VarGrad: A Low-Variance Gradient Estimator for Variational Inference

Lorenz Richter[*]
Freie Universität Berlin
BTU Cottbus-Senftenberg
dida Datenschmiede GmbH
lorenz.richter@fu-berlin.de

Ayman Boustati[*]
University of Warwick
a.boustati@warwick.ac.uk

Nikolas Nüsken
Universität Potsdam
nuesken@uni-potsdam.de

Francisco J. R. Ruiz
DeepMind
franrruiz@google.com

Ömer Deniz Akyildiz
University of Warwick
The Alan Turing Institute
omer.akyildiz@warwick.ac.uk

Abstract

We analyse the properties of an unbiased gradient estimator of the evidence lower bound (ELBO) for variational inference, based on the score function method with leave-one-out control variates. We show that this gradient estimator can be obtained using a new loss, defined as the variance of the log-ratio between the exact posterior and the variational approximation, which we call the *log-variance loss*. Under certain conditions, the gradient of the log-variance loss equals the gradient of the (negative) ELBO. We show theoretically that this gradient estimator, which we call *VarGrad* due to its connection to the log-variance loss, exhibits lower variance than the score function method in certain settings, and that the leave-one-out control variate coefficients are close to the optimal ones. We empirically demonstrate that VarGrad offers a favourable variance versus computation trade-off compared to other state-of-the-art estimators on a discrete variational autoencoder (VAE).

1 Introduction

Estimating the gradient of the expectation of a function is a problem with applications in many areas of machine learning, ranging from variational inference to reinforcement learning [Mohamed et al., 2019]. Different gradient estimators lead to different algorithms; two examples of estimators are the score function gradient [Williams, 1992] and the reparameterisation gradient [Kingma and Welling, 2014, Rezende et al., 2014, Titsias and Lázaro-Gredilla, 2014]. Many recent works develop new estimators with different properties (such as their variance); see Section 5 for a review.

We focus on variational inference (VI), where the goal is to approximate the posterior distribution $p(z \mid x)$ of a model $p(x, z)$, where x denotes the observations and z refers to the latent variables of the model [Jordan et al., 1999, Blei et al., 2017]. VI approximates the posterior using a parameterised family of distributions $q_\phi(z)$, and finds the parameters ϕ by minimising the Kullback-Leibler (KL)

---

[*]Equal contribution.

divergence from $q_\phi(z)$ to $p(z \,|\, x)$, i.e., $\text{KL}(q_\phi(z) \,||\, p(z \,|\, x))$. Since the KL is intractable, VI solves instead an equivalent problem that maximises the evidence lower bound (ELBO), given by

$$\text{ELBO}(\phi) = \mathbb{E}_{q_\phi} \left[\log \frac{p(x, z)}{q_\phi(z)} \right]. \tag{1}$$

Thus, VI casts the inference problem as an optimisation problem, which can be solved with stochastic optimisation tools when the ELBO is not available in closed form. In particular, VI forms a Monte Carlo estimator of the gradient of the ELBO, $\nabla_\phi \text{ELBO}(\phi)$.

In this paper, we analyse a multi-sample estimator of the gradient of the ELBO. In particular, we focus on an estimator first introduced by Salimans and Knowles [2014] and Kool et al. [2019], which is based on the score function method [Williams, 1992] with leave-one-out control variates.

We first show the connection between this estimator and an alternative divergence measure between the variational distribution $q_\phi(z)$ and the exact posterior $p(z \,|\, x)$. This divergence, which is different from the standard KL used in variational inference, is defined as the variance, under some arbitrary distribution $r(z)$, of the log-ratio $\log \frac{q_\phi(z)}{p(z \,|\, x)}$. We refer to this divergence as the *log-variance loss*. Inspired by Nüsken and Richter [2020], we show that we recover the gradient estimator of Salimans and Knowles [2014] and Kool et al. [2019] by taking the gradient with respect to the variational parameters ϕ of the log-variance loss and evaluating the result at $r(z) = q_\phi(z)$. Due to this property, we refer to the gradient estimator as *VarGrad*. This property also suggests a simple algorithm for computing the gradient estimator, based on differentiating through the log-variance loss.

We then study the relationship between VarGrad and the score function estimator [Williams, 1992, Carbonetto et al., 2009, Paisley et al., 2012, Ranganath et al., 2014] with optimal control variate coefficients. We show that the control variate coefficients of VarGrad are close to the (intractable) optimal coefficients. Indeed, we show both theoretically and empirically that the difference between both is small in many cases; for example when the KL from $q_\phi(z)$ to the posterior is either small or large, which is generally the case in the late and early stages of the optimisation, respectively. This explains the success of the VarGrad estimator in a variety of settings [Kool et al., 2019, 2020].

Since it is based on the score function, VarGrad is a black-box, general purpose estimator because it makes no assumptions on the model $p(x, z)$, such as differentiability with respect to the latent variables z. It introduces no additional parameters to be tuned and it is not computationally expensive. In Section 6, we show empirically that VarGrad exhibits a favourable variance versus computation trade-off compared to other unbiased gradient estimators, including the score function gradient with control variates [Williams, 1992, Ranganath et al., 2014], REBAR [Tucker et al., 2017], RELAX [Grathwohl et al., 2018], and ARM [Yin and Zhou, 2019].

2 Background

In this section, we introduce the notation and review one of the most relevant estimators in VI: the score function method. We also review its improved version based on leave-one-out control variates.

Consider a probabilistic model $p(x, z)$, where z denotes the latent variables and x are the observations. We are interested in computing the posterior $p(z \,|\, x) = p(x, z)/p(x)$, where $p(x) = \int p(x, z) \, \mathrm{d}z$ is the marginal likelihood. For most models of interest, the posterior is intractable due to the intractability of the marginal likelihood, and we resort to an approximation.

Variational inference approximates the posterior $p(z \,|\, x)$ with a parameterised family of distributions $q_\phi(z)$ (with $\phi \in \Phi$), called variational family. Variational inference finds the parameters ϕ^* that minimise the KL divergence, $\phi^* = \text{argmin}_{\phi \in \Phi} \text{KL} (q_\phi(z) \,||\, p(z \,|\, x))$. This optimisation problem is intractable because the KL itself depends on the intractable posterior. Variational inference sidesteps this problem by maximising instead the ELBO defined in Eq. 1, which is a lower bound on the marginal likelihood, since $\log p(x) = \text{ELBO}(\phi) + \text{KL} (q_\phi(z) \,||\, p(z \,|\, x))$. As the expectation in Eq. 1 is typically intractable, variational inference uses stochastic optimisation to maximise the ELBO. In particular, it forms unbiased Monte Carlo estimators of the gradient $\nabla_\phi \text{ELBO}(\phi)$.

We next review the score function method, a Monte Carlo estimator commonly used in variational inference. Instead of the ELBO, we focus on the gradients of the KL divergence with respect to the variational parameters ϕ. These gradients are equal to the gradients of the negative ELBO because the marginal likelihood $p(x)$ does not depend on ϕ; that is, $\nabla_\phi \text{KL} (q_\phi(z) \,||\, p(z \,|\, x)) = -\nabla_\phi \text{ELBO}(\phi)$.

The score function estimator [Williams, 1992, Carbonetto et al., 2009, Paisley et al., 2012, Ranganath et al., 2014], also known as Reinforce, expresses the gradient as an expectation that depends on the log-ratio $q_\phi(z)/p(x, z)$ weighted by the score function $\nabla_\phi \log q_\phi(z)$. The resulting estimator is

$$\nabla_\phi \mathrm{KL}\left[q_\phi(z) \,||\, p(z \,|\, x)\right] \approx \widehat{g}_{\mathrm{Reinforce}}(\phi) = \frac{1}{S} \sum_{s=1}^{S} \log\left(\frac{q_\phi(z^{(s)})}{p(x, z^{(s)})}\right) \nabla_\phi \log q_\phi(z^{(s)}), \qquad (2)$$

where $z^{(s)} \overset{\text{i.i.d.}}{\sim} q_\phi(z)$. Due to its high variance, the score function estimator requires additional tricks in practice; Ranganath et al. [2014] use Rao-Blackwellization and control variates. The control variates are multiples of the score function, $a \odot \nabla_\phi \log q_\phi(z)$, where $\odot$ denotes the Hadamard (element-wise) product and the coefficient a is chosen to minimise the estimator variance.

Salimans and Knowles [2014] and Kool et al. [2019] leverage the multi-sample estimator by using $S - 1$ samples to compute the control variate coefficient a and then average over the resulting estimators, obtaining a leave-one-out estimator

$$\widehat{g}_{\mathrm{LOO}}(\phi) = \frac{1}{S-1}\left(\sum_{s=1}^{S} f_\phi(z^{(s)}) \nabla_\phi \log q_\phi(z^{(s)}) - \bar{f}_\phi \sum_{s=1}^{S} \nabla_\phi \log q_\phi(z^{(s)})\right), \quad z^{(s)} \overset{\text{i.i.d.}}{\sim} q_\phi(z), \quad (3)$$

where for simplicity of notation we have defined $f_\phi(z)$ as the log-ratio $\log \frac{q_\phi(z)}{p(x,z)}$ and $\bar{f}_\phi$ as its empirical average (which is a multi-sample Monte Carlo estimate of the negative ELBO), i.e.,

$$f_\phi(z) = \log \frac{q_\phi(z)}{p(x, z)} \quad \text{and} \quad \bar{f}_\phi = \frac{1}{S} \sum_{s=1}^{S} f_\phi(z^{(s)}) \approx -\mathrm{ELBO}(\phi). \qquad (4)$$

The score function method makes no assumptions on the model $p(x, z)$ or the distribution $q_\phi(z)$; the only requirements are to be able to sample from $q_\phi(z)$ and to evaluate $\log q_\phi(z)$ and $\log p(x, z)$.

3 The Log-Variance Loss and its Connection to VarGrad

In this section, we show the connection between the leave-one-out estimator in Eq. 3 and a novel divergence, which we call the log-variance loss. We introduce the log-variance loss in Section 3.1 and show its connection to Eq. 3 in Section 3.2. We refer to the estimator in Eq. 3 as *VarGrad*.

3.1 The Log-Variance Loss

The log-variance loss is defined as the variance, under some arbitrary distribution $r(z)$, of the log-ratio $\log \frac{q_\phi(z)}{p(z \,|\, x)}$. It has the property of reproducing the gradients of the KL divergence under certain conditions (see Proposition 1 for details). We next give the precise definition of the loss.

Definition 1. *For a given distribution $r(z)$, the log-variance loss $\mathcal{L}_r(\cdot)$ is given by*

$$\mathcal{L}_r(q_\phi(z) \,||\, p(z \,|\, x)) = \frac{1}{2}\mathrm{Var}_r\left(\log\left(\frac{q_\phi(z)}{p(z \,|\, x)}\right)\right). \qquad (5)$$

We refer to the distribution $r(z)$ as the *reference distribution* under which the discrepancy between $q_\phi(z)$ and the posterior $p(z \,|\, x)$ is computed. When the support of the reference distribution contains the supports of $q_\phi(z)$ and $p(z \,|\, x)$, Eq. 5 is a divergence;[†] it is zero if and only if $q_\phi(z) = p(z \,|\, x)$. The factor $1/2$ in Eq. 5 is only included because it simplifies some expressions later in this section.

We next show that the gradient of the log-variance loss and the gradient of the standard KL divergence coincide under certain conditions. In particular, taking the gradient of Eq. 5 with respect to the variational parameters ϕ *and then* evaluating the result for a reference distribution $r(z) = q_\phi(z)$ gives the gradient of the KL. This property is detailed in Proposition 1.

[†]More technically, as we assume that $r(z)$, $p(z \,|\, x)$, and $q_\phi(z)$ admit densities, it follows that measure-zero sets of $r(z)$ are necessarily measure-zero sets of $p(z \,|\, x)$ and $q_\phi(z)$, implying that the divergence is well defined.

Proposition 1. *The gradient with respect to ϕ of the log-variance loss, evaluated at $r(z) = q_\phi(z)$, equals the gradient of the* KL *divergence,*

$$\nabla_\phi \mathcal{L}_r(q_\phi(z) \,||\, p(z \,|\, x))\Big|_{r=q_\phi} = \nabla_\phi \mathrm{KL}(q_\phi(z) \,||\, p(z \,|\, x)). \qquad (6)$$

Proof. See Appendix A.1. □

Proposition 1 implies that we can estimate the gradient of the KL divergence by estimating instead the gradient of the log-variance loss.

