Variational Learning in Graphical Models and Neural Networks

Christopher M. Bishop
Microsoft Research,
7 J. J. Thomson Avenue,
Cambridge, CB3 0FB, U.K.
cmbishop@microsoft.com
http://research.microsoft.com/∼cmbishop

Invited paper, published in *Proceedings 8th International Conference on Artificial Neural Networks, ICANN'98*, L. Niklasson *et al.* (eds), Springer (1998) **1**, 13-22.

Abstract

Variational methods are becoming increasingly popular for inference and learning in probabilistic models. By providing bounds on quantities of interest, they offer a more controlled approximation framework than techniques such as Laplace's method, while avoiding the mixing and convergence issues of Markov chain Monte Carlo methods, or the possible computational intractability of exact algorithms. In this paper we review the underlying framework of variational methods and discuss example applications involving sigmoid belief networks, Boltzmann machines and feed-forward neural networks.

1 Introduction

Probability theory provides a principled, consistent framework for quantification of uncertainty, and as such underpins much of the current research in neural computing. A central concept in probabilistic inference is that of *marginalization* involving summing (or integrating) over the distribution of unobserved variables. For many models, however, these summations are computationally intractable, and so it is necessary to resort to approximation schemes. Variational methods [**?**] provide a new framework for inference and learning in probabilistic models, which complement previous approaches and offer some specific advantages.

A probabilistic model defines a joint distribution over a set S of random variables. We can partition these variables into two groups corresponding to visible (observed) variables V and the remaining hidden (or latent) variables H. Typically the model is governed by a set of adaptive parameters $\mathbf{w}$, and is then described by a joint distribution $P(H, V|\mathbf{w})$, conditioned on the parameters. The distribution of observed variables is obtained by marginalization over the hidden variables, so that

$$L(\mathbf{w}) \equiv P(V|\mathbf{w}) = \sum_H P(H, V|\mathbf{w}) \tag{1}$$

where the sum over H is replaced by an integration in the case of continuous variables. The quantity $L(\mathbf{w})$ in (1) is the likelihood function, and maximization of the likelihood (or equivalently its logarithm) can be used to estimate a value for the parameter vector $\mathbf{w}$. The likelihood also plays an important role in a Bayesian treatment of the parameters $\mathbf{w}$. Another central quantity of interest is the posterior distribution $P(H|V)$ of the hidden variables, given the observed

variables. This is given, from Bayes' theorem, by

$$P(H|V,\mathbf{w}) = \frac{P(H,V|\mathbf{w})}{P(V|\mathbf{w})} \tag{2}$$

which again involves the likelihood function. For complex probabilistic models, the summation over H needed to evaluate $L(\mathbf{w})$ may involve an exponentially large number of terms and may therefore be computationally intractable.

Variational methods involve the introduction of an approximating distribution $Q(H|V)$ which, as we shall see shortly, provides an approximation to the true posterior distribution. Consider the following transformation applied to the log likelihood function

$$\begin{aligned} \ln P(V|\mathbf{w}) &= \ln \sum_H P(H,V|\mathbf{w}) && (3)\\ &= \ln \sum_H Q(H|V) \frac{P(H,V|\mathbf{w})}{Q(H|V)} && (4)\\ &\geq \mathcal{L}(Q,\mathbf{w}) = \sum_H Q(H|V) \ln \frac{P(H,V|\mathbf{w})}{Q(H|V)} && (5) \end{aligned}$$

where we have applied Jensen's inequality. We see that the function $\mathcal{L}(Q,\mathbf{w})$ forms a rigorous lower bound on the true log likelihood. The significance of this transformation is that, through a suitable choice for the Q distribution, the quantity $\mathcal{L}(Q,\mathbf{w})$ may be tractable to compute, even though the original log likelihood function is not. From (5) it is easy to see that the difference between the true log likelihood $\ln P(V|\mathbf{w})$ and the bound $\mathcal{L}(Q,\mathbf{w})$ is given by

$$\mathrm{KL}(Q\|P) = -\sum_H Q(H|V) \ln \frac{P(H|V,\mathbf{w})}{Q(H|V)} \tag{6}$$

which is the Kullback-Leibler (KL) divergence between the approximating distribution $Q(H|V)$ and the true posterior $P(H|V,\mathbf{w})$. The relationship between the various quantities is shown in Figure 1.

