set of exponential parameters $\{\widetilde{\theta}_s, \widetilde{\theta}_{st}\}$ obtained by incorporating the observations into the model. Explicitly, the modified terms at each node have the form $\widetilde{\theta}_{s;k} = \theta_{s;k} + \log p(y | X_s = k; \theta, \nu, \sigma)$ for each $k \in \{0, 1, \ldots, m\}$; the coupling terms remain unchanged (i.e., $\theta_{st} = \widetilde{\theta}_{st}$).

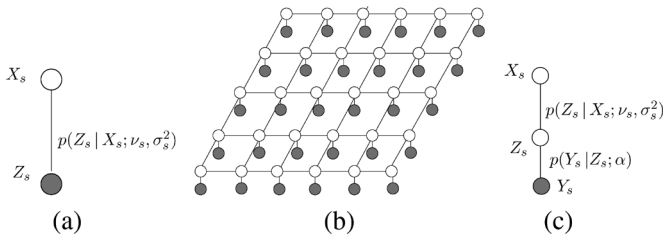

Fig. 2. (a) A simple graphical model showing a scalar mixture-of-Gaussian (MOG) model. (b) A coupled MOG model defined on a four nearest neighbor lattice. (c) Modification to each local module when given noisy observations Y_s of Z_s.

III. Exact Variational Formulation

In this section, we show how the problem of computing marginals and the log partition function can be reformulated as the solution of a low-dimensional convex optimization problem, which we refer to as a *variational formulation*. We discuss the challenges associated with an exact solution of this optimization problem for general Markov random fields.

At a high level, our strategy for obtaining the desired variational representation can be summarized as follows. Recall that the log partition function A maps parameter vectors $\theta \in \mathbb{R}^d$ to real numbers and is a convex function of θ. The convexity of A means that its epigraph $\{(\theta, t) | A(\theta) \leq t\}$ is a convex subset of $\mathbb{R}^{d+1}$, and therefore can be characterized as the intersection of all half-spaces that contain it [19]. This half-space representation of the epigraph is equivalent to saying that A can be written in a *variational fashion* as follows:

$$A(\theta) = \sup_{\mu \in \mathbb{R}^d} \{\langle \theta, \mu \rangle - A^*(\mu)\}. \tag{7}$$

Here $\langle \theta \rangle$ denotes the Euclidean inner product between the vectors $\theta, \mu \in \mathbb{R}^d$ and A^* is an auxiliary function, known as the conjugate dual, that we describe in more detail in Section III-B. In geometric terms, the dual vector μ represents the slope of the hyperplane describing a half-space, whereas the dual value $A^*(\mu)$ represents the (negative) intercept of the hyperplane. In principle, this conjugate dual relation allows us to compute $A(\theta)$ by solving the optimization problem (7). Accordingly, we first turn to an investigation of the form of the dual function A^*. Of particular importance is characterizing the subset of $\mathbb{R}^d$ on which A^* is finite-valued (known as its *effective domain*) because we can always restrict the optimization (7) from $\mathbb{R}^d$ to this set.

Although the variational principle (7) is related to the "free energy" principle of statistical physics [18], it differs from this classical approach in important ways. More specifically, it is *low-dimensional* convex problem, in which the optimization variables μ have a natural interpretation as realizable marginal probabilities. An important consequence, as we will see later in this section, is that an effective solution to the optimization problem (7) yields not only the value of the log partition function $A(\theta)$ but it also the marginal probabilities (i.e., $p(X_s = j; \theta)$ and $p(X_s = j, X_t = k; \theta)$) that we aim to compute. Moreover, in contrast to the free energy approach, this perspective clarifies that there are two distinct components to any relaxation of the variational principle (7)—namely, an approximation of the dual function and an outer bound on the effective domain.

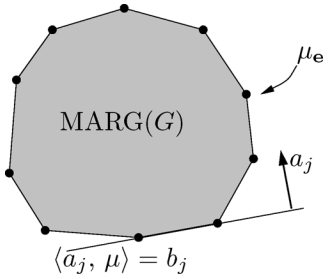

Fig. 3. Geometrical illustration of a marginal polytope. Each vertex corresponds to the mean parameter $\mu_{\mathbf{e}} := \phi(\mathbf{e})$ realized by the distribution $\delta_{\mathbf{e}}(x)$ that puts all of its mass on the configuration $\mathbf{e} \in \mathcal{X}^n$. The facets of the marginal polytope are specified by hyperplane constraints $\langle a_j \mu \rangle \leq b_j$.

A. Marginal Polytopes

We begin by defining the set that is to play the role of the constraint set in our variational principle. Recall the sufficient statistics in a discrete exponential family: they are the collection of indicator functions $\phi(x) := \{\mathbb{I}_j[x_s]\} \cup \{\mathbb{I}_j[x_s]\mathbb{I}_k[x_t]\}$. For each discrete configuration $x \in \mathcal{X}^n$, the quantity $\phi(x)$ is a $\{0\text{–}1\}$-valued vector contained within $[0,1]^d$; of interest is the convex hull of this set of vectors. More formally, we define

$$\text{MARG}(G;\boldsymbol{\phi}) := \left\{ \mu \in \mathbb{R}^d \middle| \sum_{x \in \mathcal{X}^n} \phi(x) p(x) = \mu \quad \text{for some distribution } p(\cdot) \right\} \tag{8}$$

where $p(\cdot)$ is any valid distribution. The elements of μ, which can be indexed as $\{\mu_{s;j}\} \cup \{\mu_{st;jk}\}$, have a very concrete interpretation. For instance, element $\mu_{s;j} = \sum_x p(x)\mathbb{I}_j[x_s]$ is simply the marginal probability that $X_s = j$ (under the distribution $p(\cdot)$). Similarly, element $\mu_{st;jk}$ corresponds to a particular joint marginal probability. Accordingly, we refer to the set $\text{MARG}(G;\phi)$ as the *marginal polytope* associated with the graph G and the potentials ϕ. We refer to elements $\mu \in \text{MARG}(G;\phi)$ as *mean parameters* associated with the Markov random field defined by G and ϕ. Fig. 3 provides a geometric illustration of a marginal polytope. Since it corresponds to the convex hull of a finite number of vectors, it must be a polytope (and hence can be characterized by a finite number of hyperplane constraints). Although $\text{MARG}(G;\phi)$ has a very simple definition, it is actually a rather complicated set. In particular, the number of hyperplane constraints required to specify this polytope grows at least exponentially in the graph size n (see [22] and [23] for further discussion of this point).

