# Reparameterization

*Posted on January 6, 2020 by Praveen Narayanan*

In this note, we take a look at the reparameterization trick, an idea forms the basis of the Variational Autoencoder. Nonetheless, my notes here come from the fantastic paper by Ruiz et al [1]. The main idea is that the reparameterization trick [4,5] gives us a lower variance estimator than that obtained from the score function gradient, but suffers because the class of distributions to which it can be applied is somewhat limited – Kingma and Welling discusses Bernaulli and Gaussian variants. Nevertheless, it is possible to remedy this defect, as is done in papers [1], [2], [3]. The paper by Ruiz is especially instructive. It walks us through the machinery involved in reparameterization and discusses variance reduction [Casella and Berger, Robert and Casella] for variational inference – also [BBVI]. I was originally intending on writing a much more complete summary of the papers [1], [2], [3] but ran out of juice, leaving it for the future as might transpire (or not). If I may insert inappropriately (also, perhaps to acknowledge a decade that just ended):

*What might have been is an abstraction*
*Remaining a perpetual possibility*
*Only in a world of speculation.*
*What might have been and what has been*
*Point to one end, which is always present.*

Notwithstanding such tangential musings, it behooves us to study [Casella and Berger]. The authors note that it is a 22 month long course to learn statistics the hard way. Incidentally, the style of books that have connections with George Casella (also [Robert and Casella], [Robert]) very much reminds me of [Bender and Orszag], with engaging quotes from Holmes and their straight forward slant on equations.

# The Variational Objective

Recall that in variational problems we are interested in obtaining the ELBO. We briefly derive this using Jensen's inequality ($\log E_q(.) \geq E_q \log(.)$)

Take the joint form presented in Ruiz:

$$p(x, z) = \frac{p(x, z)}{q(z; v)} q(z; v)$$

Or

$$\begin{aligned} \log \int p(x, z) dz &= \log \int \frac{p(x, z)}{q(z; v)} q(z; v) dz \\ &\geq \int \log \frac{p(x, z)}{q(z; v)} q(z; v) dz \end{aligned}$$

Or

$$\mathcal{L}_v = E_{q(z;v)}[\log p(x,z) - \log q(z;v)] = E_{q(z;v)}[f(z)] + H(q(z;v))$$

The equation written in this form is quite instructive. We want to optimize the joint $p(x, z)$ by varying the variational parameters v through the surrogate distribution $q$.

# Gradient estimation with score function

Our aim is to obtain Monte Carlo estimates of the gradient (by taking expectations), but this is not possible as is. However, it is possible to arrange this as follows:

$$\int \nabla_v q(z;v) f(z) dz = \int q(z;v) \nabla_v \log q(z;v) f(z) dz = E_q(z;v)[f(z) \log q(z;v)]$$

This is known as the score function estimator, or the REINFORCE gradient. When we view it as a discrete gradient, allowing us to take gradients of non-differentiable functions by taking samples.

$$E_q(z;v)[f(z) \log q(z;v)] = \frac{1}{L} \sum_{l=1}^{L} [f(z^l) \log q(z^l;v)]$$

However, the estimates produced tend to be noisy, and needs provisos for variance reduction – e.g. Rao-Black Wellization, control variates.

# Gradient of expectation, expectation of gradient

A principal contribution of the VAE approach is that we have an alternative way to derive the estimator, that is generally of lower variance than the score function method described above. That being said, it has the drawback that this method is not as widely applicable as the score function approach.

Recall that we would like to take gradients of the term containing the log joint in the ELBO.

$$\nabla_v \mathcal{L} = \nabla_v E_{q(z;v)}[f(z)] = \nabla_v \int q(z;v) f(z) dz + \cdots$$

In this equation, we can take samples $f(z^l)$, but as the estimator contains variational parameters $v$ within it, we cannot carry out any sort of diffentiation operations to it with respect to $v$– necessary to take gradients. The reparameterization trick gets around this problem.

