Northeastern University
College of Engineering

EECE 5698 - ST: Reinforcement Learning Spring 2023

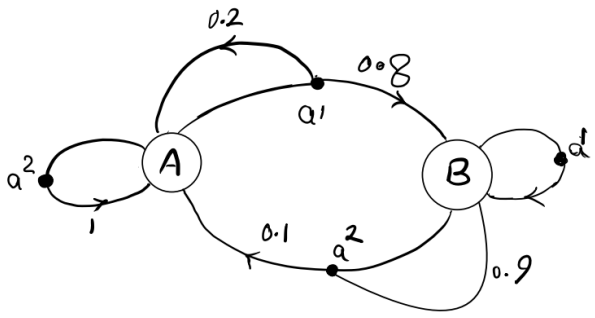
HW3

Problem 1.

consider the following system with the state space $S=\{A,B\}$, and action space $A=\{a^1, a^2\}$. The state transition diagram is shown below, where $P(s'=B \mid s=A, a=a^1)=0.8$, $P(s'=A \mid s=A, a=a^1)=0.2$.

The reward is as follows:

+2	moving to state B
0	moving to state A
-1.5	taking action a^1
-1	taking action a^2



a) Construct transition matrices $M(a^1)$, $M(a^2)$ and compute $R_s^{a^1}$, $R_s^{a^2}$.

b) Perform matrix-form policy Iteration method with initial policy $\pi^1(A)=a^2$, $\pi^1(B)=a^1$ and $\gamma=0.9$ to compute π^*.

a)

$$M(a^1)=\begin{array}{c} \\ A \\ B \end{array}\begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \end{array}$$

$$M(a^2)=\begin{array}{c} \\ A \\ B \end{array}\begin{array}{c} \begin{array}{cc} A & B \end{array} \\ \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \end{array}$$

$$R_{ss'}^{a^1} = \begin{bmatrix} -1.5 & 0.5 \\ -1.5 & 0.5 \end{bmatrix} \qquad R_{ss'}^{a^2} \begin{bmatrix} -1 & 1 \\ -1 & 1 \end{bmatrix}$$

$$R_s^{a^1} = \begin{bmatrix} R(A, a^1) \\ R(B, a^1) \end{bmatrix} = (M(a^1) \odot R_{ss'}^{a^1}) \mathbb{1}_{2\times 1}$$

$$= \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix}$$

$$R_s^{a^2} = \begin{bmatrix} R(A, a^2) \\ R(B, a^2) \end{bmatrix} = (M(a^2) \odot R_{ss'}^{a^2}) \mathbb{1}_{2\times 1}$$

$$= \begin{bmatrix} -1 \\ 0.8 \end{bmatrix}$$

b) $M(\pi^1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \quad R_s^{\pi^0} = \begin{bmatrix} -1 \\ 0.5 \end{bmatrix}$

$$v^{\pi^1} = (I - \gamma M(\pi^1))^{-1} R_s^{\pi^1}$$

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}\right)^{-1} \begin{bmatrix} -1 \\ 0.5 \end{bmatrix}$$

$$= \begin{bmatrix} -10 \\ 5 \end{bmatrix}$$

$$\pi^2 = \underset{a \in A}{\operatorname{argmax}}\, R_S^a + \gamma M(a) V^{\pi^1}$$

$$= \operatorname{argmax}\left\{ R_S^{a^1} + \gamma M(a^1) V^{\pi^1},\ R_S^{a^2} + \gamma M(a^2) V^{\pi^1} \right\}$$

$$= \operatorname{argmax}\left\{ \underbrace{\begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} -10 \\ 5 \end{bmatrix}}_{a^1},\ \underbrace{\begin{bmatrix} -0.1 \\ 0.8 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} -10 \\ 5 \end{bmatrix}}_{a^2} \right\}$$

$$= \begin{Bmatrix} a^1 \\ a^1 \end{Bmatrix}$$

$\pi^2 \neq \pi^1$, so PE continue.

