On Structured Variational Approximations

Zoubin Ghahramani

Gatsby Computational Neuroscience Unit
University College London

Center for Automated Learning and Discovery
School of Computer Science
Carnegie Mellon University
Email: zoubin@gatsby.ucl.ac.uk

20 March 2002
(This is an revised and expanded version of
University of Toronto Technical Report CRG-TR-97-1)

Abstract

The problem of approximating a probability distribution occurs frequently in many areas of applied mathematics, including statistics, communication theory, machine learning, and the theoretical analysis of complex systems such as neural networks. Saul and Jordan (1996) have recently proposed a powerful method for efficiently approximating probability distributions known as structured variational approximations. In structured variational approximations, exact algorithms for probability computation on tractable substructures are combined with variational methods to handle the interactions between the substructures which make the system as a whole intractable. In this note, I present a mathematical result which can simplify the derivation of structured variational approximations in the exponential family of distributions.

1 Introduction

Belief networks provide a well-understood graphical framework for expressing the interactions between random variables. Such networks have proven useful for modeling the causal structure of complex systems of interacting variables, such as diseases and symptoms in a medical diagnosis problem. They also provide an elegant framework for understanding the relation between neural network learning algorithms and more traditional statistical models. Finally, some would argue that belief networks themselves are an appealing model of neural computation and perceptual inference in humans.

One of the essential attributes of a belief network is that it defines a graphical structure within which to do Bayesian inference in a probabilistically consistent manner. The graphical structure specifies a set of conditional independences between the variables in the network. These independences can be exploited to derive recursive algorithms for inferring the conditional probabilities of any set of variables given any other set of variables. However, for general belief networks with arbitrary connectivity and nonlinear interactions, the problem of exact inference is computationally intractable (Cooper, 1990). Therefore, in practice this intractability must be circumvented by making use of approximate algorithms for inference. Two such classes of algorithms are Markov chain Monte Carlo methods and variational approximations, both of which were developed in large part by statistical physicists modeling systems of many interacting particles.[1] In this paper we present a simple mathematical result concerning variational approximations and discuss its applicability to the practical problem of deriving learning algorithms.

2 Structured Variational Approximations

Consider the belief network shown in Figure 1a, where the shaded node corresponds to an observed variables V and the unshaded nodes correspond to hidden variables S. The presence of directed edges in the belief network expresses a set of conditional independence relations between the variables: namely, that the variable

[1] A review of Markov Chain Monte Carlo methods is provided by Neal (1993); mean field methods in physics, which are a class of variational approximation, are discussed in Parisi (1988).

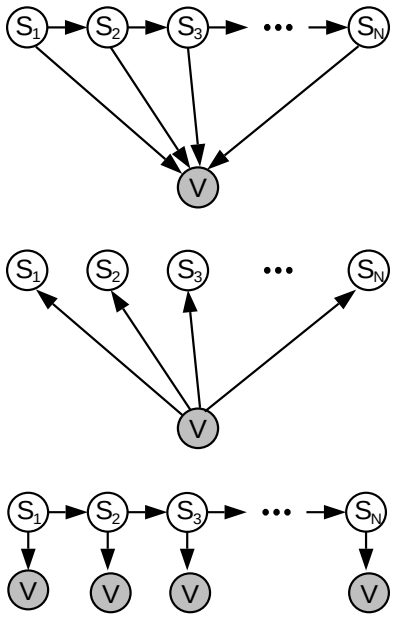

Figure 1: A belief network and two approximations.

associated with each node is conditionally independent of the variables associated with that node's non-descendants given its parents. Using these independence relations, the joint probability of S and V can be written as

$$P(S_1, \ldots, S_N, V) = P(S_1)P(S_2|S_1) \ldots P(S_N|S_{N-1})P(V|S_1, \ldots, S_N). \tag{1}$$

To illustrate the intractability of inference in this network, consider the problem of computing the conditional probability distribution of one of the hidden variables, say S_1, given the observed variable. There are at least two reasons one may want to compute this distribution. First, S_1 may be the variable of interest in an inference problem, for example, for medical diagnosis, which justifies marginalizing over the other hidden variables. Second, to estimate the parameters of the belief network using the Expectation-Maximization (EM) algorithm (Dempster et al., 1977) it is necessary to compute conditional probabilities of subsets of the hidden variables given the observed variables.

