Notes on Variational Learning

W.D. Penny and S.J. Roberts

Technical Report PARG-2000-02,
Department of Engineering Science,
Oxford University.

July 27, 2000

1 Maximum Likelihood with Hidden Variables

In probabilistic models where some variables, V, are observed and other variables, H, are hidden, learning of the model parameters, θ, can be achieved by variational learning or, in some cases, Expectation-Maximisation (EM). Both approaches rely on a fundamental relationship between log likelihood, variational free energy and the Kullback-Liebler (KL) divergence. This relationship can be derived as follows. Firstly

$$p(V \mid \theta) = \frac{p(H, V \mid \theta)}{p(H \mid V, \theta)} \tag{1}$$

This means that the log-likelihood, $L(\theta) \equiv \log p(V \mid \theta)$, can be written

$$L(\theta) = \log p(H, V \mid \theta) - \log p(H \mid V, \theta) \tag{2}$$

If we now take expectations with respect to the distribution $Q_\lambda(H)$, where λ are the parameters of the distribution (the 'variational' parameters) then we get

$$L(\theta) = \int Q_\lambda(H) \log p(H, V \mid \theta) dH - \int Q_\lambda(H) \log p(H \mid V, \theta) dH \tag{3}$$

where the term on the left is unchanged as it does not depend on H. By replacing $p(H \mid V, \theta)$ with $p(H \mid V, \theta) Q_\lambda(H)/Q_\lambda(H)$ and re-arranging we get

$$L(\theta) = F(\lambda, \theta) + D(Q \mid\mid p) \tag{4}$$

where

$$F(\lambda, \theta) = \int Q_\lambda(H) \log \frac{p(H, V \mid \theta)}{Q_\lambda(H)} dH \tag{5}$$

$$D(Q \mid\mid p) = \int Q_\lambda(H) \log \frac{Q_\lambda(H)}{p(H \mid V, \theta)} dH \tag{6}$$

The term $-F(\lambda, \theta)$ is known (to physicists) as the variational free energy and $D(Q \mid\mid p)$ is the KL-divergence. From Jensen's inequality it can be proven that $D(Q \mid\mid p) \geq 0$ with equality if $Q = p$ [7]. This means that $F(\lambda, \theta)$ is a strict lower bound on $L(\theta)$

$$L(\theta) \geq F(\lambda, \theta) \tag{7}$$

with equality if $Q_\lambda(H) = p(H \mid V, \theta)$.

1.1 Variational learning and EM

Variational learning may be viewed as consisting of two distinct steps:

Step 1: Approximate E-step With model parameters fixed at θ_{t-1}, update the variational parameters λ to maximise $F(\lambda, \theta)$.
Step 2: M-step With variational parameters fixed at λ_t, update the model parameters θ to maximise $F(\lambda, \theta)$.

These steps are iterated as necessary and are analagous to the E and M steps of EM learning. In the approximate E-step we update the parameters of the approximating density (the so-called variational parameters) and in the M-step we update the parameters of the probabilistic model.

The EM algorithm is a special case of variational learning, where in the E step we are able to set $Q_\lambda(H) = p(H \mid V, \theta)$ [14]

E-step With model parameters fixed at θ_{t-1}, set $Q_\lambda(H) = p(H \mid V, \theta)$.
M-step With variational parameters fixed at λ_t, update the model parameters θ to maximise $F(\lambda, \theta)$.

Note that the M-steps are identical, but, in variational learning, the exact E-step is replaced with an approximate E-step where we *minimise* the KL-divergence between the approximating density and the posterior over hidden variables instead of setting it to zero.

In Gaussian Mixture Models $p(H \mid V, \theta)$ is trivial to compute and in Hidden Markov Models, for example, this density can be computed from the forward-backard algorithm. For these models EM learning is therefore tractable. In more complex models (such as Independent Factor Analysis [2]), however, $p(H \mid V, \theta)$ is not easy (or efficient) to compute and in these cases a variational learning approach is appealing. See [13] for more details.

2 Bayesian Learning

Variational learning can also be used to approximate posterior densities in the Bayesian framework. The 'evidence' or 'marginal likelihood' of a probabilistic model, $p(V)$, is the likelihood of the model, $p(V|\theta)$, after its parameters θ have been integrated out.

From

$$p(V) = \frac{p(V, \theta)}{p(\theta|V)} \tag{8}$$

the log-evidence can be written as

$$\log p(V) = \log p(V, \theta) - \log p(\theta|V) \tag{9}$$

If we now take expectations wrt. the 'approximate posterior' or 'variational posterior', $q(\theta|V)$, then

$$\log p(V) = \int q(\theta|V) \log p(V, \theta) d\theta - \int q(\theta|V) \log p(\theta|V) d\theta \tag{10}$$

where the left hand side remains unchanged as it does not depend on θ. Writing $LogEv \equiv \log p(V)$ and multiplying the probability $p(\theta|V)$ top and bottom by $q(\theta|V)$ and rearranging gives

$$LogEv = F(\theta) + D(q(\theta|V)||p(\theta|V)) \tag{11}$$

where

$$F(\theta) = \int q(\theta|V) \log \frac{p(V, \theta)}{q(\theta|V)} d\theta \tag{12}$$

and $D(q(\theta|V)||p(\theta|V))$ is the KL-divergence between the approximate posterior and the true posterior. As $D(Q \mid\mid p) \geq 0$ with equality if $Q = p$ [7] this means that $F(\theta)$ is a strict lower bound on $LogEv$

$$LogEv \geq F(\theta) \tag{13}$$

with equality if the approximate posterior equals the true posterior.

