CONVEX SURROGATES AND STABLE MESSAGE-PASSING: JOINT PARAMETER ESTIMATION AND PREDICTION IN COUPLED GAUSSIAN MIXTURE MODELS

Martin J. Wainwright

Departments of EECS and Statistics, UC Berkeley
Berkeley, CA, 94720
Email: wainwrig@eecs.berkeley.edu

ABSTRACT

The coupled mixture of Gaussian (MoG) model is a graphical model useful for various applications in signal processing. The parameter estimation and prediction problems, though tractable for tree-structured graphs, are intractable when the local mixture models are coupled together with a more complex graph with cycles. We present a joint approach to parameter estimation and prediction/smoothing problems in a coupled MoG model for an arbitrary graph with cycles. Our method exploits a convex surrogate to the cumulant generating function, for which both the parameter estimation and prediction steps can be solved efficiently by a tree-reweighted sum-product algorithm. We prove that our methods are globally Lipschitz stable, and provide bounds on the increase in MSE relative to the (unattainable) Bayes optimum. We also present the results of experimental simulations that both confirm these theoretical results, and show that our method outperforms the analogous method based on the ordinary sum-product algorithm.

I. INTRODUCTION

Graphical models provide a flexible framework for capturing statistical dependencies among random variables, and are widely-used in signal processing [e.g., 1], [2]. Many tasks in statistical signal processing (e.g., classification, prediction) can be formulated as computing likelihoods and/or marginals in a graphical model. The focus of this paper is a particular type of graphical model, which we refer to as a *coupled mixture of Gaussian (MoG) model*, consisting of a collection of local Gaussian mixture models that are coupled together by a graph. Mixtures of Gaussians are well-suited to capturing the statistics of various signal classes, including natural images, speech signals, and financial time series. The coupled MoG model that we consider has been used in signal processing frameworks based on wavelets [3], for which it is natural to couple together local mixture models using the multiresolution tree associated with the wavelet transform. An attractive feature of trees is the existence of efficient dynamic-programming algorithms for computing likelihoods and marginals [4], [2]. On the other hand, tree-structured models are well-known to cause boundary artifacts in signal reconstruction, caused by pairs of nodes that are spatially close but separated by a large graph distance in the tree. Perhaps the most natural remedy is to add extra edges to the tree, so that spatially adjacent nodes are also close in graph distance. Adding such edges leads to a more general graphical model with cycles. Computing likelihoods and marginals for such models, in contrast to the tree case, is a very challenging problem. Accordingly, the goal of the current research is develop efficient algorithms both for estimating parameters and for computing approximations to marginal probabilities for such graphical models with cycles.

The belief propagation or sum-product algorithm [4], [5] is widely used to compute approximate marginal probabilities in graphical models with cycles. In previous work [6], [7], we have introduced a family of tree-reweighted sum-product algorithms, and described preliminary results on their use for approximate parameter estimation. Other researchers [e.g., 8] have explored the use of unweighted belief propagation algorithms for approximate parameter estimation. In this paper, we further develop the tree-reweighted approach for joint parameter estimation and prediction. In particular, we prove that tree-reweighted algorithms satisfy a certain Lipschitz stability condition, and for coupled MoG models, we provide theoretical bounds on the performance loss relative to Bayes optimum. Due to space constraints, this paper provides only a relatively high-level description of our methods and main results; a complete version with proofs can be found in the technical report [9].

II. BACKGROUND

Model: We focus on a *coupled mixture of Gaussians (MoG)* model, consisting of a vector Z of finite Gaussian mixtures, where the mixture components for each scalar mixture variable Z_s are indexed by a discrete random variable X_s. The random variable X_s takes values in

the set $\mathcal{X} := \{0, 1, \ldots, m-1\}$, where m corresponds to the number of mixture components. We assume that the mixture indicator vector $X := \{X_s \mid s \in V\}$ can be described as a Markov random field over a graph $G = (V, E)$; see Figure 1 for a particular illustration on a grid-structured graph. In particular, letting $\mathbb{I}_j[x_s]$ be an indicator for the event $\{x_s = j\}$, we define functions $\theta_s(x_s) := \sum_{j \in \mathcal{X}} \theta_{s;j} \mathbb{I}_j[x_s]$, and $\theta_{st}(x_s, x_t) := \sum_{(j,k) \in \mathcal{X} \times \mathcal{X}} \theta_{st;jk} \mathbb{I}_j[x_s] \mathbb{I}_k[x_t]$, associated with nodes and edges of the graph, respectively. With this notation, the distribution of X decomposes on the graph G in the form

