Variational Belief Networks for Approximate Inference

Wim Wiegerinck David Barber
Stichting Neurale Netwerken, University of Nijmegen
Nijmegen, The Netherlands
e-mail: {wimw,davidb}@mbfys.kun.nl

Abstract

Exact inference in large, densely connected probabilistic networks is computationally intractable, and approximate schemes are therefore of great importance. One approach is to use mean field theory, in which the exact log-likelihood is bounded from below using a simpler approximating distribution. In the standard mean field theory, the approximating distribution is factorial. We propose instead to use a (tractable) belief network as an approximating distribution. The resulting compact framework is analogous to standard mean field theory and no additional bounds are required, in contrast to other recently proposed extensions. We derive mean field equations which provide an efficient iterative algorithm to optimize the parameters of the approximating belief network. Simulation results indicate a considerable improvement on previously reported methods.

1 Introduction

Belief networks provide a rich framework for probabilistic modeling and reasoning [1, 2]. Their graphical structure provides an intuitively appealing modularity and is well suited to the incorporation of prior knowledge. Belief networks often used in a domain with causal structures, such as speech recognition and medical diagnosis. Thanks to the directed graphical structure in these models, exact inference requires only local computations, in contrast to undirected models, such as Boltzmann machines. However, the complexity of exact inference scales exponentially with the clique size and, as a result, large densely connected networks are intractable for exact computations. Finding good approximate techniques for dealing with such models is therefore crucial if their power is to be realised. Whilst techniques do exist to handle large complex networks, most are based on uncontrolled approximations (such as Monte Carlo methods - see, *e.g.* [3]) and we seek instead a more rigorous approach.

One such method is to use mean field theory [3] which provides a rigorous lower bound on the likelihood through the use of a simpler variational approximating distribution. A lower bound is particularly useful since maximization of the lower bound can be used as a learning procedure when the true loglikelihood is intractable. Standard mean field theory resorts to using completely factorized models as approximating distributions. We will show that this restriction is unnecessary and that mean field techniques can be pushed much further 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, as proposed by [4]. In section 3 we show how this theory generalizes in a natural way when belief networks are used and discuss the relation to other extensions proposed in the literature [5, 6, 3]. In section 4 we apply the method on a toy benchmark problem [4, 5].

2 Mean Field Theory

Consider a probability model $P(S)$ on T binary valued units, $S = \{S_1, S_2, \ldots, S_T\}$ with $S_i \in 0/1$. (The following exposition is readily generalisable to many-valued discrete units). We wish to compute the likelihood $P(S_V)$ that the set of visible variables is in state $S_V \equiv S_{v_1}, \ldots S_{v_{N_V}}$. This involves the summation over exponentially many states of the H remaining (hidden) variables, $P(S_V) = \sum_{\{S_H\}} P(S_V, S_H)$. When this summation cannot be computed efficiently, approximate methods must be introduced. We present here the mean field theory approximation in some generality before specialising to particular models in later sections.

Consider the Kullback-Leibler divergence between the conditional hidden unit distribution, $P(S_H|S_V)$ and an approximating distribution $Q(S_H)$,

$$KL = \sum_{\{S_H\}} Q(S_H) \ln Q(S_H) - \sum_{\{S_H\}} Q(S_H) \ln P(S_H|S_V) \geq 0 \tag{1}$$

Using Bayes rule, $P(S_H|S_V) = P(S_H, S_V)/P(S_V)$, we obtain the following bound,

$$\ln P(S_V) \geq - \sum_{\{S_H\}} Q(S_H) \ln Q(S_H) + \sum_{\{S_H\}} Q(S_H) \ln P(S_H, S_V) \tag{2}$$

