Variance reduction properties of the reparameterization trick

Ming Xu[†] **Matias Quiroz*** **Robert Kohn*** **Scott A. Sisson**[†]

[†]School of Mathematics and Statistics
University of New South Wales

*UNSW Business School, School of Economics
University of New South Wales

Abstract

The reparameterization trick is widely used in variational inference as it yields more accurate estimates of the gradient of the variational objective than alternative approaches such as the score function method. Although there is overwhelming empirical evidence in the literature showing its success, there is relatively little research exploring why the reparameterization trick is so effective. We explore this under the idealized assumptions that the variational approximation is a mean-field Gaussian density and that the log of the joint density of the model parameters and the data is a quadratic function that depends on the variational mean. From this, we show that the marginal variances of the reparameterization gradient estimator are smaller than those of the score function gradient estimator. We apply the result of our idealized analysis to real-world examples.

1 INTRODUCTION

Background Variational inference (VI) (Jordan et al., 1999; Ormerod and Wand, 2010; Blei et al., 2017) provides a fast and approximate alternative to exact Monte Carlo methods when performing Bayesian inference on parameters in complex statistical models. The idea of VI is to approximate the posterior density with a family of tractable densities, indexed by variational parameters, where a member of that family is referred to as a variational approximation. VI then proceeds by finding a set of variational parameters such that the variational approximation is close to the true posterior density in some sense. In machine learning, VI has been used in generative models through variational autoencoders (Kingma and Welling, 2014). In econometrics and statistics, complex regression density estimation (Nott et al., 2012), state space models (Tan and Nott, 2018), and high-dimensional time-varying parameter models (Quiroz et al., 2018) are approximated using VI. Furthermore, VI has recently been extended to cases where the likelihood is intractable (Tran et al., 2017; Ong et al., 2018a). Complex variational families have been proposed, e.g. Gaussian mixtures to account for multi-modality (Zobay, 2014; Miller et al., 2016) and Gaussian copulas (Han et al., 2016) for flexible multivariate modeling.

VI formulates the problem of approximating a probability density as an optimization problem. To implement the optimization efficiently, it is crucial to obtain an accurate estimate of the gradient when the function to be optimized is intractable but can be unbiasedly estimated. To this end, the reparameterization (RP) trick (Kingma and Welling, 2014; Rezende et al., 2014) has been useful and much more efficient than the original score function method (Williams, 1992). There is now a large literature applying the RP trick successfully in different settings and recently it has been extended to a wider range of variational approximations (Ruiz et al., 2016; Figurnov et al., 2018) and even for non-differentiable models (Lee et al., 2018). Remarkably, despite the abundance of research utilizing the RP trick, its variance reduction properties are not well studied, apart from a few exceptions, which we review in Section 3.8.

General Framework We compare the RP trick to the score function method and show that the former yields more efficient gradient estimators under certain simplifying assumptions. Our first main assumption is that the variational approximation is a mean-field Gaussian density, which is a common modelling choice that has been successfully used in many challenging applications (Kingma and Welling, 2014; Rezende et al., 2014; Kucukelbir et al., 2017, among others). Our second main assumption is that the log-joint density of the model parameters and the data

Proceedings of the 22$^{\text{nd}}$ International Conference on Artificial Intelligence and Statistics (AISTATS) 2019, Naha, Okinawa, Japan. PMLR: Volume 89. Copyright 2019 by the author(s).

is a quadratic function that varies with the variational mean. We refer to this function as the log-joint density for simplicity. For any general log-joint density, applying this assumption is the same as approximating the true log-joint density with its second-order Taylor series expansion around the variational mean.

These assumptions allow us to derive expressions for the marginal variances of the gradient estimators under the score function method and RP trick. We then show that the RP gradient estimator is more efficient than the score function estimator since it yields lower marginal variances. This is done by finding a lower bound on the score function marginal variance through applying Rao-Blackwellization. Finally, these expressions are used to understand why and when the RP trick is more efficient.

Contribution Our contribution is to both prove and understand why the RP trick yields more efficient gradients than the score function method under the simplifying assumptions above. We conclude that:

- The score function method yields an estimator containing higher order powers of θ than that of the RP trick, resulting in the score function estimator "varying" more over its "sampling region". Section 3.7 elaborates further and illustrates this with a simple example.
- The marginal variance of each element in both the score function and RP gradient increases with the local "curvature" of the log-joint density around the variational mean. Furthermore, the marginal variances under the score function method tends to be smaller when the variational mean is close to the true posterior mode. This does not occur under the RP trick.
- The marginal variances of the gradient under the score function method increase as the variational scale parameters tend to 0 unlike the RP trick.
- Section 3.6 discusses other fundamental differences between the gradients.

2 STOCHASTIC GRADIENT VARIATIONAL INFERENCE

2.1 The variational lower bound

Let $y = \{y_1, \ldots, y_n\}$ denote a dataset with n observations, where $y_i \in \mathcal{X} \subseteq \mathbb{R}^l$ for all i. Given a model parameterized by $\theta \in \Theta \subseteq \mathbb{R}^k$, with prior density $p(\theta)$, the posterior density is

$$p(\theta|y) = p(y|\theta)p(\theta)/p(y), \tag{1}$$

where $p(y|\theta)$ denotes the model likelihood, $p(y) = \int_{\theta\in\Theta} p(y,\theta)d\theta$ is the marginal likelihood or evidence and $p(y,\theta) = p(y|\theta)p(\theta)$ is the joint density of y and θ. Bayesian inference generally involves computing expectations of functions of θ with respect to (1) which usually does not belong to a known family of densities.