$$p(x; \theta) = \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) - A(\theta) \Big\}, \tag{1}$$

where

$$A(\theta) := \log \sum_{x \in \mathcal{X}^N} \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\} \tag{2}$$

is the log normalization constant or cumulant generating function.

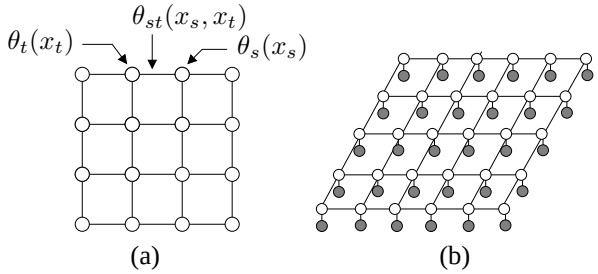

Fig. 1. (a) A Markov random field, used to capture the statistical interaction among the mixture indicator variables X_s. (b) Structure of joint model (X, Z): grid model for indicator variables (unshaded nodes) is combined with Gaussian mixture variables Z_s (shaded nodes).

Overall, the joint distribution over mixture components X and Gaussian mixture vector Z takes the form

$$p(x, z; \theta, \nu, \sigma) \quad \propto \quad \prod_{s=1}^{n} p(z_s \,|\, x_s; \nu_s, \sigma_s) \;\; p(x; \theta) \tag{3}$$

where for each $j \in \mathcal{X}$, the conditional $p(z_s \,|\, X_s = j; \nu_j, \sigma_j)$ is a Gaussian distribution with mean ν_j and variance σ_j^2. (For simplicity, we take the Gaussian means and variances to be identical for all nodes.)

Given a coupled MoG model specified by a graph with cycles, our goal is to develop approximate methods for solving the following two problems: (a) estimating model parameters from data; and (b) using the model to make predictions on the basis of noisy observations. Both of these problems are central to many signal processing applications of the coupled MoG model.

Parameter estimation: Suppose that we are given i.i.d. observations $(x^1, z^1), \ldots, (x^n, z^n)$. Since we are assuming that the mixture indicators x^i are given, estimating the Gaussian means ν_j and variances σ_j^2 is straightforward, using the classical sample mean and sample variance respectively. Accordingly, our main focus will be estimating the functions θ_s and θ_{st} that define the Markov random field over the mixture indicator vector X. The maximum likelihood (ML) estimate of θ is given by maximizing the log likelihood $L(\theta) = \frac{1}{n} \sum_{i=1}^{n} \log p(x^i; \theta)$ Equivalently, the log likelihood can be written in terms of the *empirical marginals* $\widehat{\mu}_s$ and $\widehat{\mu}_{st}$ determined by the data $x^1, \ldots, x^n$ as $L(\theta) = \langle \widehat{\mu}, \theta \rangle - A(\theta)$, where A is the cumulant generating function (2). Since it is straightforward to compute the empirical marginals $\widehat{\mu}_s$ and $\widehat{\mu}_{st}$, the challenge in ML estimation lies in computing the cumulant generating function.

Prediction: The goal of prediction is to optimally estimate the random vector Z based on noisy observations

$$Y \quad = \quad \alpha Z + \sqrt{1 - \alpha^2}\, W, \tag{4}$$

where $W \sim \mathcal{N}(0, I)$ is additive white Gaussian noise, and the parameter $\alpha \in [0, 1]$ specifies the signal-to-noise (SNR) ratio. The Bayes least square estimate of Z given the observation $Y = y$ is given by the conditional mean $\widehat{z}(y) = \mathbb{E}[Z \,|\, Y = y]$. In the coupled MoG model and under observation model (4), it is straightforward to compute the following expression for component s of this conditional mean:

