# Comparing the mean field method and belief propagation for approximate inference in MRFs

**Yair Weiss**
Computer Science Division
UC Berkeley, 485 Soda Hall
Berkeley, CA 94720-1776
Phone: 510-642-5029
yweiss@cs.berkeley.edu

Methods for approximate inference can be divided into two broad categories: Monte-Carlo methods and deterministic methods. Perhaps the most commonly used deterministic methods are belief propagation (aka probability propagation, turbo inference) and variational methods (of which the simplest version is mean field). Surprisingly, there have been little "head to head" comparisons of these two methods. A notable exception is Mackay's work on low density parity check codes [6], where he showed that belief propagation decoding outperforms variational decoding.

In this chapter, I compare the two approaches in the context of performing inference in Markov Random Fields. I am particularly interested in the kinds of MRF inference problems that arise in low-level vision. One of the advantages of both approaches in this context is that they lead to simple, local parallel algorithms, and that makes them appealing as metaphors for low-level vision in biological systems [8].

Why have there been so little head-to-head comparisons? One of the reasons is that they are typically formulated using quite different formalisms in different contexts. Pearl's belief propagation algorithm [7] was originally formulated for small, heterogeneous directed graphs, while mean field approximations are typically formulated in large (sometimes infinite) homogeneous undirected graphs (e.g. [1]). The desired outputs are also different: in the belief propagation setting the goal is typically to estimate marginal probabilities over hidden nodes (these are the "beliefs") while in the physics setting mean field approximations may be used to predict the macroscopic behavior of the system (e.g. critical temperature, correlation lengths etc.).

I will therefore start this chapter reformulating both methods in a common language. I will then show simulation results on some toy problems. I tried to set up the toy problems such that they are small enough to make exact inference possible yet capture some of the qualities of the MRFs that occur in vision applications. I hope more head to head comparisons will be done in different graphs with different parameter regimes.

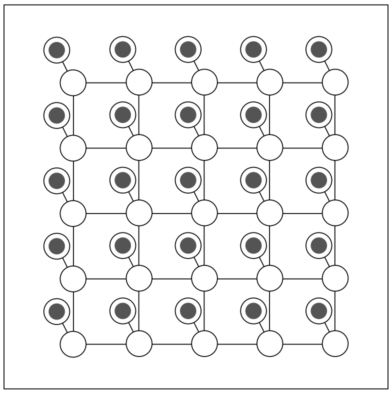


Figure 1: An example of a MRF with pairwise potentials. The hidden nodes $x_i$ are denoted with open circles and observed nodes $y_i$ are denoted with filled circles.

# 1 The setting: undirected graphs with pairwise potentials

I will assume we are dealing with pairwise Markov Random Fields (MRFs) of the form shown in figure 1. These are undirected graphs with cliques of size two. In other words, we have a set of hidden variables $x$ and a set of observed variables $y$ and the conditional probability of $x$ given $y$ is given by:

$$P(x|y) = \frac{1}{Z} \prod_{<ij>} \Psi_{ij}(x_i, x_j) \prod_i \Psi_{ii}(x_i) \tag{1}$$

where the product is taken over nodes $x_i, x_j$ that are connected in the graph. Each local potentials $\Psi_{ii}(x_i)$ capture the local interaction between $x_i$ and the observed $y_i$

Why consider only pairwise cliques? Any graphical model can be converted into this form before doing inference through a suitable clustering of nodes into large nodes [9]. Also, many of the MRFs that appear in low-level vision are of this form (e.g. [3]).

# 2 The algorithms

## 2.1 From mean field theory to an inference algorithm

There is certainly no need to introduce mean field theory in a book devoted to advanced mean field methods. I want to be clear, however, on what I mean by the mean field inference *algorithm.*. By this I mean solving for variational parameters by iterating the mean-field fixed-point equations, and using these parameters to approximate marginal probabilities.