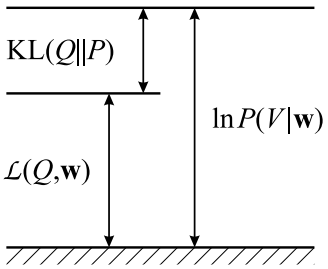

Figure 1: The quantity $\mathcal{L}(Q,\mathbf{w})$ provides a rigorous lower bound on the true log likelihood $\ln P(V|\mathbf{w})$, with the difference being given by the Kullback-Leibler divergence $\mathrm{KL}(Q\|P)$ between the approximating distribution $Q(H|V)$ and the true posterior $P(H|V,\mathbf{w})$.

The goal in a variational approach is to choose a suitable form for $Q(H|V)$ which is sufficiently simple that the lower bound $\mathcal{L}(Q,\mathbf{w})$ can readily be evaluated and yet which is sufficiently flexible that the bound is reasonably tight. We generally choose some family of Q distributions and then seek the best approximation within this family by maximizing the lower bound. Since the true log likelihood is independent of Q we see that this is equivalent to minimizing the Kullback-Leibler divergence. Note that for data sets consisting of N independent observations we need to perform the variational optimization separately for each observation.

Suppose we consider a completely free-form optimization over Q, allowing for all possible Q distributions. Using the well-known result that the KL divergence between two distributions $Q(H)$ and $P(H)$ is minimized by $Q(H) = P(H)$ we see that the optimal Q distribution is given by the true posterior, in which case the KL divergence is zero and the bound becomes exact. However, this will not lead to any simplification of the problem. In order to make progress it is necessary to consider a more restricted range of Q distributions.

One approach is to consider a parametric family of Q distributions of the form $Q(H|V, \psi)$ governed by a set of parameters ψ. We can then adapt ψ by minimizing the KL divergence to find the best approximation within this family. In Section 2 we will see that the graphical models perspective provides a natural framework for motivating suitable choices for parametric Q distributions.

In some applications an alternative approach can be adopted which is to restrict the functional form of $Q(H|V)$ by assuming that it factorizes over the component variables $\{h_i\}$ in H, so that

$$Q(H|V) = \prod_i Q_i(h_i|V). \tag{7}$$

The KL divergence can then be minimized over all possible factorial distributions by performing a free-form minimization over the Q_i, leading to the following result

$$Q_i(h_i|V) = \frac{\exp \langle \ln P(H, V|\mathbf{w}) \rangle_{k \neq i}}{\sum_j \exp \langle \ln P(H, V|\mathbf{w}) \rangle_{k \neq j}} \tag{8}$$

where $\langle \ \cdot \ \rangle_{k \neq i}$ denotes an expectation with respect to the distributions $Q_k(h_k|V)$ for all $k \neq i$.

There is an interesting relationship between this variational framework and the expectation maximization (EM) algorithm, as pointed out by Neal and Hinton [**?**]. If we return to the free-form optimization over Q, then we have noted that this involves the evaluation of the posterior distribution $P(H|V, \mathbf{w})$ for a given value of $\mathbf{w} = \mathbf{w}_{\text{old}}$ and hence corresponds to the E-step of the EM algorithm. If we back-substitute the result $Q(H|V) = P(H|V, \mathbf{w}_{\text{old}})$ into $\mathcal{L}$ then we obtain (up to an additive term independent of $\mathbf{w}$) the expected complete-data log likelihood whose maximization over $\mathbf{w}$ for fixed $\mathbf{w}_{\text{old}}$ constitutes the M-step of the standard EM algorithm [**?**]. It is well known that there is a generalized EM (GEM) algorithm in which at the M-step the expected complete-data log likelihood is only increased and not fully maximized. The above relationship to the variational lower bound demonstrates that it is also possible to generalize to a partial E-step in which the Q distribution is adjusted to increase $\mathcal{L}(Q, \mathbf{w})$ without actually reducing the KL divergence to zero. Since this generalized EM algorithm is increasing $\mathcal{L}$ at every step it is guaranteed to result in a stable algorithm.