The goal of VI is to approximate the posterior density in (1) by using an appropriate approximating family of variational densities $q(\theta;\lambda)$, where $\lambda = \{\lambda_1, \ldots, \lambda_p\}$ are the variational parameters with $\lambda_i \in \Lambda_i \subset \mathbb{R}^{p_{\lambda_i}}$ where p_{λ_i} is the number of variational parameters in parameter set i and p is the number of parameter sets in the variational approximation. For example, if $q(\theta;\lambda)$ is Gaussian, then $p = 2$, where $\lambda_1 \in \mathbb{R}^k$ is the mean and $\lambda_2 \in \mathbb{R}^{k(k+1)/2}$ are the unique elements of the covariance matrix. VI finds the optimal λ by minimizing the Kullback-Leibler (KL) divergence between the approximation and the true posterior density,

$$\begin{aligned}\mathrm{KL}(q(\theta;\lambda)\|p(\theta|y)) &= \int_{\theta\in\Theta} q(\theta;\lambda)\log\frac{q(\theta;\lambda)}{p(\theta|y)}d\theta \\ &= \mathbb{E}_q[\log q(\theta;\lambda) - \log p(\theta|y)],\end{aligned} \tag{2}$$

where $\mathbb{E}_q[\cdot]$ denotes expectation with respect to density $q(\cdot)$. The KL divergence is non-negative and is zero if and only if $q(\theta;\lambda) = p(\theta|y)$. Computing (2) requires evaluating $p(y)$, which is typically intractable. A tractable approach is obtained by maximizing an alternative objective function, which is equivalent to minimizing the KL divergence. We have that

$$\log p(y) = \mathcal{L}(\lambda) + \mathrm{KL}(q(\theta;\lambda)\|p(\theta|y)), \tag{3}$$

where

$$\begin{aligned}\mathcal{L}(\lambda) &= \int \log\left(\frac{p(y,\theta)}{q(\theta;\lambda)}\right) q(\theta;\lambda)\, d\theta \\ &= \mathbb{E}_q[h(\theta) - \log q(\theta;\lambda)],\end{aligned} \tag{4}$$

is referred to as the evidence lower bound (ELBO) because $\log p(y) \geq \mathcal{L}(\lambda)$ and $h(\theta) = \log p(y,\theta)$. Eq. (3) shows that minimizing the KL divergence is equivalent to maximizing the ELBO in (4), which does not require evaluating $p(y)$.

2.2 Stochastic gradient optimization

The gradient of the ELBO in (4) is rarely available in closed form. Stochastic gradient methods (Robbins and Monro, 1951; Bottou, 2010) are useful for optimizing an objective function whose gradient can be unbiasedly estimated. Let $\nabla_\lambda\mathcal{L}(\lambda)$ be the gradient vector of $\mathcal{L}(\lambda)$ in (4) with respect to λ. There are numerous ways to represent this gradient, each one giving a specific estimator: see Roeder et al. (2017) for some choices. We use the following representation

$$\nabla_\lambda\mathcal{L}(\lambda) = \nabla_\lambda\mathbb{E}_q[h(\theta)] + \nabla_\lambda\mathbb{H}_q[q(\theta;\lambda)], \tag{5}$$

where $\mathbb{H}_q[q(\theta;\lambda)] = -\mathbb{E}_q[\log q(\theta;\lambda)]$ is the entropy of q and is analytically solvable when the variational density is Gaussian (Assumption 1). In the rest of the article, whenever the entropy term appears in an estimator, it is evaluated explicitly.

Let $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}$ be an unbiased estimator of the gradient which we obtain by Monte Carlo simulation as follows. Suppose that the first term of $\nabla_\lambda \mathcal{L}(\lambda)$ in (5) may be written as an expectation of a function $\Delta(\theta;\lambda)$ with respect to a density $g(\theta;\lambda)$. Then, providing that sampling from $g(\theta;\lambda)$ is possible, an unbiased estimate of $\nabla_\lambda \mathcal{L}(\lambda)$ can be constructed through

$$\widehat{\nabla_\lambda \mathcal{L}(\lambda)} = \frac{1}{S}\sum_{s=1}^{S} \Delta(\theta^{(s)};\lambda) + \nabla_\lambda \mathbb{H}_q[q(\theta;\lambda)], \qquad (6)$$
$$\theta^{(s)} \sim g(\theta;\lambda),\ s = 1,\ldots,S.$$

Now, starting from $\lambda = \lambda^{(0)}$, the iteration

$$\lambda^{(t+1)} = \lambda^{(t)} + \eta_t \circ \widehat{\nabla_\lambda \mathcal{L}(\lambda^{(t)})} \qquad (7)$$

may be performed until some convergence criteria on $\mathcal{L}(\lambda)$ is met, where the vector η_t is a sequence of learning rates and $\circ$ denotes the Hadamard product (element-wise multiplication). Under certain regularity conditions, and when the learning rates satisfy the Robbins-Monro conditions

$$\sum_{t=0}^{\infty} \eta_t = \infty \quad \text{and} \quad \sum_{t=0}^{\infty} \eta_t^2 < \infty,$$

the iterates converge to a local optimum (Robbins and Monro, 1951). Adaptive learning rates are currently popular (Duchi et al., 2011; Zeiler, 2012; Kingma and Ba, 2015) and we use Adam (Kingma and Ba, 2015) in our empirical examples, but this choice does not affect our results or conclusions.

The efficiency of the optimization when iterating (7), i.e. how fast it converges, depends on how one expresses $\nabla_\lambda \mathcal{L}(\lambda)$; different parametrizations (different Δ and/or g), give rise to different estimators, all of which are unbiased but may have very different variances. For each parameterization, the accuracy of the estimator in (6) also depends on the number of Monte Carlo samples S. Our article considers three gradient estimators: the RP gradient (Kingma and Welling, 2014; Rezende et al., 2014), the score function gradient (Williams, 1992) and a Rao-Blackwellized version of the score function gradient (Ranganath et al., 2014). The Rao-Blackwellization is used to derive lower bounds for the marginal variances of the score function gradient. Under our assumptions we show that the marginal variances of the RP gradient are less than or equal to the Rao-Blackwellized score function gradient. It trivially follows that the trace of the covariance matrix is smaller for the RP gradient, explaining its superiority over the score function gradient.

3 FRAMEWORK

3.1 Structure of the variational approximation

Assumption 1. *The variational approximation is* $q(\theta;\lambda) = \mathcal{N}(\theta|\mu,\Sigma)$*, with* $\mu = (\mu_1,\ldots,\mu_k)^\top$ *and* $\Sigma = \mathrm{diag}(\exp(2\phi_1),\ldots,\exp(2\phi_k))$, $\phi_i = \log(\sigma_i), \sigma_i = \Sigma_{ii}^{1/2}$*, where* $\mathcal{N}(\cdot|\mu,\Sigma)$ *denotes the Gaussian density with mean vector* μ *and (diagonal) covariance matrix* Σ*.*

Assumption 1 implies an independence structure known as a mean-field approximation and has been extensively used in conjuction with stochastic gradient methods (Kingma and Welling, 2014; Rezende et al., 2014; Kucukelbir et al., 2017, among others). Under this assumption, the variational density takes the form

$$q(\theta;\lambda) = \prod_{i=1}^{k} \mathcal{N}(\theta_i|\mu_i, \exp(2\phi_i)), \qquad (8)$$

with variational parameters $\mu = (\mu_1,\ldots,\mu_k)^\top$ and $\phi = (\phi_1,\ldots,\phi_k)^\top$, and the vector of all variational parameters is $\lambda = (\mu^\top,\phi^\top)^\top$. There are two reasons we use ϕ_i instead of σ_i: the optimization is easier as it is unrestricted and, moreover, Assumption 2 in the next subsection becomes more plausible.

