HW4

Eduardo Miranda 95569

Rodrigo Pinto 95666

①

a) Using Q-learning

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

$$Q_{t+1}((E,1,0,1), R) = 2.0 + 0.1\left[0.2 + 0.9 \times 2.0 - 2.0\right] = 2.0$$

$$Q^{(t+1)}_{(E,1,0,1)} = \left[2.8 \quad 2.8 \quad 2.8 \quad 2.8 \quad 2.54 \quad 2.0\right]$$

b) Using SARSA

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

$$Q_{t+1}((E,1,0,1), R) = 2.0 + 0.1\left[0.2 + 0.9 \times 2.8 - 2.0\right] = 2.072$$

$$Q^{(t+1)}_{(E,1,0,1)} = \left[2.8 \quad 2.8 \quad 2.8 \quad 2.8 \quad 2.54 \quad 2.072\right]$$

c) Q-learning updates its values using the optimal policy $\left(\min_{a' \in A} Q_t(x_t, a')\right)$. Since Q-learning doesn't use the same policy as the one used to select actions, we can say that is off-policy.

SARSA updates its values using the policy used to select actions, so its on-policy

As we can see, using Q-learning the chosen action is DOWN and using SARSA the chosen action is RIGHT as chosen by the used policy.