2 Directed Graphs: Sigmoid Belief Networks

As a first application of the variational framework we consider sigmoid belief networks, which correspond to directed, acyclic graphs in which each node i represents a binary stochastic variable $S_i \in \{0, 1\}$. The joint distribution in a directed graph is defined by specifying the conditional distribution of each variable given the states of its parent variables in the graph. In the case of the sigmoid belief network, the probability of a node being 'on' is given by a sigmoidal function of a linear combination of the parent values of the form

$$P(S_i = 1|\Pi_i) = \sigma \left(\sum_j w_{ij} S_j \right) \tag{9}$$

where $\sigma(z) \equiv (1 + e^{-z})^{-1}$ is the logistic sigmoid function, Π_i denote the parents of S_i in the network, and w_{ij} represent the adaptive parameters in the model (note that we have implicitly

absorbed a bias parameter for each node into the $\{w_{ij}\}$). The variables S can be grouped into visible V and hidden H. Evaluation of the likelihood function, using (1), requires summing over all $2^{|H|}$ configurations of the states of the hidden units, where $|H|$ denotes the number of hidden variables. Although there are efficient algorithms for evaluating such sums in polynomial time for simple graphs such as trees, they are no longer applicable to densely connected graphs, and so for networks with more than a few hidden units we are forced to consider approximations.

A simple variational approach to learning in sigmoid belief networks was introduced by Saul *et al.* [**?**] in which the Q distribution is chosen to be factorized (this is called *mean field theory*). Specifically we consider a Q given by a product of Bernoulli distributions in the form

$$Q(H|V) = \prod_i \mu_i^{h_i}(1-\mu_i)^{1-h_i} \tag{10}$$

in which we have introduced a mean-field parameter μ_i corresponding to each of the hidden variables h_i. Even this severely restricted (fully factorized) choice of Q distribution does not lead to an analytically tractable expression for the summation over H in (5) and a further approximation is required. However, since this takes the form of a generalization of Jensen's inequality, it still maintains a rigorous bound on the true log likelihood. Setting the derivatives of the bound with respect to the μ_i to zero then leads to a set of re-estimation equations for μ_i which can be iterated until some convergence criterion is satisfied. This represents the E-step of a generalized EM algorithm. The corresponding M-step is obtained by computing the derivatives of the bound with respect to the parameters w_{ij} and taking a step in the negative gradient direction.

Although mean field theory leads to a workable learning algorithm, it is based on a very restricted class of Q distributions. It is natural to seek more flexible distributions, which nevertheless remain tractable, in order to obtain an improved learning algorithm. One specific limitation of a factorized approximation is that it cannot capture multi-modality, and this can be overcome by considering a probabilistic *mixture* of mean field distributions of the form

$$Q(H|V) = \sum_{m=1}^{M} \alpha_m Q(H|V,m) \tag{11}$$

in which each of the components $Q(H|V,m)$ has the form (10) with its own independent set of mean field parameters. A general framework for variational inference using mixtures was proposed by Jaakkola and Jordan [**?**] and has been applied to sigmoid belief networks by Bishop *et al.* [1]. Although this leads to a more complex framework than for simple mean field theory it is nevertheless possible to obtain a computationally tractable algorithm while preserving a rigorous lower bound on the log likelihood.

Insight into these approximations can be obtained by considering the corresponding graphical representations as shown in Figure 2. Mean field theory treats the hidden nodes as independent and is represented by a graph with no links, whereas a mixture of mean field distributions is obtained by introducing an extra discrete latent variable m, giving a tree structured graph. This graphical perspective motivates the use of other approximating distributions corresponding to simple graphical structures. For instance, the Markov chain approximation, corresponding to (d) in Figure 2, has been considered in [**?**].