3.2 Comparing gradient estimators

The gradient of $\mathcal{L}(\lambda)$ is partitioned as

$$\nabla_\lambda \mathcal{L}(\lambda) = (\nabla_\mu \mathcal{L}(\lambda)^\top, \nabla_\phi \mathcal{L}(\lambda)^\top)^\top,$$

with its estimator

$$\widehat{\nabla_\lambda \mathcal{L}(\lambda)} = \left(\widehat{\nabla_\mu \mathcal{L}(\lambda)}^\top, \widehat{\nabla_\phi \mathcal{L}(\lambda)}^\top\right)^\top, \qquad (9)$$

where $\lambda = (\mu^\top,\phi^\top)^\top \in \mathbb{R}^{2k}$ contains all of the variational parameters. The Central Limit Theorem (CLT) motivates the next assumption. Recall that the entropy term $\mathbb{H}_q[q(\theta;\lambda)]$ is assumed known.

Assumption 2. *Let*

$$\widehat{\nabla_\lambda \mathcal{L}(\lambda)} = \frac{1}{S}\sum_{s=1}^{S} \Delta(\theta^{(s)};\lambda) + \nabla_\lambda \mathbb{H}_q[q(\theta;\lambda)],$$
$$\theta^{(s)} \overset{\text{iid}}{\sim} g(\theta;\lambda), \quad \theta \in \mathbb{R}^k, \quad (10)$$

where $\Delta : \mathbb{R}^k \to \mathbb{R}^{2k}$ *and* $g(\theta;\lambda)$ *is any density. We assume that for each* $j = 1,\ldots,2k$*,*

$$\widehat{\nabla_\lambda \mathcal{L}(\lambda)}_j \sim \mathcal{N}\left(\nabla_\lambda \mathcal{L}(\lambda)_j, \frac{1}{S}\mathbb{V}_g\left(\Delta_j(\theta;\lambda)\right)\right),$$

where $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}_j$ *and* $\nabla_\lambda \mathcal{L}(\lambda)_j$ *are the* j*-th elements of the corresponding vectors, and* $\Delta_j(\theta;\lambda)$ *denotes the* j*-th element of* $\Delta(\theta;\lambda)$*.*

The CLT approximately holds even for small values of S due to independent sampling from g. We have found empirically that the transformation $\phi_i = \log(\sigma_i)$, $i = 1,\dots,k$, makes Assumption 2 more plausible in practice since it corrects for skewness.

Assumption 2 allows us to only consider the marginal variances when comparing unbiased estimators for the j-th element obtained with different Δ-functions. Balles and Hennig (2018) also consider only the marginal variances when studying the effect of the variability of the stochastic gradient on the Adam optimizer (Kingma and Ba, 2015).

To compare the efficiency of the full (vector) gradient estimator, we follow Miller et al. (2017) and consider the trace metric, which is the trace of the estimator covariance matrix, as a scalar measure of variability. This is justified by Assumption 2 and allows us to establish analytical results. Under our assumptions, this metric is smaller for the RP gradient compared to the score function gradient. Alternative scalar metrics which capture dependencies between gradient components exist. Roeder et al. (2017) use the nuclear norm of the estimator covariance matrix. Another metric is the generalized variance (Wilks, 1932), defined as the determinant of the estimator covariance matrix. However, these metrics are analytically intractable under our assumptions. Furthermore, they rely on the multivariate CLT because the covariance matrix is only useful for comparing variability between multivariate Gaussian random variables. For high-dimensional λ, the multivariate CLT requires a prohibitively large S and is therefore not appropriate in practice.

3.3 Gradient estimators

The RP trick assumes that $\theta \sim q(\theta;\lambda)$ can be written as $\theta = T(z;\lambda)$, $T : \mathbb{R}^k \to \mathbb{R}^k$, where z is a random vector (with the same dimension as θ) with density $f(z)$ which does not depend on the variational parameters λ. This describes a generative model for θ in terms of the variational parameters. For example, when $q(\theta;\lambda) \sim \mathcal{N}(\mu, \mathrm{diag}(\exp(2\phi)))$, then $T(z;\lambda) = \mu + \exp(\phi) \circ z$ with $z \sim \mathcal{N}(0, I)$, where I is the $k \times k$ identity matrix and the exponential function is applied element-wise. The gradient of the ELBO under reparameterization becomes

$$\nabla_\lambda \mathbb{E}_q[h(\theta)] = \mathbb{E}_f[\nabla_\lambda T(z;\lambda)\nabla_\theta h(\theta)|_{\theta=T(z;\lambda)}], \quad (11)$$

where $\Delta^{\mathrm{RP}}(z;\lambda) = \nabla_\lambda T(z;\lambda)\nabla_\theta h(\theta)|_{\theta=T(z;\lambda)}$ and $h : \mathbb{R}^k \to \mathbb{R}$, using RP to emphasize that it is the Δ-function in (6) (now a function of z) for the RP trick. The gradient of the ELBO under the RP trick is

$$\begin{aligned}\nabla_\lambda \mathcal{L}(\lambda)_{\mathrm{RP}} = \mathbb{E}_f[\nabla_\lambda T(z;\lambda)\nabla_\theta h(\theta)|_{\theta=T(z;\lambda)}] \\ + \nabla_\lambda \mathbb{H}_q[q(\theta;\lambda)], \quad (12)\end{aligned}$$

and an unbiased estimate is obtained by

$$\begin{aligned}\widehat{\nabla_\lambda \mathcal{L}(\lambda)}_{\mathrm{RP}} = \frac{1}{S}\sum_{s=1}^{S} \Delta^{\mathrm{RP}}(z^{(s)};\lambda) + \nabla_\lambda \mathbb{H}_q[q(\theta;\lambda)], \\ z^{(s)} \sim f(z),\ s = 1,\dots,S. \quad (13)\end{aligned}$$

