Second order approximations for probability models

Hilbert J. Kappen
Department of Biophysics
Nijmegen University
Nijmegen, the Netherlands
bert@mbfys.kun.nl

Wim Wiegerinck
Department of Biophysics
Nijmegen University
Nijmegen, the Netherlands
wimw@mbfys.kun.nl

Abstract

In this paper, we derive a second order mean field theory for directed graphical probability models. By using an information theoretic argument it is shown how this can be done in the absense of a partition function. This method is the direct generalisation of the well-known TAP approximation for Boltzmann Machines. In a numerical example, it is shown that the method greatly improves the first order mean field approximation. The computational complexity of the first (second) order method is linear (quadratic) in the network size and is exponential in the potential size. For a restricted class of graphical models, so-called single overlap graphs, the second order method has comparable complexity to the first order method.

1 Introduction

Recently, a number of authors have proposed methods for approximate inference in large graphical models. The simplest approach gives a lower bound on the probability of a subset of variables using Jenssen's inequality (Saul et al., 1996). The method involves the minimization of the KL divergence between the target probability distribution p and some 'simple' variational distribution q. The method can be applied to a large class of probability models, such as sigmoid belief networks, DAGs and Boltzmann Machines (BM).

For Boltzmann-Gibbs distributions, it is possible to derive the lower bound as the first term in a Taylor series expansion of the free energy around a factorized model. The free energy is given by $-\log Z$, where Z is the normalization constant of the Boltzmann-Gibbs distribution: $p(x) = \frac{\exp(-E(x))}{Z}$. This Taylor series can be continued and the second order term is known as the TAP correction (Plefka, 1982; Kappen and Rodríguez, 1998). The second order term significantly improves the quality of the approximation, but is no longer a bound.

For probability distributions that are not Boltzmann-Gibbs distributions, it is not obvious how to obtain the second order approximation. However, there is an alternative way to compute the higher order corrections, based on an information theoretic argument. Recently, this argument was applied to stochastic neural networks with asymmetric connectivity (Kappen and Spanjers, 1999). Here, we apply this idea to directed graphical models.

2 The method

Let $x = (x_1, \ldots, x_n)$ be an n-dimensional vector, with x_i taking on discrete values. Let $p(x)$ be a directed graphical model on x. We will assume that $p(x)$ can be written as a product of potentials in the following way:

$$p(x) = \prod_{k=1}^{n} p_k(x_k|\pi_k) = \exp \sum_{k=1}^{n} \phi_k(x^k). \tag{1}$$

Here, $p_k(x_k|\pi_k)$ denotes the conditional probability table of variable x_k given the values of its parents π_k. $x^k = (x_k, \pi_k)$ denotes the subset of variables that appear in potential k and $\phi_k(x^k) = \log p_k(x_k|\pi_k)$. Potentials can be overlapping, $x^k \cap x^l \neq \emptyset$, and $x = \cup_k x^k$.

We wish to compute the marginal probability that x_i has some specific value h_i in the presence of some evidence. We therefore denote $x = (e, h)$ where e denote the subset of variables that constitute the evidence, and h denotes the remainder of the variables. The marginal is given as

$$p(h_i|e) = \frac{p(h_i, e)}{p(e)}. \tag{2}$$

Both numerator and denominator contain sums over hidden states. These sums scale exponentially with the size of the problem, and therefore the computation of marginals is intractable.

We propose to approximate this problem by using a mean field approach. Consider a factorized distribution on the hidden variables h:

$$q(h) = \prod_i q_i(h_i) \tag{3}$$

We wish to find the factorized distribution q that best approximates $p(h|e)$. Consider as a distance measure

$$KL = \sum_h p(h|e) \log \left(\frac{p(h|e)}{q(h)} \right). \tag{4}$$

It is easy to see that the q that minimizes KL satisfies:

$$q(h_i) = p(h_i|e) \tag{5}$$

We now think of the manifold of all probability distributions of the form Eq. 1, spanned by the coordinates $\phi_k(x^k), k = 1, \ldots, m$. For each k, $\phi_k(x^k)$ is a table of numbers, indexed by x^k. This manifold contains a submanifold of factorized probability distributions in which the potentials factorize: $\phi_k(x^k) = \sum_{i, i \in k} \phi_{ki}(x_i)$. When in addition, $\sum_{k, i \in k} \phi_{ki}(x_i) = \log q_i(x_i), i \in h$, $p(h|e)$ reduces to $q(h)$.

