REINTERPRETING IMPORTANCE-WEIGHTED AUTOENCODERS

UNIVERSITY OF TORONTO COMPUTER SCIENCE

Chris Cremer, Quaid Morris, David Duvenaud

Department of Computer Science, University of Toronto

Main Idea

The standard interpretation of importance-weighted autoencoders is that they maximize a tighter lower bound on the marginal likelihood. We give an alternate interpretation of this procedure: that it optimizes the standard variational lower bound, but using a more complex distribution. In other words, the IWAE lower bound can be interpreted as the standard VAE lower bound with an implicit q_{IW} distribution:

$$L_{IWAE}[q] = L_{VAE}[q_{IW}]$$

With this interpretation in mind, we can generalize q_{IW} to be more broadly applicable to any divergence measure.

Background

The variational autoencoder (VAE; [4]) maximizes the following evidence lower bound (ELBO):

$$\begin{aligned} log(p(x)) &\geq E_{z \sim q(z|x)} \left[log \left(\frac{p(x,z)}{q(z|x)} \right) \right] \\ &= L_{VAE}[q]. \end{aligned}$$

The importance-weighted autoencoder (IWAE; [2]) maximizes the following tighter multi-sample lower bound:

$$\begin{aligned} log(p(x)) &\geq E_{z_1 \ldots z_k \sim q(z|x)} \left[log \left(\frac{1}{k} \sum_{i=1}^{k} \frac{p(x,z_i)}{q(z_i|x)} \right) \right] \\ &= L_{IWAE}[q]. \end{aligned}$$

Importance Resampling

Algorithm 1: Sampling from q_{IW}

```
k ← number of samples
for i in 1 ... k do
    z_i ~ q(z|x)
    w_i = p(x,z_i) / q(z_i|x)
Each w̃ = w_i / Σ_{i=1}^k w_i
j ~ Cat(w̃)
Return z_j
```

Algorithm 1 is the procedure to sample from $q_{IW}(z|x)$. It is equivalent to sampling-importance-resampling (SIR). $Cat(\tilde{w})$ refers to a categorical distribution parametrized by $\tilde{w}$.

Implicit q_{IW} Distribution

In this section, we derive the implicit distribution that arises from importance sampling from a distribution p using q as a proposal distribution. Given a batch of samples $z_1 \ldots z_k$ from $q(z|x)$, the following is the importance weighted q_{IW} distribution as a function of one of the samples, z_i:

$$q_{IW}(z_i|x, z_{\setminus i}) = k\tilde{w}_i q(z_i|x) = \left(\frac{\frac{p(x,z_i)}{q(z_i|x)}}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}} \right) q(z_i|x) = \frac{p(x,z_i)}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}}$$

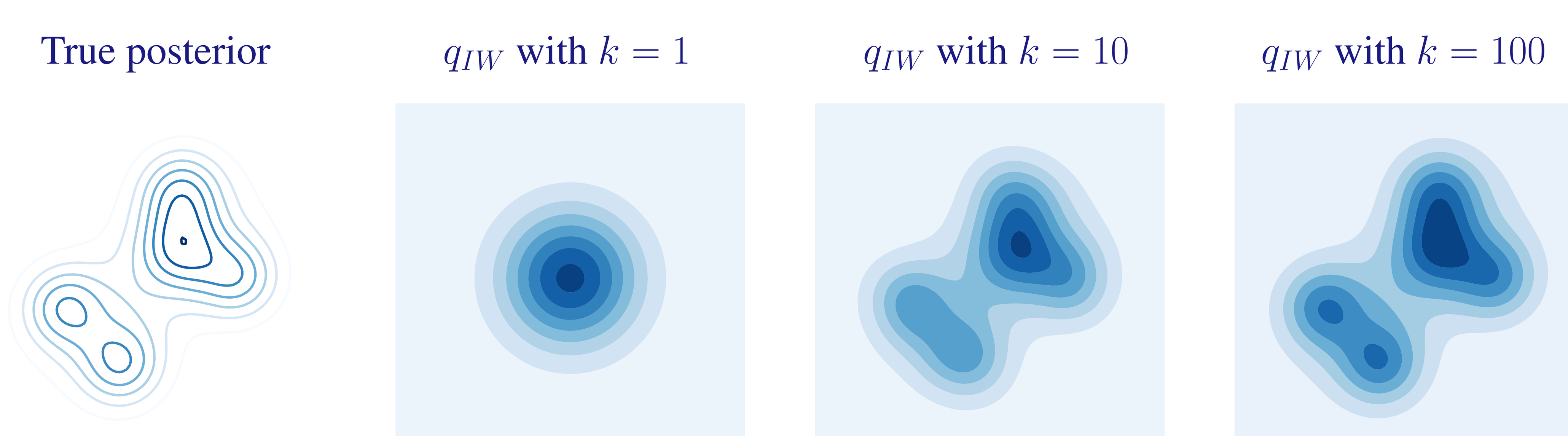
The marginal distribution $q_{IW}(z|x)$ is given by:

$$q_{IW}(z|x) = E_{z_2 \ldots z_k \sim q(\cdot|x)} \left[\frac{p(x,z)}{\frac{1}{k}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right)} \right]$$

Note that, to evaluate q_{IW}, we must approximate the expectation over batches of samples $z_2 \ldots z_k$ from $q(z|x)$. The following are properties of q_{IW}:

