A Mean Field Approach to MAP in Belief Networks

Yun Peng and Miao Jin
Department of Computer Science and Electrical Engineering
University of Maryland Baltimore County
Baltimore, MD 21250
ypeng@csee.umbc.edu

Abstract

The maximum *a posteriori* probability (MAP) problem is to find the most probable instantiation of all uninstantiated variables, given an instantiations of a set of variables in a Bayesian belief network (BBN). MAP is known to be *NP-hard.* To circumvent the high computational complexity, we propose in this paper a neural network approach based on the mean field theory to *approximate* the MAP problem. In this approach, a given BBN is treated as a neural network with an energy function defined in such a way that the MAP solution corresponds to the global minimum energy state. The mean field equation is then derived. We also propose a method called *resettling* to further improve the solution accuracy. A series of computer experiment shows that this approach may lead to effective and accurate solutions to MAP problems.

Key words: Bayesian belief networks, MAP problems, , mean field theory, neural networks, probabilistic inference.

1 Introduction

Bayesian belief networks (BBN) are a widely studies graphical model that can effectively represent uncertainty and probabilistic interdependence among variables [6]. In the most general form, the structure of a BBN of n variables is a directed acyclic graph of n nodes and a number of arcs. Nodes correspond to random variables, denoted $X_i, i = 1, \cdots, n$. The lower case x_i denotes an instantiation of X_i, and $x = \{x_1, \cdots, x_n\}$ an instantiation of the set of all variables. An arc $<X_i, X_j>$ represents a direct causal relation from X_i to X_j. Each node X_i is associated with a conditional probability distribution $P(X_i|\pi_i)$, where π_i is the set of all parents of X_i.

We pay a special attention to a class of simple BBN, known as the *noisy-or networks* of binary variables [6, 8]. Instead of $P(X_i|\pi_i)$, a noisy-or network captures the degree of uncertainty of the direct causal relation from X_i to X_j by a single probability measure, called the *causal strength* and denoted c_{ij}. The total probabilities needed are thus reduced from $2^{|\pi_i|}$ to $|\pi_i|$ for each node. The combined influence from π_i to X_i can be computed by

$$P(x_i = 1|\pi_i) = 1 - \prod_{X_k \in \pi_i} (1 - c_{ki}x_k). \tag{1}$$

One of the most important features of BBN is that the joint probability distribution of X can be computed from the conditional distributions as [6]

$$P(X = x) = \prod_{i=1}^{n} P(x_i|\pi_i). \tag{2}$$

BBN thus can support, at least in theory, any probabilistic inference based on Bayesian conditioning.

The maximum *a posteriori* probability (MAP) problem in BBN is to find the most probable instantiation of the set of all uninstantiated variables, given a set of instantiated variables. Stated formally,

> Let v^+ be an instantiation of $V \subseteq X$ in a BBN. Then a MAP problem is to find an instantiation u^* of $U = X - V$ such that $P(u^*|v^+) = \max_u P(u|v^+)$. Since v^+ is fixed, u^* also maximizes $P(u, v^+)$.

It has been established that both exact solutions and approximate solutions with a constant ratio bound of MAP are *NP-hard* [11, 2]. To circumvent the high computational complexity, researchers have attempted to develop methods for good *approximate* MAP solutions with sub-exponential complexity. These include the methods of Gibbs sampling [7] and genetic algorithms [1], to mention just a few.

In this paper, we take another, namely the neural network (NN) approach to approximating MAP in BBN. This is partly motivated by the fact that both BBN and NN are graphical representations of relationships among variables. This structural similarity may allow us to adopt some of the existing NN-based combinatorial optimization techniques for MAP. In particular, we apply the mean field theory to develop an efficient and accurate approximation method for MAP. We also propose a post-process called *resettling* to further improve the solution accuracy at a reasonable cost.

2 A Neural Network Formulation of MAP

In our approach, the belief network, besides serving as a static representational vehicle for causal and probabilistic knowledge of a domain, is treated as a neural network (NN), and thus a computing vehicle. Under this dual-role scheme, a node X_i becomes a processing unit whose activation $x_i(t)$ at time t is influenced by the activations of its neighbors (its parents and children) in a BBN. The strength of the influence is determined by the associated conditional probabilities (or causal strengths in case of noisy-or networks). When solving a MAP problem, activations of nodes in V are fixed (clamped) to their given instantiations, while activations of nodes in U are free to change. When the network stabilizes at time t, $U(t)$ emerges as the final solution . If V is empty, then $U = X$, and MAP becomes the problem of finding a instantiation x^* that maximizes the unconditional probability $P(X)$. For simplicity and without loss of generality, we present our model with empty V. We further assume that all variables are binary.