To compute this conditional probability distribution we need to sum (or integrate) over all the possible values of the hidden variables we are not directly interested in:

$$P(S_1|V) = \sum_{S_2,\ldots,S_N} P(S_1, \ldots, S_N|V) \tag{2}$$

$$= \frac{\sum_{S_2,\ldots,S_N} P(S_1, \ldots, S_N, V)}{\sum_{S_1,\ldots,S_N} P(S_1, \ldots, S_N, V)}. \tag{3}$$

For binary S_i variables, for example, this summation includes 2^N terms. Without additional constraints, there is no way of making use of the factorization in (1) to simplify this computation.

To overcome this computational cost one can approximate the conditional distribution over the hidden variables by a simpler, tractable distribution. For example,

$$Q(S_1, \ldots, S_N|V) = Q(S_1|V) \ldots Q(S_N|V)$$

assumes that given V, all the S_i are independent (Figure 1b). This complete factorization is the assumption used in simple mean field approximations in statistical mechanics. Associated with the approximating distribution Q is a vector of *variational parameters* γ, which can be optimized so as to make $Q(S|V)$ as similar as possible to $P(S|V)$. A standard measure of similarity between two probability distributions is the

Kullback-Leibler divergence (or cross-entropy):

$$KL(Q\|P) = \sum_S Q(S|V) \log \frac{Q(S|V)}{P(S|V)}. \tag{4}$$

Note that the KL-divergence is asymmetric in its two arguments; we focus on the above form of the divergence for two reasons. First, it involves averages with respect to the tractable, Q, distribution. Second, minimizing this form of the KL-divergence corresponds to maximizing a lower bound on the log likelihood, a sensible criterion for a learning algorithm (Neal and Hinton, 1998). The minimum of the KL-divergence is obtained by taking the partial derivatives of $KL(Q\|P)$ with respect to the elements of γ, which generally results in a set of fixed-point equations which can be solved iteratively.

A *structured variational approximation* is simply an approximation in which the hidden variables are not completely factorized, but rather they are related in a structured manner (Saul and Jordan, 1996). The belief network corresponding to this approximation would therefore contain some edges between the hidden variables. For example, a structured variational approximation to (1) could be

$$Q(S_1, \ldots, S_N, V) = Q(S_1)Q(S_2|S_1) \ldots Q(S_N|S_{N-1}) \cdot \frac{1}{Z} Q(V|S_1) \ldots Q(V|S_N). \tag{5}$$

The terms $Q(S_1)Q(S_2|S_1 \ldots Q(S_N|S_{N-1})$ retain the Markov chain structure connecting the hidden variables in (1); however, the rest of the terms replace the N^{th}-order interaction between the hidden variables and the observed variable by N second-order interactions. The constant Z normalizes the product of these second-order interactions so as to define a valid conditional probability of V given S_1 to S_N. A belief network representing this approximating distribution is shown in Figure 1c, which can be recognized as the belief network corresponding to a hidden Markov model (Smyth et al., 1997). However, it is a curious hidden Markov model in which a single observed variable has been replicated N times and placed at all of the visible (shaded) nodes. Regardless of the nature of these visible nodes, a fast recursive algorithm exists—the forward-backward algorithm–for calculating the posterior probabilities of the hidden variables given the visible variables.[2]

It now remains how to find parameters for Q that minimize (4). The result we present here can be used to easily determine the fixed point equations for the minimum of (4).

3 Results

Theorem 1: Exponential Families *For any distribution $P(S)$ defined over a set of variables $S = \{S_i : i \in I = \{1, \ldots N\}\}$, where $H(S)$ is defined so that*

$$P(S) = \frac{1}{Z} \exp\{-H(S)\},$$

and any approximating distribution in the exponential family parametrized by γ,

$$Q(S) = \exp\left\{ \sum_{j=1}^{K} f_j(S)\alpha_j(\gamma) + \beta(\gamma) + f_0(S) \right\} \tag{6}$$

$$= \frac{1}{Z_Q} \exp\{-H_Q(S)\}, \tag{7}$$

where $H_Q(S) = -\sum_{j=0}^{J} f_j(S)\alpha_j(\gamma)$, $\alpha_0(\gamma) = 1$ and $Z_Q = \exp\{-\beta(\gamma)\} = \sum_S \exp\{-H_Q(S)\}$, the Kullback-Leibler divergence $KL(Q\|P)$ can be minimized by iteratively solving

$$\frac{\partial \langle H_Q \rangle}{\partial \langle f_j(S) \rangle} = \frac{\partial \langle H \rangle}{\partial \langle f_j(S) \rangle} \tag{8}$$