$$\widehat{z}_s(y) = \sum_{j \in \mathcal{X}_s} p(x_s = j; \theta, y) \Big[\omega_j (y_s - \nu_j) + \nu_j \Big], \tag{5}$$

where $\omega_j(\alpha) := \frac{\alpha \sigma_j^2}{\alpha^2 \sigma_j^2 + (1 - \alpha^2)}$. This expression shows that the conditional mean for the Gaussian mixture z is a combination of linear least-squares estimators (LLSEs) for each Gaussian component, where component j is weighted by the conditional probability $p(X_s = j \,|\, y)$. Consequently, the challenge in performing optimal prediction lies in computing the vector $\mu_s(j; \theta) := p(X_s = j \,|\, y; \theta)$ of conditional probabilities. It is well-known that computing these probabilities is equivalent to computing a derivative of a cumulant generating function A of the form (2).

The preceding discussion highlights that the central object in both the parameter estimation and prediction problems is the cumulant generating function A, which is intractable to compute exactly. Accordingly, our approach is based on constructing a new function C_ρ that acts as a *convex surrogate* to the cumulant generating function A. The cumulant generating function A can be represented as the solution of a certain optimization problem, which in turn motivates different ways in which to construct approximations to A; in the following section, we provide

background on a particular convex surrogate, based on a convexified Bethe entropy, that we have described in previous work [7], [6].

III. PROPOSED APPROACH

The sum-product or belief propagation algorithm is a widely-used iterative algorithm in which nodes in the graph perform local computations, and exchange statistical information by passing messages [1], [4]. An important fact, as shown by Yedidia et al. [5], is that the sum-product algorithm can be derived as a Lagrangian method for solving the Bethe variational problem (BVP). The optimization variables in the BVP can be interpreted as *pseudomarginal* distributions τ_s and τ_{st} associated with the nodes and edges of the graph; they must belong to the polyhedral constraint set $\mathrm{LOCAL}(G)$ given by

$$\{\tau \in \mathbb{R}^d \mid \sum_{x_s} \tau_s(x_s) = 1, \quad \sum_{x'_t} \tau_{st}(x_s, x'_t) = \tau_s(x_s)\}.$$

These constraints ensure that the functions $\tau_s(\cdot)$ and τ_{st} are locally valid marginal distributions. Consequently, we can define the singleton entropy and mutual information

$$H_s(\tau_s) \quad := \quad -\sum_{x_s \in \mathcal{X}_s} \tau_s(x_s) \log \tau_s(x_s), \tag{6a}$$

$$I_{st}(\tau_{st}) \quad := \quad \sum_{x_s, x_t} \tau_{st}(x_s, x_t) \log \frac{\tau_{st}(x_s, x_t)}{\tau_s(x_s)\tau_t(x_t)}. \tag{6b}$$

In the context of the ordinary sum-product algorithm, these terms are used to define the so-called Bethe entropy. In previous work [7], we have defined a family of "convexified" Bethe entropies of the form

$$H_{Bethe}(\tau; \rho) := \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}). \tag{7}$$

To ensure convexity, the vector $\boldsymbol{\rho} := \{\rho_{st}, \ (s,t) \in E\}$ of edge weights must arise from some convex combination of spanning trees on the original graph [7], which we refer to as a *valid setting*. For example, the setting $\rho_{st} = \frac{1}{2}$ is one such valid choice for a 2-D grid with toroidal boundary conditions. The set of all valid settings of $\boldsymbol{\rho}$ is known as the spanning tree polytope, denoted by $\mathbb{T}(G)$. As a side comment, the usual Bethe entropy [5] corresponds to the special case of setting $\rho_{st} = 1$ for all edges $(s,t) \in E$; such a choice belongs to the spanning tree polytope only if G is actually a tree. Herein we restrict our attention to valid settings $\boldsymbol{\rho} \in \mathbb{T}(G)$.

With this set-up, we consider the function C_ρ defined as the optimal value of the following convexified Bethe variational problem:

$$C_\rho(\theta) \quad := \quad \max_{\tau \in \mathrm{LOCAL}(G)} \big\{ \langle \tau, \, \theta \rangle + H_{Bethe}(\tau; \boldsymbol{\rho}) \big\} \tag{11}$$

For any valid setting $\boldsymbol{\rho} > 0$, this optimization problem has a unique optimum.

