Machine learning: Advanced Probabilistic Methods (2019): Derivation of the ELBO for the simple model. P. Marttinen

Recall that variational inference is based on the decomposition

$$\log p(x) = \mathcal{L}(q) + KL(q|p),$$

where $q(\mathbf{Z})$ is any approximation to the posterior distribution $p(\mathbf{Z}|\mathbf{X})$ of unobserved variables $\mathbf{Z}$ in the model, given observed variables $\mathbf{X}$. The goal of the variational inference algorithm is to maximize the evidence lower bound (ELBO) $\mathcal{L}(q)$, or equivalently minimize the KL-divergence $KL(q|p)$ between the approximation and the true posterior. Here we show how to compute the ELBO for the 'simple model' derived earlier[1]. Briefly, the model is

$$p(x_n|\theta,\tau) = (1-\tau)N(x_n|0,1) + \tau N(x_n|\theta,1), \quad n = 1,\ldots,N.$$

The latent variable representation is given by

$$p(\mathbf{x}|\mathbf{z},\theta) = \prod_{n=1}^{N} N(x_n|0,1)^{z_{n1}} N(x_n|\theta,1)^{z_{n2}}, \tag{1}$$

and

$$p(\mathbf{z}|\tau) = \prod_{n=1}^{N} \tau^{z_{n2}}(1-\tau)^{z_{n1}}. \tag{2}$$

Priors are specified as follows:

$$p(\tau) = Beta(\tau|\alpha_0,\alpha_0) \propto \tau^{\alpha_0-1}(1-\tau)^{\alpha_0-1}$$
$$p(\theta) = N(\theta|0,\beta_0^{-1}) \propto \exp\left(-\frac{\beta_0}{2}\theta^2\right).$$

The logarithm of the joint distribution can be written as:

$$\log p(\mathbf{x},\mathbf{z},\tau,\theta) = \log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z},\theta). \tag{3}$$

We assume the mean-field approximation

$$p(\mathbf{z},\tau,\theta|\mathbf{x}) \approx q(\tau)q(\theta)\textstyle\prod_n q(z_n). \tag{4}$$

Assume that currently we have factors

$$q(z_n|r_{n1},r_{n2}) = Categorical(z_n|r_{n1},r_{n2}) = r_{n1}^{z_{n1}} r_{n2}^{z_{n2}}, \tag{5}$$
$$q(\tau) = Beta(\tau|\alpha_\tau,\beta_\tau), \tag{6}$$
$$q(\theta) = N(\theta|m_2,\beta_2^{-1}), \tag{7}$$

where r_{n1}, r_{n2}, $n = 1,\ldots,N$, α_τ, β_τ, m_2, and β_2 are so-called *variational parameters*, i.e. parameters that specify the exact distribution of the factor. Previously, in document *simple_vb_example.pdf*, we derived formulas for updating the the variational parameters of any factor conditionally on the other factors.

The general formula for the ELBO is given by

$$\begin{aligned}\mathcal{L}(q) &= \int q(\mathbf{Z}) \log \frac{p(\mathbf{X},\mathbf{Z})}{q(\mathbf{Z})} d\mathbf{Z} \\ &= E_q[\log p(\mathbf{X},\mathbf{Z})] - E_q[\log q(\mathbf{Z})],\end{aligned} \tag{8}$$

where $\mathbf{Z}$ is a generic notation that includes all unobservables. It will be left as an **exercise** to show that the ELBO can be written as:

$$\begin{aligned}\mathcal{L}(q) = {} & E_{q(\tau)}[\log p(\tau)] + E_{q(\theta)}[\log p(\theta)] + E_{q(\mathbf{z})q(\tau)}[\log p(\mathbf{z}|\tau)] \\ & + E_{q(\mathbf{z})q(\theta)}[\log p(\mathbf{x}|\mathbf{z},\theta)] - E_{q(\mathbf{z})}[\log q(\mathbf{z})] - E_{q(\tau)}[\log q(\tau)] \\ & - E_{q(\theta)}[\log q(\theta)].\end{aligned} \tag{9}$$

[1]The derivation of the ELBO for the general GMM case can be found in Bishop's book, Section 10.2.2.

