Variational Inference for Dirichlet Process Mixtures

Michael David Sekora

29 October 2007

We begin by deriving the coordinate ascent algorithm for variational inference in exponential families. Given our latent variable model, we have hyperparameters θ, observed variables $\boldsymbol{x} = \{x_1, \ldots, x_N\}$, and latent variables $\boldsymbol{w} = \{w_1, \ldots, w_M\}$. We are interested in the hidden structure of the posterior distribution but need to bound its denominator:

$$p(\boldsymbol{w}|\boldsymbol{x}, \theta) = \frac{p(\boldsymbol{w}, \boldsymbol{x}|\theta)}{\int p(\boldsymbol{w}, \boldsymbol{x}|\theta) d\boldsymbol{w}} = \frac{p(\boldsymbol{w}, \boldsymbol{x}|\theta)}{p(\boldsymbol{x}|\theta)}. \tag{1}$$

We bound the denominator by using Jensen's inequality:

$$\begin{aligned} \log p(\boldsymbol{x}|\theta) &= \log \int p(\boldsymbol{w}, \boldsymbol{x}|\theta) d\boldsymbol{w} && (2) \\ &= \log \int p(\boldsymbol{w}, \boldsymbol{x}|\theta) \frac{q(\boldsymbol{w})}{q(\boldsymbol{w})} d\boldsymbol{w} && (3) \\ &\geq \int q(\boldsymbol{w}) \log \left(\frac{p(\boldsymbol{w}, \boldsymbol{x}|\theta)}{q(\boldsymbol{w})} \right) d\boldsymbol{w} && (4) \\ &= \mathbb{E} \left(\log p(\boldsymbol{w}, \boldsymbol{x}|\theta) \right) - \mathbb{E} \left(\log q(\boldsymbol{w}) \right), && (5) \end{aligned}$$

where q is the variational distribution. The last term in the above inequality is the entropy of q. The q that optimizes the above inequality is $p(\boldsymbol{w}|\boldsymbol{x}, \theta)$. However, we cannot compute the posterior distribution directly. Instead, we optimize the above bound by restricting ourselves to the fully-factorized variational distribution of the form:

$$q_{\boldsymbol{\nu}}(\boldsymbol{w}) = \prod_{i=1}^{M} q_{\nu_i}(w_i), \tag{6}$$

where $\boldsymbol{\nu} = \{\nu_1, \ldots, \nu_M\}$ are the variational parameters and each distribution is in the exponential family. We derive a coordinate ascent algorithm in which we iteratively maximize the bound with respect to each ν_i. Using the chain rule, we write the above inequality as:

$$\begin{aligned} \log p(\boldsymbol{x}|\theta) &\geq \mathbb{E} \left(\log p(\boldsymbol{x}|\theta) + \sum_{m=1}^{M} \log p(w_m|\boldsymbol{x}, w_1, \ldots, w_{m-1}, \theta) - \sum_{m=1}^{M} \log q_{\nu_m}(w_m) \right) && (7) \\ &= \log p(\boldsymbol{x}|\theta) + \sum_{m=1}^{M} \mathbb{E} \left(\log p(w_m|\boldsymbol{x}, w_1, \ldots, w_{m-1}, \theta) \right) - \sum_{m=1}^{M} \mathbb{E} \left(\log q_{\nu_m}(w_m) \right). && (8) \end{aligned}$$

We proceed with the coordinate ascent algorithm by isolating terms that contain ν_i and reordering $\boldsymbol{w}$ such that w_i is the last in the list. We maximize the following function with respect to ν_i:

$$l_i = \mathbb{E}\left(\log p(w_i|\boldsymbol{w}_{-i}, \boldsymbol{x}, \theta)\right) - \mathbb{E}\left(\log q_{\nu_i}(w_i)\right). \tag{9}$$

This maximization is equivalent to minimizing the KL-divergence, which is the difference between two distributions:

$$\min_{\nu} \quad \mathrm{KL}\left(q_{\nu}(\boldsymbol{w})||p(\boldsymbol{w}|\boldsymbol{x}, \theta)\right). \tag{10}$$