Tree-reweighted message-passing: As with the ordinary Bethe problem and the sum-product algorithm, the optimization problem on the RHS of equation (11) can be solved efficiently using the *tree-reweighted sum-product* algorithm [7]. As with the ordinary sum-product algorithm, each node s performs local computations and relays a message vector M_{st} to each of its neighbors $t \in N(s)$. For a pairwise MRF of the form (1), the messages are updated as shown in equation (9). The case of standard belief propagation corresponds to the special case of setting all $\rho_{st} = 1$, but this is not a valid setting on a graph with cycles. For any valid set of weights ρ_{st}, the updates (9) are guaranteed to have a unique fixed point for any choice of parameter vector θ. (This statement does not hold for standard belief propagation.)

Thus, the tree-reweighted sum-product algorithm can be used to compute both the value $C_\rho(\theta)$, and the vector $\tau(\theta)$ of *pseudomarginals* that achieve the maximum in equation (11). The function C_ρ can be viewed as a surrogate to the true cumulant generating function; in particular, it can be shown [7] that in analogy to A, it is differentiable. More concretely, its first-order derivatives are given by the optimizing pseudomarginals—namely $\tau_\alpha(\theta) = \frac{\partial C_\rho}{\partial \theta_\alpha}(\theta)$.

Parameter estimation: Our convex surrogate C_ρ forms the basis for the following approach to parameter estimation. Given a collection of i.i.d. samples $x^1, \ldots, x^n$ of the discrete mixture indicator variables, we form the empirical marginals at single nodes $\widehat{\mu}_s$ for $s \in V$ and $\widehat{\mu}_{st}$ for edges $(s,t) \in E$. We use these empirical marginals and the convex surrogate C_ρ to form the following surrogate likelihood function:

$$\widetilde{L}(\theta) \quad := \quad \langle \widehat{\mu}, \, \theta \rangle - C_\rho(\theta). \tag{12}$$

Maximizing this surrogate likelihood yields an parameter estimate $\widehat{\theta}$. When the surrogate C_ρ is formed from the convexified Bethe approximation (and there is no regularization), it can be shown [7] that the approximate parameter estimate $\widehat{\theta}$ has a very simple closed-form solution, specified in terms of the weights ρ_{st} and the empirical marginals $\widehat{\mu}$, as shown in Step 1 of Figure 2. (If a regularizing term is added, these estimates no longer have a closed-form solution, but can be computed efficiently with an iterative procedure.) Moreover, by construction of the surrogate likelihood (12), the approximate parameter estimate $\widehat{\theta}$ satisfies the *matching property*: if we run the tree-reweighted (TRW) sum-product algorithm on the problem $p(x; \widehat{\theta})$, then its unique fixed point is given by the pseudomarginals $\tau(\widehat{\theta}) = \widehat{\mu}$.

Algorithm for joint parameter estimation and prediction:

1) Estimate parameters $\widehat{\theta}$ from initial data $x^1, \ldots, x^n$ as follows:

$$\widehat{\theta}_s(x_s) = \log \widehat{\mu}_s(x_s), \qquad \widehat{\theta}_s(x_s) = \rho_{st} \log \frac{\widehat{\mu}_{st}(x_s, x_t)}{\widehat{\mu}_s(x_s)\widehat{\mu}_t(x_t)}. \tag{8}$$

2) Incorporate observation y into model: $\widetilde{\theta}_s(x_s) = \widehat{\theta}_s(x_s) + \log p(y_s \,|\, x_s)$.

