Estimating the "wrong" Markov random field: Benefits in the computation-limited setting

Martin J. Wainwright
Department of Statistics, and
Department of Electrical Engineering and Computer Science
UC Berkeley, Berkeley CA 94720
wainwrig@{stat,eecs}.berkeley.edu

Abstract

Consider the problem of joint parameter estimation and prediction in a Markov random field: i.e., the model parameters are estimated on the basis of an initial set of data, and then the fitted model is used to perform prediction (e.g., smoothing, denoising, interpolation) on a new noisy observation. Working in the computation-limited setting, we analyze a joint method in which the *same convex variational relaxation* is used to construct an M-estimator for fitting parameters, and to perform approximate marginalization for the prediction step. The key result of this paper is that in the computation-limited setting, using an inconsistent parameter estimator (i.e., an estimator that returns the "wrong" model even in the infinite data limit) is provably beneficial, since the resulting errors can partially compensate for errors made by using an approximate prediction technique. En route to this result, we analyze the asymptotic properties of M-estimators based on convex variational relaxations, and establish a Lipschitz stability property that holds for a broad class of variational methods. We show that joint estimation/prediction based on the reweighted sum-product algorithm substantially outperforms a commonly used heuristic based on ordinary sum-product. [1]

Keywords: Markov random fields; variational method; message-passing algorithms; sum-product; belief propagation; parameter estimation; learning.

1 Introduction

Consider the problem of joint learning (parameter estimation) and prediction in a Markov random field (MRF): in the learning phase, an initial collection of data is used to estimate parameters, and the fitted model is then used to perform prediction (e.g., smoothing, interpolation, denoising) on a new noisy observation. Disregarding computational cost, there exist optimal methods for solving this problem (Route A in Figure 1). For general MRFs, however, optimal methods are computationally intractable; consequently, many researchers have examined various types of message-passing methods for learning and prediction problems, including belief propagation [3, 6, 7, 14], expectation propagation [5], linear response [4], as well as reweighted message-passing algorithms [10, 13]. Accordingly, it is of considerable interest to understand and quantify the performance loss incurred

[1]Work partially supported by Intel Corporation Equipment Grant 22978, an Alfred P. Sloan Foundation Fellowship, and NSF Grant DMS-0528488.

by using computationally tractable methods versus exact methods (i.e., Route B versus A in Figure 1).

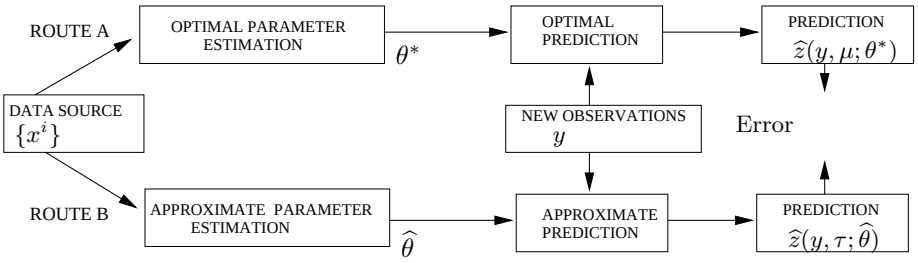

Figure 1. Route A: computationally intractable combination of parameter estimation and prediction. Route B: computationally efficient combination of approximate parameter estimation and prediction.

It is now well known that many message-passing algorithms—including mean field, (generalized) belief propagation, expectation propagation and various convex relaxations—can be understood from a variational perspective; in particular, all of these message-passing algorithms are iterative methods solving relaxed forms of an exact variational principle [12]. This paper focuses on the analysis of variational methods based *convex relaxations*, which includes a broad range of extant algorithms—among them the tree-reweighted sum-product algorithm [11], reweighted forms of generalized belief propagation [13], and semidefinite relaxations [12]. Moreover, it is straightforward to modify other message-passing methods (e.g., expectation propagation [5]) so as to "convexify" them. At a high level, the key idea of this paper is the following: given that approximate methods can lead to errors at both the estimation and prediction phases, it is natural to speculate that these sources of error might be arranged to partially cancel one another. Our theoretical analysis confirms this intuition: we show that with respect to end-to-end performance, it is in fact beneficial, even in the infinite data limit, to learn the "wrong" the model by using an *inconsistent* parameter estimator.

