

# A Mean Field Example

Noah A. Smith

Consider the following generative model.

1. Draw multinomial distribution  $q$  from a Dirichlet distribution parameterized by  $\alpha$ .
2. Draw  $y$  from  $q$ .
3. Draw  $x$  from distribution  $p(\cdot | y)$ .

This is a kind of Bayesian mixture model where the mixture components  $p$  are all known, but the mixture coefficients ( $q$ ) are treated as a random variable. (If we wanted to *estimate*  $q$  using EM, with or without specifying a prior over  $q$ , you should be able to derive the EM algorithm for this model.)

Suppose we have a collection of training examples  $\langle x_1, \dots, x_n \rangle$ , and we know  $\alpha$  and  $p$ . We wish to derive a variational inference algorithm for this model that will give us a distribution over each  $y_i$ . (It will also give us a distribution over each  $q_i$ .) To keep things simple, suppose  $n = 1$ , so there's only one observable  $x$ . Our full model says:

$$P(q, y, x | \alpha, p) = \text{Dir}(q | \alpha) \times q(y) \times p(x | y) \quad (1)$$

where  $x$ ,  $\alpha$ , and  $p$  are observed and  $q$  and  $y$  are hidden. Variational inference tells us that, for any model  $Q$  over the hidden variables  $q$  and  $y$ :

$$\log P(x | \alpha, p) \geq \mathbb{E}_Q[\log P(q, y, x | \alpha, p) - \log Q(q, y)] \quad (2)$$

The idea is to choose the parameters of  $Q$  to maximize the right-hand side above (called the variational bound). The mean-field approximation factors  $Q$  over each variable separately (so defining it will depend on how you carve up your random variables!). Here, we let

$$Q(q, y) = Q(q) \times Q(y) = \text{Dir}(q | \tilde{\alpha}) \times \tilde{q}(y) \quad (3)$$

That is, the distribution  $q$  is assumed to be drawn from a Dirichlet with parameters  $\tilde{\alpha}$  and  $y$  is assumed to be drawn from a multinomial  $\tilde{q}$ .

The rest is mechanical but error-prone.

**Step 1** Write out the variational bound for this model by plugging in  $P$  and  $Q$ :

$$\mathbb{E}_Q[\log P(q, y, x | \alpha, p) - \log Q(q, y)] \quad (4)$$

$$= \int \sum_y \text{Dir}(q | \tilde{\alpha}) \tilde{q}(y) (\log (\text{Dir}(q | \alpha) q(y) p(x | y)) - \log (\text{Dir}(q | \tilde{\alpha}) \tilde{q}(y))) dq \quad (5)$$

$$= \int \sum_y \text{Dir}(q | \tilde{\alpha}) \tilde{q}(y) (\log \text{Dir}(q | \alpha) + \log q(y) + \log p(x | y) - \log \text{Dir}(q | \tilde{\alpha}) - \log \tilde{q}(y)) dq \quad (6)$$

**Step 2** Note that expectations collapse for some of the log-terms:

$$\begin{aligned}
&= \int \text{Dir}(q \mid \tilde{\alpha}) \log \text{Dir}(q \mid \alpha) dq \\
&\quad - \int \text{Dir}(q \mid \tilde{\alpha}) \log \text{Dir}(q \mid \tilde{\alpha}) dq \\
&\quad + \int \sum_y \text{Dir}(q \mid \tilde{\alpha}) \tilde{q}(y) \log q(y) dq \\
&\quad + \sum_y \tilde{q}(y) (\log p(x \mid y) - \log \tilde{q}(y))
\end{aligned} \tag{7}$$

**Step 3** Solve each term to get rid of integrals. This requires you to remember the form of the Dirichlet:

$$\text{Dir}(r_1^m \mid \delta_1^m) = \frac{\prod_{i=1}^m \Gamma(\delta_i) r_i^{\delta_i - 1}}{\Gamma(\sum_{i=1}^m \delta_i)} \tag{8}$$

Plugging this in for the log Dir terms and rearranging slightly in term 3, our objective becomes

$$\begin{aligned}
&= - \sum_y \log \Gamma(\alpha_y) + \log \Gamma(\sum_y \alpha_y) + \sum_y (\alpha_y - 1) \int \text{Dir}(q \mid \tilde{\alpha}) \log q(y) dq \\
&\quad + \sum_y \log \Gamma(\tilde{\alpha}_y) - \log \Gamma(\sum_y \tilde{\alpha}_y) - \sum_y (\tilde{\alpha}_y - 1) \int \text{Dir}(q \mid \tilde{\alpha}) \log q(y) dq \\
&\quad + \sum_y \tilde{q}(y) \int \text{Dir}(q \mid \tilde{\alpha}) \log q(y) dq \\
&\quad + \sum_y \tilde{q}(y) (\log p(x \mid y) - \log \tilde{q}(y))
\end{aligned} \tag{9}$$

$$\begin{aligned}
&= \sum_y (\alpha_y - \tilde{\alpha}_y + \tilde{q}(y)) \int \text{Dir}(q \mid \tilde{\alpha}) \log q(y) dq \\
&\quad + \sum_y \log \Gamma(\tilde{\alpha}_y) - \log \Gamma(\sum_y \tilde{\alpha}_y) + \sum_y \tilde{q}(y) (\log p(x \mid y) - \log \tilde{q}(y)) + \text{const}(\tilde{\alpha}, \tilde{q})
\end{aligned} \tag{10}$$