3 Undirected Graphs: Boltzmann Machines

In this section we consider the application of variational methods to probabilistic models which correspond to undirected graphs, and in particular we focus on the Boltzmann machine [**?**]. The

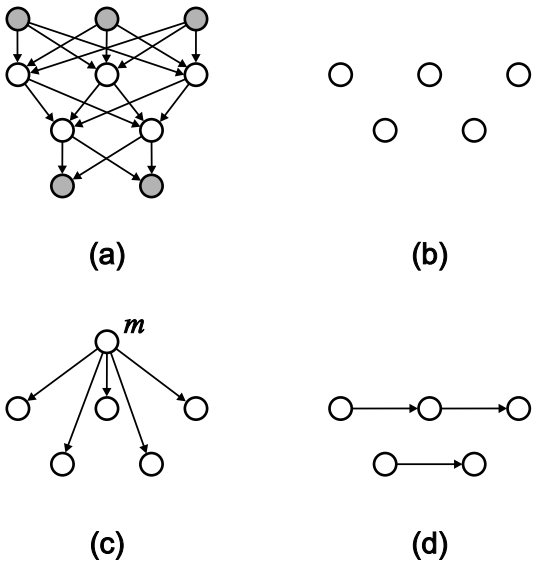

Figure 2: (a) Graphical representation of a densely connected sigmoid belief network with visible nodes in black and hidden nodes in white. (b) The corresponding mean field approximation. (c) A mixture of mean field distributions. (d) A Markov chain approximation.

nodes of a Boltzmann machine graph represent two-state stochastic variables for which the joint distribution has the Boltzmann form

$$P(S) = \frac{\exp(-E(S))}{Z} \tag{12}$$

in which $S = \{s_i\}$ denotes the set of stochastic variables, and $E(S)$ denotes the energy of a particular configuration given by a quadratic function of the states

$$E(S) = -\sum_i \sum_{j>i} w_{ij} s_i s_j. \tag{13}$$

Here $w_{ij} = 0$ for nodes which are not neighbours on the graph, and again biases are treated implicitly. The normalization factor Z^{-1} in (12) is called the *partition function* in statistical physics terminology, and is obtained by marginalizing the numerator over all configurations of states

$$Z = \sum_S \exp(-E(S)). \tag{14}$$

If there are L variables in the network, the number of configurations of states is 2^L, and so evaluation of Z may require exponential time (e.g. for fully connected models) and hence, in the worst case, is computationally intractable.

Again, we can partition the units S into visible V and hidden H. Learning in the Boltzmann machine is achieved by maximizing the likelihood (1) with respect to the parameters $\{w_{ij}\}$ using gradient methods. Differentiating the log of the likelihood (1) and using (12), (13) and (14) we obtain

$$\frac{\partial \ln P(V)}{\partial w_{ij}} = \langle s_i s_j \rangle_{\mathrm{C}} - \langle s_i s_j \rangle_{\mathrm{F}} \tag{15}$$

where $\langle \cdot \rangle_{\mathrm{C}}$ denotes an expectation with respect to the *clamped* distribution $P(H|V)$ while $\langle \cdot \rangle_{\mathrm{F}}$ denotes expectation with respect to the *free* distribution $P(H, V)$.

Evaluation of the expectations in (15) requires summing over exponentially many states, and so is intractable for densely connected models. The original learning algorithm for Boltzmann

machines made use of Gibbs sampling to generate separate samples from the joint and conditional distributions over states, and used these to evaluate the required gradients. A serious limitation of this approach, however, is that the gradient is expressed as the difference between two Monte Carlo estimates and is thus very prone to sampling error. This results in a very slow learning algorithm.

Mean field theory, equivalent to the use of a variational approximation involving a factorized Q distribution, was developed for Boltzmann machines by Peterson *et al.* [**?**]. It allows the stochastic averages $\langle s_i s_j \rangle$ to be approximated by deterministic products of the corresponding mean field parameters $\mu_i \mu_j$. Optimization of the Q distribution yields deterministic re-estimation equations for the mean field parameters, and so we again obtain a generalized EM algorithm in which optimization of Q is alternated with adaptation of the model parameters w_{ij} using gradient ascent.

