COS513: VARIATIONAL INFERENCE CONTINUED

JIA DENG AND MATT HOFFMAN

We want to infer the posterior distribution of our hidden variables $z_{1:m}$ conditioned on our observed variables $x_{1:n}$. Last time we saw that we can define a variational distribution q over our hidden parameters z, and that no matter what we choose for q the following lower bound holds (due to Jensen's inequality):

$$
\begin{aligned}
\log p(x) &= \log \int p(z)p(x|z)dz \\
&= \log \int \frac{p(z)p(x|z)q(z)}{q(z)}dz \\
&\geq \int q(z)\log p(x,z)dz - \int q(z)\log q(z)dz \\
&= \mathrm{E}_q[\log p(x,z)] - \mathrm{E}_q[\log q(z)] \qquad (1)
\end{aligned}
$$

So we can lower bound the log-likelihood of the observed data under our model by choosing some variational distribution q. It turns out that tightening this lower bound (i.e. maximizing the right side of equation 1) is equivalent to minimizing the Kullback-Leibler (KL) divergence between $q(z)$ and $p(z|x)$. This can be seen easily (after a little algebra) from the definition of KL divergence:

$$
\begin{aligned}
\mathrm{KL}(q(z)||p(z|x)) &\triangleq \int q(z)\log\frac{q(z)}{p(z|x)}dz = \mathrm{E}_q\left[\log\frac{q(z)}{p(z|x)}\right] \\
&= \mathrm{E}_q[\log q(z)] - \mathrm{E}_q[\log p(z|x)] \\
&= \mathrm{E}_q[\log q(z)] - \mathrm{E}_q\left[\log\frac{p(x,z)}{p(x)}\right] \\
&= \mathrm{E}_q[\log q(z)] - \mathrm{E}_q[\log p(x,z)] - \mathrm{E}_q[\log p(x)] \qquad (2)
\end{aligned}
$$

The third term is constant with respect to q (since $\mathrm{E}_q[\log p(x)] = \log p(x)$), and the first two terms are just the right side of equation 1 negated, so minimizing $\mathrm{KL}(q(z)||p(z|x))$ with respect to q is equivalent to maximizing the lower bound in equation 1 with respect to q.

We want to choose a form for q that is reasonably powerful (so that we can get a reasonable approximation to $p(z|x)$), but also easy to work with (so that we can actually compute the expectations in equation 1). A popular approach is to use a fully factorized form for q:

$$q(z|\nu) = q(z_1|\nu_1)q(z_2|\nu_2)\dots q(z_m|\nu_m). \tag{3}$$

If $q(z_i|\nu_i)$ is in the exponential family, then this becomes

$$q(z_i|\nu_i) = h(z_i)\exp\{\nu_i^T z_i - a(\nu_i)\}. \tag{4}$$

This form will be useful later, especially if $q(z_i)$ is of the same form as $p(z_i|z_{-i}, x)$.

We want to maximize our objective function

$$\mathcal{L} = \mathrm{E}_q \log p(z_{1:m}, x_{1:n}) - \mathrm{E}_q[\log q(z_{1:m})]. \tag{5}$$

By the chain rule, this becomes:

$$\mathcal{L} = \log p(x_{1:n}) + \sum_{i=1:m} \mathrm{E}_q[\log p(z_i|z_{1:i-1}, x_{1:n})] - \mathrm{E}_q[\log q(z_{1:m})]. \tag{6}$$

Note that we can move the expectations inside of the summations because we have chosen q to be fully factorized.

We will do coordinate ascent over each ν_i on the objective function. We can put whichever z_i we're working on at the end of the sum in equation 6 because the chain rule works regardless of order. Doing so, we define

$$l_i = \mathrm{E}_q[\log p(z_i|z_{-i}, x_{1:n})] - \mathrm{E}_q[\log q(z_i|\nu_i)]. \tag{7}$$

Since l_i is the only part of $\mathcal{L}$ that depends on z_i (once we've reordered the sum in equation 6), we only need to optimize l_i when updating ν_i.

