Black Box Variational Inference (Extra Materials)

Rajesh Ranganath **Sean Gerrish** **David M. Blei**
Princeton University, 35 Olden St., Princeton, NJ 08540
{rajeshr,sgerrish,blei}@cs.princeton.edu

Appendix: The Gradient of the ELBO

The key idea behind our algorithm is that the gradient of the ELBO can be written as an expectation with respect to the variational distribution. We start by differentiating Eq. 1,

$$\begin{aligned}
\nabla_\lambda \mathcal{L} &= \nabla_\lambda \int (\log p(x,z) - \log q(z|\lambda)) q(z|\lambda) dz \\
&= \int \nabla_\lambda [(\log p(x,z) - \log q(z|\lambda)) q(z|\lambda)] dz \\
&= \int \nabla_\lambda [\log p(x,z) - \log q(z|\lambda)] q(z|\lambda) dz \\
&\quad + \int \nabla_\lambda q(z|\lambda) (\log p(x,z) - \log q(z|\lambda)) dz \\
&= -\mathrm{E}_q [\nabla_\lambda \log q(z|\lambda)] \qquad \text{(A.1)} \\
&\quad + \int \nabla_\lambda q(z|\lambda) (\log p(x,z) - \log q(z|\lambda)) dz,
\end{aligned}$$

where we have exchanged derivatives with integrals via the dominated convergence theorem[1] [Cinlar, 2011] and used $\nabla_\lambda[\log p(x,z)] = 0$.

The first term in Eq. A.1 is zero. To see this, note

$$\begin{aligned}
\mathrm{E}_q[\nabla_\lambda \log q(z|\lambda)] &= \mathrm{E}_q \left[\frac{\nabla_\lambda q(z|\lambda)}{q(z|\lambda)} \right] = \int \nabla_\lambda q(z|\lambda) dz \\
&= \nabla_\lambda \int q(z|\lambda) dz = \nabla_\lambda 1 = 0.
\end{aligned} \tag{A.2}$$

To simplify the second term, first observe that $\nabla_\lambda[q(z|\lambda)] = \nabla_\lambda[\log q(z|\lambda)] q(z|\lambda)$. This fact gives us the gradient as an expectation,

$$\begin{aligned}
\nabla_\lambda \mathcal{L} &= \int \nabla_\lambda [q(z|\lambda)] (\log p(x,z) - \log q(z|\lambda)) dz \\
&= \int \nabla_\lambda \log q(z|\lambda) (\log p(x,z) \\
&\quad - \log q(z|\lambda)) q(z|\lambda) dz \\
&= \mathrm{E}_q [\nabla_\lambda \log q(z|\lambda) (\log p(x,z) - \log q(z|\lambda))].
\end{aligned}$$

Unnormalized joint distributions In some situations it is easier to write down an unnormalized version of the joint distribution. In this case, we can replace p with its unnormalized version u in Eq. A.1 and still have a valid gradient. More formally, let $p(x,z) = u(x,z)/C$, then our gradient becomes

$$\begin{aligned}
\nabla_\lambda \mathcal{L} &= \mathrm{E}_q [\nabla_\lambda \log q(z|\lambda) (\log p(x,z) - \log q(z|\lambda))] \\
&= \mathrm{E}_q [\nabla_\lambda \log q(z|\lambda) (\log u(x,z) - \log C - \log q(z|\lambda))] \\
&= \mathrm{E}_q [\nabla_\lambda \log q(z|\lambda) (\log u(x,z) - \log q(z|\lambda))],
\end{aligned} \tag{A.3}$$

where the last equality follows from Eq. A.2.

Supplement

Derivation of the Rao-Blackwellized Gradient To compute the Rao-Blackwellized estimators, we need to compute conditional expectations. Due to the mean field-assumption, the conditional expectation simplifies due to the factorization

$$\begin{aligned}
\mathrm{E}[J(X,Y)|X] &= \frac{\int J(x,y) p(x) p(x) dy}{\int p(x) p(y) dy} \\
&= \int J(x,y) p(y) dy = \mathrm{E}_y[J(x,y)].
\end{aligned} \tag{S.4}$$

Therefore, to construct a lower variance estimator when the joint distribution factorizes, all we need to do is integrate out some variables. In our problem this means for each component of the gradient, we should compute expectations with respect to the other factors. We present the estimator in the full mean field family of variational distributions, but note it applies to any variational approximation with some factorization like structured mean-field.