The score function method, also known as the log-derivative trick or the REINFORCE algorithm (Williams, 1992), expresses the gradient of the first term in (5) as

$$\nabla_\lambda \mathbb{E}_q[h(\theta)] = \mathbb{E}_q[h(\theta)\nabla_\lambda \log q(\theta;\lambda)].$$

For this estimator, the Δ-function in (6) is $\Delta^{\mathrm{score}}(\theta;\lambda) = h(\theta)\nabla_\lambda \log q(\theta;\lambda)$. The gradient of the ELBO under the score function method is

$$\begin{aligned}\nabla_\lambda \mathcal{L}(\lambda)_{\mathrm{score}} = \mathbb{E}_q[h(\theta)\nabla_\lambda \log q(\theta;\lambda)] \\ + \nabla_\lambda \mathbb{H}_q[q(\theta;\lambda)], \quad (14)\end{aligned}$$

and an unbiased estimate is obtained by

$$\begin{aligned}\widehat{\nabla_\lambda \mathcal{L}(\lambda)}_{\mathrm{score}} = \frac{1}{S}\sum_{s=1}^{S} \Delta^{\mathrm{score}}(\theta^{(s)};\lambda) + \nabla_\lambda \mathbb{H}_q[q(\theta;\lambda)], \\ \theta^{(s)} \sim q(\theta;\lambda),\ s = 1,\dots,S. \quad (15)\end{aligned}$$

We use a Rao-Blackwellized score function gradient estimator introduced by Ranganath et al. (2014) to find a lower bound for the marginal variances of the score function estimator and show that the corresponding variances under the RP gradient are smaller. To implement the Rao-Blackwellization (RB), suppose that the variational approximation satisfies Assumption 1 and define $h_{-i}(\theta)$ to be $h(\theta)$ with any elements not containing θ_i removed. Furthermore, denote the Markov blanket of the i-th parameter as $\theta_{(i)}$, see Section A of the supplementary material for details. The gradient in (11) may be written as an iterated conditional expectation, which for $i = 1,\dots,k$, simplifies to

$$\begin{aligned}\nabla_{(\mu_i,\phi_i)}\mathbb{E}_q[h(\theta)] = \mathbb{E}_{q_{(i)}}[h_{-i}(\theta_{(i)}) \\ \nabla_{(\mu_i,\phi_i)} \log q(\theta_i;\mu_i,\phi_i)], \quad (16)\end{aligned}$$

where $q_{(i)}$ is the density of $\theta_{(i)}$. Hence, we define $\Delta^{\mathrm{RB}}(\theta_{(i)};\lambda) = h_{-i}(\theta_{(i)})\nabla_{(\mu_i,\phi_i)} \log q(\theta_i;\mu_i,\phi_i)$ and form the Rao-Blackwellized gradient estimator for the i-th component as

$$\begin{aligned}\widehat{\nabla_{(\mu_i,\phi_i)}\mathcal{L}(\lambda)}_{\mathrm{RB}} = \frac{1}{S}\sum_{s=1}^{S} \Delta^{\mathrm{RB}}(\theta^{(s)};\lambda) + \nabla_{(\mu_i,\phi_i)}\mathbb{H}_{q_{(i)}}[\\ q(\theta;\mu_i,\phi_i)], \quad \theta^{(s)} \sim q_{(i)}(\theta;\mu_i,\phi_i),\ s = 1,\dots,S. \quad (17)\end{aligned}$$

The full estimator, i.e. $\widehat{\nabla_\lambda \mathcal{L}(\lambda)}_{\text{RB}}$, is obtained by merging (17) for $i = 1, \ldots, k$ and ordering them as $\lambda = (\mu^\top, \phi^\top)^\top$. For details and a full derivation, see Ranganath et al. (2014) and Section A of the supplementary material.

Table 1: $\mathbb{V}_q\left[\Delta^{\text{score}}(\theta;\lambda)\right]$ and $\mathbb{V}_f\left[\Delta^{\text{RP}}(z;\lambda)\right]$ estimated using $S = 10,000$ samples. The approximations deteriorate as $\sigma = \exp(\phi)$ increases. True refers to the using the true log-joint density, and approx. refers to replacing $h(\theta)$ with the quadratic approximation.

$\widehat{\nabla\mathcal{L}(\lambda)}$ \ σ	(0.1, 0.1)	(0.5, 0.5)	(1, 1)	(2,2)
Score (true)	32,459	1,648	439	229
Score (approx.)	32,459	1,659	473	369
RP (true)	0.06	1.40	3.56	7.76
RP (approx.)	0.06	1.64	5.69	24.05

3.4 Structure of the log-joint density

We now present an assumption that allows us to (i) obtain analytical expressions for the marginal variances of the score function and RP gradient estimators and (ii) understand how the RP trick reduces the variance.

Assumption 3. *Let $\mu = (\mu_1, \ldots, \mu_k)^\top$ be the variational mean and suppose that the log-joint density $h(\theta) = \log p(y, \theta)$ is given by*

$$h(\theta) = C + G(\mu)^\top(\theta - \mu) + \frac{1}{2}(\theta-\mu)^\top H(\mu)(\theta - \mu), \quad (18)$$

where C is a constant, $G(\mu)$ is a vector whose entries are functions of μ and $H(\mu)$ is a symmetric matrix.

We refer to Assumption 3 as the quadratic assumption on the log-joint density. We can liken this to a second-order Taylor series expansion of any general log-joint density around the variational mean. In this case, $G(\mu) = \nabla_\theta h(\mu)$ and $H(\mu)$ is the hessian of $h(\theta)$ evaluated at μ.

The plausibility of Assumption 3 depends on how far the sampled values of θ are from μ when evaluating the Monte Carlo gradients and to what degree the true $h(\theta)$ is quadratic in this region. We would expect that as ϕ increases, more samples lie in a region where the approximation is poor and so the corresponding estimates of the marginal variances will deteriorate.

We now introduce a simple Bayesian logistic regression model as a running example for the rest of the paper to illustrate our assumptions and findings. We generate $n = 10$ observations from a logistic regression model, with input $x \in \mathbb{R}$, response $y \in \{0, 1\}$ and $p(y|x, \theta) = p(x)^y(1-p(x))^{1-y}$, where $p(x) = 1/(1+e^{\theta_1+\theta_2 x})$. Furthermore, we set a $\mathcal{N}(0, \sigma_0^2 I)$ prior on θ where $\sigma_0 = 5$ and apply a mean-field Gaussian variational approximation $q(\theta; \mu, \phi)$. Table 1 illustrates how increasing ϕ causes the approximations to the marginal variances deteriorate in this example.

