Sparse Message Passing and Efficiently Learning Random Fields for Stereo Vision

Jerod J. Weinman
Department of Computer Science
University of Massachusetts Amherst
Amherst, MA 01003
weinman@cs.umass.edu

Chris Pal
Department of Computer Science
University of Rochester
Rochester, NY 14627
cpal@cs.rochester.edu

Daniel Scharstein
Department of Computer Science
Middlebury College
Middlebury, VT 05753
schar@middlebury.edu

Technical Report UM-CS-2007-054
October 28, 2007

Abstract

Message passing algorithms based on variational methods and belief propagation are widely used for approximate inference in a variety of directed and undirected graphical models. However, inference can become extremely slow when the cardinality of the state space of individual variables is high. In this paper we explore sparse message passing to dramatically accelerate approximate inference. We show theoretically that sparse variational message passing iteratively and monotonically minimizes the KL divergence between a variational approximation and a distribution of interest. Sparse variational message passing achieves this by optimizing a lower bound on the partition function. We present experiments confirming these results using a conditional random field for a difficult stereo vision problem. We observe dramatic reductions in inference time with no loss in approximation quality. Learning using sparse methods also improves results over prior work using graph cuts.

1 Introduction

Belief propagation [10] and variational methods [6] are widely used techniques for inference in probabilistic graphical models. Both techniques have been used for inference and learning in models with applications ranging from text processing to computer vision [3, 5]. Winn and Bishop proposed Variational Message Passing (VMP) [9] as a way to view many variational inference techniques and represents a general purpose algorithm for approximate inference. The approach is similar in nature to belief propagation (BP) in that messages propagate uncertainty in a graph and the computations of messages when viewed in this framework have a similar form to BP. However, when used for inference about hidden variables in a graphical model, unlike BP, VMP optimizes a lower bound on the log probability of observed variables in the model.

As more large scale problems are addressed with probabilistic machine learning techniques, fast inference can be particularly important. While experimental work and theoretical analysis of variational methods show that the asymptotic performance of other methods such as sampling [1] can be

superior, variational methods are comparatively fast for approximate inference. However, for many real world problems, models containing variables with high cardinality state spaces arise. Under these conditions the gains in terms of inference speed with variational methods can be diminished. We are interested here in addressing this issue through constructing sparse variational methods which also provide theoretical guarantees that the Kullback-Leibler (KL) divergence between approximate distributions and true distributions are iteratively minimized.

In our empirical investigation here we focus upon approximate inference and learning in lattice structured Conditional Random Fields (CRFs) [7] applied to a real world stereo vision problem. However, our theoretical results and some experimental insights should be applicable to CRFs, MRFs and Bayesian Networks with arbitrary structures. In our experiments we explore a CRF formulation for stereo vision in which a lattice is constructed leading to energy functions with a traditional structure—single variable terms and pairwise terms. Unlike purely energy based formulations, since we cast the stereo problem as a CRF, we are able to view approximate inference and learning in the model from the perspective of variational analysis.

The remainder of this paper is structured as follows: In section 2 we show how sparse variational message passing minimizes the KL divergence between a variational approximation and a distribution of interest. In section 3 we present a conditional random field to solve a real world stereo vision problem and show how approximate inference is used to infer depth in an image as well as for learning. Finally, in section 4 we present results comparing sparse BP and VMP with graph cuts showing how sparse message passing can lead to an order of magnitude reduction in inference time and using variational distributions for learning improve results over a point estimate given by graph cuts.

2 Sparse mean field in a CRF

In the following exposition we derive the equations for mean field inference using a variational message passing perspective [9]. We show that sparse VMP will iteratively minimize the KL divergence between an approximation Q and a true distribution P. Further, we present sparse VMP in the context of CRFs and show that the functional we optimize is an upper bound on the negative log conditional partition function.