B. Conjugate Dual Function

Given the convexity of A, it is natural to consider its conjugate dual [19], which is a function $A^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$ defined as follows:

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

Here the vector of *dual variables* $\mu \in \mathbb{R}^d$ is the same dimension as the vector θ of exponential parameters.

Our choice of notation is deliberately suggestive, in that the dual variables μ turn out to be closely associated with the marginal polytope defined in the previous section. In order to see the connection, consider the gradient of the log partition function. A straightforward calculation using the definition (2b) of A yields that

$$\nabla A(\theta) = \sum_{x \in \mathcal{X}^n} p(x;\theta)\boldsymbol{\phi}(x) = \mathbb{E}_\theta[\boldsymbol{\phi}(x)] \tag{10}$$

so that elements of this gradient correspond to particular marginal probabilities (under the distribution $p(\cdot;\theta)$). This fact implies that the image of the gradient mapping (i.e., $\nabla A(\mathbb{R}^d)$) is contained within the marginal polytope (8).

Our goal now is to compute a more explicit form for the dual function A^*. A quantity that plays a key role in this context is the *discrete entropy* [24] associated with a distribution $p(\cdot)$, defined as

$$H(p) := -\sum_{x \in^n} p(x) \log p(x). \tag{11}$$

We begin by observing that for a fixed $\mu \in \mathbb{R}^d$, the function $J(\theta) := \langle \mu, \theta \rangle - A(\theta)$ is strictly concave and differentiable. Therefore, if there exists a solution θ to the equation $\nabla J(\theta) = 0$, then the supremum (9) is attained at this point. Accordingly, we compute the gradient using (10) and set it equal to zero

$$\nabla J(\theta) = \mu - \mathbb{E}_\theta[\boldsymbol{\phi}(x)] = 0. \tag{12}$$

It can be shown [23] that this equation has a unique solution $\theta(\mu)$ whenever μ belongs to the interior of $\text{MARG}(G;\phi)$. Substituting the relation $\mu = \mathbb{E}_{\theta(\mu)}[\phi(x)]$ into the definition (9) of A^* yields, for any μ in the interior of $\text{MARG}(G;\phi)$, the important relation $A^*(\mu) = \langle \mu, \theta(\mu) \rangle - A(\theta(\mu))$. To interpret $A^*(\mu)$, we compute the negative entropy of $p(x;\theta)$ as follows:

$$\begin{aligned} -H\left(p\left(x;\theta(\mu)\right)\right) &= \sum_{x \in \mathcal{X}^n} p\left(x;\theta(\mu)\right)\{\log p\left(x;\theta(\mu)\right)\} \\ &= \sum_{x \in \mathcal{X}^n} p\left(x;\theta(\mu)\right)\{\langle \theta(\mu), \boldsymbol{\phi}(x) \rangle - A\left(\theta(\mu)\right)\} \\ &= A^*(\mu). \end{aligned} \tag{13}$$

Thus, we recognize the dual function $A^*(\mu)$ as the negative entropy specified as a function of the mean parameters.

We have established that for μ in the interior of $\text{MARG}(G;\phi)$, the value of the conjugate dual $A^*(\mu)$ is given by the negative entropy of the distribution $p(x;\theta(\mu))$, where the pair $\theta(\mu)$ and μ are dually coupled via (12). Moreover, it can be shown [23] that when μ lies outside the closure of the marginal polytope, then the value of the dual function is $+\infty$. We summarize as follows:

$$A^*(\mu) = \begin{cases} -H\left(p(x;\theta(\mu))\right), & \text{for } \mu \text{ in the interior of} \\ & \quad \text{MARG}(G;\boldsymbol{\phi}) \\ +\infty, & \text{otherwise.} \end{cases} \tag{14}$$

Since the function A is differentiable on $\mathbb{R}^d$, we are guaranteed that taking the conjugate dual twice recovers the original func-

tion [19]. By applying this fact to A^* and A, we obtain the following relation:

$$A(\theta) = \max_{\mu \in \mathrm{MARG}(G,\boldsymbol{\phi})} \{\langle \theta, \mu \rangle - A^*(\mu)\}. \tag{15}$$

Note that the optimization here, in contrast to (7), is restricted to the marginal polytope $\mathrm{MARG}(G;\phi)$, since the function A^* is infinite outside of this set. Moreover, it can be shown [23] that the optimum (15) is attained uniquely at the vector $\mu(\theta)$ of marginals associated with $p(x;\theta)$. Consequently, the optimization problem (15) is a variational representation in two senses. First, the optimal *value* of the optimization problem yields the value $A(\theta)$ of the log partition function; and second, the optimal *arguments* yield the mean parameters or marginals $\mu = \mathbb{E}_\theta[\phi(x)]$. This variational formulation plays a central role in our development in the sequel.

IV. Log-Determinant Relaxation

In this section, we derive an algorithm for approximating marginal probabilities based on the solution of a relaxed variational problem involving determinant maximization and semidefinite constraints. We first provide a high-level description of our approach. There are two challenges associated with the variational formulation (15). First, actually evaluating the dual function $A^*(\mu)$ for some vector μ of marginal probabilities is a very challenging problem, since it requires first computing the exponential family distribution $p(x;\theta(\mu))$ with those marginals, and then computing its entropy. Indeed, with the exception of trees and more general junction trees, it is typically impossible to specify an explicit form for the dual function $A^*(\mu)$. Second, for a general graph with cycles, an exact description of the marginal polytope $\mathrm{MARG}(G;\phi)$ requires a number of inequalities that grows rapidly with the size of the graph [22]. Accordingly, our approach is to relax the exact variational formulation (15) as follows: we replace the marginal polytope by a convex outer bound, and bound the intractable dual function A^* with a convex surrogate. In the following two sections, we describe each of these steps in turn.

Although the ideas and methods described here are more generally applicable, for the sake of clarity in exposition we focus here on the case of a binary random vector $X \in \{0,1\}^n$. In this case, it is necessary only to consider the indicator functions $\mathbb{I}_1[x_s] = x_s$ at each node and $\mathbb{I}_1[x_s]\mathbb{I}_1[x_t] = x_s x_t$ on each edge. Thus, for a given graph G, the parameter vector θ has a total of $d = |V| + |E|$ elements (i.e., its elements have the form $\{\theta_s, s \in V\} \cup \{\theta_{st}, (s,t) \in E\}$). We refer the reader to Appendix A for discussion of general multinomial case.

