Assignment Four

Matthew C. Scicluna
Département d'Informatique et de Recherche Opérationnelle
Université de Montréal
Montréal, QC H3T 1J4
matthew.scicluna@umontreal.ca

April 9, 2018

1 Reparameterization Trick of Variational Autoencoder

Consider a generative model that factorizes as follows $p(x,z) = p(x|z)p(z)$, with $p(x|z) = p(x|h_\theta(z))$ mapped through a neural net and θ being the set of parameters for the generative network (i.e. decoder), a simple distribution parameterized by $h(\cdot)$. In the case of Gaussian, $h_\theta(z)$ refers to the mean and variance, per dimension as it is fully factorized in the common setting. We have $z \in \mathbb{R}^K$, and $p(z) = \mathcal{N}(0, I_K)$. The framework of auto-encoding variational Bayes considers maximizing the variational lower bound on the log-likelihood $\mathcal{L}(\theta,\phi) \leq \log p(x)$, which is expressed as:

$$\mathcal{L}(\theta,\phi) = \mathbb{E}_{q_\phi}\{\log p(x|z)\} - KL\left(q_\phi || p_\theta(z)\right)$$

where ϕ is the set of parameters used for the inference network (i.e. encoder). The reparameterization trick used in the original work rewrites the random variable in the variational distribution as:

$$z = \mu(x) + \sigma(x) \odot \epsilon$$

where $\epsilon \sim \mathcal{N}(0, I)$

(a) We show that the linearly transformed standard Gaussian noise has the same mean and variance as $\mathcal{N}(z;\mu(x),\sigma^2(x))$. This can be easily seen since:

$$\begin{aligned}\mathbb{E}_\epsilon\left\{z\right\} &= \mathbb{E}_\epsilon\left\{\mu(x)+\sigma(x)\odot\epsilon\right\}\\ &= \mathbb{E}_\epsilon\left\{\mu(x)\right\}+\mathbb{E}_\epsilon\left\{\sigma(x)\odot\epsilon\right\}\\ &= \mu(x)+\sigma(x)\odot\underbrace{\mathbb{E}_\epsilon\left\{\epsilon\right\}}_{=0}\\ &= \mu(x)\end{aligned}$$

$$\begin{aligned}Var\left\{z\right\} &= Var\left\{\mu(x)+\sigma(x)\odot\epsilon\right\}\\ &= Var\left\{\sigma(x)\odot\epsilon\right\}\\ &= \sigma(x)^2\odot\underbrace{Var\left\{\epsilon\right\}}_{=I}\\ &= I\sigma(x)^2\end{aligned}$$

If we write $z=\mu(x)+S(x)\epsilon$, where $S(x)\in\mathbb{R}^{K\times K}$, the new distribution this reparameterization induces is $\mathcal{N}(z;\mu(x),S(x)S(x)^T)$. This can be computed since:

$$\begin{aligned}\mathbb{E}_\epsilon\left\{z\right\} &= \mathbb{E}_\epsilon\left\{\mu(x)+S(x)\epsilon\right\}\\ &= \mathbb{E}_\epsilon\left\{\mu(x)\right\}+\mathbb{E}_\epsilon\left\{S(x)\epsilon\right\}\\ &= \mu(x)+S(x)\underbrace{\mathbb{E}_\epsilon\left\{\epsilon\right\}}_{=0}\\ &= \mu(x)\end{aligned}$$

$$\begin{aligned}Var\left\{z\right\} &= Var\left\{\mu(x)+S(x)\epsilon\right\}\\ &= Var\left\{S(x)\epsilon\right\}\\ &= S(x)Var\left\{\epsilon\right\}S(x)^T\\ &= S(x)S(x)^T\end{aligned}$$

(b) If the traditional mean field variational method is used, i.e. if we factorize the variational distribution as a product of distributions: $q^{mf}(z_i)=\prod_j\mathcal{N}(z_{i,j}|m_{i,j},\sigma_{i,j})$ for each x_i, and we maximize the lower bound with respect to the variational parameters and model parameters iteratively, the inference network used in the variational autoencoder q_ϕ will not outperform the mean field method on the training set. This is because the model will learn, for each datapoint, the optimal mean and variance parameters for its reconstruction. Wheras with the encoder network, the mean and variance parameters will not necessarily be optimal for any particular datapoint, but will be for the training set as a whole. The advantage of using an encoder as in VAE is efficiency, the inference doesn't grow linearly with the data as it would with the traditional mean field variational method.

