Sticking the Landing: A Simple Reduced-Variance Gradient for ADVI

Geoffrey Roeder
University of Toronto
roeder@cs.toronto.edu

Yuhuai Wu
University of Toronto
ywu@cs.toronto.edu

David Duvenaud
University of Toronto
duvenaud@cs.toronto.edu

Abstract

Compared to the REINFORCE gradient estimator, the reparameterization trick usually gives lower-variance estimators. We propose a simple variant of the standard reparameterized gradient estimator for the evidence lower bound that has even lower variance under certain circumstances. Specifically, we decompose the derivative with respect to the variational parameters into two parts: a path derivative and the score function. Removing the second term produces an unbiased gradient estimator whose variance approaches zero as the approximate posterior approaches the exact posterior. We propose that the removed term has arbitrarily high variance when the variational posterior has a complex form, as when using adaptive posteriors such as given by normalizing flows or stochastic Hamiltonian inference.

1 Estimators of the Evidence Lower Bound

Variational inference posits a family of distributions Q and attempts to find an approximate posterior q_ϕ by optimizing the evidence lower bound (ELBO):

$$\mathcal{L}(\phi) = \mathbb{E}_{\mathbf{z}\sim q}[\log p(\mathbf{x}, \mathbf{z}) - \log q_\phi(\mathbf{z} \mid \mathbf{x})] \tag{ELBO}$$

An unbiased approximation of the gradient of the ELBO allows stochastic gradient descent to scalably learn complex models.

When the joint distribution $p(x, z)$ can be evaluated by $p(x|z)$ and $p(z)$ separately, the ELBO can be written in the following three forms:

$$\begin{aligned}
\mathcal{L}(\phi) &= \mathbb{E}_{\mathbf{z}\sim q}[\log p(x|z) + \log p(z) - \log q_\phi(z|x)] && (1)\\
&= \mathbb{E}_{\mathbf{z}\sim q}[\log p(x|z) + \log p(z))] + \mathbb{H}[q_\phi] && (2)\\
&= \mathbb{E}_{\mathbf{z}\sim q}[\log p(x|z)] - KL(q_\phi(z|x)||p(z)) && (3)
\end{aligned}$$

By sampling $z \sim q(z)$, equations (1), (2), and (3) can be used to construct Monte-Carlo estimates of the ELBO. When $p(z)$ and $q(z|x)$ are multivariate Gaussians, using (3) is appealing because it analytically integrates out terms that would otherwise have to be estimated by Monte Carlo. Intuitively, we might that using exact integrals wherever possible will give lower-variance estimators.

Surprisingly, though, there are circumstances under which (1), which we call the full Monte Carlo estimator of the ELBO, has lower variance than the estimator based on (3) which calculates the KL divergence exactly. Specifically, when $q(z|x) = p(z|x)$, i.e. the variational approximation is exact, then the variance of the full Monte Carlo estimator is exactly zero, since its value is a constant

independent of z:

$$\begin{aligned}
\hat{\mathcal{L}}_{MC}(\phi) &= \log p(x|z_i) + \log p(z_i) - \log q_\phi(z_i|x) \qquad z_i \overset{\text{iid}}{\sim} q(z) && (4)\\
&= \log p(x|z_i) + \log p_\theta(z_i) - \log q_\phi(z_i|x) && (5)\\
&= \log p(x, z_i) - \log q_\phi(z_i|x) && (6)\\
&= \log p(z_i|x) + \log p(x) - \log p(z_i|x) \qquad (\text{using } q(z|x) = p(z|x)\ ) && (7)\\
&= \log p(x) && (8)
\end{aligned}$$

This result suggests that using (1) should be preferred when we believe that $q(z|x) \approxeq p(z|x)$.

2 Estimators of the Gradient

What about estimating the *gradient* of the evidence lower bound? In this section, we show that the variance of the gradient of the fully Monte Carlo estimator (1) with respect to the variational parameters is not zero, even when $q(z|x) = p(z|x)$, and when using the reparameterization trick.

Using the reparameterization trick [3], we can say that a sample z is a deterministic function of a random variable, ϵ, with a fixed distribution. This can be written as $z_\phi = f(\epsilon, \phi)$. Then, gradient of the estimator based on (1) with respect to the variational parameters ϕ has the form:

$$\begin{aligned}
\hat{\nabla}_{\text{MC}} &= \nabla_\phi \left[\log p(x|z_\phi) + \log p(z_\phi) - \log q_\phi(z_\phi|x)\right] \qquad \epsilon \overset{\text{iid}}{\sim} \mathcal{N}(0, I) && (9)\\
&= \frac{\partial \log p(x|z_\phi)}{\partial \phi} + \frac{\partial \log p(z_\phi)}{\partial \phi} - \frac{\partial \log q_\phi(z_\phi|x)}{\partial \phi} && (10)\\
&= \underbrace{\frac{\partial \log p(z_\phi|x)}{\partial z_\phi}\frac{\partial z_\phi}{\partial \phi} - \frac{\partial \log q(z_\phi|x)}{\partial z_\phi}\frac{\partial z_\phi}{\partial \phi}}_{\text{path derivative}} - \underbrace{\frac{\partial \log q_\phi(z|x)}{\partial \phi}}_{\text{score function}} && (11)
\end{aligned}$$

The gradient estimate can be broken into two parts. The path derivative measures dependence on ϕ only through the sample $z_\phi = f(\epsilon, \phi)$. The score function measures the depends on $\log q_\phi$ directly, without considering how the sample z changes as a function of ϕ.

