Tighter Variational Bounds are Not Necessarily Better

Tom Rainforth* **Tuan Anh Le*** **Maximilian Igl** **Chris J. Maddison**
Yee Whye Teh **Frank Wood**
University of Oxford
{rainforth,cmaddis,y.w.teh}@stats.ox.ac.uk
{tuananh,igl,fwood}@robots.ox.ac.uk

Abstract

We provide theoretical and empirical evidence that using tighter evidence lower bounds (ELBOs) can be detrimental to the process of learning an inference network by reducing the signal-to-noise ratio of the gradient estimator. Our results call into question common implicit assumptions that tighter ELBOs are better variational objectives for simultaneous model learning and inference amortization schemes, suggesting that further investigation is required to assess the relative utility of different approaches. Our results hint that it may be beneficial to use different targets for training the generative and inference networks.

1 Introduction

Variational bounds provide tractable and state-of-the-art objectives for training deep generative models (Kingma and Welling, 2014; Rezende et al., 2014). Typically taking the form of a lower bound on the intractable model evidence, they provide surrogate targets that are more amenable to optimization. In general, this optimization requires the generation of approximate posterior samples during the model training and so a number of methods look to simultaneously learn an inference network alongside the target generative network. This assists in the training process and provides an amortized inference artifact which can be used at test time (Kingma and Welling, 2014). The performance of such methods is critically dependent upon the choice of evidence lower bound (ELBO) and formulation of the required inference, with the two often intricately linked to one another. For example, if the inference network formulation is not sufficiently expressive, this can have a knock-on effect on the generative network unless precautionary steps are taken (Burda et al., 2016).

It is often implicitly assumed in the literature that using tighter ELBOs is universally beneficial and that larger values of the ELBO indicate a better model. In this work we question these implicit assumptions by demonstrating that, although using a tighter ELBO is typically beneficial to gradient updates of the generative network, it can be detrimental to updates of the inference network, which can then impact the generative network at future iterations. Specifically, we present theoretical and empirical evidence that increasing the number of importance sampling particles, K, to tighten the bound in the importance-weighted auto-encoder (IWAE), degrades the signal-to-noise ratio (SNR) of the gradient estimates for the inference network.

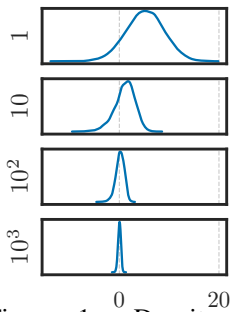


Figure 1: Density of ∇_ϕ ELBO for different K.

An intuitive demonstration of this effect is given in Figure 1. This shows a kernel density estimation for the distribution of the ELBO gradient estimator with respect to the proposal parameter A for the model discussed in Section 4 (with $D = N = 1$) and different K. We see that as we increase K, both the amplitude of the gradient and the standard deviation of the estimator decrease. However, because the former reduces faster,

* Equal Contribution

the SNR deteriorates. This is perhaps easiest to appreciate by noting that for the larger values of K, there is a roughly equal probability of the estimator being positive or negative, such that we are equally likely to increase or decrease the parameter value at the next iteration, inevitably leading to poor performance. On the other hand, when $K = 1$, it is far more likely that the gradient estimator is positive than negative and so there is clear drift to the gradient steps. Note that using a larger K should always give better performance at test time (Cremer et al., 2017) – the implication of our result is that it may be better to learn the inference network using a smaller K during training.

2 Background and Notation

Let x be an $\mathcal{X}$-valued random variable defined via a process involving an unobserved $\mathcal{Z}$-valued random variable z with joint density $p_\theta(x, z)$. Direct maximum likelihood estimation of θ is generally intractable if $p_\theta(x, z)$ is a deep generative model due to the marginalization of z. A common strategy is to instead optimize a variational lower bound on $\log p_\theta(x)$, defined via an auxiliary inference model $q_\phi(z|x)$ as follows

$$\text{ELBO}_{\text{VAE}}(\theta, \phi, x) \coloneqq \int q_\phi(z|x) \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \, \mathrm{d}z = \log p_\theta(x) - \mathrm{KL}(q_\phi(z|x)||p_\theta(z|x)). \quad (1)$$

Here q_ϕ is usually parameterized by a neural network, for which this approach is known as the variational auto-encoder (VAE) (Kingma and Welling, 2014). Optimization of $\text{ELBO}_{\text{VAE}}(\theta, \phi, x)$ is performed with stochastic gradient ascent (SGA) using unbiased estimators of $\nabla_{\theta,\phi} \text{ELBO}_{\text{VAE}}(\theta, \phi, x)$. If q_ϕ is reparameterizable (Kingma and Welling, 2014), then given a reparameterized sample $z \sim q_\phi(z|x)$, the gradients $\nabla_{\theta,\phi} (\log p_\theta(x, z) - \log q_\phi(z|x))$ can be used for the optimization.

The VAE objective places a harsh penalty on mismatch between $q_\phi(z|x)$ and $p_\theta(z|x)$; optimizing jointly in θ, ϕ can confound improvements in $\log p_\theta(x)$ with reductions in the KL (Turner and Sahani, 2011). Thus, a major research direction is the development of bounds that separate the tightness of the bound from the expressiveness of the class of q_ϕ. For example, the IWAE objectives (Burda et al., 2016), which we denote as $\text{ELBO}_{\text{IS}}(\theta, \phi, x)$, are a family of bounds defined by,

$$Q_{\text{IS}}(z_{1:K}|x) \coloneqq \prod_{k=1}^{K} q_\phi(z_k|x), \quad \hat{Z}_{\text{IS}}(z_{1:K}, x) \coloneqq \frac{1}{K} \sum_{k=1}^{K} \frac{p_\theta(x, z_k)}{q_\phi(z_k|x)}, \quad (2)$$

$$\text{ELBO}_{\text{IS}}(\theta, \phi, x) \coloneqq \int Q_{\text{IS}}(z_{1:K}|x) \log \hat{Z}_{\text{IS}}(z_{1:K}, x) \, \mathrm{d}z_{1:K} \leq \log p_\theta(x). \quad (3)$$

The IWAE objectives generalize the VAE objective ($K = 1$ corresponds to the VAE) and the bounds become strictly tighter as K increases (Burda et al., 2016). Furthermore, Burda et al. (2016) provide strong empirical evidence that setting $K > 1$ leads to significant empirical gains over the VAE in terms of learning the generative model. More generally, optimizing tighter bounds is usually empirically associated with better models p_θ in terms of marginal likelihood on held out data. Other related approaches extend this to sequential Monte Carlo (SMC) (Maddison et al., 2017; Le et al., 2017; Naesseth et al., 2017) or change the lower bound that is optimized to reduce the bias (Li and Turner, 2016; Bamler et al., 2017). A second, unrelated, approach is to tighten the bound by improving the expressiveness of q_ϕ (Salimans et al., 2015; Tran et al., 2015; Rezende and Mohamed, 2015; Kingma et al., 2016; Maaløe et al., 2016; Ranganath et al., 2016). In this work we focus on the former, algorithmic approaches to tighter bounds.

3 Assessing the Signal-to-Noise Ratio of the Gradient Estimators

Increasing K in the IWAE leads to empirical improvements in learning the generative network (Burda et al., 2016). However, as shown by, for example, Le et al. (2017), when the family of q_ϕ contains the true posteriors, the global optimum parameters $\{\theta^*, \phi^*\}$ are independent of K. Moreover, in any non-trivial model, we cannot carry out the required optimization analytically, nor even analytically calculate the gradients of the ELBO. The effectiveness of any particular choice of K is thus inextricably linked to our ability to numerically solve the resulting optimization problem. This motivates us to examine the effect K has on the variance and magnitude of the gradient estimates for the two networks. We find that increasing K does indeed improve the gradient estimates for the generative network, but that it can worsen those for the inference network. More precisely, we present a theoretical result showing that the signal-to-noise ratio of the reparameterization gradients of the inference network for the IWAE decreases with rate $O(1/\sqrt{K})$.