We assume that the variational and conditional distributions are in the exponential family:

$$q_{\nu_i} = h(w_i)\exp\left(\nu_i^T w_i - a(\nu_i)\right) \tag{11}$$

$$p(w_i|\boldsymbol{w}_{-i}, \boldsymbol{x}, \theta) = h(w_i)\exp\left(g_i(\boldsymbol{w}_{-i}, \boldsymbol{x}, \theta)^T w_i - a(g_i(\boldsymbol{w}_{-i}, \boldsymbol{x}, \theta))\right), \tag{12}$$

where $g_i(\boldsymbol{w}_{-i}, \boldsymbol{x}, \theta)$ denotes the natural parameter for the sufficient statistic w_i when conditioning on the remaining latent variables and the observations. After optimizing l with respect to ν_i, we find that the maximum is attained at:

$$\nu_i = \mathbb{E}\left(g_i(\boldsymbol{w}_{-i}, \boldsymbol{x}, \theta)\right). \tag{13}$$

We base the coordinate ascent algorithm on the above expression. Such an algorithm finds a local maximum for Eq 5 by iteratively updating ν_i for $i \in \{1, \ldots, M\}$. However, it may be possible to employ other algorithms such as the Newton, Conjugate Gradient, Gauss-Jacobi, or Gauss-Seidel Method.

Now, we apply our understanding to the Dirichlet Process (DP) mixture model, where the vector $\pi(\boldsymbol{v})$ comprises the infinite vector of mixing properties and $\{\eta_1^*, \eta_2^*, \ldots\}$ are the atoms representing the mixture components. Let z_n be an assignment variable of the mixture component with which the data point x_n is associated. The data can be described as arising from the following process:

1. Draw $v_i|\alpha \sim \mathrm{Beta}(1, \alpha), \quad i = \{1, 2, \ldots\}$
2. Draw $\eta_i^*|G_0 \sim G_o, \quad i = \{1, 2, \ldots\}$
3. For the n^{th} data point:
   - Draw $z_n|v_1, v_2, \ldots \sim \mathrm{Mult}(\pi(\boldsymbol{v}))$
   - Draw $x_n|z_n \sim p(x_n|\eta_{z_n}^*)$

We restrict ourselves to DP mixtures for which the observable data is drawn from an exponential family distribution and where the base distribution for the DP is the corresponding conjugate prior. The distribution of x_n conditional on z_n and $\{\eta_1^*, \eta_2^*, \ldots\}$ is:

$$p(x_n|z_n, \eta_1^*, \eta_2^*, \ldots) = \prod_{i=1}^{\infty} \left(h(x_n) \exp\left(\nu_i^{*T} x_n - a(\nu_i^*)\right)\right)^{\mathbf{1}[z_n=i]}. \tag{14}$$

Furthermore, the base distribution is:

$$p(\eta^*|\lambda) = h(\eta^*) \exp\left(\lambda_1^T \eta^* + \lambda_2(-a(\eta^*)) - a(\lambda)\right), \tag{15}$$

where we decompose the hyperparameters λ such that λ_1 contains the first $\dim(\eta^*)$ components and λ_2 is a scalar.

Using the Truncated Stick Breaking (TSB) representation as a model and approximation for the DP mixture, we develop a mean-field variational algorithm. In this representation the latent variables are the stick lengths, the atoms, and the cluster assignments: $\boldsymbol{w} = \{\boldsymbol{v}, \boldsymbol{\eta^*}, \boldsymbol{z}\}$. The hyperparameters are the scaling parameter and the parameter of the conjugate base distribution: $\theta = \{\alpha, \lambda\}$. T is the truncation level, which is a variational parameter that can be freely set. We write the variational bound on the log marginal probability of the data as:

$$\begin{aligned} \log p(\boldsymbol{x}|\alpha, \lambda) \quad \geq \quad & \mathbb{E}\left(\log p(\boldsymbol{v}|\alpha)\right) + \mathbb{E}\left(\log p(\boldsymbol{\eta^*}|\lambda)\right) \\ & + \sum_{n=1}^{N} \left(\mathbb{E}\left(\log p(z_n|\boldsymbol{v})\right) + \mathbb{E}\left(\log p(x_n|z_n)\right)\right) \\ & - \mathbb{E}\left(\log q(\boldsymbol{v}, \boldsymbol{\eta^*}, \boldsymbol{z})\right). \end{aligned} \tag{16}$$