Assuming that q is in the exponential family, we have

$$\begin{aligned} l_i &= \mathrm{E}_q\left[\log p(z_i|z_{-i}, x_{1:n})\right] - \mathrm{E}_q[\log h(z_i) + \nu_i^T z_i - a(\nu_i)] \\ &= \mathrm{E}_q\left[\log p(z_i|z_{-i}, x_{1:n})\right] - \left(\mathrm{E}_q[\log h(z_i)] + \nu_i^T a'(\nu_i) - a(\nu_i)\right). \end{aligned}$$

This holds because for all exponential family distributions $q(z_i|\nu_i)$ the expectation of the random variable z_i is the first derivative of the log normalizer term $a(\nu_i)$.

Take the derivative of l_i with respect to ν_i,

$$\frac{\partial l_i}{\partial \nu_i} = \frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log p(z_i|z_{-i}, x_{1:n})\right] - \frac{\partial}{\partial \nu_i}\mathrm{E}_q[\log h(z_i)] - \nu_i^T a''(\nu_i). \tag{8}$$

Set the above equation to zero:

$$\nu_i = a''(\nu_i)^{-1}\left(\frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log p(z_i|z_{-i}, x_{1:n})\right] - \frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log h(z_i)\right]\right) \tag{9}$$

We assume the conditionals $p(z_i|z_{-i}, x_{1:n})$ are in the exponential family. Moreover, we assume they are in the same exponential family as $q(z_i|\nu_i)$, that is,

$$p(z_i|z_{-i}, x_{1:n}) = h(z_i)\exp\{g_i(z_{-i}, x_{1:n})^T z_i - a(g_i(z_{-i}, x_{1:n}))\}. \tag{10}$$

$g_i(z_{-i}, x_{1:n})$ is the natural parameter to the (exponential family) posterior distribution over z_i.

Therefore,

$$\mathrm{E}_q\left[\log p(z_i|z_{-i}, x_{1:n})\right] = \mathrm{E}_q\left[\log h(z_i)\right] + \mathrm{E}_q\left[g_i(z_{-i}, x_{1:n})^T z_i\right] - \mathrm{E}_q\left[a(g_i(z_{-i}, x_{1:n}))\right]. \tag{11}$$

We observe two facts: (1) g_i doesn't depend on z_i; (2) g_i is independent with z_i. Therefore,

$$\mathrm{E}_q\left[g_i(z_{-1}, x_{1:n})^T z_i\right] = \mathrm{E}_q\left[g_i(z_{-i}, x_{1:n})^T\right] a'(\nu_i). \tag{12}$$

It follows that

$$\frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log p(z_i|z_{-i}, x_{1:n})\right] = \frac{\partial}{\partial \nu_i}\mathrm{E}_q\left[\log h(z_i)\right] + \mathrm{E}_q\left[g_i(z_{-i}, x_{1:n})\right]^T a''(\nu_i). \tag{13}$$

Substitue above to equation 9. we have

$$\nu_i = \mathrm{E}_q\left[g_i(z_{-i}, x_{1:n})^T\right] \tag{14}$$

Thus we have obtained the update equation for each iteration—we simply set ν_i to be the expectation under q of the natural parameter of the posterior distribution of $z_i|z_{-i}, x_{1:n}$. (The updates do not typically have such a simple form in a non-conjugate setting.)

It is instructive to compare these updates with Gibbs sampling. In Gibbs sampling, we sampled from the conditional distribution $p(z_i|z_{-i}, x_{1:n})$, whereas in mean-field variational inference we just set ν_i equal to the conditional expectation of the natural parameter of $p(z_i|z_{-i}, x_{1:n})$ under q. A crucial difference is that in Gibbs sampling we set the hidden variables z_i to specific values, whereas in variational inference we only consider *distributions* over them.