As estimating the ELBO requires a Monte Carlo estimation of an expectation over z, we have two samples sizes to tune for the estimate: the number of samples M used for Monte Carlo estimation of the ELBO and the number of importance samples K used in the bound construction. Here M does not change the true value of $\nabla_{\theta,\phi}$ ELBO, only our variance in estimating it, while changing K changes the ELBO itself, with larger K leading to tighter bounds (Burda et al., 2016). Presuming that reparameterization is possible, we can express our gradient estimate in the general form

$$\Delta_{M,K} := \frac{1}{M}\sum_{m=1}^{M} \nabla_{\theta,\phi} \log \hat{Z}_{m,K} \quad \text{where} \quad \hat{Z}_{m,K} = \frac{1}{K}\sum_{k=1}^{K} w_{m,k}, \quad w_{m,k} = \frac{p_\theta(z_{m,k}, x)}{q_\phi(z_{m,k}|x)}, \quad (4)$$

and each $z_{m,k} \overset{i.i.d.}{\sim} q_\phi(z_{m,k}|x)$. Note that the case $K = 1$ corresponds to using the VAE objective, while one typically uses $M = 1$ for IWAE training. We will further use $\Delta_{M,K}(\theta)$ to refer to gradient estimates with respect to θ and $\Delta_{M,K}(\phi)$ to refer to gradients estimates with respect to ϕ.

One might presume that the variance is a good barometer for the effectiveness of a gradient estimation scheme. While this is blatantly true when the expected gradient estimate is fixed, in need not be for changes that simultaneously affect both the variance and expected value of the gradient. For example, because the marginal likelihood estimates $\hat{Z}_{m,K}$ become exact (and thus independent of the proposal) as $K \to \infty$, it must be the case that $\lim_{K\to\infty} \Delta_{M,K}(\phi) = 0$. Thus as K becomes large, the expected value of the gradient must decrease along with its variance, such that the variance relative to the problem scaling need not actually improve.

To investigate this effect more concretely, we introduce the notion of a signal-to-noise-ratio:

$$\text{SNR}_{M,K}(\theta) = \left|\frac{\mathbb{E}\left[\Delta_{M,K}(\theta)\right]}{\sigma\left[\Delta_{M,K}(\theta)\right]}\right| \quad \text{and} \quad \text{SNR}_{M,K}(\phi) = \left|\frac{\mathbb{E}\left[\Delta_{M,K}(\phi)\right]}{\sigma\left[\Delta_{M,K}(\phi)\right]}\right| \quad (5)$$

where $\sigma[\cdot]$ denotes the standard deviation of a random variable and the SNR is defined separately on each dimension of the parameter vector. The SNR is thus the absolute value of the expected gradient scaled by its standard deviation, providing a measure of the relative accuracy of the gradient estimates. Though a high SNR does not always indicate a good SGA scheme (as the target objective itself might be poorly chosen), a low SNR is always problematic because it indicates that the gradient estimates are dominated by noise: if SNR $\to 0$ then the estimates become completely random. We are now ready to state our main theoretical result.

Theorem 1. *Assume that when $M = K = 1$, the expected gradients; the variances of the gradients; and the first four moments of $w_{1,1}$, $\nabla_\theta w_{1,1}$, and $\nabla_\phi w_{1,1}$ are all finite, with the variances also being non-zero. Then the signal-to-noise ratios of the gradient estimates converge at the following rates*

$$\text{SNR}_{M,K}(\theta) = \sqrt{M}\left|\frac{\sqrt{K}\,\nabla_\theta Z - \frac{1}{2Z\sqrt{K}}\nabla_\theta\left(\frac{\text{Var}[w_{1,1}]}{Z^2}\right) + O\left(\frac{1}{K^{3/2}}\right)}{\sqrt{\mathbb{E}\left[w_{1,1}^2\left(\nabla_\theta \log w_{1,1} - \nabla_\theta \log Z\right)^2\right]} + O\left(\frac{1}{K}\right)}\right| = O\left(\sqrt{MK}\right), \quad (6)$$

$$\text{SNR}_{M,K}(\phi) = \sqrt{M}\left|\frac{\nabla_\phi \text{Var}\left[w_{1,1}\right] + O\left(\frac{1}{K}\right)}{2Z\sqrt{K}\,\sigma\left[\nabla_\phi w_{1,1}\right] + O\left(\frac{1}{\sqrt{K}}\right)}\right| = O\left(\sqrt{\frac{M}{K}}\right) \quad (7)$$

where $Z := p_\theta(x)$ is the true marginal likelihood.

Proof. We give only an intuitive demonstration of the high-level result here and provide a formal proof in Appendix A. The effect of M on the SNR ratio follows straightforwardly from using the law of large numbers on the random variable $\nabla_{\theta,\phi} \log \hat{Z}_{m,K}$. Namely, the overall expectation is independent of M and the variance reduces at a rate $O(1/M)$. The effect of K is more complicated but is perhaps most easily seen by noting that (as shown by Burda et al. (2016))

$$\nabla_{\theta,\phi} \log \hat{Z}_{m,K} = \sum_{k=1}^{K} \frac{w_{m,k}}{\sum_{\ell=1}^{K} w_{m,k}} \nabla_{\theta,\phi} \log\left(w_{m,k}\right), \quad (8)$$

such that $\nabla_{\theta,\phi} \log \hat{Z}_{m,K}$ can be interpreted as a self-normalized importance sampling estimate. We can, therefore, invoke the known result (see e.g. Hesterberg (1988)) that the bias of a self-normalized importance sampler converges at a rate $O(1/K)$ and the standard deviation at a rate $O(1/\sqrt{K})$.

We thus see that the SNR converges at a rate $O((1/K)/(1/\sqrt{K})) = O(1/\sqrt{K})$ if the asymptotic gradient is 0 and $O((1)/(1/\sqrt{K})) = O(\sqrt{K})$ otherwise, giving the convergence rates in the ϕ and θ cases respectively. □

The implication of these convergence rates is that increasing M is monotonically beneficial to the SNR for both θ and ϕ, but that increasing K is beneficial to the former and detrimental to the latter. An important point of note is that, for large K, the direction the expected inference network gradient is independent of K. Namely, because we have as an intermediary result from deriving the SNRs that

$$\mathbb{E}\left[\Delta_{M,K}(\phi)\right] = -\frac{1}{2KZ^2}\nabla_\phi \text{Var}\left[w_{1,1}\right] + O\left(\frac{1}{K^2}\right), \tag{9}$$

we see that the direction of the gradient tends towards $-\nabla_\phi \text{Var}\left[w_{1,1}\right] / \left\|\nabla_\phi \text{Var}\left[w_{1,1}\right]\right\|_2$ for large K. This direction is rather interesting: it implies that as $K \to \infty$, the optimal ϕ is that which minimizes the variance of the weights. This is well known to be the optimal importance sampling distribution in terms of approximating the posterior (Owen, 2013). Though it is not also necessarily the optimal proposal in terms of estimating the ELBO,[1] this is nonetheless an interesting equivalence that complements the results of Cremer et al. (2017). It suggests that increasing K may provide a preferable target in terms of the direction of the true inference network gradients, creating a trade-off with the fact that it also diminishes the SNR, reducing the estimates to pure noise if K is set too high. In the absence of other factors, there may thus be a "sweet-spot" for setting K.