Like in many neural network models, we define an energy function $E(x(t))$ for MAP over the state of the system $x(t)$,

$$E(x(t)) = -\ln P(x(t)). \tag{3}$$

Next we define a quantity r_i, which plays a central role in our model. Consider two states related to $x(t)$ with respect to a chosen node X_i. These two states are the same as $x(t)$ except X_i is set to 1 in one state and 0 in another. Let $\bar{X}_i \equiv \{X_j, j \neq i\}$ denote the set of all variables excluding X_i, and $\bar{x}_i(t)$ its instantiation at time t. Then, the energy difference between the two states is

$$\Delta E_i(x(t)) = E(x_i = 1, \bar{x}_i(t)) - E(x_i = 0, \bar{x}_i(t)) = -\ln \frac{P(x_i = 1, \bar{x}_i(t))}{P(x_i = 0, \bar{x}_i(t))}. \tag{4}$$

The ratio of the probabilities of the two states is defined as $r_i(t) = P(x_i = 1, \bar{x}_i(t))/P(x_i = 0, \bar{x}_i(t))$. Since all variables, except x_i, have the same instantiations in both the numerator and denominator, $r_i(t)$ can be simplified to

$$r_i(t) \quad = \quad \frac{P(x_i = 1|\pi_i(t))}{P(x_i = 0|\pi_i(t))} \prod_{X_j \in Child(X_i)} \frac{P(x_j(t)|\pi_j(t))_{x_i=1}}{P(x_j(t)|\pi_j(t))_{x_i=0}}, \tag{5}$$

where $(.)_{x_i=1}$ and $(.)_{x_i=0}$ indicate x_i is set to 1 and 0, respectively The computation of $r_i(t)$ is thus completely local, involving only the parents and children of X_i.

By Eq. 4, we have

$$\ln r_i(t) = -\Delta E_i(t). \tag{6}$$

This leads directly to the following simple discrete model [9]. For a randomly selected node X_i at time t, update its activation by

$$x_i(t+1) = \begin{cases} 1 & \text{if } r_i(t) > 1 \\ x_i(t) & \text{if } r_i(t) = 1 \\ 0 & \text{if } r_i(t) < 1 \end{cases} \tag{7}$$

Eq. 7 causes the energy to monotonically decrease, and the system is thus guaranteed to settle to a local minimum energy state (thus a local optimal solution). To avoid being trapped to local optima, we can convert the discrete model to a stochastic model by simulated annealing (SA) as follows [9]. When node X_i is selected to update at time t, its activation is assigned to 1 with the probability

$$P(X_i \leftarrow 1|T(t)) = (1 + exp(\Delta E_i(t)/T(t)))^{-1} = (1 + r_i(t)^{-1/T(t)})^{-1}, \tag{8}$$

where $T(t)$ is the artificial temperature at t. It has been proven [4] that when the system arrives at thermal equilibrium at a given temperature, its states obey Boltzmann distribution,

$$P(X = x(t)|T(t)) = exp(-E(x(t))/T(t))/Z \tag{9}$$

where $Z = \sum_x exp(-E(x(t))/T(t))$. Substituting Eq. 3 to Eq. 9, we have

$$P(X = x(t)|T(t)) = exp(\ln P(x(t))/T(t))/Z = P(x(t))^{1/T(t)}/Z. \tag{10}$$

3 The Mean Field Model

The annealing process often takes a long time to complete. To overcome this difficulty, we turn to the mean field theory (MFT) and approximate the stochastic annealing process by a deterministic one. The MFT has been applied to various neural network applications since 1980's [10]. Its basic idea is to ignore the value fluctuations of random variables and replacing them by their average (*mean*) values. Recently, Haft *et al* have developed a mean field model for BBN [5]. We extend their model in several important aspects in the context of MAP problems.