More specifically, we show how any convex variational method can be used to define a surrogate likelihood function. We then investigate the asymptotic properties of parameter estimators based maximizing such surrogate likelihoods, and establish that they are asymptotically normal but inconsistent in general. We then prove that any variational method that is based on a strongly concave entropy approximation is globally Lipschitz stable. Finally, focusing on prediction for a coupled mixture of Gaussians, we prove upper bounds on the increase in MSE of our computationally efficient method, relative to the unachievable Bayes optimum. We provide experimental results using the tree-reweighted (TRW) sum-product algorithm that confirm the stability of our methods, and demonstrate its superior performance to a heuristic method based on standard sum-product.

2 Background

We begin with necessary notation and background on multinomial Markov random fields, as well as variational representations and methods.

Markov random fields: Given an undirected graph $G = (V, E)$ with $N = |V|$ vertices, we associate to each vertex $s \in V$ a discrete random variable X_s, taking values in $\mathcal{X}_s = \{0, 1 \ldots, m-1\}$. We assume that the vector $X = \{X_s \,|\, s \in V\}$ has a distribution that is

Markov with respect to the graph G, so that its distribution can be represented in the form

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

Here $A(\theta) := \log\sum_{x\in\mathcal{X}^N}\exp\big\{\sum_{s\in V}\theta_s(x_s) + \sum_{(s,t)\in E}\theta_{st}(x_s,x_t)\big\}$ is the *cumulant generating function* that normalizes the distribution, and $\theta_s(\cdot)$ and $\theta_{st}(\cdot,\cdot)$ are potential functions. In particular, we make use of the parameterization $\theta_s(x_s) := \sum_{j\in\mathcal{X}_s}\theta_{s;j}\mathbb{I}_j[x_s]$, where $\mathbb{I}_j[x_s]$ is an indicator function for the event $\{x_s = j\}$; the quantity θ_{st} is defined analogously. Overall, the family of MRFs (1) is an exponential family with canonical parameter $\theta \in \mathbb{R}^d$. Note that the elements of the canonical parameters are associated with vertices $\{\theta_{s;j},\ s\in V, j\in\mathcal{X}_s\}$ and edges $\{\theta_{st;jk},\ (s,t)\in E,\ (j,k)\in\mathcal{X}_s\times\mathcal{X}_t\}$ of the underlying graph.

Variational representation: We now describe how the cumulant generating function can be represented as the solution of an optimization problem. The constraint set is given by $\mathrm{MARG}(G;\phi) := \big\{\mu\in\mathbb{R}^d \mid \mu = \sum_{x\in\mathcal{X}^N}p(x)\phi(x) \text{ for some } p(\cdot)\big\}$, consisting of all globally realizable singleton $\mu_s(\cdot)$ and pairwise $\mu_{st}(\cdot\,,\cdot)$ marginal distributions on the graph G. For any $\mu\in\mathrm{MARG}(G;\phi)$, we define $A^*(\mu) = -\max_p H(p)$, where the maximum is taken over all distributions that have mean parameters μ. With these definitions, it can be shown [12] that A has the variational representation

$$A(\theta) \quad = \quad \max_{\mu\in\mathrm{MARG}(G;\phi)}\big\{\theta^T\mu - A^*(\mu)\big\}. \tag{2}$$

3 From convex surrogates to joint estimation/prediction

In general, solving the variational problem (2) is intractable for two reasons: (i) the constraint set $\mathrm{MARG}(G;\phi)$ is extremely difficult to characterize; and (ii) the dual function A^* lacks a closed-form representation. These challenges motivate approximations to A^* and $\mathrm{MARG}(G;\phi)$; the resulting relaxed optimization problem defines a convex surrogate to the cumulant generating function.

