1 Part 1. Q-learning and SARSA

Consider a discounted MDP with $\mathcal{S} = \{\text{A, B, C}\}$ and $\mathcal{A} = \{a, b, c\}$. We plan to use either the Q-learning or the SARSA algorithm in order to learn to control the system. We initialize the estimated Q-function as all zeros – that is:

$$Q^{(0)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \end{array} .$$

The observed trajectory is as follows (for these transitions, we are imposed a policy):

$$(?, ?, ?); (A, ?, ?); (B, a, 100); (A, b, 60); (B, c, 70); (C, b, 40); (A, a, 20); (C, c, \ldots)$$

where each triplet represents the state, the selected action, and the corresponding reward. Some of the information has been corrupted (marked with question marks) in the above sequence.

a) Before the information became corrupt, we ran the Q-learning algorithm and obtained that

$$Q^{(2)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} 11 & 0 & 0 \\ 0 & 0 & 60 \\ 0 & 0 & 0 \end{bmatrix} \end{array} .$$

The discount factor was $\lambda = 0.5$ and the learning rate was fixed to $\alpha = 0.1$. Can you infer what the corrupt information was (i.e., the first state, the first and second selected actions, and the first and second observed rewards? **Answer**:

$(\underline{B}, \underline{c}, \underline{600}); (A, \underline{a}, \underline{80}); (B, a, 100); (A, b, 60); (B, c, 70); (C, b, 40); (A, a, 20); (C, c, \ldots)$

b) Provide the updated Q-values, using the Q-learning algorithm, at the 7th iteration. Use the same values for λ and α as in a). **Answer**:

$$Q^{(7)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} \underline{12{,}127} & \underline{9} & \underline{0} \\ \underline{10{,}55} & \underline{0} & \underline{61} \\ \underline{0} & \underline{4{,}55} & \underline{0} \end{bmatrix} \end{array} .$$

c) What is the greedy policy w.r.t. the estimated Q function at the 7th iteration? $\pi(A) = \underline{a}$, $\pi(B) = \underline{c}$, $\pi(C) = \underline{b}$.

d) Provide the updated Q-values at the 7th iteration using the SARSA algorithm (initialized with $Q^{(0)}$ as all zeros). Take the first two (state, action, reward)-triplets as those given in your answer to a). Let the discount factor be $\lambda = 0.5$ and the learning rate fixed to $\alpha = 0.1$. **Answer**:

$$Q^{(7)} = \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccc} a & b & c \end{array} \\ \begin{bmatrix} \underline{9.2} & \underline{9} & \underline{0} \\ \underline{10} & \underline{0} & \underline{61} \\ \underline{0} & \underline{4{,}4} & \underline{0} \end{bmatrix} \end{array} .$$

e) What is the greedy policy at the 7th iteration? $\pi(A) = \underline{a}$, $\pi(B) = \underline{c}$, $\pi(C) = \underline{b}$.

f) (Tick the correct circle) Are the rewards deterministic? ○ Yes - ☒ No