1 The Evidence Lower Bound

We are interested in finding the distribution $q(\theta)$ that best approximates the true posterior $p(\theta|y)$, with the best approximation taken as the $q(\theta)$ that minimises the Kullback-Leibler divergence with the posterior, defined as

$$KL[q(\theta)|p(\theta|y)] = \int q(\theta) \ln\left(\frac{q(\theta)}{p(\theta|y)}\right) d\theta.$$

Beginning with the unknown constant $p(y)$, we have that

$$\begin{aligned}
\ln(p(y)) &= \int_\theta q(\theta) \ln(p(y)) d\theta \\
&= \int_\theta q(\theta) \ln\left(\frac{p(y,\theta)}{p(\theta|y)}\right) d\theta \\
&= \int_\theta q(\theta) \ln\left(\frac{p(y,\theta)}{p(\theta|y)}\frac{q(\theta)}{q(\theta)}\right) d\theta \\
&= \int_\theta q(\theta) \ln\left(\frac{q(\theta)}{p(\theta|y)}\right) + \ln\left(\frac{p(y,\theta)}{q(\theta)}\right) d\theta \\
&= \int_\theta q(\theta) \ln\left(\frac{q(\theta)}{p(\theta|y)}\right) d\theta + \int_\theta q(\theta) \ln\left(\frac{p(y,\theta)}{q(\theta)}\right) d\theta \\
&= KL[q(\theta)||p(\theta|y)] + F(q,y)
\end{aligned}$$

where

$$\begin{aligned}
F(q,y) &= \int_\theta q(\theta) \ln\left(\frac{p(y,\theta)}{q(\theta)}\right) d\theta \\
&= \int_\theta q(\theta) \ln(p(y,\theta)) d\theta - \int_\theta q(\theta) \ln(q(\theta)) d\theta.
\end{aligned}$$

As $\ln(p(y))$ is constant, the distribution $q(\theta)$ that minimises the KL divergence can be found by maximising $F(q,y)$, which is called the Evidence Lower Bound (ELBO, or $\mathcal{L}(q,y)$) in the machine learning literature. We are yet to make any assumptions about the distribution of $q(\theta)$.

2 The Mean Field Assumption

The mean-field assumption is that $q(\theta)$ can be factorised as $q(\theta) = \prod_i q_i(\theta_i)$, where each i may be a scalar or vector. The mean-field derivations follows, assuming that $q(\theta) = q_1(\theta_1)q_2(\theta_2)$.

$$
\begin{aligned}
F(q,y) &= \int_\theta q_1 q_2 \ln(p(y,\theta))d\theta - \int_\theta q_1 q_2 \ln(q_1 q_2)d\theta \\
&= \int_\theta q_1 q_2 (\ln(p(y,\theta)) - \ln(q_1))d\theta - \int_\theta q_1 q_2 \ln(q_2)d\theta \qquad (1)\\
&= \int_{\theta_1} q_1 \left(\int_{\theta_2} q_2 \ln(p(y,\theta))d\theta_2 - \ln(q_1)\right) d\theta_1 - \int_{\theta_1} q_1 \int_{\theta_2} q_2 \ln(q_2) d\theta_2 d\theta_1 \\
&= \int_{\theta_1} q_1 \ln\left(\frac{\exp(\mathbb{E}_{q_2}[\ln(p(y,\theta))])}{q_1}\right) d\theta_1 + c \qquad (2)\\
&= -KL(q_1 || \exp(\mathbb{E}_{q_2}[\ln(p(y,\theta))])) + c
\end{aligned}
$$

(1) uses the fact that

$$\int_{\theta_2} q_2 \ln(q_1) d\theta_2 = \ln(q_1),$$

and (2) uses the fact that

$$\int_{\theta_1} q_1 \int_{\theta_2} q_2 \ln(q_2) d\theta_2 d\theta_1 = \mathbb{E}_{q_1}\left[\int_{\theta_2} q_2 \ln(q_2) d\theta_2\right] = \int_{\theta_2} q_2 \ln(q_2) d\theta_2,$$

a constant term with respect to q_1.

Maximisation with respect to q_1 is simple due to the KL divergence term, as we can minimise

$$KL(q_1 || \exp(E_{q_2}(\ln(p(y,\theta)))))$$

by setting

$$q_1 \propto \exp(E_{q_2}(\ln(p(y,\theta))))$$

for every i in θ.

3 Copula Variational Bayes

If we attach a copula function to our approximation, so that $q(\theta) = q_1(\theta_1)q_2(\theta_2)c(Q_1(\theta_1), Q_2(\theta_2)$, we get the Evidence Lower Bound of

$$\begin{aligned}
F(q,y) &= \int_\theta q_1 q_2 c(Q_1,Q_2)\ln(p(y,\theta))d\theta - \int_\theta q_1 q_2 c(Q_1,Q_2)\ln(q_1 q_2 c(Q_1,Q_2))d\theta \\
&= \int_{\theta_1} q_1 c(Q_1,Q_2)\left(\int_{\theta_2} q_2 \ln(p(y,\theta))d\theta_2 - \ln(q_1)\right)d\theta_1 - \int_\theta q_1 q_2 c(Q_1,Q_2)\ln(q_2 c(Q_1,Q_2))d\theta \\
&= \int_{\theta_1} q_1 c(Q_1,Q_2)\ln\left(\frac{\exp(\mathbb{E}_{q_2}[\ln(p(y,\theta))])}{q_1}\right)d\theta_1 + f(q_1,q_2).
\end{aligned}$$

This derivation runs into two problems: the first term is no longer a KL divergence, and the second term is no longer constant with respect to q_1. If, between the first and second lines we moved $q_1 q_2 c(Q_1,Q_2)\ln(q_1 c(Q_1,Q_2))$ from the second term to the first, instead of $q_1 q_2 c(Q_1,Q_2)\ln(q_1)$ we could attempt to the match the denominator with the $q_1 c(Q_1,Q_2)$ term and make something in the form of a KL divergence.

We would have had

$$\begin{aligned}
F(q,y) &= \int_{\theta_1} q_1 c(Q_1,Q_2)\left(\int_{\theta_2} q_2 \ln(p(y,\theta))d\theta_2 - \int_{\theta_2} q_2 \ln(q_1 c(Q_1,Q_2))d\theta_2\right)d\theta_1 \\
&- \int_\theta q_1 q_2 c(Q_1,Q_2)\ln(q_2)d\theta
\end{aligned}$$

and still would not get the KL divergence term as the result used in (1) would no longer hold as

$$\int_{\theta_2} q_2 \ln(q_1 c(Q_1,Q_2))d\theta_2 \neq \ln(q_1 c(Q_1,Q_2)).$$

Even if we could do this, the second term would be

$$\begin{aligned}
\int_\theta q_1 q_2 c(Q_1,Q_2)\ln(q_2)d\theta &= \int_{\theta_1} q_1 \int_{\theta_2} q_2 c(Q_1,Q_2)\ln(q_2)d\theta_2 d\theta_1 \\
&= \mathbb{E}_{q_1}\left[\int_{\theta_2} q_2 c(Q_1,Q_2)\ln(q_2)d\theta_2\right]
\end{aligned}$$

which is not constant with respect to q_1.

Clearly the mean field approach to the derivation can not be directly adapted.