Northeastern University
College of Engineering

EECE 5698 - ST: Reinforcement Learning Spring 2023

HW4

Problem 1.

Consider the following system with two states $\{A, B\}$ and two possible actions $\{a^1, a^2\}$.

The transition probabilities can be expressed as:

$$P(s'|s,a^1) = \begin{cases} 1 & s=A,\ s'=A \\ 0 & s=A,\ s'=B \\ 0 & s=B,\ s'=A \\ 1 & s=B,\ s'=B \end{cases} \qquad P(s'|s,a^2) = \begin{cases} 0 & s=A,\ s'=A \\ 1 & s=A,\ s'=B \\ 1 & s=B,\ s'=A \\ 0 & s=B,\ s'=B \end{cases}$$

The reward function is as follow: $\begin{cases} 2 & \text{moving to state } s'=B \\ 0 & \text{moving to state } s'=A \\ -1 & \text{taking action } a^2 \\ 0 & \text{taking action } a^1 \end{cases}$

a) Consider the initial Q-value $\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix}$ and $\begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0.5 \\ 0 \end{bmatrix}$ with $\gamma=0.9$ and $\alpha=0.5$. Perform Q-Learning Algorithm with $\varepsilon=0$ (greedy policy), initial state $S_0=A$ for four steps: $(S_0=A, a_0), (S_1, a_1), (S_2, a_2), (S_3, a_3), (S_4, a_4)$. Show all intermediate Q-Values.

b) Compute the final policy $\pi(A)$, and $\pi(B)$, after all transitions in part (a).

c) If you would use SARSA istead of Q-Learning, would the intermediate Q-values and final policy be different from part (a). Justify your answer. (no computations needed in this part).

a)

episode 0

(A) $a^2 = a_0 \sim \lambda(A|\cdot)$ ε-greedy (B)

$$Q(A,a^2) = \underbrace{Q(A,a^2)}_{0.5} + \underbrace{\alpha}_{0.5}\Big[\underbrace{R_{t+1}}_{1} + \underbrace{\gamma}_{0.9}\underbrace{\max Q(B,a)}_{0} - \underbrace{Q(A,a^2)}_{0.5}\Big] = 0.75$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0.75 \\ 0 \end{bmatrix}$$

closest Deterministic policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^2 \quad \pi(B) = a^2$

episode1

(B) $\xrightarrow{a^2 = a_1 \sim \pi(B)^{\varepsilon\text{-greedy}}}$ (A)

$$Q(B,a^2) = \underbrace{Q(B,a^2)}_{0} + \alpha\left[\underbrace{R_{t+1}}_{0.5} + \underbrace{\gamma}_{-1} \underbrace{\max}_{0.9} \underbrace{Q(A,a)}_{0.75} - \underbrace{Q(B,a^2)}_{0}\right] = -0.1625$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0.75 \\ -0.1625 \end{bmatrix}$$

closest Deterministic policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^2 \quad \pi(B) = a^1$

episode2

(A) $\xrightarrow{a^2 = a_2 \sim \pi(A)^{\varepsilon\text{-greedy}}}$ (B)

$$Q(A,a^2) = \underbrace{Q(A,a^2)}_{0.75} + \alpha\left[\underbrace{R_{t+1}}_{0.5} + \underbrace{\gamma}_{1} \underbrace{\max}_{0.9} \underbrace{Q(B,a)}_{-0.1} - \underbrace{Q(A,a^2)}_{0.25}\right] = 0.83$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ -0.1 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0.5 \\ -0.1625 \end{bmatrix}$$

closest Deterministic policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^2 \quad \pi(B) = a^1$

episode 3

(B) $\xrightarrow{a^1 = a_3 \sim \pi(B)^{\varepsilon\text{-greedy}}}$ (B)

$$Q(B,a^1) = \underbrace{Q(B,a^1)}_{-0.1} + \underbrace{\alpha}_{0.5}\Big[\underbrace{R_{t+1}}_{2} + \underbrace{\gamma}_{0.9}\max Q(B,a) - \underbrace{Q(B,a^1)}_{-0.1}\Big] = 0.905$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 0.905 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0.5 \\ -0.1625 \end{bmatrix}$$