When conjugate priors are used, as is the case with the simple model, all seven terms in formula (9) can be computed analytically. Below we consider each of these terms in turn. **The ELBO can then be computed simply by plugging each of the derived terms into Equation (9).** In these derivations we will occasionally discard some terms that do not depend on the variational parameters, as our purpose of deriving the ELBO is to monitor the convergence of the VB algorithm and those terms are constant across the iterations.

1st term in (9):

$$\begin{aligned} E_{q(\tau)}[\log p(\tau)] &= E_{q(\tau)}\left[(\alpha_0 - 1)\log\tau + (\alpha_0 - 1)\log(1-\tau)\right] \\ &= (\alpha_0 - 1)E_{q(\tau)}[\log\tau] + (\alpha_0 - 1)E_{q(\tau)}[\log(1-\tau)] \\ &= (\alpha_0 - 1)[\psi(\alpha_\tau) - \psi(\alpha_t + \beta_\tau)] + (\alpha_0 - 1)[\psi(\beta_\tau) - \psi(\alpha_t + \beta_\tau)]. \end{aligned}$$

The last line above followed from the known formula for $E_{q(\tau)}[\log\tau]$ when $q(\tau)$ has the Beta distribution specified in Equation (6).

2nd term in (9):

$$\begin{aligned} E_{q(\theta)}\left[\log p(\theta)\right] &= \ldots(\textbf{exercise}) \\ &= -\frac{\beta_0}{2}\left(\beta_2^{-1} + m_2^2\right). \end{aligned}$$

The last line followed directly from the normal distribution (7) for θ.

3rd term in (9):

$$\begin{aligned} E_{q(\mathbf{z})q(\tau)}\left[\log p(\mathbf{z}|\tau)\right] &= \sum_{n=1}^{N} E_{q(z_n)q(\tau)}[\log p(z_n|\tau)] \\ &= \sum_{n=1}^{N} E_{q(z_n)q(\tau)}[z_{n2}\log\tau + z_{n1}\log(1-\tau)] \\ &= \sum_{n=1}^{N} \left\{E_{q(z_n)}[z_{n2}]E_{q(\tau)}[\log\tau] + E_{q(z_n)}[z_{n1}]E_{q(\tau)}[\log(1-\tau)]\right\} \\ &= \sum_{n=1}^{N} \left\{r_{n2}[\psi(\alpha_\tau) - \psi(\alpha_t + \beta_\tau)] + r_{n1}[\psi(\beta_\tau) - \psi(\alpha_\tau + \beta_\tau)]\right\}. \end{aligned}$$

4th term in (9):

$$\begin{aligned} E_{q(\mathbf{z})q(\theta)}\left[\log p(\mathbf{x}|\mathbf{z},\theta)\right] &= \ldots(\textbf{exercise}) \\ &= -\frac{N}{2}\log(2\pi) - \frac{1}{2}\sum_{n=1}^{N} r_{n1}x_n^2 - \frac{1}{2}\sum_{n=1}^{N} r_{n2}\left\{(x_n - m_2)^2 + \beta_2^{-1}]\right\}. \end{aligned}$$

5th term in (9):

$$\begin{aligned} E_{q(\mathbf{z})}[\log q(\mathbf{z})] &= \sum_{n=1}^{N} E_{q(z_n)}\left[z_{n1}\log r_{n1} + z_{n2}\log r_{n2}\right] \\ &= \sum_{n=1}^{N} r_{n1}\log r_{n1} + r_{n2}\log r_{n2}\ . \end{aligned}$$

6th term in (9):

$$E_{q(\tau)}\left[\log q(\tau)\right] = \log \frac{\Gamma(\alpha_\tau + \beta_\tau)}{\Gamma(\alpha_\tau)\Gamma(\beta_\tau)} + (\alpha_\tau - 1)\psi(\alpha_\tau) + (\beta_\tau - 1)\psi(\beta_\tau) - (\alpha_\tau + \beta_\tau - 2)\psi(\alpha_\tau + \beta_\tau).$$

This is just the negative entropy of the $Beta(\alpha_\tau, \beta_\tau)$ distribution (see Wikipedia).

7th term in (9):

$$E_{q(\theta)}[\log q(\theta)] = \ldots (\textbf{exercise})$$