One-Shot Marginal MAP Inference in Markov Random Fields

Hao Xiong [*], **Yuanzhen Guo** [*], **Yibo Yang** [*], **and Nicholas Ruozzi**
Erik Jonsson School of Engineering and Computer Science
The University of Texas at Dallas
Richardson, TX 75080

Abstract

Statistical inference in Markov random fields (MRFs) is NP-hard in all but the simplest cases. As a result, many algorithms, particularly in the case of discrete random variables, have been developed to perform approximate inference. However, most of these methods scale poorly, cannot be applied to continuous random variables, or are too slow to be used in situations that call for repeated statistical inference on the same model. In this work, we propose a novel variational inference strategy that is efficient for repeated inference tasks, flexible enough to handle both continuous and discrete random variables, and scalable enough, via modern GPUs, to be practical on MRFs with hundreds of thousands of random variables. We prove that our approach overcomes weaknesses of existing ones and demonstrate its efficacy on both synthetic models and real-world applications.

1 INTRODUCTION

Markov random fields (MRFs) and their conditional variants provide a general approach to probabilistic inference and learning tasks in such diverse domains as artificial intelligence, bioinformatics, and signal and image processing (Koller and Friedman, 2009; Wainwright and Jordan, 2008). MRFs encode local relationships among a collection of random variables, which can then be exploited for fast approximate inference. Given an MRF that has either been specified in advance or learned from data, we will be interested in performing statistical inference, e.g., computing the mode, marginals, or some combination of these, often in the presence of evidence. In typical AI applications these kinds of statistical queries may be performed many times, perhaps conditioned on different evidence each time.

Unfortunately, MRFs have significant practical limitations that have hindered their application on the types of large-scale prediction tasks for which deep neural networks are the current state-of-the-art. First, while neural networks operate over real variables, many of the approximate inference algorithms for MRFs operate only on discrete models. Until recently, this has meant that in order to handle prediction tasks with continuous random variables (or both discrete and continuous random variables) either the state space was discretized, which can be expensive both in terms of space and time complexity, or the set of allowable potential functions was severely restricted. The situation has improved in recent years, and a variety of new approximate inference techniques have been developed for continuous/hybrid MRFs including marginal inference (Minka, 2001; Sudderth et al., 2003; Ihler and McAllester, 2009; Noorshams and Wainwright, 2013; Lienart et al., 2015; Ruozzi, 2017; Wang et al., 2018; Guo et al., 2019), MAP inference (Wang et al., 2014; Pacheco and Sudderth, 2015; Ruozzi, 2015), and joint inference/learning (Song et al., 2011).

However, significant work remains to be done in order to make most of the above algorithms accurate enough to be competitive in practice while also being fast enough to be applicable at the scale of modern neural networks. First, while prediction is typically fast in deep neural networks, prediction in MRFs is often slow and can require solving a new, expensive inference problem every time new evidence is presented. This is especially time consuming for structured prediction, marginal MAP, or general inference on models with latent variables. As a result, the types of applications for which MRFs are a practical tool is somewhat limited. Second, inference using belief propagation style message-passing in continuous models can lead to divergent behavior if the approximation is unbounded, which can happen even in Gaussian graphical models when the precision matrix is not walk-summable

[*] Alphabetical order; equal contribution.

(Malioutov et al., 2006; Ruozzi and Tatikonda, 2013a). This limits the applicability of these methods to specific subsets of continuous graphical models.

In this work, we show that a simple variational approach, based on the same approximations as many of the above message-passing schemes, yields a scalable approximate inference procedure for MAP, marginal, and marginal MAP inference in continuous graphical models with a number of attractive features:

- Our approach is flexible enough to handle models with both discrete and continuous random variables. We provide theoretical results and detailed experiments to show that our approach does not suffer from the convergence and boundedness issues that arise with message-passing strategies in continuous models.
- After an initial inference pass, new MAP/marginal inference queries can be approximated in linear time. For applications in which many distinct queries need to be computed, we show experimentally, that one round of inference via our approach is significantly faster than the competing methods and yields comparable or, in many cases, higher accuracy than the state-of-the-art methods run for each individual query.
- Like deep neural networks, our approach is capable of taking advantage of modern GPUs to speed up inference. We show via experiments on depth estimation and optical flow problems that our method is capable of scaling to model sizes for which existing state-of-the-art methods are impractical.

In a variety of experiments, from small synthetic experiments to large scale computer vision tasks with both discrete and continuous random variables, we demonstrate that our approach yields fast, competitive approximate inference when compared to existing methods on MAP and marginal MAP inference tasks.

2 MARKOV RANDOM FIELDS

A Markov random field (MRF) is a graph together with a collection of nonnegative potential functions defined over its cliques. In this work, we focus on pairwise MRFs but note that our method can be extended to larger potential functions with minor modification. Given an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with vertex set $\mathcal{V}$ and edge set $\mathcal{E}$, such that each node $i \in \mathcal{V}$ corresponds to a random variable $x_i \in \mathcal{X}_i$, a pairwise MRF defines a joint distribution

$$p(x_{\mathcal{V}}) = \frac{1}{\mathcal{Z}} \prod_{i \in \mathcal{V}} \phi_i(x_i) \prod_{(i,j) \in \mathcal{E}} \psi_{ij}(x_i, x_j), \qquad (1)$$

where $\phi_i : \mathcal{X}_i \to \mathbb{R}_{\geq 0}$ is a potential function defined over node i and $\psi_{ij} : \mathcal{X}_i \times \mathcal{X}_j \to \mathbb{R}_{\geq 0}$ is the edge potential defined over the edge (i, j). The partition function, $\mathcal{Z}$, is a normalizing constant that ensures $p(x)$ sums/integrates to one. In this work, we will allow models that contain both continuous and discrete random variables, that is $\mathcal{X}_i$ need not be a finite set. In all cases, we assume that the corresponding integrals/sums exist.

The aim of marginal inference, a typical inference task, is to calculate the partition function $\mathcal{Z}$, and/or marginal distributions of the form $p(x_{\mathcal{A}})$ where $\mathcal{A} \subseteq \mathcal{V}$. More useful for classification tasks is maximum a posteriori (MAP) inference, which seeks to find an assignment to the set of (non-evidence) variables $x_{\mathcal{A}}$ that maximizes the probability $p(x_{\mathcal{A}}|x_{\mathcal{V}\setminus\mathcal{A}})$ for $p(x_{\mathcal{V}\setminus\mathcal{A}}) > 0$. More generally, given two disjoint sets $\mathcal{A}, \mathcal{B} \subseteq \mathcal{V}$ the marginal MAP task involves maximizing a conditional marginal distribution, $\arg\max_{x_{\mathcal{A}}} p(x_{\mathcal{A}}|x_{\mathcal{B}})$.

