THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

MSBD 5012: Machine Learning
Homework 4 Solutions

This assignment is for self-practice.

Solutions will be provided later.

Question 1: Let $p(\mathbf{x}, \mathbf{z})$ and $q(\mathbf{z})$ be two probability distributions. Show that

$$\log p(\mathbf{x}) \geq E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}|\mathbf{z}) - \mathcal{D}_{KL}(q(\mathbf{z})||p(\mathbf{z}))$$

where $\mathcal{D}_{KL}(q(\mathbf{z})|p(\mathbf{z})) = E_{\mathbf{z}\sim q(\mathbf{z})} \log q(\mathbf{z}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{z})$.

Note that the RHS of the ineqaulity is known as the **variational lower bound** of $\log p(\mathbf{x})$, or the **evidence lower bound (ELBO)**. Another way to write the inequality is as follows:

$$\log p(\mathbf{x}) \geq E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}, \mathbf{z}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log q(\mathbf{z}) = E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}, \mathbf{z}) + H(q).$$

Reflect on how the variational lower bound is used variational autoencoder. You can do this by pointing out what are the two distributions.

Solution: Consider $\mathcal{D}_{KL}(q(\mathbf{z})|p(\mathbf{z}|\mathbf{x}))$:

$$\begin{aligned}
\mathcal{D}_{KL}(q(\mathbf{z})|p(\mathbf{z}|\mathbf{x})) &= E_{\mathbf{z}\sim q(\mathbf{z})} \log q(\mathbf{z}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{z}|\mathbf{x}) \\
&= E_{\mathbf{z}\sim q(\mathbf{z})} \log q(\mathbf{z}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log \frac{p(\mathbf{x}|\mathbf{z})p(\mathbf{z})}{p(\mathbf{x})} \\
&= E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}|\mathbf{z}) + E_{\mathbf{z}\sim q(\mathbf{z})} \log q(\mathbf{z}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{z}) \\
&= E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}) - E_{\mathbf{z}\sim q(\mathbf{z})} \log p(\mathbf{x}|\mathbf{z}) + \mathcal{D}_{KL}(q(\mathbf{z})|p(\mathbf{z})).
\end{aligned}$$

The equation follows from the fact that $\mathcal{D}_{KL}(q(\mathbf{z})|p(\mathbf{z}|\mathbf{x})) \geq 0$.

In VAE, the target distribution is $p_\theta(\mathbf{x}^{(i)}|\mathbf{z})$ and the variational distribution is $q(\mathbf{z}|\mathbf{x}^{(i)})$. The first version of the inequality is used.

Question 2: Suppose two random variables x and z are related by the following equation:

$$z = x^2 + 2x + \frac{x^2}{2}\epsilon,$$

where ϵ is a random variable that follows the normal distribution $\mathcal{N}(0, 1)$, i.e.,

$$p(\epsilon) = \frac{1}{\sqrt{2\pi}} e^{\frac{-\epsilon^2}{2}}.$$

What is the density function of the conditional distribution $p(z|x)$?

(Note that the purpose of this problem is to help you understand reparameterization.)

Solution: Since ϵ is Gaussian, z is also Gaussian. Its means is $\mu_z = x^2 + 2x$, and its variance is $\sigma_z^2 = x^2/2$. Hence, its density function is:

$$p(z|x) = \frac{1}{\sqrt{\pi x^2}} e^{-\frac{(z-(x^2+2x))^2}{x^2/2}}$$

Question 3: What are the main differences between variational autoencoder (VAE) and generative adversarial network (GAN) in terms their functionalities and the ways they operate?

Solution: Both VAE and GAN take a collection $\{\mathbf{x}^{(i)}\}$ of unlabeled data as input and assume the data are generated from a random latent vector $\mathbf{z}$. VAE learns a conditional distribution $p(\mathbf{x}|\mathbf{z})$ while GAN learns a deterministic function $\mathbf{x} = g(\mathbf{z})$.

Both VAE and GAN represent the relationship between $\mathbf{x}$ and $\mathbf{z}$ using a deep neural network, which is called the generative model (aka decode in VAE). VAE learns the parameters of the generative network by minimizing the KL divergence between the real data distribution and the model distribution. The objective in intractable. So, an encoder network is introduced to provide a lower bound. The parameters of the encoder and the decoder are trained simultaneously.

GAN learns the parameters of the generative network by minimizing the Jensen-Shannon divergence between the real data distribution and the model distribution. The objective in intractable. So, a discriminator network is introduced to provide an approximation. The parameters of the generator and the discriminator are trained alternately.

Question 4: What are the objective functions for the discriminator and the generator in GAN? What are the objective functions for the critic and generator in WGAN?