The first term in this bound, $H(Q) \equiv \sum_{\{S_H\}} Q(S_H) \ln Q(S_H)$ is the entropy of the approximating distribution Q. For general probability distributions on the hidden variables, $H(Q)$ is not tractable, and we therefore restrict $Q(S_H)$ to be in a class of simpler, tractable models $\mathcal{M}$. Unfortunately, the second term $\sum_{\{S_H\}} Q(S_H) \ln P(S_H, S_V)$ may not be tractable, even if the entropy term is. In this case we assume, however, that the term is also boundable, perhaps with recourse to other variational parameters, $\sum_{\{S_H\}} Q(S_H) \log P(S_V, S_H) \geq E_V(Q, \xi)$. We then write the bound in the general form,

$$\ln P(S_V) \geq E_V(Q, \xi) - H(Q) \equiv \mathcal{F}_V(Q, \xi) \tag{3}$$

in which ξ is a free parameter vector in a domain Ξ. The bound (3) is then made as tight as possible by maximizing $\mathcal{F}_V(Q, \xi)$ with respect to $Q \in \mathcal{M}$ and $\xi \in \Xi$.

2.1 Factorized models

The simplest mean field theory restricts the class $\mathcal{M}$ to factorized distributions

$$Q(S_H) = \prod_{i \in H} Q(S_i) \tag{4}$$

The entropy $H(Q)$ decomposes nicely into a tractable sum of entropies per site,

$$\sum_{\{S_H\}} Q(S_H) \ln Q(S_H) = \sum_{i \in H} \sum_{\{S_i\}} Q(S_i) \ln Q(S_i) \tag{5}$$

Similarly, it would be nice if the energy would decouple in a corresponding way. Clearly, if the energy factorizes over the sites, $\prod_{i \in H} e_{\mu i}(S_i, \xi) \equiv e_\mu(S_H, \xi)$ then this is possible, so that a general form of "nice" energy functions is given by

$$E_V(Q, \xi) = \sum_{\mu=1}^{M} f_\mu(\langle e_\mu(S_H, \xi) \rangle_Q) \tag{6}$$

where $f_\mu(x)$ is a scalar function. The computing time needed for the energy is then linear in M, which we assume to be at most polynomial in the number of hidden units, H. We will encounter just such an energy form in relation to the approximation of sigmoid belief networks, section 4.

In order to optimize $\mathcal{F}_V(Q, \xi)$, we introduce parameters $q_i \equiv Q(S_i = 1)$. The required normalization of Q fixes the remaining probability $Q(S_i = 0) = 1 - q_i$. Setting the gradient of $\mathcal{F}_V$, (3) with respect to the q_i's equal to zero yields the mean field equations

$$q_i = \sigma\left(\nabla_i E_V(Q, \xi)\right) \tag{7}$$

where the gradient ∇_i is with respect to q_i. The sigmoid function $\sigma(x) \equiv 1/(1 + e^{-x})$ is the inverse of the gradient of the entropy. Since $\langle e_\mu(S_H, \xi) \rangle_Q$ is linear in each of the parameters q_i, computation of the gradient $\nabla_i E_V(Q, \xi)$ is straightforward. A fast, two step iterative procedure to optimize $\mathcal{F}_V(Q, \xi)$ with respect to the q_i's and ξ, is proposed in [4]. In the first step, the q_i's are optimized by iteration of the mean field equations (7) while the ξ remain fixed. In the second step, ξ is optimized directly using $\mathcal{F}_V(Q, \xi)$ while the q_i's remain fixed. The two steps are iterated until convergence is reached.

3 Mean Field Theory using Belief Networks

We now consider the much richer class of belief networks as approximating models, while using the same general energy $E_V(Q, \xi)$ as in the previous section. A belief network is defined by a factorization over conditional probability distributions,

$$Q(S_H) = \prod_{i \in H} Q(S_i | S_{\pi_i}) \tag{8}$$

in which π_i denotes the sets of parent nodes of i, see fig(1).

