REINTERPRETING IMPORTANCE-WEIGHTED AUTOENCODERS

Chris Cremer, Quaid Morris & David Duvenaud
Department of Computer Science
University of Toronto
{ccremer,duvenaud}@cs.toronto.edu
{quaid.morris}@utoronto.ca

ABSTRACT

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. We formally derive this result, and visualize the implicit importance-weighted approximate posterior.

1 BACKGROUND

The importance-weighted autoencoder (IWAE; Burda et al. (2016)) maximizes the following multi-sample evidence lower bound (ELBO):

$$log(p(x)) \geq E_{z_1 \dots z_k \sim q(z|x)} \left[log \left(\frac{1}{k} \sum_{i=1}^{k} \frac{p(x, z_i)}{q(z_i|x)} \right) \right] = L_{IWAE}[q] \qquad \text{(IWAE ELBO)}$$

which is a tighter lower bound than the ELBO maximized by the variational autoencoder (VAE; Kingma & Welling (2014)):

$$log(p(x)) \geq E_{z_1 \dots z_k \sim q(z|x)} \left[\frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x, z_i)}{q(z_i|x)} \right) \right] = L_{VAE}[q]. \qquad \text{(VAE ELBO)}$$

Here we've written the VAE bound as a multisample lower bound to compare it to the IWAE bound. The following equations are the gradients of the VAE ELBO and the IWAE ELBO, respectively:

$$\nabla_\Theta \mathcal{L}_{VAE}[q] = E_{z_1 \dots z_k \sim q(z|x)} \left[\sum_{i=1}^{k} \frac{1}{k} \nabla_\Theta log \left(\frac{p(x, z_i)}{q(z_i|x)} \right) \right] \tag{1}$$

$$\nabla_\Theta \mathcal{L}_{IWAE}[q] = E_{z_1 \dots z_k \sim q(z|x)} \left[\sum_{i=1}^{k} \tilde{w}_i \nabla_\Theta log \left(\frac{p(x, z_i)}{q(z_i|x)} \right) \right] \tag{2}$$

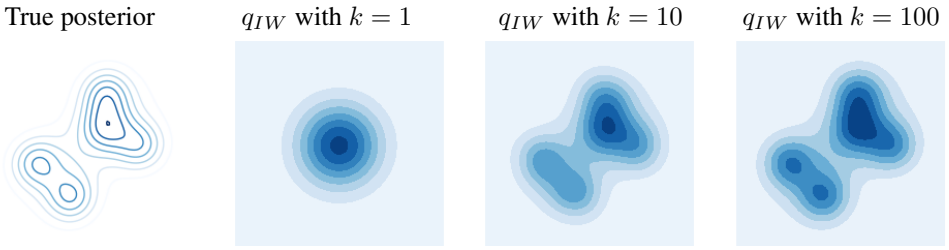

Figure 1: Approximations to a complex true distribution, defined via sampling-importance-resampling. As k grows, this approximation approaches the true distribution.

where

$$\tilde{w}_i = \frac{\frac{p(x,z_i)}{q(z_i|x)}}{\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}}.$$

From equations 1 and 2, we see that the gradient of the VAE ELBO evenly weights the samples, whereas the IWAE gradient weights the samples based on their relative importance $\tilde{w}_i$.

2 DEFINING THE IMPLICIT DISTRIBUTION q_{IW}

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 $z_1 ... z_k$ from $q(z|x)$, the following is the importance weighted q_{IW} distribution as a function of one of the samples, z_i:

$$q_{IW}(z_i|x, z_{\setminus i}) = k\tilde{w}_i q(z_i|x) = \left(\frac{\frac{p(x,z_i)}{q(z_i|x)}}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}} \right) q(z_i|x) = \frac{p(x,z_i)}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}} \tag{3}$$

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

$$q_{IW}(z|x) = E_{z_2 ... z_k \sim q(\cdot|x)} \left[\frac{p(x,z)}{\frac{1}{k}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right)} \right] \tag{4}$$

When $k = 1$, $q_{IW}(z|x)$ will be equal to $q(z|x)$. When $k > 1$, we see that the form of q_{IW} depends on the true posterior p. When $k = \infty$, $q_{IW}(z|x)$ becomes the true posterior $p(z|x)$. See the Appendix for details. Note that, to evaluate q_{IW}, we require an integral over batches of samples $z_2 ... z_k$ from $q(z|x)$. Since this integral is intractable, q_{IW} must be approximated by a finite number of batches.