Following the notation of [4], we are approximating the distribution $P(x|y)$ with a fully factorized distribution $q(x)$:

$$q(x) = \prod_i b_i(x_i) \tag{2}$$

where $b_i$ are variational parameters corresponding to marginal probabilities over node $x_i$. The variational parameters are obtained by minimizing the KL divergence between $P$ and $q$ which is easily shown to be equivalent to minimizing the mean-field free energy:

$$F_{MF}(\{b_i\}) = -\sum_{<ij>}\sum_{x_i,x_j} b_i(x_i)b_j(x_j)\ln\psi_{ij}(x_i,x_j) + \sum_i\sum_{x_i} b_i(x_i)\left[\ln b_i(x_i) - \ln\psi_i(x_i)\right] \tag{3}$$

subject to the constraint $\sum_i b_i(x_i) = 1$.

Setting the derivative with respect to $b_i(x_i)$ equal to zero gives:

$$b_i(x_i) = \alpha\Psi_{ii}(x_i)\exp\left(\sum_{j\in N_i}\sum_{x_j} b_j(x_j)\log\Psi_{ij}(x_i,x_j)\right) \tag{4}$$

where $\alpha$ is a normalization constant chosen so that $\sum_{x_i} b_i(x_i) = 1$ and $N_i$ denotes the set of all nodes neighboring $x_i$.

Henceforth when I talk of the MF algorithm this refers to iterating equation 4, i.e. setting:

$$b_i(x_i) \leftarrow \alpha\Psi_{ii}(x_i)\exp\left(\sum_{j\in N_i}\sum_{x_j} b_j(x_j)\log\Psi_{ij}(x_i,x_j)\right) \tag{5}$$

and then approximating the marginal probability $P(x_i|y)$ by the steady-state $b_i(x_i)$.

Note that this can be thought of as a parallel message-passing algorithm on the graph. Every node $x_i$ sends a message $b_i$ to its neighbors. This message is, in turn, based on the message it received from its neighbors in a previous time-step. I used an asynchronous random update schedule (synchronous updates tended not to converge). I initialized the beliefs with $b_i(x_i) = \Psi_{ii}(x_i)$. I have also tried initializing the beliefs with the results of running the iterations at a higher temperature (mean field annealing): this did not seem to have any effect on the final MF fixed-points.

### 2.2 Belief propagation algorithm

The belief propagation algorithm iterates:

$$m_{ij}(x_j) \leftarrow \alpha\sum_{x_i}\Psi_{ij}(x_i,x_j)\Psi_{ii}(x_i)\prod_{x_k\in N_i\backslash x_j} m_{ki}(x_i) \tag{6}$$

$$b_i(x_i) \leftarrow \alpha\Psi_{ii}(x_i)\prod_{x_k\in N_i} m_{ki}(x_i) \tag{7}$$

where $\alpha$ denotes a normalization constant and $N_i\backslash x_j$ means all nodes neighboring $x_i$, except $x_j$.

This can again be thought of as a parallel message-passing algorithm on the graph. Every node $x_i$ sends a *different* message $m_{ij}$ to its neighbors. This message is, in turn, based on the message it received from all the other neighbors in a previous time-step. The results in the paper are for a synchronous update schedule (I have also used asynchronous update and it gave similar results). The messages are initialized with $m_{ij}(x_j) = 1$.

The belief propagation iterations were originally developed by Pearl as a way of performing exact inference on trees. There was no energy function being minimized. In recent work with Jonathan Yedidia and Bill Freeman [10], we have shown that the belief propagation iterations converge to a zero-gradient location of the Bethe free energy:

$$\begin{aligned} F_\beta(\{b_{ij}, b_i\}) &= \sum_{ij} \sum_{x_i, x_j} b_{ij}(x_i, x_j) \left[\ln b_{ij}(x_i, x_j) - \ln \phi_{ij}(x_i, x_j)\right] \\ &\quad - \sum_i (d_i - 1) \sum_{x_i} b_i(x_i) \left[\ln b_i(x_i) - \ln \Psi_i(x_i)\right] \end{aligned} \tag{8}$$

where we define $b_{ij}(x_i, x_j)$ the pairwise belief of nodes $xi, x_j$ by $b_{ij}(x_i, x_j) = \alpha \phi_{ij}(x_i, x_j) \prod_{N(i)\setminus j} m_{ki}(x_i) \prod_{N(j)\setminus i} m_{lj}(x_j)$ and $\phi_{ij}(x_i, x_j) = \Psi_{ij}(x_i, x_j)\Psi_{ii}(x_i)\Psi_{jj}(x_j)$. The minimization of the free energy is subject to the normalization and marginalization constraints: $\sum_{x_i} b_i(x_i) = 1$, $\sum_{x_i} b_{ij}(x_i, x_j) = b_j(x_j)$. ($d_i$ is the number of neighbors of node $i$.)

in [10], we have shown that there is a one-to-one correspondence between zero gradient locations of $F_\beta$ and fixed-points of BP. The messages $m_{ij}(x_j)$ end up being monotonic transformations of the Lagrange multipliers that enforce the marginalization constraints.