Before presenting the derivation of our mean field equation, it is helpful to give a brief summary of Haft *et al*'s work. Their objective is to approximate $P(X)$, a complex probability distribution such as the one represented by a BBN, by a simpler distribution $Q(X)$ with independent variables. This is done by minimizing the Kullback-Leibler distance (cross-entropy) between $P(X)$ and $Q(X)$

$$D(Q||P) \quad = \quad \sum_x Q(x) \ln \frac{Q(x)}{P(x)} \equiv \left\langle \ln \frac{Q(x)}{P(x)} \right\rangle_{Q(X)}, \tag{11}$$

where $\langle . \rangle_{Q(X)}$ denotes the expectation or mean value under $Q(X)$. The mean field equation is derived by solving

$$\frac{\partial}{\partial Q(x_i)} \left[D(Q||P) - \lambda(\sum_{x_i} Q(x_i) - 1) \right] = 0. \tag{12}$$

The Lagrange parameter λ is used to accommodate the normalization constraint $\sum_{x_i} Q(x_i) = 1$, where the summation is over all possible instantiations of varialbe X_i. Solving this equation we have

$$Q(x_i) = \frac{1}{exp(1-\lambda)} exp\langle \ln P(x_i|\bar{X}_i) \rangle_{Q(\bar{X}_i)}. \tag{13}$$

The Lagrange normalization constant $exp(1-\lambda)$ can be computed by $exp(1-\lambda) = \sum_{x_i} exp\langle \ln P(x_i|\bar{X}_i)\rangle_{Q(\bar{X}_i)}$. When all variables are binary,

$$exp(1-\lambda) = exp\langle \ln P(x_i = 1|\bar{X}_i)\rangle_{Q(\bar{X}_i)} + exp\langle \ln P(x_i = 0|\bar{X}_i)\rangle_{Q(\bar{X}_i)},$$

and the mean field equation Eq. 13 becomes

$$Q(x_i = 1) = (1 + (\frac{exp\langle \ln P(x_i = 1|\bar{X}_i)\rangle_{Q(\bar{X}_i)}}{exp\langle \ln P(x_i = 0|\bar{X}_i)\rangle_{Q(\bar{X}_i)}})^{-1})^{-1}. \tag{14}$$

Several problems prevent us from applying the MF equations Eqs. 13 and 14 to MAP problems. First of all, these equations are used to find a *distribution* $Q(X)$ that approximates another distribution $P(X)$, they do not provide a mechanism directly applicable to finding an *optimal instantiation* to X or its approximation under $Q(X)$. Secondly, the calculation of $\langle \ln P(x_i|\bar{X}_i)\rangle_{Q(\bar{X}_i)}$ is computationally expensive as it involves calculating $P(x_i|\bar{X}_i)Q(\bar{X}_i)$ for every possible instantiation of $\bar{X}_i$.

To overcome the first difficulty, we incorporate annealing into the mean field approximation so that $Q(X|T)$ (approximately) obeys the Boltzmann distribution at any given temperature T. This is done by replacing $P(X)$ and $Q(X)$ with $P(X|T)$ and $Q(X|T)$, respectively, in Eq. 14.

$$\begin{aligned} Q(x_i = 1|T) \quad &= \quad (1 + (\frac{exp\langle \ln P(x_i = 1|\bar{X}_i, T)\rangle_{Q(\bar{X}_i|T)}}{exp\langle \ln P(x_i = 0|\bar{X}_i, T)\rangle_{Q(\bar{X}_i|T)}})^{-1})^{-1} \\ &= \quad (1 + (\frac{exp\langle \ln P(x_i = 1, \bar{X}_i|T)\rangle_{Q(\bar{X}_i|T)}}{exp\langle \ln P(x_i = 0, \bar{X}_i|T)\rangle_{Q(\bar{X}_i|T)}})^{-1})^{-1}). \end{aligned} \tag{15}$$

This is our MF equation with annealing, which has Eq. 14 of Haft *et al* as a special case of $T = 1$. In other words, Eq. 14 only approximates the distribution $P(X)$ of the given original BBN while our MF equation approximate $P(X|T)$ at all T during the annealing process. Since $P(X|T)$ obeys the Boltzmann distribution, the system will approach the (approximately) global optimal solution when T slowly decreases toward zero.