$$M(\pi^2) = \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \quad R_S^{\pi^2} = \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix}$$

$$V^{\pi^2} = (I - \gamma M(\pi^2))^{-1} R_S^{\pi^2}$$

$$= \left(\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \right)^{-1} \cdot \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix}$$

$$= \begin{bmatrix} 4.5 \\ 5 \end{bmatrix}$$

$$\pi^3 = \text{argmax}_a \{ R_S^a + \gamma M(a) V^{\pi^2} \}$$

$$= \{ R_S^{a^1} + \gamma M(a^1) V^{\pi^2},\ R_S^{a^2} + \gamma M(a^2) V^{\pi^2} \}$$

$$= \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 4.5 \\ 5 \end{bmatrix},\ \begin{bmatrix} -1 \\ 0.8 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 4.5 \\ 5 \end{bmatrix} \right\}$$

(a^1, a^2)

$$= \begin{Bmatrix} a^1 \\ a^2 \end{Bmatrix}$$

$\therefore \pi^3 \neq \pi^2$, so PE continue

$$M(\pi^3) = \begin{bmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{bmatrix} \quad R^{\pi^3} = \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix}$$

$$V^{\pi^3} = [I - \gamma M(\pi^3)]^{-1} R_S^{\pi^3}$$

$$= \left[\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} - 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0.1 & 0.9 \end{bmatrix} \right]^{-1} \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix}$$

$$= \begin{bmatrix} 6.54 \\ 7.31 \end{bmatrix}$$

$$\pi^4 = \arg\max \{ R_s^a + \gamma M(a) V^{\pi^3} \}$$

$$= \arg\max \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 6.54 \\ 7.31 \end{bmatrix}, \right.$$

a^1

$$\left. \begin{bmatrix} -1 \\ 0.8 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 6.54 \\ 7.31 \end{bmatrix} \right\}$$

a^2

$$= \begin{Bmatrix} a^1 \\ a^2 \end{Bmatrix}$$

$$\because \pi^4 = \pi^3 \quad \therefore \pi^* = \pi^4 = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$$

Problem 2.

For the system defined in Problem 2, perform matrix-form Value Iteration method with $V_0(s)=0$, $\gamma=0.9$ and $\theta=0.5$ to compute V^* and π^*.

$V_1 = \max R_s^a + \delta M(a) V_0$

$= \max \{ R_s^{a^1} + \delta M(a^1) V_0, R_s^{a^2} + \delta M(a^2) V_0 \}$

$= \max \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0.8 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\}$

$= \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix}$

$\| V_1 - V_0 \|_{\max} = \left\| \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix} - \begin{bmatrix} 0 \\ 0 \end{bmatrix} \right\|_{\max} < 0.5$ is not true

$\therefore$ continue

$V_2 = \max R_s^a + \delta M(a) V_1$

$= \max \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix}, \begin{bmatrix} -1 \\ 0.8 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix} \right\}$

$= \begin{bmatrix} 0.69 \\ 1.46 \end{bmatrix}$

$\| V_2 - V_1 \|_{\max} = \left\| \begin{bmatrix} 0.69 \\ 1.46 \end{bmatrix} - \begin{bmatrix} 0.1 \\ 0.8 \end{bmatrix} \right\| < 0.5$ is not true

$\therefore$ continue

$V_3 = \max R_S^a + \gamma \sum_{s'} P V_2$

$= \max \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 0.69 \\ 1.46 \end{bmatrix}, \begin{bmatrix} -1 \\ 0.2 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 0.69 \\ 1.46 \end{bmatrix} \right\}$

$= \begin{bmatrix} 1.28 \\ 2.05 \end{bmatrix}$

$\| V_4 - V_3 \|_{\max} = \left\| \begin{bmatrix} 1.28 \\ 2.05 \end{bmatrix} - \begin{bmatrix} 0.69 \\ 1.46 \end{bmatrix} \right\|_{\max} < 0.5$ is not true

∴ continue

$V_4 = \max R_S^a + \gamma \sum_{s'} P V_3$

$= \max \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.28 \\ 2.05 \end{bmatrix}, \begin{bmatrix} -1 \\ 0.2 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 1.28 \\ 2.05 \end{bmatrix} \right\}$