Remark 1. The result in Proposition 1 is obtained by setting $r(z) = q_\phi(z)$ *after* taking the gradient with respect to ϕ. The same result does not hold if we set $r(z) = q_\phi(z)$ before differentiating.

3.2 VarGrad: Derivation of the Gradient Estimator from the Log-Variance Loss

The leave-one-out estimator in Eq. 3 [Salimans and Knowles, 2014, Kool et al., 2019] is connected to the log-variance loss from Section 3.1 through Proposition 1. Firstly, note that the log-variance loss is intractable as it depends on the posterior $p(z \,|\, x)$. However, since the marginal likelihood $p(x)$ has zero variance, it can be dropped from the definition in Eq. 5, yielding

$$\mathcal{L}_r(q_\phi(z) \,||\, p(z \,|\, x)) = \frac{1}{2}\mathrm{Var}_r\left(\log\left(\frac{q_\phi(z)}{p(x,z)}\right)\right) = \frac{1}{2}\mathrm{Var}_r\left(f_\phi(z)\right), \qquad (7)$$

where $f_\phi(z)$ is defined in Eq. 4.

Next, we build the estimator of the log-variance loss as the empirical variance of S Monte Carlo samples,

$$\mathcal{L}_r(q_\phi(z) \,||\, p(z \,|\, x)) \approx \frac{1}{2(S-1)}\sum_{s=1}^{S}\left(f_\phi(z^{(s)}) - \bar{f}_\phi\right)^2, \quad z^{(s)} \overset{\text{i.i.d.}}{\sim} r(z). \qquad (8)$$

Applying Proposition 1 by differentiating through Eq. 8, we arrive at the VarGrad estimator, $\widehat{g}_{\mathrm{VarGrad}}(\phi) = \widehat{g}_{\mathrm{LOO}}(\phi) \approx \nabla_\phi \mathrm{KL}\left(q_\phi(z) \,||\, p(z \,|\, x)\right)$, where

$$\widehat{g}_{\mathrm{VarGrad}}(\phi) = \frac{1}{S-1}\left(\sum_{s=1}^{S} f_\phi(z^{(s)})\nabla_\phi \log q_\phi(z^{(s)}) - \bar{f}_\phi \sum_{s=1}^{S}\nabla_\phi \log q_\phi(z^{(s)})\right) \qquad (9)$$

and $z^{(s)} \overset{\text{i.i.d.}}{\sim} q_\phi(z)$.

The expression for VarGrad in Eq. 9 is identical to that of the leave-one-out estimator in Eq. 3. Thus, VarGrad is an unbiased estimator of the gradient of the KL (and equivalently the gradient of the ELBO). From a probabilistic programming perspective, setting the reference $r(z) = q_\phi(z)$ *after* differentiating w.r.t. ϕ amounts to sampling $z^{(s)} \sim q_\phi(z)$ and detaching the resulting samples from the computational graph. This suggests a novel algorithmic procedure, given in Algorithm 1. Its implementation is simple: we only need the samples $z^{(s)} \sim q_\phi(z)$ and apply the `stop_gradient` operator, evaluate the log-ratio $f_\phi(z^{(s)})$ for each sample, and then differentiate through the empirical variance of this log-ratio.

Algorithm 1 Pseudocode for VarGrad

Input: Variational parameters ϕ, data x
    for $s = 1, \ldots, S$ **do**
        $z^{(s)} \leftarrow$ `sample`$(q_\phi(\cdot))$ ▷ Sample from the approximate posterior
        $z^{(s)} \leftarrow$ `stop_gradient`$(z^{(s)})$ ▷ Detach the samples from the computational graph
        $f_\phi^{(s)} \leftarrow \log q_\phi(z^{(s)}) - \log p(x, z^{(s)})$ ▷ An estimate of the negative ELBO
    $\widehat{\mathcal{L}} \leftarrow \frac{1}{2}$`Variance`$(\{f_\phi^{(s)}\}_{s=1}^{S})$ ▷ An estimate of the log-variance loss
    return `grad`$(\widehat{\mathcal{L}})$ ▷ Differentiate through the loss w.r.t. ϕ

---

4 Analytical Results

In this section we study the properties of $\widehat{g}_{\text{VarGrad}}$ in comparison to other estimators based on the score function method. In Section 4.1, we analyse the difference δ^{CV} between the control variate coefficient of VarGrad (called a^{VarGrad}) and the optimal one. The former can be approximated cheaply and unbiasedly, while a standard Monte Carlo estimator of the latter is biased and often exhibits high variance. Furthermore, we establish that the difference δ^{CV} is negligible in certain settings, in particular when $\text{KL}(q_\phi(z) \,||\, p(z \,|\, x))$ is either very large or close to zero; thus in these settings the control variate coefficient of VarGrad is close to the optimal coefficient. In Section 4.2 we show that a simple relation between δ^{CV} and the ELBO is sufficient to *guarantee* that $\widehat{g}_{\text{VarGrad}}$ has lower variance than $\widehat{g}_{\text{Reinforce}}$ when the number of Monte Carlo samples is large enough.

4.1 Analysis of the Control Variate Coefficients

As alluded to in Section 2, Ranganath et al. [2014] proposed to modify $\widehat{g}_{\text{Reinforce}}$ using a score function control variate, that is,

$$\widehat{g}_{\text{CV}}(\phi) = \widehat{g}_{\text{Reinforce}}(\phi) - a \odot \left(\frac{1}{S} \sum_{s=1}^{S} \nabla_\phi \log q_\phi(z^{(s)}) \right), \tag{10}$$

where a is a vector chosen so as to reduce the variance of the estimator. We recover VarGrad (Eq. 9), up to a factor of proportionality, by setting the control variate coefficient $a = \bar{f}_\phi \mathbf{1}$ in Eq. 10, where $\mathbf{1}$ is a vector of ones. The proportionality relation is $\frac{S-1}{S}\widehat{g}_{\text{CV}} = \widehat{g}_{\text{VarGrad}}$. In terms of variance reduction, the coefficients of the optimal a^* are given by

$$a_i^* = \frac{\text{Cov}_{q_\phi}\left(f_\phi \partial_{\phi_i} \log q_\phi, \partial_{\phi_i} \log q_\phi\right)}{\text{Var}_{q_\phi}\left(\partial_{\phi_i} \log q_\phi\right)}. \tag{11}$$

We next show that the coefficients of VarGrad, a^{VarGrad}, are close to the optimal coefficients a^*. For this, we first relate a^{VarGrad} to a^* in Lemma 1.

Lemma 1. *We can write the optimal control variate coefficient as the expected value of $a^{VarGrad}$ plus a* control variate correction *term δ^{CV}, i.e.,*

$$a^* = \mathbb{E}_{q_\phi}[a^{VarGrad}] + \delta^{CV} = -\text{ELBO}(\phi) + \delta^{CV}, \tag{12}$$

where $a^{VarGrad} = \bar{f}_\phi$ and the components of the correction term δ^{CV} are given by

$$\delta_i^{CV} = \frac{\text{Cov}_{q_\phi}\left(f_\phi, (\partial_{\phi_i} \log q_\phi)^2\right)}{\text{Var}_{q_\phi}\left(\partial_{\phi_i} \log q_\phi\right)}. \tag{13}$$

Proof. See Appendix A.2. $\square$

According to Lemma 1, the difference between the optimal control variate coefficient and the (expected) VarGrad coefficient is equal to the correction δ^{CV}. We hypothesise that direct Monte Carlo estimation of δ^{CV} in Eq. 13 (or similarly for Eq. 11) suffers from high variance because it takes the form of a fraction[‡] (see for instance Appendix C). Moreover, estimating Eq. 13 by taking the ratio of two Monte Carlo estimators gives a biased estimate.

We next show that in certain settings the correction term δ^{CV} becomes negligible, implying that $\widehat{g}_{\text{VarGrad}}$ and $\widehat{g}_{\text{Reinforce}}$ equipped with the optimal control variate coefficients behave almost identically. We provide empirical evidence of this finding in Section 6 (and in Appendix C for the Gaussian case).

Proposition 2 (δ^{CV} is small in comparison to $\mathbb{E}_{q_\phi}[a^{\text{VarGrad}}]$ if the KL divergence between $q_\phi(z)$ and $p(z \,|\, x)$ is large or small)**.** *Assume that $q_\phi(z)$ has lighter tails than the posterior $p(z \,|\, x)$, in the sense that there exists a constant $C > 0$ such that*

$$\sup_z \frac{q_\phi(z)}{p(z \,|\, x)} < C. \tag{14}$$

[‡]Monte Carlo estimators of fractions are not straightforward. As a simple example, consider the ratio of two independent Gaussian random variables, each with zero mean and unit variance. The ratio follows a Cauchy distribution, which has infinite variance.

Furthermore, define the kurtosis of the score function,

$$\text{Kurt}[\partial_{\phi_i} \log q_\phi] = \frac{\mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^4]}{(\mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^2])^2}, \tag{15}$$

and assume that it is bounded, $\text{Kurt}[\partial_{\phi_i} \log q_\phi] < \infty$. *Then, the ratio between the control variate correction* δ^{CV} *and the expected control variate coefficient of VarGrad can be upper bounded by*

$$\left| \frac{\delta_i^{CV}}{\mathbb{E}_{q_\phi}[a^{VarGrad}]} \right| \leq \frac{2\sqrt{C \, \text{Kurt}[\partial_{\phi_i} \log q_\phi]}}{\left| \sqrt{\text{KL}(q_\phi(z) \, || \, p(z \, | \, x))} - \frac{\log p(x)}{\sqrt{\text{KL}(q_\phi(z) \, || \, p(z \, | \, x))}} \right|}. \tag{16}$$

Proof. See Appendix A.3. □

Remark 2. The variational approximation $q_\phi(z)$ typically underestimates the spread of the posterior $p(z \, | \, x)$ [Blei et al., 2017], and so the assumption in Eq. 14 is typically satisfied in practice after a few iterations of the optimisation algorithm. The kurtosis $\text{Kurt}[\partial_{\phi_i} \log q_\phi]$ quantifies the weight of the tails of the variational approximation in terms of the score function. In Appendix A.6 we analyse the kurtosis of exponential family distributions and show that it is uniformly bounded for Gaussian variational families.

Remark 3. The upper bound in Eq. 16 allows us to identify two regimes. When $\text{KL}(q_\phi(z) \, || \, p(z \, | \, x))$ is large, the bound asserts that the relative error satisfies

$$\left| \frac{\delta_i^{CV}}{\mathbb{E}_{q_\phi}[a^{VarGrad}]} \right| \lessapprox \mathcal{O}\left(\text{KL}(q_\phi(z) \, || \, p(z \, | \, x))^{-1/2} \right), \tag{17}$$

as the second term in the denominator of Eq. 16 becomes negligible. This can happen in the early stages of the optimisation process, in which case we can conclude that δ^{CV} is expected to be small. Since the KL divergence increases with the dimensionality of the latent variable z (see Appendix A.7), Eq. 16 also implies that the ratio becomes smaller as the number of latent variables grows. Moreover, if the *minimum* KL divergence between the variational family and the true posterior is large (i.e., if the best candidate in the variational family is still far away from the target), the correction term δ_i^{CV} can be negligible during the whole optimisation procedure, which is often the case in practice.

In the regime where $\text{KL}(q_\phi(z) \, || \, p(z \, | \, x))$ approaches zero (i.e., towards the end of the optimisation process if the variational family is well specified and includes the posterior), then Eq. 16 implies that

$$\left| \frac{\delta_i^{CV}}{\mathbb{E}_{q_\phi}[a^{VarGrad}]} \right| \lessapprox \mathcal{O}\left(\text{KL}(q_\phi(z) \, || \, p(z \, | \, x))^{1/2} \right). \tag{18}$$

In this regime, the error w.r.t. the optimal control variate coefficient decreases with the KL divergence. The estimates in Eq. 17 and Eq. 18 combined suggest that the relative error remains bounded throughout the optimisation. We verify this proposition experimentally in Section 6.

4.2 Variance of the Estimator

In this section we provide a result that guarantees that the variance of $\widehat{g}_{\text{VarGrad}}$ is smaller than the variance of $\widehat{g}_{\text{Reinforce}}$ when the number of Monte Carlo samples is large enough.