Unlike the MF iterations which minimize $F_{MF}$ at every iteration, the BP iterations do not necessarily minimize $F_\beta$ at every iteration. In fact, for most iterations the $b_{ij}, b_i$ that BP calculates, do not satisfy the marginalization constraints. It is only at convergence of BP that the constraints are satisified and the solution is a zero-gradient point of $F_\beta$.

What, then, do the intermediate BP values mean? In [9] I showed that the beliefs $b_i(x_i)$ at any iteration are the exact marginal probabilities for the *unwrapped graph*: a graph that has the same local topology as the original MRF but is singly connected. I refer the reader to [9] for more details about the unwrapped graph. Not surprisingly, it is closely related to the Bethe lattice.

How are $F_\beta$ and $F_{MF}$ related? First note that if the joint belief of $x_i, x_j$ factorizes (i.e. $b_{ij}(x_i, x_j) = b_i(x_i)b_j(x_j)$) then the two free energies are equivalent: $F_\beta(\{b_{ij}, b_i\}) = F_{MF}(\{b_i\})$. Second, note that if the graph is singly connected, then $F_\beta$ is exact while $F_{MF}$ is not. Finally, note that if the true $P(x_i|y)$ is tightly concentrated around a single state $P(x_i|y) = \prod_i \delta(x_i - x_i^*)$ both approximate free energies become exact. That is $F_\beta(\{b_{ij}, b_i\}) = F_{MF}(\{b_i\}) = F(P(x_i|y))$ for $b_i(x_i) = \delta(x_i - x_i^*)$ and $b_{ij}(x_i, x_j) = b_i(x_i)b_j(x_j)$.

## 3 Simulation Results

I will start with some simulations on singly connected graphs. Although inference in these graphs can be done exactly, it is interesting to compare local, parallel algorithms on these structures.

Figure 2a shows a contour of a hand. *Direction of figure* (DOF) refers to which of these two regions is figure and which is ground. A local cue for DOF is convexity - given three neighboring points along the contour we prefer the DOF that makes the angle defined by those points acute rather than obtuse. Let $x_i$ be a set of nodes

