Journal of Machine Learning Research 7 (2006) 1829-1859 Submitted 5/06; Revised 8/06; Published 9/06

Estimating the "Wrong" Graphical Model: Benefits in the Computation-Limited Setting

Martin J. Wainwright WAINWRIG@STAT.BERKELEY.EDU
Department of Statistics
Department of Electrical Engineering and Computer Sciences
University of California at Berkeley
Berkeley, CA 94720, USA

Editor: Max Chickering

Abstract

Consider the problem of joint parameter estimation and prediction in a Markov random field: that is, 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 under the restriction of limited computation, 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 convex variational methods. This stability result provides additional incentive, apart from the obvious benefit of unique global optima, for using message-passing methods based on convex variational relaxations. We show that joint estimation/prediction based on the reweighted sum-product algorithm substantially outperforms a commonly used heuristic based on ordinary sum-product.

Keywords: graphical model, Markov random field, belief propagation, variational method, parameter estimation, prediction error, algorithmic stability

1. Introduction

Graphical models such as Markov random fields (MRFs) are widely used in many application domains, including machine learning, natural language processing, statistical signal processing, and communication theory. A fundamental limitation to their practical use is the difficulty associated with computing various statistical quantities (e.g., marginals, data likelihoods etc.); such quantities are of interest both Bayesian and frequentist settings. Sampling-based methods, especially those of the Markov chain Monte Carlo (MCMC) variety (Liu, 2001; Robert and Casella, 1999), represent one approach to obtaining stochastic approximations to marginals and likelihoods. A possible disadvantage of sampling methods is their relatively high computational cost. It is thus of considerable interest for various application domains to consider less computationally intensive methods for generating approximations to marginals, log likelihoods, and other relevant statistical quantities.

Variational methods are one class of techniques that can be used to At the foundation of these methods is the fact that for a broad class of MRFs, the computation of the log likelihood and

marginal probabilities can be reformulated as a convex optimization problem; see Yedidia (2001) or Wainwright and Jordan (2003) for overviews. Although this optimization problem is intractable to solve exactly for general MRFs, it suggests a principled route to obtaining approximations—namely, by relaxing the original optimization problem, and taking the optimal solutions to the relaxed problem as approximations to the exact values. In many cases, optimization of the relaxed problem can be carried out by "message-passing" algorithms, in which neighboring nodes in the Markov random field convey statistical information (e.g., likelihoods) by passing functions or vectors (referred to as messages). Well-known examples of such variational methods include mean field algorithms, the belief propagation or sum-product algorithm, as well as various extensions including generalized belief propagation and expectation propagation.

Estimating the parameters of a Markov random field from data poses another significant challenge. A direct approach—for instance, via (regularized) maximum likelihood estimation—entails evaluating the cumulant generating (or log partition) function, which is computationally intractable for general Markov random fields. One viable option is the pseudolikelihood method (Besag, 1975, 1977), which can be shown to produce consistent parameter estimates under suitable assumptions, though with an associated loss of statistical efficiency. Other researchers have studied algorithms for ML estimation based on stochastic approximation (Younes, 1988; Benveniste et al., 1990), which again are consistent under appropriate assumptions, but can be slow to converge.

1.1 Overview

As illustrated in Figure 1, the problem domain of interest in this paper is that of joint estimation and prediction in a Markov random field. More precisely, given samples $\{X^1, \ldots, X^n\}$ from some unknown underlying model $p(\cdot\,; \theta^*)$, the first step is to form an estimate of the model parameters. Now suppose that we are given a noisy observation of a new sample path $Z \sim p(\cdot\,; \theta^*)$, and that we wish to form a (near)-optimal estimate of Z using the fitted model, and the noisy observation (denoted Y). Examples of such prediction problems include signal denoising, interpolation of missing data, and sentence parsing. Disregarding any issues of computational cost and speed, one could proceed via Route A in Figure 1—that is, one could envisage first using a standard technique (e.g., regularized maximum likelihood) for parameter estimation, and then carrying out the prediction step (which might, for instance, involve computing certain marginal probabilities) by Monte Carlo methods.

This paper, in contrast, is concerned with the *computation-limited* setting, in which both sampling or brute force methods are overly intensive. With this motivation, a number of researchers have studied the use of approximate message-passing techniques, both for problems of prediction (Heskes et al., 2003; Ihler et al., 2005; Minka, 2001; Mooij and Kappen, 2005b; Tatikonda, 2003; Wainwright et al., 2003a; Wiegerinck, 2005; Yedidia et al., 2005) as well as for parameter estimation (Leisink and Kappen, 2000; Sutton and McCallum, 2005; Teh and Welling, 2003; Wainwright et al., 2003b). However, despite their wide-spread use, the theoretical understanding of such message-passing techniques remains limited[1], especially for parameter estimation. Consequently, it is of considerable interest to characterize and quantify the loss in performance incurred by using computationally tractable methods versus exact methods (i.e., Route B versus A in Figure 1). More

1. The behavior of sum-product is relatively well understood in certain settings, including graphs with single cycles (Weiss, 2000), Gaussian models (Freeman and Weiss, 2001; Rusmevichientong and Roy, 2000) and Ising models (Tatikonda and Jordan, 2002; Ihler et al., 2005; Mooij and Kappen, 2005a). Similarly, there has been substantial progress for graphs with high girth (Richardson and Urbanke, 2001), but much of this analysis breaks down in application to graphs with short cycles.

specifically, our analysis applies to variational methods that are based on *convex relaxations*. This class includes a number of existing methods—among them the tree-reweighted sum-product algorithm (Wainwright et al., 2005), reweighted forms of generalized belief propagation (Wiegerinck, 2005), and semidefinite relaxations (Wainwright and Jordan, 2005). Moreover, it is possible to modify other variational methods—for instance, expectation propagation (Minka, 2001)—so as to "convexify" them.

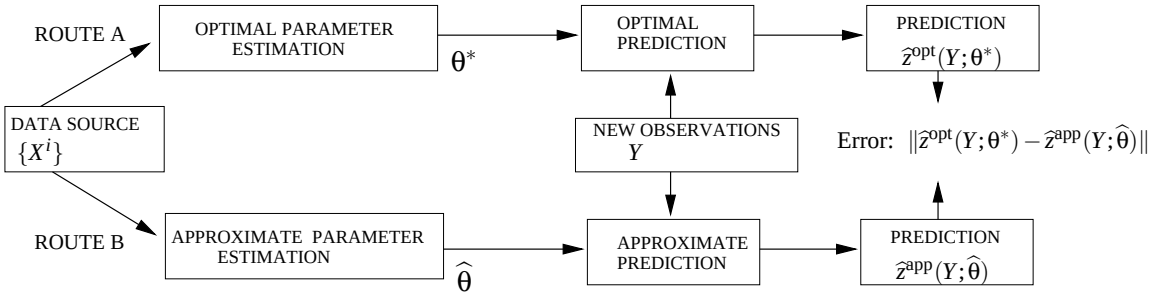


Figure 1: Route A: computationally intractable combination of parameter estimation and prediction. Route B: computationally efficient combination of approximate parameter estimation and prediction.

1.2 Our Contributions

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. The theoretical analysis of this paper 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 *inconsistent* methods for parameter estimation. 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. Such global algorithmic stability is a fundamental concern given statistical models imperfectly estimated from limited data, or for applications in which "errors" may be introduced into message-passing (e.g., due to quantization or other forms of communication constraints in sensor networks). Thus, our global stability result provides further theoretical justification—apart from the obvious benefit of unique global optima—for using message-passing methods based on convex variational relaxations. Finally, we provide some empirical results to show that joint estimation/prediction based on the reweighted sum-product algorithm substantially outperforms a commonly used heuristic based on ordinary sum-product.

The remainder of this paper is organized as follows. Section 2 provides background on Markov random fields. In Section 3, we introduce background on variational representations, including the notion of a convex surrogate to the cumulant generating function, and then illustrate this notion via the tree-reweighted Bethe approximation (Wainwright et al., 2005). In Section 4, we describe how any convex surrogate defines a particular joint scheme for parameter estimation and prediction. Section 5 provides results on the asymptotic behavior of the estimation step, as well as the stability of the prediction step. Section 6 is devoted to the derivation of performance bounds for joint estimation

and prediction methods, with particular emphasis on the mixture-of-Gaussians observation model. In Section 7, we provide experimental results on the performance of a joint estimation/prediction method based on the tree-reweighted Bethe surrogate, and compare it to a heuristic method based on the ordinary belief propagation algorithm. We conclude in Section 8 with a summary and discussion of directions for future work.

2. Background

We begin with background on Markov random fields. Consider an undirected graph $G = (V, E)$, consisting of a set of vertices $V = \{1, \ldots, N\}$ and an edge set E. We associate to each vertex $s \in V$ a multinomial random variable X_s taking values in the set $\mathcal{X}_s = \{0, 1, \ldots, m-1\}$. We use the lower case letter x_s to denote particular realizations of the random variable X_s in the set $\mathcal{X}_s$. This paper makes use of the following exponential representation of a pairwise Markov random field over the multinomial random vector $X := \{X_s,\ s \in V\}$. We begin by defining, for each $j = 1, \ldots, m-1$, the $\{0,1\}$-valued indicator function

$$\mathbb{I}_j[x_s] \quad := \quad \begin{cases} 1 & \text{if } x_s = j \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

These indicator functions can be used to define a potential function $\theta_s(\cdot) : \mathcal{X}_s \to \mathbb{R}$ via

$$\theta_s(x_s) \quad := \quad \sum_{j=1}^{m-1} \theta_{s;j} \mathbb{I}_j[x_s] \tag{2}$$

where $\theta_s = \{\theta_{s;j}, j = 1, \ldots, m-1\}$ is the vector of exponential parameters associated with the potential. Our exclusion of the index $j = 0$ is deliberate, so as to ensure that the collection of indicator functions $\phi_s(x_s) := \{\mathbb{I}_j[x_s],\ j = 1, \ldots, m-1\}$ remain affinely independent. In a similar fashion, we define for any pair $(s,t) \in E$ the pairwise potential function

$$\theta_{st}(x_s, x_t) \quad := \quad \sum_{j=1}^{m-1} \sum_{k=1}^{m-1} \theta_{st;jk} \mathbb{I}_j[x_s]\, \mathbb{I}_k[x_t],$$

where we use $\theta_{st} := \{\theta_{st;jk},\ j,k = 1, 2, \ldots, m-1\}$ to denote the associated collection of exponential parameters, and $\phi_{st}(x_s, x_t) := \{\mathbb{I}_j[x_s]\mathbb{I}_k[x_s],\ j,k = 1, 2, \ldots, m-1\}$ for the associated set of sufficient statistics.

Overall, the probability mass function of the multinomial Markov random field in exponential form can be written as

$$p(x; \theta) \quad = \quad \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) - A(\theta) \Big\}. \tag{3}$$

Here the function

$$A(\theta) \quad := \quad \log \Big[\sum_{x \in \mathcal{X}^N} \exp\Big\{ \sum_{s \in V} \theta_s(x_s) + \sum_{(s,t) \in E} \theta_{st}(x_s, x_t) \Big\} \Big] \tag{4}$$

is the logarithm of the normalizing constant associated with $p(\cdot; \theta)$.

The collection of distributions thus defined can be viewed as a regular and minimal exponential family (Brown, 1986). In particular, the exponential parameter θ and the vector of sufficient statistics ϕ are formed by concatenating the exponential parameters (respectively indicator functions) associated with each vertex and edge—viz.

$$\begin{aligned} \theta &= \{\theta_s, s \in V\} \cup \{\theta_{st},\ (s,t) \in E\} \\ \phi(x) &= \{\phi_s(x_s), s \in V\} \cup \{\phi_{st}(x_s, x_t),\ (s,t) \in E\} \end{aligned}$$

This notation allows us to write Equation (3) more compactly as $p(x;\theta) = \exp\{\langle \theta, \phi(x)\rangle - A(\theta)\}$. A quick calculation shows that $\theta \in \mathbb{R}^d$, where $d = N(m-1) + |E|\,(m-1)^2$ is the dimension of this exponential family.