Proposition 3. *Consider the two gradient estimators* $\widehat{g}_{Reinforce}(\phi)$ *and* $\widehat{g}_{VarGrad}(\phi)$, *each with S Monte Carlo samples, as defined in Eq. 2 and Eq. 9, respectively. If*

$$-\frac{\delta_i^{CV}}{\mathbb{E}_{q_\phi}[a^{VarGrad}]} = \frac{\delta_i^{CV}}{\text{ELBO}(\phi)} < \frac{1}{2} \tag{19}$$

then there exists $S_0 \in \mathbb{N}$ *such that*

$$\text{Var}\left(\widehat{g}_{VarGrad,i}(\phi)\right) \leq \text{Var}\left(\widehat{g}_{Reinforce,i}(\phi)\right), \qquad \text{for all} \quad S \geq S_0. \tag{20}$$

Proof. See Appendix A.4. □

If the correction δ^{CV} is negligible in the sense of Proposition 2, then the assumption in Eq. 19 is satisfied and Proposition 3 guarantees that VarGrad has lower variance than Reinforce when S is large enough. We arrive at the following corollary, which also considers the dimensionality of the latent variables. The main assumption -that the KL-divergence increases with the dimension of the latent space- is supported by the result in Appendix A.7.

Corollary 1. *Let S be the number of samples and D the dimension of the latent variable z. Furthermore, let the assumptions of Proposition 2 be satisfied and assume that $\mathrm{KL}(q_\phi(z) \,||\, p(z \,|\, x))$ is strictly increasing in D. Then, there exist $S_0, D_0 \in \mathbb{N}$ such that*

$$\mathrm{Var}\left(\widehat{g}_{VarGrad,i}(\phi)\right) \leq \mathrm{Var}\left(\widehat{g}_{Reinforce,i}(\phi)\right), \qquad \text{for all } S \geq S_0 \text{ and } D \geq D_0. \tag{21}$$

Proof. See Appendix A.5. □

We provide further intuition on the condition in Eq. 19 with the analysis in Appendix C.1.

5 Related Work

In the last few years, many gradient estimators of the ELBO have been proposed; see Mohamed et al. [2019] for a comprehensive review. Among those, the score function estimators [Williams, 1992, Carbonetto et al., 2009, Paisley et al., 2012, Ranganath et al., 2014] and the reparameterisation estimators [Kingma and Welling, 2014, Rezende et al., 2014, Titsias and Lázaro-Gredilla, 2014], as well as combinations of both [Ruiz et al., 2016, Naesseth et al., 2017], are arguably the most widely used. NVIL [Mnih and Gregor, 2014] and MuProp [Gu et al., 2016] are unbiased gradient estimators for training stochastic neural networks.

Other gradient estimators are specific for discrete-valued latent variables. The concrete relaxation [Maddison et al., 2017, Jang et al., 2017] described a way to form a biased estimator of the gradient, which REBAR [Tucker et al., 2017] and RELAX [Grathwohl et al., 2018] use as a control variate to obtain an unbiased estimator. Other recent estimators have been proposed by Lee et al. [2018], Peters and Welling [2018], Shayer et al. [2018], Cong et al. [2019], Yin and Zhou [2019], Yin et al. [2019], and Dong et al. [2020]. In Section 6, we compare VarGrad with some of these estimators, showing that it exhibits a favourable performance versus computational complexity trade-off.

The VarGrad estimator was first introduced by Salimans and Knowles [2014] and Kool et al. [2019]. It also relates to VIMCO [Mnih and Rezende, 2016] in that it is a leave-one-out estimator. In this paper, we have described an alternative derivation of VarGrad, based on the log-variance loss.

The log-variance loss from Section 3.1 defines an alternative divergence between the approximate and the exact posterior distributions. In the context of optimal control of diffusion processes and related forward-backward stochastic differential equations, it arises naturally to quantify the discrepancy between measures on path space [Nüsken and Richter, 2020]. Other forms of alternative divergences have also been explored in previous work; for example the χ^2-divergence [Dieng et al., 2017], the Rényi divergence [Li and Turner, 2016], the Langevin-Stein [Ranganath et al., 2016], the α-divergence [Hernández-Lobato et al., 2016], other f-divergences [Wang et al., 2018], a contrastive divergence [Ruiz and Titsias, 2019], and also the inclusive KL [Naesseth et al., 2020], see also Appendix D.

Finally, from an implementation perspective, Algorithm 1 contains a `stop_gradient` operator that resembles the method of Roeder et al. [2017]. In Roeder et al. [2017], this operator is used on the variational parameters to eliminate the entropy term in the reparameterisation gradient to reduce its variance; this has also been extended to importance weighted variational objectives [Tucker et al., 2018]. In contrast, VarGrad applies the `stop_gradient` operator on the samples and is based on the score function method.

6 Experiments

In order to verify the properties of VarGrad empirically, we test it on two popular models: a Bayesian logistic regression model on a synthetic dataset and a discrete variational autoencoder (DVAE)
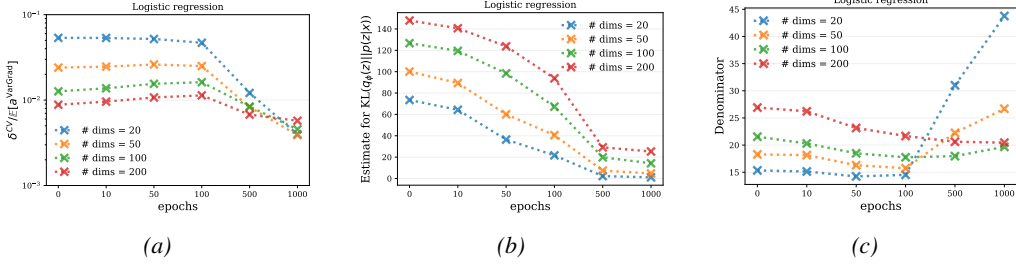
Figure 1. Illustration of Proposition 2 and Remark 3 on the logistic regression model. In (a), we show that the ratio $\left|\delta_i^{\mathrm{CV}}/\mathbb{E}_{q_\phi}[a^{\mathrm{VarGrad}}]\right|$ is small and uniformly bounded over epochs, illustrating that the VarGrad estimator stays close to the optimal control variate coefficients during the whole optimisation procedure. Additionally, this ratio decreases with increasing dimensionality of the latent variables. In (b), we display an estimate of the KL divergence across epochs and demonstrate the beneficial effect of higher dimensions, since the bound of Eq. 16 is expected to scale like $\mathcal{O}(\mathrm{KL}^{-1/2})$ in the early phase. In (c), we plot an estimate of the denominator of the bound (Eq. 16), which increases or stays constant over epochs, demonstrating that the ratio in Eq. 16 stays stable (and small) over epochs.
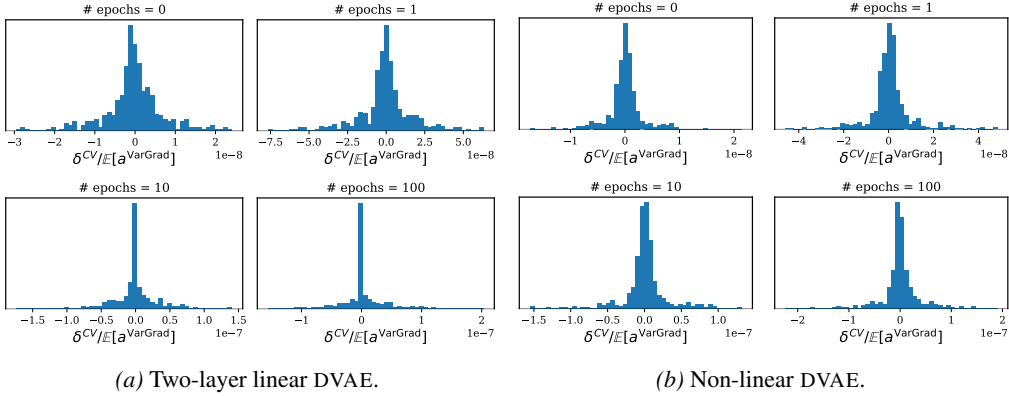

(a) Two-layer linear DVAE.   *(b)* Non-linear DVAE.

Figure 2. The distribution of $\frac{\delta_i^{\mathrm{CV}}}{\mathbb{E}[a^{\mathrm{VarGrad}}]}$ associated with the biases of two DVAE models with 200 latent dimensions trained on Omniglot using VarGrad. The estimates are obtained with 2,000 Monte Carlo samples. The ratio $\frac{\delta_i^{\mathrm{CV}}}{\mathbb{E}[a^{\mathrm{VarGrad}}]}$ is consistently small throughout the optimisation procedure.

[Salakhutdinov and Murray, 2008, Kingma and Welling, 2014] on a fixed binarisation of Omniglot [Lake et al., 2015]. All details of the experiments can be found in Appendix B.[§]

Closeness to the optimal control variate. In Section 4 we analytically showed that VarGrad is close to the optimal control variate, and in particular that the ratio $\left|\delta_i^{\mathrm{CV}}/\mathbb{E}_{q_\phi}[a^{\mathrm{VarGrad}}]\right|$ can be small over the whole optimisation procedure. This behaviour is expected to be even more pronounced with growing dimensionality of the latent space. In Figure 1, we confirm this result by showing the ratio $\left|\delta_i^{\mathrm{CV}}/\mathbb{E}_{q_\phi}[a^{\mathrm{VarGrad}}]\right|$ for the logistic regression model. We also show the KL divergence along the iterations and the denominator of the bound in Eq. 16; see Figure 1 for the details.

In Figure 2, we provide further evidence that this ratio is also small when fitting DVAEs. Indeed, we observe that the ratio $\delta_i^{\mathrm{CV}}/\mathbb{E}[a^{\mathrm{VarGrad}}]$ is typically very small and is distributed around zero during the whole optimisation procedure.

Variance reduction and computational cost. In Figure 3 we show the variance of different gradient estimators throughout the optimisation in the logistic regression setting. We realise a significant improvement of VarGrad compared to the standard Reinforce estimator (Eq. 2). In fact, we observe a small difference between the variance of VarGrad and the variance of an *oracle estimator* based on Reinforce with access to the optimal control variate coefficient a^*. Figure 3 also shows the variance

---

[§]Code in JAX [Bradbury et al., 2018, Hennigan et al., 2020] is available at `https://github.com/aboustati/vargrad`.
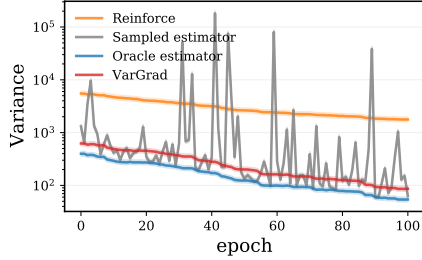
Figure 3. Estimates of the variance of the gradient component w.r.t. the posterior mean of one of the weights for the logistic regression model. The variance of VarGrad is close to the *oracle estimator* based on Reinforce with access to the optimal control variate coefficient a^*. Moreover, the *sampled estimator* (based on Reinforce with an estimate of a^*) shows the difficulty of estimating the optimal control variate coefficient in practice.
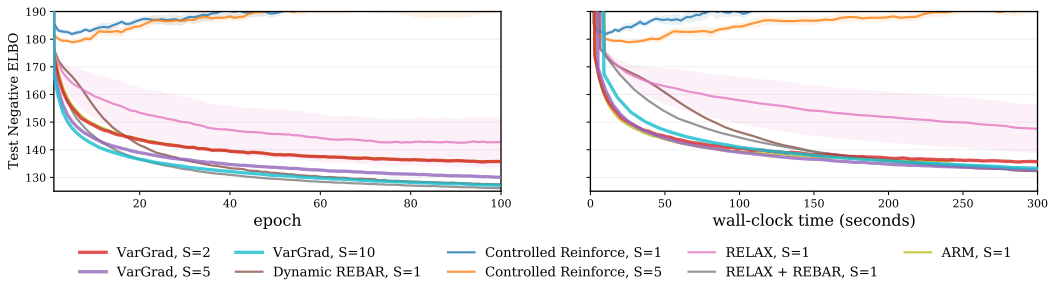

Figure 4. Optimisation trace versus epoch (left) and wall-clock time (right) for a two-layer linear DVAE on a fixed binarisation of Omniglot. The plot compares VarGrad to Reinforce with score function control variates [Ranganath et al., 2014], dynamic REBAR [Tucker et al., 2017], RELAX, RELAX + REBAR [Grathwohl et al., 2018] and ARM [Yin and Zhou, 2019]. The number of samples used to compute each gradient estimator is given in the figure legend. VarGrad demonstrates favourable scalability and performance when compared to the other estimators.

of the *sampled estimator*, which is based on Reinforce with an estimate of the optimal control variate; this confirms the difficulty of estimating it in practice. (A similar trend can be observed for the DVAE in the results in Appendix B, where VarGrad is compared to a wider list of estimators from the DVAE literature.) All methods use $S = 4$ Monte Carlo samples, and the control variate coefficient is estimated with either 2 extra samples (*sampled estimator*) or 1,000 samples (*oracle estimator*).

