# 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

$$\begin{aligned}
\nabla_\theta \widehat{L}^{\mathrm{IWAE}}(\theta, \phi; \mathbf{x}) &= \nabla_\theta \log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(\mathbf{z}, \mathbf{x}_k; \theta)}{q(\mathbf{z}_k|\mathbf{x}; \phi)}\right) \\
&= \sum_{k=1}^{K}\frac{w_k}{\sum_{k'} w_{k'}} \cdot \frac{\nabla_\theta w_k}{w_k} \\
&= \sum_{k=1}^{K}\frac{w_k}{\sum_{k'} w_{k'}} \cdot \nabla_\theta \log w_k \\
&= \sum_{k=1}^{K}\frac{w_k}{\sum_{k'} w_{k'}} \cdot \nabla_\theta \log p(\mathbf{x}, \mathbf{z}; \theta),
\end{aligned}$$

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

## 2 Proof of Proposition 2.

Note that the Chi-squared distance is defined as

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

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 $X_n \to X$ in probability and there is a continuous function $g$ with $g(x) > 0$ for large $x$ with $|x|/g(x) \to 0$ as $|x| \to \infty$ so that $\mathrm{E}g(X_n) \le C < \infty$ for all $n$. Then $\mathrm{E}X_n \to \mathrm{E}X$ as $n \to \infty$.*

Now we are ready to prove Proposition 2. Let $\mathbf{z}_1, ..., \mathbf{z}_K$ be random vectors whose density is $q(\mathbf{z})$. Using

$$\mathrm{E}_q\left(\frac{p(\mathbf{z})}{q(\mathbf{z})}\right) = 1,$$

we have

$$\chi^2(p||q) = \mathrm{E}_q\left(\frac{p(\mathbf{z})}{q(\mathbf{z})}\right)^2 - 1 = Var_q\left(\frac{p(\mathbf{z})}{q(\mathbf{z})}\right)$$

In turn, the central limit theorem implies

$$\sqrt{K}\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(\mathbf{z}_k)}{q(\mathbf{z}_k)} - 1\right) \to N(0, \chi^2(p||q))$$

and thus

$$\frac{1}{K}\sum_{k=1}^{K}\frac{p(\mathbf{z}_k)}{q(\mathbf{z}_k)} - 1 = O_p(K^{-1/2}).$$

The rest of proof consists of two steps.

**[Step 1.]** Let $Y_K := \frac{1}{K}\sum_{k=1}^{K}\left(\frac{p(\mathbf{z}_k)}{q(\mathbf{z}_k)} - 1\right)$. We are going to show that $\mathrm{E}(KY_K^3) \to 0$ 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

$$\begin{aligned}
\mathrm{E}(g(KY_K) &= \frac{1}{K^{8/3}}\mathrm{E}\left[\sum_{k=1}^{K}\left(\frac{p(\mathbf{z}_k)}{q(\mathbf{z}_k)} - 1\right)^4\right] \\
&= \frac{1}{K^{8/3}}\left[\sum_{k=1}^{K}\mathrm{E}\left[\left(\frac{p(\mathbf{z}_k)}{q(\mathbf{z}_k)} - 1\right)^4\right] + \sum_{k\neq k'}\mathrm{E}\left[\left(\frac{p(\mathbf{z}_k)}{q(\mathbf{z}_k)} - 1\right)^2\right]\mathrm{E}\left[\left(\frac{p(\mathbf{z}_{k'})}{q(\mathbf{z}_{k'})} - 1\right)^2\right]\right] \\
&= O(K^{-2/3}) < \infty.
\end{aligned}$$

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

**[Step 2.]** By Taylor's theorem, there exists $\xi_K$ between 0 and $Y_K$ such that

$$-\log\left(\frac{1}{K}\sum_{k=1}^{K}\frac{p(\mathbf{z}_k)}{q(\mathbf{z}_k)}\right) = -\log(1 + Y_K) = -Y_K + \frac{1}{2}Y_K^2 - \frac{(1+\xi_K)^{-3}}{3}Y_K^3.$$

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

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

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

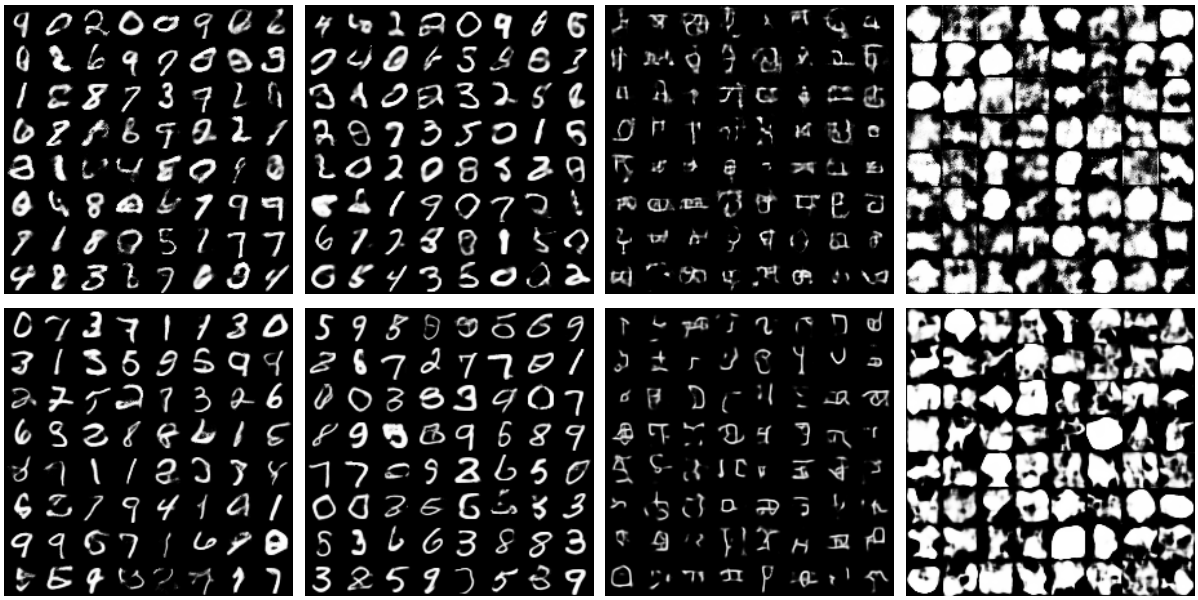

Figure 1: Randomly generated images of IWEM with (**Upper**) MLP and (**Lower**) CNN architectures over 4 datasets.