Thus, under the mean field assumption the Rao-

[1] The score function exists. The score and likelihoods are bounded.

Blackwellized estimator for the gradient becomes

$$\begin{aligned}\nabla_\lambda \mathcal{L} =& \mathrm{E}_{q_1} \dots \mathrm{E}_{q_n}[\sum_{j=1}^{n} \nabla_\lambda \log q_j(z_j|\lambda_j)(\log p(x,z) \\ &- \sum_{j=1}^{n} \log q_j(z_j|\lambda_j))]. \end{aligned} \tag{S.5}$$

Recall the definitions from Section 3 where we defined $\nabla_{\lambda_i}\mathcal{L}$ as the gradient of the ELBO with respect to λ_i, p_i as the components of the log joint that include terms from the ith factor, and $E_{q_{(i)}}$ as the expectation with respect to the set of latent variables that appear in the complete conditional for z_i. Let p_{-i} be the components of the joint that does not include terms from the ith factor respectively. We can write the gradient with respect to the ith factor's variational parameters as

$$\begin{aligned}&\nabla_{\lambda_i}\mathcal{L} \\ &= E_{q_1} \dots \mathrm{E}_{q_n}[\nabla_{\lambda_i} \log q_i(z_i|\lambda_i)(\log p(x,z) \\ &\quad - \sum_{j=1}^{n} \log q_j(z_j|\lambda_j))] \\ &= E_{q_1} \dots \mathrm{E}_{q_n}[\nabla_{\lambda_i} \log q_i(z_i|\lambda_i)(\log p_i(x,z) \\ &\quad + \log p_{-i}(x,z) - \sum_{j=1}^{n} \log q_j(z_j|\lambda_j))] \\ &= E_{q_i}[\nabla_{\lambda_i} \log q_i(z_i|\lambda_i)(E_{q_{-i}}[\log p_i(x,z_{(i)})] \\ &\quad - \log q_i(z_i|\lambda_i) + \mathrm{E}_{q_{-i}}[\log p_{-i}(x,z) \\ &\quad - \sum_{j=1,i\neq j}^{n} \log q_j(z_j|\lambda_j)]] \\ &= E_{q_i}[\nabla_{\lambda_i} \log q_i(z_i|\lambda_i)(\mathrm{E}_{q_{-i}}[\log p_i(x,z)] \\ &\quad - \log q_i(z_i|\lambda_i) + C_i)] \\ &= E_{q_i}[\nabla_{\lambda_i} \log q_i(z_i|\lambda_i)(\mathrm{E}_{q_{-i}}[\log p_i(x,z_{(i)})] \\ &\quad - \log q_i(z_i|\lambda_i))] \\ &= E_{q_{(i)}}[\nabla_{\lambda_i} \log q_i(z_i|\lambda_i)(\log p_i(x,z_{(i)}) - \log q_i(z_i|\lambda_i))], \end{aligned} \tag{S.6}$$

where we have leveraged the mean field assumption and made use of the identity for the expected score Eq. 14. This means we can Rao-Blackwellize the gradient of the variational parameter λ_i with respect to the latent variables outside of the Markov blanket of z_i without needing model specific computations.

Derivation of Stochastic Inference in Hierarchical Bayesian Models Recall the definition of a hierarchical Bayesian model with n observations given in Eq. 12

$$\begin{aligned}&\log p(x_{1\dots n}, z_{1\dots n}, \beta) \\ &\quad = \log p(\beta|\eta) + \sum_{i=1}^{n} \log p(z_i|\beta) + \log p(x_i|z_i,\beta).\end{aligned}$$

Let the variational approximation for the posterior distribution be from the mean field family. Let λ be the global variational parameter and let $\phi_{1\dots n}$ be the local variational parameters. The variational family is

$$q(\beta, z_{1\dots n}) = q(\beta|\lambda)\prod_{i=1}^{m} q(z_i|\phi_i). \tag{S.7}$$

Using the Rao Blackwellized estimator to compute noisy gradients in this family for this model gives

$$\begin{aligned}\hat{\nabla}_\lambda \mathcal{L} =& \frac{1}{S}\sum_{i=1}^{S} \nabla_\lambda \log q(\beta_s|\lambda)(\log p(\beta_s|\eta) - \log q(\beta_s|\lambda) \\ &+ \sum_{i=1}^{n}(\log p(z_{is}|\beta_s) + \log p(x_i|z_{is},\beta_s))) \\ \hat{\nabla}_{\phi_i} \mathcal{L} =& \frac{1}{S}\sum_{i=1}^{S} \nabla_{\phi_i} \log q(z_{is}|\phi_i)((\log p(z_{is}|\beta_s) \\ &+ \log p(x_i|z_{is},\beta_s) - \log q(z_{is}|\phi_i))).\end{aligned}$$