The following properties of A are well-known:

Lemma 1 *(a) The function A is convex, and strictly so when the sufficient statistics are affinely independent.*
(b) It is an infinitely differentiable function, with derivatives corresponding to cumulants. In particular, for any indices $\alpha, \beta \in \{1, \ldots, d\}$, we have

$$\frac{\partial A}{\partial \theta_\alpha} = \mathbb{E}_\theta[\phi_\alpha(X)], \qquad \frac{\partial^2 A}{\partial \theta_\alpha \partial \theta_\beta} = \mathrm{cov}_\theta\{\phi_\alpha(X), \phi_\beta(X)\},$$

where $\mathbb{E}_\theta$ and cov_θ denote the expectation and covariance respectively.

We use $\mu \in \mathbb{R}^d$ to denote the vector of *mean parameters* defined element-wise by $\mu_\alpha = \mathbb{E}_\theta[\phi_\alpha(X)]$ for any $\alpha \in \{1, \ldots, d\}$. A convenient property of the sufficient statistics ϕ defined in Equations (1) and (2) is that these mean parameters correspond to marginal probabilities. For instance, when $\alpha = (s; j)$ or $\alpha = (st; jk)$, we have respectively

$$\mu_{s;j} = \mathbb{E}_\theta[\mathbb{I}_j[x_s]] = p(X_s = j; \theta), \quad \text{and} \tag{5a}$$

$$\mu_{st;jk} = \mathbb{E}_\theta\big\{\mathbb{I}_j[x_s]\,\mathbb{I}_k[x_t]\big\} = p(X_s = j, X_t = k; \theta). \tag{5b}$$

3. Construction of Convex Surrogates

This section is devoted to a systematic procedure for constructing convex functions that represent approximations to the cumulant generating function. We begin with a quick development of an exact variational principle, one which is intractable to solve in general cases; see the papers (Pietra et al., 1997; Wainwright and Jordan, 2005) for further details. Nonetheless, this exact variational principle is useful, in that various natural relaxations of the optimization problem can be used to define convex surrogates to the cumulant generating function. After a high-level description of such constructions in general, we then illustrate it more concretely with the particular case of the "convexified" Bethe entropy (Wainwright et al., 2005).

3.1 Exact Variational Representation

Since A is a convex and continuous function (see Lemma 1), the theory of convex duality (Rockafellar, 1970) guarantees that it has a variational representation, given in terms of its conjugate dual function $A^* : \mathbb{R}^d \to \mathbb{R} \cup \{+\infty\}$, of the following form

$$A(\theta) = \sup_{\mu \in \mathbb{R}^d} \big\{\theta^T \mu - A^*(\mu)\big\}.$$

In order to make effective use of this variational representation, it remains determine the form of the dual function. A useful fact is that the exponential family (3) arises naturally as the solution of an entropy maximization problem. In particular, consider the set of linear constraints

$$\mathbb{E}_p[\phi(X)] := \sum_{x \in \mathcal{X}^N} p(x)\phi_\alpha(x) \quad = \quad \mu_\alpha \quad \text{for } \alpha = 1, \ldots, d, \tag{6}$$

where $\mu \in \mathbb{R}^d$ is a set of target mean parameters. Letting $\mathcal{P}$ denote the set of all probability distributions with support on $\mathcal{X}^N$, consider the *constrained entropy maximization problem*: maximize the entropy $H(p) := -\sum_{x \in \mathcal{X}^N} p(x) \log p(x)$ subject to the constraints (6).

A first question is when there any distributions p that satisfy the constraints (6). Accordingly, we define the set

$$\text{MARG}_\phi(G) \quad := \quad \left\{ \mu \in \mathbb{R}^d \,\middle|\, \mu = \mathbb{E}_p[\phi(X)] \quad \text{for some } p \in \mathcal{P} \right\},$$

corresponding to the set of μ for which the constraint set (6) is non-empty. For any $\mu \notin \text{MARG}_\phi(G)$, the optimal value of the constrained maximization problem is $-\infty$ (by definition, since the problem is infeasible). Otherwise, it can be shown that the optimum is attained at a unique distribution in the exponential family, which we denote by $p(\cdot\,; \theta(\mu))$. Overall, these facts allow us to specify the conjugate dual function as follows:

$$A^*(\mu) \quad = \quad \begin{cases} -H(p(\cdot\,; \theta(\mu))) & \text{if } \mu \in \text{MARG}_\phi(G) \\ +\infty & \text{otherwise.} \end{cases} \tag{7}$$

See the technical report (Wainwright and Jordan, 2003) for more details of this dual calculation. With this form of the dual function, we are guaranteed that the cumulant generating function A has the following variational representation:

$$A(\theta) \quad = \quad \max_{\mu \in \text{MARG}_\phi(G)} \left\{ \theta^T \mu - A^*(\mu) \right\}. \tag{8}$$

However, in general, solving the variational problem (8) is intractable. This intractability should not be a surprise, since the cumulant generating function is intractable to compute for a general graphical model. The difficulty arises from two sources. First, the *constraint set* $\text{MARG}_\phi(G)$ is extremely difficult to characterize exactly for a general graph with cycles. For the case of a multinomial Markov random field (3), it can be seen (using the Minkowski-Weyl theorem) that $\text{MARG}_\phi(G)$ is a polytope, meaning that it can be characterized by a finite number of linear constraints. The question, of course, is how rapidly this number of constraints grows with the number of nodes N in the graph. Unless certain fundamental conjectures in computational complexity turn out to be false, this growth must be non-polynomial; see Deza and Laurent (1997) for an in-depth discussion of the binary case. Tree-structured graphs are a notable exception, for which the junction tree theory (Lauritzen, 1996) guarantees that the growth is only linear in N.

Second, the *dual function* A^* lacks a closed-form representation for a general graph. Note in particular that the representation (7) is not explicit, since it requires solving a constrained entropy maximization problem in order to compute the value $H(p(\cdot\,; \theta(\mu)))$. Again, important exceptions to this rule are tree-structured graphs. Here a special case of the junction tree theory guarantees

that any Markov random field on a tree $T = (V, E(T))$ can be factorized in terms of its marginals as follows

$$p(x; \theta(\mu)) \quad = \quad \prod_{s \in V} \mu_s(x_s) \prod_{(s,t) \in E(T)} \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}. \tag{9}$$

Consequently, in this case, the negative entropy (and hence the dual function) can be computed explicitly as

$$-A^*(\mu; T) \quad = \quad \sum_{s \in V} H_s(\mu_s) - \sum_{(s,t) \in E(T)} I_{st}(\mu_{st}) \tag{10}$$

where $H_s(\mu_s) := -\sum_{x_s} \mu_s(x_s) \log \mu_s(x_s)$ and $I_{st}(\mu_{st}) := \sum_{x_s, x_t} \mu_{st}(x_s, x_t) \log \frac{\mu_{st}(x_s, x_t)}{\mu_s(x_s)\mu_t(x_t)}$ are the singleton entropy and mutual information, respectively, associated with the node $s \in V$ and edge $(s,t) \in E(T)$. For a general graph with cycles, in contrast, the dual function lacks such an explicit form, and is not easy to compute.

Given these challenges, it is natural to consider approximations to A^* and $\mathrm{MARG}_\phi(G)$. As we discuss in the following section, the resulting relaxed optimization problem defines a convex surrogate to the cumulant generating function.

3.2 Convex Surrogates to the Cumulant Generating Function

We now describe a general procedure for constructing convex surrogates to the cumulant generating function, consisting of two main ingredients. Given the intractability of characterizing the marginal polytope $\mathrm{MARG}_\phi(G)$, it is natural to consider a relaxation. More specifically, let $\mathrm{REL}_\phi(G)$ be a convex and compact set that acts as an outer bound to $\mathrm{MARG}_\phi(G)$. We use τ to denote elements of $\mathrm{REL}_\phi(G)$, and refer to them as *pseudomarginals* since they represent relaxed versions of local marginals. The second ingredient is designed to sidestep the intractability of the dual function: in particular, let B^* be a strictly convex and twice continuously differentiable approximation to A^*. We require that the domain of B^* (i.e., $\mathrm{dom}(B^*) := \{\tau \in \mathbb{R}^d \,|\, B^*(\tau) < +\infty\}$) be contained within the relaxed constraint set $\mathrm{REL}_\phi(G)$.

By combining these two approximations, we obtain a convex surrogate B to the cumulant generating function, specified via the solution of the following relaxed optimization problem

$$B(\theta) \quad := \quad \max_{\tau \in \mathrm{REL}_\phi(G)} \left\{ \theta^T \tau - B^*(\tau) \right\}. \tag{11}$$

Note the parallel between this definition (11) and the variational representation of A in Equation (8).

The function B so defined has several desirable properties, as summarized in the following proposition:

Proposition 2 *Any convex surrogate B defined via* (11) *has the following properties:*

(i) For each $\theta \in \mathbb{R}^d$, the optimum defining B is attained at a unique point $\tau(\theta)$.

(ii) The function B is convex on $\mathbb{R}^d$.

(iii) It is differentiable on $\mathbb{R}^d$, and more specifically:

$$\nabla B(\theta) \quad = \quad \tau(\theta).$$

Proof (i) By construction, the constraint set $\mathrm{REL}_\phi(G)$ is compact and convex, and the function B^* is strictly convex, so that the optimum is attained at a unique point $\tau(\theta)$.
(ii) Observe that B is defined by the maximum of a collection of functions linear in θ, which ensures that it is convex (Bertsekas, 1995).
(iii) Finally, the function $\theta^T\tau - B^*(\tau)$ satisfies the hypotheses of Danskin's theorem (Bertsekas, 1995), from which we conclude that B is differentiable with $\nabla B(\theta) = \tau(\theta)$ as claimed. ■

Given the interpretation of $\tau(\theta)$ as a pseudomarginal, this last property of B is analogous to the well-known cumulant generating property of A—namely, $\nabla A(\theta) = \mu(\theta)$—as specified in Lemma 1.

3.3 Convexified Bethe Surrogate

The following example provides a more concrete illustration of this constructive procedure, using a tree-based approximation to the marginal polytope, and a convexifed Bethe entropy approximation (Wainwright et al., 2005). As with the ordinary Bethe approximation (Yedidia et al., 2005), the cost function and constraint set underlying this approximation are exact for any tree-structured Markov random field.

Relaxed polytope: We begin by describing a relaxed version $\mathrm{REL}_\phi(G)$ of the marginal polytope $\mathrm{MARG}_\phi(\phi)$. Let τ_s and τ_{st} represent a collection of singleton and pairwise pseudomarginals, respectively, associated with vertices and edges of a graph G. These quantities, as locally valid marginal distributions, must satisfy the following set of local consistency conditions:

$$\mathrm{LOCAL}_\phi(G) \quad := \quad \Big\{\tau \in \mathbb{R}_+^d \,\Big|\, \sum_{x_s} \tau_s(x_s) = 1, \;\; \sum_{x_t} \tau_{st}(x_s, x_t) = \tau_s(x_s)\Big\}.$$

By construction, we are guaranteed the inclusion $\mathrm{MARG}_\phi(G) \subset \mathrm{LOCAL}_\phi(G)$. Moreover, a special case of the junction tree theory (Lauritzen, 1996) guarantees that equality holds when the underlying graph is a tree (in particular, any $\tau \in \mathrm{LOCAL}_\phi(G)$ can be realized as the marginals of the tree-structured distribution of the form (9)). However, the inclusion is strict for any graph with cycles; see Appendix A for further discussion of this issue.

Entropy approximation: We now define an entropy approximation B_ρ^* that is finite for any pseudomarginal τ in the relaxed set $\mathrm{LOCAL}_\phi(G)$. We begin by considering a collection $\{T \in \mathfrak{T}\}$ of spanning trees associated with the original graph. Given $\tau \in \mathrm{LOCAL}_\phi(G)$, there is—for each spanning tree T—a unique tree-structured distribution that has marginals τ_s and τ_{st} on the vertex set V and edge set $E(T)$ of the tree. Using Equations (9) and (10), the entropy of this tree-structured distribution can be computed explicitly. The *convexified Bethe entropy* approximation is based on taking a convex combination of these tree entropies, where each tree is weighted by a probability $\rho(T) \in [0,1]$. Doing so and expanding the sum yields

$$B_\rho^*(\tau) \quad := \quad \sum_{T \in \mathfrak{T}} \rho(T) \Big\{ \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E(T)} I_{st}(\tau_{st}) \Big\} \;=\; \sum_{s \in V} H_s(\tau_s) - \sum_{(s,t) \in E} \rho_{st} I_{st}(\tau_{st}), \quad (12)$$

where $\rho_{st} = \sum_T \rho(T) \mathbb{I}[(s,t) \in T]$ are the *edge appearance probabilities* defined by the distribution ρ over the tree collection. By construction, the function B_ρ^* is differentiable; moreover, it can be shown (Wainwright et al., 2005) that it is strictly convex for any vector $\{\rho_{st}\}$ of strictly positive edge appearance probabilities.

