Algorithm 1 Black Box Variational Inference

Input: data x, joint distribution p, mean field variational family q.
Initialize $\lambda_{1:n}$ randomly, $t = 1$.
repeat
  // Draw S samples from q
  for $s = 1$ **to** S **do**
    $z[s] \sim q$
  end for
  $\rho = t$th value of a Robbins Monro sequence (Eq. 2)
  $\lambda = \lambda + \rho \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z[s]|\lambda)(\log p(x, z[s]) - \log q(z[s]|\lambda))$
  $t = t + 1$
until change of λ is less than 0.01.

Figure 1: Taken from article [1] (note that the reference to Eq. 2 is in the original article, not this document).

Maximizing the ELBO for the simple model using stochastic gradients a.k.a. Black-box variational inference (P. Marttinen 2018)

Idea of black-box variational inference (BBVI)

Let $\mathcal{L}$ denote the ELBO, which depends on some variational parameters λ, i.e.

$$\mathcal{L}(\lambda) = E_{q(z|\lambda)}[\log p(x, z) - \log q(z|\lambda)],$$

where $q(z|\lambda)$ is the variational approximation of $p(z|x)$. To optimize the ELBO using stochastic gradient search, we iterate using

$$\lambda_{t+1} = \lambda_t + \rho_t \eta_t,$$

where η_t is a random variable whose expected value is the gradient of $\mathcal{L}$, that is

$$E(\eta_t) = \nabla_\lambda \mathcal{L},$$

and ρ_t is a suitably defined sequence of step sizes. The following formula can be derived for the gradient, see article [1]:

$$\nabla_\lambda \mathcal{L} = E_{q(z|\lambda)} \left[\nabla_\lambda \log q(z|\lambda)(\log p(x, z) - \log q(z|\lambda)) \right]. \tag{1}$$

In black-box variational inference, the expectation in Equation 1 over the distribution $q(z|\lambda)$ is approximated by generating S samples $z_s \sim q(z|\lambda)$ from the distribution, computing the term whose expectation we want to approximate using each of these samples, and averaging over the samples, i.e.

$$\nabla_\lambda \mathcal{L} \approx \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda)). \tag{2}$$

The full optimization algorithm using these stochastic gradients is shown in Figure 1. We note that the BBVI as presented in [1] includes further technical details to improve the algorithm, e.g., Rao-Blackwellization and the use of control variates, which are not discussed here for simplicity.

Stochastic gradient of the ELBO for the simple model

The logarithm of the joint distribution in the 'simple model', see *simple_vb_example.pdf*, can be written as:

$$\log p(\mathbf{x}, \mathbf{z}, \tau, \theta) = \log p(\tau) + \log p(\theta) + \log p(\mathbf{z}|\tau) + \log p(\mathbf{x}|\mathbf{z}, \theta). \tag{3}$$

We assume the mean-field approximation

$$p(\mathbf{z}, \tau, \theta|\mathbf{x}) \approx q(\tau)q(\theta)\textstyle\prod_n q(z_n),$$

with factors

$$q(z_n|r_{n1}, r_{n2}) = Categorical(z_n|r_{n1}, r_{n2}) = r_{n1}^{z_{n1}} r_{n2}^{z_{n2}}, \tag{4}$$
$$q(\tau) = Beta(\tau|\alpha_\tau, \beta_\tau), \tag{5}$$
$$q(\theta) = N(\theta|m_2, \beta_2^{-1}), \tag{6}$$

where r_{n1}, r_{n2}, $n = 1, \ldots, N$, α_τ, β_τ, m_2, and β_2 are the *variational parameters* (represented jointly by λ in the general description of BBVI above). With these assumptions, the term whose expectation we are computing in Equation 1 can be written as

$$\begin{aligned}
\nabla_\lambda \log q(z|\lambda)(\log p(x,z) - \log q(z|\lambda)) &= \ldots \textbf{exercise} \ldots \\
&= \nabla \left[\log q(\tau|\alpha_\tau, \beta_\tau) + \log q(\theta|m_2, \beta_2^{-1}) + \textstyle\sum_n \log q(z_n|r_n)\right] \\
&\times \left[\log p(\tau) + \log p(\theta) + \textstyle\sum_n \log p(z_n|\tau) + \textstyle\sum_n \log p(x_n|z_n, \theta)\right. \\
&\left.- \log q(\tau|\alpha_\tau, \beta_\tau) - \log q(\theta|m_2, \beta_2^{-1}) - \textstyle\sum_n \log q(z_n|r_n)\right].
\end{aligned} \tag{7}$$

