Variational Inference Guide

Neil Lawrence

18th December 2002

Abstract

This report is a brief introduction to variational inference for Bayesian models from the perspective of the Expectation Maximisation (EM) algorithm [1]. We start with an overview of the EM algorithm from the perspective of variational inference and then we show how approximate inference may also be performed. We discuss briefly when variational inference may be used and finally we mentione the variational importance sampler as an alternative approach.

1 Exact Variational Inference

In Bayesian inference we start with a prior distribution, $p(\boldsymbol{\theta})$ over the our parameters, $\boldsymbol{\theta}$, and a likelihood model, $p(\mathbf{X}|\boldsymbol{\theta})$ of some variables $\mathbf{X}$ given the parameters.

We wish to obtain the probability of the data, $p(\mathbf{X})$, which we find through intergrating over $\boldsymbol{\theta}$

$$p(\mathbf{X}) = \int p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}.$$

This result is often called the 'sum rule' of probability, $p(\mathbf{X})$ is then known as the marginalised likelihood of the variables $\mathbf{X}$. Consider the logarithm of this marginalised likelihood[1]

$$\log p(\mathbf{X}) = \log \int p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})\, \partial\boldsymbol{\theta}.$$

Now we introduce an undefined probability distribution $q(\boldsymbol{\theta})$.

$$\log p(\mathbf{X}) = \log \int q(\boldsymbol{\theta})\, \frac{p(\mathbf{X}|\boldsymbol{\theta})\, p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \partial\boldsymbol{\theta}.$$

Jensen's inequality states that $\log\left(\int f(x)\, dx\right) \geq \int \log(f(x))\, dx$, where $f(x)$ is some positive function of x. Here, we implement a modified form of Jensen's inequality, to obtain the following lower bound on the log likelihood

[1] Note that the logarithm is a monotonic function therefore $\log(y_1) > \log(y_2)$if and only if $y_1 > y_2$. We are often interested in comparing likelihoods of variables to see which is more likely than the other. Taking logarithms preserves this ordering.

$$\begin{aligned}
\log p(\mathbf{X}) &\geq \int q(\boldsymbol{\theta}) \log \frac{p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})}{q(\boldsymbol{\theta})} \partial\boldsymbol{\theta} && (1) \\
&\geq \int q(\boldsymbol{\theta}) \log p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \partial\boldsymbol{\theta} \\
&\quad - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \partial\boldsymbol{\theta} && (2)
\end{aligned}$$

The quality of the bound is dependent on the functional form of $q(\boldsymbol{\theta})$ as we will now demonstrate.

1.1 The quality of the bound as a function of $q(\boldsymbol{\theta}_k)$

The product rule of probability gives us

$$p(\mathbf{X}, \boldsymbol{\theta}) = p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta})$$

and

$$p(\mathbf{X}, \boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{X}) p(\mathbf{X})$$

which in turn means[2]

$$p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) = p(\boldsymbol{\theta}|\mathbf{X}) p(\mathbf{X}) \tag{3}$$

We now substitue (3) into (1) to obtain

$$\begin{aligned}
\log p(\mathbf{X}) &\geq \int q(\boldsymbol{\theta}) \log \frac{p(\boldsymbol{\theta}|\mathbf{X}) p(\mathbf{X})}{q(\boldsymbol{\theta})} \partial\boldsymbol{\theta} && (4) \\
&\geq \int q(\boldsymbol{\theta}) \log p(\mathbf{X}) \partial\boldsymbol{\theta} \\
&\quad + \int q(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}|\mathbf{X}) \partial\boldsymbol{\theta} \\
&\quad - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \partial\boldsymbol{\theta} && (5)
\end{aligned}$$

Now we note that the first term in (5) is an expectation of $\log p(\mathbf{X})$ under the distribution $q(\boldsymbol{\theta})$ and is therefore simply equal to $\log p(\mathbf{X})$. Therefore we rewrite (5)

$$\begin{aligned}
\log p(\mathbf{X}) &\geq \log p(\mathbf{X}) + \int q(\boldsymbol{\theta}) \log p(\boldsymbol{\theta}|\mathbf{X}) \partial\boldsymbol{\theta} \\
&\quad - \int q(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \partial\boldsymbol{\theta}
\end{aligned}$$

[2] This is recognised as the foundation of the proof of Bayes's theorem.