Assume now that $p(h|e)$ is somehow close to the factorized submanifold. The difference $\Delta p(h_i|e) = p(h_i|e) - q_i(h_i)$ is then small, and we can expand this small difference in terms of changes in the parameters $\Delta\phi_k(x^k) = \phi_k(x^k) - \log q(x^k), k = 1, \ldots, m$:

$$\begin{aligned} \Delta \log p(h_i|e) &= \sum_{k=1}^{n} \sum_{\bar{x}^k} \left(\frac{\partial \log p(h_i|e)}{\partial \phi_k(\bar{x}^k)} \right)_q \Delta\phi_k(\bar{x}^k) \\ &+ \frac{1}{2} \sum_{kl} \sum_{\bar{x}^k, \bar{y}^l} \left(\frac{\partial^2 \log p(h_i|e)}{\partial \phi_k(\bar{x}^k) \partial \phi_l(\bar{y}^l)} \right)_q \Delta\phi_k(\bar{x}^k) \Delta\phi_l(\bar{y}^l) \\ &+ \text{higher order terms} \end{aligned} \tag{6}$$

The differentials are evaluated in the factorized distribution q. The strategy is now the following. We are interested in the marginals $p(h_i)$. Trivially, this can be reformulated as being interested in the factorized distribution such that $q(h_i) = p(h_i)$. Now, the approach is to set the expansion

$$\Delta \log p(h_i|e) = 0 \tag{7}$$

and solve for $q(h_i)$. This factorized distribution gives the desired marginals up to the order of the expansion of $\Delta \log p(h_i|e)$.

It is straightforward to compute the derivatives:

$$\begin{aligned}
\frac{\partial \log p(h_i|e)}{\partial \phi_k(\bar{x}^k)} &= p(\bar{x}^k|h_i,e) - p(\bar{x}^k|e) \\
\frac{\partial^2 \log p(h_i|e)}{\partial \phi_k(\bar{x}^k)\partial \phi_l(\bar{y}^l)} &= p(\bar{x}^k,\bar{y}^l|h_i,e) - p(\bar{x}^k,\bar{y}^l|e) \\
&\quad -p(\bar{x}^k|h_i,e)p(\bar{y}^l|h_i,e) + p(\bar{x}^k|e)p(\bar{y}^l|e)
\end{aligned} \tag{8}$$

Using notation where $\langle \ldots \rangle_i$ and $\langle \ldots \rangle$ are the expectation values with respect to the factorized distributions $q(x|h_i,e)$ and $q(x|e)$, and $\langle\langle \ldots \rangle\rangle_i \equiv \langle \ldots \rangle_i - \langle \ldots \rangle$ is the difference between expectation values, we obtain

$$\Delta \log p(h_i|e) = \sum_k \langle\langle \Delta\phi_k \rangle\rangle_i \tag{9}$$

$$+\frac{1}{2}\sum_{k,l} \left(\langle\langle \Delta\phi_k \Delta\phi_l \rangle\rangle_i - \langle \Delta\phi_k \rangle_i \langle \Delta\phi_l \rangle_i + \langle \Delta\phi_k \rangle \langle \Delta\phi_l \rangle \right), \tag{10}$$

$$+\text{higher order terms} \tag{11}$$

To first order, setting Eq. 11 equal to zero, we obtain

$$0 = \sum_k \langle\langle \Delta\phi_k \rangle\rangle_i = \langle \log p(x) \rangle_i - \log q(h_i) + \text{const.}, \tag{12}$$

where we have absorbed all terms independent of i into a constant. Thus, we find the solution

$$q(h_i) = \frac{1}{Z_i} \exp\left(\langle \log p(x) \rangle_i \right) \tag{13}$$

in which the constants Z_i follow from normalisation. The first order term is equivalent to the standard mean field equations, obtained from the Jensen inequality.

The correction with second order terms is obtained in the same way, again dropping terms independent of i:

$$q(h_i) = \frac{1}{Z_i} \exp\left(\langle \log p(x) \rangle_i + \frac{1}{2}\sum_{k,l} \left(\langle \Delta\phi_k \Delta\phi_l \rangle_i - \langle \Delta\phi_k \rangle_i \langle \Delta\phi_l \rangle_i \right) \right) \tag{14}$$

were, again, the constants Z_i follow from normalisation. These equations, which form the main result of this paper, are generalization of the mean field equations with TAP corrections for directed graphical models.

