Deep Generative Models

Part V: improving the ELBO: better bounds

Estimators

- Starting from the simplest estimatror:

$$\hat{I}(z_{1:K}) = \sum_{k=1}^{K} p(x|z_k) \text{ with } z_k \sim p(z)$$

- The previous estimator is unbiased but can have high variance as a consequence of not taking into account the observation we would like the latent variables to explain. To incorporate that information we can use the estimator:

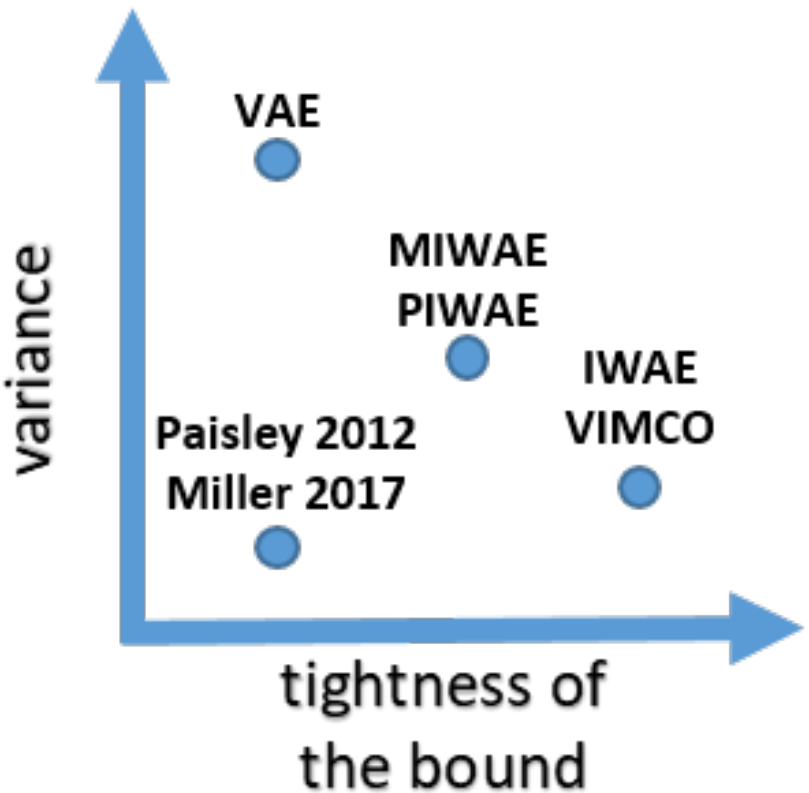
$$\hat{I}(z_{1:K}) = \frac{1}{K} \sum_{k=1}^{K} \frac{p(x, z_k)}{q(z_k|x)}$$

$$\mathbb{E}_{q(z_{1:K}|x)} \left[\log \hat{I}(z_{1:K}) \right] \leq \log \mathbb{E}_{q(z_{1:K}|x)} \left[\hat{I}(z_{1:K}) \right] = \log p(x)$$

- Note that the idea above will work if the only parameter that we are estimating is θ. Besides the variance reduction techniques that are applied to the estimator of the model parameters, as we discussed in the last week we can also apply these techniques to control the variance of the gradient estimator.

Estimators

- Tightening the bound or lowering the variance?



VAEs and Reparametrization Trick

- Given the generative process $p_\theta(x)$ and the recognition model $q_\phi(z|x)$ the ELBO cost function is:

$$\log p_\theta(x) = \log \mathbb{E}_{q_\phi(z|x)}\left[\frac{p_\theta(x,z)}{q_\phi(z|x)}\right] \geq \mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathcal{L}(x)$$

- Since q_ϕ does not depend on θ the gradient w.r.t θ can be rewritten as an expectation but for computing the gradient w.r.t ϕ reparametrization trick is used:

$$z(x,\epsilon,\phi) = \mu(x,\phi) + \Sigma(x,\phi)^{\frac{1}{2}}\epsilon \quad \text{where} \quad \epsilon \sim \mathcal{N}(0,\mathcal{I})$$

$$\nabla_{\theta,\phi}\mathcal{L}(x) = \nabla_{\theta,\phi}\mathbb{E}_{q_\phi(z|x)}\left[\log \frac{p_\theta(x,z)}{q_\phi(z|x)}\right] = \mathbb{E}_\epsilon\left[\nabla_{\theta,\phi}\log \frac{p_\theta(x,z(x,\epsilon,\phi))}{q_\phi(z(x,\epsilon,\phi)|x)}\right]$$