As all of the above inference tasks are NP-hard in general (Koller and Friedman, 2009), approximate inference routines are often necessary in practice. Many of these approximate inference schemes are modeled after the belief propagation (BP) algorithm (Pearl, 1982). BP is a message-passing procedure for marginal inference in MRFs. While the algorithm is exact on tree-structured graphs, it can also perform well on loopy graphs (Taga and Mase, 2006). BP iteratively computes a series of messages sent between neighboring nodes of the MRF. Upon convergence, the messages are used to construct beliefs, which are proportional to the true marginal distributions of the models in the case of tree-structured MRFs.

Marginal inference with BP becomes significantly more challenging with continuous variables as the message updates involve computing integrals instead of sums. The integrals generally cannot be computed in closed-form outside of special cases, e.g., Gaussian graphical models (Bickson, 2008), and require approximations. Sudderth et al. (2003) proposed approximating the messages as mixtures of Gaussians, but this approach is too expensive in practice. Ihler and McAllester (2009) and Lienart et al. (2015) proposed approximating the continuous messages with a finite number of appropriately chosen particles. At each iteration, the current set of particles for each node is resampled. While, theoretically, such an approach can be made arbitrarily accurate, in practice, a large number of particles may be needed for accurate inference, and selecting an efficient proposal distribution for the particle updates can be challenging. Noorshams and Wainwright (2013) proposed representing the messages using the top M terms of an orthogonal series expansion. Minka (2001) proposed the expectation propagation algorithm which represents the messages using a tractable family

and approximates the message updates using a moment matching procedure. Song et al. (2011) proposed a joint learning and inference strategy based on kernel methods to perform the message updates, though this approach can yield beliefs with negative values and performing MAP inference using the converged beliefs is non-trivial.

For the MAP task, the sums in the BP message-passing updates can be replaced with $\max$'s to yield the max-product message-passing algorithm. This algorithm can have convergence issues in practice, and many different alternative schemes with improved performance have been proposed (Kolmogorov and Wainwright, 2005; Kolmogorov and Rother, 2007; Globerson and Jaakkola, 2007; Werner, 2007; Ruozzi and Tatikonda, 2013b). For continuous MRFs, max-product versions of the particle methods have also been proposed (Pacheco and Sudderth, 2015).

For the discrete marginal MAP task, several approaches based on AND/OR search and NP oracles have recently been proposed (Marinescu et al., 2014, 2017; Xue et al., 2016), none of which, however, has been extended to the continuous case yet. The mixed-product algorithm (Liu and Ihler, 2013) is a hybrid message-passing algorithm corresponding to a variational approximation, which can be used to design convergent methods. In principle, mixed product could be extended to the continuous case using the same ideas as the particle methods above, but it's unclear how efficient such a procedure would be in practice.

2.1 BETHE FREE ENERGY

The converged messages in BP correspond to local optima of the Bethe free energy (BFE) optimized over the set of beliefs, which when appropriately normalized satisfy a collection of local consistency constraints (Yedidia et al., 2005). The valid beliefs in the local marginal polytope $\mathcal{M}$ are required to be nonnegative functions that marginalize to each other, i.e.,

$$\int_{x_i} b_i(x_i) = 1, \forall i \in \mathcal{V} \tag{2}$$

$$\int_{x_j} b_{ij}(x_i, x_j) = b_i(x_i), \forall (i,j) \in \mathcal{E}, \tag{3}$$

and the BFE is defined as

$$\begin{aligned}\mathcal{F}(b) = &- \sum_{i \in \mathcal{V}} \mathbb{E}_{b_i}[\log \phi_i] - \sum_{(i,j) \in \mathcal{E}} \mathbb{E}_{b_{ij}}[\log \psi_{ij}] \\ &+ \sum_{i \in \mathcal{V}} \mathbb{E}_{b_i} \log[b_i] + \sum_{(i,j) \in \mathcal{E}} \mathbb{E}_{b_{ij}} \left[\log \frac{b_{ij}}{b_i b_j}\right]. \end{aligned} \tag{4}$$

The log-partition function, $\log \mathcal{Z}$, can be approximated by minimizing (4) over beliefs in the local marginal polytope, i.e., $\log \mathcal{Z} \approx -\min_{b \in \mathcal{M}} \mathcal{F}(b)$. The optimum of the Bethe free energy yields the true log-partition function and node/edge marginals whenever the graph is a tree but only yields an approximation more generally. Convergent alternatives to message-passing algorithms have been designed using gradient descent on the BFE (Welling and Teh, 2001), and this is the approach we adopt here. Note, however, that (4) is not a convex function over general graphs and even computing the integrals in the general case could be nontrivial. Different "reweighted" entropy approximations other than the one used in (4) have also been considered in practice (Wainwright et al., 2003; Meltzer et al., 2009; London et al., 2015) and can easily be incorporated into our approach if desired.

3 ONE-SHOT INFERENCE

Our aim in this work is to show that a simple variational method can be used as an alternative to the above message-passing schemes to find local optima of the BFE in the hybrid case. To accomplish this, we restrict the beliefs to be mixtures of fully factorized distributions. Specifically,

$$b_i(x_i; \eta) = \sum_{k=1}^{K} w_k b_i^k(x_i; \eta_i^k) \tag{5}$$

$$b_{ij}(x_i, x_j; \eta) = \sum_{k=1}^{K} w_k b_i^k(x_i; \eta_i^k) b_j^k(x_j; \eta_j^k), \tag{6}$$

where K is the number of mixture components, the w_k are the mixture probabilities (which are shared across all marginals), and η is a vector of parameters used to construct each of the univariate distributions. For example, if $b_i(x_i)$ is chosen to be a mixture of normal distributions, then η_i^k would be a vector whose components are the mean and variance of the corresponding normal distribution. Under an appropriate choice of distribution, these mixtures can be made arbitrarily expressive. With these definitions, the local marginalization constraints (2)-(3) are trivially satisfied, and the BFE optimization problem is reduced to an optimization over the η and w parameters.