A. Outer Bounds on the Marginal Polytope

We first show how to derive various outer bounds on the marginal polytope. In this context, it is convenient to consider the marginal polytope $\mathrm{MARG}(K_n;\phi)$ associated with the complete graph K_n (i.e., the graph in which each node is joined by an edge to all $(n-1)$ other nodes). It should be noted that this assumption entails no loss of generality, since an arbitrary pairwise Markov random field can be embedded into the complete graph by setting to zero a subset of the $\{\theta_{st}\}$ parameters. (In particular, for a graph $G = (V,E)$, we simply set $\theta_{st} = 0$ for all pairs $(s,t) \notin E$).

On the complete graph, the model dimension is $d = n + \binom{n}{2}$. Given an arbitrary vector $\mu \in \mathbb{R}^d$, consider the following $(n+1) \times (n+1)$ matrix:

$$M_1[\mu] := \begin{bmatrix} 1 & \mu_1 & \mu_2 & \cdots & \mu_{n-1} & \mu_n \\ \mu_1 & \mu_1 & \mu_{12} & \cdots & \cdots & \mu_{1n} \\ \mu_2 & \mu_{21} & \mu_2 & \cdots & \cdots & \mu_{2n} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \mu_{n-1} & \vdots & \vdots & \vdots & \vdots & \mu_{n,(n-1)} \\ \mu_n & \mu_{n1} & \mu_{n2} & \cdots & \mu_{(n-1),n} & \mu_n \end{bmatrix}. \tag{16}$$

The motivation underlying this definition is the following: suppose that the given dual vector μ actually belongs to $\mathrm{MARG}(K_n;\phi)$, in which case there exists some distribution $p(x;\theta)$ such that $\mu_s = \sum_x p(x;\theta) x_s$ and $\mu_{st} = \sum_x p(x;\theta) x_s x_t$. Under this condition, the matrix $M_1[\mu]$ can be interpreted as the matrix of second-order moments for the vector $(1,x)$, as computed under $p(x;\theta)$. An important point here is that in computing these second-order moments, we use the fact that $x_s^2 = x_s$ when $x_s \in \{0,1\}$. It is for this reason that the elements $(\mu_1, \ldots, \mu_n)$ appear along the diagonal.

Given a symmetric matrix $S \in \mathbb{R}^{d \times d}$, we use $S \succeq 0$ to mean that S is positive semidefinite. The significance of the moment matrix $M_1[\mu]$ is illustrated in the following.

Lemma 1 [Semidefinite Outer Bound]: The binary marginal polytope $\mathrm{MARG}(K_n)$ is contained within the semidefinite constraint set $\mathrm{SDEF}_1(K_n) := \{\mu \in \mathbb{R}^d | M_1[\mu] \succeq 0\}$.

Proof: This result follows from the fact that any second-order moment matrix must be positive semidefinite, as can be verified by the following simple argument. Letting $y := (1,x)$, then for any vector $a \in \mathbb{R}^{n+1}$, we have $a^T M_1[\mu] a = a^T \mathbb{E}[yy^T] a = \mathbb{E}[\|a^T y\|^2]$, which is certainly nonnegative. □

This semidefinite relaxation can be further strengthened by including higher order terms in the moment matrices, as described by Lasserre [25].

In addition to such semidefinite constraints, there are various linear constraints that any member μ of the marginal polytope $\mathrm{MARG}(G;\phi)$ must satisfy. Here we consider some linear constraints that are relevant to the sum-product algorithm. For a given edge $(s,t) \in E$, there are three mean parameters associated with each pair of random variables (X_s, X_t)—namely, $\mu_s = \mathbb{E}[X_t]$, $\mu_t = \mathbb{E}[X_t]$ and $\mu_{st} = \mathbb{E}[X_s X_t]$. Collectively, these three mean parameters specify the joint marginal distribution over (X_s, X_t) as

$$p(x_s, x_t; \mu) = \begin{bmatrix} (1 + \mu_{st} - \mu_s - \mu_t) & \mu_t - \mu_{st} \\ \mu_s - \mu st & \mu_{st} \end{bmatrix}. \tag{17}$$

Therefore, the four inequality constraints obtained by requiring that each entry of $p(x_s, x_t; \mu)$ be nonnegative are necessary and sufficient to ensure that the pairwise marginals on each edge

(s,t) are valid. For a binary-valued Markov random field defined on a general graph with cycles, these constraints are equivalent to the constraints that are enforced by the sum-product algorithm [18]. Thus, these constraints provide a complete characterization of the marginal polytope for any tree-structured graph.

B. Gaussian Entropy Bound

We now describe how to upper bound the (negative) dual function $-A^*(\mu)$ using a Gaussian-type approximation. Our starting point is the well-known fact [24] that the (differential) entropy of any continuous random vector $\widetilde{X}$ is upper bounded as

$$h(\widetilde{X}) \leq \frac{1}{2}\log\det \mathrm{cov}(\widetilde{X}) + \frac{n}{2}\log(2\pi e). \tag{18}$$

This upper bound corresponds to the differential entropy of a Gaussian matched to the covariance matrix (denoted $\mathrm{cov}(\widetilde{X})$) of the continuous random vector $\widetilde{X}$. However, it is not directly useful in application to a discrete random vector, for which the differential entropy is not well-defined. Our approach is to "smooth" X by adding an independent noise variable, and then apply the bound (18). After some derivation, the end result is the following:

Lemma 2: The negative dual function is upper bounded as

$$-A^*(\mu) \leq \frac{1}{2}\log\det\left(M_1[\mu] + \frac{1}{12}\mathrm{blkdiag}[0, I_n]\right) + \frac{n}{2}\log(2\pi e) \tag{19}$$

where the $(n+1)\times(n+1)$ matrix $M_1[\mu]$ is defined in (16) and $\mathrm{blkdiag}[0, I_n]$ is an $(n+1)\times(n+1)$ block-diagonal matrix with a 1×1 zero block, and another $n \times n$ block with the identity matrix I_n.

See Appendix B for the proof of this claim.