There is a convenient trick that lets us eliminate the last integral:

$$\mathbb{E}_{\text{Dir}(r_1^m \mid \delta_1^m)}[\log r_i] = \Psi(\delta_i) - \Psi(\sum_{i=1}^m \delta_i) \tag{11}$$

This gives the following form to the variational bound:

$$\begin{aligned}
&= \sum_y (\alpha_y - \tilde{\alpha}_y + \tilde{q}(y)) (\Psi(\tilde{\alpha}_y) - \Psi(\sum_{y'} \tilde{\alpha}_{y'})) \\
&\quad + \sum_y \log \Gamma(\tilde{\alpha}_y) - \log \Gamma(\sum_y \tilde{\alpha}_y) + \sum_y \tilde{q}(y) (\log p(x \mid y) - \log \tilde{q}(y)) + \text{const}(\tilde{\alpha}, \tilde{q})
\end{aligned} \tag{12}$$

**Step 4** Optimizing the bound is usually done through coordinate ascent. For each variational parameter (here,  $\forall y$ ,  $\tilde{q}(y)$  and  $\tilde{\alpha}(y)$ ), we write the bound as a function solely of that variable:

$$B(\tilde{q}(y)) = \tilde{q}(y)(\Psi(\tilde{\alpha}_y) - \Psi(\sum_{y'} \tilde{\alpha}_{y'}) + \log p(x | y) - \log \tilde{q}(y)) + \lambda \left(1 - \sum_{y'} \tilde{q}(y')\right) \quad (13)$$

$$B(\tilde{\alpha}_y) = (\alpha_y - \tilde{\alpha}_y + \tilde{q}(y))(\Psi(\tilde{\alpha}_y) - \Psi(\sum_{y'} \tilde{\alpha}_{y'})) + \log \Gamma(\tilde{\alpha}_y) - \log \Gamma(\sum_y \tilde{\alpha}_y) \quad (14)$$

Note that we use a Lagrangean multiplier  $\lambda$  to enforce that the  $\tilde{q}(\cdot)$  sum to one.

Consider first  $B(\tilde{q}(y))$ , which we want to maximize. Taking the derivative:

$$\frac{\partial B}{\partial \tilde{q}(y)} = \Psi(\tilde{\alpha}_y) - \Psi(\sum_{y'} \tilde{\alpha}_{y'}) + \log p(x | y) - 1 - \log \tilde{q}(y) - \lambda \quad (15)$$

Setting equal to 0, we arrive at:

$$\begin{aligned} \tilde{q}(y) &= \frac{p(x | y) \exp\left(\Psi(\tilde{\alpha}_y) - \Psi(\sum_{y'} \tilde{\alpha}_{y'})\right)}{\exp(1 + \lambda)} \\ &= \frac{p(x | y) \exp\left(\Psi(\tilde{\alpha}_y) - \Psi(\sum_{y'} \tilde{\alpha}_{y'})\right)}{\sum_{y'} p(x | y') \exp\left(\Psi(\tilde{\alpha}_{y'}) - \Psi(\sum_{y''} \tilde{\alpha}_{y''})\right)} \end{aligned} \quad (16)$$

where we choose  $\lambda$  to enforce the sum-to-one constraint.

Now we turn to  $B(\tilde{\alpha})$ . Note that the derivative of  $\log \Gamma$  is the  $\Psi$  function, and the derivative of  $\Psi$  is denoted  $\Psi'$ . Taking the derivative:

$$\begin{aligned} \frac{\partial B}{\partial \tilde{\alpha}_y} &= (\alpha_y - \tilde{\alpha}_y + \tilde{q}(y))(\Psi'(\tilde{\alpha}_y) - \Psi'(\sum_{y'} \tilde{\alpha}_{y'})) - \Psi(\tilde{\alpha}_y) + \Psi(\sum_{y'} \tilde{\alpha}_{y'}) + \Psi(\tilde{\alpha}_y) - \Psi(\sum_{y'} \tilde{\alpha}_{y'}) \\ &= (\alpha_y - \tilde{\alpha}_y + \tilde{q}(y))(\Psi'(\tilde{\alpha}_y) - \Psi'(\sum_{y'} \tilde{\alpha}_{y'})) \end{aligned} \quad (17)$$

Following Blei, Ng, and Jordan<sup>1</sup> the derivative is zero when

$$\tilde{\alpha}_y = \alpha_y + \tilde{q}(y) \quad (18)$$

(This is intuitive; the derivative consists of two factors, the second of which will always be positive. To set the derivative equal to zero, we must force the first term to be zero, yielding equation 18.)

It is important to note that our assumption that  $n = 1$  makes all of the above easier to read but does not fundamentally change the result. If  $n > 1$ , it would be reasonable to choose  $Q(y_1^n, q_1^n) = \prod_{i=1}^n Q(y_i) \times Q(q_i)$ , taking the same form as above for each independent example.

In implementation, the variational inference algorithm would alternate between updating  $\tilde{\alpha}$  (using  $\tilde{q}$  in equation 18) and updating  $\tilde{q}$  (using  $\tilde{\alpha}$  in equation 16).

---

<sup>1</sup>David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. ‘‘Latent Dirichlet Allocation,’’ *JMLR* 3:993–1022.