Finally, Figure 4 compares VarGad with other estimators by training a DVAE on Omniglot. The figure shows the negative ELBO as a function of the epoch number (left plot) and against the wall-clock time (right plot). The negative ELBO is computed on the standard test split and the optimisation uses Adam [Kingma and Ba, 2015] with learning rate of 0.001. VarGrad achieves similar performance to state-of-the-art estimators, such as REBAR [Tucker et al., 2017], RELAX [Grathwohl et al., 2018], and ARM [Yin and Zhou, 2019], while being simpler to implement (see Algorithm 1) and without any tunable hyperparameters.

7 Conclusions

We have analysed the VarGrad estimator, an estimator of the gradient of the KL that is based on Reinforce with leave-one-out control variates, which was first introduced by Salimans and Knowles [2014] and Kool et al. [2019]. We have established the connection between VarGrad and a novel divergence, which we call the log-variance loss. We have showed theoretically that, under certain conditions, the VarGrad control variate coefficients are close to the optimal ones. Moreover, we have established the conditions that guarantee that VarGrad exhibits lower variance than Reinforce. We leave it for future work to explore the direct optimisation of the log-variance loss for alternative choices of the reference distribution $r(z)$.

Acknowledgments

Broader Impact

Variational inference algorithms are approximate inference methods used in many practical applications, including computational neuroscience, natural language processing, and computer vision; see Blei et al. [2017]. The performance of variational inference often depends on the variance of the gradient estimator of its objective. High-variance estimators can make the resulting algorithm unstable and unreliable to use in real-world deployment scenarios; hence, deeper theoretical and empirical understanding of different gradient estimators is crucial for safe applicability of these methods in real-world settings.

In our work, we provide an analysis of the VarGrad estimator. We show the connection between this estimator and the "log-variance loss" and demonstrate its relationship to the optimal score function control variate for Reinforce, providing a theoretical analysis. We support our theoretical analysis with empirical results. We believe our work contributes in the further understanding of variational inference methods and contributes to improving their safety and applicability.

While our theoretical results are suggestive, we warn that the bounds depend on the nature of the model at hand. Therefore, the use of the results here while ignoring the model-specific aspects of our assumptions may lead to some risks in application.

References

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017.

J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, and S. Wanderman-Milne. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/google/jax.

P. Carbonetto, M. King, and F. Hamze. A stochastic approximation method for inference in probabilistic graphical models. In *Advances in Neural Information Processing Systems*, 2009.

Y. Cong, M. Zhao, K. Bai, and L. Carin. GO gradient for expectation-based objectives. In *International Conference on Learning Representations*, 2019.

T. M. Cover and J. A. Thomas. *Elements of Information Theory*. John Wiley & Sons, 2012.

A. B. Dieng, D. Tran, R. Ranganath, J. Paisley, and D. M. Blei. Variational inference via χ-upper bound minimization. In *Advances in Neural Information Processing Systems*, 2017.

Z. Dong, A. Mnih, and G. Tucker. DisARM: An antithetic gradient estimator for binary latent variables. In *Advances in Neural Information Processing Systems*, 2020.

S. Ghosal, J. K. Ghosh, A. W. Van Der Vaart, et al. Convergence rates of posterior distributions. *Annals of Statistics*, 28(2):500–531, 2000.

W. Grathwohl, D. Choi, Y. Wu, G. Roeder, and D. Duvenaud. Backpropagation through the void: Optimizing control variates for black-box gradient estimation. In *International Conference on Learning Representations*, 2018.

S. Gu, S. Levine, I. Sutskever, and A. Mnih. MuProp: Unbiased backpropagation for stochastic neural networks. In *International Conference on Machine Learning*, 2016.

T. Hennigan, T. Cai, T. Norman, and I. Babuschkin. Haiku: Sonnet for JAX, 2020. URL http://github.com/deepmind/dm-haiku.

Y. Hernández-Lobato, J. M. amd Li, M. Rowland, D. Hernández-Lobato, T. Bui, and R. E. Turner. Black-box α-divergence minimization. In *International Conference on Machine Learning*, 2016.

E. Jang, S. Gu, and B. Poole. Categorical reparameterization with Gumbel-softmax. In *International Conference on Learning Representations*, 2017.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Machine Learning*, 37(2):183–233, Nov. 1999.

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

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *International Conference on Learning Representations*, 2014.

W. Kool, H. van Hoof, and M. Welling. Buy 4 REINFORCE samples, get a baseline for free! In *ICLR Workshop on Deep Reinforcement Learning Meets Structured Prediction*, 2019.

W. Kool, H. van Hoof, and M. Welling. Estimating gradients for discrete random variables by sampling without replacement. In *International Conference on Learning Representations*, 2020.

B. M. Lake, R. Salakhutdinov, and J. B. Tenenbaum. Human-level concept learning through probabilistic program induction. *Science*, 350(6266):1332–1338, 2015.

W. Lee, H. Yu, and H. Yang. Reparameterization gradient for non-differentiable models. In *Advances in Neural Information Processing Systems*, 2018.

Y. Li and R. E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, 2016.

C. J. Maddison, A. Mnih, and Y. W. Teh. The concrete distribution: A continuous relaxation of discrete random variables. In *International Conference on Learning Representations*, 2017.

A. Mnih and K. Gregor. Neural variational inference and learning in belief networks. In *International Conference on Machine Learning*, 2014.

A. Mnih and D. J. Rezende. Variational inference for Monte Carlo objectives. In *International Conference on Machine Learning*, 2016.

S. Mohamed, M. Rosca, M. Figurnov, and A. Mnih. Monte Carlo gradient estimation in machine learning. *arXiv preprint arXiv:1906.10652*, 2019.

C. Naesseth, F. J. R. Ruiz, S. Linderman, and D. M. Blei. Reparameterization gradients through acceptance-rejection methods. In *Artificial Intelligence and Statistics*, 2017.

C. A. Naesseth, F. Lindsten, and D. M. Blei. Markovian score climbing: Variational inference with kl$(p||q)$. *arXiv preprint arXiv:2003.10374*, 2020.

N. Nüsken and L. Richter. Solving high-dimensional Hamilton-Jacobi-Bellman PDEs using neural networks: perspectives from the theory of controlled diffusions and measures on path space. *arXiv preprint arXiv:2005.05409*, 2020.

J. W. Paisley, D. M. Blei, and M. I. Jordan. Variational Bayesian inference with stochastic search. In *International Conference on Machine Learning*, 2012.

J. W. T. Peters and M. Welling. Probabilistic binary neural networks. *arXiv preprint arXiv:1809.03368*, 2018.

R. Ranganath, S. Gerrish, and D. M. Blei. Black box variational inference. In *Artificial Intelligence and Statistics*, 2014.

R. Ranganath, J. Altosaar, D. Tran, and D. M. Blei. Operator variational inference. In *Advances in Neural Information Processing Systems*, 2016.

R.-D. Reiss. *Approximate distributions of order statistics: with applications to nonparametric statistics*. Springer science & business media, 2012.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, 2014.

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

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

F. J. R. Ruiz and M. K. Titsias. A contrastive divergence for combining variational inference and MCMC. In *International Conference on Machine Learning*, 2019.

F. J. R. Ruiz, M. K. Titsias, and D. M. Blei. The generalized reparameterization gradient. In *Advances in Neural Information Processing Systems*, 2016.

R. Salakhutdinov and I. Murray. On the quantitative analysis of deep belief networks. In *Proceedings of the 25th international conference on Machine learning*, pages 872–879, 2008.

T. Salimans and D. A. Knowles. On using control variates with stochastic approximation for variational bayes and its connection to stochastic linear regression. *arXiv preprint arXiv:1401.1022*, 2014.

O. Shayer, D. Levi, and E. Fetaya. Learning discrete weights using the local reparameterization trick. In *International Conference on Learning Representations*, 2018.

M. K. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational Bayes for non-conjugate inference. In *International Conference on Machine Learning*, 2014.

G. Tucker, A. Mnih, C. J. Maddison, and J. Sohl-Dickstein. REBAR: low-variance, unbiased gradient estimates for discrete latent variable models. In *International Conference on Learning Representations*, 2017.

G. Tucker, D. Lawson, S. Gu, and C. J. Maddison. Doubly reparameterized gradient estimators for Monte Carlo objectives. In *International Conference on Learning Representations*, 2018.

D. Wang, H. Liu, and Q. Liu. Variational inference with tail-adaptive f-divergence. In *Advances in Neural Information Processing Systems*, 2018.

R. J. Williams. Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3–4):229–256, 1992.

M. Yin and M. Zhou. ARM: Augment-REINFORCE-merge gradient for stochastic binary networks. In *International Conference on Learning Representations*, 2019.

M. Yin, Y. Yue, and M. Zhou. ARSM: Augment-REINFORCE-swap-merge estimator for gradient backpropagation through categorical variables. In *International Conference on Machine Learning*, 2019.

Supplementary Material

A Proof of Paper Results

A.1 Proof of Proposition 1

Proof. We first consider the gradient of the KL divergence. It is given by

$$\nabla_\phi \text{KL}(q_\phi(z) \,||\, p(z\,|\,x)) = \int \nabla_\phi q_\phi(z) \,\mathrm{d}z + \int \log\left(\frac{q_\phi(z)}{p(z\,|\,x)}\right) \nabla_\phi q_\phi(z) \,\mathrm{d}z, \qquad (22)$$

where we can drop the first term since $\int \nabla_\phi q_\phi(z) \,\mathrm{d}z = \nabla_\phi \int q_\phi(z) \,\mathrm{d}z = \nabla_\phi(1) = 0$.

We now consider the gradient of the log-variance loss. Using the definition from Eq. 5, we see that

$$\nabla_\phi \mathcal{L}_r(q_\phi(z) \,||\, p(z\,|\,x)) = \frac{1}{2}\nabla_\phi \int \log^2\left(\frac{q_\phi(z)}{p(z\,|\,x)}\right) r(z) \,\mathrm{d}z - \frac{1}{2}\nabla_\phi \left(\int \log\left(\frac{q_\phi(z)}{p(z\,|\,x)}\right) r(z) \,\mathrm{d}z\right)^2$$

$$= \int \log\left(\frac{q_\phi(z)}{p(z\,|\,x)}\right)\frac{\nabla_\phi q_\phi(z)}{q_\phi(z)} r(z) \,\mathrm{d}z - \left(\int \log\left(\frac{q_\phi(z)}{p(z\,|\,x)}\right) r(z) \,\mathrm{d}z\right)\left(\int \frac{\nabla_\phi q_\phi(z)}{q_\phi(z)} r(z) \,\mathrm{d}z\right).$$

When we evaluate the gradient at $r(z) = q_\phi(z)$, the right-most term vanishes, since $\int \frac{\nabla_\phi q_\phi(z)}{r(z)} r(z) \,\mathrm{d}z = \int \nabla_\phi q_\phi(z) \,\mathrm{d}z = 0$. Thus, the gradient of the log-variance loss becomes equal to the gradient of the KL divergence. $\square$

A.2 Proof of Lemma 1

Proof. First, notice that $\text{Var}_{q_\phi}(\partial_{\phi_i} \log q_\phi) = \mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^2]$ since $\mathbb{E}_{q_\phi}[\partial_{\phi_i} \log q_\phi] = 0$. We then compute

$$a_i^* = \frac{\mathbb{E}_{q_\phi}\left[f_\phi(\partial_{\phi_i} \log q_\phi)^2\right]}{\mathbb{E}_{q_\phi}\left[(\partial_{\phi_i} \log q_\phi)^2\right]} \qquad (23)$$

$$= \frac{\mathbb{E}_{q_\phi}\left[f_\phi(\partial_{\phi_i} \log q_\phi)^2\right] - \mathbb{E}_{q_\phi}\left[f_\phi\right]\mathbb{E}_{q_\phi}\left[(\partial_{\phi_i} \log q_\phi)^2\right] + \mathbb{E}_{q_\phi}\left[f_\phi\right]\mathbb{E}_{q_\phi}\left[(\partial_{\phi_i} \log q_\phi)^2\right]}{\mathbb{E}_{q_\phi}\left[(\partial_{\phi_i} \log q_\phi)^2\right]} \qquad (24)$$

$$= \mathbb{E}_{q_\phi}[\bar{f}_\phi] + \delta_i^{\text{CV}}. \qquad (25)$$

In the last line we have used the fact that $\mathbb{E}_{q_\phi}[f_\phi] = \mathbb{E}_{q_\phi}[\bar{f}_\phi]$. $\square$