- In practice, the expectation is approximated using M Monte Carlo samples $\epsilon_1, ..., \epsilon_M$ as following:

$$\mathbb{E}_\epsilon\left[\nabla_{\theta,\phi}\log \frac{p_\theta(x,z(x,\epsilon,\phi))}{q_\phi(z(x,\epsilon,\phi)|x)}\right] \approx \frac{1}{M}\sum_{m=1}^{M}\nabla_{\theta,\phi}\log \frac{p_\theta(x,z(x,\epsilon_i,\phi))}{q_\phi(z(x,\epsilon_i,\phi)|x)}$$

Importance Sampling and IWAE - 1

- The main idea of importance sampling is to instead of rejecting samples, weight them by their importance:

$$\mathbb{E}_p[f(x)] = \int f(x)p(x)dx = \int f(x)\frac{p(x)}{q(x)}q(x)dx = \mathbb{E}_q\left[\frac{f(x)p(x)}{q(x)}\right]$$

  - The assumption is that sampling from p is difficult but sampling from q is easy.
  - Monte Carlo estimates of q are used to to approximate $\mathbb{E}_p[f(x)]$.
- IWAE uses a different lower bound on $\log p(x)$ using the idea above:

$$\mathcal{L}_k(x) = \mathbb{E}_{z_1,\ldots,z_K \sim q_\phi(z|x)}\left[\log \frac{1}{k}\sum_{k=1}^{K}\frac{p_\theta(x, z_k)}{q_\phi(z_k|x)}\right]$$

  - This lower bound is tighter: $\log p_\theta(x) \geq \mathcal{L}_k \geq \mathcal{L}_{k-1} \geq \cdots \geq \mathcal{L}_1$.
  - As K goes to infinity $\mathcal{L}_k$ converges to $\log p(x)$.

Importance Sampling and IWAE - 2

- To compute the gradient of $\mathcal{L}_k$ we can use the reparametrization trick and use Monte Carlo samples of the function inside the expectation.
  - If K denotes the number of importance samples and M is the number of Monte Carlo samples the gradient of $\mathcal{L}_k$ is:

$$\Delta_{M,K} := \frac{1}{M}\sum_{m=1}^{M} \nabla_{\theta,\phi} \log \frac{1}{K}\sum_{k=1}^{K} w_{m,k},$$

    where $w_{m,k} = \frac{p_\theta(z_{m,k},x)}{q_\phi(z_{m,k}|x)}$, and $z_{m,k} \overset{iid}{\sim} q_\phi(z|x)$.
  - In IWAE context, M is always equal to 1.
  - When $K = 1$, we recover VAE gradient estimate.
- Since the lower bound is changed the difference between $\log p(x)$ and L_k does not represent $\mathrm{KL}(q||p)$ anymore. Instead:

$$Q_{IS}(z_{1:K}|x) := \prod_{k=1}^{K} q_\phi(z_k|x) \quad , \quad P_{IS}(z_{1:K}) = \frac{1}{K}\sum_{k=1}^{K}\left(\frac{p_\theta(z_k|x)}{q_\phi(z_k|x)}\prod_{k=1}^{K} q_\phi(z_k|x)\right)$$

$$\log p(x) - \mathcal{L}_k = \mathrm{KL}(Q_{\mathrm{IS}}||P_{\mathrm{IS}})$$

VIMCO

- IWAE does not work with discrete variables since it involves reparametrization trick. VIMCO uses the same objective as IWAE but uses the SCORE gradient idea to make it compatible with discrete variables.
- The naive Monte Carlo estimator has high variance.

$$\nabla_{\theta,\phi}\mathcal{L}_k(x) = \nabla_{\theta,\phi}\mathbb{E}_{z_1,\ldots,z_m\sim q_\phi(z|x)}\left[\log\frac{1}{k}\sum_{k=1}^{K}\frac{p_\theta(x,z_k)}{q_\phi(z_k|x)}\right]$$

$$\approx \sum_{k=1}^{K}\hat{L}(z_{1:K})\nabla_{\theta,\phi}\log q(z_k|x) + \sum_{k=1}^{K}\tilde{w}_k\nabla_{\theta,\phi}\log\frac{p(x,z_k)}{q(z_k|x)}$$