Typically when training deep generative models, one does not optimize a single ELBO but instead its average over multiple data points, i.e.

$$\mathcal{J}(\theta, \phi) := \frac{1}{N}\sum_{n=1}^{N} \text{ELBO}_{\text{IS}}(\theta, \phi, x^{(n)}). \tag{10}$$

Our results extend to this setting because the z are drawn independently for each $x^{(n)}$, so

$$\mathbb{E}\left[\frac{1}{N}\sum_{n=1}^{N}\Delta_{M,K}^{(n)}\right] = \frac{1}{N}\sum_{n=1}^{N}\mathbb{E}\left[\Delta_{M,K}^{(n)}\right], \quad \text{Var}\left[\frac{1}{N}\sum_{n=1}^{N}\Delta_{M,K}^{(n)}\right] = \frac{1}{N^2}\sum_{n=1}^{N}\text{Var}\left[\Delta_{M,K}^{(n)}\right] \tag{11}$$

We thus also see that if we are using mini-batches such that N is a chosen parameter and the $x^{(n)}$ are drawn from the empirical data distribution, then the SNRs of $\bar{\Delta}_{N,M,K} := \frac{1}{N}\sum_{n=1}^{N}\Delta_{M,K}^{(n)}$ scales as $\sqrt{N}$, i.e. $\text{SNR}_{N,M,K}(\theta) = O(\sqrt{NMK})$ and $\text{SNR}_{N,M,K}(\phi) = O(\sqrt{NM/K})$. Therefore increasing N has the same ubiquitous benefit as increasing M. In the rest of the paper, we will implicitly be considering the SNRs for $\bar{\Delta}_{N,M,K}$, but will omit the dependency on N to simplify the notation.

4 Experiments

Our convergence results hold exactly in relation to M (and N) but are only asymptomatic in K due to the higher order terms. Therefore their applicability should be viewed with a healthy degree of skepticism in the small K regime. With this in mind, we now present empirical support for our theoretical results and test how well they hold in the small K regime using a simple Gaussian model, for which we can analytically calculate the ground truth.

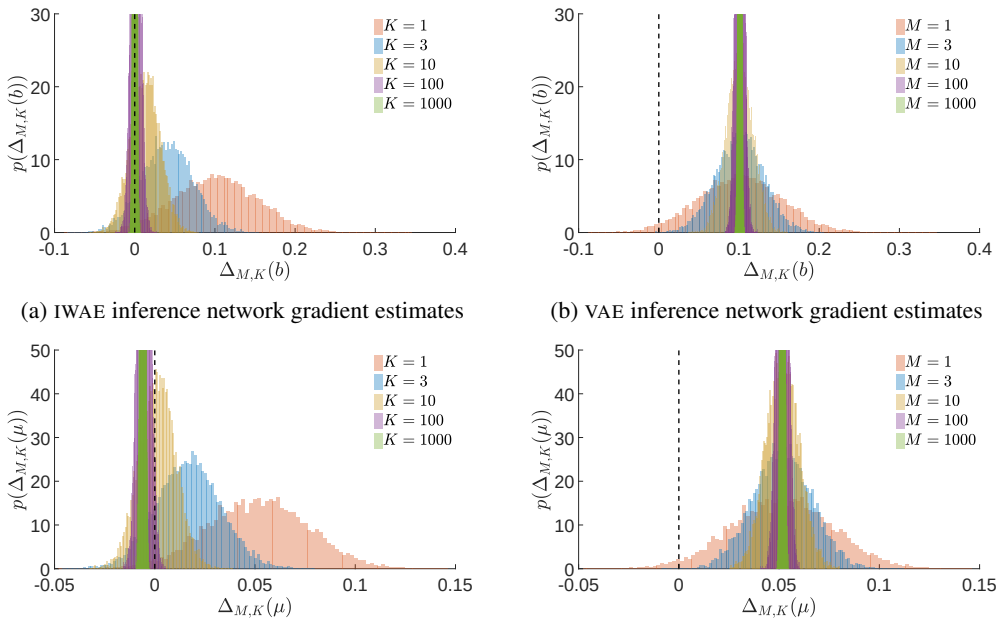
Consider a family of generative models with $\mathbb{R}^D$–valued latent variable z and observed variable x:

$$z \sim \mathcal{N}(z; \mu, I), \qquad x|z \sim \mathcal{N}(x; z, I), \tag{12}$$

which is parameterized by $\theta := \mu$. Let the inference network be parameterized by $\phi = (A, b),\ A \in \mathbb{R}^{D\times D},\ b \in \mathbb{R}^D$ where $q_\phi(z|x) = \mathcal{N}(z; Ax + b, \frac{2}{3}I)$. Given a dataset $(x^{(n)})_{n=1}^N$, we can analytically calculate the optimum of our target $\mathcal{J}(\theta, \phi)$ as explained in Appendix B, giving $\theta^* := \mu^* = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$ and $\phi^* := (A^*, b^*)$, where $A^* = I/2$ and $b^* = \mu^*/2$. For this particular problem, the optimal proposal is independent of K. This will not be the case in general unless the family of possible q_ϕ contains the the true posteriors $p_\theta(z|x^{(n)})$. Further, even for this problem, the expected gradients for the inference network still change with K.

To conduct our investigation, we randomly generated a synthetic dataset from the model with $D = 20$ dimensions, $N = 1024$ data points, and a true model parameter value μ_{true} that was itself randomly

[1] This is because the optimum importance sampling proposal for calculating expectations of a particular function is distinct to that which best approximates the posterior. See, e.g., Owen (2013) and Ruiz et al. (2016).



(a) IWAE inference network gradient estimates

(b) VAE inference network gradient estimates

(c) IWAE generative network gradient estimates

(d) VAE generative network gradient estimates

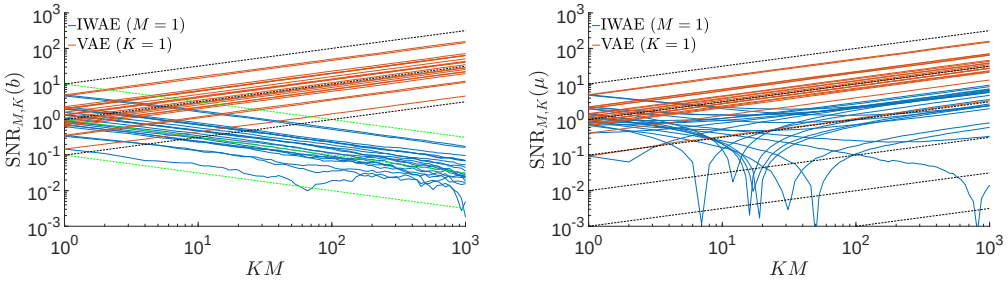
Figure 2: Histograms of gradient estimates $\Delta_{M,K}$ for the generative network and the inference network using the IWAE ($M = 1$) and VAE ($K = 1$) objectives with different values of M and K.

generated from a unit Gaussian, i.e. $\mu_{\text{true}} \sim \mathcal{N}(\mu_{\text{true}}; 0, I)$. We then considered the gradient at a random point in the parameter space close to optimum,[2] namely each dimension of each parameter was randomly offset from its optimum value using a zero-mean Gaussian with standard deviation 0.01. We then calculated empirical estimates of the ELBO gradients for the IWAE, where $M = 1$ is held fixed and we consider increasing K, and for the VAE, where $K = 1$ is held fixed and we consider increasing M. In all cases we calculated 10^4 such estimates and used these samples to provide empirical estimates for, amongst other things, the mean and standard deviation of the estimator, and thereby an empirical estimate for the SNR. For the inference network, we predominantly focused on investigating the gradients of b.