A.3 Proof of Proposition 2

Proof. Note that

$$\left|\frac{\delta_i^{\text{CV}}}{\mathbb{E}_{q_\phi}[a^{\text{VarGrad}}]}\right| = \left|\frac{\text{Cov}_{q_\phi}\left(f_\phi, (\partial_{\phi_i} \log q_\phi)^2\right)}{\mathbb{E}_{q_\phi}[f_\phi]\text{Var}_{q_\phi}(\partial_{\phi_i} \log q_\phi)}\right| = \left|\frac{\mathbb{E}_{q_\phi}\left[(f_\phi - \mathbb{E}_{q_\phi}[f_\phi])(\partial_{\phi_i} \log q_\phi)^2\right]}{\mathbb{E}_{q_\phi}[f_\phi]\mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^2]}\right|, \qquad (26)$$

where we have used the fact that $\mathbb{E}_{q_\phi}[\partial_{\phi_i} \log q_\phi] = 0$. From

$$\mathbb{E}[f_\phi] = -\text{ELBO}(\phi) = \text{KL}(q_\phi(z) \,||\, p(z|x)) - \log p(x),$$

and using the Cauchy-Schwarz inequality, Eq. 26 can be bounded from above by

$$\frac{\left(\text{Var}_{q_\phi}\left(\log\frac{q_\phi(z)}{p(z|x)}\right)\right)^{1/2}}{|\text{KL}(q_\phi(z) \,||\, p(z|x)) - \log p(x)|}\left(\frac{\mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^4]}{(\mathbb{E}_{q_\phi}[(\partial_{\phi_i} \log q_\phi)^2])^2}\right)^{1/2}. \qquad (27)$$

The second factor equals $\sqrt{\text{Kurt}[\partial_{\phi_i} \log q_\phi]}$. To bound the first factor, notice that

$$\left(\text{Var}_{q_\phi}\left(\log\frac{q_\phi(z)}{p(z|x)}\right)\right)^{1/2} \leq \left(\mathbb{E}_{q_\phi}\left[\log^2\frac{q_\phi(z)}{p(z|x)}\right]\right)^{1/2} = \left(\mathbb{E}_{q_\phi}\left[\log^2\frac{p(z|x)}{q_\phi(z)}\right]\right)^{1/2} \qquad (28)$$

$$\leq \left(2\mathbb{E}_{q_\phi}\left[\exp\left(\left|\log\frac{p(z|x)}{q_\phi(z)}\right|\right) - 1 - \left|\log\frac{p(z|x)}{q_\phi(z)}\right|\right]\right)^{1/2}, \qquad (29)$$

where we have used the estimate

$$e^x - 1 - x = \sum_{n=0}^{\infty} \frac{x^n}{n!} - 1 - x = \sum_{n=2}^{\infty} \frac{x^n}{n!} \geq \frac{1}{2}x^2, \quad x \geq 0, \tag{30}$$

with $x = \left| \log \frac{p(z|x)}{q_\phi(z)} \right|$. We now use [Ghosal et al., 2000, Lemma 8.3] to bound Eq. 29 from above by

$$2\sqrt{C}h(q_\phi(z) \,||\, p(z|x)), \tag{31}$$

where

$$h(q_\phi(z) \,||\, p(z|x)) = \sqrt{\int \left(\sqrt{q_\phi(z)} - \sqrt{p(z|x)} \right)^2 \, \mathrm{d}z} \tag{32}$$

is the Hellinger distance. From [Reiss, 2012, Lemma A.3.5] we have the bound $h(q_\phi(z) \,||\, p(z|x)) \leq \sqrt{\mathrm{KL}(q_\phi(z) \,||\, p(z|x))}$. Combining these estimates we arrive at the claimed result. $\square$

A.4 Proof of Proposition 3

Proof. We start by defining the short-cuts

$$A = f_\phi(z), \qquad B = (\partial_{\phi_i} \log q_\phi)(z). \tag{33}$$

Let us compute the difference of the variances of the estimators to leading order in S, namely

$$\mathrm{Var}(\widehat{g}_{\text{Reinforce},i}) - \mathrm{Var}(\widehat{g}_{\text{VarGrad},i}) = \frac{1}{S} \mathrm{Var}(AB) + \frac{S-2}{S(S-1)} \mathbb{E}\left[(A - \mathbb{E}[A])(B - \mathbb{E}[B]) \right]^2 \tag{34a}$$

$$- \frac{\mathrm{Var}(A)\,\mathrm{Var}(B)}{S(S-1)} - \frac{1}{S} \mathbb{E}\left[(A - \mathbb{E}[A])^2 (B - \mathbb{E}[B])^2 \right] \tag{34b}$$

$$= \frac{1}{S} \left(\mathbb{E}\left[A^2 B^2 \right] - \mathbb{E}[AB]^2 \right) + \frac{S-2}{S(S-1)} \mathbb{E}[AB]^2 \tag{34c}$$

$$- \frac{1}{S} \left(\mathbb{E}\left[A^2 B^2 \right] - 2\,\mathbb{E}[A]\,\mathbb{E}\left[AB^2 \right] + \mathbb{E}[A]^2\,\mathbb{E}\left[B^2 \right] \right) + \mathcal{O}\left(\frac{1}{S^2} \right) \tag{34d}$$

$$= -\frac{1}{S(S-1)} \mathbb{E}[AB]^2 - \frac{1}{S} \mathbb{E}[A] \left(\mathbb{E}[A]\,\mathbb{E}\left[B^2 \right] - 2\,\mathbb{E}\left[AB^2 \right] \right) + \mathcal{O}\left(\frac{1}{S^2} \right) \tag{34e}$$

$$= \frac{1}{S} \mathbb{E}[A] \left(2\,\mathbb{E}\left[AB^2 \right] - \mathbb{E}[A]\,\mathbb{E}\left[B^2 \right] \right) + \mathcal{O}\left(\frac{1}{S^2} \right) \tag{34f}$$

$$= \frac{1}{S} \mathbb{E}[A]\,\mathbb{E}\left[B^2 \right] \left(2\delta_i^{\text{CV}} + \mathbb{E}[A] \right) + \mathcal{O}\left(\frac{1}{S^2} \right) \tag{34g}$$

and we note that with $\mathbb{E}\left[B^2 \right] > 0$ the leading term is positive if

$$\mathbb{E}[A]\delta_i^{\text{CV}} + \frac{1}{2} \mathbb{E}[A]^2 > 0, \tag{34h}$$

which is equivalent to the statement in the proposition. $\square$

A.5 Proof of Corollary 1

Proof. Note that with Proposition 2 we have

$$\left| \frac{\delta_i^{\text{CV}}}{\mathbb{E}_{q_\phi}[a^{\text{VarGrad}}]} \right| \to 0 \tag{35}$$

for $D \to \infty$, assuming that $\mathrm{KL}(q_\phi(z) \,||\, p(z\,|\,x))$ is strictly increasing in D. Therefore, for large enough D, the condition from Proposition 3 (see Eq. 19), is fulfilled and the statement follows immediately. $\square$

A.6 Results on the Kurtosis of the Score for Exponential Families

Here we provide a more explicit expression for $\mathrm{Kurt}[\partial_{\phi_i} \log q_\phi]$ in the case when $q_\phi(z)$ is given by an exponential family, i.e. $q_\phi(z) = h(z) \exp\left(\phi^\top T(z) - A(\phi)\right)$, where $T(z)$ is the vector of sufficient statistics and $A(\phi)$ denotes the log-partition function. As an application, we show that in the Gaussian case, $\mathrm{Kurt}[\partial_{\phi_i} \log q_\phi]$ is uniformly bounded across the whole variational family.

Lemma 2. *Let $q_\phi(z) = h(z) \exp\left(\phi^\top T(z) - A(\phi)\right)$. Then*

$$\mathrm{Kurt}[\partial_{\phi_i} \log q_\phi] = \frac{\mathbb{E}_{q_\phi}\left[(T_i(z) - m_i)^4\right]}{\mathbb{E}_{q_\phi}\left[(T_i(z) - m_i)^2\right]^2}, \tag{36}$$

where $m_i = \mathbb{E}_{q_\phi}[T_i(z)]$ denotes the mean of the sufficient statistics. In particular, $\mathrm{Kurt}[\partial_{\phi_i} \log q_\phi]$ does not depend on $h(z)$ or $A(\phi)$.

Proof. The claim follows by direct calculation. Indeed,

$$\partial_{\phi_i} \log q_\phi(z) = T_i(z) - \frac{\partial A}{\partial \phi_i}(\phi). \tag{37}$$

It is left to show that $\frac{\partial A}{\partial \phi_i}(\phi) = \mu_i$. For this, notice that the normalisation condition

$$\int h(z) \exp\left(\phi^\top T(z) - A(\phi)\right) \mathrm{d}z = 1 \tag{38}$$

implies

$$\int h(z) \left(T_i(z) - \frac{\partial A(\phi)}{\partial \phi_i}\right) \exp\left(\phi^\top T(z) - A(\phi)\right) \mathrm{d}z = 0 \tag{39}$$

by taking the derivative w.r.t. ϕ_i. The left-hand side equals $\mathbb{E}_{q_\phi}[T_i(z)] - \frac{\partial A}{\partial \phi_i}(\phi)$, and so the claim follows. $\square$

Lemma 3. *Let $q_\phi(z)$ be the family of one-dimensional Gaussian distributions. Then there exists a constant $K > 0$ such that*

$$\mathrm{Kurt}[\partial_{\phi_i} \log q_\phi] < K \tag{40}$$

for all i and all $\phi \in \Phi$. In fact, it is possible to take $K = 15$.

Proof. For the Gaussian family, the sufficient statistics are given by $T_1(z) = z$ and $T_2(z) = z^2$. We have that

$$\frac{\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}\left[(T_1(z) - m_1)^4\right]}{\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}\left[(T_1(z) - m_1)^2\right]^2} = \frac{\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}\left[(z - \mu)^4\right]}{\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}\left[(z - \mu)^2\right]^2} = 3, \tag{41}$$

by the well-known fact the standard kurtosis of any univariate Gaussian is 3. A lengthy but straightforward computation shows that

$$\frac{\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}\left[(T_2(z) - m_2)^4\right]}{\mathbb{E}_{\mathcal{N}(\mu,\sigma^2)}\left[(T_2(z) - m_2)^2\right]^2} = \frac{3(4\mu^4 + 20\mu^2\sigma^2 + 5\sigma^4)}{(2\mu^2 + \sigma^2)^2}, \tag{42}$$

which is maximised for $\mu = 0$, taking the value 15. $\square$

Lemma 3 shows that the kurtosis term in our bound Eq. 16 can be bounded for Gaussian families. This result is expected to extend to the multivariate cases as well. We note that we observe in our experiments that the bound is finite in a variety of cases.

A.7 Dimension-dependence of the KL-divergence

The following lemma shows that the KL-divergence increases with the number of dimensions. This result follows from the chain-rule of KL divergence, see, e.g., Cover and Thomas [2012].

Lemma 4. *Let* $u^{(D)}(z_1, \ldots, z_D)$ *and* $v^{(D)}(z_1, \ldots, z_D)$ *be two arbitrary probability distributions on* $\mathbb{R}^D$. *For* $J \in \{1 \ldots, D\}$ *denote their marginals on the first* J *coordinates by* $u^{(J)}$ *and* $v^{(J)}$, *i.e.*

$$u^{(J)}(z_1, \ldots, z_J) = \int \cdots \int u^{(D)}(z_1, \ldots, z_D) \, \mathrm{d}z_{J+1} \ldots \mathrm{d}z_D, \tag{43}$$

and

$$v^{(J)}(z_1, \ldots, z_J) = \int \cdots \int v^{(D)}(z_1, \ldots, z_D) \, \mathrm{d}z_{J+1} \ldots \mathrm{d}z_D. \tag{44}$$

Then

$$\mathrm{KL}(u^{(1)} \mid\mid v^{(1)}) \leq \mathrm{KL}(u^{(2)} \mid\mid v^{(2)}) \leq \ldots \leq \mathrm{KL}(u^{(D)} \mid\mid v^{(D)}), \tag{45}$$

i.e. the function $J \mapsto \mathrm{KL}(u^{(J)} \mid\mid v^{(J)})$ *is increasing.*

B Details of Experimental Setup and Additional Results

B.1 Details of the Experiments

B.1.1 Logistic Regression

This section describes the experimental setup of the Bayesian logistic regression example which was discussed in the main text in Section 6.