Our proposed strategy is to optimize the BFE over mixtures of the above form using standard gradient descent, similar in spirit to other variational approaches employing mean-field mixtures as approximate distributions but with a different entropy approximation: Jaakkola and Jordan (1999) use local variational approximation and introduce additional variational parameters to lower-bound the mixture entropy; their optimization procedure involves iterating interdependent consistency equations, which can be hard to parallelize to take advantage of GPUs; Gershman et al. (2012) consider Gaussian mixtures with a diagonal covariances and employ Jensen's inequality to approximate the mixture entropy. It can be hard to tell

in practice which entropy approximation (Jensen'fs inequality, Bethe, etc.) works better. However, our method can exploit the model structure (due to the tree-based Bethe approximation), and has the following guarantee in tree-structured MRFs: (1) the negative BFE always lower bounds the log-partition function assuming exact integration and (2) the gap between $-\log Z$ and the optimum of the BFE becomes arbitrarily small as the number of mixture components goes to infinity. Because of this, our method should be preferred in tree-structured models; in other models, we dont have such guarantee (e.g., the BFE could be larger or smaller than $-\log \mathcal{Z}$), and itll be harder to compare the variational approximations.

Full details of the gradient computations can be found in Appendix B. Note that, technically, the mixture weights are constrained to be nonnegative and sum to one. This can either be handled by projected gradient method or by introducing a change of variables that represent the mixture weights as a softmax of unconstrained variables. Computing the gradient of the BFE requires computing expectations with respect to beliefs in (5), (6). For discrete random variables these sums can be computed exactly, but for continuous random variables, the integrals, which can be expressed as expectations with respect to the beliefs, will need to be approximated.

Any of a variety of methods can be used to compute the expectations in the BFE, e.g., sampling methods, quadrature methods (Golub and Welsch, 1969), Stein variational gradient methods (Liu et al., 2016; Wang et al., 2018). The preferred method may depend on the specific application. The complexity of gradient descent optimization on the BFE scales roughly as $O(|\mathcal{E}|K^2)$ times the cost of approximate integration per iteration (an additional $O(L^2)$ if Gauss-Hermite quadrature with L points is used), though parallelization and stochastic methods can be used to significantly reduce the per iteration complexity in practice. Additionally, the gradient computation can be easily formulated in such a way as to take advantage of modern GPU hardware (we explore this in more detail in the experimental section). We have observed that while increasing K generally yields more accurate solutions, it may increase the number of iterations required for convergence.

After performing one round of inference that yields a K component mixture model with partition function $\mathcal{Z}$, each of the typical inference tasks can be approximated directly from the mixture without needing to perform additional rounds of gradient descent on the BFE (though this can also be done if desired). For many inference tasks, this significantly reduces the computational overhead needed to perform inference after the initial mixture distribution is computed. For applications in which repeated inference on the same model is desired, this can lead to significant practical performance gains.

Marginal Inference: The marginal distribution over $A \subseteq V$ can be approximated directly from the mixture, $p_A(x_A) = \sum_{k=1}^K w_k \prod_{i \in A} b_i^k(x_i)$. Also note that the energy of this distribution can be approximated as $\mathcal{Z} \cdot p_A(x_A)$.

MAP Inference: The mode of the distribution is found by computing $\arg\max_x \sum_{k=1}^K w_k \prod_{i \in V} b_i^k(x_i)$. This can be done exactly if the number of MAP variables is small. Alternatively, we can approximate the MAP problem in one of two ways. First, under the assumption that univariate distributions are easy to sample from, we could approximate the MAP assignment via sampling, though this approach only works well if the MAP assignment occurs with relatively high probability. Alternatively, we could approximate the MAP assignment using a coordinate/gradient ascent method starting from each of the K modes of the separate mixture components. The latter approach seems to perform quite well in practice, at least in the case of isotropic Gaussian mixtures, and yields a practical method for finding all local maxima of univariate Gaussians distributions (Carreira-Perpinan, 2000). Our experimental results suggest that this also works well for other types of mixtures.

Marginal MAP Inference: A combination of the previous two inference tasks for $A \subseteq V$, $\arg\max_{x_A} \sum_{k=1}^K w_k \prod_{i \in A} b_i^k(x_i)$. The max can be approximated using the strategies discussed above.

Conditional Marginals: The conditional distribution of x_A given x_O for disjoint subsets $A, O \subseteq \mathcal{V}$ is given by

$$p_A(x_A|x_O) = \frac{\sum_{k=1}^K w_k \prod_{i \in A \cup O} b_i^k(x_i)}{\sum_{k=1}^K w_k \prod_{i \in O} b_i^k(x_i)}.$$

Sampling: New samples can easily be generated from the mixture assuming that each univariate distribution is easy to sample from, without the need for MCMC methods.

As all of the above inference tasks can be efficiently estimated, the primary question, then, is whether or not a single round of inference is good enough to yield accurate and fast predictions in practice. In Section 4, we show that this is indeed the case in a variety of applications.

3.1 BOUNDEDNESS OF THE BFE

A significant limitation of the belief propagation based approaches is that, like BP, many of them may fail to converge, even on simple models. For example, for Gaussian graphical models, the BFE optimization problem (4) is unbounded from below whenever the precision matrix is not walk-summable, and BP can fail to converge for these models (Malioutov et al., 2006; Cseke and Heskes, 2011; Ruozzi and Tatikonda, 2013a).

Here, we show that the unboundedness of the BFE approximation in the Gaussian case occurs only over the local marginal polytope - not the marginal polytope. In particular, as all of the beliefs produced by our approach must be realized as the marginals of some joint distribution, the BFE optimization problem over the set of beliefs that are Gaussian mixtures of the form (5)-(6) is bounded from below, and consequently, gradient descent on the BFE is guaranteed to converge (given proper step sizes).