Solution: The discriminator in GAN maximizing:

$$V(G, D) = E_{\mathbf{x}\sim p_r(\mathbf{x})}[\log D(\mathbf{x})] + E_{\mathbf{x}\sim p_g(\mathbf{x})}[\log(1 - D(\mathbf{x}))],$$

p_r is the real data distribution, p_g is the generator distribution, and $D(\mathbf{x})$ is the probability that $\mathbf{x}$ is a real example.

Theoretically, the generator of GAN should minimize $E_{\mathbf{x}\sim p_g(\mathbf{x})}[\log(1 - D(\mathbf{x}))]$. However, this leads to unstable training. In practice, it minimizes the following function instead:

$$-E_{\mathbf{x}\sim p_g(\mathbf{x})}[\log D(\mathbf{x})].$$

The critic of WGAN maximizes the following function:

$$L_c = E_{\mathbf{x}\sim p_r}[f(\mathbf{x})] - E_{\mathbf{x}\sim p_g}[f(\mathbf{x})],$$

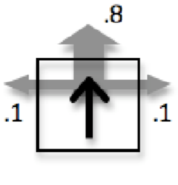
where $f(\mathbf{x})$ is the critic function represented by a neural network and the weights of the network are restricted to $[-c, c]$. The generator of WGAN minimizes:

$$-E_{\mathbf{x}\sim p_g(\mathbf{x})}[\log f^*(\mathbf{x})],$$

f^* is the critic function learning by the critic.

Question: 5 Consider an agent that acts in the gridworld shown below. The agent always starts in state $(1, 1)$, marked with the letter S. There are two terminal goal states, $(3, 2)$ with reward $+5$ and $(3, 1)$ with reward -5. Rewards are 0 in non-terminal states. (The reward for a state is received as the agent moves into the state.) The transition function is such that the intended agent movement (North, South, West, or East) happens with probability 0.8. With probability 0.1 each, the agent ends up in one of the states perpendicular to the intended direction. If a collision with a wall happens, the agent stays in the same state.

		+5
S		**-5**



The expected immediate reward function $r(s, a) = \sum_{s'} r(s, a, s')P(s'|s, a)$ is as follows:

$r(s,a)$	N	S	W	E
$(1,1)$	0	0	0	0
$(1,2)$	0	0	0	0
$(2,1)$	-0.5	-0.5	0	-4
$(2,2)$	0.5	0.5	0	4
$(3,1)$	0	0	0	0
$(3,2)$	0	0	0	0

(a) Assume the initial value function $Q_0(s,a) = 0$ for all states s and actions a. Let $\gamma = 0.9$. The Q-function Q_1 after the first value iteration is the same as $r(s,a)$. What is the Q-function Q_2 after the second value iteration? What is the greedy policy π_2 based on Q_2. In case of ties, list all tied actions.

(b) Suppose the agent does not know the transition probabilities and the reward function, and it tries to learn by interacting with the environment. Assume the Q-learning algorithm is used with $Q(s,a) = 0$ initially. Let $\alpha = 0.1$ and $\gamma = 0.9$. Update the Q-function using the following experience tuples. Show the function after each update.

s	a	r	s'
(2, 2)	E	5	(3, 2)
(2, 1)	N	0	(2, 2)
(1, 2)	E	0	(2, 2)
(1, 1)	N	0	(1, 2)

Give the greedy policy based on the latest Q function. In case of ties, list all tied actions.

Solution: (a) First, note that $V_1(s') = \max_{a'} Q_1(s',a')$ is as follows: Q_2 can be obtained from V_1 as

s'	$V_1(s')$
$(1,1)$	0
$(1,2)$	0
$(2,1)$	0
$(2,2)$	4
$(3,1)$	0
$(3,2)$	0

follows:

$$Q_2(s,a) = r(s,a) + \gamma \sum_{s'} P(s'|s,a) V_1(s').$$

Hence, we have:

$Q_2((1,1),N) = 0 + \gamma * (0.8 * V_1((1,2)) + 0.1 * V_1((1,1)) + 0.1 * V_1((2,1))) = 0,$

$Q_2((1,1),S) = 0 + \gamma * (0.8 * V_1((1,1)) + 0.1 * V_1((1,1)) + 0.1 * V_1((2,1))) = 0,$

$Q_2((1,1),W) = 0 + \gamma * (0.8 * V_1((1,1)) + 0.1 * V_1((1,1)) + 0.1 * V_1((1,2))) = 0,$

$Q_2((1,1),E) = 0 + \gamma * (0.8 * *V_1((2,1)) + 0.1 * V_1((1,2)) + 0.1 * V_1((1,1))) = 0,$

$Q_2((1,2),N) = 0 + \gamma * (0.8 * V_1((1,2)) + 0.1 * V_1((1,2)) + 0.1 * V_1((2,2))) = 0.36,$

$Q_2((1,2),S) = 0 + \gamma * (0.8 * V_1((1,1)) + 0.1 * V_1((1,2)) + 0.1 * V_1((2,2))) = 0.36,$