Figure 1 visualizes q_{IW} on a 2D distribution approximation problem. The base distribution q is a Gaussian. As we increase the number of samples k used for the sampling-resampling, the approximation approaches the true distribution. This distribution is nonparametric in the sense that, as the true posterior grows more complex, so does the shape of q_{IW}.

2.1 RECOVERING THE IWAE BOUND FROM THE VAE BOUND

Here we show that the IWAE ELBO is equivalent to the VAE ELBO, but with a more flexible q_{IW} distribution, implicitly defined by importance reweighting. First, we start by writing the VAE ELBO in its minibatch form, as an average over k samples:

$$\log p(x) \geq \mathcal{L}_{VAE}[q] \quad = E_{z \sim q(z|x)} \left[log \left(\frac{p(x,z)}{q(z|x)} \right) \right] = E_{z_1 ... z_k \sim q(z|x)} \left[\frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x,z_i)}{q(z|x)} \right) \right] \tag{5}$$

If we now set $q(z|x) = q_{IW}(z|x)$, then we recover the IWAE ELBO:

$$\mathcal{L}_{VAE}[q_{IW}] = E_{z_1 ... z_k \sim q_{IW}(z|x)} \left[\frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x,z_i)}{q_{IW}(z_i|x, z_{\setminus i})} \right) \right] \tag{6}$$

$$= E_{z_1 ... z_k \sim q(z|x)} \left[\sum_{l=1}^{k} \tilde{w}_l \frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x,z_i)}{\frac{\frac{p(x,z_i)}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}}}} \right) \right] \tag{7}$$

$$= E_{z_1 ... z_k \sim q(z|x)} \left[log \left(\frac{1}{k} \sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right) \right] = \mathcal{L}_{IWAE}[q] \tag{8}$$

Thus we see that VAE with q_{IW} is equivalent to the IWAE ELBO. For a more detailed derivation, see the Appendix.

Algorithm 1 Sampling from q_{IW}

1: $k \leftarrow$ *number of samples*
2: $q(z|x) = f_\phi(x)$
3: **for** i in $1 \dots k$ **do**
4: $\quad z_i \sim q(z|x)$
5: $\quad w_i = \frac{p(x,z_i)}{q(z_i|x)}$
6: Each $\tilde{w} = w_i / \sum_{i=1}^{k} w_i$
7: $j \sim Cat(\tilde{w})$
8: Return z_j

Figure 2: Algorithm 1 defines the procedure to sample from q_{IW}.

3 SAMPLING q_{IW}

The procedure to sample from $q_{IW}(z|x)$ is shown in Algorithm 1. It is equivalent to sampling-importance-resampling (SIR).

4 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.

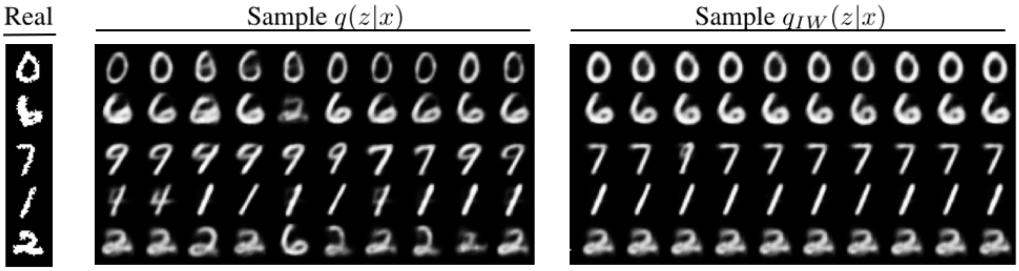

Figure 3: Reconstructions of MNIST samples from $q(z|x)$ and q_{IW}. The model was trained by maximizing the IWAE ELBO with K=50 and 2 latent dimensions. The reconstructions from $q(z|x)$ are greatly improved with the sampling-resampling step of q_{IW}.