We start by examining the qualitative behavior of the different gradient estimators as K increases as shown in Figure 2. This shows histograms of the gradient estimators for a single parameter of the inference network (top) and generative network (bottom) for the IWAE (left) and the VAE (right). We see that, as expected, the expectation of the gradients does not change with M: the only effect of increasing M is to reduce the variance. The effect of increasing K is quite different. We first see in Figure 2a that as K increases, both the magnitude and the standard deviation of the estimator decrease for the inference network, with the former decreasing faster. This matches the qualitative behavior of our theoretical result, with the SNR ratio diminishing as K increases. In particular, the probability of the gradient being positive or negative becomes roughly even for the larger values of K, meaning the optimizer is equally likely to increase as decrease the inference network parameters at the next iteration. By contrast, for the generative network, the IWAE converges towards a non-zero gradient, such that, even though the SNR initially decreases with K, it then rises again, with a very clear gradient signal for $K = 1000$.

To provide a more rigorous analysis, we next directly examine the convergence of the SNR. Figure 3 shows the convergence of the estimators with increasing M and K. The observed rates for the inference network (Figure 3a) correspond remarkably exactly to our theoretical results, with the suggested rates observed all the way back to $K = M = 1$. Thus, as expected, we see that as M increases, so does $\text{SNR}_{M,K}(b)$, but as K increases, $\text{SNR}_{M,K}(b)$ reduces. In Figure 3b, we see that the theoretical convergence for $\text{SNR}_{M,K}(\mu)$ is again observed exactly for variations in M, but a more

[2]We consider the behavior of for points far away from the optimum in Appendix C, for which the variance of the weights is substantially higher.



(a) Convergence of SNR for inference network

(b) Convergence of SNR for generative network

Figure 3: Convergence of signal-to-noise ratios of gradient estimates with total budget MK. Different lines correspond to different dimensions of the parameter vectors. Shown in blue is the convergence of the IWAE where we keep $M = 1$ fixed and increase K. Shown in red is the convergence of the VAE where $K = 1$ is fixed and we increase M. The black and green dashed lines show the expected convergence rates from our theoretical results, representing gradients of $1/2$ and $-1/2$ respectively.

unusual behavior is seen for variations in K where the SNR initially decreases before starting to increase again for large enough K, eventually exhibiting behavior consistent with the theoretical result for large enough K. The driving factor for this appears to be that, at least for this model, $\mathrm{SNR}_{M,\infty}(\mu)$ typically has a smaller magnitude (and often opposite sign) to $\mathrm{SNR}_{M,1}(\mu)$ (see Figure 2c). If we think of the estimators for all values of K as biased estimates for $\mathrm{SNR}_{M,\infty}(\mu)$, we see from our theoretical results that this bias decreases faster than the standard deviation. Consequently, if reducing this bias causes the magnitude of the expected gradient to diminish, this can mean that increasing K initially causes the SNR to reduce.

Note that this does not mean that the estimates are getting worse for the generative network. As we increase K our bound is getting tighter and our estimates closer to the true gradient for the target that we actually want to optimize, i.e. $\nabla_\mu \log Z$. It is thus perhaps better to measure the quality of the gradient estimates for the generative network by looking at the root mean squared error (RMSE) to $\nabla_\mu \log Z$, i.e. $\sqrt{\mathbb{E}\left[\|\Delta_{M,K} - \nabla_\mu \log Z\|_2^2\right]}$. The convergence of this RMSE is shown in Figure 4 where the solid lines are the RMSE estimates using 10^4 runs and the shaded regions show the interquartile range of the individual estimates. We see that increasing M in the VAE reduces the variance of the estimates but has negligible effect on the RMSE due to the fixed bias. On the other hand, we see that increasing K leads to a monotonic improvement, initially improving at a rate $O(1/K)$ (because the bias is the dominating term in this region), before settling to the standard Monte Carlo convergence rate of $O(1/\sqrt{K})$ (shown by the dashed lines).

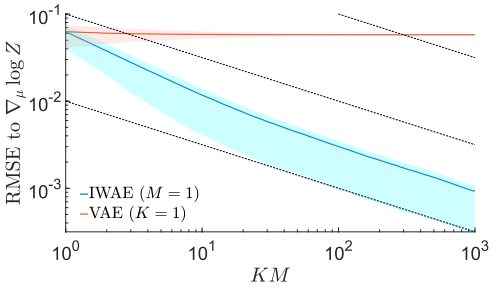
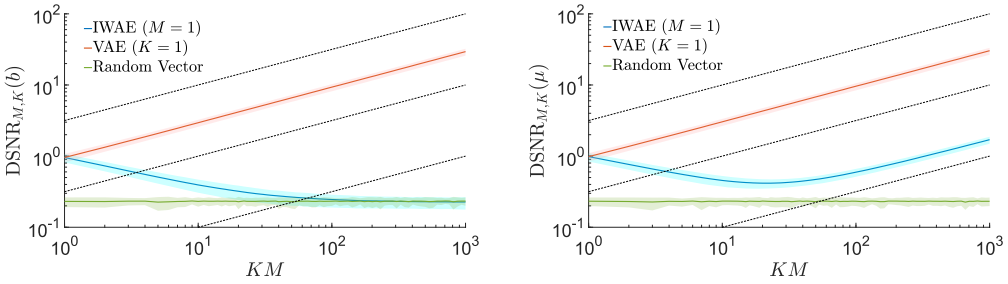


Figure 4: Root mean squared error in μ gradient estimate to $\nabla_\mu \log Z$

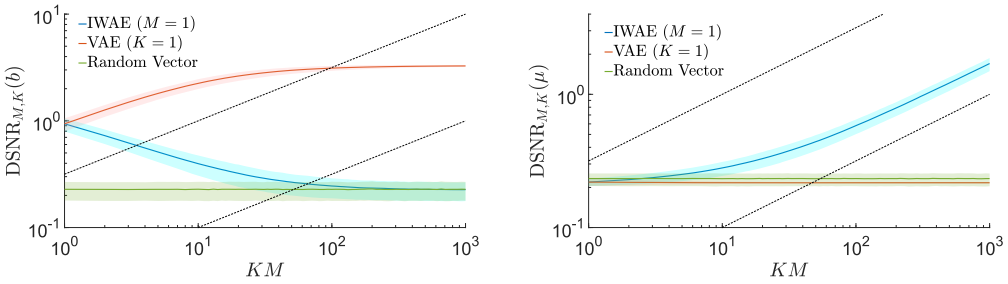
It is also the case that increasing K could be beneficial for the inference network even if it reduces the SNR by improving the direction of the expected gradient. Because each of the individual gradients tend to zero, there is no trivial equivalent test to Figure 4. However, we will return to consider a comparable metric in Figure 6, where we will see that the SNR seems to be the dominant effect for the inference network.

As a reassurance that our chosen definition of the SNR is appropriate for the problem at hand and to examine the effect of multiple dimensions explicitly, we now also consider an alternative definition of the SNR that is similar (though distinct) to that used in Roberts and Tedrake (2009). We refer to this as the "directional" SNR (DSNR). At a high-level, we define the DSNR by splitting each gradient estimate into two component vectors, one parallel to the true gradient and one perpendicular, then taking the expectation of ratio of their magnitudes. More precisely, we define $u = \mathbb{E}\left[\Delta_{M,K}\right] / \|\mathbb{E}\left[\Delta_{M,K}\right]\|_2$ as



(a) Convergence of DSNR for inference network (b) Convergence of DSNR for generative network

Figure 5: Convergence of directional signal-to-noise ratio of gradients estimates with total budget MK. The solid lines show the estimates DSNR and the shaded regions the interquartile range of in the individual ratios. Also shown for reference is the DSNR for a randomly generated vector where each component is drawn from a unit Gaussian.