The aim of Variational Bayesian (VB) learning is to maximise this lower bound and therefore make the approximate posterior as close as possible to the true posterior. By using $p(V, \theta) = p(V|\theta)p(\theta)$ we can expand equation 12 into two terms

$$F(\theta) = \int q(\theta|V) \log p(V|\theta) d\theta - D(q(\theta|V)||p(\theta)) \tag{14}$$

where the first term is the average likelihood of the data and the second term is the KL-divergence between the approximating posterior and the *prior* (in equation 11, the KL-divergence was between the approximate posterior and the true posterior). This acts as a penalty term which penalizes more complex models.

3 Bayesian Learning with Hidden Variables

If there are hidden variables then the Bayesian approach can be implemented by maximising [4]

$$F(\theta) = \int q(\theta, H|V) \log \frac{p(V, H, \theta)}{q(\theta, H|V)} d\theta dH \tag{15}$$

By assuming the approximating posterior factorises into separate distributions over the model parameters and hidden variables

$$q(\theta, H|V) = q(\theta|V)q(H|V) \tag{16}$$

and using the decomposition

$$p(V, H, \theta) = p(V, H|\theta)p(\theta) \tag{17}$$

equation 15 can be expanded into two terms

$$F(\theta) = \int q(\theta|V)q(H|V) \log \frac{p(V, H|\theta)}{q(H|V)} d\theta dH - D(q(\theta|V)||p(\theta)) \tag{18}$$

where the first term is the average likelihood and the second term is the KL-divergence between the approximate parameter posterior and the parameter prior.

4 Review of Applications

The applications of variational learning may be characterised according to what form of approximating density $Q_\lambda(\boldsymbol{w})$ is used and what probabilistic model is involved. The choice of approximating density is particularly important; we need to choose a density which can make a good approximation to the true posterior but which, at the same time, is sufficiently simple to enable the variational free energy to be computed.

To date, approaches have used three main categories of approximating density (i) a diagonal Gaussian ensemble, (ii) factorised densities $Q_\lambda(\boldsymbol{w}) = \prod_i Q_\lambda(w_i)$ and (iii) mixture densities.

In 1993 Hinton and Van Camp used a diagonal Gaussian ensemble applied to a Bayesian multilayer perceptron (MLP) [9]. Lappalainen [8] has recently used a diagonal Gaussian ensemble in a Bayesian ICA model with Gaussian mixture sources.

Factorised densities have been used by Mackay for Bayesian linear regression with known observation noise variance [11] and Bayesian Hidden Markov Models [12]. They have also been used by Bishop for Bayesian Principal Component Analysis [5]. Attias has also used factorised densities for maximum likelihood estimation in an Independent Factor Analysis (IFA) model [2] with Gaussian Mixture sources and more recently for IFA with Hidden Markov sources [1]. More recently Attias has proposed variational methods for Bayesian Mixture Models and Bayesian Blind Source Separation [3].

The use of factorised densities has attracted much attention because it is not necessary to assume any particular parametric form for each component of the approximating density; the form is imposed simply by the overall factorisation and the fact that you want to minimise the variational free energy. See for example [5] and [11].

Bishop has used mixture densities as the approximating densities in belief networks [6] and Jaakkola and Jordan examine the general framework [10].

References

[1] H. Attias. Blind Source Separation by Dynamic Graphical Models. In *International Conference on Artificial Neural Networks*, pages 126–131, 1999.

[2] H. Attias. Independent Factor Analysis. *Neural Computation*, 11:803–851, 1999.

[3] H. Attias. Inferring Parameters and Structure of Latent Variable Models by Variational Bayes. In *Proceedings of the Fifteenth Conference on Uncertainty in Artificial Intelligence*, 1999.

[4] H. Attias. A Variational Bayesian Framework for Graphical Models. In T. Leen et al, editor, *NIPS 12*, Cambridge, MA, 2000. MIT Press.

[5] C.M. Bishop. Variational Principal Components. In *International Conference on Artificial Neural Networks*, pages 509–514, 1999.

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

[7] T. M. Cover and J. A. Thomas. *Elements of Information Theory.* John Wiley, 1991.

[8] H. Lappalainen. Ensemble learning for Independent Component Analysis. In *Proceedings of ICA'99,Aussois, France.*, 1999.

[9] G.E. Hinton and D. van Camp. Keeping neural networks simple by minimizing the description length of the weights. In *Proceedings of the Sixth Annual Conference on Computational Learning Theory*, pages 5–13, 1993.

[10] T.S. Jaakkola and M.I. Jordan. Improving the mean field approximation. In M.I. Jordan, editor, *Learning in Graphical Models.* Kluwer Academic Press, 1998.

[11] D.J.C. Mackay. Ensemble Learning and Evidence Maximization. Technical report, Cavendish Laboratory, University of Cambridge, 1995.

[12] D.J.C. Mackay. Ensemble Learning for Hidden Markov Models. Technical report, Cavendish Laboratory, University of Cambridge, 1998.

[13] T.S. Jaakola M.I. Jordan, Z. Ghahramani and L.K. Saul. An Introduction to Variational Methods for Graphical Models. In M.I. Jordan, editor, *Learning in Graphical Models.* Kluwer Academic Press, 1998.

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