2.1 General mean field

Let X_i be a discrete random variable taking on values x_i from a finite alphabet $\mathcal{X} = \{0, \ldots, N-1\}$, where N is the number of states. Let $\boldsymbol{X}$ denote the concatenation of all random variables, which take on values denoted by $\boldsymbol{x}$. Let $\boldsymbol{y}$ be the observation conditioned on. Given a general conditional distribution

$$P(\boldsymbol{X} \mid \boldsymbol{y}) = \frac{1}{Z(\boldsymbol{y})} F(\boldsymbol{X}, \boldsymbol{y}) \quad \text{with} \quad Z(\boldsymbol{y}) = \sum_{\boldsymbol{x}} F(\boldsymbol{x}, \boldsymbol{y}),$$

consider minimizing the KL divergence between an approximate distribution $Q(\boldsymbol{X})$ and the true distribution $P(\boldsymbol{X} \mid \boldsymbol{y})$, which is expressed as

$$\begin{aligned} \mathrm{KL}(Q(\boldsymbol{X}) \,||\, P(\boldsymbol{X} \mid \boldsymbol{y})) &= \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \log \frac{Q(\boldsymbol{x})}{P(\boldsymbol{x} \mid \boldsymbol{y})} \\ &= \sum_{\boldsymbol{x}} Q(\boldsymbol{x}) \log \frac{Q(\boldsymbol{x}) Z(\boldsymbol{y})}{F(\boldsymbol{x}, \boldsymbol{y})} \\ &= -\langle \log F(\boldsymbol{x}, \boldsymbol{y}) \rangle_{Q(\boldsymbol{X})} - H(Q(\boldsymbol{X})) + \log Z(\boldsymbol{y}). \end{aligned}$$

Thus, a free energy $\mathcal{L}(Q(\boldsymbol{X}))$ and the KL divergence are related in the definition

$$\mathcal{L}(Q(\boldsymbol{X})) = -\langle \log F(\boldsymbol{x}, \boldsymbol{y}) \rangle_{Q(\boldsymbol{X})} - H(Q(\boldsymbol{X})), \tag{1}$$

$$\mathrm{KL}(Q(\boldsymbol{X}) \,||\, P(\boldsymbol{X} \mid \boldsymbol{y})) = \mathcal{L}(Q(\boldsymbol{X})) + \log Z(\boldsymbol{y}). \tag{2}$$

If we define the energy of a configuration $\boldsymbol{x}$ as $-\log F(\boldsymbol{x}, \boldsymbol{y})$, the free energy for our model is the expected energy under our variational distribution less the entropy of the approximating distribution

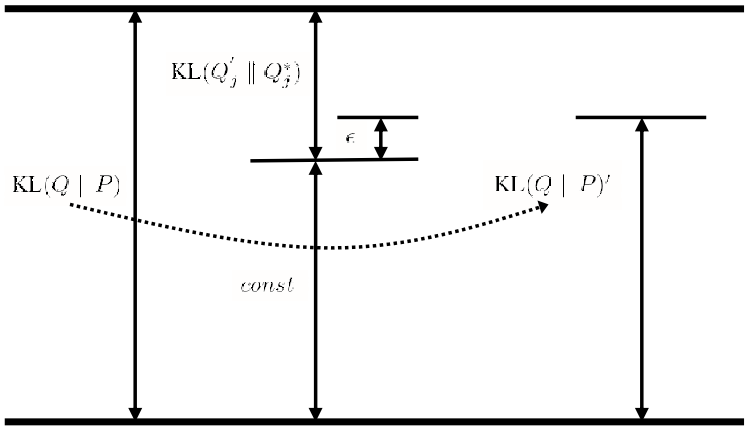

Figure 1: The complete divergence $\mathrm{KL}\left(Q \parallel P\right)$ can be decomposed into $\mathrm{KL}\left(Q_j^{'} \parallel Q_j^{*}\right) + \text{const}$. At each step of sparse variational message passing we minimize $\mathrm{KL}\left(Q_j^{'} \parallel Q_j^{*}\right)$ to within some ϵ which iteratively minimizes the global objective.

$Q\left(\boldsymbol{X}\right)$. Since the KL divergence is always greater than or equal to zero, $\mathcal{L}\left(Q\left(\boldsymbol{X}\right)\right) \geq -\log Z\left(\boldsymbol{y}\right)$ and the KL divergence is zero when the free energy equals the negative log partition function. Importantly, since $\log Z(\boldsymbol{y})$ is constant for a given observation, minimizing the free energy also minimizes the KL divergence.