2 Importance Weighted Autoencoder

When training a variational autoencoder, the standard training objective is to maximize the evidence lower bound (ELBO). Here we consider another lower bound, called the Importance Weighted Lower Bound (IWLB), a tighter bound than ELBO, defined as

$$\mathcal{L}_k = \mathbb{E}_{z_{1:k} \sim q(z|x)} \left\{ \log \frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)} \right\}$$

for an observed variable x and a latent variable z, k being the number of importance samples. The model we are considering has joint that factorizes as $p(z, x) = p(x|z)p(z)$, x and z being the observed and latent variables, respectively.

(a) We show that IWLB is a lower bound on the log likelihood $\log p(x)$. We first use Jensens inequality and the concavity of log, and marginalize to get the desired result:

$$\begin{aligned}
\mathcal{L}_k &= \mathbb{E}_{z_{1:k} \sim q(z|x)} \left\{ \log \frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)} \right\} \\
&\leq \log \mathbb{E}_{z_{1:k} \sim q(z|x)} \left\{ \frac{1}{k} \sum_{j=1}^{k} \frac{p(x, z_j)}{q(z_j|x)} \right\} \\
&= \log \frac{1}{k} \sum_{j=1}^{k} \mathbb{E}_{z_j \sim q(z|x)} \left\{ \frac{p(x, z_j)}{q(z_j|x)} \right\} \\
&= \log \frac{1}{k} \sum_{j=1}^{k} \int_{z_j} q(z_j|x) \frac{p(x, z_j)}{q(z_j|x)} dz_j \\
&= \log \frac{1}{k} \sum_{j=1}^{k} \int_{z_j} p(x, z_j) dz_j \\
&= \log \frac{1}{k} \sum_{j=1}^{k} p(x) \\
&= \log p(x)
\end{aligned}$$

(b) Given $k = 2$, we prove that $\mathcal{L}_2$ is a tighter bound than the ELBO (with $k = 1$). It is enough to show that $\mathcal{L}_2 \geq$ ELBO (since both are $\leq \log p(x)$). We show this using an argument similar to Burda et al (2016) [1] . First notice that $\mathcal{L}_1$ is equivalent to the ELBO, which is clear when we write the ELBO in the following form:

$$ELBO = \mathbb{E}_{z \sim q(z|x)} \left\{ \log \frac{p(x, z)}{q(z|x)} \right\}$$

From this, it is enough to show that $\mathcal{L}_2 \geq \mathcal{L}_1$. Notice that:

$$\begin{aligned}
\mathcal{L}_2 &= \mathbb{E}_{z_1,z_2 \sim q(z|x)} \left\{ \log \frac{1}{2} \left(\frac{p(x,z_1)}{q(z_1|x)} + \frac{p(x,z_2)}{q(z_2|x)} \right) \right\} \\
&\overset{(a)}{=} \mathbb{E}_{z_1,z_2 \sim q(z|x)} \left\{ \log \mathbb{E}_{i \sim Unif(\{1,2\})} \left\{ \frac{p(x,z_i)}{q(z_i|x)} \right\} \right\} \\
&\overset{(b)}{\geq} \mathbb{E}_{z_1,z_2 \sim q(z|x)} \left\{ \mathbb{E}_{i \sim Unif(\{1,2\})} \left\{ \log \frac{p(x,z_i)}{q(z_i|x)} \right\} \right\} \\
&= \mathbb{E}_{z_1 \sim q(z_1|x)} \left\{ \log \frac{p(x,z_1)}{q(z_1|x)} \right\}
\end{aligned}$$

Where (a) follows since $p(i=1) = p(i=2) = \frac{1}{2}$ for $i \sim Unif(\{1,2\})$ and viewing $\frac{p(x,z_i)}{q(z_i|x)}$ as a function of i. (b) follows from Jensens inequality.