C. Log-Determinant Relaxation

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

Theorem 1: Consider a binary Markov random field $p(x;\theta)$ over the vector $X \in \{0,1\}^n$. Then for any compact convex outer bound $\mathrm{OUT}(K_n)$ on the marginal polytope $\mathrm{MARG}(K_n)$, the log partition function $A(\theta)$ is upper bounded by the solution of the following variational problem:

$$A(\theta) \leq \max_{\mu\in\mathrm{OUT}(K_n), M_1[\mu]\succeq 0} \times\left\{\langle\theta,\mu\rangle + \frac{1}{2}\log\det\times\left[M_1(\mu) + \frac{1}{12}\mathrm{blkdiag}[0,I_n]\right]\right\} + \frac{n}{2}\log(2\pi e). \tag{20}$$

This problem is strictly concave, and so has a unique global optimum.

Given our development thus far, the proof is straightforward. In particular, by examining the variational representation (15) of A, we see that an upper bound on A can be obtained via an upper bound on the negative dual function $-A^*$ and an outer bound on the marginal polytope $\mathrm{MARG}(K_n;\phi)$. The bound thus follows by applying Lemmas 1 and 2. Strict concavity and uniqueness follow by standard results on the log-determinant function [20].

The simplest form of the relaxation (20) is obtained when the semidefinite constraint (see Lemma 1) $M_1[\mu] \succeq 0$ is the only constraint imposed. In this case, the relaxation has a natural interpretation as optimizing over a subset of valid covariance matrices. It is also straightforward to strengthen the relaxation via additional linear constraints on the marginal polytope, such as those associated with the Bethe problem and the sum-product algorithm [see (17)]. Overall, our approach will be to proceed in analogy to the exact variational principle (15): in particular, we will solve (20) and then make use of the optimizing arguments $\widehat{\mu}$ as *approximations* to the exact marginals. Insofar as our relaxation is relatively tight, this procedure can be expected to provide reasonable approximations, as we will see in Section V.

D. Efficient Solution of Log-Determinant Relaxation

An important fact is that the unique optimum of problem (20) can be obtained in polynomial time by interior point methods specialized for log-determinant problems [e.g., [20]]. However, the complexity of a generic interior point method is $\mathcal{O}(n^6)$, which (though polynomial time) is too large to be practically viable. Accordingly, here we describe how a suitable dual reformulation leads to very efficient methods for solving a slightly weakened form of the log-determinant relaxation in Theorem 1.

Our starting point is the observation that the log-determinant term in (20) acts naturally as a barrier function to enforce the constraint $M_1[\mu] \succeq -(1/12)\mathrm{blkdiag}[0, I_n]$, which is a somewhat weaker constraint than $M_1[\mu] \succeq 0$. This observation leads to the relaxed problem

$$\max_{\mu}\left\{\langle\theta,\mu\rangle + \frac{1}{2}\log\det\left[M_1(\mu) + \frac{1}{12}\mathrm{blkdiag}[0,I_n]\right]\right\}. \tag{21}$$

By an appropriate dual reformulation described in Appendix C, we can convert (21) into an *unconstrained* log-determinant problem that can be solved with i) complexity $\mathcal{O}(n^3)$ per iteration for a full Newton method on an arbitrary graph or ii) complexity $\mathcal{O}(n^{1.5})$ per iteration for a diagonally scaled quasi-Newton method on grid-structured problems. By comparison, the computational complexity per iteration of sum-product is $\mathcal{O}(|E|)$. As particular illustrations, this amounts to a complexity per iteration of $\mathcal{O}(n^2)$ for complete graphs and $\mathcal{O}(n)$ for grid-structured problems. The overall complexity is determined by the per iteration cost as well as the convergence rate, which determines the total number of iterations required to reach a pre-specified error tolerance. The convergence rate of the sum-product algorithm (assuming that it converges) is at best linear [26]. In contrast, an appropriately scaled gradient method, like Newton's method, has a superlinear rate of convergence [27].

V. Experimental Results

In this section, we describe experimental results of applying the log-determinant relaxation (21) to the noisy prediction

problem in a coupled mixture-of-Gaussians (MOG) model (see Section II-B). Recall that the MOG model is specified by a discrete MRF $p(x;\theta)$ over the vector X of discrete mixing variables, a vector Z of Gaussian mixture variables, and a vector Y of noisy observations, as described by (5). The noisy prediction problem is to compute for each vertex $s \in V$ the conditional mean $\widehat{z}_s(y)$, as defined in (6). We consider the approximation $\widehat{z}_s(y;\mu^{\mathrm{LD}})$ to this conditional mean, in which the true marginals $p(X_s = j|y;\theta)$ are replaced by the approximations μ^{LD} obtained from the log-determinant relaxation. For all experiments reported here, we fix the variances $\sigma_{s;0} = \sigma_{s;1} = 0.25$ and means $\nu_{s;0} = 1 = -\nu_{s;1}$ of the Gaussian mixture components. We study the behavior of the log-determinant (LD) relaxation (21) as the coupling strengths θ and the SNR α are varied. For purposes of comparison, we also show prediction results based on the approximate predictor $\widehat{z}_s(y;\mu^{\mathrm{BP}})$ computed using the vector μ^{BP} of approximate marginals obtained from the sum-product or belief propagation (BP) algorithm (see [18, (4) and (5)]).

Our experiments cover two types of graphs: the complete graph K_n, in which each node is connected to every other, and the four nearest neighbor lattice graph [see Fig. 2(b) for an illustration]. For each set of trials on a given graph, we generate the parameter θ that specifies the coupling distribution (2a) $p(X;\theta)$ in the following way. In all trials, we set the single node parameters $\theta_{s;0} = \theta_{s;1} = 0$ for all vertices $s \in V$. The edge parameters θ_{st} are chosen differently depending on the *coupling strength* $d_{\mathrm{coup}} \geq 0$; here $d_{\mathrm{coup}} = 0$ corresponds to an independence model, whereas larger d_{coup} generates increasing amounts of dependence among the indicator variables X_s. Let $\mathcal{U}[a,b]$ denote the uniform distribution on the interval $[a,b]$. For a fixed coupling strength d_{coup}, we sample γ_{st} independently from the $\mathcal{U}[0,d_{\mathrm{coup}}]$ distribution and then set

$$\theta_{st} = \begin{bmatrix} \theta_{st;00} & \theta_{st;01} \\ \theta_{st;10} & \theta_{st;11} \end{bmatrix} = \begin{bmatrix} \gamma_{st} & -\gamma_{st} \\ -\gamma_{st} & \gamma_{st} \end{bmatrix}.$$

Note that this procedure for choosing the parameter vector θ produces a distribution $p(x;\theta)$ in which each variable X_s is equally likely to be zero or one, and for which neighboring variables are more likely to take the same value. This probabilistic structure is consistent with the coupled mixture models used in practice (e.g., in wavelet denoising [9]).