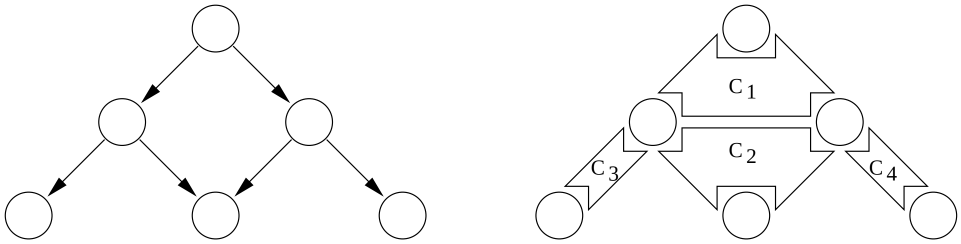
The efficiency of computation in a belief network depends on the underlying graphical structure of the model and is exponential in the maximal clique size (see, for example [2]).

We now assume that our model class $\mathcal{M}$ consists of belief networks with a fixed, tractable graphical structure. The entropy can then be computed efficiently since it decouples into a sum of averaged entropies per site i (with the convention that $Q(S_{\pi_i}) \equiv 1$ if $\pi_i = \phi$),

$$\sum_{\{S_H\}} Q(S_H) \ln Q(S_H) = \sum_{i \in H} \sum_{\{S_{\pi_i}\}} Q(S_{\pi_i}) \sum_{\{S_i\}} Q(S_i | S_{\pi_i}) \ln Q(S_i | S_{\pi_i}) \tag{9}$$

For tractable networks, the energy $E_V(Q, \xi)$ (6) is composed of terms

$$\langle e_\mu(S_H, \xi) \rangle_Q = \sum_{\{S_H\}} e_\mu(S_H, \xi) Q(S_H) = \prod_i \sum_{\{S_i\}} e_{\mu i}(S_i, \xi) Q(S_i | S_{\pi_i})$$



(a) Belief network

(b) Clique Graph

Figure 1: (a) A Belief Network. The parent set of each node i is that set of nodes which point into node i. (b) Clique graph, containing cliques of size 2 and 3.

which has exactly the same graphical structure as the belief network Q itself, (8). Therefore, computing the energy is of the same complexity as that for computing with the approximating belief network Q.

In order to optimize the bound, analogous to the factorized case, we parametrize Q via its conditional probabilities, $q_i(S_{\pi_i}) \equiv Q(S_i = 1|S_{\pi_i})$. The remaining probability $Q(S_i = 0|S_{\pi_i})$ follows from normalization. We therefore have a set $\{q_i(S_{\pi_i})|S_{\pi_i} = (0\ldots0), \ldots, (1\ldots1)\}$ of variational parameters for each node in the graph. Setting the gradient of $\mathcal{F}_V$ with respect to the $q_i(S_{\pi_i})$'s equal to zero, yields the mean field equations

$$q_i(S_{\pi_i}) = \sigma_i\left(\frac{\left(\nabla_{iS_{\pi_i}} E_V(Q,\xi)\right) + L_{iS_{\pi_i}}}{Q(S_{\pi_i})}\right) \tag{10}$$

with

$$L_{iS_{\pi_i}} = -\sum_j \sum_{S_{\pi_j}} [\nabla_{iS_{\pi_i}} Q(S_{\pi_j})] \sum_{S_j} Q(S_j|S_{\pi_j}) \ln Q(S_j|S_{\pi_j}) \tag{11}$$

The gradient $\nabla_{iS_{\pi_i}}$ is with respect to $q_i(S_{\pi_i})$. The explicit evaluation of the gradients can be performed efficiently, since all that need to be differentiated are at most scalar functions of quantities that depend again only linearly on the parameters $q_i(S_{\pi_i})$. To optimize the bound, we again use a two step iterative procedure as described in section 2.1. We see therefore, that the application of mean field theory using belief networks is analogous to that of using factorized models. However, the more powerful class of approximating distributions described by belief networks should enable a much tighter bound on the likelihood of the visible units.

3.1 Other methods

In the previous section, we described an approach in which a tractable belief network can be used, with the whole network of conditional probabilities being available as variational parameters. We briefly mention a related approach, which can be seen as a subclass of the more general algorithm previously described, and for which only a subset of the variables of a tractable distribution are varied. This modified mean field method was developed in [6].