Data. We use a synthetic dataset with $N = 100$, where input-output pairs are generated as follows: we sample a design matrix $X \in \mathbb{R}^{N \times D}$ for the inputs uniformly on $[-1, 1]$, random weights $\mathbf{w} \in \mathbb{R}^D$ from $\mathcal{N}(0, 25 \operatorname{Id}_{D \times D})$ and a random bias $b \in \mathbb{R}$ from $\mathcal{N}(0, 1)$. We set $\mathbf{p} = \sigma(X\mathbf{w} + b\mathbf{1})$, where $\mathbf{1}$ is an N-dimensional vector of ones and $\sigma(x) = \frac{1}{1+\exp(-x)}$ is the logistic sigmoid applied elementwise. Finally, we sample the outputs $Y \sim \operatorname{Bernoulli}(\mathbf{p})$.

Approximate Posterior. For this model, we set the approximate posterior to a diagonal Gaussian with free mean and log standard deviation parameters.

Training. For all the experiments listed in the main text, we use the VarGrad estimator for the gradients of the logistic regression models. We train the models using stochastic gradient descent [Robbins and Monro, 1951] with a learning rate of 0.001.

Estimation of intractable quantities. To estimate the intractable quantities in Figure 1, we use Monte Carlo sampling with 2000 samples for δ^{CV} and $\mathbb{E}_{q_\phi}[a^{\mathrm{VarGrad}}]$. We estimate the KL divergence with the identity $\operatorname{KL}(q_\phi(z) \| p(z|x)) = \log p(x) - \operatorname{ELBO}(\phi)$, where $\log p(x)$ is estimated using importance sampling with 10000 samples and $\operatorname{ELBO}(\phi)$ using standard Monte Carlo sampling with 2000 samples.

For the variance estimates in Figure 3, we use 1000 Monte Carlo samples. As explained in the main text, to estimate the control variate coefficients, we use either 2 samples for the *sampled estimator* or 1000 samples for the *oracle estimator*.

B.2 Discrete VAEs

This section describes the experimental setting for the Discrete VAE, where we closely follow the setup in Maddison et al. [2017], which was also replicated in Tucker et al. [2017] and Grathwohl et al. [2018]. As we are comparing the usefulness of different estimators in the optimisation and time their run-times, we opted to re-implement the various methods using JAX [Bradbury et al., 2018, Hennigan et al., 2020]. Extra care was taken to be as faithful as possible to the implementation description in the respective papers as well as in optimising the run-time of the implementations.

Data. We use a fixed binarisation of Omniglot [Lake et al., 2015], where we binarise at the standard cut-off of 0.5. We use the standard train/test splits for this dataset.

Model Architectures. For the DVAE experiments we use the *two layers linear* architecture, which has 2 stochastic binary layers with 200 units each, which was used in Maddison et al. [2017]. For this model, the decoders mirror the corresponding encoders. We use a Bernoulli(0.5) prior on the latent space and fix its parameters throughout the optimisation.

Approximate Posterior. We use an amortised mean-field Bernoulli approximation for the posterior.

Training. For training the models, we use the Adam optimiser [Kingma and Ba, 2015] with learning rates 0.001, 0.0005 and 0.0001.

Estimation of intractable quantities . We use Monte Carlo sampling with 2000 samples for δ^{CV} and $\mathbb{E}_{q_\phi}[a^{\mathrm{VarGrad}}]$. Due to the high memory requirements of these computations and sparsity of the weight gradients, we only compute them for the biases. To estimate the gradient variances we use 1000 Monte Carlo samples.

B.3 Additional Results for DVAEs

B.3.1 Variance Reduction

In Figure B.5 we present additional results on the practical variance reduction that VarGrad induces in the two layer linear DVAE. Here, we compare with various other estimators from the literature. VarGrad achieves considerable variance reduction over the adaptive (RELAX) and non-adaptive (Controlled Reinforce) model-agnostic estimators. Structured adaptive estimators such as Dynamic REBAR and RELAX + REBAR start with a higher variance at the beginning of optimisation, which reduces towards the end. ARM, which uses antithetic sampling, achieves the most reduction; however, it is only applicable to models with Bernoulli latent variables. Notably, the extra variance reduction seen in some of the methods does not translate to better optimisation performance on this example as seen in Appendix B.3.2.
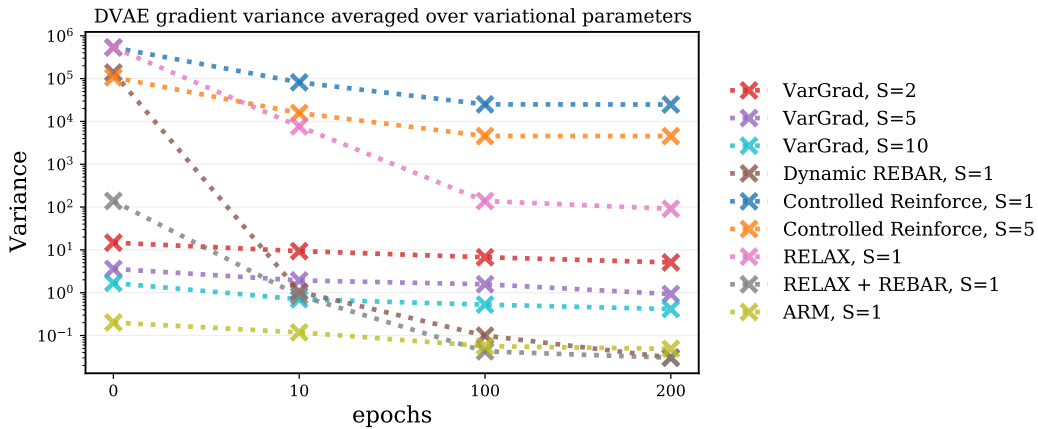

Figure B.5. Estimates of the gradient variance of the DVAE at 4 points during the optimisation for different gradient estimators. The plot compares VarGrad to Reinforce with score function control variates [Ranganath et al., 2014], dynamic REBAR [Tucker et al., 2017], RELAX, RELAX + REBAR [Grathwohl et al., 2018] and ARM [Yin and Zhou, 2019]. The number of samples used to compute each gradient estimator is given in the figure legend.

B.3.2 Performance in Optimisation

In this section we present additional results on training the DVAE with VarGrad. Figure B.6 replicates Figure 4 with a longer run-time. Figure B.7 and Figure B.8 show the optimisation traces for different Adam learning rates.
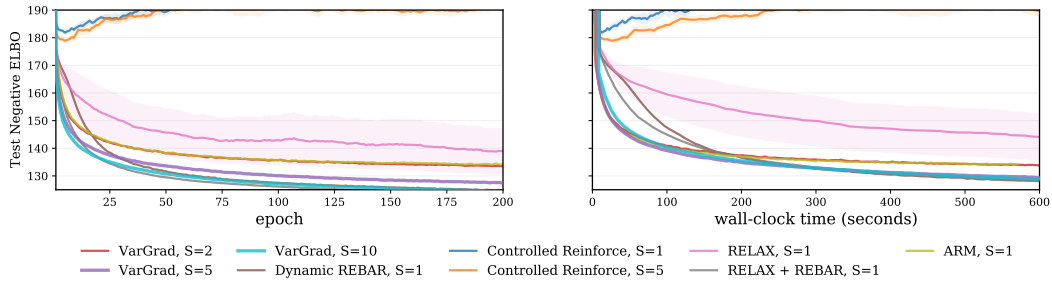
Figure B.6. Optimisation trace versus epoch (left) and wall-clock time (right) for a two-layer linear DVAE on a fixed binarisation of Omniglot, trained with Adam with a learning rate of 0.001. The plot compares VarGrad to Reinforce with score function control variates [Ranganath et al., 2014], dynamic REBAR [Tucker et al., 2017], RELAX, RELAX + REBAR [Grathwohl et al., 2018] and ARM [Yin and Zhou, 2019]. The number of samples used to compute each gradient estimator is given in the figure legend. The results here are identical to the ones in Figure 4 but with a longer run-time.
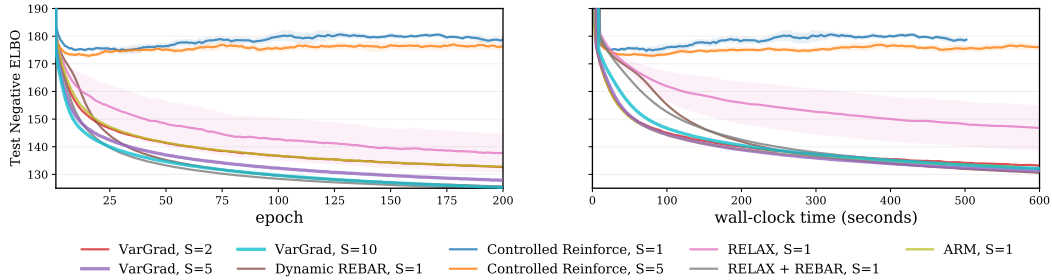

Figure B.7. Optimisation trace versus epoch (left) and wall-clock time (right) for a two-layer linear DVAE on a fixed binarisation of Omniglot, trained with Adam with a learning rate of 0.0005. The plot compares VarGrad to Reinforce with score function control variates [Ranganath et al., 2014], dynamic REBAR [Tucker et al., 2017], RELAX, RELAX + REBAR [Grathwohl et al., 2018] and ARM [Yin and Zhou, 2019]. The number of samples used to compute each gradient estimator is given in the figure legend.
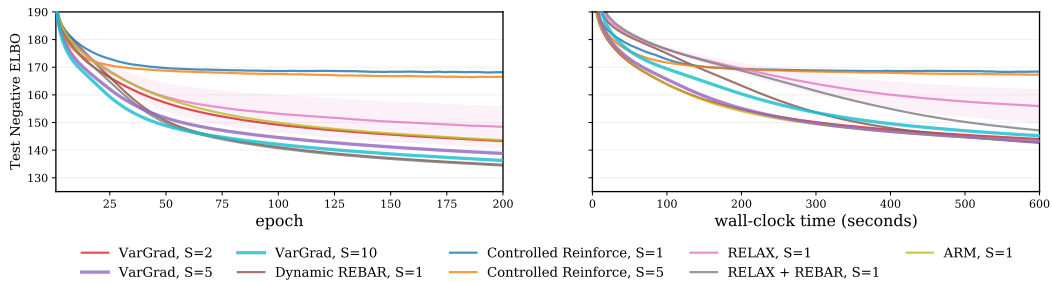

Figure B.8. Optimisation trace versus epoch (left) and wall-clock time (right) for a two-layer linear DVAE on a fixed binarisation of Omniglot, trained with Adam with a learning rate of 0.0001. The plot compares VarGrad to Reinforce with score function control variates [Ranganath et al., 2014], dynamic REBAR [Tucker et al., 2017], RELAX, RELAX + REBAR [Grathwohl et al., 2018] and ARM [Yin and Zhou, 2019]. The number of samples used to compute each gradient estimator is given in the figure legend.

C Gaussians

In the case when $q(z)$ and $p(z|x)$ are (diagonal) Gaussians we can gain some intuition on the performance of VarGrad by computing the relevant quantities analytically. The principal insights obtained from the examples presented in this section can be summarised as follows: Firstly, in certain scenarios the Reinforce estimator does indeed exhibit a lower variance in comparison with VarGrad (although the advantage is very modest and only materialises for a restricted set of parameters). This finding illustrates that the conditions in Eq. 19 and Eq. 20 (the latter referring to $S \geq S_0$) cannot be dropped without replacement from the formulation of Proposition 3. Secondly, in line with the results from Section 6, the relative error $\delta_i^{\mathrm{CV}} / \mathbb{E}[a^{\mathrm{VarGrad}}]$ decreases with increased dimensionality. Moreover, the variance associated to computing the optimal control variate coefficients a^* is significant and increases considerably with the number of latent variables.