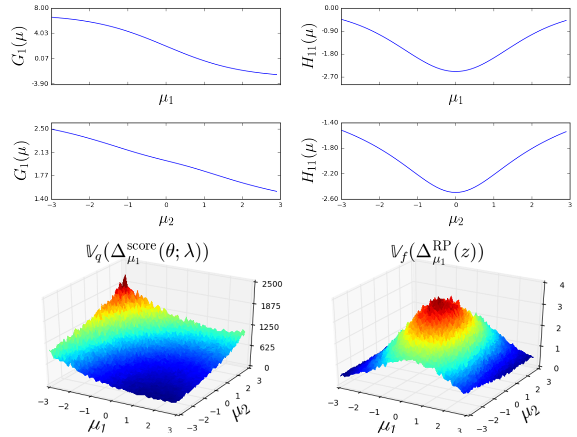

Figure 1: **Top row:** Cross section of $G_1(\mu)$ (left) and $H_{11}(\mu)$ (right) for $\mu_2 = 0$ (top) and $\mu_1 = 0$ (bottom). **Bottom row:** Simulation estimates of μ_1 gradient marginal variances ($S = 1,000$) with $\sigma_i = 1$ for $i = 1, 2$. The score function variance increases with $G_1(\mu)$ whereas the RP variance depends more on $H(\mu)$.

3.5 Results

The following proposition gives the marginal variances for the RP gradient and shows that they are smaller or equal to those of the score function gradient. Section A of the supplementary material provides a proof.

Proposition 1. *Suppose that Assumptions 1-3 hold and let $T(z; \lambda) = \mu + \sigma \circ z$, where $\mu = (\mu_1, \ldots, \mu_k)^\top$, $\sigma = (\sigma_1, \ldots, \sigma_k)^\top$, $\sigma_i = \exp(\phi_i)$ and $z = (z_1, \ldots, z_k)^\top$ with $z_i \sim \mathcal{N}(0, 1)$. Then, for $i = 1, \ldots, k$,*

(i)

$$\mathbb{V}_q\left(\Delta^{\text{score}}_{\mu_i}(\theta;\lambda)\right) = \frac{1}{\sigma_i^2}(C^2 + C\text{diag}(H(\mu)^2)^\top\sigma^2 + 2C\sigma_i^2 H_{ii}(\mu)G(\mu)^{2\top}\sigma^2 + \sigma_i^2 G_i(\mu)^2) + Q(H(\mu), \sigma) \quad (19)$$

$$\mathbb{V}_q\left(\Delta^{\text{score}}_{\phi_i}(\theta;\lambda)\right) = 3C^2 + CH_i(\mu)^\top\sigma^2 + 4C\sigma^2 H_{ii}(\mu) + 3(G(\mu)^{2\top}\sigma^2 + 4G_i(\mu)^2\sigma^2) + R(H(\mu), \sigma), \quad (20)$$

where C is a constant independent of λ and $Q(H(\mu), \sigma)$ and $R(H(\mu), \sigma)$ are second order function of elements of $H(\mu)$ and σ.

(ii)

$$\mathbb{V}_f\left(\Delta^{\text{RP}}_{\mu_i}(z;\lambda)\right) = H_i(\mu)^{2\top}\sigma^2 \quad (21)$$

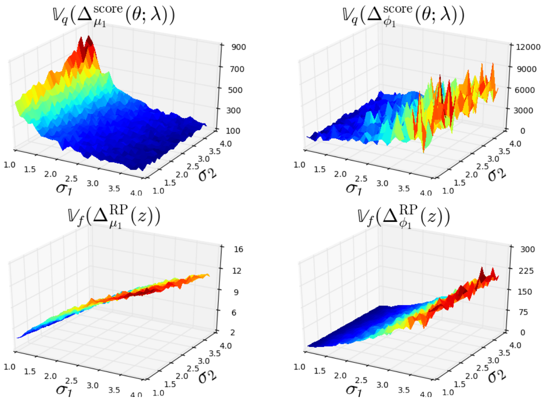

Figure 2: **Top row:** Simulation estimates of score function marginal variances ($S = 1,000$) with $\mu_i = 0$ for $i = 1, 2$. $\mathbb{V}_q(\Delta^{\text{score}}_{\mu_1}(\theta;\lambda))$ increases as $\sigma_1 \to 0$, but not when $\sigma_2 \to 0$. **Bottom row:** As per top row but for RP. No deterioration occurs as $\sigma_1 \to 0$.

$$\mathbb{V}_f\left(\Delta^{\text{RP}}_{\phi_i}(z;\lambda)\right) = \sigma_i^2\Big(H_i(\mu)^{2\top}\sigma^2 + H_{ii}(\mu)^2\sigma_i^2 + G_i(\mu)^2\Big), \quad (22)$$

(iii)

$$\mathbb{V}_f\left(\Delta^{\text{RP}}_{\mu_i}(z;\lambda)\right) \le \mathbb{V}_q\left(\Delta^{\text{score}}_{\mu_i}(\theta;\lambda)\right)$$

and

$$\mathbb{V}_f\left(\Delta^{\text{RP}}_{\phi_i}(z;\lambda)\right) \le \mathbb{V}_q\left(\Delta^{\text{score}}_{\phi_i}(\theta;\lambda)\right).$$

Corollary 1 shows that the trace of the covariance matrix of the RP gradient is smaller than that of the score function gradient.

Corollary 1. *Suppose Assumptions 1–3 hold and define*

$$\widehat{\nabla_\lambda\mathcal{L}(\lambda)}_{\text{RP}} \text{ and } \widehat{\nabla_\lambda\mathcal{L}(\lambda)}_{\text{score}}$$

as in Section 3.3. Then,

$$\operatorname{tr}\left(\operatorname{Cov}_f\left(\widehat{\nabla_\lambda\mathcal{L}(\lambda)}_{\text{RP}}\right)\right) \le \operatorname{tr}\left(\operatorname{Cov}_q\left(\widehat{\nabla_\lambda\mathcal{L}(\lambda)}_{\text{score}}\right)\right).$$

3.6 Observations on results

The expressions derived in Section 3.5 yield some intuition behind the differences in marginal variances between the score function and RP gradients. Firstly, the marginal variances of the score function gradient given in (19) and (20) depend on $G(\mu)$, meaning we would expect the marginal variance to be lowest when μ is near the true posterior mode where the gradient is 0. In contrast, the RP gradient marginal variances given in (21) and (22) have very little dependence on $G(\mu)$. Furthermore, (19) contains a $1/\sigma_i^2$, which implies that $\mathbb{V}_q\left(\Delta^{\text{score}}_{\mu_i}(\theta;\lambda)\right) \to \infty$ as $\sigma_i \to 0$. Interestingly, this is not the case for $\mathbb{V}_q\left(\Delta^{\text{score}}_{\phi_i}(\theta;\lambda)\right)$ or the RP gradients. Finally, the μ_i and ϕ_i components of the RP gradient only contain gradient component i and row i hessian terms. In contrast, the score function gradient contains all gradient and hessian components. This is due to the RP gradient taking the gradient of the log-joint density, causing all terms not containing θ_i to vanish. These observations imply that the score function gradient estimator behaves in a fundamentally different way to the RP gradient estimator. Figures 1 and 2 illustrate this for the logistic regression example presented in Section 3.4.