We solved the log-determinant relaxation (21) via Newton's method, as described in Appendix C. We used the standard parallel message-passing form of the sum-product algorithm with a damping factor[1] $\beta = 0.05$; if the sum-product algorithm failed to converge, we switched to a convergent double-loop alternative [28]. For each graph (fully connected or grid), we examined a range of coupling strengths d_{coup} and the full range of SNRs parameterized by $\alpha \in [0,1]$. For purposes of comparison, we computed the exact marginal values either by exhaustive summation on the complete graph or by applying the junction tree algorithm to grid-structured problems. Due to the computational complexity of these exact calculations, we performed our experiments on $n = 16$ nodes for complete graphs and $n = 100$ nodes for the lattices. We used the following procedure to assess the error in the approximations. Let $\mathrm{MSE}^{\mathrm{Bayes}} := \mathbb{E}\{(1/n)\sum_{s=1}^{n}[\widehat{z}_s(y) - z_s]^2\}$ denote the MSE of the optimal Bayes estimator (6). Similarly, let $\mathrm{MSE}(\mu^{\mathrm{LD}})$ and $\mathrm{MSE}(\mu^{\mathrm{BP}})$ denote the MSEs of the log-determinant and BP-based predictors, respectively. For any given experimental trial, our evaluation is based on the percentage increase in MSE; for instance, for the log-determinant predictor, we compute $100\times((\mathrm{MSE}(\mu^{\mathrm{LD}}) - \mathrm{MSEB}^{\mathrm{Bayes}})/\mathrm{MSE}^{\mathrm{Bayes}})$.

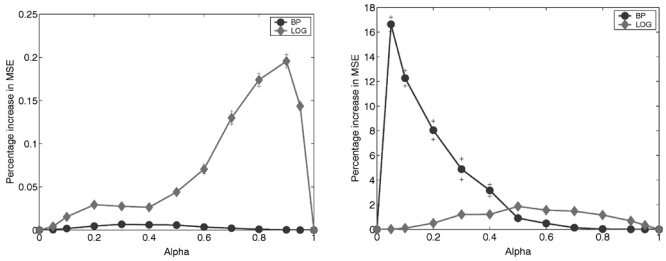

Fig. 4. Percentage increase in MSE as a function of the SNR parameter $\alpha \in [0,1]$ for a four nearest-neighbor grid with $n = 100$ nodes. Each point on each line is the average over 50 trials; the plus signs provide the standard errors. (a) Edge coupling strength $d_{\mathrm{coup}} = 0.45$. (b) Edge coupling strength $d_{\mathrm{coup}} = 0.90$. Note that there is a factor of ten difference in the vertical scales between panels (a) and (b).

Fig. 4 shows the results for the grid with $n = 100$ nodes; each plot displays the percentage increase in MSE versus the SNR parameter α for a fixed coupling strength d_{coup}. Each point in each solid line corresponds to the mean taken over 30 trials; the plus marks show the standard errors associated with these estimates. Shown in panel (a) is the case of low coupling strength $(d_{\mathrm{coup}} = 0.45)$, for which the mixing variables on the grid interact only weakly. For these types of problems, the performance of BP is slightly but consistently better than the LD method; however, note that both methods lead to a percentage increase in MSE of less than 0.25% over all α. Panel (b), in contrast, shows the case of stronger coupling $(d_{\mathrm{coup}} = 0.90)$. Here the percentage loss in MSE can be quite substantial for BP in the low SNR region (α small), whereas its behavior improves for high SNR. In contrast, the behavior of the LD method is more stable, with the percentage MSE loss remaining less than 2% over the entire range of α. The degradation of BP performance for strong couplings can be attributed to the nonconvexity of the Bethe problem and the appearance of multiple local optima.

Fig. 5 shows analogous results for the fully connected graph. In panel (a), we see the weakly coupled case $(d_{\mathrm{coup}} = 0.075)$; as with the lattice model in Fig. 4(a), the performance loss for either method is less than 2.5% when the dependencies are weak. The results in panel (b), where the couplings are stronger $(d_{\mathrm{coup}} = 0.15)$, are markedly different: for harder problems with low SNR, the BP method can show MSE percentage increases upwards of 25%. Over the same range of α, the loss in the LD method remains less than 5%. It should be noted that the relatively poor performance of BP for this fully connected graph is to be expected in some sense, since it is an approximation that is essentially tree-based. Nonetheless, it is interesting that the performance of the LD method remains reasonable

[1]To be more precise, we applied damping the log domain as $\beta \log M_{st} + (1-\beta)\log M_{st}$.

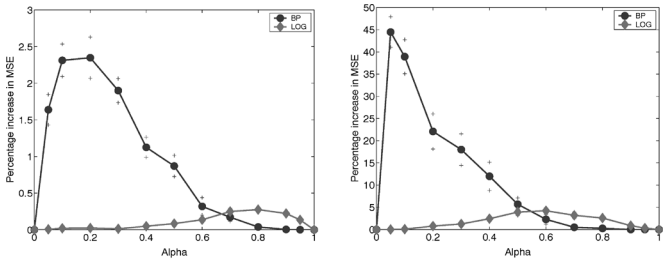

Fig. 5. Percentage increase in MSE as a function of the SNR parameter $\alpha \in [0, 1]$ for a fully connected graph K_{16}. Each point on each line is the average over 50 trials; the plus signs provide the standard errors. (a) Edge coupling strength $d_{\text{coup}} = 0.075$. (b) Edge coupling strength $d_{\text{coup}} = 0.15$. Note that there is a factor of ten difference in vertical scales between panels (a) and (b).

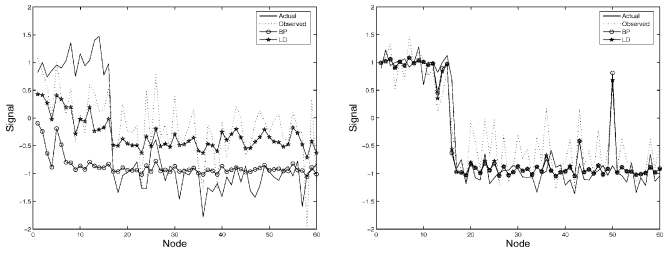

Fig. 6. Cross-sections of the actual signal z (solid line), noisy observation y (dotted lines), as well as the BP-based and LD-based signal reconstructions. (a) Low SNR example ($\alpha = 0.30$). (b) High SNR example ($\alpha = 0.70$).

even for this densely connected model over the same range of coupling strengths and SNRs.

