# MathDNN Homework 11

Department of Computer Science and Engineering
2021-16988 Jaewan Park

## Problem 1

(a) Since log is a concave function, using Jensen's inequality, we obtain the following.

$$\begin{aligned}
\mathrm{VLB}_{\theta,\phi}^{(K)}(x) &= \mathbb{E}_{Z_1,\cdots,Z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x\,|\,Z_k)p_Z(Z_k)}{q_\phi(Z_k\,|\,x)}\right] \\
&\le \log\left(\mathbb{E}_{Z_1,\cdots,Z_K\sim q_\phi(z|x)}\left[\frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x\,|\,Z_k)p_Z(Z_k)}{q_\phi(Z_k\,|\,x)}\right]\right) \\
&= \log\left(\frac{1}{K}\sum_{k=1}^{K}\mathbb{E}_{Z_k\sim q_\phi(z|x)}\left[\frac{p_\theta(x\,|\,Z_k)p_Z(Z_k)}{q_\phi(Z_k\,|\,x)}\right]\right) \\
&= \log\left(\frac{1}{K}\sum_{k=1}^{K}p_\theta(x)\right) = \log p_\theta(x)
\end{aligned}$$

(b) Using the given hint together with Jensen's inequality, we obtain the following.

$$\begin{aligned}
\mathrm{VLB}_{\theta,\phi}^{(K)}(x) &= \mathbb{E}_{Z_1,\cdots,Z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x\,|\,Z_k)p_Z(Z_k)}{q_\phi(Z_k\,|\,x)}\right] \\
&= \mathbb{E}_{Z_{i_1},\cdots,Z_{i_M}\sim q_\phi(z|x)}\left[\log\left(\mathbb{E}_{I=\{i_1,\cdots,i_M\}}\left[\frac{1}{M}\sum_{m=1}^{M}\frac{p_\theta(x|Z_{i_m})p_Z(Z_{i_m})}{q_\phi(Z_{i_m}|x)}\right]\right)\right] \\
&\ge \mathbb{E}_{Z_{i_1},\cdots,Z_{i_M}\sim q_\phi(z|x)}\left[\mathbb{E}_{I=\{i_1,\cdots,i_M\}}\left[\log\frac{1}{M}\sum_{m=1}^{M}\frac{p_\theta(x|Z_{i_m})p_Z(Z_{i_m})}{q_\phi(Z_{i_m}|x)}\right]\right] \\
&= \mathbb{E}_{I=\{i_1,\cdots,i_M\}}\left[\mathbb{E}_{Z_{i_1},\cdots,Z_{i_M}\sim q_\phi(z|x)}\left[\log\frac{1}{M}\sum_{m=1}^{M}\frac{p_\theta(x|Z_{i_m})p_Z(Z_{i_m})}{q_\phi(Z_{i_m}|x)}\right]\right] \\
&= \mathbb{E}_{I=\{i_1,\cdots,i_M\}}\left[\mathrm{VLB}_{\theta,\phi}^{(M)}(x)\right] = \mathrm{VLB}_{\theta,\phi}^{(M)}(x)
\end{aligned}$$

(c) We should choose $q_\phi$ powerful enough so that $q_\phi(Z_k\,|\,x) = p_\theta(Z_k\,|\,x)$ for all $k = 1,\cdots,K$.

$$\begin{aligned}
\mathrm{VLB}_{\theta,\phi}^{(K)}(x) &= \mathbb{E}_{Z_1,\cdots,Z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x\,|\,Z_k)p_Z(Z_k)}{q_\phi(Z_k\,|\,x)}\right] \\
&= \mathbb{E}_{Z_1,\cdots,Z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K}\frac{p_\theta(x\,|\,Z_k)p_Z(Z_k)}{p_\theta(Z_k\,|\,x)}\right] \\
&= \mathbb{E}_{Z_1,\cdots,Z_K\sim q_\phi(z|x)}\left[\log \frac{1}{K}\sum_{k=1}^{K}p_\theta(x)\right] = p_\theta(x)
\end{aligned}$$

## Problem 2

(a) Since log is a concave function, using Jensen's inequality, we obtain the following.

$$\begin{aligned}\log p_\theta(X_i) &= \log\left(\mathbb{E}_{Z\sim q_\phi(z\mid X_i)}\left[\frac{p_\theta(X_i\mid Z)r_\lambda(Z)}{q_\phi(Z\mid X_i)}\right]\right)\\ &\geq \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}\left[\log\left(\frac{p_\theta(X_i\mid Z)r_\lambda(Z)}{q_\phi(Z\mid X_i)}\right)\right] = \mathrm{VLB}_{\theta,\phi,\lambda}(X_i)\end{aligned}$$