Theorem 1. *The BFE optmization problem* (4) *is bounded below whenever* p *is a Gaussian distribution and the optimization is performed over beliefs that arise from any joint distribution* q *with finite first and second moments (for example, when* q *is a mixture of Gaussians).*

Proof. Given a Gaussian distribution over n variables $p(x) = \tilde{p}(x)/\mathcal{Z}$, with $\tilde{p}(x) = \exp(-\frac{1}{2}x^T Jx + h^T x)$, J positive definite, suppose we approximate it by a continuous distribution q, such that $\mathbb{E}_q[X] = \mu, \mathbb{V}_q[X] = \Sigma$ (which are assumed to exist). Denote mutual information by $\mathbf{I}$, entropy by $\mathbf{H}$, and the set of edges in the Gaussian MRF by $\mathcal{E}$. By simple algebra, the BFE is then

$$\begin{aligned}\mathcal{F}(q) &= \mathbb{E}_q[\frac{1}{2}x^T Jx - h^T x] + \sum_{(i,j)\in\mathcal{E}} \mathbf{I}[q_{ij}] - \sum_i \mathbf{H}[q_i] \\ &\geq \mathbb{E}_q[\frac{1}{2}x^T Jx - h^T x] - \sum_i \mathbf{H}[q_i] \\ &= \frac{1}{2}\text{Tr}[J\Sigma] + \frac{1}{2}\mu^T J\mu - h^T\mu - \sum_i \mathbf{H}[q_i],\end{aligned}$$

where the inequality follows from the fact that mutual information is always nonnegative. Since J is positive definite, the quadratic form $\frac{1}{2}\mu^T J\mu - h^T\mu$ is bounded from below. So it's sufficient to show that $g(q) \triangleq \frac{1}{2}\text{Tr}[J\Sigma] - \sum_i \mathbf{H}[q_i]$ is bounded from below.

Lemma 2. *Let* A, B *be two* $n \times n$ *real symmetric matrices, with* B *positive definite; let* $\lambda_n(A)$ *be the smallest eigenvalue of* A*. Then* $\text{Tr}[AB] \geq \lambda_n(A)\text{Tr}[B]$.

The proof of Lemma 2 can be found in Appendix C. As a result, we have

$$\begin{aligned}g(q) &\geq \frac{\lambda_n(J)}{2}\text{Tr}[\Sigma] - \sum_i \mathbf{H}[q_i] \\ &\geq \frac{\lambda_n(J)}{2}\sum_i \Sigma_{ii} - \frac{1}{2}\sum_i \log(2\pi e\Sigma_{ii}),\end{aligned}$$

where the first inequality follows from Lemma 2 and the second inequality is a consequence of the fact that differential entropy of a distribution with variance σ is maximized by a Gaussian distribution with variance σ. Finally, as $\lambda_n(J), \Sigma_{11}, \ldots, \Sigma_{nn} > 0$, we have that $(\lambda_n(J)\Sigma_{ii} - \log\Sigma_{ii})$ is bounded below for all i. □

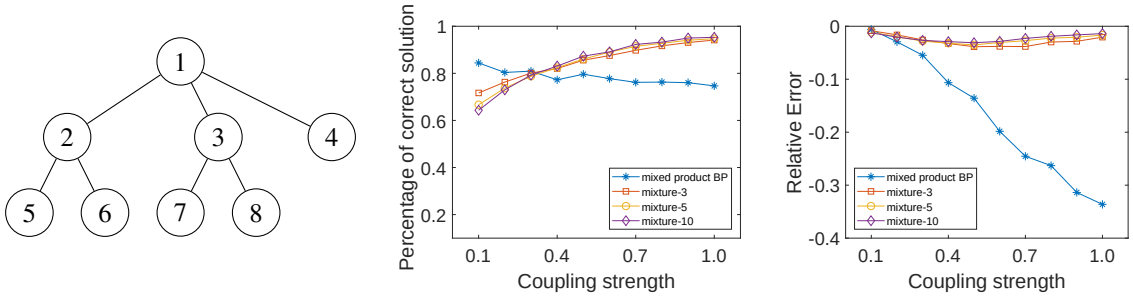
4 EXPERIMENTS

We apply our one-shot inference method (OSI) to a variety of MAP and marginal MAP (MMAP) tasks. In each setting, we compare against appropriate baselines that optimize the same objective: max-product/D-PMP for discrete/continuous MAP problems and mixed-product BP (MPBP) for discrete marginal MAP problems. All methods were implemented in MATLAB. For OSI, we used mixtures of Gaussian or Beta distributions as the approximate beliefs, and Gaussian quadrature for approximate integration (see Appendix 3). MAP inference in OSI used coordinate/gradient descent starting from the modes of the individual mixture components as discussed above. All experiments were performed on a desktop with 8 core i7-6700 CPU, except for experiments in section 4.4 which used an additional Nvidia Tesla V100 GPU.

4.1 SYNTHETIC MARGINAL MAP ON TREES

We begin with synthetic experiments on the tree model in Figure 1a with discrete random variables and pairwise factorization $p(X) = \frac{1}{\mathcal{Z}}\exp\left(\sum_{i\in V}\theta_i(x_i) + \sum_{(i,j)\in E}\theta_{ij}(x_i,x_j)\right)$. The parameters θ_i and θ_{ij} are sampled from Gaussian distributions, $\theta_i(x_i) \sim \mathcal{N}(0, 0.01)$ and $\theta_{ij}(x_i, x_j) \sim \mathcal{N}(0, \sigma^2)$, where the coupling σ is varied from 0.1 to 1.0. For each different value of σ, 100 different set of θ parameters are sampled. Out of the 8 nodes in the graph, we pick three nodes to be MAP nodes and the rest to be sum nodes. For each θ all 56 combinations of MAP/sum nodes are considered and the results are averaged. Note that inference using OSI is only run once for each θ while MPBP is rerun for each of the 56 possibilities. OSI is run with 3, 5, and 10 mixture components for 500 iterations each, starting from randomly selected initial discrete beliefs and uniform initial mixture weights. We compare against MPBP, initialized with random messages and run until convergence (at least 50 iterations, if it has not converged, we run another 200 extra iterations).

We report the average percentage of correctly identified MMAP assignments as well as the average relative error, $(p(\hat{x}_B) - p(x_B^*))/p(x_B^*)$ where $\hat{x}_B$ is the estimated MMAP assignment and x_B^* is the optimal MMAP assignment, for each method in Figure 1. The entire process using OSI for prediction finishes within 40 seconds while MPBP requires roughly 250 seconds as it must be rerun on each new inference task. For this task, OSI performed better on average with respect to both relative error and percentage of correct solutions for larger couplings. For small couplings, while OSI returned a MAP estimate that had small relative error, it did not return the exact MAP estimate. This is not entirely surprising as OSI was not



(a) Tree structure (b) Percentage of correct solutions (c) Relative error

Figure 1: Graph and inference results for a synthetic tree-structured model.

used to solve each specific instance but only to generate a good approximate distribution. From a practical perspective, it is probably sufficient to return a solution that is close enough to the true solution in energy value.

