Tighter Variational Bounds are Not Necessarily Better

1. Background: IWAE

It is possible to tighten your bound on log likelihood while arbitrarily deteriorating your inference network.

Starting from the variational lower bound,

$$\log p(x) = \log \mathbb{E}_{q(h|x)}\Big[\frac{p(x,h)}{q(h|x)}\Big] \geq \mathbb{E}_{q(h|x}\Big[\log \frac{p(x,h)}{q(h|x)}\Big] = L(x)$$

we instead use a k-sample importance weighting estimate of the log likeihood,

$$L_k(x) = \mathbb{E}_{h_1,\ldots,h_k \sim q(h|x)}\Big[\log \frac{1}{k}\sum_{i=1}^{k} \frac{p(x,h_i)}{q(h_i|x)}\Big]$$

where the term inside the sum has unnormalized importance weights for the joint distribution. Notice that the average importance weights are an unbiased estimator of $p(x)$,

$$L_k = \mathbb{E}\Big[\log \frac{1}{k}\sum_{i=1}^{k}\omega_i\Big] \leq \log \mathbb{E}\Big[\frac{1}{k}\sum_{i=1}^{k}\omega_i\Big] = \log p(x)$$

The case of $k = 1$ is the standard VAE objective. More samples improves the tightness of the bound. In fact we have a theorem that,

$$\log p(x) \geq L_{k+1} \geq L_k$$

To train this objective we draw from the inference distribution to get an unbiased estimate of the gradient of L_k. We use reparam trick to get low variance update rule. We use monte carlo estimate of the expectation of the gradient with different draws from our latent stochastic variables.

2. Why IWAE Sometimes Fails

We need to be able to numerically solve the optimization problem. We study the grad estimates of IWAE with M MC estimates built from K importance samples (particles).

Note that M doesn't change the true gradient, only the variance in estimating it. K changes the true gradient (higher K gives a tighter bound).

Define the signal to noise ratio:

$$SNR_{M,K}(\theta) = \left|\frac{\mathbb{E}[\Delta_{M,K}(\theta)]}{\sigma[\Delta_{M,K}(\theta)]}\right|$$

Main result:

$$SNR_{M,K}(\theta) = O(\sqrt{MK})$$
$$SNR_{M,K}(\phi) = O(\sqrt{M/K})$$

There is a factor in the other direction. As $K \to \infty$ the expected gradient points in the direction of -variance so the optimal ϕ will minimize the variance of the weights. So there is some optimal K.