(b) Gradients regarding $\theta$ and $\lambda$ can be easily derived as the follwing.

$$\nabla_\theta \mathrm{VLB}_{\theta,\phi,\lambda}(X_i) = \nabla_\theta \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}\left[\log\left(\frac{p_\theta(X_i\mid Z)r_\lambda(Z)}{q_\phi(Z\mid X_i)}\right)\right] = \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}[\nabla_\theta \log p_\theta(X_i\mid Z)]$$

$$\nabla_\lambda \mathrm{VLB}_{\theta,\phi,\lambda}(X_i) = \nabla_\lambda \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}\left[\log\left(\frac{p_\theta(X_i\mid Z)r_\lambda(Z)}{q_\phi(Z\mid X_i)}\right)\right] = \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}[\nabla_\lambda \log r_\lambda(Z)]$$

For gradients on $\phi$, we can use the log-derivative trick.

$$\begin{aligned}\nabla_\phi \mathrm{VLB}_{\theta,\phi,\lambda}(X_i) &= \nabla_\phi \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}\left[\log\left(\frac{p_\theta(X_i\mid Z)r_\lambda(Z)}{q_\phi(Z\mid X_i)}\right)\right] = \nabla_\phi \int \log\left(\frac{p_\theta(X_i\mid z)r_\lambda(z)}{q_\phi(z\mid X_i)}\right)q_\phi(z\mid X_i)dz\\ &= \int\left(-\frac{\nabla_\phi q_\phi(z\mid X_i)}{q_\phi(z\mid X_i)}q_\phi(z\mid X_i) + \log\left(\frac{p_\theta(X_i\mid z)r_\lambda(z)}{q_\phi(z\mid X_i)}\right)\nabla_\phi q_\phi(z\mid X_i)\right)dz\\ &= \int \log\left(\frac{p_\theta(X_i\mid z)r_\lambda(z)}{q_\phi(z\mid X_i)}\right)\frac{\nabla_\phi q_\phi(z\mid X_i)}{q_\phi(z\mid X_i)}q_\phi(z\mid X_i)dz\\ &= \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}\left[\log\left(\frac{p_\theta(X_i\mid Z)r_\lambda(Z)}{q_\phi(Z\mid X_i)}\right)\nabla_\phi \log q_\phi(Z\mid X_i)\right]\end{aligned}$$

(c) We can rewrite VLB as the following.

$$\begin{aligned}\mathrm{VLB}_{\theta,\phi,\lambda}(X_i) &= \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}\left[\log\left(\frac{p_\theta(X_i\mid Z)r_\lambda(Z)}{q_\phi(Z\mid X_i)}\right)\right]\\ &= \mathbb{E}_{Z\sim q_\phi(z\mid X_i)}[\log p_\theta(X_i\mid Z)] - D_{\mathrm{KL}}(q_\phi(z\mid X_i)\,||\,r_\lambda(z))\end{aligned}$$

Then the first term can be calculated as

$$\begin{aligned}\mathbb{E}_{Z\sim q_\phi(z\mid X_i)}[\log p_\theta(X_i\mid Z)] &= \mathbb{E}_{Z\sim\mathcal{N}(\mu_\phi(X_i),\Sigma_\phi(X_i))}\left[\log\mathcal{N}\left(f_\theta(Z),\sigma^2 I\right)\right]\\ &= \mathbb{E}_{Z\sim\mathcal{N}(\mu_\phi(X_i),\Sigma_\phi(X_i))}\Big[-\frac{1}{2}(X_i - f_\theta(Z))^\intercal\left(\sigma^2 I\right)^{-1}(X_i - f_\theta(Z))\\ &\qquad\qquad - \frac{1}{2}\log\left((2\pi)^k\left|\sigma^2 I\right|\right)\Big]\\ &= -\frac{1}{2\sigma^2}\mathbb{E}_{Z\sim\mathcal{N}(\mu_\phi(X_i),\Sigma_\phi(X_i))}\left[\|X_i - f_\theta(Z)\|^2\right] - \frac{k}{2}\log\left(2\pi\sigma^2\right)\\ &= -\frac{1}{2\sigma^2}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left[\left\|X_i - f_\theta\left(\mu_\phi(X_i) + \sqrt{\Sigma_\phi(X_i)}\varepsilon\right)\right\|^2\right] - \frac{k}{2}\log\left(2\pi\sigma^2\right).\end{aligned}$$

Using the reparametrization trick simplifies the expectation term, and makes able the gradient of this first term be directly calculated. The second term also can be calculated as the following.

$$
\begin{aligned}
&D_{\mathrm{KL}}(q_\phi(z \mid X_i) \,||\, r_\lambda(z)) \\
&\quad = \frac{1}{2}\left(\mathrm{tr}\Big(\mathrm{diag}(\lambda_2)^{-1}\Sigma_\phi(X_i)\Big) + (\lambda_1 - \mu_\phi(X_i))^\intercal \mathrm{diag}(\lambda_2)^{-1}(\lambda_1 - \mu_\phi(X_i)) - k + \log\left(\frac{\det\left(\mathrm{diag}(\lambda_2)\right)}{\det\left(\Sigma_\phi(X_i)\right)}\right)\right)
\end{aligned}
$$

The gradient of the second term can also be directly calculated, so we can obtain the gradients via backpropagation.