(a) Convergence of DSNR for inference network (b) Convergence of DSNR for generative network

Figure 6: Convergence of directional signal-to-noise ratio of gradient estimates where the true gradient is taken as $\mathbb{E}\left[\Delta_{1,1000}\right]$. Figure conventions as per Figure 5.

being the true normalized gradient direction and then the DSNR as

$$\mathrm{DSNR}_{M,K} = \mathbb{E}\left[\frac{\|\Delta_{\|}\|_2}{\|\Delta_{\perp}\|_2}\right] \quad \text{where} \quad \Delta_{\|} = \left(\Delta_{M,K}^T u\right) u \quad \text{and} \quad \Delta_{\perp} = \Delta_{M,K} - \Delta_{\|}. \tag{13}$$

The DSNR thus provides a measure of the expected proportion of the gradient that will point is the true direction. For perfect estimates of the gradients, then DSNR $\to \infty$, but unlike the SNR, arbitrarily bad estimates do not have DSNR $= 0$ because even random vectors will have a component of their gradient in the true direction.

The convergence of the DSNR is shown in Figure 5, for which the true normalized gradient u has been estimated empirically, noting that this varies with K. We see a similar qualitative behavior to the SNR, with the gradients of IWAE for the inference network degrading to having the same directional accuracy as drawing a random vector. Interestingly, the DSNR seems to be following the same asymptotic convergence behavior as SNR for the generative network and for the inference network in M (as shown by the dashed lines), even though we have no theoretical result to suggest this should occur.

As our theoretical results suggest that the direction of the true gradients correspond to targeting an improved objective as K increases, we now examine whether this or the changes in the SNR is the dominant effect. To this end, we repeat our calculations for the DSNR but take u as the true direction of the gradient for $K = 1000$. This provides a measure of how varying M and K affects the quality of the gradient directions as biased estimators for $\mathbb{E}\left[\Delta_{1,1000}\right] / \|\mathbb{E}\left[\Delta_{1,1000}\right]\|_2$. As shown in Figure 6, increasing K is still detrimental for the inference network by this metric, even though it brings the expected gradient estimate closer to the true gradient. By contrast, increasing K is now monotonically beneficial for the generative network. Increasing M with $K = 1$ leads to initial improvements for the inference network before plateauing due to the bias of the estimator. For the generative network, increasing M has little impact, with the bias being the dominant factor throughout. Though this

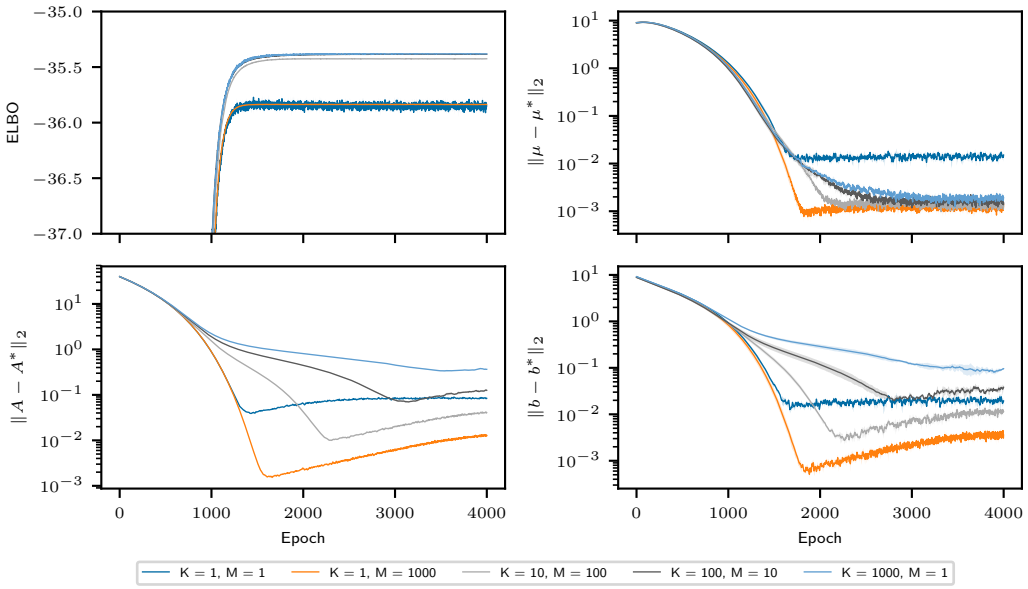


Figure 7: Convergence of optimization for different values of K and M. *(Top, left)* ELBO_{IS} during training (note this represents a different metric for different K). *(Top, right)* L_2 distance of the generative network parameters from the true maximizer. *(Bottom)* L_2 distance of the inference network parameters from the true maximizer. Plots show means over 3 repeats with ± 1 standard deviation. Optimization is performed using the Adam algorithm with all parameters initialized by sampling from the uniform distribution on $[1.5, 2.5]$.

metric is not an absolute measure of performance of the SGA scheme, e.g. because high bias may be more detrimental than high variance, it is nonetheless a powerful result in suggesting that increasing K can be detrimental to learning the inference network.

We finish our experiments by assessing the effect of the outlined changes in the quality of the gradient estimates on the final optimization problem. Figure 7 shows the convergence of running Adam (Kingma and Ba, 2014) to optimize μ, A, and b. This suggests that the effects observed predominantly transfer to the overall optimization problem. Interestingly, setting $K = 1$ and $M = 1000$ gave the best performance on learning not only the inference network parameters, but also the generative network parameters.

5 Conclusions

We have provided theoretical and empirical evidence that algorithmic approaches of increasing the tightness of the ELBO independently to the expressiveness of the inference network can be detrimental to learning. Namely, we have shown for the case of IWAE that the signal-to-noise ratio, namely the magnitude of expected value divided by the standard deviation, of the inference network gradients estimates decreases as we increase the number of importance samples K. However, we also showed that increasing K can provide a better target for the inference network if gradients can be calculated exactly, suggesting that there is trade-off involved in setting K. Experiments on a simple latent variable model support our findings. Our results qualify recent developments with regards to learning inference networks and instigate further investigations regarding desired properties of variational objective functions.

A natural conclusion from our results is that it may be beneficial to use different objectives for learning the generative and inference networks. For example, one might look to use a tighter bound for the generative network than the inference network. Naturally, doing this might introduce its own complications, but it forms a tantalizing possible line of inquiry for future work nonetheless.

Acknowledgments

TR and YWT are supported in part by the European Research Council under the European Union's Seventh Framework Programme (FP7/2007–2013) / ERC grant agreement no. 617071. TAL is supported by a Google studentship, project code DF6700. MI is supported by the UK EPSRC CDT in Autonomous Intelligent Machines and Systems. CJM is funded by a DeepMind Scholarship. FW is supported under DARPA PPAML through the U.S. AFRL under Cooperative Agreement FA8750-14-2-0006, Sub Award number 61160290-111668.

References

R. Bamler, C. Zhang, M. Opper, and S. Mandt. Perturbative black box variational inference. *arXiv preprint arXiv:1709.07433*, 2017.

Y. Burda, R. Grosse, and R. Salakhutdinov. Importance weighted autoencoders. In *ICLR*, 2016.

C. Cremer, Q. Morris, and D. Duvenaud. Reinterpreting importance-weighted autoencoders. *arXiv preprint arXiv:1704.02916*, 2017.

G. Fort, E. Gobet, and E. Moulines. Mcmc design-based non-parametric regression for rare event. application to nested risk computations. *Monte Carlo Methods and Applications*, 23(1):21–42, 2017.

