Reproduction: Tighter Variational Bounds Are Not Necessarily Better

Department of Computer Science, University of Oxford

Variational Autoencoders and Theoretical Findings

Variational Autoencoders (VAEs) model the probability density of a sampled distribution $p(x)$ by modelling x with stochastic latent variables z, using a learnt prior $p_\theta(z)$ and likelihood $p_\theta(x \mid z)$. This is done through variational inference over the posterior $q_\phi(z \mid x)$.
It is classically optimised using the Evidence-Lower-Bound (ELBO) loss, here also called VAE loss,

$$-\mathcal{L}_{\text{VAE}}(x) = \mathbb{E}_{z\sim q(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z \mid x)}\right] \leq \log p_\theta(x)$$

This expectation is then estimated using M samples. A generalisation of ELBO known as IWAE tightens the bound with $\log p(x)$ by averaging the term inside the $\log$ over K samples of z and estimating the expectation with a single sample.
The paper [1] finds that this tighter bound comes at a cost. It considers the *Signal-to-Noise Ratio* (SNR) of the gradient estimates for both the inference and generative networks, defined to be the ratio of the expectation and standard deviation of the gradient estimates. It finds that the SNR of the gradient estimates for the generative network is $O(\sqrt{MK})$. However the inference gradient estimates have an SNR of $O(\sqrt{M}/\sqrt{K})$, meaning IWAE has difficulties learning the posterior parameters.
The authors propose three new estimates to help alleviate this drawback.

- MIWAE, which allows both $M > 1$ and $K > 1$. This way, one compromise between SNR of the posterior gradient estimates and the tightness of the bound
- CIWAE(β), with loss function $\mathcal{L}_{\text{CIWAE}} = \beta\mathcal{L}_{\text{VAE}} + (1-\beta)\mathcal{L}_{\text{IWAE}}$
- PIWAE, in which p_θ is optimised using IWAE and q_ϕ is optimised using MIWAE. This should allow p_θ to be optimised with the tightest bound without severely decreasing the SNR for q_ϕ

Furthermore in [2], the authors show that by applying the reparameterization trick to the IWAE estimator, the resulting DReG estimator has a gradient with reduced variance compared to the IWAE estimator. They also show that the DReG estimator has an SNR of $O(\sqrt{MK})$ for both the inference and generative network gradients.

Our Contributions

- Reproduction of all experiments on the toy model used to confirm the theoretical findings
- Reproduction of the main results testing the proposed estimates on MNIST data, and compared their respective Effective Sample Sizes
- Extension of the MNIST experiments with a IWAE gradient estimator called DReG by Tucker et al. in [2]

Toy Model Findings

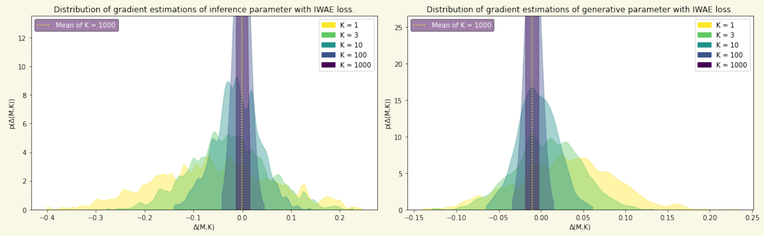

Figure: Gradient estimates $\Delta(M, K)$ of a single parameter in the inference and generative network using the IWAE objective for different values of K

Figure: Convergence of SNR of gradient estimators for inference and generative networks with increasing $M \cdot K$. Shown in red is VAE ($K = 1$), and shown in blue is IWAE ($M = 1$).

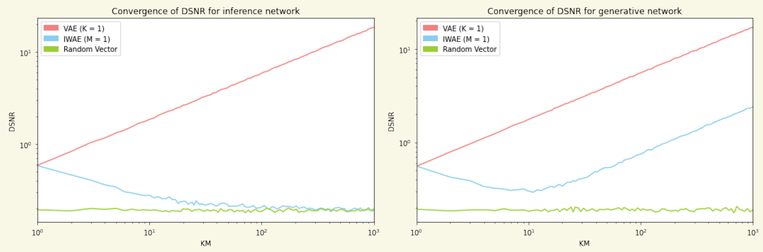

Figure: Convergence of sample DSNR of gradient estimators for inference and generative networks, with reference vector $\mathbb{E}[\Delta_{M,K}]$.