4.2 MARGINAL MAP ON UCI DATASETS

In a second set of experiments, we evaluated OSI on MMAP tasks over several UCI repository datasets (Dheeru and Karra Taniskidou, 2017): Iris, Letter, Solar Flare, Mammographic masses (M.M.), Tic-Tac-Toe and Yacht Hydrodynamics (Y.H.). For each dataset, we learn a discrete tree-structured distribution using the method of Chow and Liu (1968). The number of random variables in these models varies from 5 to 17.

We compared MPBP with OSI using 3, 5, and 10 mixture components; on each model, the MMAP performance was assessed by computing the relative MAP error (against ground truth) averaged over all possible subsets of 3 MAP nodes (the remaining nodes were summed over). Each method was run with 5 random initialization; Table 1 shows the resulting mean and standard deviation of MMAP performance. Note that, while increasing the number of mixture components seems to increase the number of iterations necessary for OSI to converge, having more mixture components tends to reduce the error of the marginal MAP assignment. Again our procedure matches or outperforms MPBP while only being run once - not being tailored to a specific marginal MAP instance.

4.3 INFERENCE IN CYCLIC GRAPHS

Continuing in the same vein as the previous experiments, we apply our method for MAP and marginal MAP inference on larger graphs that contain cycles. We consider MMAP problems on an OCR data set, MNIST, and instan ecsselected from the set of UAI challenge problems.

Image Completion: For the image completion task we considered two data sets: an OCR dataset collected by Kassel (1995) that consists of 16×8 binary images of handwritten letters and the MNIST dataset (LeCun et al., 1998) consisting of 28×28 grayscale hand-written digits. For both models we trained a simple MRF on a subset of the data using maximum likelihood estimation whose structure contains a single label node connected to all of the observed pixels and the observed pixels are connected via a grid structure. Given the trained models, we took a collection of data points, removed part of the input image (the top or bottom half) and then performed conditional MAP inference to recover the missing pixels from the remaining observations.

For the letter data, the model contained only discrete variables and was trained on 7427 images consisting of 'f' and 'h' labels (it achieves 93% accuracy on the classification task). For evaluation we deleted the top half of all of the training examples, ran OSI with three mixture components and standard max-product message passing to complete the images, and stored the energy of the MAP solution of each completion. On average, OSI produced a completion with an average energy value of -146.77 ± 12.51 while max-product returned a slightly worse average energy value of -146.09 ± 13.33. Although OSI does not achieve a lower energy on every instance in the data set, the energy values were lower on average and when max-product outperformed OSI it did so only marginally, while when OSI outperformed max-product the gap was typically much larger.

For MNIST, the model contained 784 continuous variables, one discrete variable, and 2296 edges (1512 between continuous variables and 784 hybrid discrete-continuous edges). The model was specified by a simple exponential family, whose log potential functions correspond to over-complete features for the discrete node, second-order polynomials for the continuous nodes, and products of node features for all the edges. The model was trained on all MNIST training images of digits 1 and 9 for 300 iterations using maximum likelihood (with OSI performing the required inference) to a training accuracy

Table 1: Relative error of OSI and mixed-product BP for a marginal MAP task on various UCI datasets.

Dataset	mixed-BP	mixture-3	mixture-5	mixture-10
Iris	-0.1848 ± 0.0743	-0.1242 ± 0.1238	-0.1473 ± 0.1326	$\mathbf{-0.1188 \pm 0.0566}$
Letter	-0.0236 ± 0.0380	-0.0253 ± 0.0312	-0.0223 ± 0.0290	$\mathbf{-0.0216 \pm 0.0273}$
Solar Flare	-0.0404 ± 0.0620	-0.0370 ± 0.0654	-0.0369 ± 0.0664	$\mathbf{-0.0367 \pm 0.0627}$
M.M.	-0.1736 ± 0.2014	-0.1884 ± 0.1892	$\mathbf{-0.1637 \pm 0.1614}$	-0.1732 ± 0.1784
Tic-Tac-Toe	-0.0789 ± 0.0768	-0.1078 ± 0.0750	-0.0791 ± 0.0562	$\mathbf{-0.0757 \pm 0.0778}$
Y.H.	-0.0590 ± 0.1371	-0.0245 ± 0.0538	-0.0245 ± 0.0538	$\mathbf{-0.0211 \pm 0.0514}$

(a) (b) (c) (d)

Figure 2: Example of the image completion task on MNIST, where: (a) is the partial image given (along with the label), (b) is the ground truth, (c) is the OSI solution (0.17 s/image), (d) is the D-PMP solution (3.42 s/image).

of 96%. We randomly sampled 100 images, hid either the top or bottom half, performed MAP inference over the hidden pixels using OSI or D-PMP (Pacheco and Sudderth, 2015), and evaluated the energy of the solutions under the given model, as before.

For OSI, we used mixtures of Beta distributions for the beliefs of continuous random variables (this seems like an appropriate choice as the grayscale input only varies from 0 to 1); we ran inference for 250 iterations in the learned hybrid MRF to obtain a surrogate mixture for image completion, on which gradient ascent was then run to maximize the probability of the hidden pixels given each incomplete image; the completions had an average energy value of 3042.5 ± 58.2 and MSE (mean squared error) of 0.0193 ± 0.0082. For comparison, D-PMP was run in the reduced MRF over the remaining pixels given each incomplete observation (since D-PMP only handles continuous MRFs), using 20 particles, Gaussian random walk proposal, for 10 iterations (this resulted in full convergence). D-PMP obtained worse average completion energy of 3044.7 ± 56.0 and MSE of 0.0226 ± 0.0102, while requiring 20 times more CPU time per image (we used the optimized D-PMP implementation provided by the authors; neither method used parallelization). See Figure 2 for examples of completions.