3) Compute approximate marginals τ^{TRW} by tree-reweighted message-passing:

$$M_{st}(x_t) \leftarrow \sum_{x_s} \exp\big\{\widetilde{\theta}_s(x_s) + \frac{1}{\rho_{st}}\widehat{\theta}_{st}(x_s, x_t)\big\} \frac{\prod_{u \in N(s)\setminus t}[M_{us}(x_s)]^{\rho_{us}}}{[M_{ts}(x_s)]^{1-\rho_{ts}}}, \quad \tau_s(x_s) \propto \exp(\widetilde{\theta}_s(x_s)) \prod_{t \in N(s)} [M_{ts}(x_s)]^{\rho_{ts}}. \tag{9}$$

4) Construct prediction $\widehat{z}(y;\tau)$ of z based on the observation y and pseudomarginals τ:

$$\widehat{z}_s(y;\tau) = \sum_{j \in \mathcal{X}_s} \tau_s(j|y;\widetilde{\theta})\Big[\omega_j(y_s - \nu_j) + \nu_j\Big]. \tag{10}$$

Fig. 2. Sequence of steps performed for joint parameter estimation and prediction. (1) The parameters $\widehat{\theta}$ are estimated based on an initial set of i.i.d. data $x^1, \ldots, x^n$. (2) Given a new set of noisy observations y of z, they are incorporated into the model by modifying the potential functions $\widehat{\theta}_s(x_s)$ by addition of the term $\log p(y_s|x_s)$. (3) Approximate marginals are computed by TRW message-passing. (4) These approximate marginals are used as weights in the estimator of z.

Our overall joint approach to parameter estimation and prediction is detailed in Figure 2. In Step 1, we compute the parameter estimate $\widehat{\theta}$. Step 2 involves modifying the original parameter vector $\widehat{\theta} \to \widetilde{\theta}$ in order to account for new noisy observations. In Step 3, we apply the TRW algorithm to the modified problem $p(x;\widetilde{\theta})$ in order to compute a vector of approximate marginals $\tau(\widetilde{\theta})$. Finally, we use these pseudomarginals as weights in the combination of linear least-squares estimators given in Step 4.

IV. THEORETICAL GUARANTEES

On the theoretical side, we wish to understand the loss in performance when using our combined estimation-prediction method relative to the Bayes optimal predictor (5). This Bayes optimal method is unachievable because it has access to both the exact parameter value, and the exact marginals μ (which are intractable to compute exactly). In order to relate our method to this unachievable optimum, we need to understand the link between the approximate marginals τ and the true marginals μ.

Lipschitz stability of TRW message-passing: An important property of exact marginalization is that small changes to the model lead to correspondingly small changes to the marginals. The following result shows that the approximate marginals computed by the TRW algorithm obey an analog of this property:

Proposition 1. *For a given parameter $\theta \in \mathbb{R}^d$, let $\tau(\theta)$ denote the pseudomarginals computed by the TRW algorithm with a valid choice of weights ρ_{st}. Then there exists an absolute constant R such that $\|\tau(\theta+\delta) - \tau(\theta)\| \leq R\|\delta\|$ for all $\theta, \delta \in \mathbb{R}^d$.*

We refer the reader to the technical report [9] for the proof of this result. The stability ensured by this result is a global property, in that it holds for all parameter vectors $\theta \in \mathbb{R}^d$. In contrast, it should be noted that due to non-convexity of the Bethe variational problem, the ordinary sum-product algorithm does not satisfy an analogous claim. Indeed, in general, the algorithm has multiple fixed points, and its behavior can be highly unstable around phase transition points. The uniqueness of fixed points and Lipschitz stability of our message-passing algorithm plays an important role in the bounds that we establish in the following section.

Bounds on performance: We now turn to a comparison of the mean-squared error (MSE) of the Bayes optimal predictor $\widehat{z}(Y;\mu)$, defined in equation (5), to the MSE of the TRW-based predictor $\widehat{z}(Y;\tau)$ defined in equation (10). More specifically, we provide an upper bound on the increase in MSE, where the bound is specified in terms of the SNR parameter α defining the observation model (see equation (4)), as well as on the coupling strength.

Although results of this nature can be derived more generally, for simplicity in notation, we focus on the case of two mixture components ($m = 2$). Moreover, we consider the asymptotic setting, in which the number of data samples (used to perform the parameter estimation) tends to infinity, so that from the law of large numbers, the empirical marginals $\widehat{\mu}$ converge to the exact marginal

distributions μ^*. By standard results [10], the ML estimator (based on maximizing the true likelihood) converges to the true parameter value θ^*; moreover, it can be shown that our approximate parameter estimate (based on the surrogate likelihood) converges to a fixed quantity $\widetilde{\theta}$. By construction, we have the relations $\nabla C_\rho(\widetilde{\theta}) = \mu^* = \nabla A(\theta^*)$.

