Mathematical Foundations of Deep Neural Networks, M1407.001200
E. Ryu
Fall 2022

Homework 11
Due 5pm, Wednesday, November 23, 2022

Problem 1: *VLB for IWAE.* The standard variational lower bound (VLB) of VAE is

$$\log(p_\theta(x)) \geq \mathrm{VLB}_{\theta,\phi}(x) = \mathbb{E}_{Z\sim q(z|x)}\left[\log\left(\frac{p_\theta(x \mid Z)p_Z(Z)}{q_\phi(Z \mid x)}\right)\right],$$

where $p_\theta(z \mid x)$ is the true posterior and $q_\phi(z \mid x)$ is the approximate posterior. Define

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

to be the VLB for importance weighted autoencoders (IWAE) [1]. To clarify, $Z_1,\dots,Z_K$ are sampled independently from $q_\phi(z \mid x)$. Note that IWAE with $K = 1$ coincides with the standard VAE, and $\mathrm{VLB}^{(1)}_{\theta,\phi} = \mathrm{VLB}_{\theta,\phi}$. Show:

(a) $\log p_\theta(x) \geq \mathrm{VLB}^{(K)}_{\theta,\phi}(x)$ for all x and $K \geq 1$.

(b) If $K \geq M$, then $\mathrm{VLB}^{(K)}_{\theta,\phi}(x) \geq \mathrm{VLB}^{(M)}_{\theta,\phi}(x)$ for all x.

(c) Let $X_1,\dots,X_N$ be data for training the IWAE. Show that if q_ϕ is "powerful enough", then

$$\underset{\theta\in\Theta}{\text{maximize}} \quad \sum_{i=1}^{N}\log p_\theta(X_i) \quad = \quad \underset{\theta\in\Theta,\phi\in\Phi}{\text{maximize}} \quad \sum_{i=1}^{N}\mathrm{VLB}^{(K)}_{\theta,\phi}(X_i).$$

What should be the precise meaning of "powerful enough"?

Hint. For (a), use the Jensen's inequality. For (b), let $I \subset \{1,\dots,K\}$ with $|I| = M$ be a uniformly distributed subset of distinct indices from $\{1,\dots,K\}$. Then, $\mathbb{E}_{I=\{i_1,\dots,i_M\}}\left[\frac{a_{i_1}+\cdots+a_{i_M}}{M}\right] = \frac{a_1+\cdots+a_K}{K}$ for any sequence of numbers $a_1,\dots,a_K$.

Remark. This analysis shows that $\mathrm{VLB}^{(K)}_{\theta,\phi}$ provides a tighter approximation of the log likelihood than $\mathrm{VLB}_{\theta,\phi}$. However, using $\mathrm{VLB}^{(K)}_{\theta,\phi}$ requires more computation than $\mathrm{VLB}_{\theta,\phi}$.

Solution. (a) follows from Jensen's inequality:

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

(b) follows from

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

(c) if the approximate posterior q_ϕ is powerful enough to exactly represent the true posterior, then there exists a $\phi^\star$ such that $q_{\phi^\star}(Z_k \mid x) = p_\theta(Z_k \mid x)$ for $k = 1, \ldots, K$. Then

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

■

Problem 2: *VAE with trainable prior.* In this problem, we consider the setup of training a VAE with a trainable prior. Specifically, we assume $Z \sim r_\lambda(z)$, where λ is a trainable parameter, and $X \sim p_\theta(x \mid Z)$. Let $q_\phi(z \mid X)$ be the approximate posterior. Let

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

(a) Show that $\log p_\theta(X_i) \ge \mathrm{VLB}_{\theta,\phi,\lambda}(X_i)$.

(b) Describe how to evaluate stochastic gradients of $\mathrm{VLB}_{\theta,\phi,\lambda}(X_i)$ using the log-derivative trick.

(c) Assume $r_\lambda = \mathcal{N}(\lambda_1, \text{diag}(\lambda_2))$, where $\lambda_1, \lambda_2 \in \mathbb{R}^k$, $q_\phi(z \mid X_i) = \mathcal{N}(\mu_\phi(X_i), \Sigma_\phi(X_i))$ with diagonal Σ_ϕ, and $p_\theta(X_i \mid z) = \mathcal{N}(f_\theta(z), \sigma^2 I)$. Describe how to evaluate stochastic gradients of $\text{VLB}_{\theta,\phi,\lambda}(X_i)$ using the reparameterization trick.