Bethe surrogate and reweighted sum-product: We use these two ingredients—the relaxation $\mathrm{LOCAL}_\phi(G)$ of the marginal polytope, and the convexified Bethe entropy approximation (12)—to define the following convex surrogate

$$B_\rho(\theta) \quad := \quad \max_{\tau \in \mathrm{LOCAL}_\phi(G)} \left\{ \theta^T \tau - B^*_\rho(\tau) \right\}. \tag{13}$$

Since the conditions of Proposition 2 are satisfied, we are guaranteed that B_ρ is convex and differentiable on $\mathbb{R}^d$, and moreover that $\nabla B_\rho(\theta) = \tau(\theta)$, where (for each $\theta \in \mathbb{R}^d$) the quantity $\tau(\theta)$ denotes the unique optimum of problem (13). Perhaps most importantly, the optimizing pseudo-marginals $\tau(\theta)$ can be computed efficiently using a *tree-reweighted variant* of the sum-product message-passing algorithm (Wainwright et al., 2005). This method operates by passing "messages", which in the multinomial case are simply m-vectors of non-negative numbers, along edges of the graph. We use $M_{ts} = \{M_{ts}(i), i = 0, \ldots, m-1\}$ to represent the message passed from node t to node s. In the tree-reweighted variant, these messages are updated according to the following recursion

$$M_{ts}(x_s) \quad \leftarrow \quad \sum_{x_t} \exp\left\{ \theta_t(x_t) \frac{\theta_{st}(x_s, x_t)}{\rho_{st}} \right\} \frac{\prod_{u \in \Gamma(t) \setminus s} \left[M_{ut}(x_t) \right]^{\rho_{ut}}}{\left[M_{st}(x_t) \right]^{1-\rho_{st}}}. \tag{14}$$

Here $\Gamma(t)$ denotes the set of all neighbors of node t in the graph. Upon convergence of the updates, the fixed point messages M^* yield the unique global optimum of the optimization problem (13) via the following equations

$$\tau_s(x_s; \theta) \quad \propto \quad \exp\left\{ \theta_s(x_s) \right\} \prod_{u \in \Gamma(s)} \left[M_{us}(x_s) \right]^{\rho_{us}}, \quad \text{and} \tag{15a}$$

$$\tau_{st}(x_s, x_t; \theta) \quad \propto \quad \exp\left\{ \theta_s(x_s) + \theta_t(x_t) + \frac{\theta_{st}(x_s, x_t)}{\rho_{st}} \right\} \frac{\prod_{u \in \Gamma(s)} \left[M_{us}(x_s) \right]^{\rho_{us}} \prod_{v \in \Gamma(s)} \left[M_{vs}(x_s) \right]^{\rho_{vs}}}{M_{st}(x_t)\, M_{ts}(x_s)} \tag{15b}$$

Further details on these updates and their properties can be found in Wainwright et al. (2005).

4. Joint Estimation and Prediction Using Surrogates

We now turn to consideration of how convex surrogates, as constructed by the procedure described in the previous section, are useful for both approximate parameter estimation as well as prediction.

4.1 Approximate Parameter Estimation

Suppose that we are given i.i.d. samples $\{X^1, \ldots, X^n\}$ from an MRF of the form (3), where the underlying true parameter θ^* is unknown. One standard way in which to estimate θ^* is via maximum likelihood (possibly with an additional regularization term); in this particular exponential family setting, it is straightforward to show that the (normalized) log likelihood takes the form

$$\ell(\theta) \quad = \quad \langle \widehat{\mu}^n, \theta \rangle - A(\theta) - \lambda^n R(\theta)$$

where function R is a regularization term with an associated (possibly data-dependent) weight λ^n. The quantities $\widehat{\mu}^n := \frac{1}{n} \sum_{i=1}^n \phi(X^i)$ are the empirical moments defined by the data. For the indicator-based exponential representation (5), these empirical moments correspond to a set of singleton and pairwise marginal distributions, denoted $\widehat{\mu}^n_s$ and $\widehat{\mu}^n_{st}$ respectively.

It is intractable to maximize the regularized likelihood directly, due to the presence of the cumulant generating function A. Thus, a natural thought is to use the convex surrogate B to define an alternative estimator obtained by maximizing the regularized *surrogate likelihood*:

$$\ell_B(\theta) \quad := \quad \langle \widehat{\mu}^n, \theta \rangle - B(\theta) - \lambda^n R(\theta). \tag{16}$$

By design, the surrogate B and hence the surrogate likelihood ℓ_B, as well as their derivatives, can be computed in a straightforward manner (typically by some sort of message-passing algorithm). It is thus straightforward to compute the parameter $\widehat{\theta}^n$ achieving the maximum of the regularized surrogate likelihood (for instance, gradient descent would a simple though naive method).

For the tree-reweighted Bethe surrogate (13), we have shown in previous work (Wainwright et al., 2003b) 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 $\widehat{\mu}$. (We make use of this closed form in our experimental comparison in Section 7; see Equation (32).) If a regularizing term is added, these estimates no longer have a closed-form solution, but the optimization problem (16) can still be solved efficiently using the tree-reweighted sum-product algorithm (Wainwright et al., 2003b, 2005).

4.2 Joint Estimation and Prediction

Using such an estimator, we now consider a joint approach to estimation and prediction. Recalling the basic set-up, we are given an initial set of i.i.d. samples $\{x^1, \ldots, x^n\}$ from $p(\cdot\,; \theta^*)$, where the true model parameter θ^* is unknown. These samples are used to form an estimate of the Markov random field. We are then given a noisy observation y of a new sample $z \sim p(\cdot\,; \theta^*)$, and the goal is to use this observation in conjunction with the fitted model to form a near-optimal estimate of z. The key point is that the same convex surrogate B is used both in forming the surrogate likelihood (16) for approximate parameter estimation, and in the variational method (11) for performing prediction.

For a given fitted model parameter $\theta \in \mathbb{R}^d$, the central object in performing prediction is the posterior distribution $p(z \mid y; \theta) \propto p(z; \theta)\; p(y \mid z)$. In the exponential family setting, for a fixed noisy observation y, this posterior can always be written as a new exponential family member, described by parameter $\theta + \gamma(y)$. (Here the term $\gamma(y)$ serves to incorporate the effect of the noisy observation.) With this set-up, the procedure consists of the following steps:

Joint estimation and prediction:

1. Form an approximate parameter estimate $\widehat{\theta}^n$ from an initial set of i.i.d. data $\{x^1, \ldots, x^n\}$ by maximizing the (regularized) surrogate likelihood ℓ_B.

2. Given a new noisy observation y (i.e., a contaminated version of $z \sim p(\cdot\,; \theta^*)$) specified by a factorized conditional distribution of the form $p(y \mid z) = \prod_{s=1}^N p(y_s \mid z_s)$, incorporate it into the model by forming the new exponential parameter

$$\widehat{\theta}^n_s(\,\cdot\,) + \gamma_s(y)$$

where $\gamma_s(y)$ merges the new data with the fitted model $\widehat{\theta}^n$. (The specific form of γ depends on the observation model.)

3. Using the message-passing algorithm associated with the convex surrogate B, compute approximate marginals $\tau(\widehat{\theta}+\gamma)$ for the distribution that combines the fitted model with the new observation. Use these approximate marginals to construct a prediction $\widehat{z}(y;\tau)$ of z based on the observation y and pseudomarginals τ.

Examples of the prediction task in the final step include smoothing (e.g., denoising of a noisy image) and interpolation (e.g., in the presence of missing data). We provide a concrete illustration of such a prediction problem in Section 6 using a mixture-of-Gaussians observation model. The most important property of this joint scheme is that the *convex surrogate B* underlies both the parameter estimation phase (used to form the surrogate likelihood), and the prediction phase (used in the variational method for computing approximate marginals). It is this matching property that will be shown to be beneficial in terms of overall performance.

5. Analysis

In this section, we turn to the analysis of the surrogate-based method for estimation and prediction. We begin by exploring the asymptotic behavior of the parameter estimator. We then prove a Lipschitz stability result applicable to any variational method that is based on a strongly concave entropy approximation. This stability result plays a central role in our subsequent development of bounds on the performance loss in Section 6.

5.1 Estimator Asymptotics

We begin by considering the asymptotic behavior of the parameter estimator $\widehat{\theta}^n$ defined by the surrogate likelihood (16). Since this parameter estimator is a particular type of M-estimator (Serfling, 1980), its asymptotic behavior can be investigated using standard methods, as summarized in the following:

Proposition 3 *Recall the cumulant generating function A defined in Equation* (4)*. Let B be a strictly convex surrogate for A, defined via Equation* (11) *with a strictly concave entropy approximation $-B^*$. Consider the sequence of parameter estimates $\{\widehat{\theta}^n\}$ given by*

$$\widehat{\theta}^n \quad := \quad \arg\max_{\theta\in\mathbb{R}^d} \{\langle \widehat{\mu}^n, \theta\rangle - B(\theta) - \lambda^n R(\theta)\} \tag{17}$$

where R is a non-negative and convex regularizer, and the regularization parameter satisfies $\lambda^n = o(\frac{1}{\sqrt{n}})$.

Then for a general graph with cycles, the following results hold:

(a) we have $\widehat{\theta}^n \xrightarrow{p} \widehat{\theta}$, where $\widehat{\theta}$ is (in general) distinct from the true parameter θ^*.

(b) the estimator is asymptotically normal:

$$\sqrt{n}\big[\widehat{\theta}^n - \widehat{\theta}\big] \quad \xrightarrow{d} \quad N\Big(0,\, \big(\nabla^2 B(\widehat{\theta})\big)^{-1} \nabla^2 A(\theta^*) \big(\nabla^2 B(\widehat{\theta})\big)^{-1}\Big)$$

Proof By construction, the convex surrogate B and the (negative) entropy approximation B^* are a Fenchel-Legendre conjugate dual pair. From Proposition 2, the surrogate B is differentiable.

Moreover, the strict convexity of B and B^* ensure that the gradient mapping ∇B is one-to-one and onto the relative interior of the constraint set $\mathrm{REL}_\phi(G)$ (see Section 26 of Rockafellar (1970)). Moreover, the inverse mapping $(\nabla B)^{-1}$ exists, and is given by the dual gradient ∇B^*.

Let μ^* be the moment parameters associated with the true distribution θ^*—that is, $\mu^* = \mathbb{E}_{\theta^*}[\phi(X)]$. In the limit of infinite data, the asymptotic value of the parameter estimate is defined by

$$\nabla B(\widehat{\theta}) = \mu^*. \tag{18}$$

Note that μ^* belongs to the relative interior of $\mathrm{MARG}_\phi(G)$, and hence to the relative interior of $\mathrm{REL}_\phi(G)$. Therefore, Equation (18) has a unique solution $\widehat{\theta} = \nabla^{-1}B(\mu^*)$.