An important factor in our bound is the quantity

$$L(\theta^*; \widetilde{\theta}) := \sup_{\delta \in \mathbb{R}^d} \sigma_{\max}\big(\nabla^2 A(\theta^* + \delta) - \nabla^2 C_\rho(\widetilde{\theta} + \delta)\big),$$

where $\sigma_{\max}$ denotes the maximal singular value. Using Proposition 1 and the fact that $\nabla^2 A(\theta^* + \delta)$ is the covariance matrix of a multinomial random vector, it can be seen that $L(\theta^*; \widetilde{\theta})$ is finite. Two additional quantities that play a role in our bound are the differences

$$\Delta_\omega(\alpha) := \omega_1(\alpha) - \omega_0(\alpha), \quad \text{and} \tag{13a}$$

$$\Delta_\nu(\alpha) := [1 - \omega_1(\alpha)]\nu_1 - [1 - \omega_0(\alpha)]\nu_0, \tag{13b}$$

where the weights $\omega_j(\alpha)$ are defined below equation (5), and ν_0, ν_1 are the means of the two Gaussian components. Finally, we define $\gamma(Y; \alpha) \in \mathbb{R}^d$ with components $\log \frac{p(Y_s|X_s=1)}{p(Y_s|X_s=0)}$ for $s \in V$, and zeroes otherwise. With this notation, we have the following:

Theorem 1. *Let* $\mathrm{MSE}(\tau)$ *and* $\mathrm{MSE}(\mu)$ *denote the mean-squared prediction errors of the approximate TRW estimator* $\widehat{z}(y; \tau)$*, and the Bayes optimal estimator* $\widehat{z}(y; \mu)$ *respectively. The increase in MSE* $\mathcal{I}(\alpha) := \frac{1}{N}\big[\mathrm{MSE}(\tau) - \mathrm{MSE}(\mu)\big]$ *of the TRW method relative to Bayes optimal is upper bounded as*

$$\mathcal{I}(\alpha) \leq \mathbb{E}\Bigg\{ \Omega^2(\alpha)\Delta_\nu^2(\alpha) + \Omega(\alpha)\Big[\Delta_\omega^2(\alpha)\sqrt{\frac{\sum_s Y_s^4}{N}} + 2|\Delta_\nu(\alpha)|\,|\Delta_\omega(\alpha)|\sqrt{\frac{\sum_s Y_s^2}{N}}\Big]\Bigg\} \tag{14}$$

where $\Omega(\alpha) := \min\{1, L(\theta^*; \widetilde{\theta})\|\frac{\gamma(Y;\alpha)}{N}\|\}$.

A proof of this result can be found in the technical report [9]. Since $\gamma(Y; \alpha)$ (and hence $\Omega(\alpha)$) both converge to 0 as $\alpha \to 0^+$, the bound shows that $\mathcal{I}(\alpha) \to 0$ as $\alpha \to 0^+$, so that the TRW method is asymptotically optimal for low SNR. At the other end, it can be seen that the functions $\Delta_\nu(\alpha)$ and $\Delta_\omega(\alpha)$ both converge to zero as $\alpha \to 1-$, which guarantees asymptotic optimality in the high SNR regime. The behavior of the bound in the intermediate regime is controlled by the balance between these two terms.

V. EXPERIMENTAL RESULTS

In order to test our joint estimation/prediction procedure, we have applied it to coupled MoG models on various types of graphs, a range of coupling strengths, observation SNRs, and different types of Gaussian mixtures. Although our methods are more generally applicable, here we show representative results for $m = 2$ mixture components. We consider two different mixture types: ensemble (a) has mean and variance components $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$, whereas ensemble (b) has components $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_1, \sigma_1^2) = (0, 9)$.