Convex surrogates: Let $\mathrm{REL}(G;\phi)$ be a compact and convex outer bound to the marginal polytope $\mathrm{MARG}(G;\phi)$, and let B^* be a strictly convex and twice continuously differentiable approximation to the dual function A^*. We use these approximations to define a convex surrogate B via the relaxed optimization problem

$$B(\theta) \quad := \quad \max_{\tau\in\mathrm{REL}(G;\phi)}\big\{\theta^T\tau - B^*(\tau)\big\}. \tag{3}$$

The function B so defined has several desirable properties. First, since B is defined by the maximum of a collection of functions linear in θ, it is convex [1]. Moreover, by the strict convexity of B^* and compactness of $\mathrm{REL}(G;\phi)$, the optimum is uniquely attained at some $\tau(\theta)$. Finally, an application of Danskin's theorem [1] yields that B is differentiable, and that $\nabla B(\theta) = \tau(\theta)$. Since $\tau(\theta)$ has a natural interpretation as a *pseudomarginal*, this last property of B is analogous to the well-known cumulant generating property of A—namely, $\nabla A(\theta) = \mu(\theta)$.

One example of such a convex surrogate is the tree-reweighted Bethe free energy considered in our previous work [11]. For this surrogate, the relaxed constraint set $\mathrm{REL}(G;\phi)$ takes the form $\mathrm{LOCAL}(G;\phi) := \big\{\tau\in\mathbb{R}_+^d \mid \sum_{x_s}\tau_s(x_s) = 1, \sum_{x_t}\tau_{st}(x_s,x_t) = \tau_s(x_s)\big\}$, whereas the entropy approximation B^* is of the "convexified" Bethe form

$$-B^*(\tau) \quad = \quad \sum_{s\in V}H_s(\tau_s) - \sum_{(s,t)\in E}\rho_{st}I_{st}(\tau_{st}). \tag{4}$$

Here H_s and I_{st} are the singleton entropy and edge-based mutual information, respectively, and the weights ρ_{st} are derived from the graph structure so as to ensure convexity (see [11] for more details). Analogous convex variational formulations underlie the reweighted generalized BP algorithm [13], as well as a log-determinant relaxation [12].

Approximate parameter estimation using surrogate likelihoods: Consider the problem of estimating the parameter θ using i.i.d. samples $\{x^1, \ldots, x^n\}$. For an MRF of the form (1), the maximum likelihood estimate (MLE) is specified using the vector $\widehat{\mu}$ of empirical marginal distributions (singleton $\widehat{\mu}_s$ and pairwise $\widehat{\mu}_{st}$). Since the likelihood is intractable to optimize (due to the cumulant generating function A), it is natural to use the convex surrogate B to define an alternative estimator obtained by maximizing the regularized *surrogate likelihood*:

$$\widehat{\theta}^n \quad := \quad \arg\max_{\theta \in \mathbb{R}^d} \left\{ L_B(\theta; \widehat{\mu}) - \lambda^n R(\theta) \right\} \; = \; \arg\max_{\theta \in \mathbb{R}^d} \left\{ \theta^T \widehat{\mu} - B(\theta) - \lambda^n R(\theta) \right\}. \quad (5)$$

Here $R : \mathbb{R}^d \to \mathbb{R}_+$ is a regularization function (e.g., $R(\theta) = \|\theta\|^2$), whereas $\lambda^n > 0$ is a regularization coefficient. For the tree-reweighted Bethe surrogate, we have shown in previous work [10] that in the absence of regularization, the optimal parameter estimates $\widehat{\theta}^n$ have a very simple closed-form solution, specified in terms of the weights ρ_{st} and the empirical marginals $\widehat{\mu}$. If a regularizing term is added, these estimates no longer have a closed-form solution, but the optimization problem (5) can still be solved efficiently by message-passing methods.