T. C. Hesterberg. *Advances in importance sampling*. PhD thesis, Stanford University, 1988.

D. Kingma and J. Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

D. P. Kingma and M. Welling. Auto-encoding variational Bayes. In *ICLR*, 2014.

D. P. Kingma, T. Salimans, and M. Welling. Improving variational inference with inverse autoregressive flow. *arXiv preprint arXiv:1606.04934*, 2016.

T. A. Le, M. Igl, T. Jin, T. Rainforth, and F. Wood. Auto-encoding sequential Monte Carlo. *arXiv preprint arXiv:1705.10306*, 2017.

Y. Li and R. E. Turner. Rényi divergence variational inference. In *Advances in Neural Information Processing Systems*, pages 1073–1081, 2016.

L. Maaløe, C. K. Sønderby, S. K. Sønderby, and O. Winther. Auxiliary deep generative models. *arXiv preprint arXiv:1602.05473*, 2016.

C. J. Maddison, D. Lawson, G. Tucker, N. Heess, M. Norouzi, A. Mnih, A. Doucet, and Y. W. Teh. Filtering variational objectives. *arXiv preprint arXiv:1705.09279*, 2017.

C. A. Naesseth, S. W. Linderman, R. Ranganath, and D. M. Blei. Variational sequential Monte Carlo. *arXiv preprint arXiv:1705.11140*, 2017.

A. B. Owen. *Monte Carlo theory, methods and examples*. 2013.

T. Rainforth, R. Cornish, H. Yang, A. Warrington, and F. Wood. On the opportunities and pitfalls of nesting Monte Carlo estimators. *arXiv preprint arXiv:1709.06181*, 2017.

R. Ranganath, D. Tran, and D. Blei. Hierarchical variational models. In *International Conference on Machine Learning*, pages 324–333, 2016.

D. Rezende and S. Mohamed. Variational inference with normalizing flows. In *International Conference on Machine Learning*, pages 1530–1538, 2015.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *ICML*, 2014.

J. W. Roberts and R. Tedrake. Signal-to-noise ratio analysis of policy gradient algorithms. In *Advances in Neural Information Processing Systems*, pages 1361–1368, 2009.

F. J. Ruiz, M. K. Titsias, and D. M. Blei. Overdispersed black-box variational inference. *arXiv preprint arXiv:1603.01140*, 2016.

T. Salimans, D. Kingma, and M. Welling. Markov chain Monte Carlo and variational inference: Bridging the gap. In *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*, pages 1218–1226, 2015.

D. Tran, R. Ranganath, and D. M. Blei. The variational Gaussian process. *arXiv preprint arXiv:1511.06499*, 2015.

R. E. Turner and M. Sahani. Two problems with variational expectation maximisation for time-series models. *Bayesian Time series models*, pages 115–138, 2011.

A Proof of SNR Convergence Rates

Theorem 1. *Assume that when $M = K = 1$, the expected gradients; the variances of the gradients; and the first four moments of $w_{1,1}$, $\nabla_\theta w_{1,1}$, and $\nabla_\phi w_{1,1}$ are all finite, with the variances also being non-zero. Then the signal-to-noise ratios of the gradient estimates converge at the following rates*

$$\text{SNR}_{M,K}(\theta) = \sqrt{M}\left|\frac{\sqrt{K}\,\nabla_\theta Z - \frac{1}{2Z\sqrt{K}}\nabla_\theta\left(\frac{\text{Var}[w_{1,1}]}{Z^2}\right) + O\left(\frac{1}{K^{3/2}}\right)}{\sqrt{\mathbb{E}\left[w_{1,1}^2\left(\nabla_\theta \log w_{1,1} - \nabla_\theta \log Z\right)^2\right]} + O\left(\frac{1}{K}\right)}\right| = O\left(\sqrt{MK}\right), \tag{6}$$

$$\text{SNR}_{M,K}(\phi) = \sqrt{M}\left|\frac{\nabla_\phi \text{Var}\left[w_{1,1}\right] + O\left(\frac{1}{K}\right)}{2Z\sqrt{K}\,\sigma\left[\nabla_\phi w_{1,1}\right] + O\left(\frac{1}{\sqrt{K}}\right)}\right| = O\left(\sqrt{\frac{M}{K}}\right) \tag{7}$$

where $Z := p_\theta(x)$ is the true marginal likelihood.

Proof. We start by considering the variance of the estimators. We will first exploit the fact that each $\hat{Z}_{m,K}$ is independent and identically distributed and then apply Taylor's theorem[3] to $\log \hat{Z}_{m,K}$ about Z, using $R_1(\cdot)$ to indicate the remainder term, as follows.

$$\begin{aligned}
M \cdot \text{Var}\left[\Delta_{M,K}\right] &= \text{Var}\left[\Delta_{1,K}\right] = \text{Var}\left[\nabla_{\theta,\phi}\left(\log Z + \frac{\hat{Z}_{1,K} - Z}{Z} + R_1\left(\hat{Z}_{1,K}\right)\right)\right] \\
&= \text{Var}\left[\nabla_{\theta,\phi}\left(\frac{\hat{Z}_{1,K} - Z}{Z} + R_1\left(\hat{Z}_{1,K}\right)\right)\right] \\
&= \mathbb{E}\left[\left(\nabla_{\theta,\phi}\left(\frac{\hat{Z}_{1,K} - Z}{Z} + R_1\left(\hat{Z}_{1,K}\right)\right)\right)^2\right] - \left(\mathbb{E}\left[\nabla_{\theta,\phi}\left(\frac{\hat{Z}_{1,K} - Z}{Z} + R_1\left(\hat{Z}_{1,K}\right)\right)\right]\right)^2 \\
&= \mathbb{E}\left[\left(\frac{1}{K}\sum_{k=1}^{K}\frac{Z\nabla_{\theta,\phi}w_{1,k} - w_{1,k}\nabla_{\theta,\phi}Z}{Z^2} + \nabla_{\theta,\phi}R_1\left(\hat{Z}_{1,K}\right)\right)^2\right] \\
&\quad - \left(\nabla_{\theta,\phi}\mathbb{E}\left[\frac{\hat{Z}_{1,K} - Z}{Z}\right]^{0} + \mathbb{E}\left[\nabla_{\theta,\phi}R_1\left(\hat{Z}_{1,K}\right)\right]\right)^2 \\
&= \frac{1}{KZ^4}\mathbb{E}\left[\left(Z\nabla_{\theta,\phi}w_{1,1} - w_{1,1}\nabla_{\theta,\phi}Z\right)^2\right] + \text{Var}\left[\nabla_{\theta,\phi}R_1\left(\hat{Z}_{1,K}\right)\right] \\
&\quad + 2\mathbb{E}\left[\left(\nabla_{\theta,\phi}R_1\left(\hat{Z}_{1,K}\right)\right)\left(\frac{1}{K}\sum_{k=1}^{K}\frac{Z\nabla_{\theta,\phi}w_{1,k} - w_{1,k}\nabla_{\theta,\phi}Z}{Z^2}\right)\right]
\end{aligned}$$

Now we have by the mean-value form of the remainder that for some $\tilde{Z}$ between Z and $\hat{Z}_{1,K}$

$$R_1\left(\hat{Z}_{1,K}\right) = -\frac{\left(\hat{Z}_{1,K} - Z\right)^2}{2\tilde{Z}^2}$$

and therefore

$$\nabla_{\theta,\phi}R_1\left(\hat{Z}_{1,K}\right) = -\frac{\tilde{Z}\left(\hat{Z}_{1,K} - Z\right)\nabla_{\theta,\phi}\left(\hat{Z}_{1,K} - Z\right) - \left(\hat{Z}_{1,K} - Z\right)^2\nabla_{\theta,\phi}\tilde{Z}}{\tilde{Z}^3}.$$