Mean field updates will minimize $\mathrm{KL}\left(Q\left(\boldsymbol{X}\right) \parallel P\left(\boldsymbol{X} \mid \boldsymbol{y}\right)\right)$ for a factored distribution $Q\left(\boldsymbol{X}\right) = \prod_i Q\left(X_i\right)$. Under this Q, we can express our objective as

$$\begin{aligned} \mathcal{L}\left(Q\left(\boldsymbol{X}\right)\right) &= -\sum_{\boldsymbol{x}} \prod_i Q\left(x_i\right) \log F\left(\boldsymbol{x}, \boldsymbol{y}\right) + \sum_i \sum_{x_i} Q\left(x_i\right) \log Q\left(x_i\right) & (3) \\ &= -\sum_{\boldsymbol{x}} Q\left(x_j\right) \left\langle \log F\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\rangle_{\prod_{i \neq j} Q(X_i)} - H(Q(X_j)) - \sum_{i \neq j} H\left(Q\left(X_i\right)\right), & (4) \end{aligned}$$

where we have factored out one of the $Q\left(X_j\right)$. Adding Lagrange multipliers to form a new functional and solving for our update $Q^{*}\left(x_j\right)$ we have

$$Q^{*}\left(x_j\right) = \frac{1}{Z_j} \exp\left(\left\langle \log F\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\rangle_{\prod_{i \neq j} Q(X_i)}\right) \quad (5)$$

where Z_j is a normalization constant computed for each update. (See Appendix I for the complete derivation). Finally, iteratively updating each $Q\left(X_j\right)$ in this manner will monotonically decrease $\mathcal{L}\left(Q\left(\boldsymbol{X}\right)\right)$ and minimize our KL divergence.

2.2 Sparse updates

Given (2), (4) and (5) we can express $\mathrm{KL}\left(Q\left(\boldsymbol{X}\right) \parallel P\left(\boldsymbol{X} \mid \boldsymbol{y}\right)\right)$ as a function of a *sparse* update $Q^{'}\left(X_j\right)$, mean field update $Q^{*}\left(X_j\right)$ and the other $Q\left(X_i\right)$, where $i \neq j$:

$$\mathrm{KL}\left(Q\left(\boldsymbol{X}\right) \parallel P\left(\boldsymbol{X} \mid \boldsymbol{y}\right)\right) = \mathrm{KL}\left(Q'\left(X_j\right) \parallel Q^{*}\left(X_j\right)\right) + \log Z_j + \log Z\left(\boldsymbol{y}\right) - \sum_{i \neq j} H\left(Q\left(X_i\right)\right) \quad (6)$$

Since the last three terms of (6) are constant with respect to our update $Q^{'}\left(X_j\right)$, $\mathrm{KL}\left(Q\left(\boldsymbol{X}\right) \parallel P\left(\boldsymbol{X} \mid \boldsymbol{y}\right)\right)$ is minimized when $Q^{'}\left(X_j\right) = Q^{*}\left(X_j\right)$.

If each X_j is restricted to the subset of values $x_j \in \mathcal{X}_j \subseteq \mathcal{X}$, we may define sparse updates $Q^{'}\left(X_j\right)$ in terms of the original update $Q^{*}\left(X_j\right)$ and the indicator function $\mathbf{1}_{\mathcal{X}_j}\left(x_j\right)$ for the restricted range:

$$Q'\left(x_j\right) = \frac{\mathbf{1}_{\mathcal{X}_j}\left(x_j\right)}{Z_j'} Q^{*}\left(x_j\right) \quad (7)$$

where the new normalization constant is

$$Z'_j = \sum_{x_j} Q'(x_j) = \sum_{x_j \in \mathcal{X}_j} Q^*(x_j). \tag{8}$$

Thus the divergence between a sparse update and the original is

$$\begin{aligned} \mathrm{KL}\left(Q'(X_j) \,||\, Q^*(X_j)\right) &= \sum_x \frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j) \log\left(\left(\frac{\mathbf{1}_{\mathcal{X}_j}(x_j)}{Z'_j} Q^*(x_j)\right) \Big/ Q^*(x_j)\right) \\ &= -\log Z'_j \frac{1}{Z'_j} \sum_{x \in \mathcal{X}_j} Q^*(x_j) \\ &= -\log Z'_j. \end{aligned} \tag{9}$$