Next we simplify the computation of Eq. 15 for noisy-or Networks. From Eqs. 10 and 3, we have

$$\ln P(x_i, \bar{X}_i|T) = \ln P(x_i, \bar{X}_i)^{\frac{1}{T}}/Z = \ln \prod_{X_j \in X} P(x_j|\pi_j)^{\frac{1}{T}}/Z = \frac{1}{T} \sum_{X_j \in X} \ln P(x_j|\pi_j) - \ln Z.$$

Substituting it into Eq. 15, we have

$$\begin{aligned}
Q(x_i = 1|T) &= (1 + (\frac{exp\langle \frac{1}{T} \sum_{X_j \in X} \ln P(x_j|\pi_j) - \ln Z\rangle_{Q(\bar{X}_i|T), x_i=1}}{exp\langle \frac{1}{T} \sum_{X_j \in X} \ln P(x_j|\pi_j) - \ln Z\rangle_{Q(\bar{X}_i|T), x_i=0}})^{-1})^{-1} \\
&= (1 + (\prod_{X_j \in X} \frac{exp(\frac{1}{T}\langle \ln P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=1}/Z)}{exp(\frac{1}{T}\langle \ln P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=0}/Z)})^{-1})^{-1} \\
&= (1 + (\prod_{X_j \in X} \frac{exp\langle \ln P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=1}}{exp\langle \ln P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=0}})^{-\frac{1}{T}})^{-1}. \qquad (16)
\end{aligned}$$

The computation of $\langle \ln P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=1}$ for every $X_j \in X$ is still expensive when π_j's are large. To simplify this computation, the following approximation assumption is made.

$$\langle \ln P(x_j|\pi_j)\rangle_{Q(X|T)} \approx \ln\langle P(x_j|\pi_j)\rangle_{Q(X|T)}. \qquad (17)$$

This assumption can be justified by the following arguments. By Jesen's inequality, $\langle \ln P(x_j|\pi_j)\rangle_{Q(X|T)} \leq \ln\langle P(x_j|\pi_j)\rangle_{Q(X|T)}$. Therefore, the approximation errors caused by Eq. 17 to both the numerator and the denominator of Eq. 16 can partially cancel each other since both are an increase from their respective original values. Also, the errors will be reduced when T is approaching zero. This is because $P(X|T)_{Q(X|T)}$ approximately obeys Boltzmann distribution, and thus when T is small, a single $P(x_j|\pi_j^*)_{Q(\pi_j|T)}$ with π_j^* consistent with the most probable x^* will likely dominate strongly the rest of $P(x_j|\pi_j)_{Q(\pi_j|T)}$'s with all other possible instantiatins of π_j. Therefore, $\langle \ln P(x_j|\pi_j)\rangle_{Q(X|T)} \approx \ln P(x_j|\pi_j)_{x^*} \approx \ln\langle P(x_j|\pi_j))\rangle_{Q(X)}$.

With this new assumption, Eq. 16 changes to

$$\begin{aligned}
Q(x_i = 1|T) &\approx (1 + (\prod_{X_j \in X} \frac{exp(\ln\langle P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=1})}{exp(\ln\langle P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=0})})^{-\frac{1}{T}})^{-1} \\
&= (1 + (\prod_{X_j \in X} \frac{\langle P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=1}}{\langle P(x_j|\pi_j)\rangle_{Q(\bar{X}_i|T), x_i=0}})^{-\frac{1}{T}})^{-1}. \qquad (18)
\end{aligned}$$

From Eq. 1 of noisy-or networks, we have,

$$\langle P(x_i = 1|\pi_i)\rangle_{Q(\bar{X}_i|T)} = \langle 1 - \prod_{X_k \in \pi_i} (1 - c_{ki} x_k)\rangle_{Q(\bar{X}_i|T)} = 1 - \prod_{X_k \in \pi_i} (1 - c_{ki}\langle x_k\rangle_{Q(\bar{X}_i|T} = 1 - \prod_{X_k \in \pi_i} (1 - c_{ki}\hat{x}_k). \qquad (19)$$

The second equality comes from the fact that all x_k are independent under $Q(.)$. For brievity, in the last equality we denote the expectation $\langle x_k\rangle_{Q(\bar{X}_i|T)}$ by $\hat{x}_k$. Similarly, we have

$$\langle P(x_i = 0|\pi_i)\rangle_{Q(\bar{X}_i|T)} = \prod_{X_k \in \pi_i} (1 - c_{ki}\hat{x}_k). \qquad (20)$$

And for $j \neq i$, we have

$$\begin{aligned}
\langle P(x_j|\pi_j)\rangle_{Q(X|T)} &= Q(x_j = 1)(1 - \prod_{X_k \in \pi_j} (1 - c_{kj}\hat{x}_k))_{x_i} + Q(x_j = 0) \prod_{X_k \in \pi_j} (1 - c_{kj}\hat{x}_k)_{x_i} \\
&= \hat{x}_i(1 - \prod_{X_k \in \pi_j} (1 - c_{kj}\hat{x}_k))_{x_i} + (1 - \hat{x}_i) \prod_{X_k \in \pi_j} (1 - c_{kj}\hat{x}_k)_{x_i} \qquad (21)
\end{aligned}$$

where in the last equality we denote $Q(x_i = 1)$ by $\hat{x}_i$, the mean value of X_i, as all variables are binary.