To simplify notation, we introduce a function f:

$$f(\alpha_\tau, \beta_\tau, m_2, \beta_2^{-1}, \mathbf{r}) \triangleq \log q(\tau|\alpha_\tau, \beta_\tau) + \log q(\theta|m_2, \beta_2^{-1}) + \textstyle\sum_n \log q(z_n|r_n). \tag{8}$$

On the second line of Equation 7 we compute the gradient of f, and for this we need to differentiate $f(\alpha_\tau, \beta_\tau, m_2, \beta_2^{-1}, \mathbf{r})$ with respect to each variational parameter (r_{n1}, $n = 1, \ldots, N$, α_τ, β_τ, m_2, and β_2)[1]. Computing these partial derivatives becomes easy by noting that each variational parameter appears in only one term in f, so we can discard other terms when computing the respective derivative. For example, when differentiating w.r.t. α_τ, we only need to consider the first term in Equation 8 because the second and third terms do not depend on α_τ. Therefore

$$\begin{aligned}
\frac{\partial f}{\partial \alpha_\tau} &= \frac{\partial}{\partial \alpha_\tau} \log q(\tau|\alpha_\tau, \beta_\tau) \\
&= \frac{\partial}{\partial \alpha_\tau} \left[\log \Gamma(\alpha_\tau + \beta_\tau) - \log \Gamma(\alpha_\tau) - \log \Gamma(\beta_\tau) + (\alpha_\tau - 1) \log \tau + (\beta_\tau - 1) \log(1 - \tau)\right] \\
&= \psi(\alpha_\tau + \beta_\tau) - \psi(\alpha_\tau) + \log \tau,
\end{aligned}$$

where we used the fact that $\frac{\partial}{\partial x} \log \Gamma(x) = \psi(x)$. The rest of the partial derivatives, $\frac{\partial f}{\partial \beta_\tau}$, $\frac{\partial f}{\partial m_2}$, $\frac{\partial f}{\partial \beta_2}$, $\frac{\partial f}{\partial r_{n1}}$, are similar short calculations, and they are left as an **exercise**. The other terms that appear in Equation 7 are just log-densities of the (conditional) distributions of the model itself (the third line in Equation 7) or log-densities of the approximate model (the fourth line of Equation 7), and, because all of these distributions are known and of standard form, they are straightforward to compute.

Summary of computing stochastic gradients of ELBO for the simple model

1. Assume some current values for all the variational parameters $\alpha_\tau, \beta_\tau, m_2, \beta_2, r_{n1}$, $n = 1, \ldots, N$

2. Simulate the unobserved variables in the model, $z_n^{(s)}, n = 1, \ldots, n$ $, \tau^{(s)}, \theta^{(s)}$, using the current approximate distributions specified in Equations 4, 5, 6.

3. Plug-in the simulated values $\mathbf{z}^{(s)}, \tau^{(s)}, \theta^{(s)}$ into Equation 7, and compute the terms as described in the previous section. This gives $\nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda))$ in Equation 2.

4. Repeat steps 2 and 3 S times, and average the computed values for $\nabla_\lambda \log q(z_s|\lambda)(\log p(x, z_s) - \log q(z_s|\lambda))$.

The outcome of these steps is an estimate of the stochastic gradient as specified in Equation 2, which can then be used in the stochastic optimization algorithm presented in Figure 1 to learn the values of the variational parameters that maximize the ELBO.

[1] We only need to differentiate w.r.t. r_{n1} and not r_{n2}, because $r_{n2} = 1 - r_{n1}$.

References

[1] Rajesh Ranganath, Sean Gerrish, and David Blei. Black Box Variational Inference. In Samuel Kaski and Jukka Corander, editors, *Proceedings of the Seventeenth International Conference on Artificial Intelligence and Statistics*, volume 33 of *Proceedings of Machine Learning Research*, pages 814–822, Reykjavik, Iceland, 22–25 Apr 2014. PMLR.