Here we show results for a a coupled MoG model defined on a 2-D grid with $N = 64$ nodes. Since the mixture variables have $m = 2$ states, the coupling distribution (1) can be written as $p(x; \theta) \propto \exp\big\{\sum_{s \in V} \theta_s x_s + \sum_{(s,t) \in E} \theta_{st} x_s x_t\big\}$. where $x \in \{-1, +1\}^N$ are spin variables indexing the mixture components. In all trials, we chose $\theta_s = 0$ for all nodes $s \in V$, which ensures uniform marginal distributions $p(x_s; \theta)$ at each node. For each coupling strength $\gamma \in [0, 1]$, we chose edge parameters as $\theta_{st} \sim \mathcal{U}[0, \gamma]$, and we varied the SNR parameter α controlling the observation model (4) in $[0, 1]$. We evaluated the following three methods based on their increase in mean-squared error (MSE) over the Bayes optimal predictor (5): (a) As a baseline, we used the *independence model* for the mixture components: parameters are estimated $\theta_s(x_s) = \log \widehat{\mu}_s(x_s)$, and setting coupling terms $\theta_{st}(x_s, x_t)$ equal to zero. The prediction step reduces to performing LLSE at each node independently. (b) The *standard belief propagation* (BP) approach is based on estimating parameters (see step (1) of Figure 2) using $\rho_{st} = 1$ for all edges (s, t), and using BP to perform the prediction. (c) The *tree-reweighted method* (TRW) is based on estimating parameters using $\rho_{st} = \frac{1}{2}$ for all edges (s, t), and using the associated TRW algorithm for prediction.

Figure 3 shows 2-D surface plots of the percentage increase in MSE as a function of the coupling strength $\gamma \in [0, 1]$ and the observation SNR parameter $\alpha \in [0, 1]$ for the independence model (left column), BP approach (middle column) and TRW method (right column). The qualitative patterns of results are similar for both mixture ensembles (a) and (b), as shown in the top and bottom rows respectively of Figure 3. Note that for relatively low coupling strengths γ, all three methods—including the independence model—perform quite well. This behavior is to be expected, since there is little to be gained from exchanging information among nodes when the dependencies are weak. Although not obvious in these plots, the performance of BP is better than TRW for weak coupling; however, both methods incur a performance loss less than 1% in this regime. As the coupling strength is increased, the performance of the BP method eventually deteriorates quite seriously; indeed, for large enough coupling and low/intermediate SNR, its performance can be worse than the naive independence model. In contrast, the behavior of the TRW method is extremely stable. More precisely,