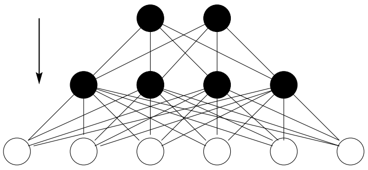


Figure 2: Graphical structure of the 2-4-6 nodes sigmoid belief network. Open circles: visible units S_V. Filled circles: hidden units S_H. Maximum clique size: 6.

Here the class of approximating distributions is given by

$$Q(S_H) = \frac{1}{Z_Q} \prod_{i \in H} \exp(\lambda_i(S_i)) \tilde{P}(S_H) \tag{12}$$

in which $\tilde{P}(S_H)$ is a tractable belief network which is preset by hand and remains fixed during optimization. Z_Q is a normalization factor. The idea is that $\tilde{P}(S_H)$ mimics the original intractable probability distribution P as much as possible. For example, deleting certain connections in the original model P may render the graph tractable. In order to compensate for this simplification, the functions $\lambda_i(S_i)$ can be varied. Since the freedom $\lambda_i(S_i)$ does not affect the graphical structure, the distributions Q are tractable belief networks as well. However, in contrast to our approach, only a small subset of tractable belief networks is explored during the optimization since only the λ_i's are varied while $\tilde{P}(S_H)$ remains fixed. In [3] the method is applied in factorial hidden Markov models, in which $\tilde{P}(S_H)$ is a product of Markov Chains.

Recently [5, 3] have proposed to use mixtures of factorized models. 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. This leads to a rather large number of variational parameters, arguably because the approximating distribution is not faithful to the structure of the original model.

4 Application to Sigmoid Belief Networks

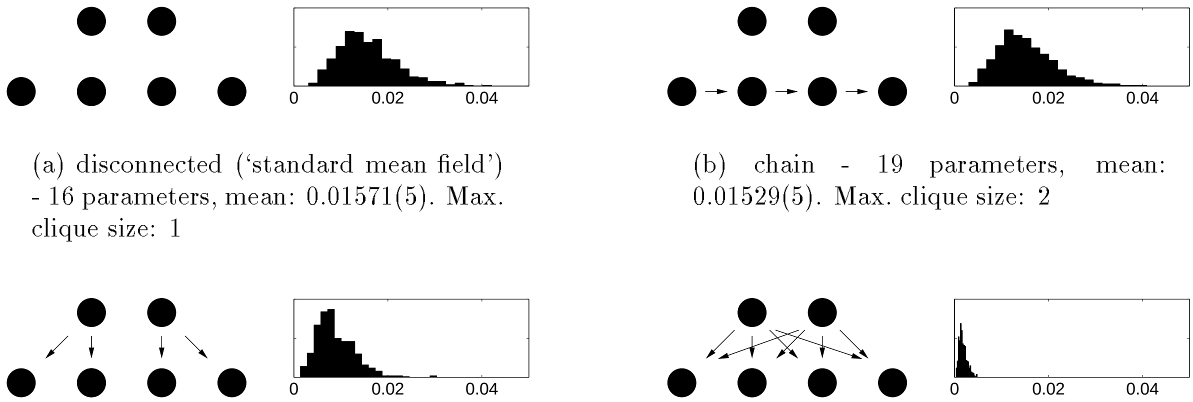
We applied mean field theory using belief networks on a toy benchmark problem to compare its performance with previously reported methods. Following [4, 5] we consider a problem in a three layer (2-4-6 nodes) sigmoid belief network in which the last 6 nodes are visible, fig. 2. In a sigmoid belief network [7] with binary units ($S_i = 0/1$) the conditional probability that the variable $S_i = 1$ given its parents S_{π_i} is

$$P(S_i = 1 | S_{\pi_i}) = \sigma(z_i) \tag{13}$$

with $z_i \equiv \sum_j J_{ij} S_j + h_i$. The weights J_{ij} (with $J_{ij} = 0$ for $j \notin \pi_i$) and biases h_j are the parameters of the network. In attempting to compute the lower bound, (2), unfortunately, the average of $P(S_H, S_V)$ is not tractable, since $\langle \ln [1 + e^z] \rangle$ does not decouple into a polynomial number of single site averages. We make use instead of the further bound proposed by [4]

$$\langle \ln [1 + e^z] \rangle \leq \xi \langle z \rangle + \ln \left\langle e^{-\xi z} + e^{(1-\xi) z} \right\rangle \tag{14}$$

