Hierarchical sampling for BBVI

Allison J.B. Chaney

September 8, 2015

1 BBVI

- Variational inference minimizes the KL divergence from an approximating distribution q to the true posterior p.
- This is equivalent to maximizing the ELBO: $\mathscr{L}(\lambda) \triangleq \mathbb{E}_{q_\lambda(z)}[\log p(x,z) - \log q_\lambda(z)]$
- We use stochastic optimization to update λ. This means we need a noisy, unbiased gradient that we can compute using samples from the posterior.
- To do so, we write the gradient of the ELBO as an expectation with respect to the variational distribution:

$$\begin{aligned}
\nabla_\lambda \mathscr{L} &= \nabla_\lambda \mathbb{E}_{q_\lambda(z)}[\log p(x,z) - \log q_\lambda(z)] \\
&= \nabla_\lambda \int_z [\log p(x,z) - \log q_\lambda(z)]\, q_\lambda(z) dz \\
&= \int_z \nabla_\lambda \left([\log p(x,z) - \log q_\lambda(z)]\, q_\lambda(z) \right) dz \\
&= \int_z q_\lambda(z) \nabla_\lambda [\log p(x,z) - \log q_\lambda(z)] + [\log p(x,z) - \log q_\lambda(z)]\, \nabla_\lambda q_\lambda(z) dz \\
&= \int_z -q_\lambda(z) \nabla_\lambda \log q_\lambda(z) + [\log p(x,z) - \log q_\lambda(z)]\, \nabla_\lambda q_\lambda(z) dz \\
&= \int_z \left(-q_\lambda(z) \nabla_\lambda \log q_\lambda(z) + [\log p(x,z) - \log q_\lambda(z)]\, \nabla_\lambda q_\lambda(z) \right) \frac{q_\lambda(z)}{q_\lambda(z)} dz \\
&= \mathbb{E}_{q_\lambda(z)} \left[\left(-q_\lambda(z) \nabla_\lambda \log q_\lambda(z) + (\log p(x,z) - \log q_\lambda(z))\, \nabla_\lambda q_\lambda(z) \right) \frac{1}{q_\lambda(z)} \right] \\
&= \mathbb{E}_{q_\lambda(z)} \left[-\nabla_\lambda \log q_\lambda(z) + (\log p(x,z) - \log q_\lambda(z)) \frac{\nabla_\lambda q_\lambda(z)}{q_\lambda(z)} \right] \\
&= \mathbb{E}_{q_\lambda(z)} \left[-\nabla_\lambda \log q_\lambda(z) + (\log p(x,z) - \log q_\lambda(z))\, \nabla_\lambda \log q_\lambda(z) \right] \\
&= \mathbb{E}_{q_\lambda(z)} \left[\nabla_\lambda \log q_\lambda(z) (\log p(x,z) - \log q_\lambda(z) - 1) \right]
\end{aligned}$$

- Now we take S samples $z_s \sim q(z \,|\, \lambda)$ and compute a noisy unbiased gradient

$$\hat{\nabla}_\lambda \mathscr{L} = \frac{1}{S} \sum_{s=1}^{S} \nabla_\lambda \log q(z_s \,|\, \lambda) (\log p(x, z_s) - \log q(z_s \,|\, \lambda) - 1)$$

2 BBVI with hierarchical sampling

- We want to maximize a new variant of the ELBO:

$$\begin{aligned}\mathscr{L}(\alpha) &\triangleq \mathbb{E}_{q_\alpha(z)}[\log p(x,z) - \log q_\alpha(z)] \\ &= \mathbb{E}_{q_\alpha(z)}[\log p(x,z) - \log(q(z\,|\,\lambda)q(\lambda\,|\,\alpha))] \\ &= \mathbb{E}_{q_\alpha(z)}[\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)]\end{aligned}$$

- Like before, we use stochastic optimization to update α. Again, we need a noisy, unbiased gradient that we can compute using samples from the posterior.
- To do so, we write the gradient of the new ELBO as an expectation with respect to the variational distribution:

$$\begin{aligned}\nabla_\lambda \mathscr{L} &= \nabla_\alpha \mathbb{E}_{q_\alpha(z)}[\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)] \\ &= \nabla_\alpha \int_z [\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)]\, q_\alpha(z) dz \\ &= \int_z \nabla_\alpha \left([\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)]\, q_\alpha(z)\right) dz \\ &= \int_z q_\alpha(z)\nabla_\alpha [\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)] + [\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)]\, \nabla_\alpha q_\alpha(z) dz \\ &= \int_z -q_\alpha(z)\nabla_\alpha \log q(\lambda\,|\,\alpha) + [\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)]\, \nabla_\alpha q_\alpha(z) dz \\ &= \int_z -q_\alpha(z)\nabla_\alpha \log q(\lambda\,|\,\alpha) + [\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)]\, \nabla_\alpha q_\alpha(z) dz \\ &= \int_z (-q_\alpha(z)\nabla_\alpha \log q(\lambda\,|\,\alpha) + [\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha)]\, \nabla_\alpha q_\alpha(z))\, \frac{q_\alpha(z)}{q_\alpha(z)} dz \\ &= \mathbb{E}_{q_\alpha(z)}\left[(-q_\alpha(z)\nabla_\alpha \log q(\lambda\,|\,\alpha) + (\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha))\, \nabla_\alpha q_\alpha(z))\, \frac{1}{q_\alpha(z)}\right] \\ &= \mathbb{E}_{q_\alpha(z)}\left[-\nabla_\alpha \log q(\lambda\,|\,\alpha) + (\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha))\, \frac{\nabla_\alpha q_\alpha(z)}{q_\alpha(z)}\right] \\ &= \mathbb{E}_{q_\alpha(z)}[-\nabla_\alpha \log q(\lambda\,|\,\alpha) + (\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha))\, \nabla_\alpha \log q_\alpha(z)] \\ &= \mathbb{E}_{q_\alpha(z)}[\nabla_\alpha \log q(\lambda\,|\,\alpha)\, (\log p(x,z) - \log q(z\,|\,\lambda) - \log q(\lambda\,|\,\alpha) - 1)]\end{aligned}$$

- Now we take S samples $\lambda_s \sim q(\lambda\,|\,\alpha)$ then $z_s \sim q(z\,|\,\lambda_s)$ and compute a noisy unbiased gradient

$$\hat{\nabla}_\alpha \mathscr{L} = \frac{1}{S}\sum_{s=1}^{S} \nabla_\alpha \log q(\lambda_s\,|\,\alpha)\, (\log p(x,z_s) - \log q(z_s\,|\,\lambda_s) - \log q(\lambda_z\,|\,\alpha) - 1)$$