It follows that the $\nabla_{\theta,\phi}R_1\left(\hat{Z}_{1,K}\right)$ terms are dominated as each of $\left(\hat{Z}_{1,K} - Z\right)\nabla_{\theta,\phi}\left(\hat{Z}_{1,K} - Z\right)$ and $\left(\hat{Z}_{1,K} - Z\right)^2$ vary with the square of the estimator error, whereas other comparable terms vary

[3] This approach follows similar lines to the derivation of nested Monte Carlo convergence bounds in Rainforth et al. (2017) and Fort et al. (2017), and the derivation of the mean squared error for self-normalized importance sampling, see e.g. Hesterberg (1988).

only with the unsquared difference. The assumptions on moments of the weights and their derivatives further guarantee that these terms are finite. More precisely, we have $\tilde{Z} = Z + \alpha\left(\hat{Z}_{1,K} - Z\right)$ for some $0 < \alpha < 1$ where $\nabla_{\theta,\phi}\alpha$ must be bounded with probability 1 as $K \to \infty$ to maintain our assumptions. It follows that $\nabla_{\theta,\phi} R_1\left(\hat{Z}_{1,K}\right) = O\left(\left(\hat{Z}_{1,K} - Z\right)^2\right)$ and thus that

$$\text{Var}\left[\Delta_{M,K}\right] = \frac{1}{MKZ^4}\mathbb{E}\left[\left(Z\nabla_{\theta,\phi}w_{1,1} - w_{1,1}\nabla_{\theta,\phi}Z\right)^2\right] + \frac{1}{M}O\left(\frac{1}{K^2}\right) \tag{14}$$

using the fact that the third and fourth order moments of a Monte Carlo estimator both decrease at a rate $O(1/K^2)$.

Considering now the expected gradient estimate and again using a Taylor's theorem, this time to a higher number of terms, we have

$$\begin{aligned}
\mathbb{E}\left[\Delta_{M,K}\right] &= \mathbb{E}\left[\Delta_{1,K}\right] = \mathbb{E}\left[\Delta_{1,K} - \nabla_{\theta,\phi}\log Z\right] + \nabla_{\theta,\phi}\log Z \\
&= \nabla_{\theta,\phi}\mathbb{E}\left[\log Z + \frac{\hat{Z}_{1,K} - Z}{Z} - \frac{\left(\hat{Z}_{1,K} - Z\right)^2}{2Z^2} + R_2\left(\hat{Z}_{1,K}\right) - \log Z\right] + \nabla_{\theta,\phi}\log Z \\
&= -\frac{1}{2}\nabla_{\theta,\phi}\mathbb{E}\left[\left(\frac{\hat{Z}_{1,K} - Z}{Z}\right)^2\right] + \nabla_{\theta,\phi}\mathbb{E}\left[R_2\left(\hat{Z}_{1,K}\right)\right] + \nabla_{\theta,\phi}\log Z \\
&= -\frac{1}{2}\nabla_{\theta,\phi}\left(\frac{\text{Var}\left[\hat{Z}_{1,K}\right]}{Z^2}\right) + \nabla_{\theta,\phi}\mathbb{E}\left[R_2\left(\hat{Z}_{1,K}\right)\right] + \nabla_{\theta,\phi}\log Z \\
&= -\frac{1}{2K}\nabla_{\theta,\phi}\left(\frac{\text{Var}\left[w_{1,1}\right]}{Z^2}\right) + \nabla_{\theta,\phi}\mathbb{E}\left[R_2\left(\hat{Z}_{1,K}\right)\right] + \nabla_{\theta,\phi}\log Z.
\end{aligned} \tag{15}$$

Using a similar process as in variance case, it is now straightforward to show that $\nabla_{\theta,\phi}\mathbb{E}\left[R_2\left(\hat{Z}_{1,K}\right)\right] = O(1/K^2)$, which is thus similarly dominated (also giving us (9)).

Finally, by combing (14) and (15) and noting that $\sqrt{\frac{A}{K} + \frac{B}{K^2}} = \frac{A}{\sqrt{K}} + \frac{B}{2AK^{3/2}} + O\left(\frac{1}{K^{(5/2)}}\right)$ we have

$$\text{SNR}_{M,K}(\theta) = \left|\frac{\nabla_{\theta,\phi}\log Z - \frac{1}{2K}\nabla_{\theta,\phi}\left(\frac{\text{Var}[w_{1,1}]}{Z^2}\right) + O\left(\frac{1}{K^2}\right)}{\sqrt{\frac{1}{MKZ^4}\mathbb{E}\left[\left(Z\nabla_{\theta,\phi}w_{1,1} - w_{1,1}\nabla_{\theta,\phi}Z\right)^2\right] + \frac{1}{M}O\left(\frac{1}{K^2}\right)}}\right| \tag{16}$$

$$= \sqrt{M}\left|\frac{Z^2\sqrt{K}\left(\nabla_\theta\log Z - \frac{1}{2K}\nabla_\theta\left(\frac{\text{Var}[w_{1,1}]}{Z^2}\right)\right) + O\left(\frac{1}{K^{3/2}}\right)}{\sqrt{\mathbb{E}\left[\left(Z\nabla_\theta w_{1,1} - w_{1,1}\nabla_\theta Z\right)^2\right] + O\left(\frac{1}{K}\right)}}\right| \tag{17}$$

$$= \sqrt{M}\left|\frac{\sqrt{K}\ \nabla_\theta Z - \frac{1}{2Z\sqrt{K}}\nabla_\theta\left(\frac{\text{Var}[w_{1,1}]}{Z^2}\right) + O\left(\frac{1}{K^{3/2}}\right)}{\sqrt{\mathbb{E}\left[w_{1,1}^2\left(\nabla_\theta\log w_{1,1} - \nabla_\theta\log Z\right)^2\right] + O\left(\frac{1}{K}\right)}}\right| = O\left(\sqrt{MK}\right). \tag{18}$$

For ϕ, then because $\nabla_\phi Z = 0$, we instead have

$$\text{SNR}_{M,K}(\phi) = \sqrt{M}\left|\frac{\nabla_\phi\text{Var}\left[w_{1,1}\right] + O\left(\frac{1}{K}\right)}{2Z\sqrt{K}\ \sigma\left[\nabla_\phi w_{1,1}\right] + O\left(\frac{1}{\sqrt{K}}\right)}\right| = O\left(\sqrt{\frac{M}{K}}\right) \tag{19}$$

and we are done. □

B Derivation of Optimal Parameters for Gaussian Experiment

To derive the optimal parameters for the Gaussian experiment we first note that

$$\mathcal{J}(\theta,\phi) = \frac{1}{N}\log\prod_{n=1}^{N} p_\theta(x^{(n)}) - \frac{1}{N}\sum_{n=1}^{N}\mathrm{KL}\left(Q_\phi(z_{1:K}|x^{(n)})\middle\|P_\theta(z_{1:K}|x^{(n)})\right) \quad \text{where}$$