UAI challenge problems: In a second set of experiments on loopy graphs we considered MMAP estimation on models and potentials obtained from various UAI challenge problems. For each of these discrete models, we considered three different configurations of MAP and sum nodes (see Appendix A). For each configuration, we ran OIS with 5 mixture components and used the resulting mixture and MPBP to predict the MMAP solution. To evaluate the quality of the predictions, we used exact inference (variable elimination). The results are described in Table 2. Our method significantly outperforms MPBP in this case. One possible explanation for this performance is that OSI, if a good mixture is obtained in the inference phase, is actually making fewer approximations than MPBP on loopy models. In particular, the beliefs returned by our method are always realizable as the marginals of some joint distribution over all of the variables whereas this is not the case with MPBP.

4.4 COMPUTER VISION TASKS

We conclude the experimental section with two computer vision tasks: optical flow estimation and stereo depth estimation. The scale of the MRFs for these tasks is significantly larger than those considered in the previous experiments as these models can contain hundreds of thousands of nodes and edges. For these tasks, our aim is to show that, like deep neural networks, our inference method can scale to such problems by taking advantage of modern GPU hardware. Since the BFE and its gradient only involve expectations with respect to cliques in the MRF, the required computation is embarrassingly parallelizable, and can be easily distributed across GPU cores, e.g., with MATLAB's `arrayfun`; alternatively, the computation can be highly vectorized to utilize efficient GPU primitives for tensor operations (see Appendix B for a discussion). As such, we implemented our method using MATLAB's built-in GPU computing support.

4.4.1 Optical Flow

Optical flow estimation attempts to recover 2D pixel motion from a sequence of images. A typical approach models the flow field using a pairwise MRF with node potentials that enforce data constancy and edge potentials that penalize discrepancies between adjacent pixels. The flow at each pixel is defined as a vector (u, v), with scalars u and v representing horizontal and vertical speed respectively. We adopt the same potentials as the Classic-C

Table 2: The marginal MAP value produced by mixed-product and OSI on several UAI challenge problems.

	Combination #1		Combination #2		Combination #3	
Dataset	**mixed-BP**	**mixture-5**	**mixed-BP**	**mixture-5**	**mixed-BP**	**mixture-5**
Grids26	4378.0 ± 149.9	$\mathbf{4963.0 \pm 40.6}$	3927.0 ± 310.3	$\mathbf{4621.6 \pm 60.9}$	3530.5 ± 276.1	$\mathbf{4189.1 \pm 63.8}$
Grids28	6661.1 ± 244.4	$\mathbf{7442.5 \pm 29.2}$	5484.1 ± 412.6	$\mathbf{6944.6 \pm 56.3}$	4851.9 ± 271.6	$\mathbf{6285.6 \pm 16.6}$
Grids29	2280.8 ± 85.4	$\mathbf{2472.2 \pm 24.4}$	2000.9 ± 111.7	$\mathbf{2290.8 \pm 22.5}$	1897.9 ± 63.6	$\mathbf{2087.7 \pm 12.1}$
Grids30	4661.2 ± 54.0	$\mathbf{5078.7 \pm 31.8}$	3784.3 ± 318.0	$\mathbf{4606.0 \pm 42.0}$	3230.5 ± 558.9	$\mathbf{4264.6 \pm 24.6}$

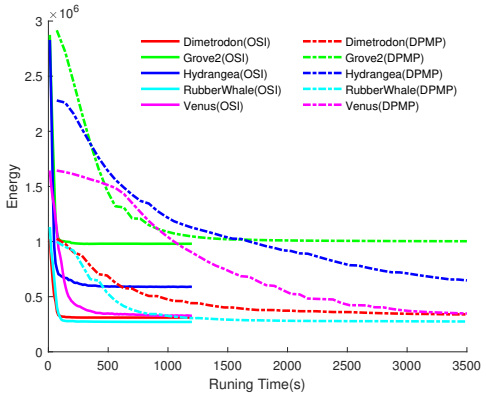


Figure 3: The running time of D-PMP versus GPU-accelerated OSI on 5 images from the Middlebury dataset.

model (Sun et al., 2014) and apply bicubic interpolation for continuous coordinates of pixels (see Appendix D).

We consider the optical flow estimation problem on 5 pairs of image sequences from the Middlebury optical flow dataset (Baker et al., 2011). We compare OSI with the base-line method Classic-C (Sun et al., 2014) for pixel level models and with particle based max-product method D-PMP on a super pixel version of the problem as in (Pacheco and Sudderth, 2015). Note that the standard implementation of D-PMP is impractical to run at the pixel level, and it's less clear how to take advantage of GPU acceleration in the particle case. For OSI, we used a Gaussian mixture with 3 components as the approximate distribution, and 9 quadrature points for approximate integration. We explored multiplying the entropy term in the BFE by a constant that was halved every 500 iterations (approximating the zero temperature limit); it gave similar performance to optimizing the BFE without modification.

The average endpoint error (AEPE) and converged energies are reported in Table 3. Classic-C, which applies a median filter on the intermediate flow results after every warping iteration, is included for comparison purposes. At the super pixel level, OSI performs comparably to D-PMP on both metrics while running for a fraction of the time - estimated energy per iteration can be found in Figure 3. Figure 4 visualizes the flow results generated by the two methods. The extension to the continuous case provides a smoother estimate of the ground truth when run at the pixel level instead of at the super pixel level, yielding the lower energy and AEPE reported in Table 3.

4.4.2 Stereo Depth Estimation

Finally, in order to further demonstrate the utility of continuous models in practice, we consider the stereo depth estimation problem: given two images taken from sightly different angles, the goal of stereo estimation is to estimate the depth $d(i, j)$ of each pixel in the image. As in optical flow, bicubic interpolation was used to extend the discretized problem to the continuous case.

We evaluate OSI on the Teddy and Cones images from the Middlebury stereo dataset (Scharstein and Szeliski, 2003). The quarter-size images in this dataset are 450×375. Ground truth disparities d_T in the data use quarter-pixel accuracy in the range $[0.25, 63.75]$, where 0 indicates an unknown value. A quantitative comparison of OSI (using the previous setup), graph cuts (min-cut with alpha-expansion), and tree-reweighted belief propagation on this dataset in terms of percentage of bad matching pixels on non-occluded area can be found in Table 4. Note, $Bad\% = \frac{1}{N}\sum(|(d(i,j) - d_T(i,j)| > t)$, where $t = 1, 2$.

In Table 5, we report the average energy estimate and the corresponding running time for each of the methods. OSI's convergence rate on the continuous inference task is comparable to that of graph cuts on the discrete inference task, while producing better depth estimates not only quantitatively but also qualitatively (see Figure 5).