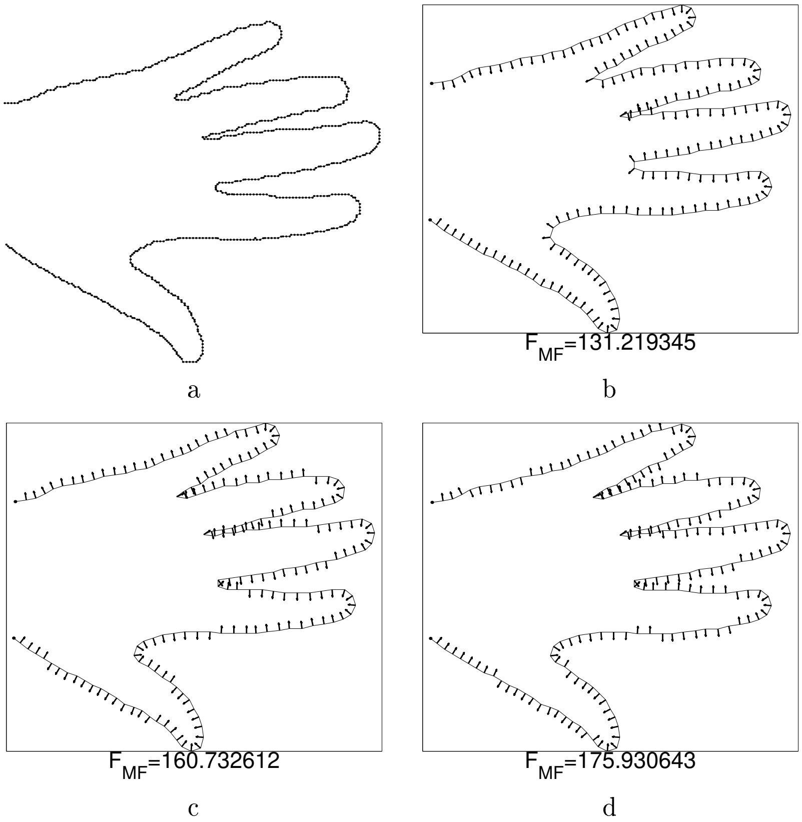



Figure 2: **a.** A contour of a hand. **b.** The predicted direction of figure using belief propagation. The arrows point in the direction of the figure (away from the ground region). **c-d.** DOF estimates using the mean field iterations. In all the simulations, the mean field iterations converged to a local minimum that was worse (in terms of $F_{MF}$) than the belief propagation estimate.

along the contour that can take on one of two possible values (the two directions perpendicular to the contour). The probability of $x$ is given by equation 1 with $\Psi_{ii}$ related to the local angle in the two directions. Let $\alpha_i$ be the local angle at location $i$ if the DOF is in direction 1 and $\beta_i = 360 - \alpha_i$ be the local angle if the DOF is in direction 2 then $\Psi_{ii}(x_i) = (\exp(-\alpha_i), \exp(-\beta_i))$. For the pairwise potential we set:

$$\Psi_{i,i+1} = \begin{pmatrix} 0.99 & 0.01 \\ 0.01 & 0.99 \end{pmatrix} \tag{9}$$

Thus $\Psi_{ii}$ reflect the preference towards convex figures and $\Psi_{i,i+1}$ reflects the preference for neighboring locations to have the same direction of figure.

Figure 2b shows the predicted direction of figure using BP. Since the graph is singly connected there is a unique fixed-point to the iterations, and the beliefs are equal to the true $P(x_i|y)$. Note that in the true posteriors all arrows point inward. Figure 2c-d shows the predicted direction of figure using MF. These are two fixed-points obtained in different runs. Note that these local minima of the mean field free energy $F_{MF}$ that are worse than the solution given by belief propagation. I reran the mean-field iterations 10 times and in all cases MF converged to a bad local minimum.

Maybe the problem is in the free energy approximation? To check this, I initialized the MF with the BP beliefs. The iterations converged in a few steps to a state where all arrows point inwards. The resulting mean field free energy $F_{MF}$ is shown in figure 2b: it is lower than the free energy obtained with any other fixed point. Thus an algorithm that was better in optimizing $F_{MF}$ compared to simply iterating the fixed point equations, would give good approximate inference for this case.

The graph in the previous example, was not only singly-connected but also very sparse: each node is connected to at most two neighbors. Will MF perform better in graphs with higher connectivity? To address this, I created the Caylee tree graph shown in figure 3. Here, all nodes (except the leaves) have four neighbors. Only the hidden nodes are shown: each hidden node has a random observed node connected to it. The nodes can take on one of two discrete values. I set $\Psi_{ij}$ to be the same as in equation 9 and the potentials $\Psi_{ii}(x_i)$ were set randomly using a uniform distribution between zero and one. These represent the interaction between $x_i$ and the observation node connected to it.

These types of graphs are used in image-processing, where a two-dimensional grid is modeled using a tree structure [5]. In these applications, the leaves of the graph correspond to pixels on a grid.

Figure 4a shows the results of BP on this problem. The figure shows the belief of being in state 1 for a node in the bottom layer: white pixels correspond to probabilities close to 1 and dark pixels correspond to probabilities close to 0. The nodes are arranged in a two-dimensional lattice for display purposes only. As in the previous example, the graph is singly connected so that the BP solution is unique and exact. Figure 4b shows the fixed-point obtained by initializing the MF with the BP beliefs. The beliefs are similar but more confident: there are less gray pixels and more black or white ones in the mean-field results.

