THE HONG KONG UNIVERSITY OF SCIENCE & TECHNOLOGY

MSBD 5012: Machine Learning

Homework 4

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.

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.,

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

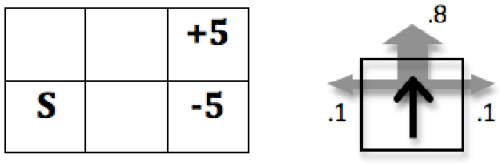
What is the density function of the conditional distribution $p(z|x)$?

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

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?

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?

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.



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.