Assignment 4 — Theoretical Part
Generative Models

Joey Litalien*
IFT6135 Representation Learning, Winter 2018
Université de Montréal
Prof. Aaron Courville
`joey.litalien@umontreal.ca`

1 Reparameterization Trick of Variational Autoencoder

(a) *Proof.* This is rather straightforward. The linearly transformed standard Gaussian noise is given by

$$\mathbf{z} = \mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \boldsymbol{\epsilon}, \qquad \boldsymbol{\epsilon} \sim \mathcal{N}(0, \mathbf{I}). \tag{1}$$

Clearly,

$$\begin{aligned} \mathbb{E}[\mathbf{z}] &= \mathbb{E}[\mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \boldsymbol{\epsilon}] = \mathbb{E}[\mu(\mathbf{x})] + \mathbb{E}[\sigma(\mathbf{x}) \odot \boldsymbol{\epsilon}] \\ &= \mu(\mathbf{x}) \end{aligned}$$

and similarly,

$$\begin{aligned} \sigma^2[\mathbf{z}] &= \mathbb{E}\left[(\mathbf{z} - \mu(\mathbf{x}))^2\right] = \mathbb{E}\left[\left(\mu(\mathbf{x}) + \sigma(\mathbf{x}) \odot \boldsymbol{\epsilon} - \mu(\mathbf{x})\right)^2\right] \\ &= \mathbb{E}\left[(\sigma(\mathbf{x}) \odot \boldsymbol{\epsilon})^2\right] \\ &= \sigma^2(\mathbf{x}). \end{aligned}$$

Hence, (1) has the same mean and variance as $\mathcal{N}(\mu(\mathbf{x}), \sigma^2(\mathbf{x}))$, as desired. □

$\boxed{\text{Do not forget 2nd part}}$

(b)

2 Importance Weighted Autoencoder

(a) *Proof.* We show that IWLB is a lower bound on the log likelihood $\log p(\mathbf{x})$. To simplify notation, let $\mathbf{w}_i = p(\mathbf{x}, \mathbf{z}_i)/q(\mathbf{z}_i \,|\, \mathbf{x})$ denote the unnormalized importance weights for the joint distribution. Using Jensen's inequality and the fact that the average importance weights are an unbiased estimator of $p(\mathbf{x})$, we have that

$$\mathcal{L}_k = \mathbb{E}\left[\log \frac{1}{k} \sum_{i=1}^{k} \mathbf{w}_i\right] \leq \log \mathbb{E}\left[\frac{1}{k} \sum_{i=1}^{k} \mathbf{w}_i\right] = \log p(\mathbf{x}),$$

where the expectations are taken with respect to $q(\mathbf{z} \,|\, \mathbf{x})$. □

*Student ID P1195712

(b) *Proof.* We want to show that ELBO $= \mathcal{L}_1 \leq \mathcal{L}_2 \leq \log p(\mathbf{x})$. Using the fact that $\mathbb{E}_i[\mathbf{w}_i] = \frac{1}{2}(\mathbf{w}_1 + \mathbf{w}_2)$, we have that

$$\begin{aligned}
\mathcal{L}_2 &= \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}\left[\log \frac{1}{2}\left(\frac{p(\mathbf{x},\mathbf{z}_1)}{q(\mathbf{z}_1\,|\,\mathbf{x})} + \frac{p(\mathbf{x},\mathbf{z}_2)}{q(\mathbf{z}_2\,|\,\mathbf{x})}\right)\right] \\
&= \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}\left[\log \mathbb{E}_i\left[\frac{p(\mathbf{x},\mathbf{z}_i)}{q(\mathbf{z}_i\,|\,\mathbf{x})}\right]\right] \\
&\geq \mathbb{E}_{\mathbf{z}_1,\mathbf{z}_2}\left[\mathbb{E}_i\left[\log \frac{p(\mathbf{x},\mathbf{z}_i)}{q(\mathbf{z}_i\,|\,\mathbf{x})}\right]\right] \\
&= \mathbb{E}_{\mathbf{z}}\left[\log \frac{p(\mathbf{x},\mathbf{z})}{q(\mathbf{z}\,|\,\mathbf{x})}\right] \\
&= \mathcal{L}_1,
\end{aligned}$$

where we used Jensen's inequality on the third line. Using the same heuristic, one can actually show that $\mathcal{L}_k \geq \mathcal{L}_m$ for any $k \geq m$.