[2] Another issue in an approximation like this one is that, while the posterior probabilities of the hidden variables might be easy to compute, the likelihood $Q(V)$ might still be intractable. In this case, computing $Q(V)$ still involves summing over all the 2^N states of the hidden variables. However, for inference, $Q(V)$ need not be computed, and during learning the algorithm will still maximize a lower bound on the likelihood, $P(V)$.

for all $j = 0, \ldots J$, where $\langle \cdot \rangle$ denotes expectation over the approximating distribution Q.

Proof. We start from the definition of the KL-divergence

$$KL(Q\|P) = \sum_S Q(S) \log \frac{Q(S)}{P(S)} \tag{9}$$

$$= \langle H \rangle - \langle H_Q \rangle + \log Z - \log Z_Q. \tag{10}$$

Expanding the four terms using the chain rule and the definitions of the relevant quantities we obtain

$$\frac{\partial \langle H \rangle}{\partial \gamma} = \sum_j \frac{\partial \langle H \rangle}{\partial \langle f_j(S) \rangle} \frac{\partial \langle f_j(S) \rangle}{\partial \gamma} \tag{11}$$

$$\frac{\partial \langle H_Q \rangle}{\partial \gamma} = -\frac{\partial}{\partial \gamma} \sum_S \sum_j f_j(S) \alpha_j(\gamma) Q(S) \tag{12}$$

$$= -\frac{\partial}{\partial \gamma} \sum_j \alpha_j(\gamma) \sum_S f_j(S) Q(S) \tag{13}$$

$$= -\sum_j \frac{\partial \alpha_j(\gamma)}{\partial \gamma} \langle f_j(S) \rangle - \sum_j \alpha_j(\gamma) \frac{\partial \langle f_j(S) \rangle}{\partial \gamma} \tag{14}$$

$$\frac{\partial \log Z}{\partial \gamma} = 0 \tag{15}$$

$$\frac{\partial \log Z_Q}{\partial \gamma} = \frac{1}{Z_Q} \frac{\partial}{\partial \gamma} \sum_S \exp \left\{ \sum_j f_j(S) \alpha_j(\gamma) \right\} \tag{16}$$

$$= \sum_j \frac{\partial \alpha_j(\gamma)}{\partial \gamma} \langle f_j(S) \rangle \tag{17}$$

Combining terms we obtain

$$\frac{\partial KL}{\partial \gamma} = \sum_j \left[\alpha_j(\gamma) + \frac{\partial \langle H \rangle}{\partial \langle f_j(S) \rangle} \right] \frac{\partial \langle f_j(S) \rangle}{\partial \gamma}. \tag{18}$$

Using $\alpha_j(\gamma) = -\frac{\partial \langle H_Q \rangle}{\partial \langle f_j(S) \rangle}$ we get that the zeros of the system of equations defined by (8) are also zeros of the system of equations defined by (18). QED

Corollary. If $H_Q(S)$ is an m^{th}-order polynomial in S,

$$H_Q(S) = \sum_{i_1 \in I} \gamma_{i_1} S_{i_1} + \sum_{i_1, i_2 \in I} \gamma_{i_1, i_2} S_{i_1} S_{i_2} \ldots + \sum_{i_1, \ldots i_m \in I} \gamma_{i_1, \ldots i_m} S_{i_1} \ldots S_{i_m},$$

then the variational fixed point equations set the coefficients of H_Q equal to the corresponding derivatives of $\langle H \rangle$.

$$\gamma_{i_1} = \frac{\partial \langle H \rangle}{\partial \langle S_{i_1} \rangle} \tag{19}$$

$$\gamma_{i_1, i_2} = \frac{\partial \langle H \rangle}{\partial \langle S_{i_1} S_{i_2} \rangle} \tag{20}$$

$$\vdots$$

$$\gamma_{i_1, \ldots i_m} = \frac{\partial \langle H \rangle}{\partial \langle S_{i_1} \ldots S_{i_m} \rangle}, \tag{21}$$

Remark 1. The exponential family of distributions includes many models of interest, e.g., Boltzmann machines, graphical Gaussian models, hidden Markov models, decision trees with multinomial variables.

