Tractable Undirected Approximations for Graphical Models

David Barber*and Wim Wiegerinck†
RWCP‡ Theoretical Foundation SNN § University of Nijmegen
6525 EZ Nijmegen, The Netherlands.

Abstract

Graphical models provide a broad framework for probabilistic inference, with application to such diverse areas as speech recognition (Hidden Markov Models), medical diagnosis (Belief networks) and artificial intelligence (Boltzmann Machines). However, the computing time is typically exponential in the number of nodes in the graph. We present a general framework for a class of approximating models, based on the Kullback-Leibler divergence between an approximating graph and the original graph. We concentrate here on *undirected* approximations of both intractable directed and undirected graphical models. Simulation results on a small benchmark problem suggest that this method compares favourably against others previously reported in the literature.

1 Introduction

Graphical models have recently drawn a great deal of interest, being recognised as a powerful framework for probabilistic inference[1]. Unfortunately, in general, large graphs are computationally intractable and approximations need to be employed. Recently, variational approximations have been popular[2, 3, 4, 5], and have the advantage of providing rigorous bounds on quantities of interest, such as the data likelihood, in contrast to other approximate procedures such as Monte Carlo methods[1]. In the neural networks community, one of the original models, the Boltzmann machine (BM), belongs to the class of *undirected* graphical models although the lack of a suitable algorithm has hindered their application to larger problems. In principle, the framework we present here is applicable to all undirected approximations, although the simplest, non-trivial of these are BMs, on which we focus here, section (2). We re-derive one of the main variational approximations in section (2.1) before generalising this to arbitrary tractable approximations in section (3). In section (3.1) we show how this can be used to accurately approximate BMs. We apply this framework to approximate directed graphs in section (3.2) by introducing extra variational parameters, and include results on a toy benchmark problem.

*`http://www.mbfys.kun.nl/~davidb`
†`http://www.mbfys.kun.nl/~wimw`
‡Real World Computing Partnership
§Foundation for Neural Networks

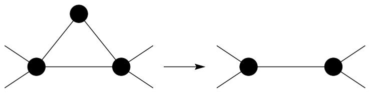


Figure 1: A decimation rule for BMs. We remove the upper node on the left so that the partition function of the reduced graph is the same. This requires a simple change in the parameters J, h coupling the two nodes on the right.

2 Boltzmann Machines

Boltzmann machines describe probability distributions over binary variables $s_i \in \{0,1\}, i = 1..N$, of the form

$$P(s) = \frac{1}{Z} \exp\left(\sum_i h_i s_i + \sum_{ij} J_{ij} s_i s_j\right) \tag{1}$$

where the normalization constant, commonly called the partition function is

$$Z = \sum_s \exp\left(\sum_i h_i s_i + \sum_{ij} J_{ij} s_i s_j\right) \tag{2}$$

All quantities of interest can be derived from the partition function, or minor variants of it. For general connection structures, J, computing Z is intractable as it involves a sum over 2^N states; however, not all Boltzmann machines are intractable. A standard class of tractable structures is described by a set of so-called decimation rules in which nodes from the graph can be removed one by one, fig(1). Provided that appropriate local changes are made to the BM parameters, the partition function of the reduced graph remains unaltered (see eg [2]). By repeated application of such rules, the partition function is calculable in linear time.

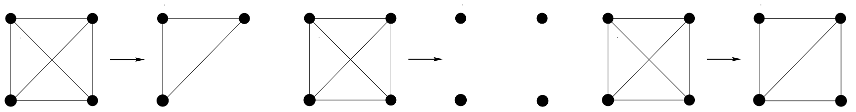
2.1 Node elimination variational approximation

For comparison, we re-derive briefly the lower bound approximation of Jaakkola et al. (see for example, [3]). The central idea is to strip away enough nodes of a general graph to reveal a tractable (decimatable) subgraph, see fig(2a), whilst retaining a bound on the partition function. Without loss of generality, we will remove node 1. Consider the following representation of P

$$P(s_1 \dots s_n) = P(s_1 | s_2 \dots s_n) P(s_2 \dots s_n) \tag{3}$$

We find an approximating distribution $Q(s_1)$ to $P(s_1|s_2 \dots s_n)$ by minimizing the Kullback-Leibler divergence, $KL = \sum_{s_1=0,1} Q(s_1) \ln(Q(s_1)/P(s_1|s_2 \dots s_n))$,

$$KL = H(\mu_1) + \ln\left(1 + e^{h_1 + \sum_j J_{1j} s_j}\right) - \mu_1 \left(\sum_j J_{1j} s_j + h_1\right) \geq 0 \tag{4}$$



(a) Node removal (b) Mean Field (c) Tractable graph

Figure 2: A fully connected 4 node BM is not decimatable. Variational approximations correspond to decimatable subgraphs of varying complexity.

where the binary entropy is given by $H(\mu) = -\mu \ln \mu - (1-\mu)\ln(1-\mu)$ and $\mu_1 = Q(s_1 = 1)$. Since $KL \geq 0$, we obtain a bound on the term,

$$\ln\left(1 + e^{h_1 + \sum_j J_{1j} s_j}\right) > H(\mu_1) + \mu_1 \left(\sum_j J_{1j} s_j + h_1\right) \tag{5}$$