We confirmed the theoretical bounds for SNR by constructing a VAE model simple enough to analytically derive an optimal parameterisation, and sampled gradient estimates from a point slightly offset from this optimum. The results closely followed the theory, with the SNR of the generative network under IWAE loss being particularly interesting as it initially degrades before matching the asymptotic bounds later on.
This behaviour was also seen in similar observations using the so-called Directional SNR (DSNR) metric, which is intriguing as there are no known theoretical bounds for this metric.

MNIST Model Findings

Metric	IWAE	PIWAE(8, 8)	MIWAE(8, 8)	CIWAE0.5	VAE	IWAE-DReG
IWAE-64	-89.13	-89.11	**-88.86**	-89.53	-89.96	-88.99
$\log \hat{p}(x)$	**-87.58**	-87.82	-87.95	-88.71	-89.29	-87.44
$-\text{KL}(Q\|P)$	-1.56	-1.43	-0.91	-0.83	**-0.67**	-1.54

Table: Result of training for 3000 epochs given each loss. Red denotes the best result, with bold denoting the best result from the original paper (which excludes IWAE-DReG)

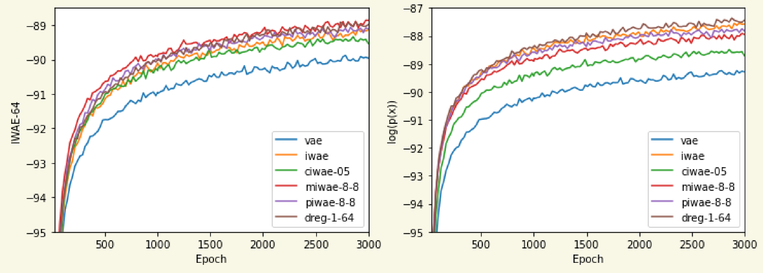

Figure: Convergence of IWAE-64 and $\log \hat{p}(x)$ metrics over 3000 epochs for the VAE, IWAE, CIWAE, MIWAE, PIWAE, and DReG

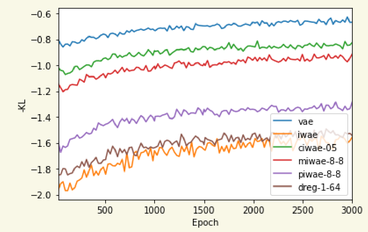

Figure: Convergence of the $-\text{KL}(Q \,||\, P)$ metric over 3000 epochs.

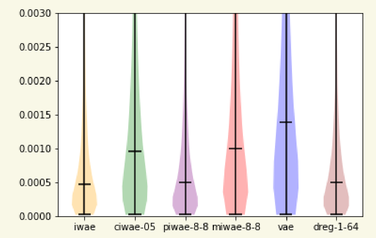

Figure: Violin plots of normalised ESS estimates for each image of MNIST, a measure of how many unweighted samples a weighted sample is worth on average

Training a model for 3000 epochs under each of the optimization regimes in the source paper showed that each proposal was approximately as good as IWAE, with MIWAE outperforming IWAE under the IWAE-64 metric, despite being a biased estimator, simply due to having lower variance. However, in our findings, as with [1], IWAE still was the most performant loss function under the $\log \hat{p}(x)$ metric. However, DReG outperformed all other estimators on the $\log \hat{p}(x)$ metric, while also improving on IWAE on the IWAE-64 metric, showing that the reduced SNR of the gradient estimator led to better training performance.

Citations

1. Tom Rainforth, Adam R. Kosiorek, Tuan Anh Le, Chris J. Maddison, Maximilian Igl, Frank Wood, and Yee Whye Teh. Tighter variational bounds are not necessarily better, 2018.
2. George Tucker, Dieterich Lawson, Shixiang Gu, and Chris J. Maddison. Doubly reparameterized gradient estimators for monte carlo objectives. CoRR, abs/1810.04152, 2018.