However, it does not include mixture models (unless the mixture component is explicitly represented by a hidden random variable) or sigmoid belief networks, for example.

Remark 2. Expressing $\langle H \rangle$ in terms of the $\langle f_j(S) \rangle$ may sometimes be difficult.

Theorem 2: Further Factorizations *If*

$$P(S,V) = \frac{1}{Z} \prod_i f(C_i)$$

where C_i are (possibly overlapping) subsets of variables $C_i \subseteq \{S,V\}$ and

$$Q(S) = \prod_j Q_j(K_j)$$

where K_j are (non-overlapping) subsets of variables $K_j \subseteq \{S\}$, then finding a variational approximation that maximizes

$$F_1(Q) = \sum_s Q(S) \log \frac{P(S,V)}{Q(S)}$$

is equivalent to maximizing

$$F_2(\tilde{Q}) = \sum_s \tilde{Q}(S) \log \frac{P(S,V)}{\tilde{Q}(S)}$$

where

$$\tilde{Q}(S) = \frac{1}{\tilde{Z}} \prod_{i,j} \tilde{Q}_{ij}(C_{ij})$$

where $C_{ij} = C_i \cap K_j$.

Proof. Writing out the variational lower bound:

$$F_1 = \sum_S \prod_j Q_j(K_j) [\sum_i f_i(C_i) - \log Z_P - \sum_j \log Q_j(K_j)]$$

taking partial derivatives with respect to one of the distributions:

$$\frac{\partial F_1}{\partial Q_j(K_j = k)} = \sum_{S \setminus K_j} \prod_{j' \neq j} Q_{j'}(K_{j'}) \sum_i f_i(C_i) - \log Q_j(K_j = k) - 1 + \lambda_j$$

where λ_j is a Lagrange multiplier ensuring that Q_j sums to one. Define

$$f_{ij}(C_i j) \equiv \sum_{S \setminus K_j} \prod_{j' \neq j} Q_{j'}(K_{j'}) f_i(C_i)$$

then

$$\frac{\partial F_1}{\partial Q_j(K_j = k)} = \sum_{i : C_i \cap K_j \neq \emptyset} f_{ij}(C_i j) + \text{const} - \log Q_j(K_j = k)$$

Solving we get:

$$Q_j(K_j) = \frac{1}{Z} \prod_i Q_{ij}(C_i j)$$

which is the same solution we would have gotten had we maximized F_2 w.r.t. $\tilde{Q}$. QED.

4 Discussion

Several points are important to make. First, this result is especially useful if both $P(S)$ and $Q(S)$ can be defined as ploynomials in S. In this case, the fixed point equations can be obtained almost by inspection, simply by equating terms with correspondings powers of S. Second, there may be multiple parametrizations of the approximating distribution in terms of polynomials in S. For example, while the result is expressed for a model in which there are interactions of every order up to the m^{th}, such models can be written in terms of only m^{th} order interactions by subsuming the effect of lower order parameters into the higher order parameters. Finally, this result is meant as a practical tool. More laborious alternative derivations of variational approximations are also generally possible.

References

Cooper, G. F. (1990). The computational complexity of probabilistic inference using Bayesian belief networks. *Artificial Intelligence*, 42(2-3):393–405.

Dempster, A., Laird, N., and Rubin, D. (1977). Maximum likelihood from incomplete data via the EM algorithm. *J. Royal Statistical Society Series B*, 39:1–38.

Neal, R. M. (1993). Probabilistic inference using Markov chain monte carlo methods. Technical Report CRG-TR-93-1, Department of Computer Science, University of Toronto.

Neal, R. M. and Hinton, G. E. (1998). A new view of the EM algorithm that justifies incremental, sparse, and other variants. In Jordan, M. I., editor, *Learning in Graphical Models*. Kluwer Academic Press.

Parisi, G. (1988). *Statistical Field Theory*. Addison-Wesley, Redwood City, CA.

Saul, L. and Jordan, M. I. (1996). Exploiting tractable substructures in Intractable networks. In Touretzky, D., Mozer, M., and Hasselmo, M., editors, *Advances in Neural Information Processing Systems 8*. MIT Press.

Smyth, P., Heckerman, D., and Jordan, M. I. (1997). Probabilistic independence networks for hidden Markov probability models. *Neural Computation*, 9:227–269.