Extracting the dependence on node 1 in (2) and summing over the two states

$$Z = \sum_{\{s_k, k \neq 1\}} \left(1 + e^{h_1 + \sum_j J_{1j} s_j}\right) \exp\left(\sum_{i \neq 1} h_i s_i + \sum_{ij \neq 1} J_{ij} s_i s_j\right) \tag{6}$$

Using the bound on $1 + e^{h_1 + \sum_j J_{1j}}$ in (5), we can then write

$$\ln Z > H(\mu_1) + h_1 \mu_1 + \ln \sum_{\{s_k, k \neq 1\}} \exp\left(\sum_{i \neq 1} \tilde{h}_i s_i + \sum_{ij \neq 1} J_{ij} s_i s_j\right) \tag{7}$$

where the new biases are given by $\tilde{h}_i = h_i + \mu_1 J_{1i}$. We then repeatedly remove nodes in a similar manner until a tractable (sub)structure is found. The bound (an extended form of (7)) is subsequently optimized with respect to the variational parameters $\{\mu_i\}$. If we continue to eliminate all the nodes in the graph, this reduces to the "naive" mean field theory, which is equivalent to a factored approximation (see fig(2b)), $Q(s) = \prod_i {\mu_i}^{s_i} (1-\mu_i)^{1-s_i}$ [6].

3 General Kullback-Leibler Approximation

Both the node elimination and naive mean field theory use the KL divergence bound for tractable variational subgraphs in which any connections J are the same as in the original graph, but with potentially adaptable biases. Here, instead of searching for a partially "factored" approximation to P, as in section (2.1), we consider a general distribution, $Q(s)$, in which *all* biases and weights are adaptable. To find the best approximation, we need to calculate the KL divergence. For convenience, we write the intractable distribution in the form,

$P(s|h,J) = e^{\phi_p}/Z_p$, and the approximating distribution $Q(s|\hat{h},\hat{J}) = e^{\phi_q}/Z_q$. The KL divergence, measuring the discrepancy between Q and P, is

$$KL = \sum_S \left(Q \ln Q - Q \ln P\right) = \langle \phi_q \rangle - \ln Z_q - \langle \phi_p \rangle + \text{const.} \tag{8}$$

where $\langle \cdots \rangle$ denotes averages over $Q(s)$.

3.1 Approximating Undirected Graphs

Since, for BMs, each term ϕ_q and ϕ_p is a quadratic function of the variables s, we need to be able to calculate first and second order correlations, $\langle s_i \rangle$, $\langle s_i s_j \rangle$. Under the representation $s \in \{0,1\}$,

$$\begin{aligned} \langle s_i s_j \rangle = Q\left(s_i = 1, s_j = 1\right) &= \sum_{s_k, k \neq i,j} Q\left(s_1, \ldots s_i = 1, \ldots s_j = 1, \ldots s_N\right) \\ &= e^{\left(\hat{h}_i + \hat{h}_j + 2\hat{J}_{ij}\right)} Z_{-[i,j]}\left(h', \hat{J}\right) / Z\left(\hat{h}, \hat{J}\right), \end{aligned} \tag{9}$$

where $h'_k = \hat{h}_k + 2\left(\hat{J}_{ik} + \hat{J}_{jk}\right)$. The partition function $Z_{-[i,j]}$ relates to a graph in which nodes i, j have been removed. The correlations $\langle s_i \rangle$ can be computed similarly[1]. With these tools, we can approximate any BM by another, tractable BM by minimizing (8) with respect to $\hat{J}$, $\hat{h}$. For example, one can approximate a fully connected graph by the largest decimatable (sub)structure, fig(2c). In contrast to the node elimination scheme, all the connection strengths and biases of the approximating graph are adaptable leading, in general, to a more powerful approximation.

3.2 Approximating Directed Graphs

Directed probability distributions have the form $P(s) = \prod_i P(s_i|\text{pa}\left(s_i\right))$, where $\text{pa}\left(s_i\right)$ is the parent set of node i[1]. For concreteness, we consider here sigmoid belief networks in which each local probability has the form

$$P(s_i = 1|\text{pa}\left(s_i\right)) = \sigma\left(\sum_j J_{ij} s_j + h_i\right) \tag{10}$$

where $\sigma\left(z\right) = 1/\left(1 + e^{-z}\right)$. We use the convention that the connection J_{ij} is directed *from* node j *to* node i with no self-interactions, $J_{ii} = 0$. The network is composed of visible (V) and hidden units (H), and the quantity of interest is the likelihood of the visible units,

$$P(V) = \sum_H P(H,V) \tag{11}$$

[1] There is a faster way to compute the term $\sum_S Q \ln Q$. Consider the partition function $\ln Z_q\left(\lambda\right) = \sum_S e^{\lambda(\mathbf{h}\cdot\mathbf{s} + \mathbf{s}\cdot J\mathbf{s})}$. Then $\sum_S Q \ln Q = \frac{\partial}{\partial \lambda} \ln Z_q\left(\lambda\right)\Big|_{\lambda=1} - \ln Z_q$. This derivative can be approximated numerically extremely accurately using only two partition functions.