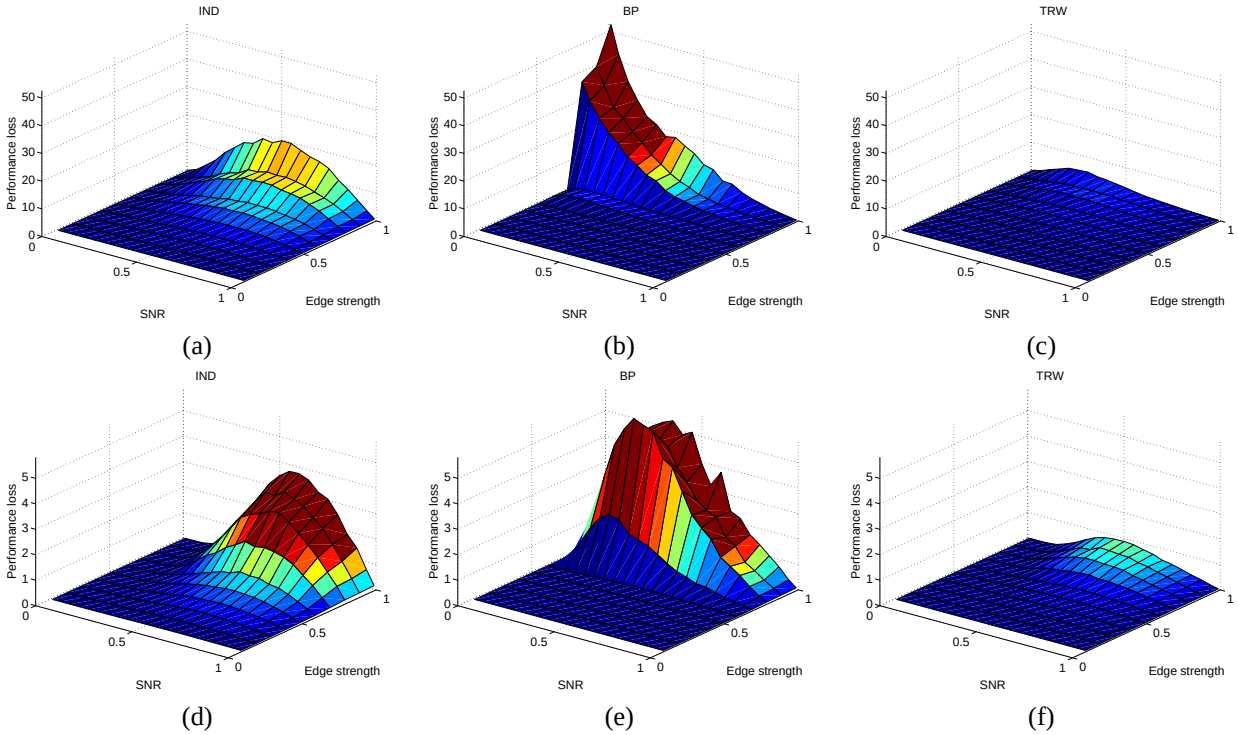

Fig. 3. Surface plots of the percentage increase in MSE relative to Bayes optimum for different methods as a function of observation SNR and coupling strength. Top row: Gaussian mixture with components $(\nu_0, \sigma_0^2) = (-1, 0.5)$ $(\nu_1, \sigma_1^2) = (1, 0.5)$. Bottom row: Gaussian mixture with components $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_0, \sigma_1^2) = (0, 9)$. Left column: independence model (IND). Center column: ordinary belief propagation (BP). Right column: tree-reweighted algorithm (TRW).

the performance loss in MSE for the TRW method relative to the unattainable Bayes optimum remains less than 5% for ensemble (a) (respectively 2% for ensemble (b)) over the entire range of coupling and SNR. This empirical demonstration of stability is consistent with our theoretical results.

VI. CONCLUSION

We have described a combined method for parameter estimation and prediction/smoothing based on a convex surrogate to the cumulant generating function. Both the estimation and prediction steps can be solved efficiently by tree-reweighted message-passing in the underlying graphical model. The combined method, when applied to coupled mixture of Gaussian model, yields prediction results close to the Bayes optimum, as shown by both theoretical analysis and experimental simulations, and outperforms an analogous method based on standard belief propagation. Our current set-up and analysis has focused on the case of fully observed data; it remains to explore extensions of these ideas to the partially observed case.

VII. REFERENCES

[1] H. A. Loeliger, "An introduction to factor graphs," *IEEE Signal Proc. Mag.*, vol. 21, pp. 28–41, 2004.

[2] A. S. Willsky, "Multiresolution Markov models for signal and image processing," *Proc. IEEE*, vol. 90, no. 8, pp. 1396–1458, 2002.

[3] M.S. Crouse, R.D. Nowak, and R.G. Baraniuk, "Wavelet-based statistical signal processing using hidden Markov models," *IEEE Trans. Signal Proc.*, vol. 46, pp. 886–902, April 1998.

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

[5] J. S. Yedidia, W. T. Freeman, and Y. Weiss, "Generalized belief propagation," in *NIPS 13*. 2001, pp. 689–695, MIT Press.

[6] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "A new class of upper bounds on the log partition function," *IEEE Trans. Info. Theory*, To appear in July 2005.

[7] M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky, "Tree-reweighted belief propagation algorithms and approximate ML estimation by pseudomoment matching," in *Workshop on AI and Statistics*, January 2003.

[8] Y. W. Teh and M. Welling, "On improving the effi ciency of the iterative proportional fi tting procedure," in *Workshop on AI and Statistics*, 2003.

[9] M. J. Wainwright, "Stable message-passing and convex surrogates: Joint parameter estimation and prediction," Tech. Rep 690., Dept. of Statistics, UC Berkeley, May 2005.

[10] A. W. van der Vaart, *Asymptotic statistics*, Cambridge University Press, Cambridge, UK, 1998.