As a consequence, it is straightforward and efficient to compute a sparse $Q'(X_j)$ such that $\mathrm{KL}\left(Q'(X_j) \,||\, Q^*(X_j)\right) \leq \epsilon$ by sorting the $Q^*(x_j)$ values and performing a sub-linear search to satisfy the inequality. For example, if we wish to preserve 99% of the probability mass in the sparse approximation we may set $\epsilon = -\log 0.99 \approx .01$. Figure 1 illustrates the way in which sparse VMP iteratively minimizes the $\mathrm{KL}(Q \,||\, P)$ after each iteration of message passing. Because of the resulting sparsity of $Q(X_j)$, subsequent messages remain sparse. In section 4 we show how the increase in inference speed can be dramatic.

3 Stereo vision and CRFs

The stereo vision problem is to estimate the *disparity* (horizontal displacement) at each pixel given a rectified pair of images. It is common in MRF-based stereo vision methods to work with functions of the form

$$-\log F(\boldsymbol{x}, \boldsymbol{y}) = \sum_i U(x_i, \boldsymbol{y}) + \sum_{i \sim j}, V(x_i, x_j, \boldsymbol{y}) \tag{10}$$

where U is a *data term* that measures the compatibility between disparities $\boldsymbol{x}_i$ and observed intensities y, and V is a *smoothness term* between disparities at neighboring locations $i \sim j$ [4].

We can construct a formal probabilistic model for stereo based on this function using a CRF with the following construction

$$-\log P(\boldsymbol{X}|\boldsymbol{y}) = -\log F(\boldsymbol{X}, \boldsymbol{y}) + \log Z(\boldsymbol{y}), \tag{11}$$

where

$$F(\boldsymbol{X}, \boldsymbol{Y}) = \prod_i \Phi(X_i, \boldsymbol{y}) \prod_{i \sim j} \Psi(X_i, X_j, \boldsymbol{y}) \tag{12}$$

such that our CRF has the standard form

$$P(\boldsymbol{X} \mid \boldsymbol{y}) = \frac{1}{Z(\boldsymbol{y})} \prod_i \Phi(X_i, \boldsymbol{y}) \prod_{i \sim j} \Psi(X_i, X_j, \boldsymbol{y}). \tag{13}$$

3.1 Specific model

The CRF of (13) is a general form. Here we present the specific CRF we used for our experiments in section 4, following the model proposed by [8]. The data term U simply measures the absolute intensity difference between corresponding pixels, summed over all color bands. We use the measure of Birchfield and Tomasi [2] for invariance to image sampling. The smoothness term V is a gradient-modulated Potts model [4, 8] with $K = 3$ parameters:

$$V(x_i, x_j, \boldsymbol{y}) = \begin{cases} 0 & \text{if } x_i = x_j \\ \theta_k & \text{if } x_i \neq x_j \text{ and } g_{ij} \in B_k \end{cases} \tag{14}$$

Here g_{ij} is the color gradient between neighboring pixels i and j, and the B_k are three consecutive gradient bins with interval breakpoints from the set $\{0, 4, 8, \infty\}$. Let Θ_v denote all the parameters.

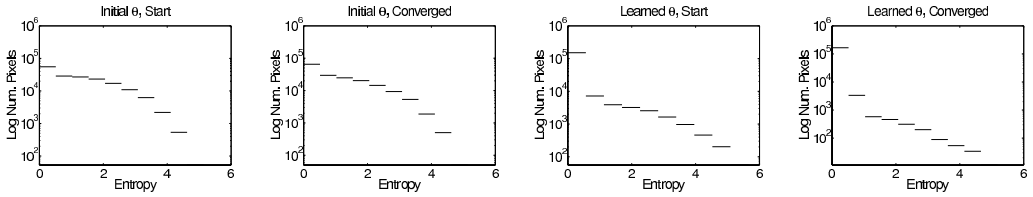

Figure 2: Histograms of approximate marginal entropies $H\left(Q\left(X_i\right)\right)$ from the variational distributions for each pixel after the first round and convergence of the mean field updates; values at the initial and learned parameters Θ_v are shown.