Figure 4c,d show two results of MF starting with random initial conditions. These results are typical: different runs give different local minima and these local minima

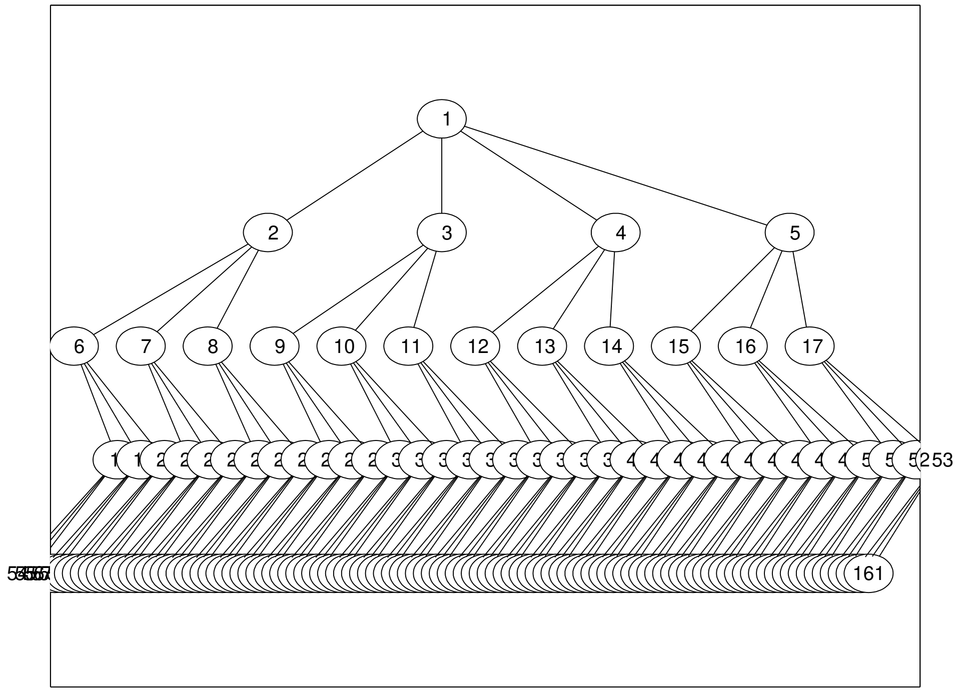



Figure 3: A Caylee tree. All hidden nodes (except leaf nodes) have degree 4, and the graph is singly connected. Each hidden node has an observation node connected to it that is not shown for clarity. Similar graphs are used in image-processing where the layers in the bottom node correspond to pixels.