Solution

(a) Using Jensen's inequality,

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

(b) Using log-derivative trick(details are in problem 1 of HW 10):

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

(c) VLB can be reformulated by:

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

The first term is as follows:

$$\mathbb{E}_{Z\sim q_\phi(z|X_i)}\left[\log p_\theta\left(X_i \mid Z\right)\right] = -\frac{1}{2\sigma^2}\mathbb{E}_{Z\sim\mathcal{N}\left(\mu_\phi(X_i),\Sigma_\phi(X_i)\right)}\left\|X_i - f_\theta(Z)\right\|^2.$$

But in this equation, gradient of inside of expectation is not able to calculated directly. So if we use reparametrization-trick, then the first term becomes:

$$\frac{1}{2\sigma^2}\mathbb{E}_{\varepsilon\sim\mathcal{N}(0,I)}\left\|X_i - f_\theta\left(\mu_\phi\left(X_i\right) + \Sigma_\phi^{1/2}(X_i)\varepsilon\right)\right\|^2$$

In this equation, after sampling, the gradient can be calculated directly. Therefore, backprop becomes possible. The second term can be calculated by problem 5 of HW9:

$$\begin{aligned}&D_{\text{KL}}\left(q_\phi\left(z|X_i\right) \| r_\lambda(z)\right) \\ &= \frac{1}{2}\left(\text{tr}\left([\text{diag}(\lambda_2)]^{-1}\Sigma_\phi(X_i)\right) + (\lambda_1 - \mu_\phi(X_i))^T[\text{diag}(\lambda_2)]^{-1}(\lambda_1 - \mu_\phi(X_i)) - k + \log\frac{\det(\text{diag}(\lambda_2))}{\det(\Sigma_\phi(X_i))}\right).\end{aligned}$$

The gradient of the second term can be calculated directly. Therefore, backpropagation is possible for both terms.

Problem 3: *Anomaly detection via flow models.* Assume we have a trained flow model that we use to evaluate the likelihood function p_θ. In this problem, you will use this trained flow model to perform anomaly detection between the MNIST and KMNIST datasets. In step 1, load the MNIST and KMNIST datasets, and split the MNIST test dataset into "validation" and "test" sets. In step 2, define the flow model. In step 3, load the trained flow model. In step 4, calculate the mean and standard deviation of

$$\{\log p_\theta(Y_i)\}_{i=1}^M,$$

where $Y_1, \ldots, Y_M$ are the validation data. Define a threshold to be mean $-$ 3 standard deviations, and define inputs with log likelihood below this threshold to be anomalies. In step 5, check how many of the MNIST images within the test set are classified as anomalies and report the type I error rate. In step 6, check how many of the KMNIST images are classified as non-anomalies and report the type II error rate. Download the starter code `flow_anomaly.py`, which provides the implementation of steps 1–3. Complete the implementation of steps 4–6.

Remark. In this problem, we split the test data into validation and test sets because the entire training set was already used to train the flow model. If we were to train the flow model from scratch, it would be better to split the training set into the training and validation sets to set aside the validation data for step 3.

Solution. See `flow_anomaly_sol.py`. ■

Problem 4: *Rock paper scissors and minimiax optimization.* Consider a game of rock paper scissors between players A and B. Players A and B play randomized strategies with

$$p_A = \begin{bmatrix} \mathbb{P}(A \text{ plays rock}) \\ \mathbb{P}(A \text{ plays paper}) \\ \mathbb{P}(A \text{ plays scissors}) \end{bmatrix}, \qquad p_B = \begin{bmatrix} \mathbb{P}(B \text{ plays rock}) \\ \mathbb{P}(B \text{ plays paper}) \\ \mathbb{P}(B \text{ plays scissors}) \end{bmatrix}.$$

Define

$$\Delta^3 = \{p = (p_1, p_2, p_3) \in \mathbb{R}^3 \,|\, p_1, p_2, p_3 \geq 0,\ p_1 + p_2 + p_3 = 1\}$$

so that $p_A, p_B \in \Delta^3$. In the game, a player receives 1 point for a win, -1 points for a loss, and 0 points for a draw. Consider the minimax problem

$$\underset{p_A \in \Delta^3}{\text{minimize}} \quad \underset{p_B \in \Delta^3}{\text{maximize}} \quad \mathbb{E}_{p_A, p_B}[\text{points for } B].$$

(a) Show that

$$p_A^\star = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}, \qquad p_B^\star = \begin{bmatrix} 1/3 \\ 1/3 \\ 1/3 \end{bmatrix}$$

is the unique solution to the minimax problem.

(b) Note that if $p_B = (1/3, 1/3, 1/3)$, then $\mathbb{E}_{p_A, p_B}[\text{points for } B] = 0$ regardless of how A plays. Does this mean any strategy $p_A \in \Delta^3$ is optimal for player A? (Here, the word "optimal" is used informally. Think about whether any $p_A \in \Delta^3$ is a best strategy for A.)

Clarification. We say $(\theta^\star, \phi^\star)$ is a solution to the minimax problem

$$\underset{\theta \in \Theta}{\text{minimize}} \quad \underset{\phi \in \Phi}{\text{maximize}} \quad L(\theta, \phi)$$

if $\theta^\star \in \Theta$, $\phi^\star \in \Phi$, and

$$L(\theta^\star, \phi) \leq L(\theta^\star, \phi^\star) \leq L(\theta, \phi^\star)$$

for all $\theta \in \Theta$ and $\phi \in \Phi$, i.e., unilaterally deviating from $\theta^\star$ increases the value of L and unilaterally deviating from $\phi^\star$ decreases the value of L.