5 DISCUSSION

We have proposed a method for one-shot inference in discrete, continuous, and hybrid graphical models that is especially practical in situations requiring repeated inference on the same model. We also showed that, even as a stand-alone inference procedure, our approach can be implemented efficiently on modern GPUs - allowing us to tackle problem sizes that would be challenging for competing general purpose MRF inference techniques. The approach retains these advantages on MRFs that may contain both discrete and continuous variables and is essentially potential function independent, i.e., the procedure

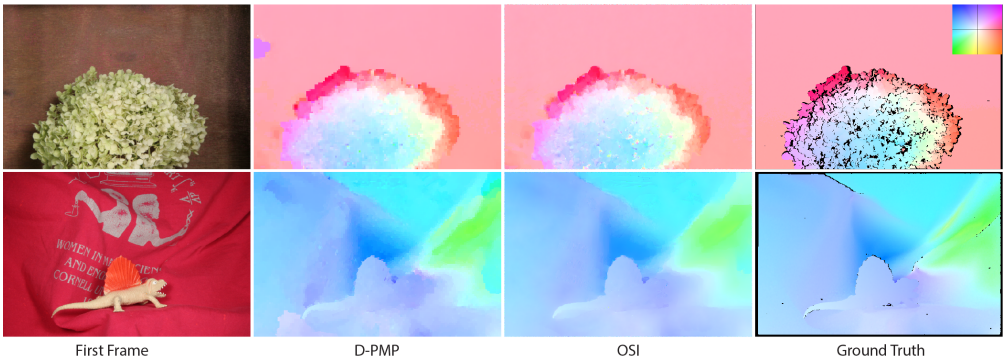


Figure 4: Optical flow estimation on Hydrangea (top) and Dimetrodon (bottom) image sequences by D-PMP (super pixel level) and OSI (pixel level). The color key in the upper right corner encodes the flow vector for each pixel.

Table 3: Energy and AEPE of the three methods for optical flow estimation on Middlebury training set on both the super pixel level (top) and the pixel level (bottom).

		Dimetrodon	**RubberWhale**	**Hydrangea**	**Venus**	**Grove2**	**Avg.**
OSI (super)	*Energy*	3.107E5	2.707E5	5.881E5	3.267E5	9.797E5	4.952E5
	AEPE	0.157	0.133	0.222	0.331	0.169	0.202
D-PMP (super)	*Energy*	3.221E5	2.714E5	5.667E5	3.239E5	9.981E5	4.964E5
	AEPE	0.203	0.137	0.245	0.357	0.168	0.222
OSI	*Energy*	3.386E5	2.950E5	8.413E5	5.420E5	9.139E5	5.862E5
	AEPE	0.157	0.132	0.220	0.330	0.162	0.200
Classic-C	*Energy*	3.472E5	3.208E5	7.228E5	4.118E5	11.87E5	5.979E5
	AEPE	0.162	0.110	0.194	0.286	0.183	0.187

Table 4: Performance on non-occlusion area for disparity estimation on the Teddy and Cones datasets.

Bad %	Teddy		Cones	
	t=1	t=2	t=1	t=2
OSI	14.1%	11.2%	10.7%	6.7%
GC	29.3%	11.4%	12.6%	7.0%
BP	16.3%	N/A	10.6%	N/A

Table 5: Converged energy (10^6) and run time (seconds) on "Teddy" image by OSI, TRW for tree reweighted BP, MP for max-product, and GC for graph cuts.

	OSI	TRW	MP	GC
Energy	**1.328**	1.366	1.402	1.365
Run Time	**30**	246	207	32

is generic enough to apply in many situations of interest without significant modification. Further, it does not suffer from the kind of convergence and unboundedness issues that can arise even in simple continuous models.

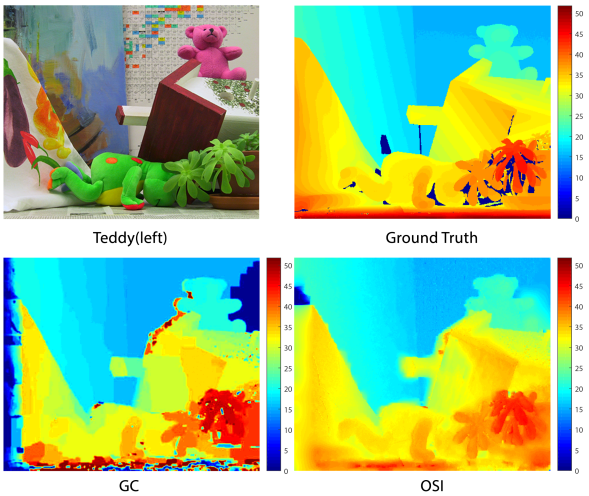


Figure 5: Stereo depth estimation results on "Teddy" using graph cuts and OSI. The color map encodes pixel disparities: hotter color means larger disparity (less depth).

Acknowledgments

This work was supported, in part, by the DARPA Explainable Artificial Intelligence (XAI) program under contract number N66001-17-2-4032 and NSF grant III-1527312.

References

S. Baker, D. Scharstein, J. P. Lewis, S. Roth, M. J. Black, and R. Szeliski. A database and evaluation methodology for optical flow. *International Journal of Computer Vision*, 92(1):1–31, 2011.

D. Bickson. *Gaussian belief propagation: Theory and application*. PhD thesis, Hebrew University of Jerusalem, 2008.

M. A. Carreira-Perpinan. Mode-finding for mixtures of Gaussian distributions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(11):1318–1323, 2000.

C. Chow and C. Liu. Approximating discrete probability distributions with dependence trees. *IEEE transactions on Information Theory*, 14(3):462–467, 1968.

B. Cseke and T. Heskes. Properties of Bethe free energies and message passing in Gaussian models. *Journal of Artificial Intelligence Research*, pages 1–24, 2011.

D. Dheeru and E. Karra Taniskidou. UCI machine learning repository, 2017. URL `http://archive.ics.uci.edu/ml`.

S. J. Gershman, M. D. Hoffman, and D. M. Blei. Nonparametric variational inference. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 235–242, 2012.

A. Globerson and T. S. Jaakkola. Fixing max-product: Convergent message passing algorithms for MAP LP-relaxations. In *Proc. 21st Neural Information Processing Systems (NIPS)*, Vancouver, B.C., Canada, 2007.