C.1 Comparing the Variances of Reinforce and VarGrad

In order to understand when the variance of VarGrad is smaller than the variance of the Reinforce estimator we first consider the one-dimensional Gaussian case $q(z) = \mathcal{N}(z; \mu, \sigma^2)$ and $p(z|x) = \mathcal{N}(z; \widetilde{\mu}, \widetilde{\sigma}^2)$ and analyse the derivative w.r.t. μ. A lengthy calculation shows that

$$\Delta_{\mathrm{Var}}(\mu, \widetilde{\mu}, \sigma^2, \widetilde{\sigma}^2, S) := \mathrm{Var}(\widehat{g}_{\mathrm{Reinforce},\mu}) - \mathrm{Var}(\widehat{g}_{\mathrm{VarGrad},\mu}) \tag{46a}$$

$$= \frac{1}{4S\sigma^4\widetilde{\sigma}^2} \left((\mu - \widetilde{\mu})^4 + 2(\mu - \widetilde{\mu})^2 \left(\frac{3S-7}{S-1}\sigma^2 - 3\widetilde{\sigma}^2 \right) + \frac{5S-7}{S-1} \left(\sigma^2 - \widetilde{\sigma}^2 \right)^2 \right) \tag{46b}$$

$$\approx \frac{1}{4S\sigma^4\widetilde{\sigma}^2} \left(\Delta_\mu^4 + 6\Delta_\mu^2 \Delta_{\sigma^2} + 5\Delta_{\sigma^2}^2 \right), \tag{46c}$$

where the last line holds for large S with $\Delta_\mu := \mu - \widetilde{\mu}$ and $\Delta_{\sigma^2} := \sigma^2 - \widetilde{\sigma}^2$.

For an illustration, let us vary the above parameters. First, let us fix $\sigma^2 = \widetilde{\sigma}^2 = 1$. We note from Eq. 46c that in this case we expect VarGrad to have lower variance regardless of Δ_μ as long as S is large enough. In Figure C.9 we see that this is in fact the case, however a different result can be observed for small S, which is again in accordance with Eq. 46b.

Next, we consider arbitrary σ^2 and $\widetilde{\sigma}^2$, but fixed $\mu = 1, \widetilde{\mu} = 2$. In Figure C.10 we observe that the variance of VarGrad is smaller for most values of σ^2 and $\widetilde{\sigma}^2$. However, even for large S there remains a region where the Reinforce estimator is superior. In fact, one can compute the condition for this to happen to be $\Delta_{\sigma^2} \in \left[-\Delta_\mu^2, -\frac{1}{5}\Delta_\mu^2 \right]$, which can be compared with the condition in Eq. 19 in Proposition 3.

In Figure C.11 we display the variance differences Δ_{Var} as functions of Δ_μ and Δ_{σ^2}, approximated according to Eq. 46c, for the same fixed values as before and see that they are bounded from below, but not from above.

For a D-dimensional Gaussian it is hard to compute the condition from Eq. 19 in full generality, but we can derive the following stronger criterion that can guarantee better performance of VarGrad when assuming that $\mathrm{ELBO}(\phi) \leq 0$ (which for instance holds in the discrete-data setting).

Lemma 5. *Assume* $\mathrm{ELBO}(\phi) \leq 0$ *and*

$$\mathrm{Cov}\left(f_\phi, \left(\partial_{\phi_i} \log q_\phi \right)^2 \right) > 0. \tag{47}$$

Then there exists $S_0 \in \mathbb{N}$ *such that*

$$\mathrm{Var}\left(\widehat{g}_{\mathit{VarGrad},i}(\phi) \right) \leq \mathrm{Var}\left(\widehat{g}_{\mathit{Reinforce},i}(\phi) \right), \qquad \text{for all} \quad S \geq S_0. \tag{48}$$

Proof. With $\mathrm{ELBO}(\phi) \leq 0$ we have

$$\mathrm{Cov}\left(f_\phi, \left(\partial_{\phi_i} \log q_\phi \right)^2 \right) \leq \mathbb{E}_{q_\phi}\left[f_\phi \left(\partial_{\phi_i} \log q_\phi \right)^2 \right] - \frac{1}{2} \mathbb{E}_{q_\phi}\left[f_\phi \right] \mathbb{E}_{q_\phi}\left[\left(\partial_{\phi_i} \log q_\phi \right)^2 \right] \tag{49}$$

$$= \mathbb{E}_{q_\phi}\left[\left(\partial_{\phi_i} \log q_\phi \right)^2 \right] \left(\delta_i^{\mathrm{CV}} - \frac{1}{2}\mathrm{ELBO}(\phi) \right). \tag{50}$$

If now

$$\mathrm{Cov}\left(f_\phi, \left(\partial_{\phi_i} \log q_\phi \right)^2 \right) > 0, \tag{51}$$
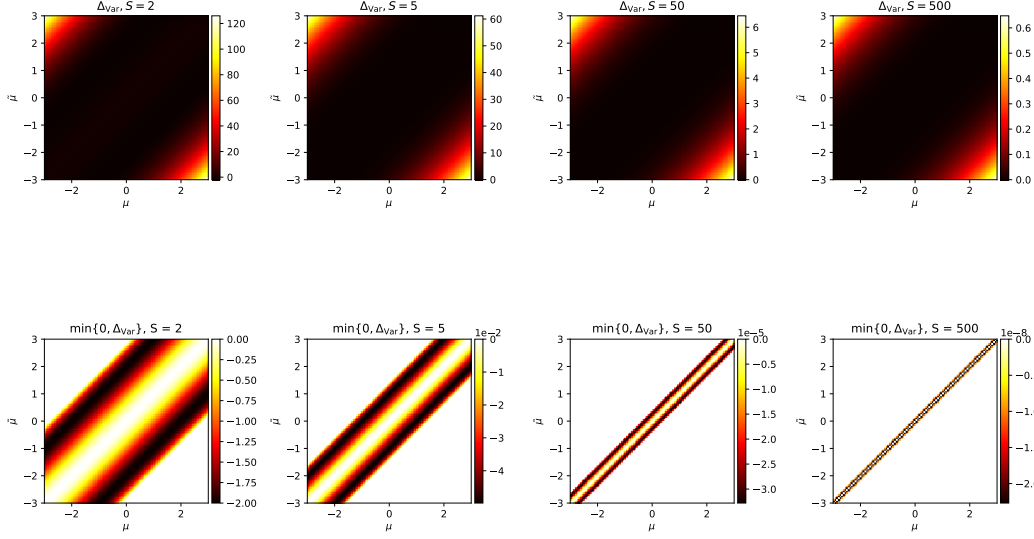
Figure C.9. We compare the variance of the reinforce estimator with the variance of VarGrad. VarGrad is often better even for small S – for large S this can be guaranteed with Proposition 3.

then also

$$\delta_i^{\text{CV}} - \frac{1}{2}\text{ELBO}(\phi) > 0, \tag{52}$$

and the statement follows by Proposition 3. $\square$

The condition from Eq. 47 gives another guarantee for VarGrad having smaller variance than the Reinforce estimator. However, we note that the converse statement is not necessarily true, i.e. if the condition does not hold, VarGrad can still be better. The advantage of Eq. 47, however, is that it can be verified more easily in certain settings, as for instance done for D-dimensional diagonal Gaussians in the following lemma.

Lemma 6 (Covariance term for diagonal Gaussians)**.** *Let $q(z)$ and $p(z|x)$ be diagonal D-dimensional Gaussians with means μ and $\widetilde{\mu}$ and covariance matrices $\Sigma = \text{diag}(\sigma_1^2, \ldots, \sigma_D^2)$ and $\widetilde{\Sigma} = \text{diag}(\widetilde{\sigma}_1^2, \ldots, \widetilde{\sigma}_D^2)$. Then*

$$\text{Cov}_{q_\phi}\left(f_\phi, (\partial_{\phi_k} \log q_\phi)^2\right) = \frac{1}{\widetilde{\sigma}_k^2} - \frac{1}{\sigma_k^2} \tag{53}$$

for $k \in \{1, \ldots, D\}$ and

$$\text{Cov}_{q_\phi}\left(f_\phi, (\partial_{\phi_k} \log q_\phi)^2\right) = \frac{1}{\sigma_k^2}\left(\frac{1}{\widetilde{\sigma}_k^2} - \frac{1}{\sigma_k^2}\right) \tag{54}$$

for $k \in \{D+1, \ldots, 2D\}$ with $\phi = (\mu_1, \ldots, \mu_D, \sigma_1^2, \ldots, \sigma_D^2)^\top$ and

$$\text{Cov}_{q_\phi}\left(f_\phi, (\partial_{\phi_k} \log q_\phi)^2\right) = \frac{1}{\widetilde{\sigma}_k^2} - \frac{1}{\sigma_k^2} \tag{55}$$

for $k \in \{D+1, \ldots, 2D\}$ with $\phi = (\mu_1, \ldots, \mu_D, \log \sigma_1^2, \ldots, \log \sigma_D^2)^\top$.

Proof. We compute

$$f_\phi = -\frac{1}{2}\sum_{i=1}^{D}\log\left(\frac{\sigma_i^2}{\widetilde{\sigma}_i^2}\right) - \frac{1}{2}\sum_{i=1}^{D}\frac{(z_i - \mu_i)^2}{\sigma_i^2} + \frac{1}{2}\sum_{i=1}^{D}\frac{(z_i - \widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2} \tag{56}$$
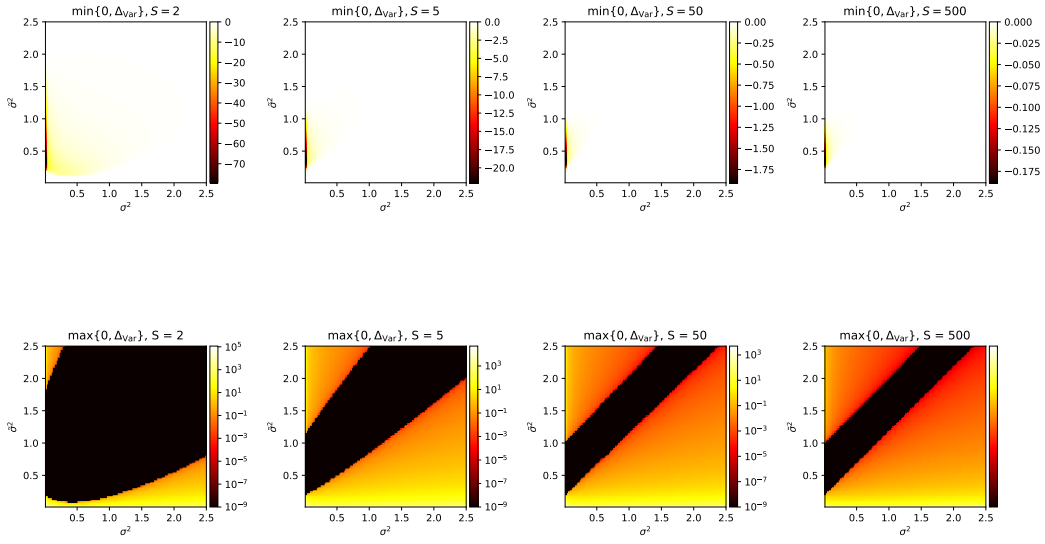
Figure C.10. Variance comparison with varying σ^2 and $\widetilde{\sigma}^2$. VarGrad only wins outside a certain region, however, if so, then potentially by orders of magnitude.
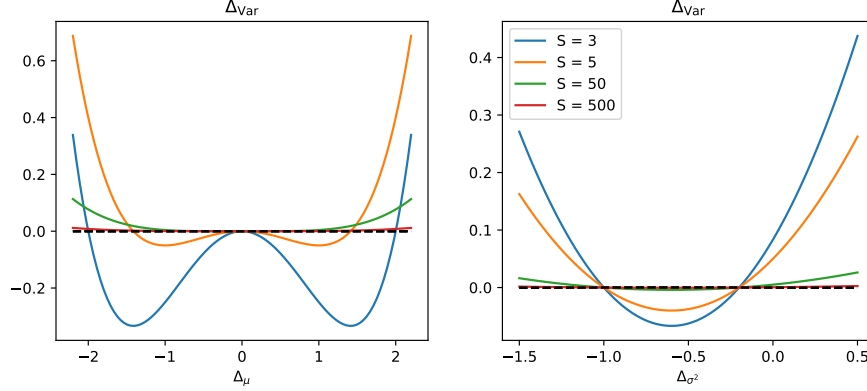

Figure C.11. Variance differences of the reinforce estimator and VarGrad with varying Δ_μ and Δ_{σ^2} for different sample sizes S.