In Fig. 3, we demonstrate 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, similar to Appendix C in Burda et al. (2016). When we sample from $q(z|x)$ and reconstruct the samples, we see a number of anomalies. However, if we perform the sampling-resampling step (Algo. 1), then the reconstructions are much more accurate. The intuition here is that we trained the model with q_{IW} with $K = 50$ then sampled from $q(z|x)$ (q_{IW} with $K = 1$), which are very different distributions, as seen in Fig. 1.

5 DISCUSSION

Bachman & Precup (2015) 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 other words, the IWAE lower bound can be interpreted as the standard VAE lower bound with an implicit q_{IW} distribution. We build on this idea by further examining q_{IW} and by providing visualizations to help better grasp the interpretation. 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 (Jimenez Rezende & Mohamed (2015)), Variational Boosting (Miller et al. (2016)) or Hamiltonian variational inference (Salimans et al. (2015)). With this interpretation in mind, we can generalize q_{IW} to be more broadly applicable to any divergence measure. An interesting avenue of future work is the comparison of IW-based variational families with alpha-divergences or operator variational objectives.

ACKNOWLEDGMENTS

We'd like to thank an anonymous ICLR reviewer for providing insightful future directions for this work.

REFERENCES

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

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

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

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

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

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

6 APPENDIX

6.1 DETAILED DERIVATION OF EQUIVALENCE OF VAE AND IWAE BOUND.

First, we start by writing the VAE ELBO in its minibatch form, as an average over k samples:

$$\log p(x) \geq \mathcal{L}_{VAE}[q] = E_{z\sim q(z|x)}\left[log\left(\frac{p(x,z)}{q(z|x)}\right)\right] \quad = E_{z_1\ldots z_k\sim q(z|x)}\left[\frac{1}{k}\sum_{i=1}^{k} log\left(\frac{p(x,z_i)}{q(z|x)}\right)\right] \tag{9}$$

If we now set $q(z|x) = q_{IW}(z|x)$, then we recover the IWAE ELBO:

$$L_{VAE}[q_{IW}] = E_{z_1 \dots z_k \sim q_{IW}(z|x)} \left[\frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x, z_i)}{q_{IW}(z_i|x, z_{\setminus i})} \right) \right] \tag{10}$$

$$= E_{z_1 \dots z_k \sim q_{IW}(z|x)} \left[\frac{1}{k} \sum_{i=1}^{k} log \left(\frac{p(x, z_i)}{\frac{p(x,z_i)}{\frac{1}{k}\sum_{j=1}^{k} \frac{p(x,z_j)}{q(z_j|x)}}} \right) \right] \tag{11}$$

$$= E_{z_1 \dots z_k \sim q_{IW}(z|x)} \left[\frac{1}{k} \sum_{i=1}^{k} log \left(\frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)} \right) \right] \tag{12}$$

$$= E_{z_1 \dots z_k \sim q_{IW}(z|x)} \left[log \left(\frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)} \right) \right] \tag{13}$$

$$= E_{z_1 \dots z_k \sim q(z|x)} \left[\sum_{l=1}^{k} \tilde{w}_l \left(log \left(\frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)} \right) \right) \right] \tag{14}$$

$$= E_{z_1 \dots z_k \sim q(z|x)} \left[log \left(\frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)} \right) \right] = \mathcal{L}_{IWAE}[q] \tag{15}$$

Eqn. 13 follows Eqn. 12 since index i is not present within the sum over i. Similarly, from Eqn. 14 to Eqn. 15, index l is not present within the sum over l and $\sum_{l=1}^{k} \tilde{w}_l$ sums to one. Eqn. 15 is the IWAE ELBO, thus we see that VAE with q_{IW} is equivalent to the IWAE ELBO.

6.2 Proof that q_{IW} is closer to the true posterior than q

Section 6.1 showed that $\mathcal{L}_{IWAE}(q) = \mathcal{L}_{VAE}(q_{IW})$. That is, the IWAE ELBO with the base q is equivalent to the VAE ELBO with the importance weighted q_{IW}. Due to Jensen's Inequality and as shown in Burda et al. (2016), we know that the IWAE ELBO is an upper bound of the VAE ELBO: $L_{IWAE}(q) \geq L_{VAE}(q)$. Furthermore, the log marginal likelihood can be factorized into: $log(p(x)) = L_{VAE}(q) + KL(q||p)$, and rearranged to: $KL(q||p) = log(p(x)) - L_{VAE}(q)$.