closest Deterministic policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^2 \quad \pi(B) = a^1$

episode 4

(B) $\xrightarrow{a^1 = a_4 \sim \pi(B)^{\varepsilon\text{-greedy}}}$ (B)

$$Q(B,a^1) = \underbrace{Q(B,a^1)}_{0.905} + \underbrace{\alpha}_{0.5}\left[\underbrace{R_{t+1}}_{2} + \underbrace{\gamma}_{0.9}\underbrace{\max Q(B,a)}_{0.905} - \underbrace{Q(B,a^1)}_{0.905}\right] = 1.86$$

$$\begin{bmatrix} Q(A,a^1) \\ Q(B,a^1) \end{bmatrix} = \begin{bmatrix} 0 \\ 1.86 \end{bmatrix} \quad \begin{bmatrix} Q(A,a^2) \\ Q(B,a^2) \end{bmatrix} = \begin{bmatrix} 0.85 \\ -0.1625 \end{bmatrix}$$

closest Deterministic policy to $\pi^{\varepsilon\text{-greedy}}$

$\pi(A) = a^2 \quad \pi(B) = a^1$

b) Q-learning

$\pi(A) = \arg\max_{a \in A} Q(A,a) = a^2$

Q-learning

$\pi(B) = \arg\max_{a \in A} Q(B,a) = a^1$

c) It would be same which is because in this question we use $\varepsilon = 0$ (greedy policy) all the time, so the $\max Q(S,a)$ in Q learning is equal to $Q(S,a)$ in the SARSA.

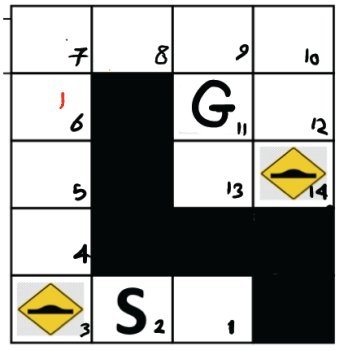
Problem 2.

Consider the following maze problem with 14 states and four actions $A=\{U, R, D, L\}$. The state transitions are deterministic, and reward for taking any action is -1 and moving to the goal state 100 and bump -10.

a) Set all initial Q-values to zero; $Q(s,a)=0$ for all $s \in S$, $a \in A$, $\alpha=0.5$ and $\gamma=0.9$. Run one episode of Q-Learning Algorithm when agent starts from state 5 and follows greedy policy. Note that in the case of equal Q-values, the tie break for actions will be in the following order U, R, D, L. For example

$$\underset{a \in \{U,R,D,L\}}{\operatorname{argmax}} Q(s,a) = \begin{cases} U & \text{if } Q(s,U)=Q(s,R)=Q(s,D)=Q(s,L) \\ R & \text{if } Q(s,R)=Q(s,D)=Q(s,L) > Q(s,U) \\ D & \text{if } Q(s,L)=Q(s,D) > Q(s,U), Q(s,R) \end{cases}$$

b) Show step by step transitions, Q-values and final policy obtained in this single episode.



a)

a) (5) $U=\lambda(5)$ ε-greedy (6)

$Q(5,U) = \underbrace{Q(5,U)}_{0} + \alpha[R + \gamma \max_{a \in A} Q(6,a) - \underbrace{Q(5,U)}_{0}] = -0.5$

$0.5 \quad -1 \quad 0.9 \quad \underbrace{\quad}_{0}$

⑥ $\xrightarrow{U=\lambda(6)}$ (ε-greedy) ⑦

$$Q(6,U) = \underset{0}{Q(6,U)} + \underset{0.5}{\alpha}\left[\underset{-1}{R} + \underset{0.9}{\delta}\max(7,a) - \underset{0}{Q(6,U)}\right] = -0.5$$

⑦ $\xrightarrow{U=\lambda(7)}$ (ε-greedy) ⑦

$$Q(7,U) = \underset{0}{Q(7,U)} + \underset{0.5}{\alpha}\left[\underset{-1}{R} + \underset{0.9}{\delta}\max(7,a) - \underset{0}{Q(7,U)}\right] = -0.5$$