For most applications it is likely that the distribution in the unclamped phase of a Boltzmann machine will be strongly multi-modal, and so mean field theory is likely to provide a poor framework. In this case we can expect a significant improvement over mean field theory to be obtained from the use of a mixture representation. Lawrence *et al.* [2] have successfully applied mixtures of mean field distributions to Boltzmann machines, for which the second-order expectation $\langle s_i s_j \rangle$ is approximated by the deterministic expression $\sum_{m=1}^{M} \alpha_m \mu_i^m \mu_j^m$. They demonstrated significantly improved inference compared with standard mean field theory, as illustrated in Figure 3.

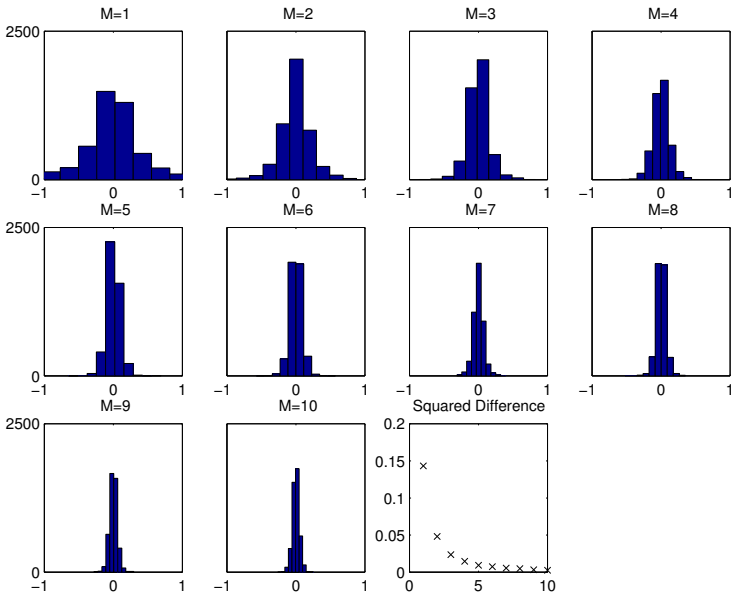

Figure 3: Histograms of the differences between the true free expectation $\langle s_i s_j \rangle_{\mathrm{F}}$ and the mean field approximation $\sum_{m=1}^{M} \alpha_m \mu_i^m \mu_j^m$ for 100 randomly generated networks each having 55 independent parameters, for different numbers M of components in the mixture approximation, together with a summary of the dependence of the sum-of-squares of the differences on M.

One limitation of the variational approach in the context of undirected graphs arises from the presence of the partition function Z (absent in directed graphical models) which leads to the derivatives with respect to model parameters involving the difference of two expectations, as in (15). Since the difference in two bounds is no longer a bound, we lose this elegant aspect of the variational approach. Nevertheless, improved inference can still lead to improved learning, as demonstrated for the case of hand-written digits in [2].

4 Neural Networks

So far we have considered maximum likelihood techniques which estimate specific values for the model parameters $\mathbf{w}$. In a Bayesian treatment, prior distributions are defined over the parameters, and the parameters are marginalized out. From the graphical models perspective this corresponds to viewing the parameters as additional nodes, and hence they can be treated on the same footing as other stochastic variables. Since the marginalization over model parameters is often analytically intractable it can again prove useful to consider variational methods.