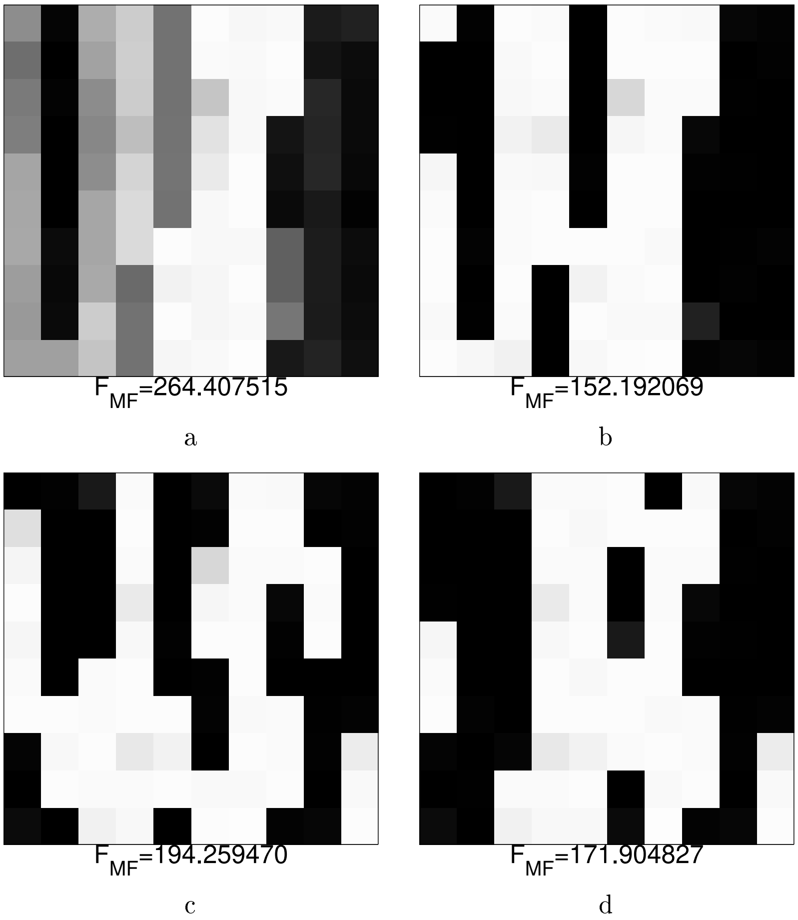



Figure 4: **a.** The beliefs at the leaf nodes of the Cayley tree obtained using belief propagation. Since the graph is singly-connected these beliefs are equal to the marginals $P(x_i|y)$. The nodes are arranged in a 2D grid for display purposes only. The brightness of a pixel is proportional to the probability of being in state 1. **b.** The approximate marginals obtained by iterating the mean-field iterations starting with the belief propagation beliefs. The marginals are qualitatively similar but overly confident: less gray pixels and more black or white ones. **c-d.** Two representative results of marginals obtained by the mean-field iterations with random initial conditions. A local minimum of the free energy is obtained.

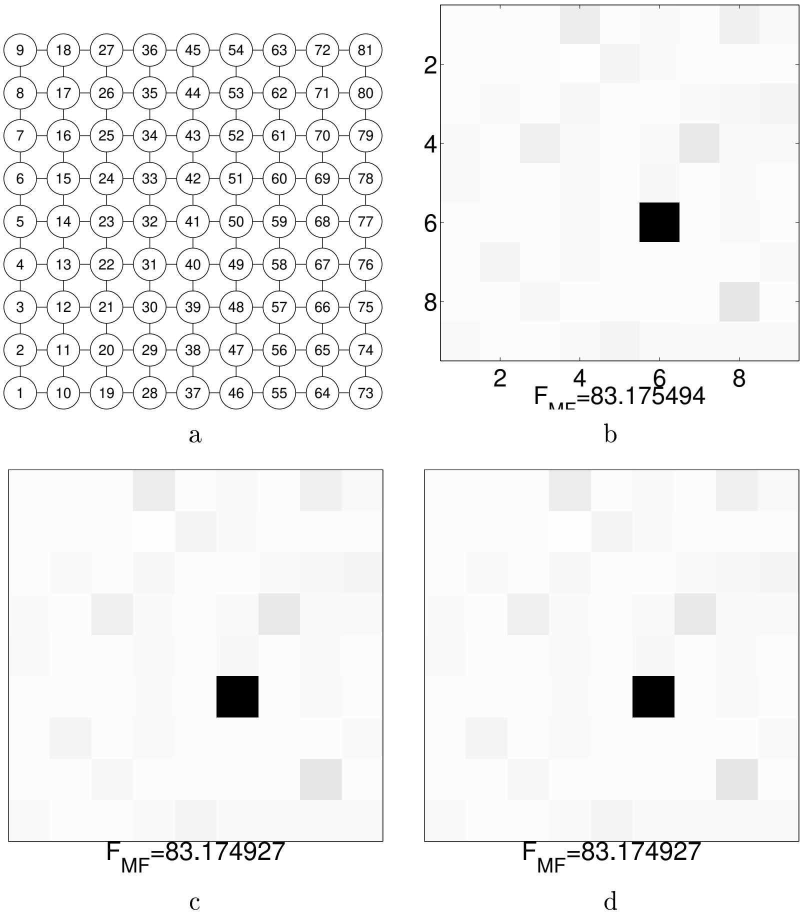



Figure 5: **a.** A toroidal $9x9$ grid. **b.** The exact marginals. **c.** The beliefs estimated by BP. **d.** the marginals estimated by MF initialized by BP.

are always worse (in terms of $F_{MF}$) than the minimum obtained by starting with the BP solution. Again, this suggests that an algorithm that could optimize $F_{MF}$ would do a good job of approximating the true posteriors for this case.