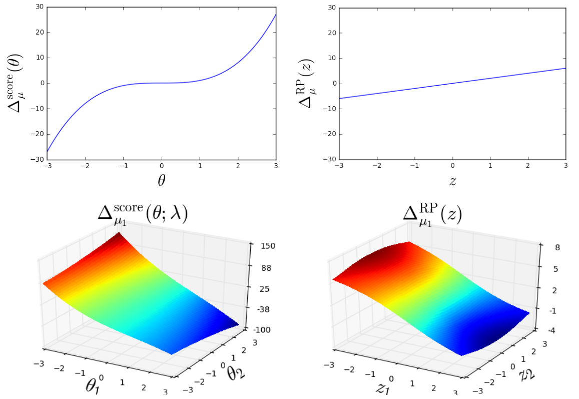

Figure 3: **Top row:** Δ-functions for $h(\theta) = \theta^2$ and $q(\theta) \sim \mathcal{N}(0,1)$. Note $\Delta^{\text{RP}}(z;\lambda)$ varies less than $\Delta^{\text{score}}(\theta;\lambda)$ over the sampling region. **Bottom row:** Δ-functions for the logistic regression example with $\mu = (0,0)$ and $\sigma = (1,1)$. Again, notice the higher variation of $\Delta^{\text{score}}(\theta;\lambda)$ over the sampling region.

3.7 Insights on the reparameterization trick

Some papers in the literature explain the success of the RP trick as due to its efficient use of gradient information from the log-joint density (Titsias and Lázaro-Gredilla, 2014; Tan and Nott, 2018; Quiroz et al., 2018) without elaborating further.

We argue that since the RP trick allows us to take the gradient of the log-joint density $h(\theta)$ with respect to θ when constructing an estimator, it yields an estimator containing lower order terms with respect to $\theta = T(z;\lambda)$ compared to the score function method. Specifically, $\Delta^{\text{score}}_{\lambda_i}(\theta;\lambda)$ contains higher orders of θ whereas $\Delta^{\text{RP}}_{\lambda_i}(z;\lambda)$ contains lower orders of z. Let $B_q \subset \Theta$ be a compact subset of Θ that contains a large proportion of the samples from q used to evaluate the Monte Carlo estimate of the gradient. We refer to this as the "sampling region" of q. Similarly, B_f refers to the sampling region of f for the RP gradient. For example, if $q(\theta;\lambda) = \mathcal{N}(0,2)$ then $B_q = [-6,6]$ and $B_f = [-3,3]$ are appropriate since 99.7% of the sam-

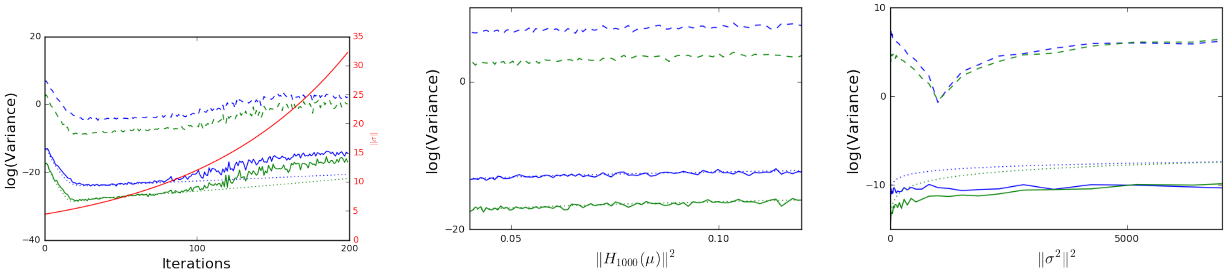

Figure 4: Bayesian multinomial logistic regression example under Experiments 1–3 (from left to right), see Section 4.1. Legend: μ_{1000} (blue), ϕ_{1000} (green), Score (dashed), RP (solid), approximations (21) and (22) (dotted). **Left:** Poor approximations at iteration 100 are due to high values in σ. **Middle:** The variance increases with $\|H_{1000}(\mu)\|$. **Right:** The variance increases with $\|\sigma\|^2$ despite the poor accuracy of the approximation.

ples lie in these intervals. The reason why the score function gradient tends to have higher variance is because the image of B_q under $\Delta^{\text{score}}_{\lambda_i}(\theta;\lambda)$ tends to have a larger range compared to the image of B_f under $\Delta^{\text{RP}}_{\lambda_i}(z;\lambda)$. We call this having a "higher variation" in the sampling region of the estimator.

To illustrate, suppose $h(\theta) = \theta^2$ and $q(\theta;\mu) \sim \mathcal{N}(\mu, 1)$. We can use (14) to show that $\Delta^{\text{score}}_{\mu}(\theta;\lambda) = \theta^3 - \theta^2\mu$, which contains a third order power of θ. From this, $\mathbb{V}_q(\Delta^{\text{score}}_{\mu}(\theta;\lambda)) = \mu^4 + 14\mu^2 + 15$. In contrast, the RP gradient estimator is given by $\Delta^{\text{RP}}_{\mu}(z;\lambda) = 2(\mu + z)$ hence $\mathbb{V}_f(\Delta^{\text{RP}}_{\mu}(z;\lambda)) = 4$. We see a large difference in variance that appears to be driven by the fact that the RP gradient estimator's leading term is at least two orders lower than that of the score function gradient estimator. Consequently, the score function estimator has higher variation over its sampling region compared to the RP gradient estimator. Figure 3 illustrates this for the example above, as well as for the logistic regression example discussed in Section 3.4. Note that these observations hold for the gradient with respect to ϕ as well and readily extends to the multivariate case. Despite many log-joint density functions not being polynomials, we can find a reasonable polynomial approximation over the sampling region using the Stone-Weierstrass theorem (Stone, 1948).