and

$$\partial_{\mu_k} \log q_\phi = \frac{z_k - \mu_k}{\sigma_k^2}. \tag{57}$$

We again use the short-cuts

$$A = f_\phi(z), \qquad B = \left(\partial_{\mu_k} \log q_\phi\right)(z), \tag{58}$$

and obtain

22

$$\mathrm{Cov}_{q_\phi}(A, B^2) = \mathbb{E}_{q_\phi}\left[AB^2\right] - \mathbb{E}_{q_\phi}[A]\,\mathbb{E}_{q_\phi}\left[B^2\right] \tag{59}$$

$$= \mathbb{E}_{q_\phi}\left[\left(-\frac{1}{2}\sum_{i=1}^{D}\log\left(\frac{\sigma_i^2}{\widetilde{\sigma}_i^2}\right) - \frac{1}{2}\sum_{i=1}^{D}\frac{(z_i-\mu_i)^2}{\sigma_i^2} + \frac{1}{2}\sum_{i=1}^{D}\frac{(z_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2}\right)\left(\frac{z_k-\mu_k}{\sigma_k^2}\right)^2\right] \tag{60}$$

$$- \mathbb{E}_{q_\phi}\left[\left(-\frac{1}{2}\sum_{i=1}^{D}\log\left(\frac{\sigma_i^2}{\widetilde{\sigma}_i^2}\right) - \frac{1}{2}\sum_{i=1}^{D}\frac{(z_i-\mu_i)^2}{\sigma_i^2} + \frac{1}{2}\sum_{i=1}^{D}\frac{(z_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2}\right)\right]\mathbb{E}_{q_\phi}\left[\left(\frac{z_k-\mu_k}{\sigma_k^2}\right)^2\right]. \tag{61}$$

$$= -\frac{1}{2}\left(\frac{3}{\sigma_k^2} + \frac{D-1}{\sigma_k^2}\right) + \frac{1}{2}\left(\frac{1}{\sigma_k^2}\sum_{\substack{i=1\\i\neq k}}^{D}\frac{\sigma_i^2+(\mu_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2} + \frac{1}{\sigma_k^2\widetilde{\sigma}_k^2}\left(3\sigma_k^2+(\mu_k-\widetilde{\mu}_k)^2\right)\right) \tag{62}$$

$$- \left(-\frac{D}{2} + \frac{1}{2}\sum_{i=1}^{D}\frac{\sigma_i^2+(\mu_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2}\right)\frac{1}{\sigma_k^2} \tag{63}$$

$$= -\frac{1}{\sigma_k^2} + \frac{1}{2\sigma_k^2\widetilde{\sigma}_k^2}\left(3\sigma_k^2+(\mu-\widetilde{\mu})^2\right) - \frac{1}{2\sigma_k^2\widetilde{\sigma}_k^2}\left(\sigma_k^2+(\mu-\widetilde{\mu})^2\right) \tag{64}$$

$$= \frac{1}{\widetilde{\sigma}_k^2} - \frac{1}{\sigma_k^2}. \tag{65}$$

For the terms with the partial derivative w.r.t. σ_k^2 we first note that

$$\partial_{\sigma_k^2}\log q_\phi = -\frac{1}{2\sigma_k^2} + \frac{(z_k-\mu_k)^2}{2\sigma_k^4} = \frac{1}{\sigma_k^2}\partial_{\log\sigma_k^2}\log q_\phi. \tag{66}$$

We compute

$$\mathbb{E}_{q_\phi}\left[AB^2\right] = \mathbb{E}_{q_\phi}\left[\frac{(z_k-\mu_k)^2}{4\sigma_k^6}\sum_{i=1}^{D}\frac{(z_i-\mu_i)^2}{\sigma_i^2} - \frac{(z_k-\mu_k)^4}{8\sigma_k^8}\sum_{i=1}^{D}\frac{(z_i-\mu_i)^2}{\sigma_i^2}\right. \tag{67}$$

$$\left. - \frac{(z_k-\mu_k)^2}{4\sigma_k^6}\sum_{i=1}^{D}\frac{(z_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2} + \frac{(z_k-\mu_k)^4}{8\sigma_k^8}\sum_{i=1}^{D}\frac{(z_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2}\right] \tag{68}$$

$$= -\frac{1}{8\sigma_k^4}(D+8) + \frac{1}{8\sigma_k^4\widetilde{\sigma}_k^2}\left(9\sigma_k^2+(\mu_k-\widetilde{\mu}_k)^2\right) + \frac{1}{8\sigma_k^4}\sum_{\substack{i=1\\i\neq k}}^{d}\frac{\sigma_i^2+(\mu_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2}, \tag{69}$$

and similarly

$$\mathbb{E}_{q_\phi}[A]\,\mathbb{E}_{q_\phi}\left[B^2\right] = -\frac{D}{8\sigma_k^4} + \frac{1}{8\sigma_k^4\widetilde{\sigma}_k^2}\left(\sigma_k^2+(\mu_k-\widetilde{\mu}_k)^2\right) + \frac{1}{8\sigma_k^4}\sum_{\substack{i=1\\i\neq k}}^{D}\frac{\sigma_i^2+(\mu_i-\widetilde{\mu}_i)^2}{\widetilde{\sigma}_i^2}. \tag{70}$$

We therefore get the result by again computing $\mathrm{Cov}_{q_\phi}(A,B^2) = \mathbb{E}_{q_\phi}\left[AB^2\right] - \mathbb{E}_{q_\phi}[A]\,\mathbb{E}_{q_\phi}\left[B^2\right]$. The partial derivative w.r.t. $\log\sigma_k^2$ can be recovered from Eq. 66. $\square$

C.2 Optimal Control Variates in the Gaussian Case

In the diagonal Gaussian case we can also easily analytically compute the optimal control variate coefficients from Eq. 11, along the lines of the proof of Lemma 6. Our setting is again $q(z) = \mathcal{N}(z;\mu,\Sigma)$, $p(z|x) = \mathcal{N}(z;\widetilde{\mu},\widetilde{\Sigma})$ with $\Sigma = \mathrm{diag}\left(\sigma_1^2,\ldots,\sigma_D^2\right)$, $\widetilde{\Sigma} = \mathrm{diag}\left(\widetilde{\sigma}_1^2,\ldots,\widetilde{\sigma}_D^2\right)$. In Figure C.12 we plot the variances of four different gradient estimators with varying sample size S, namely $\widehat{g}_{\mathrm{Reinforce}}$, $\widehat{g}_{\mathrm{VarGrad}}$, as well as the Reinforce estimator augmented with the optimal control variate, once computed analytically and once sampled using S samples. The variance depends on the mean and the covariance matrix; here we choose $\mu_i = 3$, $\sigma_i^2 = 3$, $\widetilde{\mu}_i = 1$, $\widetilde{\sigma}_i^2 = 1$ for all $i \in \{1,\ldots,D\}$. We observe that the VarGrad estimator is close to the analyitcal optimal control variate, and that the
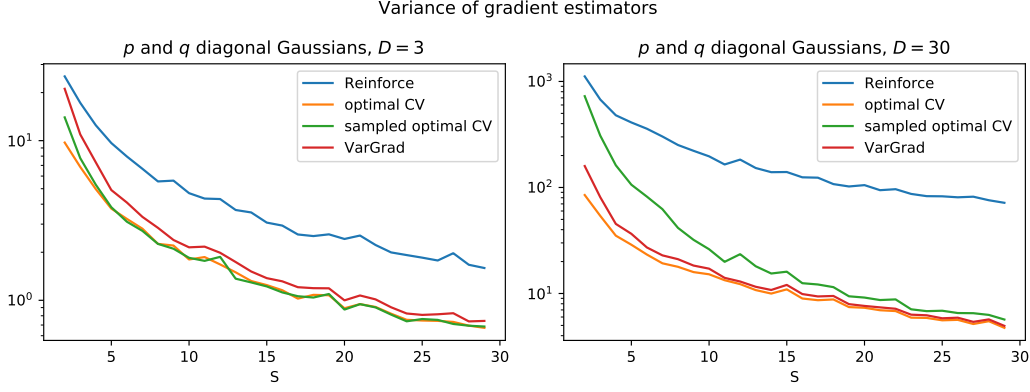
Figure C.12. Comparion of the variances of the different gradient estimators $\widehat{g}_{\text{Reinforce}}$, $\widehat{g}_{\text{VarGrad}}$, as well as the reinforce estimator with the optimal control variate coefficient, once computed analytically and once sampled with S samples, for dimensions $D = 3$ and $D = 30$.
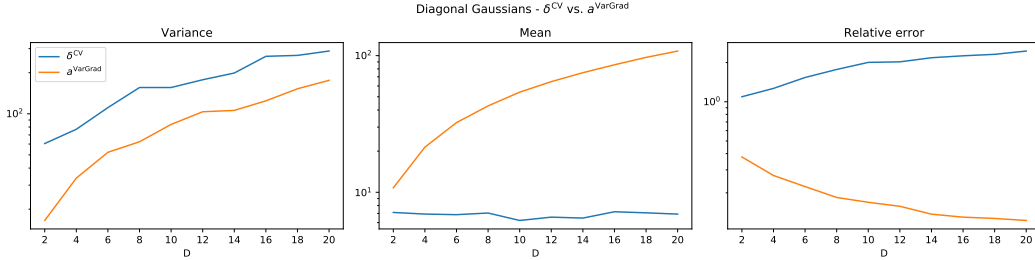

Figure C.13. Mean, variance and relative errors associated to the two contributions to the optimal control variate coefficient, δ_i^{CV} and $a^{\text{VarGrad}} = \bar{f}_\phi$.

sampled optimal control variate performs significantly worse in a small sample size regime. These observations get more pronounced in higher dimensions and indicate that the variance of the sampled optimal control variate can itself be high, showing that using it might not always be beneficial in practice.

Let us additionally investigate the optimal control variate correction term δ_i^{CV} as defined in Eq. 13 for D-dimensional Gaussians $q(z)$ and $p(z|x)$ as considered above. In Figure C.13 we display the variances, means and relative errors of δ_i^{CV} and $a^{\text{VarGrad}} = \bar{f}_\phi$ and realise that indeed the ratio of those two converges to zero when D gets larger. Furthermore we notice that the relative error of δ^{CV} increases with the dimension, explaining the difficulties when estimating the optimal control variate coefficients from samples. Finally, we plot a histogramm of δ_i^{CV} (varying across i) in Figure C.14, showing that δ_i^{CV} is small in comparison to $\mathbb{E}\left[a^{\text{VarGrad}}\right]$ and distributed around zero.

D   Connections to Other Divergences

The Reinforce gradient estimator from Eq. 2 can as well be derived from the *moment loss*

$$\mathcal{L}_r^{\text{moment}}(q_\phi(z)\|p(z|x)) = \frac{1}{2}\,\mathbb{E}_{r(z)}\left[\log^2\left(\frac{q_\phi(z)}{p(z|x)}\right)\right], \tag{71}$$

namely

$$\nabla_\phi \mathcal{L}_r^{\text{moment}}(q_\phi(z)\|p(z|x))\Big|_{r=q_\phi} = \mathbb{E}_{q_\phi}\left[\log\left(\frac{q_\phi(z)}{p(z|x)}\right)\nabla_\phi \log q_\phi(z)\right]. \tag{72}$$
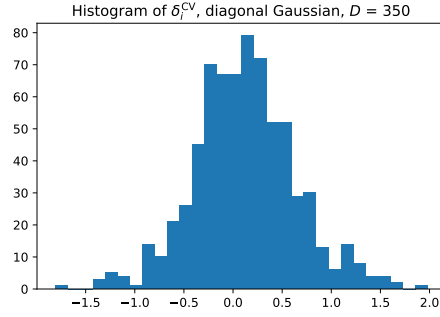
Figure C.14. The histogram of δ_i^{CV} shows that it is usually rather small in comparison to $\mathbb{E}\left[a^{\mathrm{VarGrad}}\right]$, which is roughly 700 here, and that it fluctuates around zero.

In the log-variance loss, one can omit the logarithm to obtain the *variance loss*

$$\mathcal{L}_r^{\mathrm{Var}}(p(z|x)\|q_\phi(z)) = \frac{1}{2}\,\mathrm{Var}_{r(z)}\left(\frac{p(z|x)}{q_\phi(z)}\right) = \frac{1}{2}\,\mathbb{E}_{r(z)}\left[\left(\frac{p(z|x)}{q_\phi(z)}\right)^2 - 1\right], \tag{73}$$

which with $r = q_\phi$ coincides with the χ^2-divergence. The potential of using the latter in the context of variational inference was suggested in Dieng et al. [2017]. We note that again one is in principle free in choosing $r(z)$, but that unlike the log-variance loss, this loss in not symmetric with respect to $q_\phi(z)$ and $p(z|x)$. An analysis in Nüsken and Richter [2020] (for distributions on path space) however suggests that the variance loss (unlike the log-variance loss) scales unfavourably in high-dimensional settings in terms of the variance associated to standard Monte Carlo estimators.