Fig. 6 shows two cross-sections taken from a grid for coupling strength $d_{\text{coup}} = 0.90$; panel (a) shows the low SNR behavior ($\alpha = 0.30$), whereas panel (b) shows the behavior for higher SNR ($\alpha = 0.70$). The cross-sections were chosen to illustrate a step discontinuity in the underlying signal, at which point the mixing variable switches from $X_s = 0$ (Gaussian mean $\nu_{s;0} = 1$) to $X_s = 1$ (Gaussian mean $\nu_{s;1} = -1$). In the low SNR example, note that the BP algorithm outputs approximates marginals that are strongly skewed toward state $X_s = 1$. As a result, the noisy signal reconstruction is biased toward $\nu_{s;1} = -1$. In contrast, the LD method returns approximate marginals that are more balanced between the two mixture components, and thus reconstructs a smoother version of the step that is centered about zero. For the high SNR example in panel (b), both methods perform quite well as would be expected.

VI. Discussion

Graphical models have proven useful in a variety of signal processing applications. Although exact algorithms for cycle-free graphs are widely used, the algorithmic treatment of graphs with cycles presents a number of challenges that must be addressed in order for these models to be applied to signal processing problems. The foundation of this paper is an exact variational representation of the problem of computing marginals in general Markov random fields. We demonstrated the utility of semidefinite constraints in developing convex relaxations of this exact principle, which can be solved efficiently to obtain approximations of marginal distributions. The method presented here is based on a Gaussian entropy bound in conjunction with both linear and semidefinite constraints on the marginal polytope. An attractive feature of the resulting log-determinant maximization problem is that it can be solved rapidly by efficient interior point methods [20]. Moreover, we showed how a slightly modified log-determinant relaxation can be solved even more quickly by conventional Newton-like methods. As an illustration of our methods, we applied our relaxation to a noisy prediction problem in a coupled mixture-of-Gaussians problem and found that it performed well over a range of coupling strengths and SNR settings. An important open question, not addressed in this paper, is how to estimate model parameters for such prediction problems.

There are a number of additional research directions suggested by the methods proposed here. It remains to develop a deeper understanding of the interaction between the two choices involved in these approximations (i.e., the entropy bound, and the outer bound on the marginal polytope), as well as how to tailor approximations to particular graph structures. It is certainly possible to combine semidefinite constraints with entropy approximations (preferably convex) other than the Gaussian bound used in this paper. For instance, it would be interesting to investigate the behavior of "convexified" Bethe/Kikuchi entropy approximations [29] in conjunction with semidefinite constraints.

Appendix

A. Generalization to Multinomial States

This Appendix discusses the extension of our techniques to discrete spaces $\mathcal{X}_s = \{0, 1, \ldots, m-1\}$ with $m > 2$; our treatment is necessarily brief due to space constraints. Given a random variable $X_s \in \mathcal{X}_s$, one set of sufficient statistics is the vector of monomials $f(x_s) := \{x_s^j, j = 1, \ldots, m-1\}$. The overall distribution on the random vector X can be represented in terms of these monomials, and the cross-terms $f(x_s) \otimes f(x_t) := \{x_s^j x_t^k, j, k = 1, \ldots, m-1\}$ for each edge $(s, t) \in E$. Note that this is a natural generalization of the representation $\{x_s, x_s x_t\}$ that we used in the binary case. Let $\mu_s^j = \mathbb{E}[X_s^j]$ and $\mu_{st}^{jk} = \mathbb{E}[X_s^j X_t^k]$ be the associated mean parameters. Let us define a $q \times q$ dimensional matrix, where $q = 1 + (m-1)n$, by taking the covariance $\text{cov}\{1, f(X_1), \ldots, f(X_n)\}$. The elements of this matrix are given in terms of the mean parameters defined above; it is the natural generalization of the matrix $M_1[\mu]$ defined previously (16) for binary variables. As in Lemma 1, imposing a semidefinite constraint on this matrix generates an outer bound on the associated marginal polytope. Similarly, we can also derive an upper bound on the entropy of X based on the characterization of the Gaussian distribution as the maximum entropy distribution for a fixed covariance matrix. A first step in doing so is the observation that $H(X_1, \ldots, X_n) = H(f(X_1), \ldots, f(X_n))$; we can then upper bound the entropy in terms of the covariance matrix $\text{cov}\{1, f(X_1), \ldots, f(X_n)\}$ as we did in the binary case.

B. Proof of Lemma 2

We first convert our discrete random vector X (taking values in $\{0,1\}^n$) to a continuous version $\widetilde{X}$ with an equivalent differential entropy. Define a new continuous random vector via $\widetilde{X} := X + U$, where U is a random vector independent of X,[2] with each element independently and identically distributed as $U_s \sim \mathcal{U}[-(1/2), 1/2]$.

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

Proof: Throughout this proof, we use $p(\cdot)$ to denote the probability density function of the continuous random vector $\widetilde{X}$ and $P(\cdot)$ to denote the probability mass function of the discrete random vector X. By definition [24], the differential entropy of $\widetilde{X}$ is given by the integral $h(\widetilde{X}) := -\int_{\mathcal{S}} p(\widetilde{x}) \log p(\widetilde{x}) d\widetilde{x}$, where $\mathcal{S} = \{\widetilde{x} \in \mathbb{R}^n | p(\widetilde{x}) > 0\}$ is the support of $\widetilde{X}$. By our construction of $\widetilde{X}$, the support set $\mathcal{S}$ can be decomposed into a disjoint union of hyperboxes $B(\mathbf{e})$ of unit volume, one for each configuration $\mathbf{e} \in \{0,1\}^n$. Using this decomposition, we write the differential entropy as

$$h(\widetilde{X}) = -\sum_{\mathbf{e} \in \{0,1\}^n} \int_{B(\mathbf{e})} p(\widetilde{x}) \log p(\widetilde{x}) d\widetilde{x}.$$

Now by our construction of $\widetilde{X}$, the quantity $p(\widetilde{x}) \log p(\widetilde{x})$ is equal to the constant $P(\mathbf{e}) \log P(\mathbf{e})$ for all $\widetilde{x}$ in the hyperbox $B(\mathbf{e})$, where $P(\mathbf{e})$ is the probability of the discrete configuration $\mathbf{e} \in \{0,1\}^n$. Accordingly, we have $h(\widetilde{X}) = -\sum_{\mathbf{e} \in \{0,1\}^n} P(\mathbf{e}) \log P(\mathbf{e}) [\int_{B(\mathbf{e})} 1 d\widetilde{x}]$, which is equal to $H(X)$, since the volume of each box $B(\mathbf{e})$ is unity. □