3.2 Parameter learning

Since the function $F\left(\boldsymbol{x}, \boldsymbol{y}\right)$ is parameterized by Θ_v, these parameters may be learned in a maximum-likelihood framework with labeled training pairs $\left(\boldsymbol{x}, \boldsymbol{y}\right)$. The objective function and gradient for one such example is

$$\begin{aligned} \mathcal{O}\left(\Theta_v\right) &= \log P\left(\boldsymbol{x} \mid \boldsymbol{y}; \Theta_v\right) && (15) \\ &= \log F\left(\boldsymbol{x}, \boldsymbol{y}; \Theta_v\right) - \log Z\left(\boldsymbol{y}\right) && (16) \\ \nabla\mathcal{O}\left(\Theta_v\right) &= \left\langle \log F\left(\boldsymbol{x}, \boldsymbol{y}; \Theta_v\right)\right\rangle_{P\left(\boldsymbol{X} \mid \boldsymbol{y}; \Theta_v\right)} - \log F\left(\boldsymbol{x}, \boldsymbol{y}; \Theta_v\right). && (17) \end{aligned}$$

The particular factorization of $F\left(\boldsymbol{x}, \boldsymbol{y}\right)$ in (10) allows the expectation in (17) to be decomposed into a sum of expectations over each term $U\left(x_i, \boldsymbol{y}\right)$ and $V\left(x_i, x_j, \boldsymbol{y}\right)$ using the corresponding marginals $P\left(X_i \mid \boldsymbol{Y}; \Theta_v\right)$ and $P\left(X_i, X_j \mid \boldsymbol{y}; \Theta_v\right)$, respectively.

In previous work [8], graph cuts was used to find the most likely configuration and a point estimate for $P\left(\boldsymbol{X} \mid \boldsymbol{y}; \Theta_v\right)$ was used to calculate the gradient. This could potentially be problematic for learning when the posterior is multi-modal or diffuse and unlike a delta function. Fortunately, the variational distribution $Q\left(\boldsymbol{X}\right)$ provides approximate marginals that may be used in a straightforward manner to calculate an approximate gradient. We show in our experiments that using these marginals for learning is better than using a point estimate in situations when there is greater uncertainty in the model.

4 Experiments

In this section we present the results of two experiments. The first compares sparse and traditional mean field methods for approximate inference, showing how sparse message passing can greatly accelerate free energy minimization. The second is an experiment in learning that compares the use of approximate marginals from sparse mean field with a point estimate of the posterior marginals from graph cuts.

As training and test data we use 6 stereo datasets with ground-truth disparities from the Middlebury stereo database (`http://vision.middlebury.edu/stereo/data`). These images are roughly 450×370 pixels and have discretized disparities with $N = 80$ states. Thus, when there are more than 600,000 messages of length N to send in any round of mean field updates for one image, shortening these to only a few states for most messages can dramatically reduce computation time.

4.1 Inference

During inference, the variational distribution $Q\left(\boldsymbol{X}\right)$ provides approximate marginals $Q\left(X_i\right)$ that may be used for calculating an approximate likelihood for a training instance. These marginals are also used to calculate the mean field updates during free energy minimization. If these marginals have many states with very low probability, discarding them will have minimal effect on the update.

Figure 2 shows histograms of the marginal entropies $H\left(Q\left(X_i\right)\right)$ during free energy minimization with two sets of parameters, the initial parameters, $\Theta_v = \mathbf{1}$, and the learned Θ_v. We initialize the variational distributions $Q\left(X_i\right)$ to uniform and perform one round of updates. Although most pixels