By strict convexity, the regularized surrogate likelihood (17) has a unique global maximum. Let us consider the optimality conditions defining this unique maximum $\widehat{\theta}^n$; they are given by $\nabla B(\widehat{\theta}^n) = \widehat{\mu}^n - \lambda^n \partial R(\widehat{\theta}^n)$, where $\partial R(\widehat{\theta}^n)$ denotes an arbitrary element of the subdifferential of the convex function R at the point $\widehat{\theta}^n$. We can now write

$$\nabla B(\widehat{\theta}^n) - \nabla B(\widehat{\theta}) \quad = \quad [\widehat{\mu}^n - \mu^*] - \lambda^n \partial R(\widehat{\theta}^n). \tag{19}$$

Taking inner products with the difference $\widehat{\theta}^n - \widehat{\theta}$ yields

$$0 \stackrel{(a)}{\leq} \left[\nabla B(\widehat{\theta}^n) - \nabla B(\widehat{\theta})\right]^T \left[\widehat{\theta}^n - \widehat{\theta}\right] \leq [\widehat{\mu}^n - \mu^*]^T \left[\widehat{\theta}^n - \widehat{\theta}\right] + \lambda^n \partial R(\widehat{\theta}^n)^T \left[\widehat{\theta} - \widehat{\theta}^n\right], \tag{20}$$

where inequality (a) follows from the convexity of B. From the convexity and non-negativity of R, we have

$$\lambda^n \partial R(\widehat{\theta}^n)^T \left[\widehat{\theta} - \widehat{\theta}^n\right] \leq \lambda^n \left[R(\widehat{\theta}) - R(\widehat{\theta}^n)\right] \leq \lambda^n R(\widehat{\theta}).$$

Applying this inequality and Cauchy-Schwartz to Equation (20) yields

$$0 \leq \left[\nabla B(\widehat{\theta}^n) - \nabla B(\widehat{\theta})\right]^T \left[\frac{\widehat{\theta}^n - \widehat{\theta}}{\|\widehat{\theta}^n - \widehat{\theta}\|}\right] \leq \|\widehat{\mu}^n - \mu^*\| + \lambda^n R(\widehat{\theta})$$

Since $\lambda^n = o(1)$ by assumption and $\|\widehat{\mu}^n - \mu^*\| = o_p(1)$ by the weak law of large numbers, the quantity $\left[\nabla B(\widehat{\theta}^n) - \nabla B(\widehat{\theta})\right]^T \left[\frac{\widehat{\theta}^n - \widehat{\theta}}{\|\widehat{\theta}^n - \widehat{\theta}\|}\right]$ converges in probability to zero. By the strict convexity of B, this fact implies that $\widehat{\theta}^n$ converges in probability to $\widehat{\theta}$, thereby completing the proof of part (a).

To establish part (b), we observe that $\sqrt{n}[\widehat{\mu}^n - \mu^*] \xrightarrow{d} N(0, \nabla^2 A(\theta^*))$ by the central limit theorem. Using this fact and applying the delta method to Equation (19) yields that

$$\sqrt{n}\nabla^2 B(\widehat{\theta}) \left[\widehat{\theta}^n - \widehat{\theta}\right] \xrightarrow{d} N\left(0, \nabla^2 A(\theta^*)\right),$$

where we have used the fact that $\sqrt{n}\lambda^n = o(1)$. The strict convexity of B guarantees that $\nabla^2 B(\widehat{\theta})$ is invertible, so that claim (b) follows. ■

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

5.2 Global 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 sum-product message-passing algorithm, although stable for weakly coupled MRFs (Ihler et al., 2005; Mooij and Kappen, 2005b,a; Tatikonda and Jordan, 2002; Tatikonda, 2003), can be highly unstable in other regimes due to the appearance of multiple local optima in the non-convex Bethe problem. However, previous experimental work has shown that methods based on convex relaxations, including the reweighted sum-product (or belief propagation) algorithm (Wainwright et al., 2003b), reweighted generalized BP (Wiegerinck, 2005), and log-determinant relaxations (Wainwright and Jordan, 2005) appear to be *globally stable*—that is, even for very strongly coupled problems. For instance, Figure 2 provides a simple illustration of the instability of the ordinary sum-product algorithm, contrasted with the stability of the tree-reweighted updates. Wiegerinck (2005) provides similar results for reweighted forms of the generalized belief propagation. 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. This stability property plays a fundamental role in providing a performance guarantee on joint estimation/prediction methods.

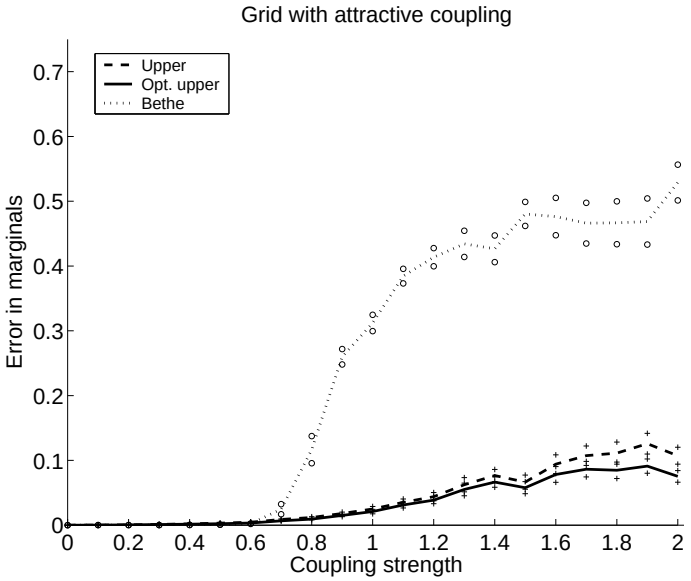


Figure 2: Contrast of the instability of the ordinary sum-product algorithm with the stability of the tree-reweighted version (Wainwright et al., 2005). Results shown with a grid with $N = 100$ nodes over a range of attractive coupling strengths. The ordinary sum-product undergoes a phase transition, after which the quality of marginal approximations degrades substantially. The tree-reweighted algorithm, shown for two different settings of the edge weights ρ_{st}, remains stable over the full range of coupling strengths. See Wainwright et al. (2005) for full details.

We begin by noting that for a multinomial Markov random field (3), the computation of the exact marginal probabilities is a globally Lipschitz operation:

Lemma 4 *For any discrete Markov random field* (3)*, there is a constant $L < +\infty$ such that*

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

This lemma, which is proved in Appendix B, guarantees that small changes in the problem parameters—that is, "perturbations" δ—lead to correspondingly small changes in the computed marginals.

Our goal is to establish analogous Lipschitz properties for variational methods. In particular, it turns out that any variational method based on a suitably concave entropy approximation satisfies such a stability condition. More precisely, 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 having the eigenspectrum of the Hessian $\nabla^2 f(x)$ be uniformly bounded below by c. With this definition, we have:

Proposition 5 *Consider any strictly convex surrogate B based on a strongly concave entropy approximation $-B^*$. 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 From Proposition 2, 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 $\gamma \in \mathbb{R}^d$. Since B and B^* are a strictly convex Legendre pair, we have $\nabla^2 B(\theta) = [\nabla^2 B^*(\tau(\theta))]^{-1}$. By the strong convexity of B^*, we are guaranteed that the spectral norm of $\nabla^2 B^*(\tau)$ is uniformly bounded away from zero, which yields the claim. ■

A number of existing entropy approximations can be shown to be strongly concave. In Appendix C, we provide a detailed proof of this fact for the convexified Bethe entropy (12).

Lemma 6 *For any set $\{\rho_{st}\}$ of strictly positive edge appearance probabilities, the convexified Bethe entropy* (12) *is strongly concave.*

We note that the same argument can be used to establish strong concavity for the reweighted Kikuchi approximations studied by Wiegerinck (2005). Moreover, it can be shown that the Gaussian-based log-determinant relaxation proposed by Wainwright and Jordan (2006) is also strongly concave. For all of these variational methods, then, Proposition 5 guarantees that the pseudomarginal computation is globally Lipschitz stable, thereby providing theoretical confirmation of previous experimental results (Wiegerinck, 2005; Wainwright et al., 2005; Wainwright and Jordan, 2006). The entropy approximations that underlie other variational methods (e.g., expectation-propagation Minka, 2001) can also be modified so as to be strongly concave; Proposition 5 provides further justification—in addition to the obvious benefit of unique global optima—for such "convexification" of entropy approximations.

6. Performance Bounds

In this section, we develop theoretical bounds on the performance loss of our approximate approach to joint estimation and prediction, relative to the unattainable Bayes optimum. So as not to unnecessarily complicate the result, we focus on the performance loss in the infinite data limit[2] (i.e., for which the number of samples $n = +\infty$).

In the infinite data setting, the Bayes optimum is unattainable for two reasons:

1. it is based on knowledge of the exact parameter θ^*, which is not easy to obtain.

2. it assumes (in the prediction phase) that computing exact marginal probabilities μ of the Markov random field is feasible.

Of these two difficulties, it is the latter assumption—regarding the computation of marginal probabilities—that is the most serious. As discussed earlier, there do exist computationally tractable estimators of θ^* that are consistent though not statistically efficient under appropriate conditions; one example is the pseudolikelihood method (Besag, 1975, 1977) mentioned previously. On the other hand, MCMC methods may be used to generate stochastic approximations to marginal probabilities, but may require greater than polynomial complexity.

Recall from Proposition 3 that the parameter estimator based on the surrogate likelihood ℓ_B is *inconsistent*, in the sense that the parameter vector $\widehat{\theta}$ returned in the limit of infinite data is generally not equal to the true parameter θ^*. Our analysis in this section will demonstrate that this inconsistency is beneficial.

6.1 Problem Set-up

Although the ideas and techniques described here are more generally applicable, we focus here on a special observation model so as to obtain a concrete result.

Observation model: In particular, we assume that the multinomial random vector $X = \{X_s,\ s \in V\}$ defined by the Markov random field (3) is a label vector for the components in a finite mixture of Gaussians. For each node $s \in V$, we specify a new random variable Z_s by the conditional distribution

$$p(Z_s = z_s \,|\, X_s = j) \sim N(\nu_j, \sigma_j^2) \quad \text{for } j \in \{0, 1, \ldots, m-1\},$$

so that Z_s is a mixture of m Gaussians. Such Gaussian mixture models are widely used in spatial statistics as well as statistical signal and image processing (Crouse et al., 1998; Ripley, 1981; Titterington et al., 1986).

Now suppose that we observe a noise-corrupted version of z_s—namely, a vector Y of observations with components of the form

$$Y_s = \alpha Z_s + \sqrt{1-\alpha^2} W_s, \tag{21}$$

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. Note that $\alpha = 0$ corresponds to pure noise, whereas $\alpha = 1$ corresponds to completely uncorrupted observations.

2. Note, however, that modified forms of the results given here, modulo the usual $O(1/n)$ corrections, hold for the finite data setting.

Optimal prediction: Our goal is to compute an optimal estimate $\widehat{z}(y)$ of z as a function of the observation $Y = y$, using the mean-squared error as the risk function. The essential object in this computation is the posterior distribution $p(x \mid y; \theta^*) \propto p(x; \theta^*)\, p(y \mid x)$, where the conditional distribution $p(y \mid x)$ is defined by the observation model (21). As shown in the sequel, the posterior distribution (with y fixed) can be expressed as an exponential family member of the form $\theta^* + \gamma(y)$ (see Equation (26a)). Disregarding computational cost, it is straightforward to show that the optimal Bayes least squares estimator (BLSE) takes the form

$$\widehat{z}_s^{\text{opt}}(Y; \theta^*) \quad := \quad \sum_{j=0}^{m-1} \mu_{s;j}(\theta^* + \gamma(Y)) \Big[\omega_j(\alpha) \big(Y_s - \alpha \nu_j \big) + \nu_j \Big], \tag{22}$$

where $\mu_{s;j}(\theta^* + \gamma)$ denotes the marginal probability associated with the posterior distribution $p(x; \theta^* + \gamma)$, and

$$\omega_j(\alpha) \quad := \quad \frac{\alpha \sigma_j^2}{\alpha^2 \sigma_j^2 + (1 - \alpha^2)} \tag{23}$$

is the usual BLSE weighting for a Gaussian with variance σ_j^2.

Approximate prediction: Since the marginal distributions $\mu_{s;j}(\theta^* + \gamma)$ are intractable to compute exactly, it is natural to consider an approximate predictor, based on a set τ of pseudomarginals computed from a variational relaxation. More explicitly, we run the variational algorithm on the parameter vector $\widehat{\theta} + \gamma$ that is obtained by combining the new observation y with the fitted model $\widehat{\theta}$, and use the outputted pseudomarginals $\tau_s(\cdot; \widehat{\theta} + \gamma)$ as weights in the approximate predictor

$$\widehat{z}_s^{\text{app}}(Y; \widehat{\theta}) \quad := \quad \sum_{j=0}^{m-1} \tau_{s;j}(\widehat{\theta} + \gamma(Y)) \Big[\omega_j(\alpha) \big(Y_s - \alpha \nu_j \big) + \nu_j \Big], \tag{24}$$

where the weights ω are defined in Equation (23).