Now to establish Lemma 2, let $\mu \in \text{MARG}(K_n)$ and let X be a random vector with these marginals. Consider the continuous-valued random vector $\widetilde{X} := X + U$. From Lemma 3, we have $H(X) = h(\widetilde{X})$. Combining this equality with the Gaussian entropy bound (18) yields the upper bound

$$H(X) \leq \frac{1}{2} \log \det \text{cov}(\widetilde{X}) + \frac{n}{2} \log(2\pi e).$$

We now express the log-determinant quantity in an alternative form. First, using the independence of X and U, we write $\text{cov}(\widetilde{X}) = \text{cov}(X) + (1/12) I_n$, where we have used the fact that the covariance matrix of an independent uniform random vector U on a unit box is $(1/12) I_n$. Next we use the Schur complement formula [20] to express $\log \det \text{cov}(\widetilde{X})$ in terms of the second-order moment matrix $M_1[\mu]$ defined in (16) as follows:

$$\log \det \left[\text{cov}(X) + \frac{1}{12} I_n \right] = \log \det \left\{ M_1[\mu] + \frac{1}{12} \text{blkdiag}[0, I_n] \right\} \quad (22)$$

where $\text{blkdiag}[0, I_n]$ is an $(n+1) \times (n+1)$ block-diagonal matrix. Combining (22) with the Gaussian upper bound on $H(X) = -A^*(\mu)$ yields $-A^*(\mu) \leq (1/2) \log \det(M_1[\mu] + (1/12)\text{blkdiag}[0, I_n]) + (n/2) \log(2\pi e)$, which is the statement of Lemma 2.

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

C. Derivation of Dual Updates

To derive Newton updates for the Lagrangian dual of the weakened relaxation described in Section IV-D, it is more convenient to work with the same relaxation, but as applied to "spin" variables $V \in \{-1, +1\}^n$. The interaction among these spin variables can be captured by a Markov random field of the form $p(v; \gamma) \propto \exp\{\sum_{s \in V} \gamma_s v_s + \sum_{(s,t) \in E} \gamma_{st} v_s v_t\}$. Let $\eta_s = \mathbb{E}[v_s]$ and $\eta_{st} = \mathbb{E}[v_s v_t]$ be the associated moments of the spin vector. Let $M_1[\eta]$ denote the second-order moment matrix associated with V; it has the form of the matrix in (16), but with diagonal elements are all equal to one, since $v_s^2 = 1$ for spin variables in $\{-1, +1\}$. In this spin representation, we convert to a continuous version $\widetilde{V} := (1/2)V + U$, where again U is uniformly distributed on $[-(1/2), + (1/2)]$. (The rescaling by 1/2 is necessary to adjust the impulses to be within distance one of each other.) The log-determinant of the covariance of $\widetilde{V}$ takes the form $\log \det\{M_1[\eta] + (1/3)\text{blkdiag}[0, I_n]\} + n \log(1/4)$. Introducing the matrix-variable $Y := M_1[\eta] + (1/3)\text{blkdiag}[0, I_n]$, the weakened semidefinite relaxation corresponds to $\max_{Y \succeq 0} \{\langle -A, Y \rangle + \log \det Y\}$ such that $\text{diag}(Y) = d$, where $d = [1\ 4/3\ \ldots\ 4/3]^T$, A is an $(n+1) \times (n+1)$ matrix involving the weight vector γ, and $\langle A, B \rangle := \text{trace}(AB)$ is the Frobenius inner product. (As a sidenote, note that we have dropped a multiplicative factor of 1/2, as well as the additive constants in this form of the cost function. Moreover, the motivation for introducing a negative sign in A will be clear momentarily.)

Let $\lambda \in \mathbb{R}^{n+1}$ be a vector of Lagrange multipliers associated with the linear constraints $\text{diag}(Y) = d$. Computing the (Lagrangian) dual function of the weakened relaxation yields

$$Q(\lambda) = -(n+1) - \log \det [A + \text{diag}(\lambda)] + \langle \text{diag}(\lambda), \text{diag}(d) \rangle. \quad (23)$$

Thus, the dual problem corresponds to optimizing a convex and continuously differentiable function over $\mathbb{R}^{n+1}$, which can be performed very efficiently by Newton's method [27]. In particular, using well-known properties of matrix derivatives, we can compute the gradient $\nabla Q(\lambda) = -\text{diag}[A + \text{diag}(\lambda)]^{-1} + d$ and Hessian $\nabla^2 Q(\lambda) = -([A + \text{diag}(\lambda)]^{-1}) \odot ([A + \text{diag}(\lambda)]^{-1})$, where $\odot$ denotes Hadamard product. Thus, we can apply Newton's method or other scaled gradient methods to solve the dual problem. Given the optimal dual solution λ^*, we obtain the optimal primal solution as $Y^* = (A + \text{diag}(\lambda))^{-1}$. Finally, the optimal moment matrix $M_1[\eta^*]$ is given by $M_1[\eta^*] = Y^* - (1/3)\text{blkdiag}[0, I_n]$.

In the absence of any particular graph structure, the overall computational complexity of full Newton updates is $\mathcal{O}(n^3)$, which arises from the inversion of $(n+1) \times (n+1)$ of matrices. If we restrict our attention to a grid-structured graphical model, then the matrix $[A + \text{diag}(\lambda)]$ will be sparse and grid-structured, and thus can be inverted with complexity $\mathcal{O}(n^{1.5})$ using nested dissection [30]. Thus, we can perform diagonally scaled gradient descent with a complexity of $\mathcal{O}(n^{1.5})$ on grid-structured problems.

ACKNOWLEDGMENT

The authors would like to thank S. Boyd, C. Caramanis, L. El Ghaoui, and L. Vandenberghe for helpful discussions.

REFERENCES

[1] H. Brehm and W. Stammler, "Description and generation of spherically invariant speech-model signals," *Signal Process.*, vol. 12, no. 2, pp. 119–141, Mar. 1987.

[2] R. F. Engle, "Autoregressive conditional heteroscedasticity with estimates of the variance of UK inflation," *Econometrica*, vol. 50, no. 4, pp. 987–1008, Jul. 1982.