- The main sources of the variance are the following:
  - The learning signal $\hat{L}$ is unbounded.
  - The learning signal scales all the samples equally.
- The proposed estimator fixes the issues mentioned above:

$$\nabla_{\theta,\phi}\mathcal{L}_k(x) \approx \sum_{k=1}^{K}\hat{L}(z_k|z_{-k})\nabla_{\theta,\phi}\log q(z_k|x) + \sum_{k=1}^{K}\tilde{w}_k\nabla_{\theta,\phi}\log\frac{p(x,z_k)}{q(z_k|x)}$$

where $\hat{L}(z_k|z_{-k}) = \hat{L}(z_{1:K}) - \log\frac{1}{K}\left(\sum_{i\neq k}\frac{p(x,z_i)}{q(z_i|x)} + \frac{1}{K-1}\sum_{i\neq k}\frac{p(x,z_i)}{q(z_i|x)}\right)$

Inference vs. Max Likelihood

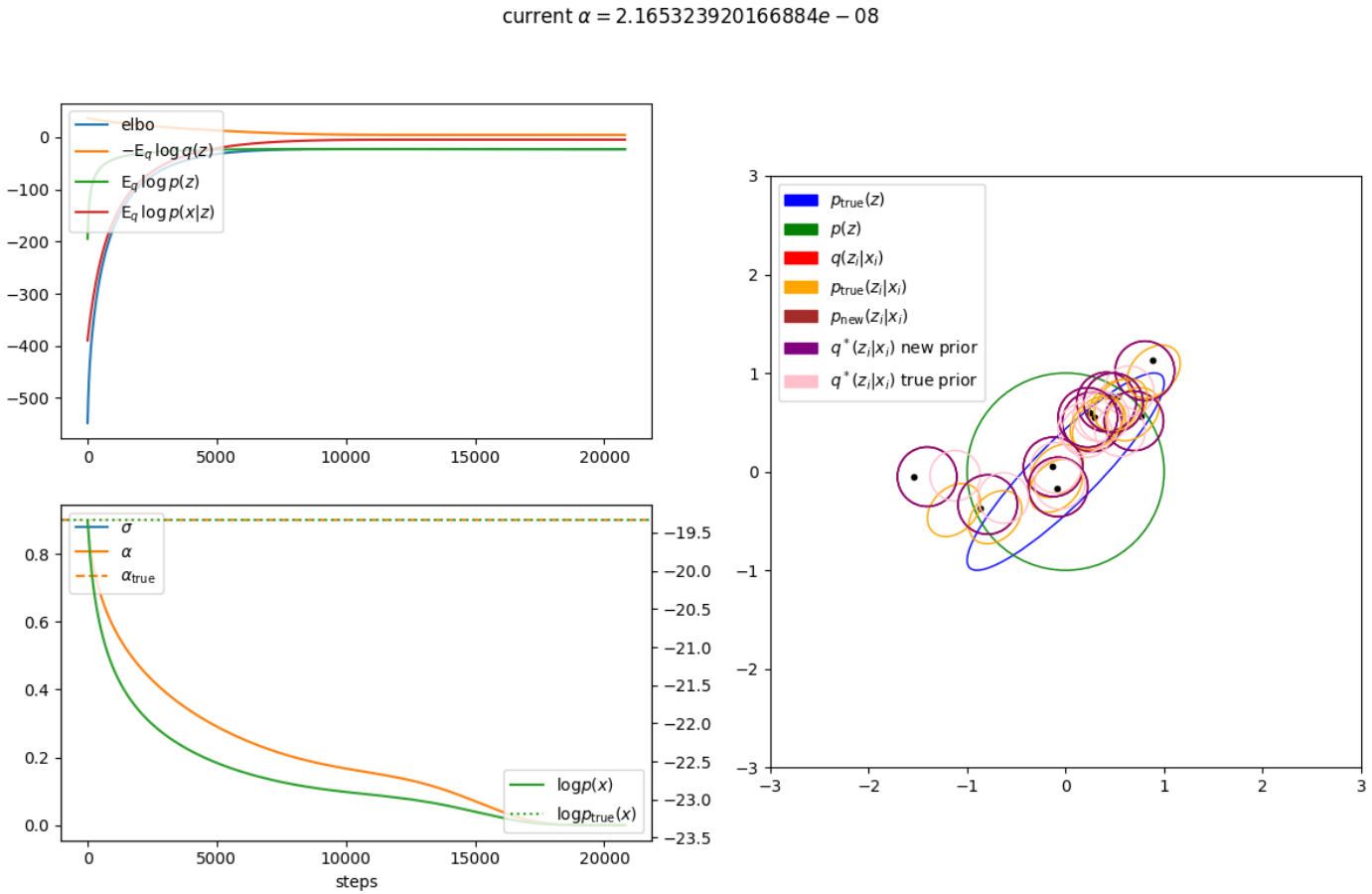
- Different cost functions solve different problems.
  - ELBO with fixed θ does variational inference, i.e. it will recover a distribution that is close to the true posterior.
  - As the cost function approaches $\log p_\theta(x)$ the problem becomes more similar to maximum likelihood and the resulting $q(z)$ does not necessarily reflect the true posterior.
- Do we really need tighter bounds?
  - If the true posterior is included in the approximating family even the VAE gradient can recover the true posterior.

