Supplementary material for
“On casting importance weighted autoencoder to an EM algorithm to learn deep generative models”

1 Proof of Proposition 1.

We can easily check that

$$\nabla_\theta \hat{L}_{\text{IWAE}}(\theta, \phi; x) = \nabla_\theta \log \left( \frac{1}{K} \sum_{k=1}^{K} \frac{p(z, x_k; \theta)}{q(z_k| \mathbf{x}; \phi)} \right)$$

$$= \sum_{k=1}^{K} \frac{w_k}{\sum_{k'} w_{k'}} \cdot \nabla_\theta w_k$$

$$= \sum_{k=1}^{K} \frac{w_k}{\sum_{k'} w_{k'}} \cdot \nabla_\theta \log w_k$$

$$= \sum_{k=1}^{K} \frac{w_k}{\sum_{k'} w_{k'}} \cdot \nabla_\theta \log p(x, z; \theta),$$

where $w_k$ is the weight defined in (2) of the paper, and thus the proof is done. \qed

2 Proof of Proposition 2.

Note that the Chi-squared distance is defined as

$$\chi^2(p||q) = E_{z \sim p} \left( \frac{p(z)}{q(z)} \right) - 1$$

for given two density functions $p$ and $q$. We remind the dominated convergence theorem (DCT) in the sense of the convergence in probability which is summarized in the following lemma.

**Lemma 1.** Suppose $X_n \rightarrow X$ in probability and there is a continuous function $g$ with $g(x) > 0$ for large $x$ with $|x|/g(x) \rightarrow 0$ as $|x| \rightarrow \infty$ so that $E g(X_n) \leq C < \infty$ for all $n$. Then $E X_n \rightarrow EX$ as $n \rightarrow \infty$. 

1
Now we are ready to prove Proposition 2. Let \( z_1, \ldots, z_K \) be random vectors whose density is \( q(z) \). Using

\[
E_q \left( \frac{p(z)}{q(z)} \right) = 1,
\]

we have

\[
\chi^2(p||q) = E_q \left( \frac{p(z)}{q(z)} \right)^2 - 1 = Var_q \left( \frac{p(z)}{q(z)} \right)
\]

In turn, the central limit theorem implies

\[
\sqrt{K} \left( \frac{1}{K} \sum_{k=1}^{K} \frac{p(z_k)}{q(z_k)} - 1 \right) \to N(0, \chi^2(p||q))
\]

and thus

\[
\frac{1}{K} \sum_{k=1}^{K} \frac{p(z_k)}{q(z_k)} - 1 = O_p(K^{-1/2}).
\]

The rest of proof consists of two steps.

[Step 1.] Let \( Y_K := \frac{1}{K} \sum_{k=1}^{K} \left( \frac{p(z_k)}{q(z_k)} - 1 \right) \). We are going to show that \( E(KY_K^3) \to 0 \) as \( K \to \infty \). First, note that \( KY_K^3 \) converges to 0 in probability since \( Y_K = O_p(K^{-1/2}) \). Let \( g(x) := |x|^{4/3} \), then

\[
E(g(Y_K)) = \frac{1}{K^{8/3}} \sum_{k=1}^{K} \left( \frac{p(z_k)}{q(z_k)} - 1 \right)^4 + \sum_{k \neq k'} \frac{1}{K^{8/3}} \sum_{k=1}^{K} \left( \frac{p(z_k)}{q(z_k)} - 1 \right)^4 + \sum_{k \neq k'} \sum_{k'=1}^{K} \frac{1}{K^{8/3}} \sum_{k=1}^{K} \left( \frac{p(z_k)}{q(z_k)} - 1 \right)^2 \frac{1}{K^{8/3}} \sum_{k'=1}^{K} \left( \frac{p(z_{k'})}{q(z_{k'})} - 1 \right)^2
\]

\[
= O(K^{-2/3}) < \infty.
\]

Thus we conclude that \( E(KY_K^3) \to 0 \) by Lemma 1.

[Step 2.] By Taylor’s theorem, there exists \( \xi_K \) between 0 and \( Y_K \) such that

\[
-\log \left( \frac{1}{K} \sum_{k=1}^{K} \frac{p(z_k)}{q(z_k)} \right) = -\log(1 + Y_K) = -Y_K + \frac{1}{2} Y_K^2 - \frac{(1 + \xi_K)^3}{3} Y_K^3.
\]

Since \( \xi_K \) is bounded, there exists a positive constant \( C > 0 \) such that

\[
-\frac{1}{2} Y_K^2 - CY_K^3 \leq -\log(1 + Y_K) \leq -\frac{1}{2} Y_K^2 + CY_K^3.
\]

By taking expectation and multiplying \( 2K \) and using the result of Step 1, we have

\[
2K \cdot D^{IW}(q||p) = \chi^2(p||q) + o(1),
\]

thus the proof is done. \( \square \)
3 Image generation of IWEM

Figure 1: Randomly generated images of IWEM with (Upper) MLP and (Lower) CNN architectures over 4 datasets.