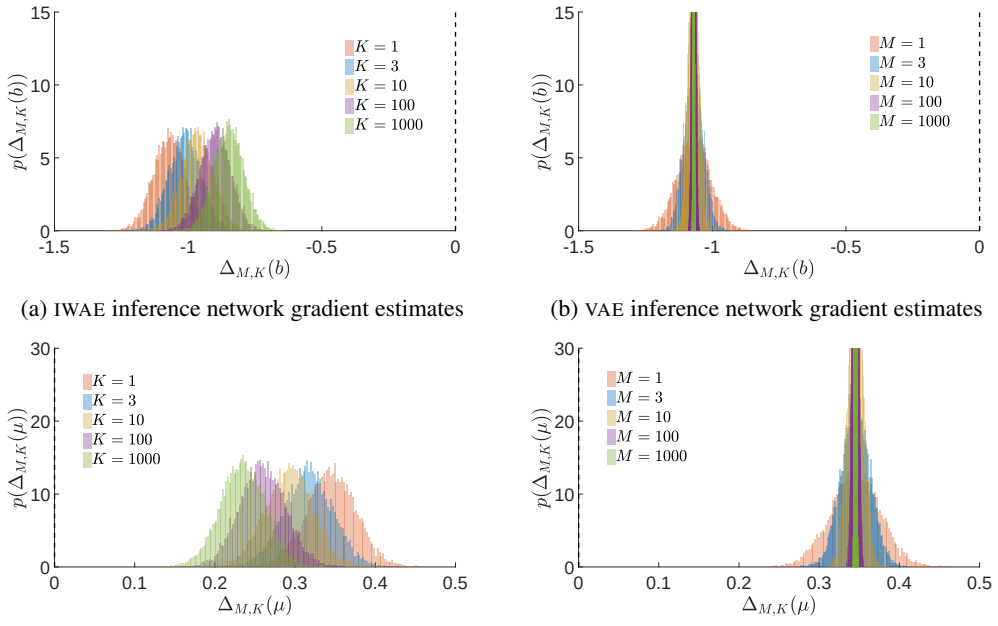
$$P_\theta(z_{1:K}|x^{(n)}) = \frac{1}{K}\sum_{k=1}^{K} q_\phi(z_1|x^{(n)})\ldots q_\phi(z_{k-1}|x^{(n)})p_\theta(z_k|x^{(n)})q_\phi(z_{k+1}|x^{(n)})\ldots q_\phi(z_K|x^{(n)}),$$

$Q_\phi(z_{1:K}|x^{(n)})$ is as per (2) and the form of the Kullback-Leibler (KL) is taken from Le et al. (2017). Next, we note that ϕ only controls the mean of the proposal so, while it is not possible to drive the KL to zero, it will be minimized for any particular θ when the means of $q_\phi(z|x^{(n)})$ and $p_\theta(z|x^{(n)})$ are the same. Furthermore, the corresponding minimum possible value of the KL is independent of θ and so we can calculate the optimum pair (θ^*,ϕ^*) by first optimizing for θ and then choosing the matching ϕ. The optimal θ maximizes $\log\prod_{n=1}^{N} p_\theta(x^{(n)})$, giving $\theta^* := \mu^* = \frac{1}{N}\sum_{n=1}^{N} x^{(n)}$. As we straightforwardly have $p_\theta(z|x^{(n)}) = \mathcal{N}(z;\left(x^{(n)}+\mu\right)/2, I/2)$, the KL is then minimized when $A = I/2$ and $b = \mu/2$, giving $\phi^* := (A^*, b^*)$, where $A^* = I/2$ and $b^* = \mu^*/2$.

C Experimental Results for High Variance Regime

We now present empirical results for a case where our weights are higher variance. Instead of choosing a point close to the optimum by offsetting parameters with a standard deviation of 0.01, we instead offset using a standard deviation of 0.5. We further increased the proposal covariance to I to make it more diffuse. This is now a scenario where the model is far from its optimum and the proposal is a very poor match for the model, giving very high variance weights.

We see that the behavior is the same for variation in M, but somewhat distinct for variation in K. In particular, the SNR and DSNR only decrease slowly with K for the inference network, while increasing K no longer has much benefit for the SNR of the inference network. It is clear that, for this setup, the problem is very far from the asymptotic regime in K such that our theoretical results no longer directly apply. Nonetheless, the high-level effect observed is still that the SNR of the inference network deteriorates, albeit slowly, as K increases.



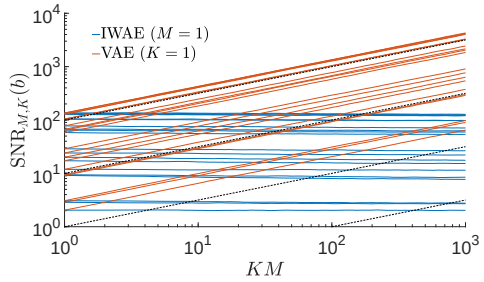
(a) IWAE inference network gradient estimates

(b) VAE inference network gradient estimates

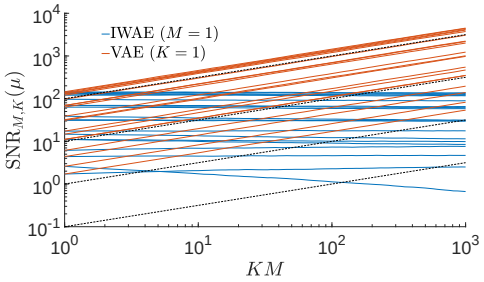
(c) IWAE generative network gradient estimates

(d) VAE generative network gradient estimates

Figure 8: Histograms of gradient estimates as per Figure 2.

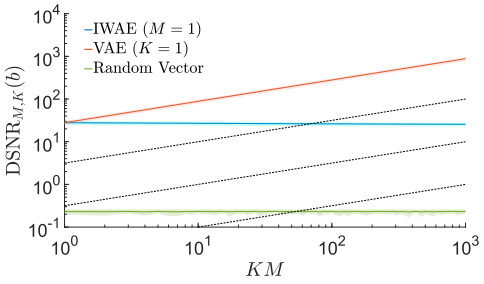


(a) Convergence of SNR for inference network

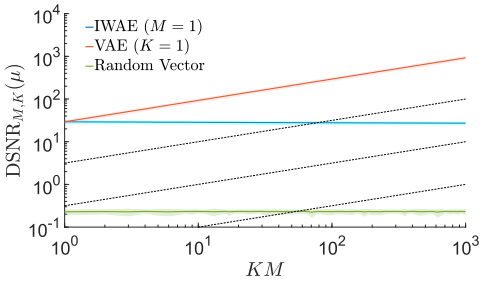


(b) Convergence of SNR for generative network

Figure 9: Convergence of signal-to-noise ratios of gradient estimates as per Figure 3.

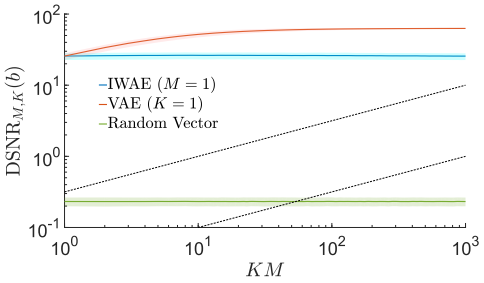


(a) Convergence of DSNR for inference network

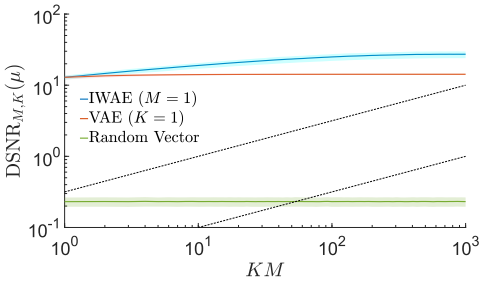


(b) Convergence of DSNR for generative network

Figure 10: Convergence of directional signal-to-noise ratio of gradients estimates as per Figure 5.



(a) Convergence of DSNR for inference network



(b) Convergence of DSNR for generative network

Figure 11: Convergence of directional signal-to-noise ratio of gradient estimates where the true gradient is taken as $\mathbb{E}\left[\Delta_{1,1000}\right]$ as per Figure 11.