What happens if we go to multiply connected graphs? Figure 5a shows a $9x9$ toroidal 2D grid. The potentials $\Psi_{ij}$ were set in the same way as the previous examples: $\Psi_{ij}$ using equation 9 and $\Psi_{ii}$ randomly between zero and one. Figure 5b shows the exact marginal probability of being in state 1 calculated using the junction tree algorithm. Figure 5b shows the approximate marginals obtained using BP. Note that despite the large number of small loops, the BP results are nearly identical to the exact solution. Figure 5c shows the results of iterating MF starting with the BP beliefs. The iterations converge immediately and barely change.

Figure 6 shows four runs of MF starting with random initial conditions. These are typical results: in 90% of the runs, MF converged to a local minimum that was worse than the true answer. In the other 10% however, MF did converge to the solution shown in figure 5d.

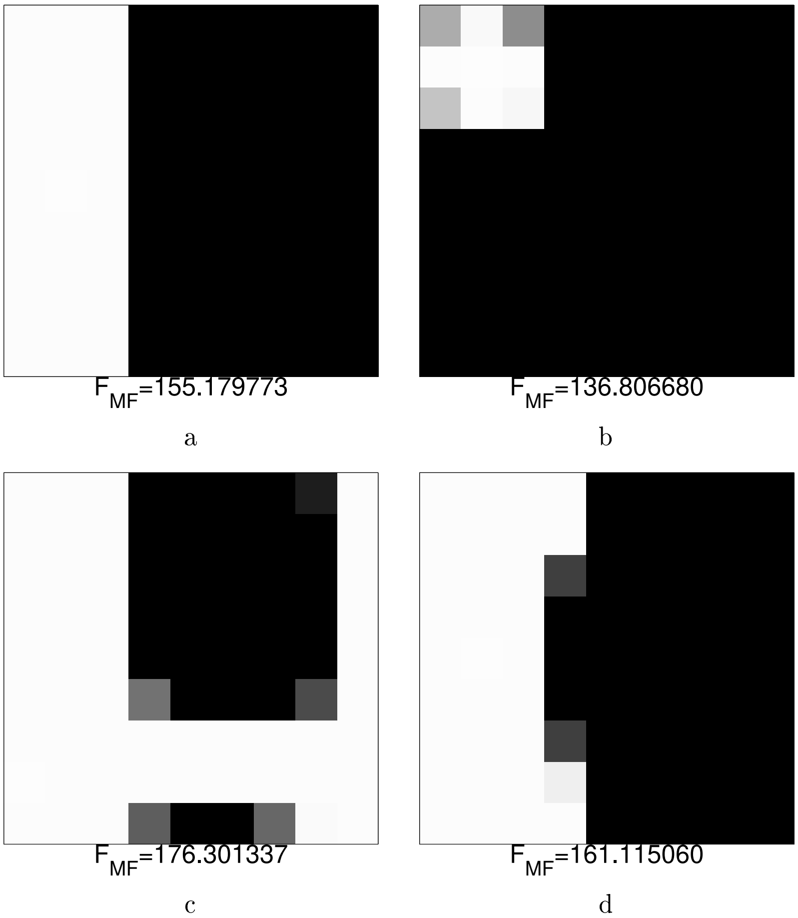
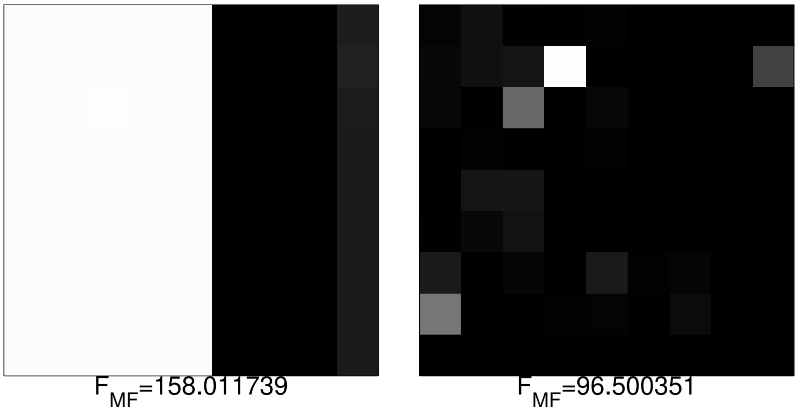



Figure 6: Four local minima obtained by MF using random initial conditions. Only in 10% of the runs were the results similar to the exact answer.




a b

Figure 7: When BP is run with random initial conditions, in 90% of the runs, the same beliefs are obtained as in figure 5. In 10% of the runs, other local minima of the Bethe free energy are obtained.