Additionally, we propose the following factorized family of variational distributions for mean-field variational inference:

$$q(\boldsymbol{v}, \boldsymbol{\eta^*}, \boldsymbol{z}) = \prod_{t=1}^{T-1} q_{\gamma_t}(v_t) \prod_{t=1}^{T} q_{\tau_t}(\eta_t^*) \prod_{t=1}^{N} q_{\phi_n}(z_n), \tag{17}$$

where $q_{\gamma_t}(v_t)$ are beta distributions, $q_{\tau_t}(\eta_t^*)$ are exponential family distributions with natural parameters τ_t, and $q_{\phi_n}(z_n)$ are multinomial distributions. Furthermore, the variational parameters are:

$$\boldsymbol{\nu} = \{\gamma_1, \ldots, \gamma_{T-1}, \tau_1, \ldots, \tau_T, \phi_1, \ldots, \phi_N\}. \tag{18}$$

Before applying the coordinate ascent algorithm to the above bound, we note that all of the terms in the bound involve standard computations in the exponential family except for the third term. Therefore, we rewrite the third term using indicator random variables:

$$\begin{aligned}
\mathbb{E}\left(\log p(z_n|\boldsymbol{v})\right) &= \mathbb{E}\left(\log\left(\prod_{i=1}^{\infty}(1-v_i)^{\mathbf{1}[z_n>i]}v_i^{\mathbf{1}[z_n=i]}\right)\right) && (19)\\
&= \sum_{i=1}^{\infty}\left(q(z_n>i)\mathbb{E}\left(\log(1-v_i)\right)+q(z_n=i)\mathbb{E}\left(\log v_i\right)\right) && (20)\\
&= \sum_{i=1}^{T}\left(q(z_n>i)\mathbb{E}\left(\log(1-v_i)\right)+q(z_n=i)\mathbb{E}\left(\log v_i\right)\right), && (21)
\end{aligned}$$

because $\mathbb{E}\left(\log(1-v_T)\right)=0$, $q(z_n>T)=0$, and we can truncate the summation at $t=T$. In the above expressions:

$$\begin{aligned}
q(z_n=i) &= \phi_{n,i} && (22)\\
q(z_n>i) &= \sum_{j=i+1}^{T}\phi_{n,j} && (23)\\
\mathbb{E}\left(\log v_i\right) &= \Psi(\gamma_{i,1})-\Psi(\gamma_{i,1}+\gamma_{i,2}) && (24)\\
\mathbb{E}\left(\log(1-v_i)\right) &= \Psi(\gamma_{i,2})-\Psi(\gamma_{i,1}+\gamma_{i,2}). && (25)
\end{aligned}$$

Using the value for which l_i was maximized: $\nu_i=\mathbb{E}\left(g_i(\boldsymbol{w}_{-i},\boldsymbol{x},\theta)\right)$, the mean-field coordinate ascent algorithm yields:

$$\begin{aligned}
\gamma_{t,i} &= 1+\sum_n\phi_{n,t} && (26)\\
\gamma_{t,2} &= \alpha+\sum_n\sum_{j=t+1}^{T}\phi_{n,j} && (27)\\
\tau_{t,1} &= \lambda_1+\sum_n\phi_{n,t}x_n && (28)\\
\tau_{t,2} &= \lambda_2+\sum_n\phi_{n,t} && (29)\\
\phi_{n,t} &\propto \exp(S_t) && (30)\\
S_t &= \mathbb{E}\left(\log v_t\right)+\sum_{i=1}^{t-1}\mathbb{E}\left(\log(1-v_i)\right)+\mathbb{E}\left(\eta_t^*\right)^T x_n-\mathbb{E}\left(a(\eta_t^*)\right). && (31)
\end{aligned}$$

for $t\in\{1,\dots,T\}$ and $n\in\{1,\dots,N\}$.