$$\text{argmin}_\phi \mathcal{L}_{\text{VAE}} = \text{argmin}_\phi \text{KL}(q_\phi(z)||p_\theta(z|x)) = p_\theta(z|x)$$

$$\text{argmin}_\theta \mathcal{L}_{\text{VAE}} = \text{argmin} \text{KL}(q(z)||p(z|x)) - \log p_\theta(x) = \text{argmax}_\theta p_\theta(x)$$

Inference vs. Max Likelihood

- What will happen if the true posterior is NOT included in the approximating family?



[Martin et al 2018]

What is wrong with IWAE?

So far, IWAE has been great – mostly for the generative model.

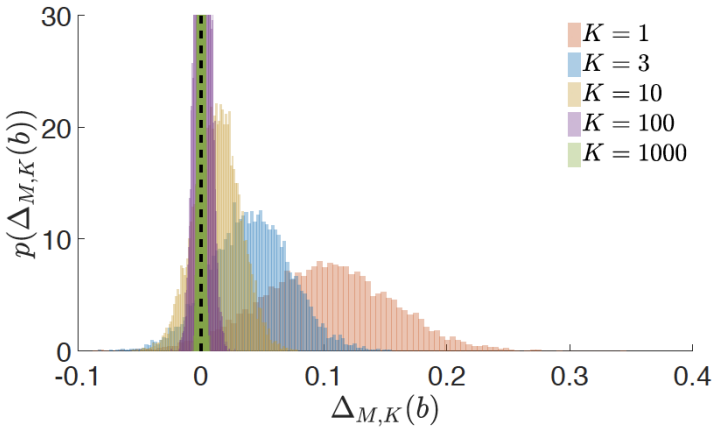
- We have a tighter lower bound: $\log p_\theta(x) \geq \mathcal{L}_k \geq \mathcal{L}_{k-1} \geq \cdots \geq \mathcal{L}_1$
- We have a more representative q

Are we missing anything? ... for the inference model

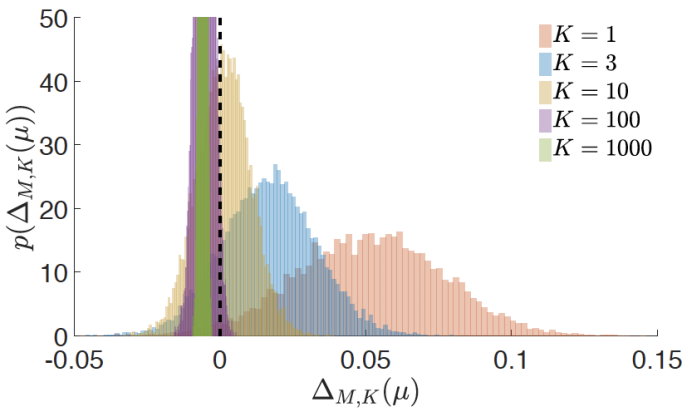
- When the number of particles K goes to infinity, the bound is tight, i.e., the marginal likelihood estimate becomes exact.
  - The choice of inference model q_ϕ does not matter
  - The gradient of $\mathcal{L}_k(\approx \log p_\theta(x))$ with respect to proposal parameters ϕ goes to zero
- What does a tighter bound mean for the inference network?
  - We have a different KL objective
  - Can optimizing this different KL objective push q to the true posterior?
  - Or, is the role of the inference network still posterior approximation?