In fact, one could also drop the last term in Eq. 14 because of the identity:

$$\sum_{k,l} \langle \Delta\phi_k \rangle_i \langle \Delta\phi_l \rangle_i = \left(\sum_k \langle\langle \Delta\phi_k \rangle\rangle_i + \langle \Delta\phi_k \rangle \right)^2 = \text{const.}$$

However, numerical experiments show that this has a strong negative effect on the convergence of the fixed point iteration.

The complexity of the mean field equations (13) is exponential in the number of variables in the potentials ϕ_k of P: if the maximal clique size is c, then for each i we need of the order of $n_i \exp(c)$ computations, where n_i is the number of cliques that contain node i.

The second term scales worse, since one must compute averages over the union of two overlapping cliques and because of the double sum. However, things are not so bad. First of all, notice that the sum over k and l can be restricted to overlapping cliques ($k \cap l \neq \emptyset$) and that i must be in either k or l or both ($i \in k \cup l$). Denote by n^k the number of cliques that have at least one variable in common with clique k and denote by $n_{\text{overlap}} = \max_k n_k$ Then, the sum over k and l contains not more than $n_i n_{\text{overlap}}$ terms.

Each term is an average over the union of two cliques, which can be worse case of size $2c-1$ (when only one variable is shared). However, since $\langle \Delta\phi_k \Delta\phi_l \rangle_i = \langle \langle \Delta\phi_k \rangle_{k \cap l} \Delta\phi_l \rangle_i$ we can precompute $\langle \Delta\phi_k \rangle_{k \cap l}$ for all pairs of overlapping cliques k, l. Thus, the worse case complexity of the second order term is less than $n_i n_{\text{overlap}} \exp(c)$. Thus, we see that the second order method has the same exponential complexity as the first order method, but with a different polynomial prefactor. Therefore, the first or second order method can be applied to directed graphical models as long as the number of parents is reasonably small.

For implementational purposes, we can therefore write Eq. 14 also as

$$\begin{aligned} q(h_i) &= \frac{1}{Z_i} \exp\left(\sum_{k, i \in k} \left\langle \log p(x^k) \right\rangle_i + \frac{1}{2} \sum_{k, i \in k} \left(\left\langle (\Delta\phi_k)^2 \right\rangle_i - \langle \Delta\phi_k \rangle^2 \right) \right. \\ &+ \sum_{k, i \in k} \sum_{l > k, i \in l} \left(\langle \Delta\phi_k \Delta\phi_l \rangle_i - \langle \Delta\phi_k \rangle_i \langle \Delta\phi_l \rangle_i \right) \\ &+ \left. \sum_{k, i \in k} \sum_{l \neq k, i \notin l} \left(\langle \Delta\phi_k \Delta\phi_l \rangle_i - \langle \Delta\phi_k \rangle_i \langle \Delta\phi_l \rangle_i \right) \right) \end{aligned} \tag{15}$$

3 Complexity and single-overlap graphs

The fact that the second order term has a worse complexity than the first order term is in contrast to Boltzmann machines, in which the TAP approximation has the same complexity as the standard mean field approximation. This phenomenon also occurs for a special class of DAGs, which we call single-overlap graphs. These are graphs in which the potentials ϕ_k share at most one node. Figure 1 shows an example of a single-overlap graph. For a single overlap graph, one can rewrite Eq. 14 as

$$\begin{aligned} q(h_i) &= \frac{1}{Z_i} \exp\left(\langle \log p(x) \rangle_i + \frac{1}{2} \sum_{l, i \in l} \left(\left\langle (\Delta\phi_l)^2 \right\rangle_i - \langle \Delta\phi_l \rangle_i^2 \right) \right. \\ &- \left. \sum_{l, i \in l} \sum_{j' \neq i} \langle \langle \langle \Delta\phi_l \rangle \rangle_{j'} \Delta\phi_l \rangle_i + \sum_{l, i \in l} \langle \langle \Delta\phi_l \rangle \rangle_i \langle \Delta\phi_l \rangle_i \right) \end{aligned} \tag{16}$$

which has a complexity that is of order $n_i(c-1)\exp(c)$. For instance, for Boltzmann Machines $n_i = n_{\text{overlap}} = n - 1$ and $c = 2$.