In the context of regression, a neural network is used to define a conditional Gaussian distribution of the form $p(t|\mathbf{x}, \mathbf{w}, \beta) = \mathcal{N}(f(\mathbf{x}, \mathbf{w}), \beta^{-1})$ where $\mathbf{x}$ is the input vector, and β is the inverse 'noise' variance [3]. The neural network function $f(\mathbf{x}, \mathbf{w})$ maps the input vector to the mean of the distribution and is governed by a vector $\mathbf{w}$ of weights and biases. In a Bayesian treatment we define a prior over $\mathbf{w}$ given, for instance, by a Gaussian $p(\mathbf{w}|\alpha) = \mathcal{N}(0, \alpha^{-1})$ where α is the inverse variance. Hyper-priors are also defined over the hyper-parameters α and β. Given a data set of labelled observations $D = \{\mathbf{x}_n, t_n\}_{n=1}^N$ the posterior distribution over parameters is given by

$$p(\mathbf{w}|D) \propto p(\mathbf{w}|\alpha) \prod_{n=1}^{N} p(t_n|\mathbf{x}_n, \mathbf{w}, \beta). \tag{16}$$

Predictive distributions are then obtained by marginalizing over $\mathbf{w}$, so that for instance the predictive mean, for a new value of $\mathbf{x}$, is given by

$$\langle t|\mathbf{x}\rangle = \int f(\mathbf{x}, \mathbf{w}) p(\mathbf{w}|D)\, d\mathbf{w}. \tag{17}$$

For non-linear network models the integration over $\mathbf{w}$ is analytically intractable. One approach [**?**] is to use a Gaussian approximation to the posterior around a mode, in which the covariance is determined by the local curvature of the posterior at the mode. However, a more general treatment can be obtained using variational methods.

The first variational treatment of neural networks was given by Hinton and van Camp [**?**] who considered a variational distribution $Q(\mathbf{w})$ given by a Gaussian distribution which was assumed to factorize over the components of $\mathbf{w}$. This was extended to Gaussian Q distributions having a general covariance matrix by Barber and Bishop [4]. A comparison of the approximations obtained using these three schemes is shown in Figure 4. This variational approach can readily

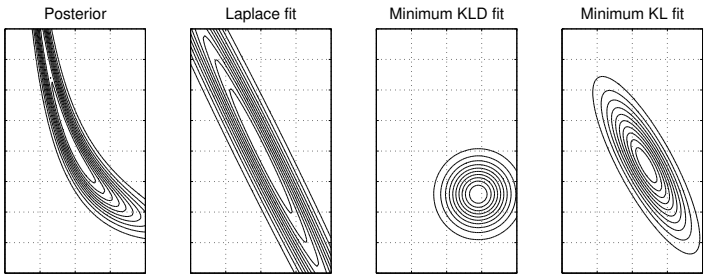

Figure 4: Laplace and variational Gaussian fits to the posterior for a two-parameter synthetic regression problem. The Laplace method underestimates the local posterior mass by basing the covariance matrix on the mode alone, and has KL value 41. The variational Gaussian fit with a diagonal covariance matrix (KLD) gives a residual KL value of 4.6, while the variational Gaussian with full covariance matrix achieves a KL value of 3.9.

be extended to a mixture of Gaussians using the framework discussed earlier in the context of graphical models.

5 Discussion

In this paper we have reviewed the general framework of variational methods, and outlined some applications to graphical models and neural networks. Currently the choice of an appropriate family of Q distributions for a specific application is largely a matter of judgement, and no systematic procedures have so far been developed. Furthermore, although variational methods often provide the reassurance of a rigorous bound on quantities of interest (in contrast to many other approximation schemes) the accuracy of this bound can be difficult to quantify. In spite of these limitations, however, variational methods have already found several successful applications and are likely to be widely used in the future.

Acknowledgements

I would like to thank Brendan Frey, Tommi Jaakkola, Michael Jordan, Neil Lawrence, David MacKay and Michael Tipping for helpful discussions regarding variational methods.

References

[1] 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, pages 416–422, 1997.

[2] N. Lawrence, C. M. Bishop, and M. Jordan. Mixture representations for inference and learning in Boltzmann machines. In *Uncertainty in Artificial Intelligence*, volume 14, pages 320–327. Morgan Kaufmann, 1998.

[3] C. M. Bishop. *Neural Networks for Pattern Recognition*. Oxford University Press, 1995.

[4] D. Barber and C. M. Bishop. Ensemble learning in Bayesian neural networks. In C. M. Bishop, editor, *Generalization in Neural Networks and Machine Learning*, pages 215–237. Springer Verlag, 1998.