⑦ $\xrightarrow{R=\lambda(7)}$ (ε-greedy) ⑧

$$Q(7,R) = \underset{0}{Q(7,R)} + \alpha\left[\underset{-1}{R} + \underset{0.9}{\delta}\max(8,a) - \underset{0}{Q(7,R)}\right] = -0.5$$

⑧ $\xrightarrow{U=\lambda(8)}$ (ε-greedy) ⑧

$$Q(8,U) = \underset{0}{Q(8,U)} + \alpha\left[\underset{-1}{R} + \underset{0.9}{\delta}\max(8,a) - \underset{0}{Q(8,U)}\right] = -0.5$$

⑧ $\xrightarrow{R=\lambda(8)}$ (ε-greedy) ⑨

$$Q(8,R) = \underset{0}{Q(8,R)} + \alpha\left[\underset{-1}{R} + \underset{0.9}{\delta}\max(9,a) - \underset{0}{Q(8,R)}\right] = -0.5$$

(9) $\xrightarrow{U = \pi(9)}$ (9) ε-greedy

$$Q(9,U) = \underbrace{Q(9,U)}_{0} + \alpha\left[\underbrace{R}_{-1} + \underbrace{\delta}_{0.9}\max(9,a) - \underbrace{Q(9,U)}_{0}\right] = -0.5$$

(9) $\xrightarrow{R = \pi(9)}$ (10) ε-greedy

$$Q(9,R) = \underbrace{Q(7,R)}_{0} + \alpha\left[\underbrace{R}_{-1} + \underbrace{\delta}_{0.9}\max(8,a) - \underbrace{Q(7,R)}_{0}\right] = -0.5$$

(10) $\xrightarrow{U = \pi(10)}$ (10) ε-greedy

$$Q(10,U) = \underbrace{Q(10,U)}_{0} + \alpha\left[\underbrace{R}_{-1} + \underbrace{\delta}_{0.9}\max(10,a) - \underbrace{Q(10,U)}_{0}\right] = -0.5$$

(10) $\xrightarrow{R = \pi(10)}$ (10) ε-greedy

$$Q(10,R) = \underbrace{Q(10,R)}_{0} + \alpha\left[\underbrace{R}_{-1} + \underbrace{\delta}_{0.9}\max(10,a) - \underbrace{Q(10,R)}_{0}\right] = -0.5$$

(10) $\xrightarrow{D = \pi(10)}$ (12) ε-greedy

$Q(10,D) = \underbrace{Q(10,D)}_{0} + \alpha[\underbrace{R}_{-1} + \underbrace{\delta}_{0.9}\max(12,a) - \underbrace{Q(10,D)}_{0}] = -0.5$

(12) $\xrightarrow{U = \pi(12)}$ (10) ε-greedy

$Q(12,U) = \underbrace{Q(12,U)}_{0} + \alpha[\underbrace{R}_{-1} + \underbrace{\delta}_{0.9}\max(10,a) - \underbrace{Q(12,U)}_{0}] = -0.5$

(10) $\xrightarrow{L = \pi(10)}$ (9) ε-greedy

$Q(10,L) = \underbrace{Q(10,L)}_{0} + \alpha[\underbrace{R}_{-1} + \underbrace{\delta}_{0.9}\max(9,a) - \underbrace{Q(10,L)}_{0}] = -0.5$

(9) $\xrightarrow{P = \pi(9)}$ (6) ε-greedy

$Q(9,D) = \underbrace{Q(9,D)}_{0} + \alpha[\underbrace{R}_{49} + \underbrace{\delta}_{0.9}\max(6,a) - \underbrace{Q(9,D)}_{0}] = 49.5$

b)