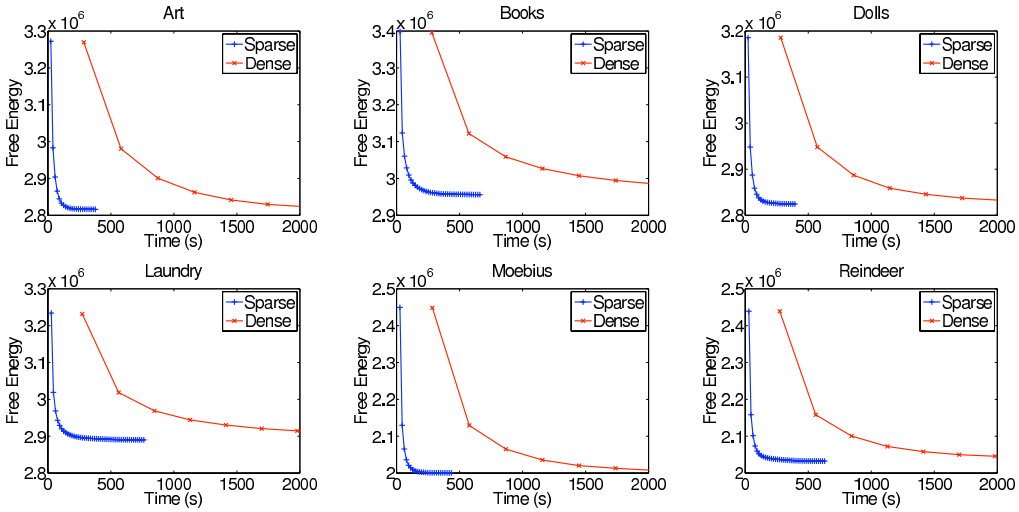

Figure 3: Comparison of CPU time for free energy minimization with sparse and dense mean field updates using learned parameters Θ_v.

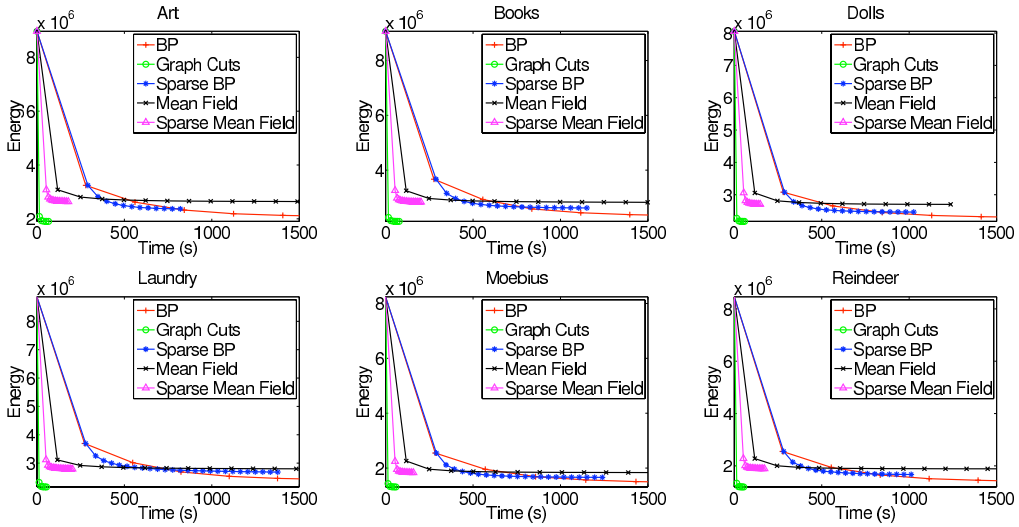

Figure 4: CPU time versus energy for graph cuts, sum-product belief propagation, and mean field using learned parameters Θ_v. The latter two do not explicitly minimize the energy; we use the *maximum posterior marginal* (MPM) prediction using the current approximate marginal at each iteration.

have very low entropy, the initial model still has several with 2-4 bits of uncertainty. Once the model parameters are learned, the marginal entropies after one round of mean field updates are much lower. By the time the mean field updates converge and free energy is minimized, only a small percentage (less than three percent) have more than a half bit of uncertainty.

Because the variational distribution has many states carrying low probability, we may greatly accelerate the update calculations by dropping these states according to our bound (9). Figure 3 shows the free energy after each round of updates for both sparse and dense mean field. In all cases, sparse mean field has nearly reached the free energy minimum before one round of dense mean field updates is done. Importantly, the minimum free energy found with sparse updates is roughly the same as its dense counterpart.