When $q(z|x) = p(z|x)$ for all z, the path derivative is identically zero for all z. However, the score function is not necessarily zero for any z, meaning that the above gradient estimator (9) will have non-zero variance even when the q matches the exact posterior everywhere.

3 A Path-Derivative of the ELBO Gradient

Could we get rid of the score function term from the gradient estimate? For stochastic gradient descent to converge, we require that our gradient estimate is unbiased. By construction, the gradient estimate given by (9) is unbiased. Luckily, the score function has zero expectation, meaning that if we simply remove that term, we still have an unbiased gradient estimator:

$$\hat{\nabla}_{\text{PD}} = \frac{\partial \log p(z_\phi|x)}{\partial z_\phi}\frac{\partial z_\phi}{\partial \phi} - \frac{\partial \log q(z_\phi|x)}{\partial z_\phi}\frac{\partial z_\phi}{\partial \phi} \tag{12}$$

This estimator, which we call the path-derivative gradient estimator, is simply the standard gradient estimate with the score function term removed, which has the desirable property that as $q(z|x)$ approaches $p(z|x)$, the variance of this estimator goes to zero. Figure 1 shows the impact that this has on optimization when the variational family contains the true posterior, in a toy example.

4 Practical Implications

When should we prefer the path derivative gradient estimator? Its variance near the optimum depends on the variance of the score function (also known as the Fisher information) of q. We conjecture that complex q distributions, such as those specified by adaptive inference schemes such as normalizing flows [6] or Hamiltonian variational inference [9], will have high Fisher information, increasing as the approximate posterior becomes more detailed. Notably, in these cases the exact entropy and KL

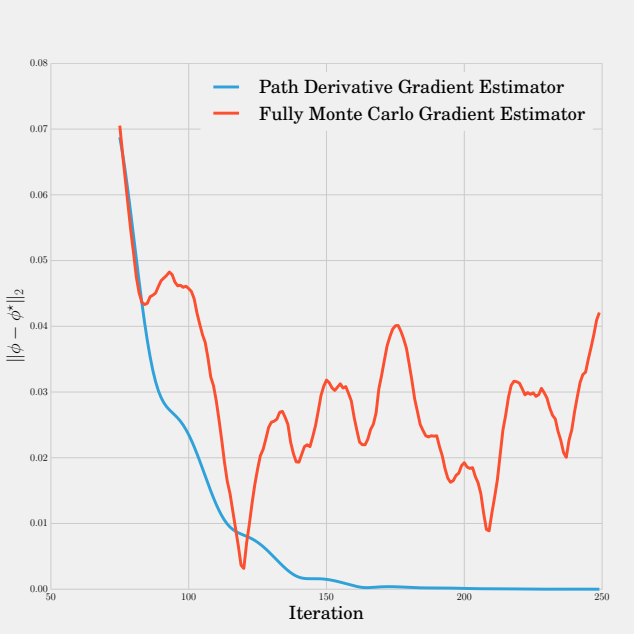


Figure 1: Optimization of variational parameters using the standard gradient estimate versus new gradient estimator. In this toy example, the true posterior and the variational family were both isotropic 2D Gaussians. However, without decaying the optimization step-size, the naïve estimator bounces around the true optimum, while the path-derivative estimator converges quickly and stays at the optimum.

are intractable, so the estimators given by (2) and (3) cannot be used. Relying on stochastic estimates of the KL or exact entropy brings us back into the Fully Monte Carlo gradient paradigm, which our proposed gradient estimator improves.

The variance of the path derivative gradient estimator can be higher in some cases where the variational approximation is far from the true parameters if the control variate is negatively correlated with the path derivative. This has the potential to slow down learning, because removing the score function gradient will increase the variance of the stochastic estimates. To address this problem, we can estimate an adataptive optimal scaling constant c^* as the ratio of the covariance of the two gradient components divided by the variance of the score function [7]. When the variational approximation is exact, we have shown that $c^* = 1$ is optimal. When the variational approximation is not exact, an estimate of c^* based on the current minibatch will change sign and magnitude depending on the positive or negative correlation of the score function with the path derivative. Minibatch estimation of the optimal scale was introduced used by [5] to reduce the high variance of REINFORCE-style gradient estimates.