$\text{argmax}\, Q(1,a) = U$

$\text{argmax}\, Q(2,a) = U$

$\text{argmax}\, Q(3,a) = U$

$\arg\max Q(4,a) = U$

$\arg\max Q(5,a) = R$

$\arg\max Q(6,a) = R$

$\arg\max Q(7,a) = D$

$\arg\max Q(8,a) = D$

$\arg\max Q(9,a) = D$

$\arg\max Q(10,a) = U$

$\arg\max Q(11,a) = U$

$\arg\max Q(12,a) = R$

$\arg\max Q(13,a) = U$

$\arg\max Q(14,a) = U$

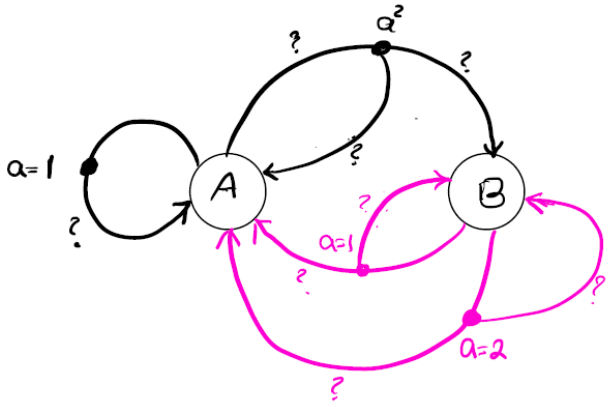
Problem 3.

Consider the following system with two states $\{A, B\}$ and two actions $\{a^1, a^2\}$. The system state transition is unknown and learning should be achieved through interactions. Consider the following state-action-reward obtained through Softmax policy in Actor-Critic algorithm.

$(S_0 = A, a_0 = a^1, r = 10), (S_1 = A, a_1 = a^2, r = -5), (S_2 = B, a_2 = a^1, r = 40),$

$(S_3 = A, a_3 = a^2, r = -5), (S_4 = B, a_4 = a^2, r = 20), (S_4 = A, a_4 = a^1, r = +10), S_5 = A$

Set the initial preferences and state values to zero. Use $\alpha = 0.5$, $\beta = 0.1$ and $\gamma = 0.9$ and show all intermediate preferences, state values and policies.



Questions about the HW should be directed to TA, Begum Taskazan, at taskazan.b@northeastern.edu.

episode 0

$S_0 = A \xrightarrow{a^1 = \pi(A)} A$

$\pi(a^1|A) = \frac{e^{H(A,a^1)}}{e^{H(A,a^1)} + e^{H(A,a^1)}} = \frac{e^0}{e^0 + e^0} = 0.5 \quad a^1$

$\pi(a^2|A) = \frac{e^{H(A,a^2)}}{e^{H(A,a^1)} + e^{H(A,a^1)}} = \frac{e^0}{e^0 + e^0} = 0.5 \quad a^2$

$\delta_t = R_{t+1} + \delta V(S_{t+1}) - V(S_t) = 10 + \delta V(A) - V(A) = 10$

$V(A) = V(A) + \underbrace{\alpha}_{0.5} \delta_t = 5$

$H(A,a^1) = H(A,a^1) + \underbrace{\beta}_{0.1} \times \delta_t (1 - \pi(a^1|A)) = 0.5$

episode 1

$S_1 = A \xrightarrow{a^2 = \pi(A)} B$

$\pi(a^1|A) = \frac{e^{H(A,a^1)}}{e^{H(A,a^1)} + e^{H(A,a^1)}} = \frac{e^{0.5}}{e^0 + e^{0.5}} = 0.623 \quad a^1$

$\pi(a^2|A) = \frac{e^{H(A,a^2)}}{e^{H(A,a^1)} + e^{H(A,a^1)}} = \frac{e^0}{e^0 + e^{0.5}} = 0.377 \quad a^2$

$\delta_t = R_{t+1} + \delta V(S_{t+1}) - V(S_t) = -5 + \delta V(B) - V(A) = -10$

$V(A) = V(A) + \underbrace{\alpha}_{0.5} \delta_t = 0$

$H(A,a^2) = H(A,a^2) + \underbrace{\beta}_{0.1} \times \delta_t (1 - \pi(a^2|A)) = -0.623$

episode 2

$S_2 = B \xrightarrow{a^1 = \pi(B)} A$

$\pi(a^1|B) = \frac{e^{H(B,a^1)}}{e^{H(B,a^1)} + e^{H(B,a^2)}} = \frac{e^0}{e^0 + e^0} = 0.5 \quad a^1$