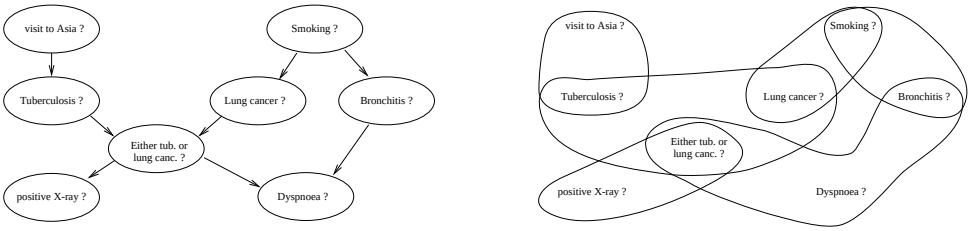

Figure 1: An example of a single-overlap graph. Left: The chest clinic model (ASIA)(Lauritzen and Spiegelhalter, 1988). Right: nodes within one potential a re grouped together, showing that potentials share at most one node.

Node	Exact	MF	TAP
visit to Asia?	0.010	0.010	0.010
Smoking?	0.500	0.420	0.524
Tuberclosis?	0.010	0.000	0.000
Lung cancer?	0.055	0.000	0.000
Bronchitis?	0.450	0.264	0.410
Either t or l?	0.065	0.000	0.000
positive X ray?	0.110	0.050	0.050
Dyspnoea?	0.436	0.223	0.497

Table 1: Marginal probabilities of states being $true$ obtained in the chest clinic model (ASIA). First column: exact marginals. Second column: marginals computed using first order approximation (mean field). Third column: marginals computed using an approximation up to second order (TAP).

4 Numerical results

We illustrate the theory by a toy problem, which is inference in Lauritzen's chest clinic model (ASIA), defined on 8 binary variables $\{A, T, S, L, B, E, X, D\}$ (see figure 1, and (Lauritzen and Spiegelhalter, 1988) for more details about the model). We computed exact marginals and approximate marginals using the approximating methods up to first and second order respectively. The approximate marginals are determined by sequential iteration of (13) and (14), starting at $q(x_i) = 0.5$ for all variables i. Results are shown in table 1.

5 Discussion

In this paper, we computed a second order mean field approximation for directed graphical models. We show that the second order approximation gives a significant improvement over the first order result. This suggests that this method can be of significant practical value.

The derivation in this paper does not use directly that the graph must be directed. Therefore, we expect that this result is equally true for Markov graphs.

Whereas the complexity of the first order approximation is of $\mathcal{O}(\exp K)$, with K the maximum number of variables in a potential of p, we find that in general the second order approximation requires $\mathcal{O}(\exp 2K)$ computations. We define a new class of graphs, single overlap graphs, for which this increase in complexity does not occur. Examples of such graphs are Boltzmann distributions with second order interactions as well as the ASIA

network, used as a numerical example in this paper. In any case, for large K, additional approximations are required, as was proposed by (Saul et al., 1996) for the first order mean field equations. It is evident, that such additional approximations are then also required for the second order mean field equations.

It has been reported (Barber and Wiegerinck, 1999; Wiegerinck and Kappen, 1999) that similar numerical improvements can be obtained by using a very different approach, which is to use an approximating distribution q that is not factorized, but still tractable. A promising way to proceed is therefore to combine both approaches and to do a second order expansion aroud a manifold of distributions with non-factorized yet tractable distributions. In this approach the sufficient statistics of the tractable structure is expanded, rather than the marginal probabilities.

Acknowledgments

This research was supported in part by the Dutch Technology Foundation (STW).

References

Barber, D. and Wiegerinck, W. (1999). Tractable variational structures for approximating graphical models. In Kearns, M., Solla, S., and Cohn, D., editors, *Advances in Neural Information Processing Systems*, volume 11 of *Advances in Neural Information Processing Systems*, pages 183–189. MIT Press. R-98-031 SNN-98-015.

Kappen, H. and Rodríguez, F. (1998). Effi cient learning in Boltzmann Machines using linear response theory. *Neural Computation*, 10:1137–1156. SNN–97–001, F–97–005.

Kappen, H. and Spanjers, J. (1999). Mean fi eld theory for asymmetric neural networks. *Physical Review E*, 61:5658–5663.

Lauritzen, S. and Spiegelhalter, D. (1988). Local computations with probabilties on graphical structures and their application to expert systems. *J. Royal Statistical society B*, 50:154–227.

Plefka, T. (1982). Convergence condition of the TAP equation for the infi nite-range Ising spin glass model. *Journal of Physics A*, 15:1971–1978.

Saul, L., Jaakkola, T., and Jordan, M. (1996). Mean fi eld theory for sigmoid belief networks. *Journal of artificial intelligence research*, 4:61–76.

Wiegerinck, W. and Kappen, H. (1999). Approximations of bayesian networks through kl minimisation. *New Generation Computing*, 18:167–175.