Assessing the gradient estimators

What do we care about the gradient estimator?



(a) IWAE inference network gradient estimates



(b) IWAE generative network gradient estimates

[Rainforth et al. 2018]

- If it's magnitude is close to zero, we want proportionally small variance to estimate it accurately.
- Otherwise, it just degrades to pure noise around 0 – not providing fidelity in directions towards improvements.

  We care about the relative variance – Signal to noise ratio (SNR).

SNR

- Use K particles to construct IWAE's ELBO:

$$\text{ELBO}_{\text{IS}}(\theta, \phi, x) := \int Q_{\text{IS}}(z_{1:K}|x) \log \hat{Z}_{\text{IS}}(z_{1:K}, x) dz_{1:K}$$

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

- Assuming that reparameterization is possible, we construct the gradient estimate from M averaging samples:

$$\Delta_{M,K} := \frac{1}{M} \sum_{m=1}^{M} \nabla_{\theta,\phi} \log \frac{1}{K} \sum_{k=1}^{K} w_{m,k},$$

where $w_{m,k} = \frac{p_\theta(z_{m,k}, x)}{q_\phi(z_{m,k}|x)}$, and $z_{m,k} \overset{iid}{\sim} q_\phi(z_{m,k}|x)$.

-

$$\text{SNR}_{M,K}(\theta) := |\mathbb{E}[\Delta_{M,K}(\theta)]/\sigma[\Delta_{M,K}(\theta)]|$$

$$\text{SNR}_{M,K}(\phi) := |\mathbb{E}[\Delta_{M,K}(\phi)]/\sigma[\Delta_{M,K}(\phi)]|$$

SNR theoretical results - Part 1

$$\text{SNR}_{M,K}(\theta) := |\mathbb{E}[\Delta_{M,K}(\theta)]/\sigma[\Delta_{M,K}(\theta)]| = O(\sqrt{MK}).$$

$$\text{SNR}_{M,K}(\phi) := |\mathbb{E}[\Delta_{M,K}(\phi)]/\sigma[\Delta_{M,K}(\phi)]| = O(\sqrt{\frac{M}{K}}).$$

- Intuitive proof.
  Denote $\hat{Z}_{m,K} = \frac{1}{K}\sum_{k=1}^{K} w_{m,k}$ as the marginal likelihood estimate, then $\Delta_{M,K} = \frac{1}{M}\sum_{m=1}^{M} \nabla_{\theta,\phi} \log \hat{Z}_{m,K}$.
  - The effect of M on SNR – just apply law of large number
    - The expectation is independent of M.
    - The variance reduces at a rate $O(\frac{1}{M})$.

SNR theoretical results - Part 2

- Intuitive proof (effect of K).
  - Note that

$$\Delta_{M,K} = \frac{1}{M}\sum_{m=1}^{M} \nabla_{\theta,\phi} \log \hat{Z}_{m,K}$$

$$\nabla_{\theta,\phi} \log \hat{Z}_{m,K} = \nabla_{\theta,\phi} \frac{1}{K}\sum_{k=1}^{K} w_{m,k} = \sum_{k=1}^{K} \tilde{w}_{m,k} \nabla_{\theta,\phi} \log(w_{m,k}),$$