which means that the difference between the lower bound obtained through our modified form of Jensen's inequality and the true likelihood is simply

$$\begin{aligned}\mathrm{KL}\left(q\left(\boldsymbol{\theta}\right)||p\left(\boldsymbol{\theta}|\mathbf{X}\right)\right) &= \int q\left(\boldsymbol{\theta}\right)\log q\left(\boldsymbol{\theta}\right)\partial\boldsymbol{\theta} \\ &\quad -\int q\left(\boldsymbol{\theta}\right)\log p\left(\boldsymbol{\theta}|\mathbf{X}\right)\partial\boldsymbol{\theta}\end{aligned}$$

which is known as the Kullback-Leibler (KL) divergence between the two distributions [5]. This divergence is always positive unless $q\left(\boldsymbol{\theta}\right) = p\left(\boldsymbol{\theta}|\mathbf{X}\right)$ when it is zero. When the divergence is equal to zero, the lower bound on the likelihood above becomes an equality. In other words if we take $q\left(\boldsymbol{\theta}\right) = p\left(\boldsymbol{\theta}|\mathbf{X}\right)$ then our bound becomes an equality. This process is known as a *free form optimisation* of the bound with respect to the distribution $q\left(\boldsymbol{\theta}\right)$.

2 Approximate Variational Inference

In the outline above, we allowed the distribution $q\left(\boldsymbol{\theta}\right)$ to have any functional form. We were, thereby, able to recover the marginalised likelihood. The algorithm outlined above is the expectation step of the Expectation-Maximisation algorithm[3] [1]. In practice, determining $p\left(\boldsymbol{\theta}|\mathbf{X}\right)$ may be a problem, more precisely, if we are unable to obtain the integral in (1) then, because the posterior distribution is given by

$$p\left(\boldsymbol{\theta}|\mathbf{X}\right) = \frac{p\left(\mathbf{X},\boldsymbol{\theta}\right)}{p\left(\mathbf{X}\right)},$$

we will be unable to compute the posterior distribuiton.

In these circumstances, to make progress, we must place further constraints on the functional form of the distribution. In variational inference [4], the option we consider is to assume that the 'q-distribution' factorises across *disjoint* subsets of the parameters:

$$q\left(\boldsymbol{\theta}\right) = q\left(\boldsymbol{\theta}_1\right)q\left(\boldsymbol{\theta}_2\right),$$

where $\boldsymbol{\theta}_1$ and $\boldsymbol{\theta}_2$ are disjoint sub-sets of the full parameter set $\boldsymbol{\theta}$. Substituting these distributions into (5) we recover

$$\begin{aligned}\log p\left(\mathbf{X}\right) &\geq \int q\left(\boldsymbol{\theta}_1\right)q\left(\boldsymbol{\theta}_2\right)\log\frac{p\left(\mathbf{X}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)p\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)}{q\left(\boldsymbol{\theta}_1\right)q\left(\boldsymbol{\theta}_2\right)}\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_2 \\ &\geq \int q\left(\boldsymbol{\theta}_1\right)q\left(\boldsymbol{\theta}_2\right)\log p\left(\mathbf{X}|\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)p\left(\boldsymbol{\theta}_1,\boldsymbol{\theta}_2\right)\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_2 \\ &\quad -\int q\left(\boldsymbol{\theta}_1\right)q\left(\boldsymbol{\theta}_2\right)\log q\left(\boldsymbol{\theta}_1\right)q\left(\boldsymbol{\theta}_2\right)\partial\boldsymbol{\theta}_1\partial\boldsymbol{\theta}_2\end{aligned}$$

Consider the dependence on one of these factors $q\left(\boldsymbol{\theta}_1\right)$

[3] The maximisation step, which is not of interest here, consists of maximising (1) with respect to a further set of parameters which we are not considering.

$$
\begin{aligned}
\log p(\mathbf{X}) &\geq \int q(\boldsymbol{\theta}_1) q(\boldsymbol{\theta}_2) \log \frac{p(\mathbf{X}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2)}{q(\boldsymbol{\theta}_1) q(\boldsymbol{\theta}_2)} \partial\boldsymbol{\theta}_1 \partial\boldsymbol{\theta}_2 \\
&\geq \int q(\boldsymbol{\theta}_1) \int q(\boldsymbol{\theta}_2) \log p(\mathbf{X}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \partial\boldsymbol{\theta}_1 \partial\boldsymbol{\theta}_2 \\
&\quad - \int q(\boldsymbol{\theta}_1) \log q(\boldsymbol{\theta}_1) \partial\boldsymbol{\theta}_1 + \text{const} \\
&\geq \int q(\boldsymbol{\theta}_1) \log \left[\exp \left\langle \log p(\mathbf{X}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right\rangle_{q(\boldsymbol{\theta}_2)} \right] \partial\boldsymbol{\theta}_1 \partial\boldsymbol{\theta}_2 \\
&\quad - \int q(\boldsymbol{\theta}_1) \log q(\boldsymbol{\theta}_1) \partial\boldsymbol{\theta}_1 + \text{const} \\
&\geq \text{const} \; - \text{KL} \left(q(\boldsymbol{\theta}_1) \| \frac{\exp \left\langle \log p(\mathbf{X}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right\rangle_{q(\boldsymbol{\theta}_2)}}{Z} \right).
\end{aligned}
$$

