

Mean Field Theory based on Belief Networks for Approximate Inference

Wim Wiegerinck*[†] and David Barber[‡]

Stichting Neurale Netwerken, University of Nijmegen
Nijmegen, The Netherlands

Abstract

Exact inference in large, densely connected belief networks is computationally intractable, and approximate schemes are therefore of great importance. In the context of approximate inference in sigmoid belief networks, mean field theory has received much interest. In this method the exact log-likelihood is bounded from below using a mean field approximating distribution. In the standard mean field theory, the approximating distribution is assumed to be factorial. In this paper we propose to use a (tractable) belief network as an approximating distribution. We show that belief networks fit very well into mean field theory, and no additional bounds are required. We derive mean field equations which provide an efficient iterative algorithm to optimize the parameters of the approximating distribution. Simulation results on an inference problem indicates a considerable improvement over existing mean field methods.

1 Introduction

Belief networks provide a rich framework for probabilistic modeling and reasoning [1]. Due to the directed graphical structure in these models, exact inference requires only local computations, in contrast to undirected approximations. In practice, this means that networks of reasonable size are tractable, as long as the ‘neighborhoods’ are small. However, the complexity of inference scales exponentially with the size of the neighborhoods and, as a result, large densely connected networks can only be handled with approximate methods. As the size of conditional probability tables also scales with the size of the neighborhoods, it is convenient to parametrize large models in a compact way, e.g. as noisy-OR networks[1] or sigmoid belief networks[2].

In mean field approximations of large networks this compact parametrization is exploited[3]. However, we will show that the graphical structure of the model can be pushed much further to provide a more accurate, bounded approximation for inference, without incurring much more computational overhead.

The paper is organized as follows. In section 2 we review the standard mean field theory using factorized models to approximate sigmoid belief networks, as

*<http://www.mbfys.kun.nl/~wimw>

[†]This research is supported by the Technology Foundation STW, applied science division of NWO and the technology programme of the Ministry of Economic Affairs

[‡]<http://www.mbfys.kun.nl/~davidb> Supported by the Real World Computing Project.

proposed by [3]. In section 3 we show how this theory can be extended in a natural way using belief networks. In section 4 we apply the method on a toy problem from [3, 4].

2 Mean Field Theory

Given a probability model $P(S)$, we wish to compute the likelihood $P(S_V)$ that the set of visible variables $V \equiv v_1, \dots, v_N$ is in state $S_V \equiv S_{v_1}, \dots, S_{v_N}$. This involves the summation over exponentially many states of the remaining (hidden) variables H , $P(S_V) = \sum_{\{S_H\}} P(S_V, S_H)$. In a sigmoid belief network [2] with binary units ($S_i = 0/1$) the probability that the variable $S_i = 1$ is

$$P(S_i = 1 | \text{pa}(S_i)) = \sigma \left(\sum_j J_{ij} S_j + h_i \right) \quad (1)$$

where $\sigma(z) \equiv (1 + e^{-z})^{-1}$. The parents of S_i are denoted $\text{pa}(S_i)$; the biases are h_j and the weights are J_{ij} such that $J_{ij} = 0$ for $S_j \notin \text{pa}(S_i)$. Mean field theory for such networks is based on the following lower bound¹ on the log likelihood [3] for any approximating distribution $Q(S_H)$ and parameters, $\xi_i \in \mathbb{R}$,

$$\begin{aligned} \ln P(S_V) \geq \mathcal{F}_V[Q, \xi] &= \sum_{ij} J_{ij} \langle S_i S_j \rangle_Q + \sum_i (h_i - \sum_j \xi_j J_{ji}) \langle S_i \rangle_Q - \sum_i h_i \xi_i \\ &\quad - \sum_i \ln \langle e^{-\xi_i z_i} + e^{(1-\xi_i) z_i} \rangle_Q - \sum_{\{S_H\}} Q(S_H) \ln Q(S_H) \end{aligned} \quad (2)$$

where $z_i = \sum_j J_{ij} S_j + h_i$ and $\langle \cdot \rangle_Q$ is the average with respect to the mean field distribution $Q(S_H)$. Since this inequality holds for any Q and ξ , one can make the bound as tight as possible by optimizing $\mathcal{F}_V[Q, \xi]$ with respect to Q and ξ .