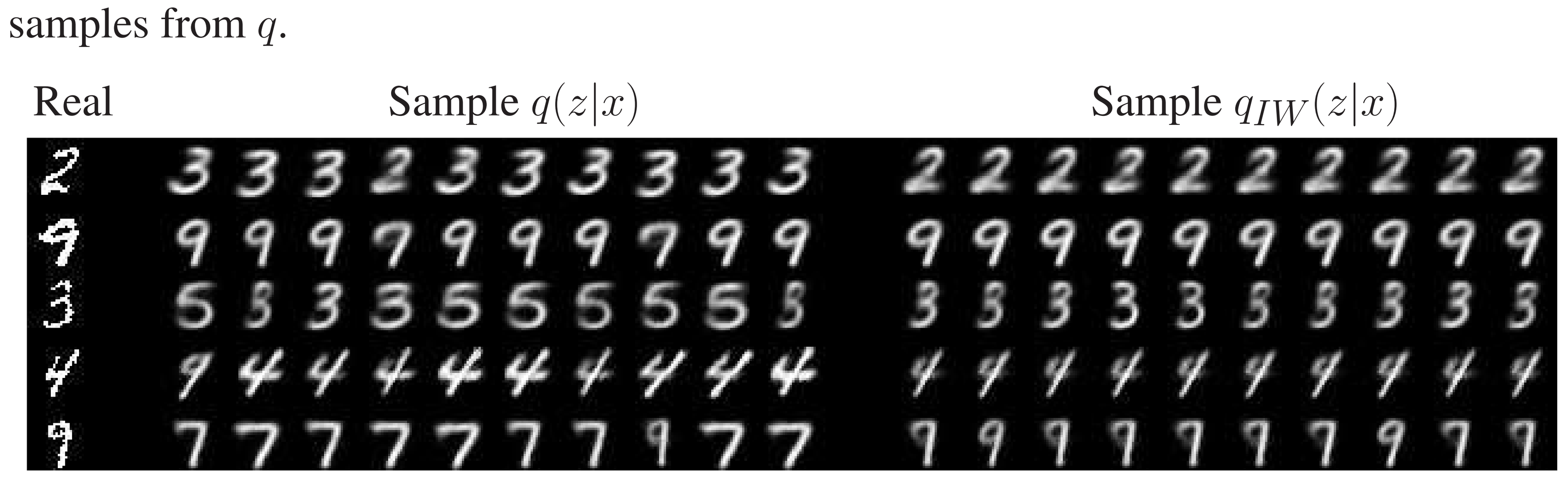
- When $k = 1$, $q_{IW}(z|x)$ will be equal to $q(z|x)$.
- When $k > 1$, the form of q_{IW} depends on the true posterior $p(z|x)$.
- When $k = \infty$, $q_{IW}(z|x)$ becomes the true posterior $p(z|x)$.

Below, we visualize the approximation of q_{IW} to the true posterior with varying number of samples k.



Resampling for Prediction

During training, we sample the q distribution and implicitly weight them with the IWAE ELBO. After training, we need to explicitly reweight samples from q.



The above figure demonstrates the need to sample from q_{IW} rather than $q(z|x)$ for reconstructing MNIST digits. We trained the model to maximize the IWAE ELBO with K=50 and 2 latent dimensions. When we reconstruct samples from $q(z|x)$, we see a number of anomalies. However, if we perform the importance-resampling step, the reconstructions become much more accurate.

VAE to IWAE Bound Proof

If we set the q distribution of the VAE ELBO to $q_{IW}(z|x)$, then we recover the IWAE ELBO:

$$\begin{aligned} &\mathcal{L}_{VAE}[q_{IW}] \\ &= E_{z_1 \ldots z_k \sim q_{IW}(z|x)} \left[\frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x,z_i)}{q_{IW}(z_i|x, z_{\setminus i})} \right) \right] \\ &= E_{z_1 \ldots z_k \sim q(z|x)} \left[\sum_{l=1}^{k} \tilde{w}_l \frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x,z_i)}{\frac{p(x,z_i)}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}}} \right) \right] \\ &= E_{z_1 \ldots z_k \sim q(z|x)} \left[log \left(\frac{1}{k} \sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right) \right] \\ &= \mathcal{L}_{IWAE}[q] \end{aligned}$$

Thus we see that VAE with q_{IW} is equivalent to the IWAE ELBO.

Discussion

Bachman and Precup [1] also showed that the IWAE objective is equivalent to stochastic variational inference with a proposal distribution corrected towards the true posterior via normalized importance sampling.

In light of this, IWAE can be seen as increasing the complexity of the approximate distribution q, similar to other methods that increase the complexity of q, such as Normalizing Flows ([3]), Variational Boosting ([5]) or Hamiltonian variational inference ([6]). An interesting avenue of future work is the comparison of IW-based variational families with alpha-divergences or operator variational objectives.

References

[1] Philip Bachman and Doina Precup. Training Deep Generative Models: Variations on a Theme. *NIPS Approximate Inference Workshop*, 2015.

[2] Yuri Burda, Roger Grosse, and Ruslan Salakhutdinov. Importance weighted autoencoders. *In ICLR*, 2016.

[3] Danilo Jimenez Rezende and Shakir Mohamed. Variational Inference with Normalizing Flows. *In ICML*, 2015.

[4] Diederik P. Kingma and Max Welling. Auto-Encoding Variational Bayes. *In ICLR*, 2014.

[5] Andrew C. Miller, Nicholas Foti, and Ryan P. Adams. Variational Boosting: Iteratively Refining Posterior Approximations. *Advances in Approximate Bayesian Inference, NIPS Workshop*, 2016.

[6] Tim Salimans, Diederik P. Kingma, and Max Welling. Markov chain monte carlo and variational inference: Bridging the gap. *In ICML*, 2015.