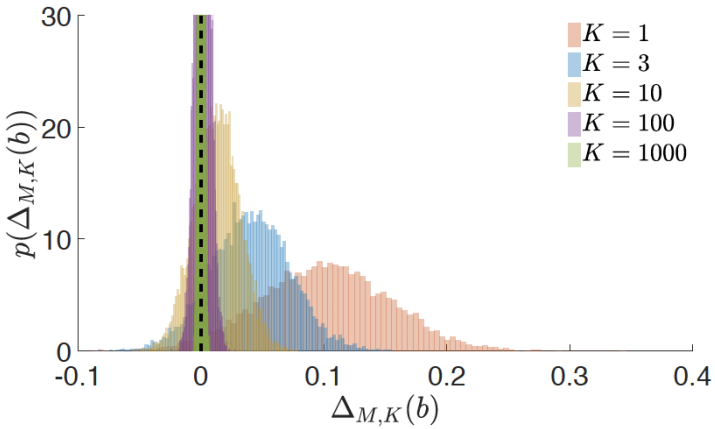
    where $\tilde{w}_{m,k} = \frac{w_{m,k}}{\sum_{i=1}^{K} w_{m,i}}$ are normalized weights.
  - $\nabla_{\theta,\phi} \log \hat{Z}_{m,K}$ can be seen as a self-normalized importance sampling estimate.
  - This implies the bias of the estimate is $O(\frac{1}{K})$, and the standard deviation is $O(\frac{1}{\sqrt{K}})$.
  - Hence,

$$\text{SNR}(\phi,\theta) = \sqrt{M}\left|\frac{\nabla_{\theta,\phi} \log Z + O(\frac{1}{K})}{O(\frac{1}{\sqrt{K}})}\right|,$$

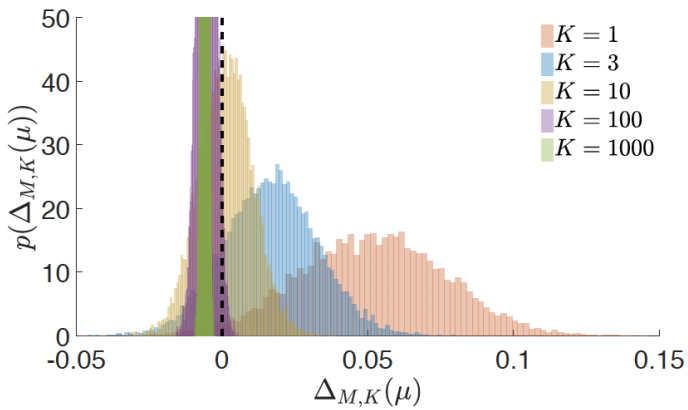
    where $Z = p_\theta(x)$.
  - So we have $\text{SNR}(\theta) = O(\sqrt{MK})$,

    and $\text{SNR}(\phi) = O(\sqrt{\frac{M}{K}})$ since $\nabla_\phi \log Z = 0$,

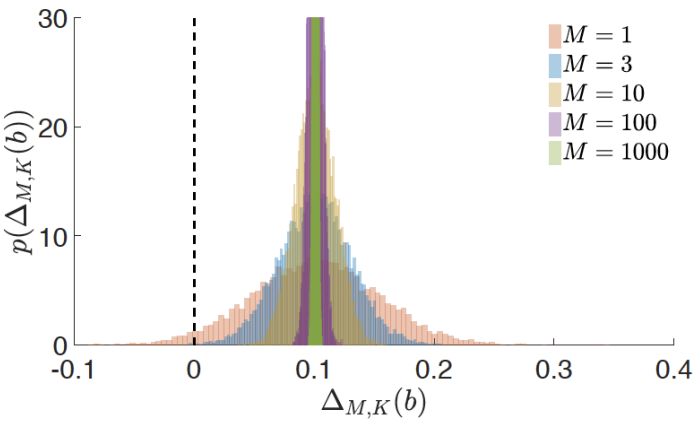
SNR experiment results – Part 1



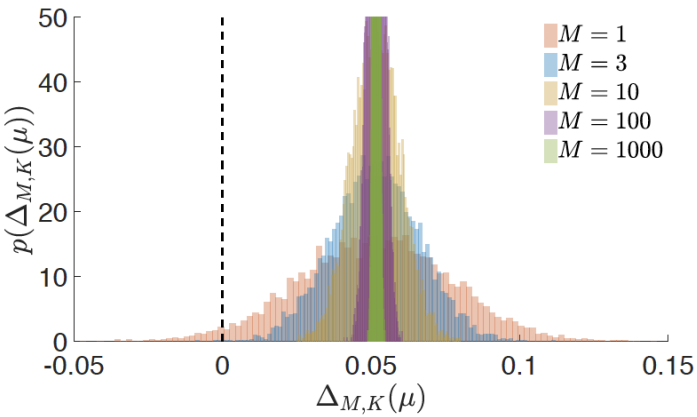
(a) IWAE inference network gradient estimates