# Problem 4

(a) Let $p_A = (p_{A1}, p_{A2}, p_{A3})$ and $p_B = (p_{B1}, p_{B2}, p_{B3})$. Then

$$\mathbb{E}_{p_A, p_B}[\text{points for } B] = p_{A1}p_{B2} + p_{A2}p_{B3} + p_{A3}p_{B1} - p_{A1}p_{B3} - p_{A2}p_{B1} - p_{A3}p_{B2}.$$

Suppose $p_A^* = (p_{A1}^*, p_{A2}^*, p_{A3}^*)$, $p_B^* = (p_{B1}^*, p_{B2}^*, p_{B3}^*)$ is the solution for the given problem. Then we have

$$
\begin{aligned}
&p_{A1}^*p_{B2} + p_{A2}^*p_{B3} + p_{A3}^*p_{B1} - p_{A1}^*p_{B3} - p_{A2}^*p_{B1} - p_{A3}^*p_{B2} \\
&\qquad \le p_{A1}^*p_{B2}^* + p_{A2}^*p_{B3}^* + p_{A3}^*p_{B1}^* - p_{A1}^*p_{B3}^* - p_{A2}^*p_{B1}^* - p_{A3}^*p_{B2}^* \\
&\qquad\qquad \le p_{A1}p_{B2}^* + p_{A2}p_{B3}^* + p_{A3}p_{B1}^* - p_{A1}p_{B3}^* - p_{A2}p_{B1}^* - p_{A3}p_{B2}^*.
\end{aligned}
$$

for all $p_A, p_B \in \Delta^3$. If $p_A^* = p_B^* = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, all three terms are 0, so it is a solution of the problem. Now we should show that this is the only solution for the problem. Suppose $p_A^* \neq \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ and generally let $p_{A1}^* < p_{A2}^*$. Now substitute $p_A = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ and $p_B = (0, 0, 1)$, then

$$
\begin{aligned}
&p_{A1}^*p_{B2} + p_{A2}^*p_{B3} + p_{A3}^*p_{B1} - p_{A1}^*p_{B3} - p_{A2}^*p_{B1} - p_{A3}^*p_{B2} = p_{A2}^* - p_{A1}^* > 0 \\
&p_{A1}p_{B2}^* + p_{A2}p_{B3}^* + p_{A3}p_{B1}^* - p_{A1}p_{B3}^* - p_{A2}p_{B1}^* - p_{A3}p_{B2}^* = 0
\end{aligned}
$$

so the inequality becomes false. Similarly, suppose $p_B^* \neq \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$ and generally let $p_{B1}^* < p_{B2}^*$. Now substitute $p_A = (0, 0, 1)$ and $p_B = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, then

$$
\begin{aligned}
&p_{A1}^*p_{B2} + p_{A2}^*p_{B3} + p_{A3}^*p_{B1} - p_{A1}^*p_{B3} - p_{A2}^*p_{B1} - p_{A3}^*p_{B2} = 0 \\
&p_{A1}p_{B2}^* + p_{A2}p_{B3}^* + p_{A3}p_{B1}^* - p_{A1}p_{B3}^* - p_{A2}p_{B1}^* - p_{A3}p_{B2}^* = p_{B1}^* - p_{B2}^* < 0
\end{aligned}
$$

so the inequality also becomes false. Therefore always $p_A^* = p_B^* = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, so it is the unique solution for the problem.

(b) If $B$ chooses $p_B$ as given, the expected points for $B$ is always 0 regardless of $A$, so $A$ can choose any strategy. However, if $B$ chooses strategies other than $p_B = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right)$, choosing any strategy may not be optimal for $A$. Choosing $p_B = (1, 0, 0), (0, 1, 0), (0, 0, 1)$ results in $\mathbb{E}_{p_A, p_B}[\text{points for } B] > 0$ each when $A$ chooses strategies such that $p_{A3} > p_{A2}$, $p_{A1} > p_{A3}$, $p_{A2} > p_{A1}$.