$\square$

3 MAXIMUM LIKELIHOOD FOR GENERATIVE ADVERSARIAL NETWORKS

Proof. We wish to derive the maximum likelihood learning rule for GANs. In particular, we have an objective function for the generator network G given by

$$J^{(G)} = \mathbb{E}_{\mathbf{x}\sim p_{\mathsf{gen}}} f(\mathbf{x}), \tag{2}$$

and we want to find a function f such that (2) yields maximum likelihood. We start by showing that

$$\frac{\partial}{\partial \boldsymbol{\theta}} J^{(G)} = \mathbb{E}_{\mathbf{x}\sim p_{\mathsf{gen}}} f(\mathbf{x}) \frac{\partial}{\partial \boldsymbol{\theta}} \log p_{\mathsf{gen}}(\mathbf{x}). \tag{3}$$

To do so, we write the expectation as an integral and use Leibniz rule to obtain

$$\frac{\partial}{\partial \boldsymbol{\theta}} J^{(G)} = \frac{\partial}{\partial \boldsymbol{\theta}} \int f(\mathbf{x}) p_{\mathsf{gen}}(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int f(\mathbf{x}) \frac{\partial}{\partial \boldsymbol{\theta}} p_{\mathsf{gen}}(\mathbf{x})\,\mathrm{d}\mathbf{x}\,. \tag{4}$$

Assuming that $p_{\mathsf{gen}} > 0$ everywhere, we can use the identity $g' = g(\log g)'$ for $g > 0$ to rewrite the right-hand side of (4) as

$$\frac{\partial}{\partial \boldsymbol{\theta}} J^{(G)} = \int f(\mathbf{x}) p_{\mathsf{gen}}(\mathbf{x}) \frac{\partial}{\partial \boldsymbol{\theta}} \log p_{\mathsf{gen}}(\mathbf{x})\,\mathrm{d}\mathbf{x}\,, \tag{5}$$

which is just (3). This gives us an expression where we can relate the gradients of the likelihood with the samples generated by G. However, we would like to have these gradients in terms of samples that came from the real data distribution p_{data}. To fix this, we can perform a simple importance sample trick by setting

$$f(\mathbf{x}) = -\frac{p_{\mathsf{data}}(\mathbf{x})}{p_{\mathsf{gen}}(\mathbf{x})}$$

to reweight our sampling. Using our assumption that the discriminator D is optimal with

$$D^*(\mathbf{x}) = \frac{p_{\mathsf{data}}(\mathbf{x})}{p_{\mathsf{data}}(\mathbf{x}) + p_{\mathsf{gen}}(\mathbf{x})},$$

which we can write as $D^*(\mathbf{x}) = \sigma(a(\mathbf{x}))$ for some network output $a(\mathbf{x})$ (and σ is the logistic sigmoid function), we can solve for f directly:

$$\begin{aligned}
\sigma(a(\mathbf{x})) &= \frac{1}{1+\exp(-a(\mathbf{x}))} = \frac{p_{\mathsf{data}}(\mathbf{x})}{p_{\mathsf{data}}(\mathbf{x}) + p_{\mathsf{gen}}(\mathbf{x})} = \frac{1}{1 + p_{\mathsf{gen}}(\mathbf{x})/p_{\mathsf{data}}(\mathbf{x})} \\
&= \frac{1}{1 - f^{-1}}.
\end{aligned}$$

Therefore, $f(\mathbf{x}) = -\exp(a(\mathbf{x}))$. We conclude that the objective function maximizing likelihood must be given by

$$J^{(G)} = \frac{1}{2}\,\mathbb{E}_{\mathbf{z}} \exp\!\big(\sigma^{-1}(D(G(\mathbf{z})))\big) = \frac{1}{2}\,\mathbb{E}_{\mathbf{z}}\left[\frac{D(G(\mathbf{z}))}{1 - D(G(\mathbf{z}))}\right].$$

$\square$