5 Related Work

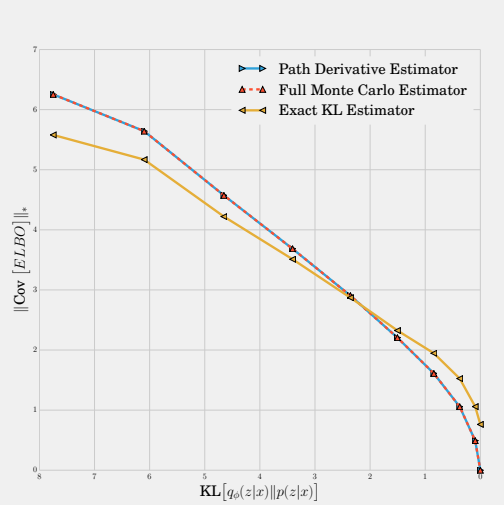
Our modification of the standard reparameterized gradient estimate can also be viewed as adding a control variate, and in fact [5] investigated the use of the score function as a control variate in the context of non-reparameterized variational inference.

The variance-reduction effect we use to motivate our general gradient estimator has been noted in the special cases of Gaussian distributions with sparse precision matrices and Gaussian copula inference in [10] and [2] respectively. In particular, [10] observes that by eliminating certain terms from a gradient estimator for Gaussian families parametrized by sparse precision matrices, two lower-variance unbiased gradient estimators may be derived. Our work is a generalization to any variational family by analyzing the cause of the variance reduction which provides a framework for easily implementing the technique in existing software packages that provide automatic differentiation. This

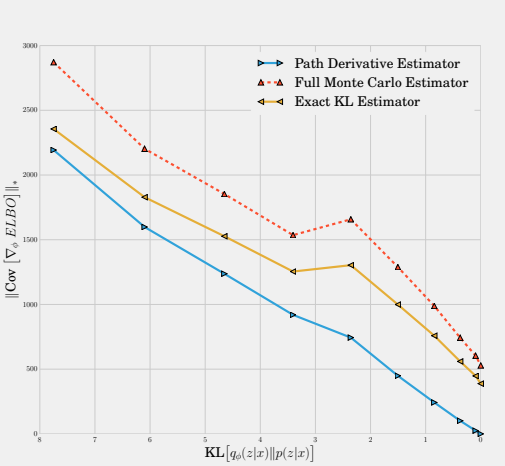
can save time in implementation by eliminating the need for detailed analyses of the functional form of the variational posterior in special cases.

An innovation by Ruiz et al. [8] introduces the generalized reparameterization gradient (GRG) which unifies the REINFORCE-style and reparameterization gradients. This technique uses a reparameterization that allows at most weak dependence on the latent variables, meaning that at least the first moment has no dependence on the latent variables. Their estimator improves on the variance of the score-function gradient estimator in BBVI without the use of Rao-Blackwellization, although a term in their estimator behaves like a control variate. The present study, by contrast, shows a simple control variate for the naive reparameterization gradient that can be easily implemented to improve existing algorithms.

6 Experiments



(a) The variance of different ELBO estimators as a function of KL divergence from the true posterior. The variance of the fully Monte Carlo estimator goes to zero as the divergence goes to zero.



(b) The nuclear norm of the covariance matrix of different ELBO gradient estimators, as a function of KL divergence from the true posterior. The variance of the fully Monte Carlo gradient estimator does not go to zero as the divergence goes to zero, but the variance of the path derivative estimator does.

Figure 2: Variance of ELBO and ELBO gradient estimates on a toy Gaussian example.

We include an empirical analysis of the different ELBO and ELBO gradient estimators in Figure 2. In this experiment, the path derivative gradient estimator dominates the fully Monte Carlo estimator, but presumably has higher variance than the exact KL estimator when the approximate posterior is far from the exact posterior.

7 Future Work

We plan to examine the performance of our new gradient estimator on representative problems, and examine the empirical Fisher information of complex approximating distributions. We also plan to examine the properties of the path derivative gradient estimator in the multi-sample setting, relating it to both importance-weighted autoencoders (IWAE) [1], and the variational inference for Monte Carlo objectives (VIMCO) [4] gradient estimator.

References

[1] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *arXiv preprint arXiv:1509.00519*, 2015.

[2] Shaobo Han, Xuejun Liao, David B Dunson, and Lawrence Carin. Variational gaussian copula inference. *arXiv preprint arXiv:1506.05860*, 2015.

[3] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[4] Andriy Mnih and Danilo J Rezende. Variational inference for monte carlo objectives. *arXiv preprint arXiv:1602.06725*, 2016.

[5] Rajesh Ranganath, Sean Gerrish, and David M Blei. Black box variational inference. In *AISTATS*, pages 814–822, 2014.

[6] Danilo Rezende and Shakir Mohamed. Variational inference with normalizing flows. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1530–1538, 2015.

[7] Sheldon Ross. Chapter 9 - variance reduction techniques. In Sheldon Ross, editor, *Simulation (Fifth Edition)*, pages 153 – 231. Academic Press, fifth edition edition, 2013.

[8] Francisco JR Ruiz, Michalis K Titsias, and David M Blei. The generalized reparameterization gradient. *arXiv preprint arXiv:1610.02287*, 2016.

[9] Tim Salimans, Diederik Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 1218–1226, 2015.

[10] Linda SL Tan and David J Nott. Gaussian variational approximation with sparse precision matrix. *arXiv preprint arXiv:1605.05622*, 2016.