We can then define the energy function



(a) disconnected ('standard mean field') - 16 parameters, mean: 0.01571(5). Max. clique size: 1

(b) chain - 19 parameters, mean: 0.01529(5). Max. clique size: 2

(c) trees - 20 parameters, mean: 0.0089(1). Max. clique size: 2

(d) network - 28 parameters, mean: 0.00183(1). Max. clique size: 3

Figure 3: Graphical structures of the approximating distributions on S_H (cf. fig. 2). For each structure, histograms of the relative error between the true log likelihood and the lower bound is plotted. The horizontal scale has been fixed to [0,0.05] in all plots. The maximum clique size refers to the complexity of computation for each approximation, which is exponential in this quantity.

$$E_V(Q,\xi) = \sum_{ij} J_{ij} \langle S_i S_j \rangle_Q + \sum_i (h_i - \sum_j \xi_j J_{ji}) \langle S_i \rangle_Q \\ - \sum_i h_i \xi_i - \sum_i \ln \left\langle e^{-\xi_i z_i} + e^{(1-\xi_i) z_i} \right\rangle_Q \qquad (15)$$

where the variational parameters $\xi \in [0,1]$. Note that this is of the class of energy functions defined by (6).

In order to test our method numerically, we generated 500 networks with parameters $\{J_{ij}, h_j\}$ drawn randomly from the uniform distribution over $[-1,1]$. The lower bounds $\mathcal{F}_V$ for several approximating structures (including 'standard mean field', section 2.1) 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 ($\approx$ 80 variational parameters) yields $\mathcal{E} = 0.01139$ on this problem [5]. These results suggest therefore that exploiting knowledge of the graphical structure of the model is useful in practice. 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 tree 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. In contrast to a previous mean field theory using constrained substructures [6], we search in the whole, unconstrained space defined by the graphical structure of the approximating model. Since factorized models and constrained belief networks are both subsets of unconstrained belief networks, the latter class of models is guaranteed to give the tighter bound on the likelihood.

In conclusion, we have described a general approach that functions with simple, or complex approximating distributions, depending on the computational resources of the user. We believe that this approach will prove beneficial for learning and inference in large real-world belief networks [8].

Acknowledgements

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

References

[1] J. Pearl. *Probabilistic Reasoning in Intelligent systems: Networks of Plausible Inference.* Morgan Kaufmann Publishers, Inc., 1988.

[2] E. Castillo, J. M. Gutierrez, and A. S. Hadi. *Expert Systems and Probabilistic Network Models.* Springer, 1997.

[3] M. I. Jordan, editor. *Learning in Graphical Models*, volume 89 of *NATO ASI, Series D: Behavioural and Social Sciences.* Kluwer, 1998.

[4] 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.

[5] C.M. Bishop, N. Lawrence, T. Jaakkola, and M. I. Jordan. Approximating posterior distributions in belief networks using mixtures. In M.I. Jordan, M.J. Kearns, and S.A. Solla, editors, *Advances in Neural Information Processing Systems*, volume 10, pages 416–422. MIT Press, 1998.

[6] 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.

[7] R. Neal. Connectionist learning of belief networks. *Artificial Intelligence*, 56:71–113, 1992.

[8] W. Wiegerinck *et. al.* Variational approximations in a broad and detailed probabilistic model for medical diagnosis. In *Tenth Netherlands/Belgium Conference on Artificial Intelligence (NAIC'98).* CWI, 1998. in press.