Following the observations above and substituting q_{IW} for q:

$$KL(q_{IW}||p) = log(p(x)) - L_{VAE}(q_{IW}) \tag{16}$$

$$= log(p(x)) - L_{IWAE}(q) \tag{17}$$

$$\leq log(p(x)) - L_{VAE}(q) = KL(q||p) \tag{18}$$

Thus, $KL(q_{IW}||p) \leq KL(q||p)$, meaning q_{IW} is closer to the true posterior than q in terms of KL divergence.

Another perspective is in the limit of $k = \infty$. Recall that the marginal likelihood can be approximated by importance sampling:

$$p(x) = E_{q(z|x)} \left[\frac{p(x, z)}{q(z|x)} \right] \approx \frac{1}{k} \sum_{i}^{k} \frac{p(x, z_i)}{q(z_i|x)} \tag{19}$$

where z_i is sampled from $q(z_i|x)$. Thus, the denominator of Eqn. 21 is approximating $p(x)$. As k approaches infinity, q_{IW} approaches the true posterior $p(z|x)$.

Another way to write it:

$$q_{IW}(z_i|x, z_{\setminus i}) = \frac{p(x)}{\frac{1}{k}\left(\frac{p(x,z_i)}{q(z_i|x)} + \sum_{j=2}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right)} p(z_i|x) \tag{20}$$

$$q_{IW}(z|x) = E_{z_2 \dots z_k \sim q(\cdot|x)} \left[\frac{p(x)}{\frac{1}{k}\left(\frac{p(x,z)}{q(z|x)} + \sum_{j=2}^{k} \frac{p(x,z_j)}{q(z_j|x)} \right)} \right] p(z|x) \tag{21}$$

7 VISUALIZING q_{IW} IN 1D

We can look at the intermediate variational distributions with different numbers of samples k in 1 dimension. Fig. 4 demonstrates how the approximate posterior approaches the true posterior as k increases.

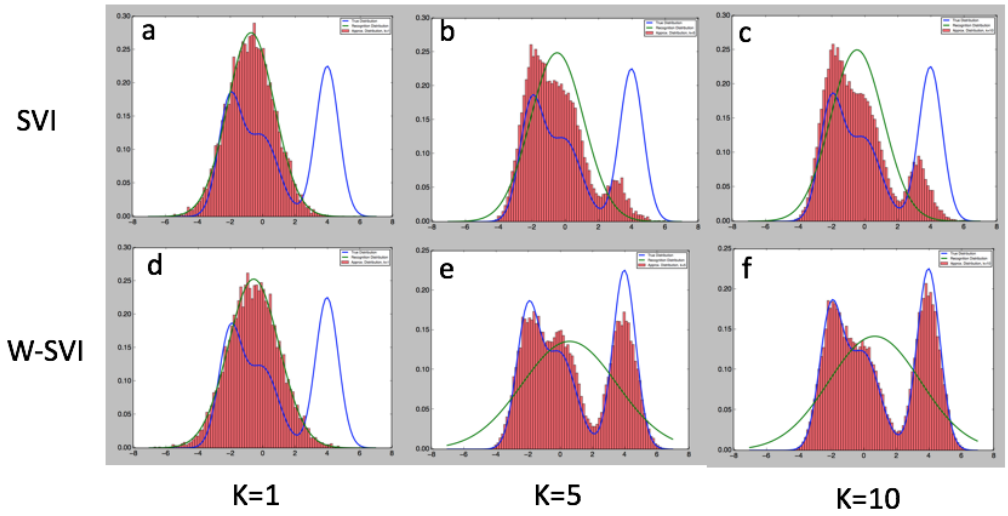

Figure 4: Visualization of the importance weighted posterior. The blue distribution is the intractable distribution that we are trying to approximate. The green distribution is the variational distribution. The variational distributions of a, b, and c were optimized via SVI, whereas d, e, and f were optimized with SVI with the IWAE ELBO. The red histograms are importance weighted samples from the variational distribution.