3.8 Related work

Fan et al. (2015) show that if a function $g : \mathbb{R}^k \to \mathbb{R}$ is Lipschitz continuous with constant L, and $z \sim \mathcal{N}(0, I_k)$, then $\mathbb{V}[g(z)] \leq L^2$. In addition, they claim that in practice the variance is highly sensitive to L. This is similar to the intuition we develop since L tends to give a rough indication of the variation of g which drives the variance. The limitation of using the Lipschitz constant is that even for basic models such as Bayesian linear regression, the log-joint density is not Lipschitz continuous and so these results are not immediately useful. In our work, we apply a more specific simplifying assumption to the log-joint density instead, which allows us to look at specific properties around the variance reduction of the RP gradient in a region where the function is locally quadratic.

Gal (2016, Chapter 3.1.2) shows that given a univariate $\theta \sim \mathcal{N}(\mu, \sigma^2)$ and assuming certain conditions on h hold, the RP gradient estimator has smaller marginal variances than the corresponding estimator under the score function method. While Gal (2016, Chapter 3.1.2) defines a set of conditions and proves that the RP gradient has lower marginal variance given these conditions, limited insight is provided around when the RP trick works well in practice. The results are also restricted to a univariate posterior. We tackle this problem in the multivariate case, and offer a set of simplifying assumptions that are reasonable for certain classes of models. Furthermore, we discuss the intuition behind the drivers of the variance of the gradient estimators and why the RP gradient is more efficient than the score function gradient.

Finally, we note that there are no guarantees that the RP trick is more efficient in the general case. A counterexample corresponding to a highly multimodal log-joint density ($h(\theta) = \sin(10\theta)$) is given in Gal (2016, Chapter 3.1.2). Parmas et al. (2018) also shows a counterexample in the reinforcement learning context. This highlights the fact that we need to make reasonable simplifying assumptions on h to be able to theoretically conclude that the RP gradient is more efficient than the score function gradient.

4 EXAMPLES

This section studies whether our results and insights from Sections 3.5, 3.6 and 3.7 derived under the quadratic assumption of the log-joint density are useful in cases where the assumption does not reasonably hold. We show that our expressions for the marginal

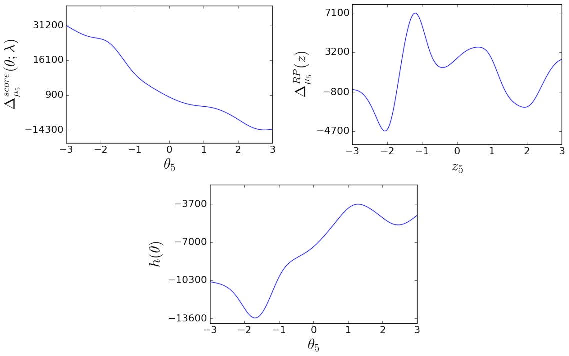

Figure 5: **Top row:** Cross section of Δ_{μ_5} functions for the Bayesian Neural Network model with $q(\theta;\lambda) \sim \mathcal{N}(0, I)$. The RP estimator varies less over its sampling region. Simulations ($S = 10,000$) yield $\mathbb{V}_q(\Delta_{\mu_5}^{\text{score}}(\theta;\lambda)) = 2.53e11$ and $\mathbb{V}_f(\Delta_{\mu_5}^{\text{RP}}(z;\lambda)) = 6.60e7$. **Bottom row:** Cross section of $h(\theta)$. The quadratic assumption is clearly inappropriate here.

variances in Section 3.5 capture the behaviour of the marginal variances of a high-dimensional multinomial logistic regression model. Furthermore, we show that our intuition regarding the difference in variation of the estimators over the sampling region explains the variance reduction properties of the RP gradient for a simple two layer Bayesian neural network model where we expect the quadratic assumption would not hold. We apply a mean-field Gaussian variational approximation in both examples.

4.1 Bayesian multinomial logistic regression

The MNIST database of handwritten digits (LeCun et al., 1998) contains 60,000 training observations and 10,000 test observations of 28×28 images with 10 prediction classes. We fit a Bayesian multinomial logistic (or softmax) regression model for classification with a $\mathcal{N}(0, \sigma_0^2 I)$ prior over the regression coefficients with $\sigma_0 = 40$. The elements of the score function and RP gradient estimators corresponding to parameters μ_{1000} and ϕ_{1000} were analyzed by conducting three experiments. In each case we evaluated the log of the marginal variance for each element.

Experiment 1 We ran the optimization for 200 iterations with $\sigma = \exp(\phi)$ initialized with very small values and observed that all elements of σ gradually increased due to the high dimensionality of the posterior relative to the number of observations.

Experiment 2 We held ϕ fixed and increased the value of μ_i while fixing μ_j for $j \neq i$. This had the effect of varying elements of $H_i(\mu)$. We expect from (21) and (22) that the marginal variance of the gradient for both μ_i and ϕ_i will increase with $\|H_i(\mu)\|^2$ where $\|\cdot\|$ is the Euclidean norm of the corresponding vector.

Experiment 3 We held μ (and therefore $H_i(\mu)$) fixed and increased ϕ_j for all j. This was designed to measure the effect of ϕ on the marginal variance.

Figure 4 shows the results of Experiments 1–3. When σ is small the quadratic assumption yields reasonable estimates for the marginal variances of the RP gradient, but it deteriorates as σ increases. In addition, the marginal variance of the score function and RP gradient clearly increases with both $\|H_i(\mu)\|^2$ and $\|\sigma^2\|^2$. Remarkably, this is consistent with both (21) and (22), despite these formulas yielding poor estimates of the true marginal variances.

4.2 Bayesian neural network

We follow Duvenaud and Adams (2015) and apply a simple Bayesian neural network on 40 simulated observations. The density of observation y_i given input $x_i \in \mathbb{R}$ and neural network weights $\mathbf{w}$ is $p(y_i|\mathbf{w}, x_i, \sigma_{\text{err}}^2) = \mathcal{N}(y_i|\text{NN}(x_i;\mathbf{w}), \sigma_{\text{err}}^2)$, where $\text{NN}(x_i;\mathbf{w})$ is a neural network with two hidden layers of size 20 with tanh activations and $\sigma_{\text{err}}^2 = 1$. A $\mathcal{N}(0, \sigma_0^2 I)$ prior is set over $\mathbf{w}$ where $\sigma_0 = 40$. Figure 5 illustrates the highly non-quadratic properties of the log-joint density of a neural network. Nevertheless, the variance of the gradient estimators mainly depends on the variation of the estimator over its sampling region.