We derive an alternative estimator by transforming to a distribution that does not depend on $v$ so that we can now take the gradient operator inside the expectation. For this to work, we rely on what they call a 'standardization' operation to transform $q(z;v)$ into another distribution $q(\epsilon)$ independent of $v$ (and other terms containing $v$). In the end, we want to have something like this:

$$\nabla_v E_q(z;v) f(z) = \nabla_v E_{q(\epsilon)} f(\tau(\epsilon;v)) g(v;\epsilon) = E_{q(\epsilon)} \nabla_v (\cdots)$$

We have pushed the gradient operator inside the expectation which allows us to take samples and allow taking gradients from it.

That's a lot of words. Let us derive this estimator to put it more concretely.

We assume that there exists an invertable transformation $z = \tau(\epsilon; v), \epsilon = \tau^{-1}(z; v)$ with pdfs $q(z; v), q_\epsilon(\epsilon; v)$. In some cases, it is possible to find a transformation such that the reparameterized distribution is independent of the variational parameters $v$. For example, consider the standard normal distribution:

$$q(\epsilon) = \frac{1}{\sqrt{2\pi}} e^{-\epsilon^2/2}$$

We can consider this as the standardized version of a normal distribution $N(\mu, \sigma)$.

$$\begin{aligned} z = \tau(\epsilon; \mu, \sigma) &= \mu + \epsilon\sigma \\ \epsilon &= \frac{z - \mu}{\sigma} \end{aligned}$$

Transform (pretend 1D for now):

$$\begin{aligned} q(z) = q_\epsilon(\epsilon)|\frac{d\epsilon}{dz}| &= \frac{1}{\sigma} q_\epsilon(\epsilon) \\ &= \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{z-\mu}{\sigma})^2} = N(\mu, \sigma) \end{aligned}$$

For more general cases, the differential is replaced by a Jacobian:

$$\begin{aligned} J(\epsilon; v) &= |\det \nabla_\epsilon \tau(\epsilon; v)| \\ q(\epsilon; v) &= q(\tau(\epsilon; v)) J(\epsilon; v) \end{aligned}$$

After standardization, we lose dependence on $v$: $q(\epsilon; v) = q_\epsilon(\epsilon)$ so that

$$\begin{aligned} \int q(z; v) f(z) dz &= \int q_\epsilon(\epsilon)) J f(\tau(\epsilon; v)) d\epsilon J^{-1} \\ &= \int q_\epsilon(\epsilon) f(\tau(\epsilon; v)) d\epsilon \end{aligned}$$

Now we can take gradient and move it inside the expectation:

$$\begin{aligned} \nabla_v \int q(z; v) f(z) dz &= \nabla_v \int q_\epsilon(\epsilon) f(\tau(\epsilon; v)) d\epsilon \\ &= \int q_\epsilon(\epsilon) \nabla_v f(\tau(\epsilon; v)) d\epsilon \\ &= E_{q(\epsilon)} \nabla_v f(\tau(\epsilon; v)) \end{aligned}$$

# Reparameterization in more general cases

The standardization procedure is now extended so that $\epsilon$ is weakly dependent on $v$ – it has zero mean, but it's first moment does not depend on $v$. Nevertheless, $q_\epsilon(\epsilon; v)$ has dependence on the variational parameters. In this case, the expectation will have to be evaluated term by term with chain rule

$$\nabla_v E_{q(z;v)}[f(z)] = \nabla_v E_{q_\epsilon(\epsilon;v)}[f(\tau(\epsilon; v))] = \nabla_v \int q_\epsilon(\epsilon; v) f(\tau(\epsilon; v)) d\epsilon$$

$$\nabla_v E_{q(z;v)}[f(z)] = \int q_\epsilon(\epsilon; v) \nabla_v f(\tau(\epsilon; v)) d\epsilon + \int q_\epsilon(\epsilon; v) f(\tau(\epsilon; v)) \nabla_v \log q_\epsilon(\epsilon; v) d\epsilon$$