$\pi(a^2|B) = \frac{e^{H(B,a^2)}}{e^{H(B,a^2)} + e^{H(B,a^1)}} = \frac{e^0}{e^0 + e^0} = 0.5 \quad a^2$

$\delta_t = R_{t+1} + \delta V(S_{t+1}) - V(S_t) = 40 + \underbrace{\delta}_{0.6} V(A) - V(B) = 40$

$$V(B) = V(B) + \underbrace{\alpha}_{0.5}\delta_t = 20$$

$$H(B,a') = H(B,a') + \underbrace{\beta}_{0.1} \times \delta_t (1 - \pi(a'|B)) = 2$$

episode 3

$$S_3 = A \xrightarrow{a^2 = \pi(A)} B$$

$$\pi(a'|A) = \frac{e^{H(A,a')}}{e^{H(A,a')} + e^{H(B,a')}} = \frac{e^{0.5}}{e^{0.5} + e^{0.623}} = 0.755$$

$$\pi(a^2|A) = \frac{e^{H(A,a^2)}}{e^{H(A,a')} + e^{H(B,a')}} = \frac{e^{0.623}}{e^{0.5} + e^{0.623}} = 0.245$$

$$\delta_t = R_{t+1} + \underbrace{\delta}_{0.9} V(S_{t+1}) - V(S_t) = -5 + \underbrace{\delta}_{0.9} V(B) - \underbrace{V(A)}_{0} = 13$$

$$V(A) = V(A) + \underbrace{\alpha}_{0.5}\delta_t = 6.5$$

$$H(A,a^2) = H(A,a^2) + \underbrace{\beta}_{0.1} \times \delta_t (1 - \pi(a^2|A)) = 0.3585$$

episode 4

$$S_4 = B \xrightarrow{a = \pi(B)} A$$

$$\pi(a'|B) = \frac{e^{H(B,a')}}{e^{H(B,a')} + e^{H(B,a')}} = \frac{e^2}{e^0 + e^2} = 0.881\ a'$$

$$\pi(a^2|B) = \frac{e^{H(B,a^2)}}{e^{H(B,a')} + e^{H(B,a')}} = \frac{e^0}{e^2 + e^0} = 0.119\ a^2$$

$$\delta_t = R_{t+1} + \underbrace{\delta}_{0.9} V(S_{t+1}) - V(S_t) = 20 + \delta V(A) - V(B) = 5.85$$

$$V(B) = V(B) + \underbrace{\alpha}_{0.5}\delta_t = 22.925$$

$$H(B,a^2) = H(B,a^2) + \underbrace{\beta}_{0.1} \times \delta_t (1 - \pi(a^2|B)) = 0.515$$

episode 5

$S_5 = A \xrightarrow{a' = \pi(A)} A$

$\pi(a^1|A) = \frac{e^{H(A,a^1)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = \frac{e^{0.5}}{e^{0.5} + e^{0.1375}} = 0.5303 \; a^1$

$\pi(a^2|A) = \frac{e^{H(A,a^2)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = \frac{e^{0.3585}}{e^{0.5} + e^{0.3585}} = 0.4697 \; a^2$

$\delta_t = R_{t+1} + \delta V(S_{t+1}) - V(S_t) = 10 + \delta V(A) - V(A) = 9.35$

$V(A) = V(A) + \underset{0.5}{\alpha} \delta_t = 11.175$

$H(A,a^1) = H(A,a^1) + \underset{0.1}{\beta} \times \delta_t (1 - \pi(a^1|A)) = 0.934$

$H(A,a^1) = 0.934 \qquad H(B,a^1) = 2$

$H(A,a^2) = 0.3585 \qquad H(B,a^2) = 0.515$

$\pi(a^1|A) = \frac{e^{H(A,a^1)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = 0.64$

$\pi(a^2|A) = \frac{e^{H(A,a^2)}}{e^{H(A,a^1)} + e^{H(A,a^2)}} = 0.36$

$\pi(a^1|B) = \frac{e^{H(B,a^1)}}{e^{H(B,a^1)} + e^{H(B,a^2)}} = 0.82$

$\pi(a^2|B) = \frac{e^{H(B,a^2)}}{e^{H(B,a^1)} + e^{H(B,a^2)}} = 0.18$