where $Z = \int \exp \left\langle \log p(\mathbf{X}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right\rangle_{q(\boldsymbol{\theta}_2)} \partial\boldsymbol{\theta}_1$ and const is a term constant in $\boldsymbol{\theta}_2$ and we have introduced the notation $\langle \cdot \rangle_{p(\cdot)}$ to denote an expectation under the distribution $p(\cdot)$. This KL-divergence is minimised when

$$
q(\boldsymbol{\theta}_1) \propto \exp \left\langle \log p(\mathbf{X}|\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) p(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) \right\rangle_{q(\boldsymbol{\theta}_2)},
$$

where the constant of proportionality is given by Z.

A similar equation may be derived for $q(\boldsymbol{\theta}_2)$. Indeed, in general, for J subsets of the parameters $\boldsymbol{\theta}_j$, if we assume that the posterior approximation factorises

$$
q(\boldsymbol{\theta}) = \prod_{j=1}^{J} q(\boldsymbol{\theta}_j)
$$

we may obtain through a similar analysis to the above

$$
q(\boldsymbol{\theta}_k) \propto \exp \left\langle \log p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \right\rangle_{\prod_{j \neq k} q(\boldsymbol{\theta}_j)}. \tag{6}
$$

3 When May we Perform Variational Inference?

We are still somewhat restricted in our implementation of variational inference in that to compute the distribution over $\boldsymbol{\theta}_k$ specified in (6) we are required to obtain the constant of proportionality which is given by

$$
Z = \int \exp \left\langle \log p(\mathbf{X}|\boldsymbol{\theta}) p(\boldsymbol{\theta}) \right\rangle_{\prod_{j \neq k} q(\boldsymbol{\theta}_j)} \partial\boldsymbol{\theta}_k. \tag{7}
$$

In [3] Ghahramani and Beal explore a class of distributions for which this integral will be soluble known as the conjugate-exponential family.

3.1 What do we do if we cannot compute the integral

In the case that the integral is not soluble, one approach is to seek an alternative functional form for $q(\boldsymbol{\theta}_k)$, $q'(\boldsymbol{\theta}_k)$ for which the integral will be tractable and attempt to minimise the following KL-divergence

$$\begin{aligned} \mathrm{KL}\left(q'(\boldsymbol{\theta}) \,||\, q(\boldsymbol{\theta})\right) &= \int q'(\boldsymbol{\theta}) \log q'(\boldsymbol{\theta}) \, \partial\boldsymbol{\theta} \\ &\quad - \int q'(\boldsymbol{\theta}) \log q(\boldsymbol{\theta}) \, \partial\boldsymbol{\theta} \end{aligned}$$

however it is not obvious what this functional form should be, nor is it clear when this step may be performed. The alternative we propose is, whilst we may not be able to compute the integral in Z we may be able to obtain samples from (6) using Monte Carlo techniques and estimate expectations of interest using the sampler. One example of this is the variational importance sampler.

4 The Variational Importance Sampler

The variational importance sampler is a general solution to handling an intractability in (6). First, note from (6) that each distribution $q(\boldsymbol{\theta}_k)$ is dependent on expectations under the other factors of our posterior approximations. In other words, if we can obtain these expectations, or estimates of these expectations, for the distribution which is intractable we may still estimate the other distributions. As its name suggests, the variational importance sampler obtains these estimates of the expectations under the intractable distributions via importance sampling[4].

Importance sampling has been mentioned in the context of variational inference by other authors, in particular Ghahramani and Beal [2], but as a method of estimating the true marginalised log likelihood, not as a method of handling intractable variational approximations.

References

[1] A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, B*, 39(1):1–38, 1977.

[2] Z. Ghahramani and M. J. Beal. Graphical models and variational methods. In M. Opper and D. Saad, editors, *Advanced Mean Field Methods — Theory and Practice*. MIT Press, 2001.

[3] Z. Ghahramani and M. J. Beal. Propagation algorithms for variational Bayesian learning. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors,

[4] It is possible to envisage various other methods of sampling from the intractable distribution to obtain the same expectations.

Advances in Neural Information Processing Systems, volume 13, Cambridge, MA, 2001. MIT Press.

[4] M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. In M. I. Jordan, editor, *Learning in Graphical Models*, volume 89 of *Series D: Behavioural and Social Sciences*, pages 105–162, Dordrecht, The Netherlands, 1998. Kluwer.

[5] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22:79–86, 1951.