2.1 Factorized models

To make the bound $\mathcal{F}_V[Q, \xi]$ tractable, standard mean field theory restricts itself to factorized models

$$Q(S_H) = \prod_{i \in H} Q(S_i) \quad (3)$$

Using the parametrization, $q_i \equiv Q(S_i = 1)$, all terms of (2) are easy to compute, e.g. $\langle S_i \rangle_Q = q_i$, $\langle S_i S_j \rangle_Q = q_i q_j$, and $\langle e^{-\xi_i z_i} \rangle_Q = e^{-\xi_i h_i} \prod_j (1 - q_j + q_j e^{-\xi_i J_{ij}})$. The entropy factorizes nicely into a tractable sum of entropies per site,

$$\sum_{\{S_H\}} Q(S_H) \ln Q(S_H) = \sum_i q_i \ln q_i + (1 - q_i) \ln(1 - q_i) \quad (4)$$

¹This bound results from the Kullback-Leibler divergence between the true distribution and an approximating distribution on the hidden units.

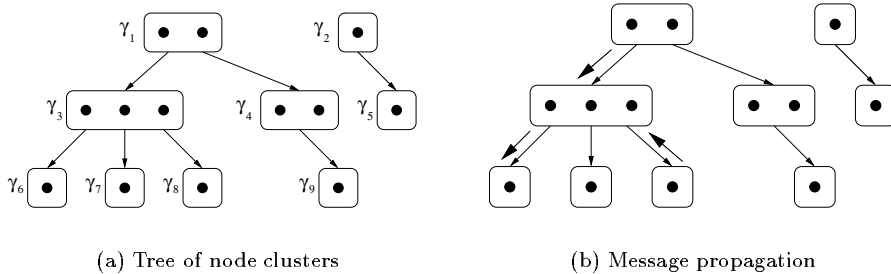


Figure 1: Graphical tree structure of the approximating distributions Q .

A fast, two step iterative procedure to optimize $\mathcal{F}_V[Q, \xi]$ with respect to Q and ξ , is the following [3]. First the gradient with respect to the q_i 's is set equal to zero, which yields the mean field equations

$$q_i = \sigma \left(h_i + \sum_j [J_{ij}q_j + J_{ji}(q_j - \xi_j) + K_{ji}] \right) \quad (5)$$

where $J_{ij}^s = J_{ij} + J_{ji}$, and

$$K_{ij} = -\frac{\partial}{\partial q_j} \ln \left\langle e^{-\xi_i z_i} + e^{(1-\xi_i)z_i} \right\rangle_Q. \quad (6)$$

The q_i 's are optimized by iteration of the mean field equations (5) while the ξ_i 's remain fixed. In the second step, the ξ_i 's are optimized using (2) while the q_i 's remain fixed. Note that the optimization of each ξ_i can be performed by one-dimensional optimizations.

Recently [5, 4] have proposed to improve the bound (2) using a mixture model. Unfortunately, the entropy term $\sum_{\{S_H\}} Q(S_H) \ln Q(S_H)$ is not tractable for mixtures, and an additional bound is needed. In the following section we propose an alternative class of models which generalize the standard mean field theory straightforwardly, without requiring any additional bound.

3 Mean Field Theory Using Belief Networks

We show here how mean field theory can be extended in a natural way by using tractable belief networks for the approximating distribution Q . For convenience we restrict ourselves to trees of clusters of nodes²,

$$Q(S_H) = \prod_{\gamma} Q(S_{\gamma} | S_{\text{pa}_{\gamma}}) \quad (7)$$

²This does not preclude disconnected branches - see fig. 1(a).

where disjoint clusters $\gamma \subset H$ form a partition of H . The clusters $\{\gamma\}$ are ordered $\gamma_1 < \gamma_2 < \dots < \gamma_k$ and $\text{pa}\gamma$ is either empty, or contained in *one* of the predecessors of γ (see fig. 1(a)). The tractability of these models is determined by the size of the γ 's, which we assume to be small. For each cluster γ , the conditional distributions in (7) contain $(2^{|\gamma|} - 1) \times 2^{|\text{pa}\gamma|}$ parameters.