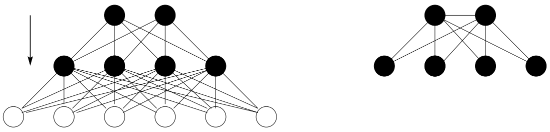


Figure 3: Directed graph toy problem (left). The hidden units (black) are approximated by a BM (right), one of many possible tractable structures.

Since this involves summing over all the states of the hidden units, we attempt to apply the framework of section (3) for approximating partition functions. We write the directed graph probability distribution as[2]

$$P(H, V) = \prod_i \exp\left[z_i s_i - \ln \psi_i\right] \tag{12}$$

where $z_i = \sum_j J_{ij} s_j + h_i$ and the local normalization functions are $\psi_i = 1 + e^{z_i}$. The first quadratic term in each exponential factor of (12) is equivalent to a BM factor, but the local normalisations ψ_i, prevent the distribution being exactly representable by a BM. In calculating the KL divergence between an approximating BM and $P(H, V)$, we need to compute averages over $\ln \psi_i$, which cannot be written in a compact form. We make use instead of the approximation[3] [4]

$$\langle \ln\left[1 + e^z\right]\rangle \leq \xi \langle z \rangle + \ln \left\langle e^{-\xi z} + e^{(1-\xi) z}\right\rangle \tag{13}$$

where ξ is a variational parameter that lies in the interval $[0, 1]$. Application of (13) in (8) results in a tractable expression for the KL divergence[4]. We then optimize the parameters of the BM, $\hat{J}, \hat{h}$, together with ξ, by minimizing the KL divergence numerically[5].

3.3 A toy benchmark problem

A toy problem presented in [4] provides a suitable benchmark for our method. Layered networks with 6 hidden and 6 visible units (see fig(3)) are randomly generated, with their parameters J_{ij} and h drawn from a uniform distribution over $[-1, 1]$. In all cases, the visible units have their states clamped to zero and we wish to approximate the likelihood $P(V)$. The true likelihood of the visible units can be computed exactly for this small problem. Naive mean field applied using the bound (13) has a mean relative error of 0.0156[4]. Mixtures of mean field[6] was used in [5], which gave a mean relative error for a five component

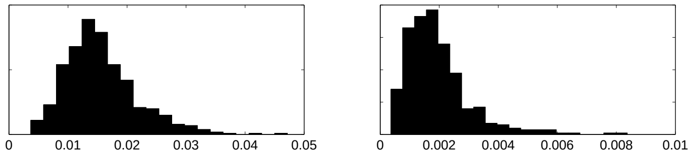
[2] In the notation of section (3), $\phi_p = \sum_i \left[z_i s_i - \ln \psi_i\right]$, and $Z_p = 1$.

[3] A more accurate bound is possible using a quadratic (tractable) function of z.

[4] It is possible to first approximate sigmoid belief networks by BMs and subsequently this by a tractable BM. We believe our approach to be more elegant and also more accurate.

[5] Fixed point equations can be derived from solving for zero derivatives of the KL divergence, although they are not presented here.

[6] In this case $Q(s) = \sum_j \lambda_j Q(s|\boldsymbol{\mu}^j)$ where $Q(s|\boldsymbol{\mu}) = \prod_i {\mu_i}^{s_i} \left(1 - \mu_i\right)^{1-s_i}$ and $\sum_j \lambda_j = 1$.



(a) Mean Field

(b) Tractable graph

Figure 4: Histogram of relative error $\ln P_{approx}(V)/\ln P_{exact}(V) - 1$ for 500 random networks - note the different scales. Mean error: (a) 0.0156 (b) 0.0020

mixture (corresponding to roughly 80 variational parameters) of 0.0114. In our approach, we use a BM to approximate the hidden unit distribution. Many choices are possible, and the one we use is displayed in fig(3), a decimatable structure. This 15 parameter BM, in conjunction with the 10 bound parameters ξ, gives a much lower mean relative error of 0.0020 when used in our method.

4 Conclusion

We have elucidated a general class of tractable undirected approximations of graphical models, based on the Kullback-Leibler divergence. Our method is complementary to other approaches since, for example, mixtures of this method are possible. We believe, however, that it is important to employ mixtures of 'strong' components, having maximal similarity with the original graph. We have also developed a similar approach based on *directed* graphical approximations, which has roughly the same accuracy, although possibly improved convergence properties [7].

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

[2] L. K. Saul and M. I. Jordan. Boltzmann Chains and Hidden Markov Models. In G. Tesauro, D. S. Touretzky, and T. K. Leen, editors, Advances in Neural Information Processing Systems, pages 435–442. MIT Press, 1995. NIPS 7.

[3] T. Jaakkola. *Variational Methods for Inference and Estimation in Graphical Models.* PhD thesis, Massachusetts Institute of Technology, 1997.

[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. MIT Press, 1998. NIPS 10.

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

[7] W. Wiegerinck and D. Barber. Mean Field Theory based on Belief Networks for Approximate Inference. 1998. ICANN 98.