As a comparison, we show in Figure 4 the true energy $\log F(\boldsymbol{x}, \boldsymbol{y})$ on several images during each iteration of several methods. It is important to note that only graph cuts explicitly minimizes this energy, but it is demonstrative of the relative speed and behavior of the methods.

Table 1: Comparison of learning with sparse mean field and graph cuts. *Err%* is the percentage of pixels with an incorrect disparity and *RMS* is the root mean square error. See text for details.

		Training			Testing		
Training Method	Metric	Art	Books	Dolls	Laundry	Moebius	Reindeer
Initial Parameters	Err%	41.4	57.3	28.4	55.1	36.1	33.1
	RMS	12.6	16.1	8.56	13.2	12.5	11.1
Graph Cuts (One Iter.)	Err%	22.7	25.7	11.5	29.2	16.9	16.5
	RMS	6.83	10.4	3.29	7.91	9.05	4.50
Graph Cuts (Convergence)	Err%	22.0	21.7	11.7	24.9	14.7	15.8
	RMS	6.80	9.03	3.25	6.91	5.97	4.58
Sparse Mean Field (One Iter.)	Err%	16.5	18.2	10.6	22.2	12.2	16.6
	RMS	6.26	8.67	2.48	5.84	4.64	3.62

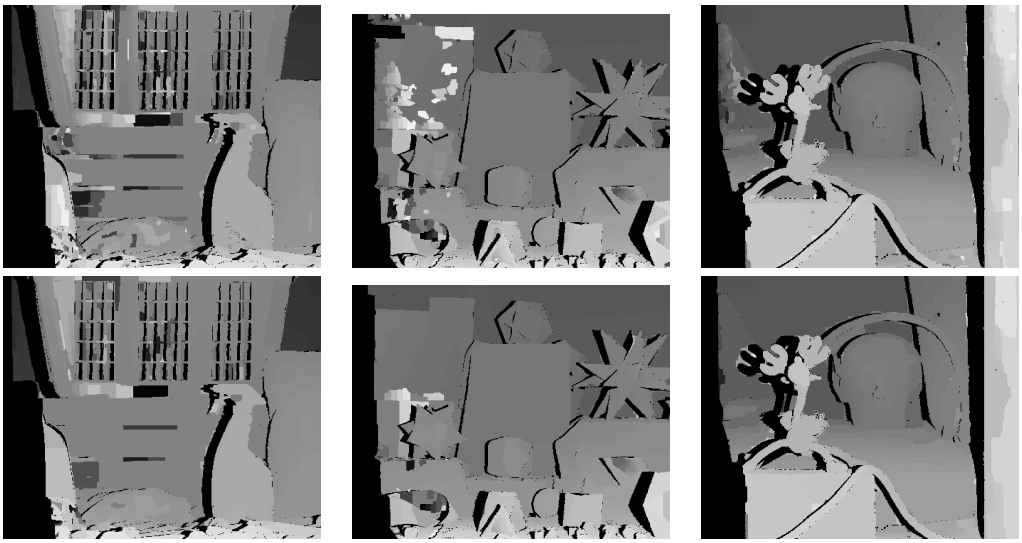

Figure 5: Test images comparing prediction (using graph cuts) after one round of learning with graph cuts (top) or sparse mean field (bottom). Occluded areas (black) are excluded during training and testing since the stereo model does not account for occlusions.

4.2 Learning

As Figure 4 shows, graph cuts does a very good job of finding a minimum energy configuration. This is useful for making a prediction in a good model. However, maximizing the log likelihood (16) for learning requires marginals on the lattice. When the model is initialized, these marginals have higher entropy (Figure 2) representing the uncertainty in the model. At this stage of learning, the point estimate resulting from an energy minimization may not be a good approximation to the posterior marginals. In fact, with initial parameters $\Theta_v = \mathbf{1}$, sparse mean field finds a lower *free* energy than the graph cuts solution, if we take the graph cuts solution as a point estimate distribution having zero entropy. We compare the results of learning using two methods for calculating the gradient: sparse mean field and graph cuts. As demonstrated earlier, the model has the highest uncertainty at the beginning of learning. It is at this point when sparse mean field has the greatest potential for improvement over graph cuts. For learning we use an initial step size of 1×10^{-4} and a simple gradient descent algorithm with an adaptive rate. For prediction evaluation, we use graph cuts to find the most probable labeling, regardless of training method.