Substituting Eqs. 19, 20, and 21 into 18, and changing $\approx$ to $=$, we have our MF model for MAP

$$\hat{x}_i(t) = (1 + \hat{r}(t)^{-1/T})^{-1}, \quad \text{where} \qquad (22)$$

$$\hat{r}_i(t) = \frac{1-\prod_{X_k\in\pi_i}(1-c_{ki}\hat{x}_k(t))}{\prod_{X_k\in\pi_i}(1-c_{ki}\hat{x}_k(t))} \times \prod_{X_j\in Child(X_i)} \frac{\hat{x}_j(t)(1-\prod_{X_k\in\pi_j}(1-c_{kj}\hat{x}_k(t))_{x_i=1}+(1-\hat{x}_j(t))\prod_{X_k\in\pi_j}(1-c_{kj}\hat{x}_k(t))_{x_i=1}}{\hat{x}_j(t)(1-\prod_{X_k\in\pi_j}(1-c_{kj}\hat{x}_k(t))_{x_i=0}+(1-\hat{x}_j(t))\prod_{X_k\in\pi_j}(1-c_{kj}\hat{x}_k(t))_{x_i=0}}. \quad (23)$$

Eq. 22 looks very much like Eq. 8 of the SA model except that 1) all binary variables x_k are replaced by their mean value $\hat{x}_k$, and 2) the activation rule is used to determine the value of $\hat{x}_i(t)$, not the probability of x_i being set to 1, and the process is thus deterministic, not stochastic.

4 Resettling Process

Suppose the final solution $\hat{x}$ obtained by the MF model is not the global optimal solution x^*, then at least one node, say X_k, has been instantiated to a wrong value in the computation. That is, the correct instantiation of X_k should have been $1 - \hat{x}_k$, not $\hat{x}_k$. This leads to the following *resettling process*. After $\hat{x}$ is computed, we perform n more computations. For the k^{th} of the n computations, we randomly assign initial values to all variables except X_k, which is assigned and *clamped* to $1 - \hat{x}_k$. The computation then proceeds according to the MF model until the system settles down to a stable state $\hat{x}^k$ (thus the name of *re*-settling). After the n resettling processes are completed, the n resulted solutions $\hat{x}^k$ are compared with each other and with $\hat{x}$, and the best one is selected as the final answer. The price we pay for a possibly better result than $\hat{x}$ is an increase of computation time by a factor of n.

5 The Computer Experiments

A series of computer simulation experiments were conducted over two examples of noisy-or BBN to test the validity, accuracy, and efficiency of the proposed MF model. The first BBN was a two-layer network of 10 roots and 10 leaves, with 40 randomly generated arcs between them. The second one, was a larger network with a more complex structure, known as the Alarm belief network [3]. It has 37 nodes (11 of which are leaves), 46 arcs, and depth 10. Although V can be any subset of X, we chose V to be the set of all leaf nodes of a network in these experiments. Each of the $2^{|V|}$ possible instantiations of V was used as v^+ to define a MAP problem instance. Therefore, 1024 problem instances for the first BBN and 2048 for the second were tested. For efficiency reason, we chose to simultaneously update activations of *all* node at each iteration, instead of updating one node at a time as in the discrete and SA models. The results from the MF model were compared with those of the simple discrete model (Eq. 7) and SA model (Eq. 8). We report the results of resettling with only the MF model because resettling had only a marginal effect with the discrete model and the SA model with resettling could not be completed within the time frame we had.

The results from the three models were compared with the first, second, and third globally best instantiations of $U = X - V$ obtained by some exact method. The experiment results with the two BBN, together with their average computing time, were given in Tables 1 and 2, respectively, where the failure rate is defined as the percentage of cases whose results were worse than the third globally best solutions, and rows A, B, C, and D in the two tables are for the discrete model, the SA model, the MF model without resettling, and MF model with resettling, respectively.

Table 1: Results with the 10 by 10 network.