We now turn to a comparison of the Bayes least-squares estimator (BLSE) defined in Equation (22) to the surrogate-based predictor (24). Since (by definition) the BLSE is optimal for the mean-squared error (MSE), using the surrogate-based predictor will necessarily lead to a larger MSE. Our goal is to prove an upper bound on the maximal possible increase in this MSE, where the bound is specified in terms of the underlying model θ^* and the SNR parameter α. More specifically, for a given problem, we define the mean-squared errors

$$\mathrm{R}^{\text{opt}}(\alpha, \theta^*) := \frac{1}{N} \mathbb{E} \| \widehat{z}^{\text{opt}}(Y; \theta^*) - Z \|^2, \quad \text{and} \quad \mathrm{R}^{\text{app}}(\alpha, \widehat{\theta}) := \frac{1}{N} \mathbb{E} \| \widehat{z}^{\text{app}}(Y; \widehat{\theta}) - Z \|^2,$$

of the Bayes-optimal and surrogate-based predictors respectively, where the expectation is taken over the joint distribution of (Y, Z). We seek upper bounds on the increase $\Delta \mathrm{R}(\alpha, \theta^*, \widehat{\theta}) := \mathrm{R}^{\text{app}}(\alpha, \widehat{\theta}) - \mathrm{R}^{\text{opt}}(\alpha, \theta^*)$ of the approximate predictor relative to Bayes optimum.

6.2 Role of Stability

Before providing a technical statement and proof, we begin with some intuition underlying the bounds, and the role of Lipschitz stability. First, consider the low SNR regime ($\alpha \approx 0$) in which the observation Y is heavily corrupted by noise. In the limit $\alpha = 0$, the new observations are pure

noise, so that the prediction of Z should be based simply on the estimated model—namely, the true model $p(\cdot;\theta^*)$ in the Bayes optimal case, and the "incorrect" model $p(\cdot;\widehat{\theta})$ for the method based on surrogate likelihood. The key point here is the following: by properties of the MLE and surrogate-based estimator, the following equalities hold:

$$\nabla A(\theta^*) \stackrel{(a)}{=} \mu(\theta^*) \stackrel{(b)}{=} \mu^* \stackrel{(c)}{=} \tau(\widehat{\theta}) \stackrel{(d)}{=} \nabla B(\widehat{\theta}).$$

Here equality (a) follows from Lemma 1, whereas equality (b) follows from the moment-matching property of the MLE in exponential families. Equalities (c) and (d) hold from the Proposition 2 and the pseudomoment-matching property of the surrogate-based parameter estimator (see proof of Proposition 3). As a key consequence, it follows that the combination of surrogate-based estimation and prediction is *functionally indistinguishable* from the Bayes-optimal behavior in the limit of $\alpha = 0$. More specifically, in the limiting case, the errors systematically introduced by the inconsistent learning procedure are cancelled out exactly by the approximate variational method for computing marginal distributions. Of course, exactness for $\alpha = 0$ is of limited interest; however, when combined with the Lipschitz stability ensured by Proposition 5, it allows us to gain good control of the low SNR regime. At the other extreme of high SNR ($\alpha \approx 1$), the observations are nearly perfect, and hence dominate the behavior of the optimal estimator. More precisely, for α close to 1, we have $\omega_j(\alpha) \approx 1$ for all $j = 0, 1, \ldots, m-1$, so that $\widehat{z}^{\text{opt}}(Y;\theta^*) \approx Y \approx \widehat{z}^{\text{app}}(Y;\widehat{\theta})$. Consequently, in the high SNR regime, accuracy of the marginal computation has little effect on the accuracy of the predictor.

6.3 Bound on Performance Loss

Although bounds of this nature can be developed in more generality, for simplicity in notation we focus here on the case of $m = 2$ mixture components. We begin by introducing the factors that play a role in our bound on the performance loss $\Delta R(\alpha, \theta^*, \widehat{\theta})$. First, the Lipschitz stability enters in the form of the quantity:

$$L(\theta^*;\widehat{\theta}) \;:=\; \sup_{\delta \in \mathbb{R}^d} \sigma_{\max}\big(\nabla^2 A(\theta^* + \delta) - \nabla^2 B(\widehat{\theta} + \delta)\big), \tag{25}$$

where $\sigma_{\max}$ denotes the maximal singular value. Following the argument in the proof of Proposition 5, it can be seen that $L(\theta^*;\widehat{\theta})$ is finite.

Second, in order to apply the Lipschitz stability result, it is convenient to express the effect of introducing a new observation vector y, drawn from the additive noise observation model (21), as a perturbation of the exponential parameterization. In particular, for any parameter $\theta \in \mathbb{R}^d$ and observation y from the model (21), the conditional distribution $p(x|y;\theta)$ can be expressed as $p(x;\theta + \gamma(y,\alpha))$, where the exponential parameter $\gamma(y,\alpha)$ has components[3]

$$\gamma_s = \frac{1}{2}\left\{ \log \frac{\alpha^2\sigma_0^2 + (1-\alpha^2)}{\alpha^2\sigma_1^2 + (1-\alpha^2)} + \frac{(y_s - \alpha\nu_0)^2}{\alpha^2\sigma_0^2 + (1-\alpha^2)} - \frac{(y_s - \alpha\nu_1)^2}{\alpha^2\sigma_1^2 + (1-\alpha^2)} \right\} \qquad \forall\, s \in V. \tag{26a}$$

$$\gamma_{st} = 0 \qquad \forall\, (s,t) \in E. \tag{26b}$$

See Appendix D for a derivation of these relations.

3. For consistency in notation with the general $m > 2$ case, these components should be labeled as $\gamma_{s;1}$ and $\gamma_{st;11}$, but we drop the additional indices for simplicity.

Third, it is convenient to have short notation for the Gaussian estimators of each mixture component:

$$g_j(Y_s;\alpha) \quad := \quad \omega_j(\alpha)\,(Y_s - \alpha\nu_j) + \nu_j \qquad \text{for } j = 0,1,$$

With this notation, we have the following

Theorem 7 *The MSE increase* $\Delta\mathrm{R}(\alpha,\theta^*,\widehat{\theta}) := \mathrm{R}(\alpha,\widehat{\theta}) - \mathrm{R}(\alpha,\theta^*)$ *is upper bounded by*

$$\Delta\mathrm{R}(\alpha,\theta^*,\widehat{\theta}) \quad \leq \quad \mathbb{E}\left\{ \min\left(1,\, L(\theta^*;\widehat{\theta})\frac{\|\gamma(Y;\alpha)\|_2}{\sqrt{N}}\right) \sqrt{\frac{\sum_{s=1}^N |g_1(Y_s) - g_0(Y_s)|^4}{N}} \right\}. \qquad (27)$$

Before proving the bound (27), we illustrate it by considering its behavior in some special cases.

6.3.1 Superiority to True Model

Theorem 7 can be used to establish that applying an approximate message-passing algorithm to the "incorrect" model yields prediction results superior to those obtained by applying the same message-passing algorithm to the true underlying model. To see one regime in which this claim is true, consider the low SNR limit in which $\alpha \to 0^+$. In this limit, it can be seen that $\|\gamma(Y;\alpha)\| \to 0$, so that the the overall bound $\Delta\mathrm{R}(\alpha)$ tends to zero. That is, the combination of approximate estimation and approximate prediction is asymptotically optimal in the low SNR limit. In sharp contrast, this claim need not be true when approximate prediction is applied to the true underlying model. As a particular example, consider a Gaussian mixture with $m = 2$ components, with equal variances but distinct means (say $\nu_0 = -1$ and $\nu_1 = 1$). Moreover, suppose that the mixture indicator vectors $X \in \{0,1\}^N$ are sampled from an underlying distribution $p(x;\theta^*)$, and let $\mu_s = [\mu_{s;0}\ \mu_{s;1}]$ denote the marginal distributions associated with this underlying model. In the limit of zero SNR (i.e., $\alpha = 0$), it is straightforward to see that the BLSE of Z is simply its mean, given (for component $s \in V$) by

$$\mathbb{E}[Z_s \mid Y] \;=\; \mathbb{E}[Z_s] \quad = \quad \mu_{s;0}\nu_0 + \mu_{s;1}\nu_1 \;=\; \mu_{s;1} - \mu_{s;0},$$

for the two-component mixture specified above. Now suppose that *when applied to this true model, the approximate message-passing algorithm yields an incorrect set of singleton pseudomarginals*—say $\tau_s \neq \mu_s$. Since standard message-passing algorithms are rarely (if ever) exactly correct on non-trivial models with cycles, this assumption is more than reasonable. Consequently applying the approximate predictor to the true model will yield an estimate of Z which is incorrect, even in the zero SNR limit; in particular, the approximate estimate is given by

$$\widehat{Z}(Y;\theta^*) \quad = \quad \tau_{s;0}\nu_0 + \tau_{s;1}\nu_1 \;=\; \tau_{s;1} - \tau_{s;0} \;\neq\; \mathbb{E}[Z].$$

Thus, in contrast to the combination of approximate estimation with approximate estimation, applying the approximate message-passing algorithm to the true model fails to be exact even in the limit of zero SNR. In fact, our later experimental results show that the superiority of using the "wrong" model holds for a broader range of SNRs as well (see Figure 4).

We conclude by turning to the high SNR limit as $\alpha \to 1^-$, in which we see that $\omega_j(\alpha) \to 1$ for $j = 0,1$, which drives the differences $|g_1(Y_s) - g_0(Y_s)|$, and in turn the overall bound $\Delta\mathrm{R}(\alpha)$ to zero. Thus, the surrogate-based method is optimal in both the low and high SNR regimes; its behavior in the intermediate regime is governed by the balance between these two terms.

6.3.2 EFFECT OF EQUAL VARIANCES

Now consider the special case of equal variances $\sigma^2 \equiv \sigma_0^2 = \sigma_1^2$, in which case $\omega(\alpha) \equiv \omega_0(\alpha) = \omega_1(\alpha)$. Thus, the difference $g_1(Y_s,\alpha) - g_0(Y_s,\alpha)$ simplifies to $(1-\alpha\omega(\alpha))(\nu_1-\nu_0)$, so that the bound (27) reduces to

$$\Delta R(\alpha,\theta^*,\widehat{\theta}) \;\leq\; (1-\alpha\omega(\alpha))^2(\nu_1-\nu_0)^2\,\mathbb{E}\left\{\min\left(1,\, L(\theta^*;\widehat{\theta})\frac{\|\gamma(Y;\alpha)\|_2}{\sqrt{N}}\right)\right\}. \tag{28}$$

As shown by the simpler expression (28), for $\nu_1 \approx \nu_0$, the MSE increase is very small, since such a two-component mixture is close to a pure Gaussian.

6.3.3 EFFECT OF MEAN DIFFERENCES

Finally consider the case of equal means $\nu \equiv \nu_0 = \nu_1$ in the two Gaussian mixture components. In this case, we have $g_1(Y_s,\alpha) - g_0(Y_s,\alpha) = [\omega_1(\alpha) - \omega_0(\alpha)]\,[Y_s - \alpha\nu]$, so that the bound (27) reduces to

$$\Delta R(\alpha,\theta^*,\widehat{\theta}) \;\leq\; [\omega_1(\alpha)-\omega_0(\alpha)]^2\;\mathbb{E}\left\{\min\left(1,\, L(\theta^*;\widehat{\theta})\frac{\|\gamma(Y;\alpha)\|_2}{\sqrt{N}}\right)\sqrt{\frac{\sum_s (Y_s-\alpha\nu)^4}{N}}\right\}.$$

Here the MSE increase depends on the SNR α and the difference

$$\omega_1(\alpha)-\omega_0(\alpha) \;=\; \frac{\alpha\sigma_1^2}{\alpha^2\sigma_1^2+(1-\alpha^2)} - \frac{\alpha\sigma_0^2}{\alpha^2\sigma_0^2+(1-\alpha^2)} \;=\; \frac{(1-\alpha^2)\,(\sigma_1^2-\sigma_0^2)}{\left[\alpha^2\sigma_0^2+(1-\alpha^2)\right]\,\left[\alpha^2\sigma_1^2+(1-\alpha^2)\right]}.$$

Observe, in particular, that the MSE increases tends to zero as the difference $\sigma_1^2-\sigma_0^2$ decreases, as should be expected intuitively.