Remark. In the setup of GANs (which is what this problem is intended to prepare you for), if the generator is perfect, the discriminator cannot do better than a 50-50 guess in detecting fakes. However, the discriminator is still forced to learn to distinguish imperfect fakes, as otherwise, the generator can take advantage of the discriminator.

Solution.

(a) Let

$$p_A = \begin{bmatrix} P_{Ar} \\ P_{Ap} \\ P_{As} \end{bmatrix}, \qquad p_B = \begin{bmatrix} P_{Br} \\ P_{Bp} \\ P_{Bs} \end{bmatrix}, \qquad \mathbf{M} = \begin{bmatrix} 0 & -1 & 1 \\ 1 & 0 & -1 \\ -1 & 1 & 0 \end{bmatrix}.$$

Then, minimax problem becomes

$$\underset{p_A \in \Delta^3}{\text{minimize}} \quad \underset{p_B \in \Delta^3}{\text{maximize}} \quad \mathbb{E}_{p_A, p_B}[\text{points for } B] = p_A{}^\intercal \mathbf{M} p_B.$$

Let's prove

$$\underset{p_A \in \Delta^3}{\text{minimize}} \quad \underset{p_B \in \Delta^3}{\text{maximize}} \quad p_A{}^\intercal \mathbf{M} p_B \;=\; \underset{p_B \in \Delta^3}{\text{maximize}} \quad \underset{p_A \in \Delta^3}{\text{minimize}} \quad p_A{}^\intercal \mathbf{M} p_B.$$

By

$$p_A{}^\intercal \mathbf{M} p_B \;\geq\; \underset{p_A \in \Delta^3}{\text{minimize}} \;\; p_A{}^\intercal \mathbf{M} p_B,$$

$$\underset{p_A \in \Delta^3}{\text{minimize}} \;\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; p_A{}^\intercal \mathbf{M} p_B \;\geq\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; \underset{p_A \in \Delta^3}{\text{minimize}} \;\; p_A{}^\intercal \mathbf{M} p_B.$$

Also,

$$\begin{aligned} &\underset{p_A \in \Delta^3}{\text{minimize}} \;\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; p_A{}^\intercal \mathbf{M} p_B \;\leq\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; p_A{}^\intercal \mathbf{M} p_B \\ =\; &\underset{p_A \in \Delta^3}{\text{minimize}} \;\; p_A{}^\intercal \mathbf{M} p_B \;\leq\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; \underset{p_A \in \Delta^3}{\text{minimize}} \;\; p_A{}^\intercal \mathbf{M} p_B. \end{aligned}$$

Since $\mathbf{M} = -\mathbf{M}^\intercal$, and

$$\begin{aligned} L^\star = \underset{p_A \in \Delta^3}{\text{minimize}} \;\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; p_A{}^\intercal \mathbf{M} p_B &= \underset{p_A \in \Delta^3}{\text{minimize}} \;\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; (p_A{}^\intercal \mathbf{M} p_B)^\intercal \\ &= \underset{p_A \in \Delta^3}{\text{minimize}} \;\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; -p_B{}^\intercal \mathbf{M} p_A \\ &= - \underset{p_A \in \Delta^3}{\text{maximize}} \;\; \underset{p_B \in \Delta^3}{\text{minimize}} \;\; p_B{}^\intercal \mathbf{M} p_A \\ &= - \underset{p_A \in \Delta^3}{\text{minimize}} \;\; \underset{p_B \in \Delta^3}{\text{maximize}} \;\; p_A{}^\intercal \mathbf{M} p_B \\ &= -L^\star \end{aligned}$$

Therefore, $L^\star = 0$.

If $\mathbf{M} p_B{}^\star \neq \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$, at least one of the element of $\mathbf{M} p_B{}^\star = \begin{bmatrix} P_{Bs} - P_{Bp} \\ P_{Br} - P_{Bs} \\ P_{Bp} - P_{Br} \end{bmatrix}$ should be negative, since there is a order of three probabilities. Therefore, $\underset{p_A \in \Delta^3}{\text{minimize}} \;\; p_A{}^\intercal (\mathbf{M} p_B{}^\star) < 0$, which contradicts the fact that $L^\star = 0$. Therefore, $\mathbf{M} P_B{}^\star = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}$. By solving, $P_{Br}^\star = P_{Bp}^\star = P_{Bs}^\star = \frac{1}{3}$. Similarly, $\mathbf{M} P_A{}^\star = 0$, so $P_{Ar}^\star = P_{Ap}^\star = P_{As}^\star = \frac{1}{3}$.

(b) The answer is negative. If B fixes its strategy to equalize each probability, A can choose any strategy because $\mathbb{E}_{p_A, p_B}[\text{points for } B]$ is always 0. However, when A chooses $p_A \neq \begin{bmatrix} \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \end{bmatrix}$, then B can choose one of the strategy $P_B = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$ to make $\mathbb{E}_{p_A, p_B}[\text{points for } B] > 0$. Therefore, in minimax problem, considering only one variable while the other one fixed is not a good method to find optimal solution. ■

References

[1] Y. Burda, R. Grosse, and R. Salakhutdinov, Importance weighted autoencoders, *ICLR*, 2016.