DDA 4230 Assignment 3

Name: Xiang Fei ; ID: 12009414

P1.

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

For (s_1, a, s_2) : $Q(s_1,a) = 0 + 0.1 \times (10 + 0.9 \times 0 - 0) = 1$

	a	b	c	d
s_1	1	0		
s_2			0	0

For (s_2, d, s_2) : $Q(s_2,d) = 0 + 0.1 \times (10 + 0.9 \times 0 - 0) = 1$

	a	b	c	d
s_1	1	0		
s_2			0	1

For (s_2, c, s_1) : $Q(s_2,c) = 0 + 0.1 \times (-10 + 0.9 \times 1 - 0) = -0.91$

$\pi(s_1) = b$, $\pi(s_2) = c$

P2.

The maximum sum of rewards is 6.1

When state 2 → state 1, we can get the max rewards 3 in one step. After that, there is at least one step to wait and then we can perform this step again. Therefore, for 5 steps, the transition can be excuted twice. For 4 steps, the max sum of rewards is 6. Then, the best reward from state 1 to state 0 is 0.1, so the max sum of rewards of 5 steps is 6.1 following the path 0 → 2 → 1 → 2 → 1 → 0

P3.

1. This part of code is to perform epsilon-greedy algorithm

firstly, I use random.random() to obtain a number between 0 and 1, then compare it with epsilon to simulate the probability. if the number is smaller than epsilon, then directly choose the action; otherwise, choose the action based on the Q values.

Here, epsilon serves as the exploration parameter, it allows the algorithm to explore and its value balances exploration and exploitation, therefore, epsilon is essential.

2. The used parameters:

learning rate: 2.5×10^{-4}

batch size: 128

buffer size: 10000

The learning rate affects the speed and convergence of the learning process, intuitively, 1×10^{-4} is a frequently used value, and experiment shows that when choosing 2.5×10^{-4}, the network converges relatively fast and smoothly. For batch size, normally we have 32, 64, 128, I choose 128 and it performs well.

3. My code can run without errors and I finished all the parts which are required for us to code. But it seems that my codes still have some small problems since the variable infos is empty, that's why I can't draw a reasonable graph. Hope that you can give points as appropriate. I have also put a lot of effort into the code.