Unfortunately, this estimator requires iterating over every data point to compute noisy realizations of the gradient. We can mitigate this by subsampling observations. If we let $i \sim Unif(1 \dots n)$, then we can write down a noisy gradient for the ELBO that does not need to iterate over every observation; this noisy gradient is

$$\begin{aligned}\hat{\nabla}_\lambda \mathcal{L} =& \frac{1}{S}\sum_{i=1}^{S} \nabla_\lambda \log q(\beta_s|\lambda)(\log p(\beta_s|\eta) - \log q(\beta_s|\lambda) \\ &+ n(\log p(z_{is}|\beta_s) + \log p(x_i|z_{is},\beta_s))) \\ \hat{\nabla}_{\phi_i} \mathcal{L} =& \frac{1}{S}\sum_{i=1}^{S} \nabla_{\phi_i} \log q(z_{is}|\phi_i)(n(\log p(z_{is}|\beta_s) \\ &+ \log p(x_i|z_{is},\beta_s) - \log q(z_{is}|\phi_i))) \\ \hat{\nabla}_{\phi_j} \mathcal{L} =& 0 \text{ for all } j \neq i.\end{aligned}$$

The expected value of this estimator with respect to the samples from the variational distribution and the sampled data point is the gradient of the ELBO. This means we can use it define a stochastic optimization procedure to maximize the ELBO. We can lower the variance of the above estimator by introducing control variates. For the dth dimension of the respective parameters, the control variates are

$$\begin{aligned}f_{\lambda_d}(\beta, z_i) =& \nabla_{\lambda_d} \log q(\beta|\lambda)(\log p(\beta|\eta) - \log q(\beta|\lambda) \\ &+ n(\log p(z_i|\beta) + \log p(x_i|z_i,\beta))) \\ h_{\lambda_d}(\beta) =& \nabla_{\lambda_d} \log q(\beta|\lambda) \\ f_{\phi_{id}}(\beta, z_i) =& \nabla_{\phi_{id}} \log q(z|\phi_i)(n(\log p(z_i|\beta) \\ &+ \log p(x_i|z_i,\beta) - \log q(z_i|\phi_i))) \\ h_{\phi_{id}}(z_i) =& \nabla_{\phi_{id}} \log q(z_i|\phi_i).\end{aligned} \tag{S.8}$$

We can estimate the optimal scalings given by Eq. 8 for the control variates using a small number of samples. This gives the following lower variance noisy gradient that does not need to iterate over all of the observations at each update

$$\begin{aligned}
\hat{\nabla}_{\lambda_d}\mathcal{L} =& \frac{1}{S}\sum_{i=1}^{S} \nabla_{\lambda_d} \log q(\beta_s|\lambda)(\log p(\beta_s|\eta) - \log q(\beta_s|\lambda) \\
& - a_{\hat{\lambda}_d} + n(\log p(z_{i\,s}|\beta_s) + \log p(x_i|z_{i\,s}, \beta_s))) \\
\hat{\nabla}_{\phi_{i\,d}}\mathcal{L} =& \frac{1}{S}\sum_{i=1}^{S} \nabla_{\phi_{i\,d}} \log q(z_{i\,s}|\phi_i)(-a_{\hat{\phi}_{i\,d}} + n(\log p(z_{i\,s}|\beta_s) \\
& + \log p(x_i|z_{i\,s}, \beta_s) - \log q(z_{i\,s}|\phi_i))) \\
\hat{\nabla}_{\phi_j}\mathcal{L} =& 0 \text{ for all } j \neq i,
\end{aligned} \tag{S.9}$$

where $a_{\hat{\lambda}_d}$ and $a_{\hat{\phi}_{i\,d}}$ are the respective control variate scalings.

Gamma parameterization equivalence The shape α and rate β parameterization can be written in terms of the mean μ and variance σ^2 of the gamma as

$$\alpha = \frac{\mu^2}{\sigma^2}, \quad \beta = \frac{\mu}{\sigma^2}. \tag{S.10}$$

References

E. Cinlar. *Probability and Stochastics.* Springer, 2011.