Computing the ELBO of a Dirichlet distribution

Philip Schulz

last modified: February 28, 2017

Abstract

This short note describes how to compute the value of the evidence lower bound (ELBO) of the Dirichlet distribution for fixed variational parameters. The derivation extends to a large class of models, including Bayesian non-parametric models of discrete data where the ELBO contains many Dirichlet terms.

1 Computing the Dirichlet ELBO

We assume a data categorical (or multinomial) set $(x_1, \ldots, x_n)$ where each $x_i,\ 1 \leq i \leq n$ represents a group of data points. These groups may vary in size so that we have $|x_i| = N_i,\ N_i \in \mathbb{N}$. We model each group x_i as a draw from a cateogorical (or multinomial) distribution with parameter vector $\theta_i \in \mathbb{R}^m$. The parameter vectors θ_i are in turn assumed to be i.i.d. draws from a Dirichlet distribution with parameter vector α. This leads to the following likelihood.

$$\mathrm{L}(\theta_1^n, \alpha) = \prod_{i=1}^{n} p(\theta_i|\alpha) \prod_{j=1}^{N_i} P(x_{ij}|\theta_i) \tag{1}$$

At this point it is convenient to remind ourselves of the exponential family form of the Dirichlet distribution which is given below.

$$p(\theta|\alpha) = \exp\left(\sum_{j=1}^{N} \log(\theta_j)(\alpha - 1) - \left(\sum_{j=1}^{N} \log\Gamma(\alpha_j) - \log\Gamma\left(\sum_{j=1}^{N} \alpha_j\right)\right)\right) \tag{2}$$

Because the group-specific categorical distributions were drawn from the same Dirichlet our data points have become marginally dependent, making exact inference in this model impossible. We therefore wish to approximate the log-likelihood of our model with a tractable lower bound so as to optimise that lower bound instead of the log-likelihood itself. Using standard results

from variational inference, the evidence lower bound fulfills these desiderata. For our model it is

$$\text{ELBO} = \sum_{i=1}^{n} \mathbb{E}_{q_i} \left[\log \left(p(\theta_i|\alpha) \prod_{j=1}^{N_i} P(x_{ij}|\theta_i) \right) \right] + \mathbb{H}(q_i) \tag{3}$$

where $q_i = \text{Dir}(\gamma_i)$ is the variational distribution for the i^{th} group which is also assumed to be Dirichlet.[1] Since the Elbo factorises over groups, we only derive it for one group here. Let us first deal with the expectation $\mathbb{E}_{q_i}\left[\log\left(\prod_{j=1}^{N_i} P(x_{ij}|\theta_i\right)\right]$. Since the expected value of the sufficient statistics of any exponential family distribution is equal to the first derivate of its log-normaliser[2], we conclude from Equation (2) that

$$\mathbb{E}_{q_i}\left[\log\left(P(x_{ij}|\theta_i)\right)\right] = \mathbb{E}_{q_i}\left[\log(\theta_{ij})\right] = \Psi(\gamma_{ij}) - \Psi\left(\sum_{j=1}^{N_i} \gamma_{ij}\right) \tag{4}$$

where $\Psi = \frac{\Gamma'}{\Gamma}$ is the digamma function.

Next, we need to compute $\mathbb{E}_{q_i}\left[\log(p(\theta_i|\alpha)\right] + \mathbb{H}(q_i)$ which for convenience we rewrite as $\mathbb{E}_{q_i}\left[\log(p(\theta_i|\alpha)) - \log(q_i(\theta_i|\gamma_i))\right]$. Using the exponential family formulation of the Dirichlet one more time, computing this quantity amounts to taking the (expected) difference between two exponential families which share their sufficient statistics. Notice that these (expected) sufficient statistics are exactly $\forall j\ \mathbb{E}_{q_i}\left[\log(\theta_{ij})\right]$ from Equation (2). In order to simplify notation, we solely derive the difference between the exponential families without reference to a particular group.

$$\begin{aligned}
&\mathbb{E}_q\left[\log(p(\theta|\alpha)) - \log(q(\theta|\gamma))\right] = \\
&\sum_{j=1}^{N} \mathbb{E}_q\left[\log(\theta_j)\right](\alpha - 1) - \left(\sum_{j=1}^{N} \log\Gamma(\alpha_j) - \log\Gamma\left(\sum_{j=1}^{N} \alpha_j\right)\right) \\
&- \sum_{j=1}^{N} \mathbb{E}_q\left[\log(\theta_j)\right](\gamma - 1) + \left(\sum_{j=1}^{N} \log\Gamma(\gamma_j) - \log\Gamma\left(\sum_{j=1}^{N} \gamma_j\right)\right) \\
&= \left\{\sum_{j=1}^{N} \mathbb{E}_q\left[\log(\theta_j)\right](\alpha_j - \gamma_j) + \log\Gamma(\gamma_j) - \log\Gamma(\alpha_j)\right\} \\
&+ \log\Gamma\left(\sum_{j=1}^{N} \alpha_j\right) - \log\Gamma\left(\sum_{j=1}^{N} \gamma_j\right)
\end{aligned} \tag{5}$$

[1] Throughout this note we make a mean field assumption meaning that the variational distribution factorises fully over all nodes in our graphical model.

[2] The log-normaliser in Equation (2) is the second term of the difference in the exponent.