6.4 Proof of Theorem 7

We now turn to the proof of the main bound (27). By the Pythagorean relation that characterizes the Bayes least squares estimator $\widehat{z}^{\,\mathrm{opt}}(Y;\theta^*) = \mathbb{E}_{(Z\,|\,Y,\theta^*)}[Z]$, we have

$$\begin{aligned}\Delta R(\alpha;\theta^*,\widehat{\theta}) \;&:=\; \frac{1}{N}\mathbb{E}\|\widehat{z}^{\,\mathrm{app}}(Y;\widehat{\theta}) - Z\|_2^2 - \frac{1}{N}\mathbb{E}\|\widehat{z}^{\,\mathrm{opt}}(Y;\theta^*) - Z\|_2^2 \\ &=\; \frac{1}{N}\mathbb{E}\|\widehat{z}^{\,\mathrm{app}}(Y;\widehat{\theta}) - \widehat{z}^{\,\mathrm{opt}}(Y;\theta^*)\|_2^2.\end{aligned}$$

Using the definitions of $\widehat{z}^{\,\mathrm{app}}(Y;\widehat{\theta})$ and $\widehat{z}^{\,\mathrm{opt}}(Y;\theta^*)$, some algebraic manipulation yields

$$\begin{aligned}\left[\widehat{z}_s^{\,\mathrm{app}}(Y;\widehat{\theta}) - \widehat{z}_s^{\,\mathrm{opt}}(Y;\theta^*)\right]^2 \;&=\; \left[\tau_s(\widehat{\theta}+\gamma) - \mu_s(\theta^*+\gamma)\right]^2\,[g_1(Y_s)-g_0(Y_s)]^2 \\ &\leq\; \left|\tau_s(\widehat{\theta}+\gamma) - \mu_s(\theta^*+\gamma)\right|\,[g_1(Y_s)-g_0(Y_s)]^2,\end{aligned}$$

where the second inequality uses the fact that $|\tau_s - \mu_s| \leq 1$ since τ_s and μ_s are marginal probabilities. Next we write

$$\begin{aligned}\frac{1}{N}\|\widehat{z}^{\,\mathrm{app}}(Y;\widehat{\theta}) - \widehat{z}^{\,\mathrm{opt}}(Y;\theta^*)\|_2^2 \;&\leq\; \frac{1}{N}\sum_{s=1}^N \left|\tau_s(\widehat{\theta}+\gamma) - \mu_s(\theta^*+\gamma)\right|\,[g_1(Y_s)-g_0(Y_s)]^2 \qquad (29)\\ &\leq\; \frac{1}{\sqrt{N}}\|\tau(\widehat{\theta}+\gamma) - \mu(\theta^*+\gamma)\|_2\,\sqrt{\frac{\sum_{s=1}^N |g_1(Y_s)-g_0(Y_s)|^4}{N}}\end{aligned}$$

where the last line uses the Cauchy-Schwarz inequality.

It remains to bound the 2-norm $\|\tau(\widehat{\theta}+\gamma)-\mu(\theta^*+\gamma)\|_2$. An initial naive bound follows from the fact $\tau_s, \mu_s \in [0,1]$ implies that $|\tau_s - \mu_s| \leq 1$, whence

$$\frac{1}{\sqrt{N}}\|\tau-\mu\|_2 \leq \quad 1. \tag{30}$$

An alternative bound, which will be better for small perturbations γ, can be obtained as follows. Using the relation $\tau(\widehat{\theta}) = \mu(\theta^*)$ guaranteed by the definition of the ML estimator and surrogate estimator, we have

$$\begin{aligned} \|\tau(\widehat{\theta}+\gamma)-\mu(\theta^*+\gamma)\|_2 &= \left\| \left[\tau(\widehat{\theta}+\gamma)-\tau(\widehat{\theta})\right] + [\mu(\theta^*)-\mu(\theta^*+\gamma)] \right\|_2 \\ &= \left\| \left[\nabla^2 B(\widehat{\theta}+s\gamma) - \nabla^2 A(\theta^*+t\gamma)\right]\gamma \right\|_2, \end{aligned}$$

for some $s,t \in [0,1]$, where we have used the mean value theorem. Thus, using the definition (25) of L, we have

$$\frac{1}{\sqrt{N}}\,\|\tau(\widehat{\theta}+\gamma)-\mu(\theta^*+\gamma)\|_2 \quad \leq \quad L(\theta^*;\widehat{\theta})\frac{\|\gamma(Y;\alpha)\|_2}{\sqrt{N}}. \tag{31}$$

Combining the bounds (30) and (31) and applying them to Equation (29), we obtain

$$\frac{1}{N}\|\widehat{z}^{\text{app}}(Y;\widehat{\theta}) - \widehat{z}^{\text{opt}}(Y;\theta^*)\|_2^2 \quad \leq \quad \min\left\{1,\, L(\theta^*;\widehat{\theta})\frac{\|\gamma(Y;\alpha)\|_2}{\sqrt{N}}\right\}\sqrt{\frac{\sum_{s=1}^N |g_1(Y_s)-g_0(Y_s)|^4}{N}}.$$

Taking expectations of both sides yields the result.

7. 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. Here we describe both experimental results to quantify the performance loss of the tree-reweighted sum-product algorithm (Wainwright et al., 2005), and compare it to both a baseline independence model, as well as a closely related heuristic method that uses the ordinary sum-product (or belief propagation) algorithm.

7.1 Methods

In Section 4.2, we described a generic procedure for joint estimation and prediction. Here we begin by describing the special case of this procedure when the underlying variational method is the tree-reweighted sum-product algorithm (Wainwright et al., 2005). Any instantiation of the tree-reweighted sum-product algorithm is specified by a collection of edge weights ρ_{st}, one for each edge (s,t) of the graph. The vector of edge weights must belong to the spanning tree polytope; see Wainwright et al. (2005) for further background on these weights and the reweighted algorithm. Given a fixed set of edge weights ρ, the joint procedure based on the tree-reweighted sum-product algorithm consists of the following steps:

1. Given an initial set of i.i.d. data $\{X^1, \ldots, X^n\}$, we first compute the empirical marginal distributions

$$\widehat{\mu}_{s;j} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_s^i = j], \qquad \widehat{\mu}_{st;jk} := \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}[X_s^i = j]\, \mathbb{I}[X_t^i = k],$$

and use them to compute the approximate parameter estimate

$$\widehat{\theta}^n_{s;j} := \log \widehat{\mu}_{s;j}, \qquad \widehat{\theta}^n_{s;j} := \rho_{st} \log \frac{\widehat{\mu}_{st;jk}}{\widehat{\mu}_{s;j}\widehat{\mu}_{t;k}}. \tag{32}$$

As shown in our previous work (Wainwright et al., 2003b), the estimates (32) are the global maxima of the surrogate likelihood (16) based on the convexified Bethe approximation (12) without any regularization term (i.e., $R = 0$).

2. Given the new noisy observation Y of the form (21), we incorporate it by by forming the new exponential parameter

$$\widehat{\theta}^n_s + \gamma_s(Y),$$

where Equation (26a) defines γ_s for the Gaussian mixture model under consideration.

3. We then compute approximate marginals $\tau(\widehat{\theta} + \gamma)$ by running the TRW sum-product algorithm with edge appearance weights ρ_{st}, using the message updates (14), on the graphical model distribution with exponential parameter $\widehat{\theta} + \gamma$. We use the approximate marginals (see Equation (15)) to construct the prediction $\widehat{z}^{\text{app}}$ in Equation (24).

We evaluated the tree-reweighted sum-product based on its increase in mean-squared error (MSE) over the Bayes optimal predictor (22). Moreover, we compared the performance of the tree-reweighted approach to the following alternatives:

(a) As a baseline, we used the *independence model* in which the mixture distributions at each node are all assumed to be independent. In this case, ML estimates of the parameters are given by $\widehat{\theta}_{s;j} = \log \widehat{\mu}_{s;j}$, with all of the coupling terms $\widehat{\theta}_{st;jk}$ equal to zero. The prediction step reduces to computing the Bayes least squares estimate at each node independently, based only on the local data y_s.

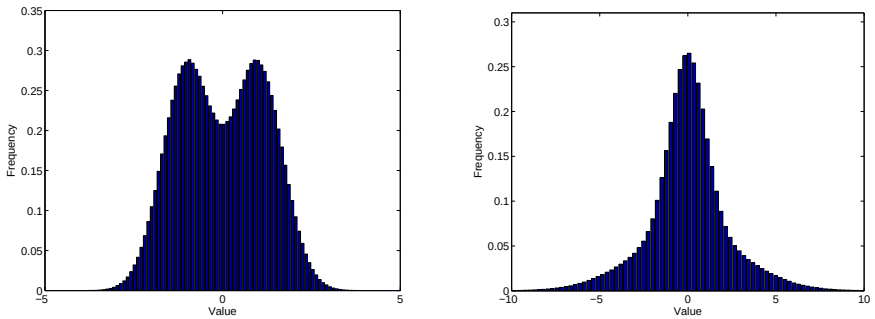
(b) The *standard sum-product or belief propagation* (BP) approach is closely related to the tree-reweighted sum-product method, but based on the edge weights $\rho_{st} = 1$ for all edges. In particular, we first form the approximate parameter estimate $\widehat{\theta}$ using Equation (32) with $\rho_{st} = 1$. As shown in our previous work (Wainwright et al., 2003b), this approximate parameter estimate uniquely defines the Markov random field for which the empirical marginals $\widehat{\mu}_s$ and $\widehat{\mu}_{st}$ are fixed points of the ordinary belief propagation algorithm. We note that a parameter estimator of this type has been used previously by other researchers (Freeman et al., 2000; Ross and Kaebling, 2005). In the prediction step, we then use the ordinary belief propagation algorithm (i.e., again with $\rho_{st} = 1$) to compute approximate marginals of the graphical model with parameter $\widehat{\theta} + \gamma$. Finally, based on these approximate BP marginals, we compute the approximate predictor using Equation (24).

Although our methods are more generally applicable, here we show representative results for $m = 2$ components, and two different types of Gaussian mixtures.

(a) Mixture ensemble A is bimodal, with components $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$.

(b) Mixture ensemble B was constructed with mean and variance components $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_1, \sigma_1^2) = (0, 9)$; these choices serve to mimic heavy-tailed behavior.

In both cases, each mixture component is equally weighted; see Figure 3 for histograms of the resulting mixture ensembles.



(a) (b)

Figure 3: Histograms of different Gaussian mixture ensembles. (a) Ensemble A: a bimodal ensemble with $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$. (b) Ensemble B: mimics a heavy-tailed distribution, with $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_1, \sigma_1^2) = (0, 9)$.

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 (except those in Section 7.2), we chose $\theta_s^* = 0$ for all nodes $s \in V$, which ensures uniform marginal distributions $p(x_s; \theta^*) = [0.5 \;\; 0.5]^T$ at each node. We tested two types of coupling in the underlying Markov random field:

(a) In the case of *attractive coupling*, for each coupling strength $\beta \in [0, 1]$, we chose edge parameters as $\theta_{st}^* \sim \mathcal{U}[0, \beta]$.

(b) In the case of *mixed coupling*, for each coupling strength $\beta \in [0, 1]$, we chose edge parameters as $\theta_{st}^* \sim \mathcal{U}[-\beta, \beta]$.

Here $\mathcal{U}[a, b]$ denotes a uniform distribution on the interval $[a, b]$. In all cases, we varied the SNR parameter α, as specified in the observation model (21), in the interval $[0, 1]$.