3.1 Computing the mean field bound

We now show that $\mathcal{F}_V[Q, \xi]$ in (2) is tractable and computable by *local* computations and simple message propagations (we refer the reader to standard texts, such as [1]). First of all, the marginal probability distributions on γ , can be computed using the recursion

$$Q(S_\gamma) = \sum_{S_{\text{pa}\gamma}} Q(S_\gamma | S_{\text{pa}\gamma}) Q(S_{\text{pa}\gamma}) \quad (8)$$

The terms $\langle S_i \rangle_Q$ with $i \in \gamma$, and $\langle S_i S_j \rangle_Q$ with $i, j \in \gamma$ can be computed by summation of $Q(S_\gamma)$ over all states S_γ with $S_j = 1$ and $S_i = 1, S_j = 1$ respectively. To compute terms of the form $\langle S_i S_j \rangle_Q$ for which $i \in \gamma_i$ and $j \in \gamma_j$, with $\gamma_i \neq \gamma_j$ having a common ancestor γ_0 , standard message passing algorithms can be used[1], see fig. 1(b).

We write the terms $\langle e^{-\xi_i z_i} \rangle_Q = e^{-\xi_i h_i} \sum_{\{S_H\}} R(S_H)$, where $R(S_H) = \prod_\gamma R(S_\gamma | S_{\text{pa}\gamma})$ and $R(S_\gamma | S_{\text{pa}\gamma}) \equiv Q(S_\gamma | S_{\text{pa}\gamma}) \exp(\sum_{j \in \gamma} -\xi_j J_{ij} S_j)$. Note that R and Q have similar graphical structures, and we can therefore use message propagation techniques again to compute $\langle e^{-\xi_i z_i} \rangle_Q$. The last term to consider is the entropy term, which decouples into a sum of averaged entropies per γ ,

$$\sum_{\{S_H\}} Q(S_H) \ln Q(S_H) = \sum_\gamma \sum_{S_{\text{pa}\gamma}} Q(S_{\text{pa}\gamma}) \sum_{S_\gamma} Q(S_\gamma | S_{\text{pa}\gamma}) \ln Q(S_\gamma | S_{\text{pa}\gamma}) \quad (9)$$

We conclude that all terms in $\mathcal{F}_V[Q, \xi]$ are tractable without the need of additional approximations.

3.2 Mean field equations

To derive mean field equations, we differentiate (2) with respect to the parameters $Q(S_\gamma^i | S_{\text{pa}\gamma})$, *i.e.*, the i -th state of the conditional probability distribution for cluster γ (of which there are $n_\gamma = 2^{|\gamma|} - 1$ states), analogous to section 2.1,

$$\begin{aligned} & (Q(S_\gamma^1 | S_{\text{pa}\gamma}) \dots Q(S_\gamma^{n_\gamma} | S_{\text{pa}\gamma})) \\ & = \bar{\sigma} \left(\frac{\sum_{i,j} J_{ij}^s \nabla_\gamma \langle S_i S_j \rangle_Q + \sum_i (h_i - \sum_j \xi_j J_{ji}) \nabla_\gamma \langle S_i \rangle_Q + K + L}{Q(S_{\text{pa}\gamma})} \right) \end{aligned} \quad (10)$$

where the gradient ∇_γ is with respect to $Q(S_\gamma^1 | S_{\text{pa}\gamma}) \dots Q(S_\gamma^{n_\gamma} | S_{\text{pa}\gamma})$. Furthermore,

$$K = -\nabla_\gamma \sum_i \ln \langle e^{-\xi_i z_i} + e^{(1-\xi_i) z_i} \rangle_Q \quad (11)$$

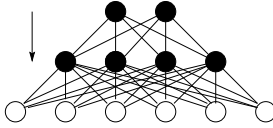


Figure 2: Graphical structure of the 2-4-6 nodes sigmoid belief network. Open circles: visible units V . Filled circles: hidden units H .

and

$$L = - \sum_{\gamma'} \sum_{S_{\text{pa}\gamma'}} [\nabla_{\gamma} Q(S_{\text{pa}\gamma'})] \sum_{S_{\gamma'}} Q(S_{\gamma'} | S_{\text{pa}\gamma'}) \ln Q(S_{\gamma'} | S_{\text{pa}\gamma'}) \quad (12)$$

$\vec{\sigma}(x_1, \dots, x_{n_\gamma}) \equiv (1 + \sum_j e^{x_j})^{-1} (e^{x_1}, \dots, e^{x_{n_\gamma}})$ is the generalized sigmoid function. Finally, $Q(S_{\text{pa}\gamma}) \equiv 1$ if $\text{pa}\gamma = \phi$. The explicit evaluation of the gradients can be performed efficiently, again using standard message propagation.

To optimize the bound, we again use a two step iterative procedure as described in section 2.1. Note that the optimization with respect to the ξ_i 's in the second step remains decoupled.