[3] D. J. Field, "Relations between the statistics of natural images and the response properties of cortical cells," *J. Opt. Soc. Amer. A*, vol. 4, no. 12, pp. 2379–2394, Dec. 1987.

[4] E. P. Simoncelli, "Statistical models for images: Compression, restoration and synthesis," in *IEEE 31st Asilomar Conf. Signals, Systems, Computers*, Nov. 1997, pp. 673–678.

[5] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, pp. 1396–1458, Aug. 2002.

[6] H. A. Loeliger, "An introduction to factor graphs," *IEEE Signal Process. Mag.*, vol. 21, no. 1, pp. 28–41, Jan. 2004.

[7] J. Pearl, *Probabilistic Reasoning in Intelligent Systems*. San Mateo, CA: Morgan Kaufman, 1988.

[8] M. Reyes, B. Raj, and D. Ellis, "Multi-channel source separation by factorial HMMS," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 1, Apr. 2003, pp. 664–667.

[9] M. Crouse, R. Nowak, and R. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Process.*, vol. 46, no. 4, pp. 886–902, Apr. 1998.

[10] L. R. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, pp. 257–285, Feb. 1989.

[11] G. Fan and X.-G. Xia, "Improved hidden Markov models in the wavelet-domain," *IEEE Trans. Signal Process.*, vol. 49, no. 1, pp. 115–120, Jan. 2001.

[12] J.-B. Durand, P. Gonçalvès, and Y. Guédon, "Computational methods for hidden Markov tree models—An application to wavelet trees," *IEEE Trans. Signal Process.*, vol. 52, no. 9, pp. 2551–2560, Sep. 2004.

[13] F. Kschischang, B. Frey, and H.-A. Loeliger, "Factor graphs and the sum-product algorithm," *IEEE Trans. Inf. Theory*, vol. 47, pp. 498–519, Feb. 2001.

[14] E. Sudderth, M. J. Wainwright, and A. S. Willsky, "Embedded trees: Estimation of Gaussian processes on graphs with cycles," *IEEE Trans. Signal Process.*, vol. 52, no. 11, pp. 3136–3150, Nov. 2004.

[15] K. Murphy, "Dynamic Bayesian networks: Representation, inference and learning," Ph.D. dissertation, Computer Science Division, Univ. of California Berkeley, May 2002.

[16] S. L. Lauritzen, *Graphical Models*. Oxford, U.K.: Oxford Univ. Press, 1996.

[17] R. Koetter, B. J. Frey, N. Petrovic, and J. D. C. Munson, "Unwrapping phase images by propagating probabilities across graphs," in *Proc. Int. Conf. Acoustics, Speech, Signal Processing (ICASSP)*, vol. 3, May 2001, pp. 1845–1848.

[18] J. Yedidia, W. T. Freeman, and Y. Weiss, "Constructing free energy approximations and generalized belief propagation algorithms," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2282–2312, Jul. 2005, to be published.

[19] G. Rockafellar, *Convex Analysis*. Princeton, NJ: Princeton Univ. Press, 1970.

[20] L. Vandenberghe, S. Boyd, and S. Wu, "Determinant maximization with linear matrix inequality constraints," *SIAM J. Matrix Anal. Applicat.*, vol. 19, no. 2, pp. 499–533, Mar. 1998.

[21] R. G. Cowell, A. P. Dawid, S. L. Lauritzen, and D. J. Spiegelhalter, *Probabilistic Networks and Expert Systems*, ser. Statistics for Engineering and Information Science. Berlin, Germany: Springer-Verlag, 1999.

[22] M. Deza and M. Laurent, *Geometry of Cuts and Metric Embeddings*. New York: Springer-Verlag, 1997.

[23] M. J. Wainwright and M. I. Jordan, "Graphical models, exponential families, and variational inference," Dept. of Statistics, Univ. of California Berkeley, Tech. Rep. 649, Sep. 2003.

[24] T. Cover and J. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.

[25] J. B. Lasserre, "Global optimization with polynomials and the problem of moments," *SIAM J. Optim.*, vol. 11, no. 3, pp. 796–817, Jul. 2001.

[26] S. Tatikonda and M. I. Jordan, "Loopy belief propagation and Gibbs measures," in *Proc. 18th Annu. Conf. Uncertainty Artif. Intell. (UAI-02)*, vol. 18, San Francisco, CA, Aug. 2002, pp. 493–500.

[27] D. Bertsekas, *Nonlinear Programming*. Belmont, MA: Athena Scientific, 1995.

[28] T. Heskes, K. Albers, and B. Kappen, "Approximate inference and constrained optimization," in *Proc. 19th Annu. Conf. Uncertainty Artificial Intelligence (UAI-03)*, San Francisco, CA, 2003, pp. 313–320.

[29] M. J. Wainwright, T. S. JaakkoJa, and A. S. Willsky, "A new class of upper bounds on the log partition function," *IEEE Trans. Inf. Theory*, vol. 51, no. 7, pp. 2313–2335, Jul. 2005.

[30] A. George and J. W.-H. Liu, Eds., *Computer Solution of Large Sparse Positive Definite Systems*. Englewood Cliffs, NJ: Prentice-Hall, 1981.

Martin J. Wainwright received the Bachelor's degree in mathematics from the University of Waterloo, Waterloo, ON, Canada, and the Ph.D. degree in electrical engineering and computer science from the Massachusetts Institute of Technology (MIT), Cambridge.

He is an Assistant Professor at the University of California at Berkeley, with a joint appointment between the Department of Statistics and the Department of Electrical Engineering and Computer Sciences. His research interests include statistical signal and image processing, error-control coding and data compression, and machine learning.

Prof. Wainwright received the George M. Sprowls Award (2002) from the MIT Department of Electrical Engineering and Computer Sciences for his doctoral dissertation and an Alfred P. Sloan Foundation Fellowship (2005).

Michael I. Jordan (M'99–SM'99–F'05) received the Master's degree from Arizona State University, Tempe, and the Ph.D. degree from the University of California, San Diego.

He is a Professor in the Department of Electrical Engineering and Computer Science and the Department of Statistics, University of California at Berkeley. He was a Professor at the Massachusetts Institute of Technology for 11 years. He has published more than 200 research papers on topics in computer science, electrical engineering, statistics, computational biology, and cognitive science. His research in recent years has focused on probabilistic graphical models, kernel machines, and applications of statistical machine learning to problems in bioinformatics, information retrieval, and signal processing.