Is the BP solution the only zero gradient point of the Bethe free energy? To check this, I reran the BP iterations with random initial conditions. In 90% of the runs, the fixed-point shown in figure 5b was obtained. However, in 10% of the runs, other fixed-points were obtained. Figure 7 shows two of these fixed-points. Thus not all zero gradient points of $F_\beta$ give good approximation to the true marginals. Note that we are in the regime where the $b_{ij}$ obtained by BP are approximately factorized so that $F_{MF}$ and $F_\beta$ give similar results: the bad fixed-points of BP are also bad in terms of $F_\beta$.

## 4 Discussion

The mean field free energy approximates the true free energy with a much simpler form involving only marginals over single variables. There have therefore been many attempts to increase the accuracy of the approximation by using more complicated forms of free energies. The Bethe approximation uses marginals over pairs of variables and I am sure others writing in this volume will discuss higher order corrections to the mean field free energy.

My experience with MRFs arising in low-level vision, however, is that the main difficulty with the mean field free energy is the difficulty of optimizing it: it is simply too easy to get stuck in local minima. This was also seen in the toy problems discussed in this chapter: in all these cases, mean field iterations starting with a good initial condition gave reasonable approximation to the true marginals. This is not to say that one cannot improve on the MF approximation: typically the MF marginals are overconfident (closer to zero and one than the true marginals are). Nevertheless, for many applications where the estimated marginals are used to make a hard decision such a result would be quite satisfactory. The hard part is getting MF to converge to the global minimum rather than a local minimum.

The Bethe free energy seems to have less local minima (indeed for singly-connected graphs there is always a unique minimum), but even in cases where there are multiple local minima, the BP iterations starting from the all ones messages appears to find a good local minimum. This may be due to the fact that the BP iterations calculate exact marginal probabilities for the unwrapped graph: a graph that has the same local topology as the original MRF but is singly connected. To the best of my knowledge, no similar result can be shown for the MF iterations. I think that the tremendous success of BP in applications such as error-correcting codes and image processing [2] stems not only from the use of a more complicated free energy but also from the algorithm it uses to optimize the free energy.

**Acknowledgements**

Support by MURI-ARO-DAAH04-96-1-0341, MURI N00014-00-1-0637 and NSF IIS-9988642 is gratefully acknowledged. I thank P. Dayan, K. Murphy, J. Yedidia and B. Freeman for stimulating discussions.

## References

[1] R.H. Baxter. *Exactly solved models in statistical mechanics.* academic press, 1982.

[2] W.T. Freeman and E.C. Pasztor. Learning to estimate scenes from images. In M.S. Kearns, S.A. Solla, and D.A. Cohn, editors, *Adv. Neural Information Processing Systems 11.* MIT Press, 1999.

[3] D. Geiger and F. Girosi. Parallel and deterministic algorithms from MRFs: surface reconstruction. *IEEE Trans. PAMI*, 13(5):401–412, 1991.

[4] M.I. Jordan, Z. Ghahramani, T. Jaakkola, and L. Saul. An introduction to variational methods for graphical models. In M.I. Jordan, editor, *Learning in Graphical Models.* MIT Press, 1998.

[5] Mark R. Luettgen, W. Clem Karl, and Allan S. Willsky. Efficient multiscale regularization with application to the computation of optical flow. *IEEE Transactions on image processing*, 3(1):41–64, 1994.

[6] D.J.C. Mackay. Good error-correcting codes based on sparse matrices. *IEEE Transactions on Information Theory*, 45(2):399–431, 1999.

[7] Judea Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference.* Morgan Kaufmann, 1988.

[8] Y. Weiss. Interpreting images by propagating Bayesian beliefs. In M. Mozer, M. Jordan, and T. Petsche, editors, *Advances in Neural Information Processing Systems 9*, 1996.

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

[10] J. Yedidia, W. Freeman, and Y. Weiss. Generalized belief propagation. *NIPS 2000*, 2000. submitted.