5 CONCLUSION AND FUTURE RESEARCH

We have studied the variance reduction properties of the reparameterization trick under certain simplifying assumptions. We argue that its success depends on the fact that it generally results in an expression that has lower variation over the sampling region of the variational distribution compared to the score function method. Finally, we showed that our conclusions in Sections 3.6 and 3.7 are useful in describing cases where our assumptions are not perfectly satisfied.

Future extensions include relaxing the mean-field assumption by considering more flexible covariance structures as in Tan and Nott (2018); Ong et al. (2018b); Quiroz et al. (2018). Variational families other than the Gaussian density may also be considered, for example a mean-field approximation with a mixture of normal and Gamma components like in Ranganath et al. (2014). Finally, alternative scalar measures of variability such as the ones discussed in Section 3.2 can be employed to assess the efficiency of the gradient estimators.

Acknowledgements

The authors were supported by the Australian Centre of Excellence in Mathematical and Statistical Frontiers (ACEMS, grant CE140100049). SAS is also supported by the Australia Research Council Discovery Projects Scheme (grant DP160102544).

References

Balles, L. and Hennig, P. (2018). Dissecting Adam: The sign, magnitude and variance of stochastic gradients. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pages 404–413.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112:859–877.

Bottou, L. (2010). Large-scale machine learning with stochastic gradient descent. In Lechevallier, Y. and Saporta, G., editors, *Proceedings of the 19th International Conference on Computational Statistics (COMPSTAT'2010)*, pages 177–187. Springer.

Duchi, J., Hazan, E., and Singer, Y. (2011). Adaptive subgradient methods for online learning and stochastic optimization. *Journal of Machine Learning Research*, 12:2121–2159.

Duvenaud, D. and Adams, R. P. (2015). Black-box stochastic variational inference in five lines of python. In *NIPS Workshop on Black-box Learning and Inference*.

Fan, K., Wang, Z., Beck, J., Kwok, J., and Heller, K. A. (2015). Fast second order stochastic backpropagation for variational inference. In *Advances in Neural Information Processing Systems*, pages 1387–1395.

Figurnov, M., Mohamed, S., and Mnih, A. (2018). Implicit reparameterization gradients. *arXiv preprint arXiv:1805.08498*.

Gal, Y. (2016). *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge.

Han, S., Liao, X., Dunson, D. B., and Carin, L. C. (2016). Variational Gaussian copula inference. In Gretton, A. and Robert, C. C., editors, *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, volume 51, pages 829–838, Cadiz, Spain. JMLR Workshop and Conference Proceedings.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Machine Learning*, 37:183–233.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations (ICLR)*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational Bayes. *Proceedings of the 2nd International Conference on Learning Representations (ICLR)*.

Kucukelbir, A., Tran, D., Ranganath, R., Gelman, A., and Blei, D. M. (2017). Automatic differentiation variational inference. *The Journal of Machine Learning Research*, 18:430–474.

LeCun, Y., Bottou, L., Bengio, Y., and Haffner, P. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86:2278–2324.

Lee, W., Yu, H., and Yang, H. (2018). Reparameterization gradient for non-differentiable models. *arXiv preprint arXiv:1806.00176*.

Miller, A., Foti, N., D'Amour, A., and Adams, R. P. (2017). Reducing reparameterization gradient variance. In *Advances in Neural Information Processing Systems*, pages 3708–3718.

Miller, A. C., Foti, N., and Adams, R. P. (2016). Variational boosting: Iteratively refining posterior approximations. arXiv: 1611.06585.

Nott, D. J., Tan, S. L., Villani, M., and Kohn, R. (2012). Regression density estimation with variational methods and stochastic approximation. *Journal of Computational and Graphical Statistics*, 21:797–820.

Ong, V. M., Nott, D. J., Tran, M.-N., Sisson, S. A., and Drovandi, C. C. (2018a). Variational Bayes with synthetic likelihood. *Statistics and Computing*, 28:971–988.

Ong, V. M.-H., Nott, D. J., and Smith, M. S. (2018b). Gaussian variational approximation with a factor covariance structure. *Journal of Computational and Graphical Statistics*, To appear.

Ormerod, J. T. and Wand, M. P. (2010). Explaining variational approximations. *The American Statistician*, 64:140–153.

Parmas, P., Rasmussen, C. E., Peters, J., and Doya, K. (2018). PIPPS: Flexible model-based policy search robust to the curse of chaos. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4065–4074, Stockholmsmssan, Stockholm Sweden. PMLR.

Quiroz, M., Nott, D. J., and Kohn, R. (2018). Gaussian variational approximation for high-dimensional state space models. *arXiv preprint arXiv:1801.07873v2*.

Ranganath, R., Gerrish, S., and Blei, D. (2014). Black box variational inference. In *Artificial Intelligence and Statistics*, pages 814–822.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *Proceedings of the 31st International Conference on Machine Learning*, volume 32, pages 1278–1286. PMLR.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22:400–407.

Roeder, G., Wu, Y., and Duvenaud, D. K. (2017). Sticking the landing: Simple, lower-variance gradient estimators for variational inference. In *Advances in Neural Information Processing Systems*, pages 6925–6934.

Ruiz, F. R., Titsias, M. K., and Blei, D. (2016). The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, pages 460–468.

Stone, M. H. (1948). The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21:237–254.

Tan, L. S. and Nott, D. J. (2018). Gaussian variational approximation with sparse precision matrices. *Statistics and Computing*, 28:259–275.

Titsias, M. and Lázaro-Gredilla, M. (2014). Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, pages 1971–1979.

Tran, M.-N., Nott, D. J., and Kohn, R. (2017). Variational Bayes with intractable likelihood. *Journal of Computational and Graphical Statistics*, 26:873–882.

Wilks, S. S. (1932). Certain generalizations in the analysis of variance. *Biometrika*, 24:471–494.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. In *Reinforcement Learning*, pages 5–32. Springer.

Zeiler, M. D. (2012). Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

Zobay, O. (2014). Variational Bayesian inference with Gaussian-mixture approximations. *Electronic Journal of Statistics*, 8:355–389.