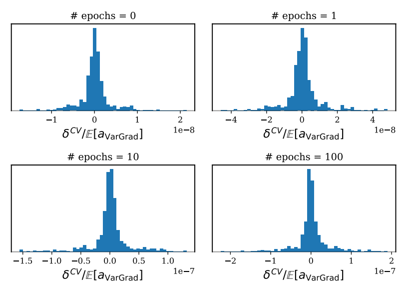
We thank all the reviewers for their stimulating comments and engaging questions. We are happy that they appreciated our theoretical analyses and argued that those, as well as VarGrad's practical usefulness, are of benefit to the NeurIPS community. We now address some questions and comments in the sequel.

R1, R2: The gradient estimator was already known. We agree and did not claim otherwise. We also appreciate that most reviewers did not find this a concern. We emphasise that the novelty in our work is: (i) the derivation of this estimator in terms of the log-variance loss, and (ii) the theoretical analysis of its variance. (i) This new perspective is simpler and allows for a natural interpretation in terms of divergences (which is beneficial in other ML areas, e.g., reinforcement learning). We also note that the connection with the log-variance loss is of practical interest, as it enables a simple implementation algorithm based on automatic differentiation. Moreover, we believe this connection opens the door for further research. (ii) Our work is the first to provide a theoretical analysis of the variance of this estimator. This analysis shows that VarGrad's control variate coefficients are close to the optimal ones for the score function control variate.

R1: What class of distributions satisfies the kurtosis condition of Proposition 2? In general, the kurtosis will not negatively affect the bound as long as the tails of the variational distribution do not become too heavy during the optimisation. More precisely, in Lemma 2 we translated the kurtosis condition to a condition that is more easily verifiable in the case of exponential families and in Lemma 3 we explicitly proved that the condition is fulfilled in the Gaussian case (where the 1-dimensional case carries over to the multidimensional case). In practice, q is often modelled by a Gaussian (potentially with a neural network outputting its mean and covariance). The analysis for the Gaussian case can be found in Section C. Note that in higher dimensions (usually implying that $\mathrm{KL}(q\|p)$ is large, as elaborated in Lemma 4) the kurtosis is allowed to be large (but must be bounded—see also Corollary 1). More importantly, our numerical experiments suggest that the kurtosis condition is satisfied for more complicated cases (see Figures 1, 2).

R1: The order notation might hide large constants; elaborate the significance of Corollary 1. Numerical evidence suggests that, in practice, the ratio $\delta^{\mathrm{CV}}/\mathbb{E}[a_{\mathrm{VarGrad}}]$ is usually very small (e.g., Figure B.4 shows that the order of magnitude is 10^{-8}, and we have observed similar results in all other experiments). Corollary 1 offers a rigorous comparison of the variances of Reinforce and VarGrad that does not depend on assumptions that might be hard to check.

R1, R2, R3, R4: Empirical evaluation. Please see the included figure for an empirical validation of $\delta^{\mathrm{CV}}/\mathbb{E}[a_{\mathrm{VarGrad}}]$ on a non-linear (deep) DVAE model used in Tucker, et. al. [2017] on Omniglot. In the attached plot this quantity is typically very small and fluctuates around zero, validating our analytical results in the challenging non-linear setting (a similar result is shown in Figure B.4 for the two-layer linear case with an in-depth discussion that follows for the non-linear case too). For the non-linear case, we also observe practical performance gains similar to the linear case in the trace plots. The new results have now been added to the manuscript.



R1: Is the log-variance well-defined? Since p, q and r are assumed to admit densities, it follows (under the support condition stated in the paper) that measure-zero sets of r are necessarily measure-zero sets of p and q, implying the divergence property. This has now been clarified in the paper.

R3: VarGrad does not account for model structure. The advantage of VarGrad is that it is easily implementable and applicable to a large array of models (black-box). In Figures 3, B.7 and B.8 we demonstrate that it is still competitive w.r.t. to alternative (tuned and structured) estimators.

R3: Deeper analysis w.r.t. the choice of r. We agree; however, as our work was focused on the theoretical analysis of the induced ELBO gradient estimator rather than a new variational objective, we left this question out intentionally for subsequent studies, in order to not confuse the presentation. Taking $r = q$ after differentiation reproduces the gradient of the KL divergence/ELBO (Proposition 1) and this choice was shown to be effective in the empirical evaluation (Nüsken & Richter [2020] provide further arguments for this choice).

R4: Comparison to reparametrisation. The reparametrisation trick is often not applicable (e.g., in discrete models), whereas VarGrad is general-purpose. That said, in Figures 3, B.7 and B.8 we compared against RELAX + REBAR, which use reparametrisation estimators as control variates.

R4: Application to mixtures of non-diagonal Gaussians. Indeed this is a potential application, since VarGrad can be applied easily to many variational families including non-diagonal mixtures of Gaussians. We will explore this further.

R1: Figure 1 error bars and clarification of the oracle estimator. We have now added the error bars. The oracle estimator in Figure 2 takes 1000 *extra* samples. We have now clarified this in the figure caption.

R1: Why did you focus on the ELBO in the presentation? As per your suggestion, we have now shifted the presentation on the ELBO to streamline the background section. Extra details were moved to the appendix.