7.2 Comparison between "Incorrect" and True Model

We begin with an experimental comparison to substantiate our earlier claim that applying an approximate message-passing algorithm to the "incorrect" model yields prediction results superior to

those obtained by applying the same message-passing algorithm to the true underlying model. As discussed earlier in Section 6.3.1, for any underlying model $p(x;\theta^*)$ in which approximate message-passing yields the incorrect marginals (without any additional observations), there exists a range of SNR around $\alpha \approx 0$ for which this superior performance will hold.

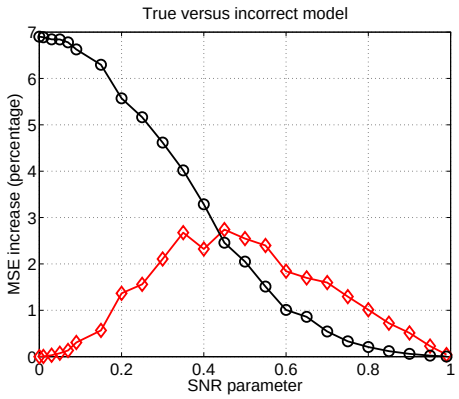


Figure 4: Line plots of percentage increase in MSE relative to Bayes optimum for the TRW method applied to the true model (black circles) versus the approximate model (red diamonds) as a function of observation SNR for grids with $N = 64$ nodes, and attractive coupling $\beta = 0.70$. As predicted by theory, using the "incorrect" model leads to superior performance, when prediction is performed using the approximate TRW method, for a range of SNR.

Figure 4 provides an empirical demonstration of this claim, when the TRW algorithm for prediction is applied to a grid with $N = 64$ nodes and attractive coupling strength $\beta = 0.70$, and the node observations chosen randomly as $\theta_s^* \sim N(0, 0.5)$. Plotted versus the SNR parameter α is the percentage increase in MSE performance relative to the Bayes optimal baseline. Note that for all SNR parameters up to $\alpha \approx 0.40$, applying the TRW algorithm to the true model yields worse performance than applying it to the "incorrect model". Beyond this point, the pattern reverses, but any differences between the two methods are rather small for $\alpha > 0.40$.

7.3 Comparison between Tree-reweighted and Ordinary Sum-product

We now compare the performance of the prediction method based on tree-reweighted sum-product (TRW) message-passing to that based on ordinary sum-product or belief propagation (BP) message-passing. Shown in Figure 5 are 2-D surface plots of the average percentage increase in MSE, taken over 100 trials, as a function of the coupling strength $\beta \in [0,1]$ 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 top two rows show performance for attractive coupling, for mixture ensemble A ((a) through (c)) and ensemble B ((d) through (f)), whereas the bottom two row show performance for mixed coupling, for mixture ensemble A ((g) through (i)) and ensemble B ((j) through (l)).

First, observe that for weakly coupled problems ($\beta \approx 0$), whether attractive or mixed coupling, all three methods—including the independence model—perform quite well, as should be expected

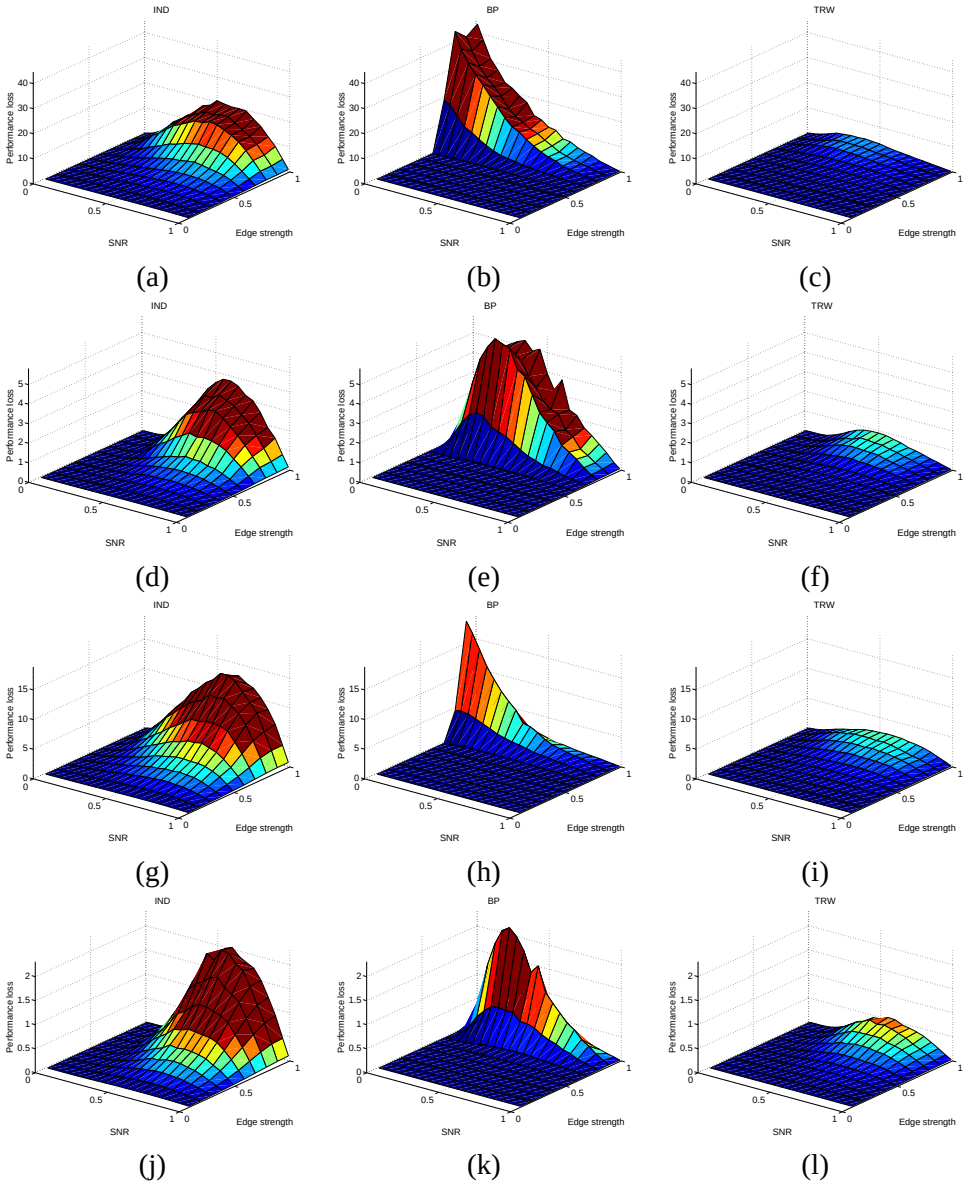


Figure 5: Surface plots of the percentage increase in MSE relative to Bayes optimum for different methods as a function of observation SNR for grids with $N = 64$ nodes. Left column: independence model (IND). Center column: ordinary belief propagation (BP). Right column: tree-reweighted algorithm (TRW). First row: Attractive coupling and a Gaussian mixture with components $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$. Second row: Attractive coupling and a Gaussian mixture with components $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_0, \sigma_1^2) = (0, 9)$. Third row: Mixed coupling and a Gaussian mixture with components $(\nu_0, \sigma_0^2) = (-1, 0.5)$ and $(\nu_1, \sigma_1^2) = (1, 0.5)$. Fourth row: Mixed coupling and a Gaussian mixture with components $(\nu_0, \sigma_0^2) = (0, 1)$ and $(\nu_0, \sigma_1^2) = (0, 9)$.

given the weak dependency between different nodes in the Markov random field. Although not clear in these plots, the standard BP method outperforms the TRW-based method for weak coupling; however, both methods lose less 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 (IND) model. This deterioration is particularly severe for the case of mixture ensemble A with attractive coupling, where the percentage loss in BP can be as high as 50%. Note that the degradation is *not* caused by failure of the BP algorithm to converge. Rather, by looking at alternative models (in which phase transitions are known), we have found that this type of rapid degradation coincides with the appearance of multiple fixed points for the BP algorithm. In contrast, the behavior of the TRW method is extremely stable, which is consistent with our theoretical results.

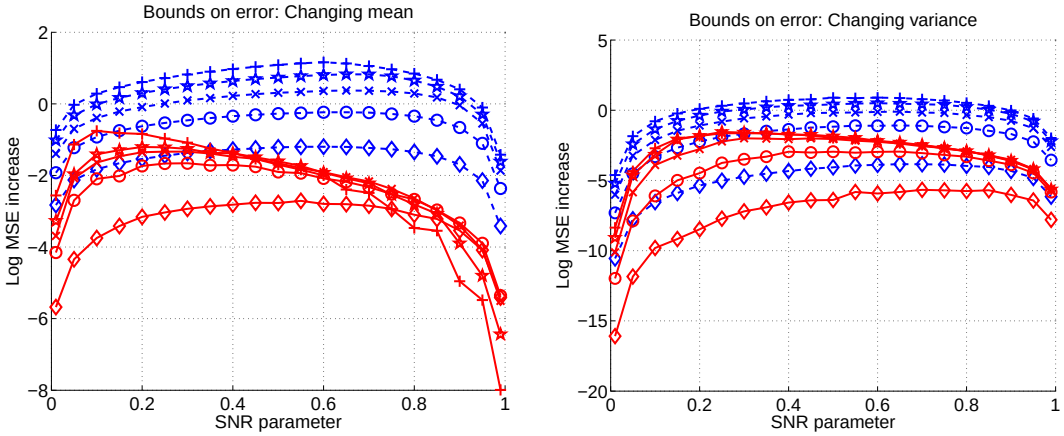
7.4 Comparison between Theory and Practice

We now compare the practical behavior of the tree-reweighted sum-product algorithm to the theoretical predictions from Theorem 7. In general, we have found that in quantitative terms, the bounds (27) are rather conservative—in particular, the TRW sum-product method performs much better than the bounds would predict. However, here we show how the bounds can capture qualitative aspects of the MSE increase in different regimes.

Figure 6 provides plots of the actual MSE increase for the TRW algorithm (solid red lines), compared to the theoretical bound (27) (dotted blue lines), for the grid with $N = 64$ nodes, and attractive coupling of strength $\beta = 0.70$. For all comparisons in both panels, we used $L = 0.10$, which numerical calculations showed to be a reasonable choice for this coupling strength. (Overall, changes in the constant L primarily cause the bounds to shift up and down on the log scale, and so do not overly affect the qualitative comparisons given here.) Panel (a) provides the comparison ensembles of type A, with fixed variances $\sigma_0^2 = \sigma_1^2 = 0.5$ and mean vectors (ν_0, ν_1) ranging from $(-0.5, 0.5)$ to $(-2.5, 2.5)$. Note how the bounds capture the qualitative behavior for low SNR, for which the difficulty of the problem increases as the mean separation is increased. In contrast, in the high SNR regime, the bounds are extremely conservative, and fail to predict that the sharp drop-off in error as the SNR parameter α approaches one. This drop-off is particularly pronounced for the ensemble with largest mean separation (marked with $+$). Panel (b) provides a similar comparison for ensembles of type B, with fixed mean vectors $\nu_0 = \nu_1 = 0$, and variances (σ_0^1, σ_1^2) ranging from $(1, 1.25)$ to $(1, 25)$. In this case, although the bounds are still very conservative in quantitative terms, they reasonably capture the qualitative behavior of the error over the full range of SNR.

8. Discussion

Key challenges in the application of Markov random fields include the estimation (learning) of model parameters, and performing prediction using noisy samples (e.g., smoothing, interpolation, denoising). Both of these problems present substantial computational challenges for general Markov random fields. In this paper, we have described and analyzed methods for joint estimation and prediction that are based on convex variational methods. Our central result is that using inconsistent parameter estimators can be beneficial in the computation-limited setting. Indeed, our results provide rigorous confirmation of the fact that using parameter estimates that are "systematically incorrect" is helpful in offsetting the error introduced by using an approximate method during the prediction step. Moreover, our analysis establishes an additional benefit—aside from the obvious one of ensuring



(a) (b)

Figure 6: Comparison of actual MSE increase and upper bounds for grid with $N = 64$ nodes with attractive coupling. (a) Equal variances $\sigma_0^2 = \sigma_1^2 = 0.5$, and mean vectors (ν_0, ν_1) ranging from $(-0.5, 0.5)$ to $(-2.5, 2.5)$. (b) Equal mean vectors $\nu_0 = \nu_1 = 0$, and variances (σ_0^2, σ_1^2) ranging from $(1, 1.25)$ to $(1, 25)$.

unique global optima—to using variational methods based on convex approximations. In particular, we established a global Lipschitz stability property that applies to any message-passing algorithm that is based on a strongly concave entropy approximation. This type of global stability is a key consideration for algorithms that are applied to models estimated from data, or in which errors might be introduced during message-passing (e.g., due to quantization or other forms of communication constraints). Our empirical results showed that a joint prediction/estimation method using the tree-reweighted sum-product algorithm yields good performance across a wide range of experimental conditions. Although our work has focused on a particular scenario, we suspect that similar ideas and techniques will be useful in related applications of approximate methods for learning combined with prediction and/or classification.