As we can see, the first term is the regular reparameterization gradient. The second term is the score function estimator, a correction term for this version of this standardization setup.

$$g^{rep} = E_{q_\epsilon(\epsilon;v)} \nabla_v f(\tau(\epsilon; v))$$
$$g^{corr} = E_{q_\epsilon(\epsilon;v)} f(\tau(\epsilon; v)) \nabla_v \log q_\epsilon(\epsilon; v)$$
$$\mathcal{L}_v = g^{rep} + g^{corr} + \nabla_v H[q(z; v)]$$

In the case of the normal distribution, the second term $g^{corr}$ vanishes.

# Interpretation

The terms are massaged so as to look like control variates, an idea used in Monte Carlo variance reduction. The authors note that while Rao-Blackwellization is not used in the paper, it is perfectly reasonable to use the setup in conjuction with it, as is done in Black Box Variational Inference [BBVI], where both Rao-Blackwellization and control variates is used to reduce the variance of the estimator.

The basic idea of control variates is as follows ([Casella and Berger] – Chapter 7 on "Point Estimation"). Given an estimator $W$ satisfying $E_\theta W = \tau(\theta)$, we seek to find another estimator $\phi_a$ of lower variance, using an estimator $U$ with $E_\theta U = 0$:

$$\phi_a = W + aU$$

The variance for this estimator is (Casella and Berger):

$$Var_\theta \phi_a = Var_\theta(W + aU) = Var_\theta W + 2aCov_\theta(W, U) + a^2 Var_\theta U$$

We would get lower variance for $\phi_a$ than $W$ if we could find $a$ such that $2aCov_\theta(W, U) + a^2 Var_\theta U < 0$.

To get back to our variational estimator, rewrite as follows for it to be interpretable as control variates (see [1]):

$$\begin{aligned}\nabla_v E_{q(z;v)}[f(z)] &= E_{q(z;v)}[f(z) \nabla_v \log q(z; v)] \\ &+ E_{q(z;v)}[\nabla_z f(z) h(\tau^{-1}(z; v); v)] \\ &+ E_{q(z;v)}[f(z)(\nabla_z \log q(z; v) h(\tau^{-1}(z; v); v) + u(\tau^{-1}(z; v); v))]\end{aligned}$$

In the first line, we have the score function expression, which is modified in subsequent lines. It is not entirely clear to me how the expectation of the terms that correct the noisy gradient is zero, but I suppose we will take it in the spirit with which it was intended.

# References

[1] Ruiz et al: The generalized reparameterization gradient: https://arxiv.org/abs/1610.02287

[2] Naesseth et al: Reparameterization Gradients through Acceptance-Rejection Sampling Algorithms: https://arxiv.org/abs/1610.05683

[3] Figurnov et al: Implicit Reparameterization Gradients: https://arxiv.org/abs/1805.08498

[4] Kingma and Welling: Autoencoding Variational Bayes: https://arxiv.org/abs/1312.6114

[5] Rezende, Mohamed and Wierstra: Stochastic Backpropagation and Approximate Inference in Deep Generative Models: https://arxiv.org/abs/1401.4082

[BBVI] Black Box Variational Inference: https://arxiv.org/abs/1401.0118

[Casella and Berger]: Statistical Inference: https://www.amazon.com/Statistical-Inference-George-Casella/dp/0534243126

[Robert and Casella]: Monte Carlo Statistical Methods: https://www.amazon.com/Monte-Statistical-Methods-Springer-Statistics/dp/1441919392

[Robert]: The Bayesian Choice: https://www.amazon.com/Bayesian-Choice-Decision-Theoretic-Computational-Implementation/dp/0387715983

[Bender and Orszag]: Advanced Mathematical Methods for Scientists and Engineers: https://www.amazon.com/Advanced-Mathematical-Methods-Scientists-Engineers/dp/0387989315