3 Maximum Likelihood for Generative Adversarial Networks

The original GAN objective is the following:

$$\begin{aligned}
&\max_D \mathbb{E}_{x \sim p_{data}} \{\log D(x)\} + \mathbb{E}_{z \sim p_z} \{\log(1 - D(G(z)))\} \\
&\max_G \mathbb{E}_{z \sim p_z} \{\log D(G(z))\}
\end{aligned}$$

This generator objective can be generalized by replacing the log with a general function f:

$$\max_G \mathbb{E}_{z \sim p_z} \{f(D(G(z)))\}$$

We find a function f such that the objective corresponds to maximum likelihood, assuming the discriminator is optimal, i.e. that:

$$D(x) = \frac{p_{data}(x)}{p_{data}(x) + p_{gen}(x)}$$

Where p_{gen} is the probability distribution of generated samples $G(z),\ z \sim p_z$. Following an appoach similar to Goodfellow (2014) [2], we claim that the following f satisfies the condition:

$$f(D(G(z))) = \exp(\sigma^{-1}(D(G(z))))$$

Where σ is the logistic sigmoid (supposing that D applies σ at its topmost layer). We show the equivalence by showing that the gradients for Maximum Likelihood are the same in expectation as those for GANs under the aforementioned conditions. The derivative of the log likelihood is:

$$\frac{\partial}{\partial\theta} \sum_{i=1}^{N} \log p_{gen}(x_i) = \mathbb{E}_{x \sim p_{data}} \left\{ \frac{\partial}{\partial\theta} \log p_{gen}(x) \right\}$$

The derivative of the generator objective with the chosen f is:

$$\begin{aligned}
\frac{\partial}{\partial\theta}\mathbb{E}_{x\sim p_{gen}}\left\{f(x)\right\} &= \frac{\partial}{\partial\theta}\int_x f(x)p_{gen}(x)dx \\
&\overset{(a)}{=} \int_x f(x)\frac{\partial}{\partial\theta}p_{gen}(x)dx \\
&\overset{(b)}{=} \int_x f(x)\frac{\partial}{\partial\theta}\exp(\log(p_{gen}(x)))dx \\
&= \int_x f(x)\exp(\log(p_{gen}(x)))\frac{\partial}{\partial\theta}\log(p_{gen}(x))dx \\
&= \int_x f(x)\frac{\partial}{\partial\theta}\log(p_{gen}(x))p_{gen}(x)dx \\
&= \mathbb{E}_{x\sim p_{gen}}\left\{f(x)\frac{\partial}{\partial\theta}\log p_{gen}(x)\right\}
\end{aligned}$$

Where (a) follows from Leibnitz's rule if we assume that p_{gen} and its derivative are continuous and (b) follows if we assume that $p_{gen}(x) \geq 0$ everywhere. If we set $f(x) = \frac{p_{data}(x)}{p_{gen}(x)}$ we see that:

$$\mathbb{E}_{x\sim p_{gen}}\left\{f(x)\frac{\partial}{\partial\theta}\log p_{gen}(x)\right\} = \mathbb{E}_{x\sim p_{data}}\left\{\frac{\partial}{\partial\theta}\log p_{gen}(x)\right\}$$

Note that p_{gen} found in f is a copy of the actual p_{gen} – this is to ensure that $\frac{\partial}{\partial\theta}f = 0$ in the above equations. If we assume an optimal discriminator, we can get the importance sampling ratio above from D:

$$\begin{aligned}
D(x) &= \frac{p_{data}(x)}{p_{data}(x) + p_{gen}(x)} = \sigma(a(x)) \\
\Rightarrow a(x) &= \log\frac{p_{data}(x)}{p_{gen}(x)}
\end{aligned}$$

We show how our definition of f from before is equivalent to the importance sampling ratio needed to make the gradients equal:

$$\begin{aligned}
f(D(x)) &= \exp(\sigma^{-1}(D(x))) \\
&= \exp(a(x)) \\
&= \exp\log\frac{p_{data}(x)}{p_{gen}(x)} \\
&= \frac{p_{data}(x)}{p_{gen}(x)}
\end{aligned}$$

References

[1] Y. Burda, R. B. Grosse, and R. Salakhutdinov, “Importance weighted autoencoders,” *CoRR*, vol. abs/1509.00519, 2015.

[2] I. J. Goodfellow, “On distinguishability criteria for estimating generative models,” *ArXiv e-prints*, Dec. 2014.