$= \begin{bmatrix} 1.81 \\ 2.58 \end{bmatrix}$

$$\|V_4 - V_3\|_{max} = \left\| \begin{bmatrix} 1.81 \\ 2.58 \end{bmatrix} - \begin{bmatrix} 1.28 \\ 2.05 \end{bmatrix} \right\|_{max} < 0.5 \text{ is}$$

not true $\therefore$ continue

$$V_5 = \max R_s^a + \gamma M(a) V_4$$

$$= \max \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 1.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1.81 \\ 2.58 \end{bmatrix}, \begin{bmatrix} -1 \\ 0.8 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 2.1 & 0.9 \end{bmatrix} \begin{bmatrix} 1.81 \\ 2.58 \end{bmatrix} \right\}$$

$$= \begin{bmatrix} 2.28 \\ 3.05 \end{bmatrix}$$

$$\|V_5 - V_4\|_{max} = 0.47 < 0.5 \quad \therefore V^* = V_5 = \begin{bmatrix} 2.28 \\ 3.05 \end{bmatrix}$$

$$\pi^* = \underset{a \in A}{\operatorname{argmax}} \left\{ R_s^{a_1} + \gamma M(a_1) V^*, R_s^{a_2} + \gamma M(a_2) V^* \right\}$$

$$= \arg\max \left\{ \begin{bmatrix} 0.1 \\ 0.5 \end{bmatrix} + 0.9 \begin{bmatrix} 0.2 & 0.8 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 2.2 \\ 3.5 \end{bmatrix}, \begin{bmatrix} -1 \\ 0.8 \end{bmatrix} + 0.9 \begin{bmatrix} 1 & 0 \\ 0.1 & 0.9 \end{bmatrix} \begin{bmatrix} 2.2 \\ 3.5 \end{bmatrix} \right\}$$

a^1 a^2

$$= \begin{Bmatrix} a^1 \\ a^2 \end{Bmatrix}$$

$$\therefore \pi^* = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$$

Problem 3.

Consider an MDP with two states $\{A, B\}$ and two actions $\{a^1, a^2\}$. The system state transitions are governed through the following transition matrices:

$$M(a^1) = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad M(a^2) = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}.$$

The reward is as follows:

$$\begin{cases} +5 & \text{moving to State } B \\ 0 & \text{moving to State } A \\ -1 & \text{taking action } a^2 \\ 0 & \text{taking action } a^1 \end{cases}$$

Consider an initial policy $\pi^0 = \begin{bmatrix} \pi^0(A) \\ \pi^0(B) \end{bmatrix} = \begin{bmatrix} a^1 \\ a^2 \end{bmatrix}$, $\gamma = 0.9$ and episode length 5. Perform Monte Carlo Policy Iteration method to obtain the best policy.

*you need to show all trajectories, the approximation of Q-values and Policy Improvement till the time that Policies in two Consecutive Iterations Stays the same.