G. H. Golub and J. H. Welsch. Calculation of Gauss quadrature rules. *Mathematics of computation*, 23(106): 221–230, 1969.

Y. Guo, H. Xiong, and N. Ruozzi. Marginal inference in continuous Markov random fields using mixtures. In *AAAI*, 2019.

A. T. Ihler and D. A. McAllester. Particle belief propagation. In *Twelfth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 256–263, 2009.

T. S. Jaakkola and M. I. Jordan. Improving the mean field approximation via the use of mixture distributions. In M. I. Jordan, editor, *Learning in Graphical Models*. Cambridge: MIT Press, 1999.

R. H. Kassel. *A comparison of approaches to on-line handwritten character recognition*. PhD thesis, Massachusetts Institute of Technology, 1995.

D. Koller and N. Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

V. Kolmogorov and C. Rother. Minimizing nonsubmodular functions with graph cuts-a review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29 (7):1274–1279, July 2007.

V. Kolmogorov and M. Wainwright. On the optimality of tree-reweighted max-product message-passing. In *Proceedings of the Twenty-First Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 316–323, Arlington, Virginia, 2005.

Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

T. Lienart, Y. W. Teh, and A. Doucet. Expectation particle belief propagation. In *Advances in Neural Information Processing Systems (NIPS)*, pages 3609–3617, 2015.

Q. Liu and A. Ihler. Variational algorithms for marginal map. *The Journal of Machine Learning Research*, 14 (1):3165–3200, 2013.

Q. Liu, J. Lee, and M. Jordan. A kernelized Stein discrepancy for goodness-of-fit tests. In *International Conference on Machine Learning (ICML)*, pages 276–284, 2016.

B. London, B. Huang, and L. Getoor. The benefits of learning with strongly convex approximate inference. In *International Conference on Machine Learning (ICML)*, pages 410–418, 2015.

D. M. Malioutov, J. K. Johnson, and A. S. Willsky. Walk-sums and belief propagation in Gaussian graphical models. *Journal of Machine Learning Research*, 7: 2031–2064, 2006.

R. Marinescu, R. Dechter, and A. Ihler. And/or search for marginal MAP. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 563–572. AUAI Press, 2014.

R. Marinescu, J. Lee, A. T. Ihler, and R. Dechter. Anytime best+depth-first search for bounding marginal MAP. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

T. Meltzer, A. Globerson, and Y. Weiss. Convergent message passing algorithms: a unifying view. In *Proc. 25th Uncertainty in Artifical Intelligence (UAI)*, Montreal, Canada, 2009.

T. P. Minka. Expectation propagation for approximate Bayesian inference. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence (UAI)*, pages 362–369, 2001.

N. Noorshams and M. J. Wainwright. Belief propagation for continuous state spaces: stochastic message-passing with quantitative guarantees. *Journal of Machine Learning Research (JMLR)*, 14(1):2799–2835, 2013.

J. Pacheco and E. Sudderth. Proteins, particles, and pseudo-max-marginals: a submodular approach. In *International Conference on Machine Learning (ICML)*, pages 2200–2208, 2015.

J. Pearl. *Reverend Bayes on inference engines: A distributed hierarchical approach*. Cognitive Systems Laboratory, School of Engineering and Applied Science, University of California, Los Angeles, 1982.

N. Ruozzi. Exactness of approximate MAP inference in continuous MRFs. In *Advances in Neural Information Processing Systems (NIPS)*, pages 2332–2340, 2015.

N. Ruozzi. A lower bound on the partition function of attractive graphical models in the continuous case. In *Artificial Intelligence and Statistics (AISTATS)*, 2017.

N. Ruozzi and S. Tatikonda. Message-passing algorithms for quadratic minimization. *Journal of Machine Learning Research*, 14:2287–2314, 2013a.

N. Ruozzi and S. Tatikonda. Message-passing algorithms: Reparameterizations and splittings. *IEEE Transactions on Information Theory*, 59(9):5860–5881, Sept. 2013b.

D. Scharstein and R. Szeliski. High-accuracy stereo depth maps using structured light. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 1, pages I–I. IEEE, 2003.

L. Song, A. Gretton, D. Bickson, Y. Low, and C. Guestrin. Kernel belief propagation. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 707–715, 2011.

E. B. Sudderth, A. T. Ihler, M. Isard, W. T. Freeman, and A. S. Willsky. Nonparametric belief propagation. In *Computer Vision and Pattern Recognition (CVPR), IEEE Computer Society Conference on*, 2003.

D. Sun, S. Roth, and M. J. Black. A quantitative analysis of current practices in optical flow estimation and the principles behind them. *International Journal of Computer Vision*, 106(2):115–137, 2014.

N. Taga and S. Mase. On the convergence of loopy belief propagation algorithm for different update rules. *IEICE Trans. Fundam. Electron. Commun. Comput. Sci.*, E89-A(2):575–582, Feb. 2006.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Foundations and Trends in Machine Learning*, 1(1-2):1–305, 2008.

M. J. Wainwright, T. S. Jaakkola, and A. S. Willsky. Tree-based reparameterization framework for analysis of sum-product and related algorithms. *IEEE Transactions on information theory*, 49(5):1120–1146, 2003.

D. Wang, Z. Zeng, and Q. Liu. Stein variational message passing for continuous graphical models. In *International Conference on Machine Learning (ICML)*, 2018.

S. Wang, A. Schwing, and R. Urtasun. Efficient inference of continuous Markov random fields with polynomial potentials. In *Advances in neural information processing systems (NIPS)*, pages 936–944, 2014.

M. Welling and Y. W. Teh. Belief optimization for binary networks: A stable alternative to loopy belief propagation. In *Proceedings of the Seventeenth conference on Uncertainty in Artificial Intelligence (UAI)*, pages 554–561. Morgan Kaufmann Publishers Inc., 2001.

T. Werner. A linear programming approach to max-sum problem: A review. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 29(7):1165–1179, 2007.

Y. Xue, Z. Li, S. Ermon, C. P. Gomes, and B. Selman. Solving marginal map problems with NP oracles and parity constraints. In *Advances in Neural Information Processing Systems (NIPS)*, pages 1127–1135, 2016.

J. S. Yedidia, W. T. Freeman, and Y. Weiss. Constructing free-energy approximations and generalized belief propagation algorithms. *Information Theory, IEEE Transactions on*, 51(7):2282 – 2312, July 2005.