4 Simulations

To compare the method with existing mean field approximations [3, 4], we examined a toy benchmark problem in a three layer (2-4-6 nodes) sigmoid belief network in which the last 6 nodes are visible, fig. 2. We generated 500 networks with parameters $\{J_{ij}, h_j\}$ drawn randomly from the uniform distribution over $[-1, 1]$. The lower bound values \mathcal{F}_V for several approximating structures (including ‘standard mean field’) are compared with the true log likelihood, using the relative error $\mathcal{E} = \mathcal{F}_V / \ln P(S_V) - 1$, fig. 3. These show that considerable improvements can be obtained when belief networks are used. Note that a 5 component mixture model (≈ 80 variational parameters) yields $\mathcal{E} = 0.01139$ on this problem [4]. The results also suggest that one should exploit knowledge about the graphical structure of the model. For instance, the chain (fig. 3(b)) with no graphical overlap with the original graph shows hardly any improvement over the standard mean field approximation. On the other hand, the trees model (fig. 3(c)), which has about the same number of parameters, but a larger overlap with the original graph, does improve considerably over the mean field approximation (and even over the 5 component mixture model). By increasing the overlap, as in fig. 3(d), the improvement gained is even greater.

5 Discussion

The use of tractable belief networks fits well in mean field theory. Their ability to exploit the graphical structure of the model seems very powerful. Note that our approach (see also [6]) is very different from suggested methods [7] of using mean field theory to strip away intractable parts of the graph and compute

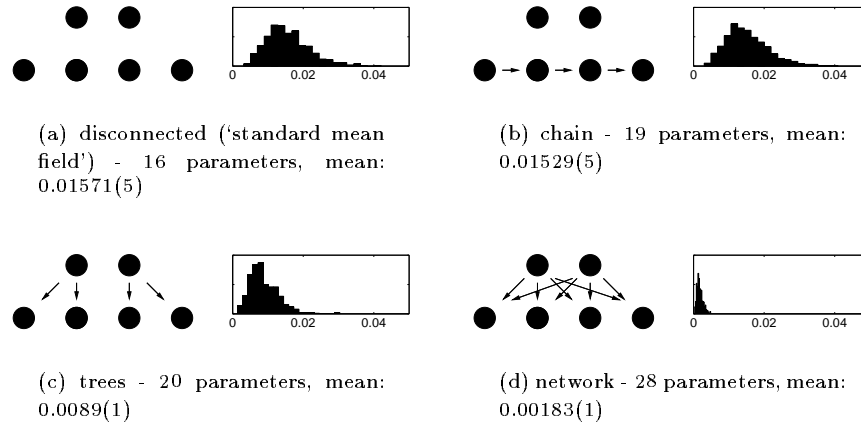


Figure 3: Graphical structures of the approximating distributions on H (cf. fig. 2). For each structure, histograms of the relative error between true log likelihood and the lower bound is plotted. Horizontal scale have been fixed to $[0,0.05]$ in all plots.

with the remaining *fixed* tractable substructure. In contrast, we use *variational* tractable structures together with mean field theory which leads, in general, to a more powerful approximation. We believe that such approaches will prove beneficial for learning large belief networks.

We have also developed a similar approach based on *undirected* graphical approximations, which have roughly the same accuracy, although the optimization procedure is implemented in a different way [6].

- [1] J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference*. Morgan Kaufmann Publishers, Inc., 1988.
- [2] R. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.
- [3] L.K. Saul, T. Jaakkola, and M.I. Jordan. Mean field theory for sigmoid belief networks. *Journal of Artificial Intelligence Research*, 4:61–76, 1996.
- [4] C.M. Bishop, N. Lawrence, T. Jaakkola, and M. I. Jordan. Approximating Posterior Distributions in Belief Networks using Mixtures. In *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1998. In press.
- [5] T.S. Jaakkola and M.I. Jordan. Approximating posteriors via mixture models. In M.I. Jordan, editor, *Proceedings NATO ASI Learning in Graphical Models*. Kluwer, 1997.
- [6] D. Barber and W. Wiering. Tractable undirected approximations for graphical models. In *ICANN'98: International Conference on Artificial Neural Networks, Skövde*, 1998.
- [7] L. K. Saul and M. I. Jordan. Exploiting Tractable Substructures in Intractable Networks. In D. S. Touretzky, M. C. Mozer, and M. E. Hasselmo, editors, *Advances in Neural Information Processing Systems*, volume 8. MIT Press, 1996.