	1st	2nd	3rd	others	failure rate	average time
A	370	58	39	557	54.4%	6ms
B	780	61	19	164	16%	1014ms
C	888	56	7	73	7.1%	31ms
D	1023	0	0	1	0.098%	312ms

Table 2: Results with the Alarm network

	1st	2nd	3rd	others	failure rate	average time
A	256	51	49	1682	82.1%	7ms
B	1101	91	94	762	37.2%	2511ms
C	1155	39	8	846	41.3%	49ms
D	1665	42	191	150	7.3%	1289ms

As expected, the experiments results show that the discrete model was the fastest but with the worst accuracy, with the failure rates too high to be acceptable in most applications. Although the SA model achieved good solution accuracy, its high computing time may limit its applicability with large BBNs. In contrast, the MF model achieves solution accuracy comparable to that of the SA model with only a small fraction of the computing time.

The experiment results also suggest that the resettling process with the MF model was well worth its cost as it improved the solution accuracy significantly. The results for the 10 by 10 network is very impressive as it obtained the globally best solutions for all but one of the 1024 cases. More importantly, resettling was able to reduce the failure rate for the more complex Alarm network from as high as 41.3% to only 7.3% while the computing time was still less than half of the time the SA model took without resettling. This suggests that the MF model with resettling can be applied to solve MAP problems in large, complex BBN with acceptable accuracy and time.

6 The Discussions

In this paper we took a neural network approach to solving the difficult probabilistic inference problem of MAP in belief networks. In particular, we combined the mean field theory and the simulated annealing technique to approximate the solution for MAP problems. As a mean field approximation, it replaces the stochastic variables with their mean values, thus avoiding time-consuming process of reaching thermal equilibrium. At the same time, as with the simulated annealing, the system's states (approximately) obeys the Boltzmann distribution, it is thus guaranteed, at least in theory, that the system will settle down to a state corresponding to a global optimal solution. The resulting MF model for MAP is shown by experiments to be both fast and fairly accurate. Moreover, With a tolerable additional cost, the solution accuracy of the MF model can be further improved significantly by the resettling process.

One direction of future research is to investigate the scalability of the proposed MF model. Further theoretical analysis and additional experiments are needed to see if and to what extent the *time complexity* and the *solution accuracy* are affected by the size and the structural complexity of the network.

Another promising direction of future research is to extend this model to develop MFT-based methods to approximate other types of time-consuming probabilistic inference as well as BBN learning.

Acknowledgement: This work was supported in part by NSF Grant IRI-9309136.

References

[1] A.M. Abdelbar and S.M. Hedetniemi "A Parallel Hybrid Genetic Algorithm Simulated Annealing Approach to Finding Probable Explanations on Bayesian Belief Networks", in *Proceedings IEEE International Conference on Neural Networks*, **1**, 1997, 450-455.

[2] A.M. Abdelbar and S.M. Hedetniemi "Approximating MAP for Belief Networks is NO-hard and Other Theorems", *Artificial Intelligence*, bf 102(1-2), 1998, 21-38.

[3] G. Cooper and E. Herskovits, "A Bayesian Method for the Induction of Probabilistic Networks from Data", *Machine Learning*, **9**, 1992, 309-347.

[4] S. Geman and D. Geman, "Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of a Images", *IEEE Transactions on Pattern REcognition and Machine Intelligence*, **PAMI-6**, 1984, 721-741.

[5] M. Haft, R. Hofmann, and V. Tresp, "Mean Field Inference in a General Probabilistic Setting", *Proceedings of The 7th International Workshop on Artificial Intelligence And Statistics*, Ft. Lauderdale, Jan. 1999.

[6] J. Pearl, *Probabilistic Reasoning in Intelligent S ystems: Networks of Plausible Inference*, Morgan Kauffman Publishers, San Mateo, CA, 1988.

[7] J. Pearl, "Evidential Reasoning Using Stochastic Simulation of Casal models", *Artificial Intelligence* **32**, 1987, 245-257.

[8] Y. Peng and J. Reggia, *Abductive Inference Models for Diagnostic Problem-Solving*, Springer-Verlag, New York, 1990.

[9] Y. Peng, M. Jin, and K. Chen, "A Neural Network Approach to MAP in Belief Networks", in *Proceedings of the International Joint Conference in Neural Networks*, Washington, DC, July 10-16, 1999, #335.

[10] C. Peterson and J.R. Anderson, "A Mean Field Theory Learning Algorithm for Neural Networks", *Complex Systems*, **1**, 1987, 995-1019.

[11] S.E. Shimony, "Finding MAPs for belief networks is NP-hard", *Artificial. Intelligence*, **68**, 1994, 399-410.