Joint estimation/prediction: Using such an estimator, we now consider the joint approach to estimation and prediction illustrated in Figure 2. Using an initial set of i.i.d. samples, we first use the surrogate likelihood (5) to construct a parameter estimate $\widehat{\theta}^n$. Given a new noisy or incomplete observation y, we wish to perform near-optimal prediction or data fusion using the fitted model (e.g., for smoothing or interpolation of a noisy image). In order to do so, we first incorporate the new observation into the model, and then use the message-passing algorithm associated with the convex surrogate B in order to compute approximate pseudomarginals τ. These pseudomarginals can then be used to construct a prediction $\widehat{z}(y; \tau)$, where the specifics of the prediction depend on the observation model. We provide a concrete illustration in Section 5 using a mixture-of-Gaussians observation model.

4 Analysis

Asymptotics of estimator: We begin by considering the asymptotic behavior of the parameter estmiator $\widehat{\theta}^n$ defined by the surrogate likelihood (5). Since this parameter estimator is a particular type of M-estimator, the following result follows from standard techniques [8]:

Proposition 1. *For a general graph with cycles, $\widehat{\theta}^n$ converges in probability to some fixed $\widehat{\theta} \neq \theta^*$; moreover, $\sqrt{n}[\widehat{\theta}^n - \widehat{\theta}]$ is asymptotically normal.*

A key property of the estimator is its *inconsistency*—i.e., the estimated model $\widehat{\theta}$ differs from the true model θ^* even in the limit of large data. Despite this inconsistency, we will see that $\widehat{\theta}^n$ is useful for performing prediction.

Algorithmic stability: A desirable property of any algorithm—particularly one applied to statistical data—is that it exhibit an appropriate form of stability with respect to its inputs. Not all message-passing algorithms have such stability properties. For instance, the standard BP algorithm, although stable for relatively weakly coupled MRFs [3, 6], can be highly unstable due to phase transitions. Previous experimental work has shown that methods based on convex relaxations, including reweighted belief propagation [10],

Generic algorithm for joint parameter estimation and prediction:

1. Estimate parameters $\widehat{\theta}^n$ from initial data $x^1, \ldots, x^n$ by maximizing surrogate likelihood L_B.
2. Given a new set of observations y, incorporate them into the model:

$$\widetilde{\theta}_s(\,\cdot\,; y_s) = \widehat{\theta}^n_s(\,\cdot\,) + \log p(y_s \,|\, \cdot\,). \tag{6}$$

3. Compute approximate marginals τ by using the message-passing algorithm associated with the convex surrogate B. Use approximate marginals to construct prediction $\widehat{z}(y; \tau)$ of z based on the observation y and pseudomarginals τ.

Figure 2. Algorithm for joint parameter estimation and prediction. Both the learning and prediction steps are approximate, but the key is that they are both based on the same underlying convex surrogate B. Such a construction yields a provably beneficial cancellation of the two sources of error (learning and prediction).

reweighted generalized BP [13], and log-determinant relaxations [12] appear to be very stable. Here we provide theoretical support for these empirical observations: in particular, we prove that, in sharp contrast to non-convex methods, any variational method based on a strongly convex entropy approximation is globally stable.

A function $f : \mathbb{R}^n \to \mathbb{R}$ is *strongly convex* if there exists a constant $c > 0$ such that $f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{c}{2}\|y - x\|^2$ for all $x, y \in \mathbb{R}^n$. For a twice continuously differentiable function, this condition is equivalent to the eigenspectrum of the Hessian $\nabla^2 f(x)$ being uniformly bounded away from zero by c. With this definition, we have:

Proposition 2. *Consider any variational method based on a strongly concave entropy approximation $-B^*$; moreover, for any parameter $\theta \in \mathbb{R}^d$, let $\tau(\theta)$ denote the associated set of pseudomarginals. If the optimum is attained interior of the constraint set, then there exists a constant $R < +\infty$ such that*

$$\|\tau(\theta + \delta) - \tau(\theta)\| \quad \leq \quad R\|\delta\| \qquad \text{for all } \theta, \delta \in \mathbb{R}^d.$$