Table 1 gives an evaluation of four parameter values. First, the results with the initial parameters are given. Next, we show the results after one step of gradient descent using graph cuts to calculate the gradient. In the third entry, the model is trained to convergence (the norm of the gradient is sufficiently small) using graph cuts as the inference method throughout learning. Finally, we show the results after one step of gradient descent using sparse mean field to calculate the gradient.

After one iteration, the training error with sparse mean field is markedly lower than that of the model fully trained with graph cuts for inference. The test RMS is uniformly lower with sparse mean field by at least one pixel. Although one image tests better in terms of absolute correctness (0/1-Loss), learning with sparse mean field provides predictions that are closer to the true disparity. Figure 5 shows the corresponding images

5 Conclusions

In this paper, we have provided a framework for sparse variational message passing. We have shown that calculating sparse updates to the approximating variational distribution can greatly reduce the time required for inference in models with large state spaces. Furthermore, we can put a variational bound on the cost of our approximation.

Not only is inference time reduced, but the resulting marginals provide better parameter estimates when used for learning in a maximimum likelihood framework. Graph cuts is often the best at finding a low energy solution in a given model. However, for model learning, a distribution over configurations is required. We find that in models where there is more uncertainty (as in the early stages of learning), sparse mean field provides a lower free energy than graph cuts. Thus, mean field methods are a better approximation to the probability distribution used in learning.

References

[1] C. Andrieu, N. de Freitas, A. Doucet, and M. Jordan. An introduction to mcmc for machine learning. *Machine Learning*, 50:5–43, 2003.

[2] S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE TPAMI*, 20(4):401–406, 1998.

[3] D. Blei, A. Ng, and M. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[4] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE TPAMI*, 23(11):1222–1239, 2001.

[5] B. J. Frey and N. Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27(9), Sept 2005.

[6] M. I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. Introduction to variational methods for graphical models. *Machine Learning*, 37:183–233, 1999.

[7] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, pages 282–289, 2001.

[8] D. Scharstein and C. Pal. Learning conditional random fields for stereo. In *Proc. CVPR*, 2007.

[9] J. Winn and C. Bishop. Variational message passing. *Journal of Machine Learning Research*, 6:661–694, 2005.

[10] J. Yedidia, W. Freeman, and Y. Weiss. *Understanding Belief Propagation and Its Generalizations*, chapter 8, pages 239–236. January 2003.

Appendix I

To compute our update we use a Lagrange multiplier form a new functional, such that

$$\widetilde{\mathcal{L}}\left(Q\left(\boldsymbol{X}\right)\right) = \mathcal{L}\left(Q\left(\boldsymbol{X}\right)\right) + \sum_{i} \lambda_i \left[\sum_{x_i} Q(x_i) - 1\right] \tag{18}$$

In this form, one can see that taking the functional derivatives with respect to each value $Q\left(x_j\right)$ we have

$$\frac{\partial}{\partial Q\left(x_j\right)} \widetilde{\mathcal{L}}\left(Q\left(\boldsymbol{X}\right)\right) = -\left\langle \log F\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\rangle_{\prod_{i \neq j} Q(X_i)} + \log Q\left(x_j\right) + N + \lambda_j, \tag{19}$$

Setting (19) to zero, we have

$$\log Q\left(x_j\right) = -\left\langle \log F\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\rangle_{\prod_{i \neq j} Q(X_i)} - N - \lambda_j, \tag{20}$$

where each λ_j thus has the form

$$\lambda_j = N + \log\left(\sum_{x_j} \exp\left\langle \log F\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\rangle_{\prod_{i \neq j} Q(X_i)}\right)$$

such that

$$Q^*\left(x_j\right) = \frac{1}{Z_j} \exp\left(\left\langle \log F\left(\boldsymbol{x}, \boldsymbol{y}\right)\right\rangle_{\prod_{i \neq j} Q(X_i)}\right) \tag{21}$$