Acknowledgments

We thank the reviewers for helpful comments and suggestions to improve the initial draft of this manuscript. We would like to acknowledge support for this project from the National Science Foundation (NSF Grant DMS-0528488), an Alfred P. Sloan Foundation Fellowship, an Okawa Foundation Research Fellowship, and an Intel Corporation Equipment Grant.

Appendix A. Tree-Based Relaxation

As an illustration on the single cycle on 3 vertices, the pseudomarginal vector with elements

$$\tau_s(x_s) = \begin{bmatrix} 0.5 \\ 0.5 \end{bmatrix} \quad \text{for } s = 1,2,3 \quad \text{and} \quad \tau_{st}(x_s, x_t) = \begin{bmatrix} \alpha_{st} & 0.5 - \alpha_{st} \\ 0.5 - \alpha_{st} & \alpha_{st} \end{bmatrix}$$

belongs to $\mathrm{LOCAL}_\phi(G)$ for all choices $\alpha_{st} \in [0, 0.5]$, but fails to belong to $\mathrm{MARG}_\phi(G)$, for instance, when $\alpha_{12} = \alpha_{23} = \alpha_{13} = 0$.

Appendix B. Proof of Lemma 4

Using Lemma 1 and the mean value theorem, we write

$$\begin{aligned} \mu(\theta+\delta) - \mu(\theta) &= \nabla A(\theta+\delta) - \nabla A(\theta) \\ &= \nabla^2 A(\theta + t\delta)\delta \end{aligned}$$

for some $t \in (0,1)$. Hence, it suffices to show that the eigenspectrum of the Hessian $\nabla^2 A(\theta) = \mathrm{cov}_\theta\{\phi(X)\}$ is uniformly bounded above by $L < +\infty$. The functions ϕ are all 0-1 valued indicator functions, so that the diagonal elements of $\mathrm{cov}_\theta\{\phi(X)\}$ are bounded above—in particular, $\mathrm{var}(\phi_\alpha(X)) \leq \frac{1}{4}$ for any index $\alpha \in \{1, \ldots, d\}$. Consequently, we have

$$\lambda_{\max}(\mathrm{cov}_\theta\{\phi(X)\}) \leq \sum_{\alpha=1}^{d} \lambda_\alpha(\mathrm{cov}_\theta\{\phi(X)\} = \mathrm{trace}(\mathrm{cov}_\theta\{\phi(X)\} = \frac{d}{4}$$

as required.

Appendix C. Proof of Lemma 6

Consider a spanning tree T of G with edge set $E(T)$. Given a vector $\tau \in \mathrm{LOCAL}_\phi(G)$, we associate with T a subvector $\tau(T)$ formed by those components of τ associated with vertices V and edges $E(T)$. Note that by construction $\tau(T) \in \mathrm{LOCAL}_\phi(T) = \mathrm{MARG}_\phi(T)$. The mapping $\tau \mapsto \tau(T)$ can be represented by a projection matrix $\Pi^T \in \mathbb{R}^{d(T) \times d}$ with the block structure

$$\Pi^T := \begin{bmatrix} I_{d(T)\times d(T)} & 0_{d(T)\times(d-d(T))} \end{bmatrix}.$$

In this definition, we are assuming for convenience that τ is ordered such that the $d(T)$ components corresponding to the tree T are placed first. With this notation, we have $\Pi^T \tau = \begin{bmatrix} \tau(T) & 0 \end{bmatrix}'$.

By our construction of the function B_ρ, there exists a probability distribution $\rho := \{\rho(T) \mid T \in \mathfrak{T}\}$ such that $B_\rho(\tau) = \sum_{T \in \mathfrak{T}} \rho(T) A^*(\tau(T))$, where $A^*(\tau(T))$ denotes the negative entropy of the tree-structured distribution defined by the vector of marginals $\tau(T)$. Hence, the Hessian of B_ρ has the decomposition

$$\nabla^2 B_\rho(\tau) = \sum_{T \in \mathfrak{T}} \rho(T) (\Pi^T)' \nabla^2 A^*(\tau(T)) (\Pi^T). \tag{33}$$

To check dimensions of the various quantities, note that $\nabla^2 A^*(\tau(T))$ is a $d(T) \times d(T)$ matrix, and recall that each matrix $\Pi^T \in \mathbb{R}^{d(T)\times d}$.

Now by Lemma 4, the eigenvalues of the $\nabla^2 A$ are uniformly bounded above; hence, the eigenvalues of $\nabla^2 A^*$ are uniformly bounded away from zero. Hence, for each tree T, there exists a constant C_T such that for all $z \in \mathbb{R}^d$

$$z'(\Pi^T)' \nabla^2 A^*(\tau(T))(\Pi^T) z \geq C_T \|\Pi^T z\|^2 = C_T \|z_T\|^2,$$

where we use $z_T = \Pi^T z$ as a shorthand for the projection of z onto the indices associated with T. Substituting this relation into our decomposition (33) and expanding the sum over T yields

$$\begin{aligned} z'\nabla^2 B_\rho(\tau) z &\geq \sum_{T \in \mathfrak{T}} \rho(T) C_T \|z_T\|^2 \\ &= \Big[\sum_{T \in \mathfrak{T}} \rho(T) C_T\Big] \sum_{s \in V} \|z_s\|^2 + \sum_{(s,t) \in E} \Big[\sum_{T \in \mathfrak{T}} \rho(T) C_T \mathbb{I}[(s,t) \in E(T)]\Big] \|z_{st}\|^2. \end{aligned} \quad (34)$$

Defining $C^* := \min_{T \in \mathfrak{T}} C_T$, we have the lower bounds

$$\begin{aligned} \Big[\sum_{T \in \mathfrak{T}} \rho(T) C_T\Big] &\geq C^* \sum_{T \in \mathfrak{T}} \rho(T) = C^* > 0 \\ \sum_{T \in \mathfrak{T}} \rho(T) C_T \mathbb{I}[(s,t) \in E(T)] &\geq C^* \sum_{T \in \mathfrak{T}} \rho(T) \mathbb{I}[(s,t) \in E(T)] = C^* \rho_{st} \geq C^* \rho^* > 0, \end{aligned}$$

where $\rho^* := \min_{(s,t) \in E} \rho_{st} > 0$. Applying these bounds to Equation (34) yields the final inequality

$$z'\nabla^2 B_\rho(\tau) z \geq C^* \rho^* \|z\|^2 \quad \forall z \in \mathbb{R}^d$$

with $C^* \rho^* > 0$, which establishes that the eigenvalues of $\nabla^2 B_\rho(\tau)$ are bounded away from zero.

Appendix D. Form of Exponential Parameter

Consider the observation model $y_s = \alpha z_s + \sqrt{1-\alpha^2} v_s$, where $v_s \sim N(0,1)$ and z_s is a mixture of two Gaussians (ν_0, σ_0^2) and (ν_1, σ_1^2). Conditioned on the value of the mixing indicator $X_s = j$, the distribution of y_s is Gaussian with mean $\alpha \nu_j$ and variance $\alpha^2 \sigma_j^2 + (1-\alpha^2)$.

Let us focus on one component $p(y_s \,|\, x_s)$ in the factorized conditional distribution $p(y \,|\, x) = \prod_{s=1}^n p(y_s \,|\, x_s)$. For $j = 0, 1$, it has the form

$$p(y_s \,|\, X_s = j) = \frac{1}{\sqrt{2\pi\big[\alpha^2\sigma_j^2 + (1-\alpha^2)\big]}} \exp\Big\{-\frac{1}{2\big[\alpha^2\sigma_j^2 + (1-\alpha^2)\big]}(y_s - \alpha\nu_j)^2\Big\}.$$

We wish to represent the influence of this term on x_s in the form $\exp(\gamma_s x_s)$ for some exponential parameter γ_s. We see that γ_s should have the form

$$\begin{aligned} \gamma_s &= \log p(y_s \,|\, X_s = 1) - \log p(y_s \,|\, X_s = 0) \\ &= \frac{1}{2} \log \frac{\big[\alpha^2\sigma_0^2 + (1-\alpha^2)\big]}{\big[\alpha^2\sigma_1^2 + (1-\alpha^2)\big]} + \frac{(y_s - \alpha\nu_0)^2}{2\big[\alpha^2\sigma_0^2 + (1-\alpha^2)\big]} - \frac{(y_s - \alpha\nu_1)^2}{2\big[\alpha^2\sigma_1^2 + (1-\alpha^2)\big]}. \end{aligned}$$

References

A. Benveniste, M. Metivier, and P. Priouret. *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, New York, NY, 1990.

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

J. Besag. Statistical analysis of non-lattice data. *The Statistician*, 24(3):179–195, 1975.

J. Besag. Efficiency of pseudolikelihood estimation for simple Gaussian fields. *Biometrika*, 64(3): 616–618, 1977.

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

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

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

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

W. T. Freeman and Y. Weiss. On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. *IEEE Trans. Info. Theory*, 47:736–744, 2001.

T. Heskes, K. Albers, and B. Kappen. Approximate inference and constrained optimization. In *Uncertainty in Artificial Intelligence*, volume 13, pages 313–320, July 2003.

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.

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

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

J. S. Liu. *Monte Carlo strategies in Scientific Computing*. Springer-Verlag, New York, NY, 2001.

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

J. M. Mooij and H. J. Kappen. On the properties of the Bethe approximation and loopy belief propagation on binary networks. *Journal of Statistical Mechanics: Theory and Experiment*, P11012: 407–432, 2005a.

J. M. Mooij and H. J. Kappen. Sufficient conditions for convergence of loopy belief propagation. Technical Report arxiv:cs.IT:0504030, University of Nijmegen, April 2005b. Submitted to IEEE Trans. Info. Theory.

S. Della Pietra, V. Della Pietra, and J. Lafferty. Inducing features of random fields. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 19(4):380–393, 1997.

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

B. D. Ripley. *Spatial statistics*. Wiley, New York, 1981.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer-Verlag, New York, NY, 1999.

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

M. Ross and L. P. Kaebling. Learning static object segmentation from motion segmentation. In *20th National Conference on Artificial Intelligence*, 2005.

P. Rusmevichientong and B. Van Roy. An analysis of turbo decoding with Gaussian densities. In *NIPS 12*, pages 575–581. MIT Press, 2000.

R. J. Serfling. *Approximation Theorems of Mathematical Statistics*. Wiley Series in Probability and Statistics. Wiley, 1980.

C. Sutton and A. McCallum. Piecewise training of undirected models. In *Uncertainty in Artificial Intelligence*, July 2005.

S. Tatikonda. Convergence of the sum-product algorithm. In *Information Theory Workshop*, April 2003.

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.

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

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

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

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 2003b.

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.

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

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.

M. J. Wainwright and M. I. Jordan. Log-determinant relaxation for approximate inference in discrete Markov random fields. *IEEE Trans. Signal Processing*, 54(6):2099–2109, June 2006.

Y. Weiss. Correctness of local probability propagation in graphical models with loops. *Neural Computation*, 12:1–41, 2000.

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

J. Yedidia. An idiosyncratic journey beyond mean field theory. In M. Opper and D. Saad, editors, *Advanced mean field methods: Theory and practice*, pages 21–36. MIT Press, 2001.

J.S. 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.

L. Younes. Estimation and annealing for Gibbsian fields. *Ann. Inst. Henri Poincare*, 24(2):269–294, 1988.