Proof. By our construction of the convex surrogate B, we have $\tau(\theta) = \nabla B(\theta)$, so that the statement is equivalent to the assertion that the gradient ∇B is a Lipschitz function. Applying the mean value theorem to ∇B, we can write $\nabla B(\theta + \delta) - \nabla B(\theta) = \nabla^2 B(\theta + t\delta)\delta$ where $t \in [0, 1]$. Consequently, in order to establish the Lipschitz condition, it suffices to show that the spectral norm of $\nabla^2 B(\gamma)$ is uniformly bounded above over all $\gamma \in \mathbb{R}^d$. Differentiating the relation $\nabla B(\theta) = \tau(\theta)$ yields $\nabla^2 B(\theta) = \nabla \tau(\theta)$. Now standard sensitivity analysis results [1] yield that $\nabla \tau(\theta) = [\nabla^2 B^*(\tau(\theta)]^{-1}$. Finally, our assumption of strong convexity of B^* yields that the spectral norm of $\nabla^2 B^*(\tau)$ is uniformly bounded away from zero, which yields the claim. □

Many existing entropy approximations, including the convexifed Bethe entropy (4), can be shown to be strongly concave [9].

5 Bounds on performance loss

We now turn to theoretical analysis of the joint method for parameter estimation and prediction illustrated in Figure 2. Note that given our setting of limited computation, the Bayes optimum is unattainable for two reasons: (a) it has knowledge of the exact parameter value θ^*; and (b) the prediction step (7) involves computing exact marginal probabilities μ. Therefore, our ultimate goal is to bound the performance loss of our method relative to the unachievable Bayes optimum. So as to obtain a concrete result, we focus on the special case of joint learning/prediction for a mixture-of-Gaussians; however, the ideas and techniques described here are more generally applicable.

Prediction for mixture of Gaussians: Suppose that the discrete random vector is a label vector for the components in a finite mixture of Gaussians: i.e., for each $s \in V$, the random variable Z_s is specified by $p(Z_s = z_s \,|\, X_s = j; \theta^*) \sim N(\nu_j, \sigma_j^2)$, for $j \in \{0, 1, \ldots, m-1\}$. Such models are widely used in statistical signal and image processing [2]. Suppose that we observe a noise-corrupted version of Z_s—namely $Y_s = \alpha Z_s + \sqrt{1-\alpha^2} W_s$, where $W_s \sim N(0,1)$ is additive Gaussian noise, and the parameter $\alpha \in [0,1]$ specifies the signal-to-noise ratio (SNR) of the observation model. (Here $\alpha = 0$ corresponds to pure noise, whereas $\alpha = 1$ corresponds to completely uncorrupted observations.)

With this set-up, it is straightforward to show that the optimal Bayes least squares estimator (BLSE) of Z takes the form

$$\widehat{z}_s(y; \mu) \quad := \quad \sum_{j=0}^{m-1} \mu_s(j; \theta^*) \Big[\omega_j(\alpha) \big(y_s - \nu_j \big) + \nu_j \Big], \tag{7}$$

where $\mu_s(j; \theta^*)$ is the exact marginal of the distribution $p(y \,|\, x) p(x; \theta^*)$; and $\omega_j(\alpha) := \frac{\alpha \sigma_j^2}{\alpha^2 \sigma_j^2 + (1-\alpha^2)}$ is the usual BLSE weighting for a Gaussian with variance σ_j. For this set-up, the approximate predictor $\widehat{z}_s(y; \tau)$ defined by our joint procedure in Figure 2 corresponds to replacing the exact marginals μ with the pseudomarginals $\tau_s(j; \widetilde{\theta})$ obtained by solving the variational problem with $\widetilde{\theta}$.