$Q_2((1,2),W) = 0 + \gamma * (0.8 * V_1((1,2)) + 0.1 * V_1((1,1)) + 0.1 * V_1((1,2))) = 0,$

$Q_2((1,2),E) = 0 + \gamma * (0.8 * V_1((2,2)) + 0.1 * V_1((1,1)) + 0.1 * V_1((1,2))) = 2.88,$

$Q_2((2,1),N) = -0.5 + \gamma * (0.8 * V_1((2,2)) + 0.1 * V_1((1,1)) + 0.1 * V_1((3,1)) = 2.38,$

$Q_2((2,1),S) = -0.5 + \gamma * (0.8 * V_1((2,1)) + 0.1 * V_1((1,1)) + 0.1 * V_1((3,1))) = -0.5$

$Q_2((2,1),W) = 0 + \gamma * (0.8 * V_1((1,1)) + 0.1 * V_1((2,1)) + 0.1 * V_1((2,2))) = 0.36$

$Q_2((2,1),E) = -4 + \gamma * (0.8 * V_1((3,1)) + 0.1 * V_1((2,1)) + 0.1 * V_1((2,2))) = -3.64,$

$Q_2((2,2),N) = 0.5 + \gamma * (0.8 * V_1((2,2)) + 0.1 * V_1((1,2)) + 0.1 * V_1((3,2))) = 3.38,$

$Q_2((2,2),S) = 0.5 + \gamma * (0.8 * V_1((2,1)) + 0.1 * V_1((1,2)) + 0.1 * V_1((3,2))) = 0.5,$

$Q_2((2,2),W) = 0 + \gamma * (0.8 * V_1((1,2)) + 0.1 * V_1((2,2)) + 0.1 * V_1((2,1))) = 0.36,$

$Q_2((2,2),E) = 4 + \gamma * (0.8 * V_1((3,2)) + 0.1 * V_1((2,2)) + 0.1 * V_1((2,1))) = 4.36.$

$Q_2(s,a)$	N	S	W	E
$(1,1)$	0	0	0	0
$(1,2)$	0.36	0.36	0	2.88
$(2,1)$	2.38	-0.5	0.36	-3.64
$(2,2)$	3.38	0.5	0.36	4.36
$(3,1)$	0	0	0	0
$(3,2)$	0	0	0	0

In summary, we have
The greedy policy $\pi(s) = \arg\max_a Q_2(s,a)$ is as follows: $\pi_2((1,1)) = \{N, S, W, E\}$, $\pi_2((1,2)) = E$, $\pi_2((2,1)) = N$, $\pi_2((2,2)) = E$.

(b) The Q-learning update rule is:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)].$$

For $s = (2,2), a = E, r = 5, s' = (3,2)$, $Q((2,2),E) = 0 + 0.1 * (5 + 0.9 * 0 - 0) = 0.5$.

$Q(s,a)$	N	S	W	E
$(1,1)$	0	0	0	0
$(1,2)$	0	0	0	0
$(2,1)$	0	0	0	0
$(2,2)$	0	0	0	0.5
$(3,1)$	0	0	0	0
$(3,2)$	0	0	0	0

For $s = (2,1), a = N, r = 0, s' = (2,2)$, $Q((2,1),N) = 0 + 0.1 * (0 + 0.9 * 0.5 - 0) = 0.045$.

$Q(s,a)$	N	S	W	E
$(1,1)$	0	0	0	0
$(1,2)$	0	0	0	0
$(2,1)$	0.045	0	0	0
$(2,2)$	0	0	0	0.5
$(3,1)$	0	0	0	0
$(3,2)$	0	0	0	0

For $s = (1,2), a = E, r = 0, s' = (2,2)$, $Q((1,2),E) = 0 + 0.1 * (0 + 0.9 * 0.5 - 0) = 0.045$.

$Q(s,a)$	N	S	W	E
$(1,1)$	0	0	0	0
$(1,2)$	0	0	0	0.045
$(2,1)$	0.045	0	0	0
$(2,2)$	0	0	0	0.5
$(3,1)$	0	0	0	0
$(3,2)$	0	0	0	0

For $s = (1,1), a = N, r = 0, s' = (1,2)$, $Q((1,1),N) = 0 + 0.1 * (0 + 0.9 * 0.045 - 0) = 0.00405$.
The greedy policy $\pi(s) = \arg\max_a Q(s,a)$ is as follows: $\pi((1,1)) = N$, $\pi((1,2)) = E$, $\pi((2,1)) = N$, $\pi_2((2,2)) = E$.

$Q(s,a)$	N	S	W	E
$(1,1)$	0.00405	0	0	0
$(1,2)$	0	0	0	0.045
$(2,1)$	0.045	0	0	0
$(2,2)$	0	0	0	0.5
$(3,1)$	0	0	0	0
$(3,2)$	0	0	0	0