(b) IWAE generative network gradient estimates

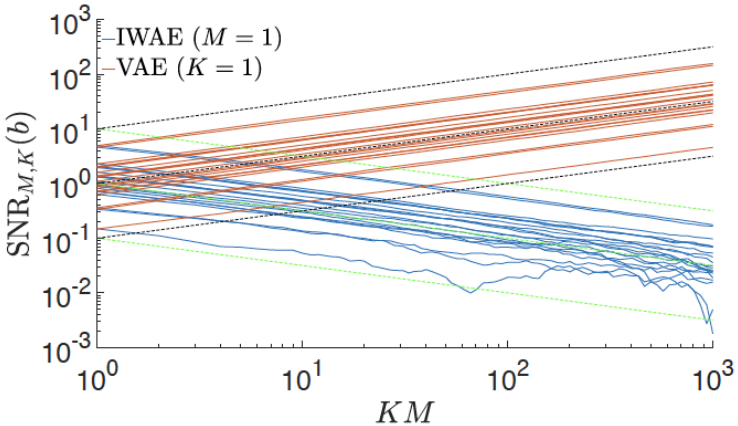


(a) VAE inference network gradient estimates

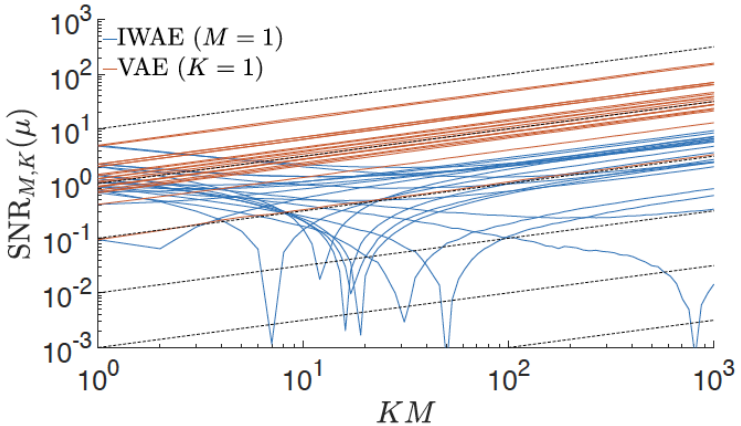


(b) VAE generative network gradient estimates

SNR experiment results – Part 2



(a) Convergence of SNR for inference network



(b) Convergence of SNR for generative network

$$\mathsf{SNR}(\phi, \theta) = \sqrt{M} \left| \frac{\nabla_{\theta,\phi} \log Z + O(\frac{1}{K})}{O(\frac{1}{\sqrt{K}})} \right|,$$

where $Z = p_\theta(x)$.

More questions on the inference network

- Before, we discuss about the problems of degrading gradients in inference network.
- Now, let's focus on **the role of inference network**:
- What does optimizing a tighter bound mean for the inference network?
  - $$KL(Q_{IS}||P_{IS}) := \log p_\theta(x) - \mathcal{L}_K$$
  - Are we still doing posterior approximation?

Asymptotic directions

- Let's take a closer look at the expected gradient estimator w.r.t. ϕ,

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

.

- The inference network is optimized along the direction that minimizes the weight's variance, as $K \to \infty$.
- This is actually pushing $q_\phi$to the optimal importance sampling distribution
  - "optimal" in terms of estimating the marginal likelihood
  - the role of inference network is to estimate the marginal likelihood
- However, if we are lucky that q_ϕ has large enough capacity, in the sense that the variance of $w_{11} = \frac{p_\theta(x,z)}{q_\phi(z|x)}$ can be minimized to 0, we are pushing q_ϕ to the true posterior.

