ADI | Homework 5.
Reinforcement Learning

Exercise 1

a) A Q-Learning update is given by:

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t \left[c_t + \gamma \min_{a' \in \mathcal{A}} Q_t(x_{t+1}, a') - Q_t(x_t, a_t) \right]$$

Where t is the timestep, x_t, a_t and c_t are the position, the action and the cost at timestep t respectively, γ is the discount and α_t is the step-size.

- x_t = (16, 3)
- a_t = D
- c_t = 0.05
- γ = 0.95
- α_t = 0.1
- x_{t+1} = (16, 4)

Q_{t+1}((16, 3), D) = 0.32 + 0.1*(0.05 + 0.95*0.29 - 0.32) = 0.32055

b) A SARSA update is given by:

$$Q_{t+1}(x_t, a_t) = Q_t(x_t, a_t) + \alpha_t \left[c_t + \gamma Q_t(x_{t+1}, a_{t+1}) - Q_t(x_t, a_t) \right]$$

Where t is the timestep, x_t, a_t and c_t are the position, the action and the cost at timestep t respectively, γ is the discount and α_t is the step-size.

- x_t = (16, 3)
- a_t = D
- c_t = 0.05
- γ = 0.95
- α_t = 0.1
- x_{t+1} = (16, 4)
- a_{t+1} = L

Q_{t+1}((16, 3), D) = 0.32 + 0.1*(0.05 + 0.95*0.36 - 0.32) = 0.3272

c) In the off-policy learning, the Q-values are updated using the Q-value of the next state x_{t+1} and the greedy action a', as can be seen by $\min_{a' \in \mathcal{A}} Q_t(x_{t+1}, a')$. In other words, it estimates the *return* (total discounted future reward) for state-action pairs assuming a greedy policy was followed despite the fact that it's not following a greedy policy.

In the on-policy learning, the Q-values are updated using the Q-value of the next state x_{t+1} and the current policy's action a'', as can be seen by $Q_t(x_{t+1}, a_{t+1})$. It estimates the *return* for the state-action pairs assuming the current policy continues to be followed.

This means that if the current policy is a greedy policy, the distinction between the two disappears.

In the case of questions a) and b) the difference is using the minimum Q-value, which corresponds to the Q-value of actions U or R, in a) and using the Q-value of the next action, which corresponds to the Q-value of action L, in b).