episode 1

$A, \pi(A) = a^1, A, r = 0$

$A, \pi(A) = a^1, A, r = 0$

$A, \pi(A) = a^1, A, r = 0$

$A, \pi(A) = a^1, A, r = 0$

$A, \pi(A) = a^1, A, r = 0$

episode 2

$B, \pi(B) = a^2, A, r = -1$

$A, \pi(A) = a^1, A, r = 0$

$A, \pi(A) = a^1, A, r = 0$

$A, \pi(A) = a^1, A, r = 0$

$A, \pi(A) = a^1, A, r = 0$

episode 3

A, $\lambda(A)=a^2$, B, r=4

B, $\lambda(B)=a^2$, A, r=-1

A, $\lambda(A)=a^1$, A, r=0

A, $\lambda(A)=a^1$, A, r=0

A, $\lambda(A)=a^1$, A, r=0

episode 4

B, $\lambda(B)=a^1$, B, r=5

B, $\lambda(B)=a^2$, A, r=-1

A, $\lambda(A)=a^1$, A, r=0

A, $\lambda(A)=a^1$, A, r=0

A, $\lambda(A)=a^1$, A, r=0

$G^1(A,a^1)=0$ $G^2(A,a^1)=0$ $G^3(A,a^1)=0$ $G^4(A,a^1)=0$

$Q^{\pi^0}(A,a^1)=0$

$G^2(B,a^2)=-1$ $G^3(B,a^2)=-1$ $G^4(B,a^2)=-1$

$Q^{\pi^0}(B,a^2)=-1$

$G^3(A,a^2)=4+5(-1)=3\cdot1$

$Q^{\pi^0}(A,a^2)=3\cdot1$

$G^4(B,a^1)=5+6(-1)=4\cdot1$ $Q^{\pi^0}(B,a^1)=4\cdot1$

$$\pi'(s) = \underset{a \in A}{\operatorname{argmax}}\, Q_{\pi^0}(s, a)$$

$$\therefore \pi' = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$$

episode 1

A, $\pi(A) = a^1$, A, $r = 0$

A, $\pi(A) = a^2$, B, $r = 4$

B, $\pi(B) = a^1$, B, $r = 5$

B, $\pi(B) = a^1$, B, $r = 5$

B, $\pi(B) = a^1$, B, $r = 5$

episode 2

B, $\pi(B) = a^2$, A, $r = -1$

A, $\pi(A) = a^2$, B, $r = 4$

B, $\pi(B) = a^1$, B, $r = 5$

B, $\pi(B) = a^1$, B, $r = 5$

B, $\pi(B) = a^1$, B, $r = 5$

episode 3

episode 4

A, $\lambda(A)=a^2$, B, $r=4$ B, $\lambda(B)=a^1$, B, $r=5$

B, $\lambda(B)=a^1$, B, $r=5$ B, $\lambda(B)=a^1$, B, $r=5$

B, $\lambda(B)=a^1$, B, $r=5$ B, $\lambda(B)=a^1$, B, $r=5$

B, $\lambda(B)=a^1$, B, $r=5$ B, $\lambda(B)=a^1$, B, $r=5$

B, $\lambda(B)=a^1$, B, $r=5$ B, $\lambda(B)=a^1$, B, $r=5$

$G^1(A, a^1)=0+5\cdot4+5^2\cdot5+5^3\cdot5+5^4\cdot5=14.58$

$Q(A, a^1)=14.58$

$G^2(B, a^2)=-1+5\cdot4+5^2\cdot5+5^3\cdot5+5^4\cdot5=13.58$

$Q(B, a^2)=13.58$

$G^1(A, a^2)=4+5\cdot5+5^2\cdot5+5^3\cdot5=16.195$

$G^2(A, a^2)=4+5\cdot5+5^2\cdot5+5^3\cdot5=16.195$

$$G^3(A,a^2) = 4 + 6 \cdot 5 + 6^2 \cdot 5 + 6^3 \cdot 5 + 6^4 \cdot 5 = 19.48$$

$$Q(A,a^2) = \frac{G^1(A,a^2) + G^2(A,a^2) + G^3(A,a^2)}{3} = 17.29$$

$$G^1(B,a') = 5 + 5 \cdot 5 + 6^2 \cdot 5 = 13.55$$

$$G^2(B,a') = 5 + 6 \cdot 5 + 6^2 \cdot 5 = 13.55$$

$$G^3(B,a') = 5 + 6 \cdot 5 + 6^2 \cdot 5 + 6^3 \cdot 5 = 17.195$$

$$G^4(B,a') = 5 + 6 \cdot 5 + 6^2 \cdot 5 + 6^3 \cdot 5 + 6^4 \cdot 5 = 20.4755$$

$$Q(B,a') = \frac{G^1(B,a') + G^2(B,a') + G^3(B,a') + G^4(B,a')}{4}$$

$$= 16.19$$

$$\pi^2 = \arg\max Q(s,a)$$

$\therefore \pi^2 = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$

$\sim\!\!\!\sim \pi^2 = \pi^1$

$\therefore$ the optimal policy is equal to $\pi^2 = \begin{bmatrix} a^2 \\ a^1 \end{bmatrix}$