Bounds on performance loss: We now turn to a comparison of the mean-squared error (MSE) of the Bayes optimal predictor $\widehat{z}(Y; \mu)$ to the MSE of the surrogate-based predictor $\widehat{z}(Y; \tau)$. More specifically, we provide an upper bound on the increase in MSE, where the bound is specified in terms of the coupling strength and the SNR parameter α. Although results of this nature can be derived more generally, for simplicity we focus on the case of two mixture components ($m = 2$), and consider the asymptotic setting, in which the number of data samples $n \to +\infty$, so that the law of large numbers [8] ensures that the empirical marginals $\widehat{\mu}^n$ converge to the exact marginal distributions μ^*. Consequently, the MLE converges to the true parameter value θ^*, whereas Proposition 1 guarantees that our approximate parameter estimate $\widehat{\theta}^n$ converges to the fixed quantity $\widehat{\theta}$. By construction, we have the relations $\nabla B(\widehat{\theta}) = \mu^* = \nabla A(\theta^*)$.

An important factor in our bound is the quantity

$$L(\theta^*; \widehat{\theta}) \quad := \quad \sup_{\delta \in \mathbb{R}^d} \sigma_{\max} \big(\nabla^2 A(\theta^* + \delta) - \nabla^2 B(\widehat{\theta} + \delta) \big), \tag{8}$$

where $\sigma_{\max}$ denotes the maximal singular value. Following the argument in the proof of Proposition 2, it can be seen that $L(\theta^*; \widehat{\theta})$ is finite. Two additional quantities that play a role in our bound are the differences

$$\Delta_\omega(\alpha) := \omega_1(\alpha) - \omega_0(\alpha), \quad \text{and} \quad \Delta_\nu(\alpha) := [1 - \omega_1(\alpha)]\nu_1 - [1 - \omega_0(\alpha)]\nu_0,$$

where 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 state the following result (see the technical report [9] for the proof):

Theorem 1. *Let* $\mathrm{MSE}(\tau)$ *and* $\mathrm{MSE}(\mu)$ *denote the mean-squared prediction errors of the surrogate-based predictor* $\widehat{z}(y; \tau)$*, and the Bayes optimal estimate* $\widehat{z}(y; \mu)$ *respectively. The MSE increase* $\mathcal{I}(\alpha) := \frac{1}{N} \big[\mathrm{MSE}(\tau) - \mathrm{MSE}(\mu) \big]$ *is upper bounded by*

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

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

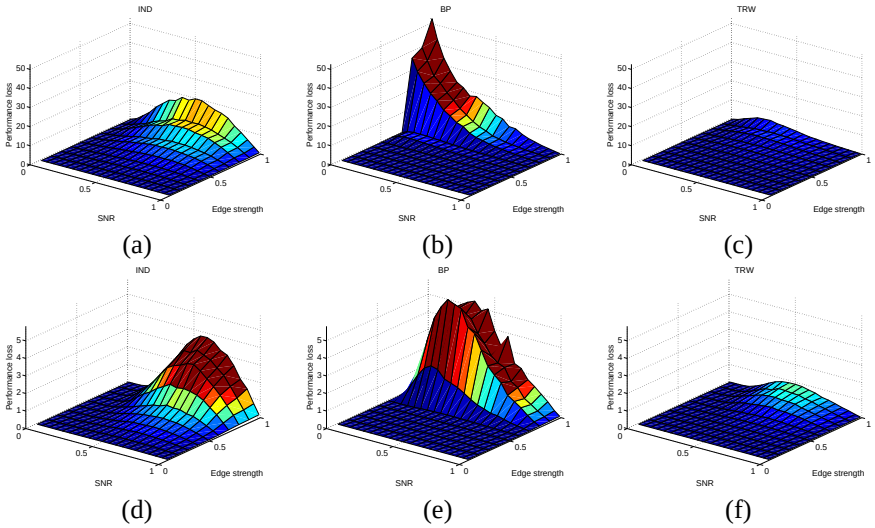

Figure 3. Surface plots of the percentage increase in MSE relative to Bayes optimum for different methods as a function of observation SNR and coupling strength. Top row: Gaussian mixture with components $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$. Bottom row: Gaussian mixture 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).

It can be seen that $\mathcal{I}(\alpha) \to 0$ as $\alpha \to 0^+$ and as $\alpha \to 1^-$, so that the surrogate-based method is asymptotically optimal for both low and high SNR. The behavior of the bound in the intermediate regime is controlled by the balance between these two terms.

