Tighter variational bounds are not necessarily better

Rainforth, et al. 2019

Motivation

- ELBO is lower bound of the log-probability term. Hence, maximizing it is not the same as maximizing the log-probability.
- Approaches such as the importance weighted auto-encoder (IWAE) hope to obtain tighter bounds on the log-probability with the hope of improving the performance of the VAE.
- This paper talks about the inference/recognition/encoder network, and how tighter bounds affect its fidelity.

Background

- $x \in \mathcal{X}$: Random variable (r.v.) whose distribution we wish to model. $z \in \mathcal{Z}$: Latent variable. Joint distribution $p_\theta(x, z)$.
- Vanilla VAE:
  - $q_\phi(z|x)$: Approximate inference model, realized using a NN with parameters ϕ.
  - ELBO:

$$\mathcal{L}_0(\theta, \phi, x) = \mathbb{E}_{q_\phi(z|x)} \log \frac{p_\theta(x, z)}{q_\phi(z|x)} \tag{1}$$

  - VAE trained by maximizing $\mathcal{L}_0$ using estimates of $\nabla_{\theta,\phi}\mathcal{L}_0(\theta, \phi, x)$ after reparametrizing q_ϕ.

Background: Importance weighted autoencoder (IWAE)

- IWAE builds tighter lower bounds to $\log p_\theta(x)$ by considering the following loss term:

$$\mathcal{L}_{IWAE}(z_{1:K}, x) = \mathbb{E}_Q\left[\log \hat{Z} dz_{1:K}\right] \leq \log p_\theta(x) \tag{2}$$

where

$$Q(z_{1:K}|x) = \prod_{k=1}^{K} q_\phi(z_k|x) \tag{3}$$

$$\hat{Z}(z_{1:K}, x) = \frac{1}{K}\sum_{k=1}^{K} \frac{p_\theta(x, z_k)}{q_\phi(z_k|x)} \tag{4}$$

z_k are iid samples from q_ϕ.

- As seen in the IWAE paper, $K > 1$ is good for generative performance.

Contributions of this paper

- Lower bound gets tighter, but how are gradient updates affected?
- Gradient estimate over M samples:

$$\Delta_{M,K} = \frac{1}{M} \sum_{m=1}^{M} \nabla_{\theta,\phi} \log \frac{1}{K} \sum_{k=1}^{K} w_{m,k} \tag{5}$$

  where $w_{m,k} = \frac{p_\theta(x, z_{m,k})}{q_\phi(z_{m,k}|x)}$.
- Simple case: $M = 1, K \to +\infty$: $\hat{Z} \to p_\theta(x)$. Therefore, both mean and variance of the gradient update with respect to ϕ, $\Delta_{M,K}(\phi)$ go to zero.

Contributions of this paper

- Need to assess relative strength of gradient update vs noise in it.
- Define (elementwise) signal to noise ratio in the gradient update:

$$SNR_{M,K}(\theta) = \left| \frac{\mathbb{E}[\Delta_{M,K}(\theta)]}{\sigma[\Delta_{M,K}(\theta)]} \right| \tag{6}$$

$$SNR_{M,K}(\phi) = \left| \frac{\mathbb{E}[\Delta_{M,K}(\phi)]}{\sigma[\Delta_{M,K}(\phi)]} \right| \tag{7}$$

$$\tag{8}$$

- The paper shows that

$$SNR_{M,K}(\theta) = O(\sqrt{MK}) \tag{9}$$

$$SNR_{M,K}(\phi) = O(\sqrt{M/K}). \tag{10}$$

Contributions of the paper

- Effect of M: This corresponds to the outer average, hence by law of large numbers, variance reduces at $O(1/M)$ rate.
- Effect of K: Prior work shows that the bias of a self-normalized importance sampler converges at $O(1/K)$ rate and standard deviation converges at $O(1/\sqrt{K})$ rate. Therefore, if the mean is 0, SNR goes down as $O(1/\sqrt{K})$. If the mean is non-zero, SNR goes up at a rate $O(\sqrt{K})$. Hence the difference in behavior in the gradient updates of ϕ and θ.

References

- Rainforth, Tom, et al. "Tighter variational bounds are not necessarily better." International Conference on Machine Learning. PMLR, 2018.
- Kingma, Diederik P., and Max Welling. "Auto-encoding variational bayes." arXiv preprint arXiv:1312.6114 (2013).
- Importance sampling: stanford statistics notes: `https://statweb.stanford.edu/ owen/mc/Ch-var-is.pdf`