Addressing the issue: New estimators

- In terms of SNR,
  - Large K is good for the generative network,
  - but bad for the inference network.
- **Solutions**:
  1. Control this trade-off between the needs of different networks (a combo of VAE and IWAE)
     - MIWAE: Use $M > 1$ and $K > 1$
     - CIWAE: $\text{ELBO}_{\text{CIWAE}} = \beta\text{ELBO}_{\text{AVE}} + (1-\beta)\text{ELBO}_{\text{IWAE}}$.
  2. Use different objectives for different networks
     - PIWAE:

- Use IWAE target for the generative network

$$\Delta_K^C(\theta) = \nabla_\theta \log \frac{1}{K} \sum_{k=1}^{K} w_k$$

- Use MIWAE target for the inference network

$$\Delta_{M,K}(\phi) = \frac{1}{M} \sum_{m=1}^{M} \nabla_\phi \log \frac{1}{L} \sum_{l=1}^{L} w_{m,l}.$$

Comparison results

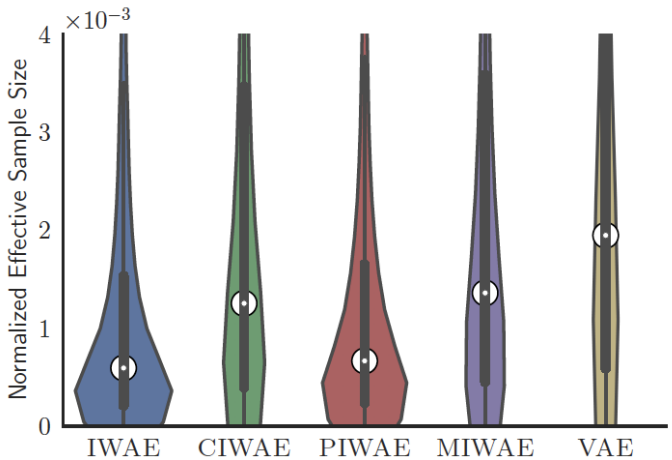


Figure 7: Violin plots of ESS estimates for each image of MNIST, normalized by the number of samples drawn. A violin plot uses a kernel density plot on each side – thicker means more MNIST images whose q_ϕ achieves that ESS.

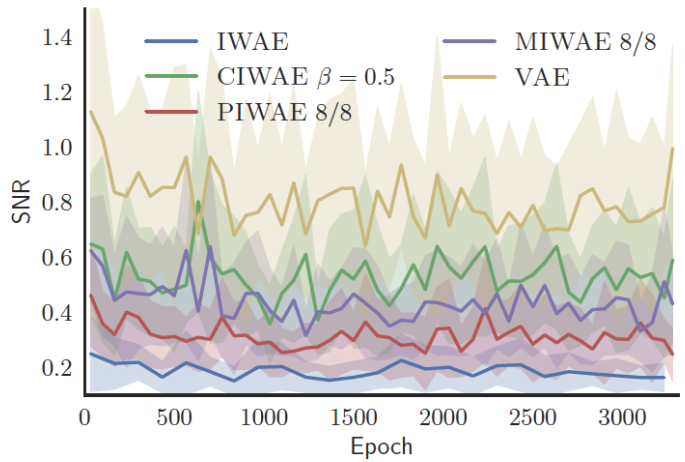


Figure 8: SNR of inference network weights during training. All lines are mean $\pm$ standard deviation over 20 randomly chosen weights per layer.

Some Take-home messages

- This paper suggests that, tighter bounds are not necessarily better
  - Large K is bad for inference network learning
  - If we care about representation learning, be careful choosing K.

Some Take-home messages

- In general, think about two different objectives when doing variational inference...
  - Maximum likelihood $\rightarrow$ learning the generative model
    - Approach: maximizing tighter lower bound to get a more accurate estimate
    - The inference model is an intermediate step for the likelihood estimate
    - In IWAE case, it serves as a proposal distribution
  - Inference $\rightarrow$ learning the inference model
    - Approach: minimizing the KL divergence
    - The goal of the inference model is to approximate true posterior
    - The role of generative model is to assess the reconstruction error
- For VAE, two objectives happen to be equivalent.
- For IWAE, the ultimate goal is to have a better generative model.
  - We are lucky that $KL(Q_{\text{IS}}||P_{\text{IS}}) = 0$ if and only if $KL(q||p) = 0$ [Le et al 2017]
  - However, if q is correct, we can directly use VAE's ELBO to get the tight bound.