Experimental results: In order to test our joint estimation/prediction procedure, we have applied it to coupled Gaussian mixture models on different graphs, coupling strengths, observation SNRs, and mixture distributions. Although our methods are more generally applicable, here we show representative results for $m = 2$ components, and two different mixture types. The first ensemble, constructed with mean and variance components $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_1, \sigma_1^2) = (0, 9)$, mimics heavy-tailed behavior. The second ensemble is bimodal, with components $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$. In both cases, each mixture component is equally weighted. Here we show results for a 2-D grid with $N = 64$ nodes. Since the mixture variables have $m = 2$ states, the coupling distribution 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 in $[0, 1]$. We evaluated the following three methods based on their increase in mean-squared error (MSE) over the Bayes optimal predictor (7): (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 BLSE 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 compute the pseudomarginals. (c) The *tree-reweighted method* (TRW) is based on estimating parameters using the tree-reweighted surrogate [10] with weights $\rho_{st} = \frac{1}{2}$ for all edges (s, t), and using the TRW sum-product algorithm to compute the pseudomarginals.

Shown in Figure 3 are 2-D surface plots of the average percentage increase in MSE, taken over 100 trials, 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). For weak coupling ($\gamma \approx 0$), all three methods—including the independence model—perform quite well, as should be expected given the weak dependency. Although not clear in these plots, BP outperforms TRW for weak coupling; however, both methods lose than than 1% in this regime. As the coupling is increased, the BP method eventually deteriorates quite seriously; indeed, for large enough coupling and low/intermediate SNR, its performance can be worse than the independence model. Looking at alternative models (in which phase transitions are known), we have found that this rapid degradation co-incides with the appearance of multiple fixed points. In contrast, the behavior of the TRW method is extremely stable, consistent with our theory.

6 Conclusion

We have described and analyzed joint methods for parameter estimation and prediction/smoothing using variational methods that are based on convex surrogates to the cumulant generating function. Our results—both theoretical and experimental—confirm the intuition that in the computation-limited setting, in which errors arise from approximations made both during parameter estimation and subsequent prediction, it is *provably beneficial* to use an inconsistent parameter estimator. Our experimental results on the coupled mixture of Gaussian model confirm the theory: the tree-reweighted sum-product algorithm yields prediction results close to the Bayes optimum, and substantially outperforms an analogous but heuristic method based on standard belief propagation.

References

[1] D. Bertsekas. *Nonlinear programming*. Athena Scientific, Belmont, MA, 1995.

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

[3] A. Ihler, J. Fisher, and A. S. Willsky. Loopy belief propagation: Convergence and effects of message errors. *Journal of Machine Learning Research*, 6:905–936, May 2005.

[4] M. A. R. Leisink and H. J. Kappen. Learning in higher order Boltzmann machines using linear response. *Neural Networks*, 13:329–335, 2000.

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

[6] S. Tatikonda and M. I. Jordan. Loopy belief propagation and Gibbs measures. In *Proc. Uncertainty in Artificial Intelligence*, volume 18, pages 493–500, August 2002.

[7] Y. W. Teh and M. Welling. On improving the efficiency of the iterative proportional fitting procedure. In *Workshop on Artificial Intelligence and Statistics*, 2003.

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

[9] M. J. Wainwright. Joint estimation and prediction in Markov random fields: Benefits of inconsistency in the computation-limited regime. Technical Report 690, Department of Statistics, UC Berkeley, 2005.

[10] 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 Artificial Intelligence and Statistics*, January 2003.

[11] 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*, 51(7):2313–2335, July 2005.

[12] M. J. Wainwright and M. I. Jordan. A variational principle for graphical models. In *New Directions in Statistical Signal Processing*. MIT Press, Cambridge, MA, 2005.

[13] W. Wiegerinck. Approximations with reweighted generalized belief propagation. In *Workshop on Artificial Intelligence and Statistics*, January 2005.

[14] J. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free energy approximations and generalized belief propagation algorithms. *IEEE Trans. Info. Theory*, 51(7):2282–2312, July 2005.