INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

Stamp / Signature of the Invigilator

EXAMINATION (Mid Semester)	**SEMESTER (Autumn 2023-2024)**

Roll Number										**Section**		**Name**	

Subject Number	C	S	6	0	0	7	7	**Subject Name**	*REINFORCEMENT LEARNING*

Department / Center of the Student		**Additional sheets**	

Important Instructions and Guidelines for Students

1. You must occupy your seat as per the Examination Schedule/Sitting Plan.
2. Do not keep mobile phones or any similar electronic gadgets with you even in the switched off mode.
3. Loose papers, class notes, books or any such materials must not be in your possession, even if they are irrelevant to the subject you are taking examination.
4. Data book, codes, graph papers, relevant standard tables/charts or any other materials are allowed only when instructed by the paper-setter.
5. Use of instrument box, pencil box and non-programmable calculator is allowed during the examination. However, exchange of these items or any other papers (including question papers) is not permitted.
6. Write on both sides of the answer script and do not tear off any page. **Use last page(s) of the answer script for rough work.** Report to the invigilator if the answer script has torn or distorted page(s).
7. It is your responsibility to ensure that you have signed the Attendance Sheet. Keep your Admit Card/Identity Card on the desk for checking by the invigilator.
8. You may leave the examination hall for wash room or for drinking water for a very short period. Record your absence from the Examination Hall in the register provided. Smoking and the consumption of any kind of beverages are strictly prohibited inside the Examination Hall.
9. Do not leave the Examination Hall without submitting your answer script to the invigilator. **In any case, you are not allowed to take away the answer script with you.** After the completion of the examination, do not leave the seat until the invigilators collect all the answer scripts.
10. During the examination, either inside or outside the Examination Hall, gathering information from any kind of sources or exchanging information with others or any such attempt will be treated as **'unfair means'**. Do not adopt unfair means and do not indulge in unseemly behavior.

Violation of any of the above instructions may lead to severe punishment.

Signature of the Student

To be filled in by the examiner

Question Number	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Marks Obtained											

Marks obtained (in words)	**Signature of the Examiner**	**Signature of the Scrutineer**

Indian Institute of Technology Kharagpur
Department of Computer Science and Engineering

Mid-Semester Exam | **Reinforcement Learning (CS60077)** | **Autumn 2023-2024**

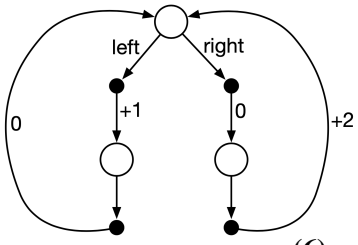
Date: 26-Sep-2023 (FN) | **Answer *all* questions.** | **Maximum Marks:** 60

— Write your answers at indicated places inside the question paper. —

— This page is kept blank intentionally. —

— The question paper starts from the next page. —

Q1. [MDP and Optimality of Policy] **18 marks**



(a) Consider the continuing MDP shown to the right. The only decision to be made is that in the *top* state (say, s_0), where two actions are available, `left` and `right`. The numbers show the rewards that are received deterministically after each action. There are exactly two deterministic policies, $\pi_{\texttt{left}}$ and $\pi_{\texttt{right}}$. Calculate and show which policy will be the optimal:

(i) if $\gamma = 0$; (ii) if $\gamma = 0.9$; (iii) if $\gamma = 0.5$. **(6)**

Answer:

In an infinite-horizon MDP, the expected return from state s following a deterministic policy π is given as,

$$\mathbb{E}_\pi\big[G_t \mid s_t = s\big] \;=\; R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \;=\; \sum_{k=0}^{\infty} \gamma^k R_{t+k}$$

Here, for two deterministic policy $\pi_{\texttt{left}}$ and $\pi_{\texttt{right}}$, the expected returns from state s_0 are:

$$\mathbb{E}_{\pi_{\texttt{left}}}\big[G_0 \mid s_0\big] \;=\; 1 + \gamma\cdot 0 + \gamma^2\cdot 1 + \gamma^3\cdot 0 + \cdots \;=\; \sum_{k=0}^{\infty} \gamma^{2k}\big(1 + \gamma\cdot 0\big) \;=\; \frac{1}{1-\gamma^2}$$

$$\mathbb{E}_{\pi_{\texttt{right}}}\big[G_0 \mid s_0\big] \;=\; 0 + \gamma\cdot 2 + \gamma^2\cdot 0 + \gamma^3\cdot 2 + \cdots \;=\; \sum_{k=0}^{\infty} \gamma^{2k}\big(0 + \gamma\cdot 2\big) \;=\; \frac{2\gamma}{1-\gamma^2}$$

Putting the values of γ, we get

$$\mathbb{E}_{\pi_{\texttt{left}}}\big[G_0 \mid s_0\big] = \begin{cases} 1, & \text{if } \gamma = 0 \\ 5.26, & \text{if } \gamma = 0.9 \\ 1.33, & \text{if } \gamma = 0.5 \end{cases} \quad \text{and} \quad \mathbb{E}_{\pi_{\texttt{right}}}\big[G_0 \mid s_0\big] = \begin{cases} 0, & \text{if } \gamma = 0 \\ 9.47, & \text{if } \gamma = 0.9 \\ 1.33, & \text{if } \gamma = 0.5 \end{cases}$$

Hence, the optimal policies are:

(i) $\pi_{\texttt{left}}$ for $\gamma = 0$,

(ii) $\pi_{\texttt{right}}$ for $\gamma = 0.9$,

(iii) Both $\pi_{\texttt{left}}$ and $\pi_{\texttt{right}}$ for $\gamma = 0.5$.

(b) Consider an MDP with an infinite set of states $S = \{1, 2, 3, \ldots\}$. The start state is $s = 1$. Each state $s \in S$ allows a continuous set of actions $a \in [0, 1]$. The transition probabilities are given by:

$$\mathbb{P}[s+1 \mid s, a] = a, \quad \mathbb{P}[s \mid s, a] = 1 - a; \qquad \forall\, s \in S,\ \forall\, a \in [0, 1]$$

For all states $s \in S$ and actions $a \in [0, 1]$, transitioning from s to $s+1$ results in a reward of $1+a$; and transitioning from s to s results in a reward of $1-a$. The discount factor $\gamma = 0.5$. Calculate the optimal value function, $V^*(s)$, and the optimal deterministic policy, $\pi^*(s)$, for all $s \in S$. **(4)**

Answer:

Since this is an infinite horizon MDP and since each state has identical state transition probabilities and identical reward function, each state would have the same value for the optimal state-value function. Let us refer to this common value as V^*.

Recall the Bellman optimality equation where we have:

$$v^*(s) = \max_{a \in A} \Big[R(s,a) + \gamma \sum_{s' \in S} \mathbb{P}[s' \mid s, a]\, v^*(s') \Big] = \max_{a \in A} \Big[\sum_{s' \in S} \mathbb{P}[s' \mid s, a] \Big(R(s,a,s') + \gamma v^*(s') \Big) \Big]$$

Here, $V^* = \max_{a \in [0,1]} \Big[a\Big((1+a) + 0.5V^*\Big) + (1-a)\Big((1-a) + 0.5V^*\Big) \Big]$

Moving V^* from the RHS to the LHS, we get:

$$\begin{aligned} V^* - 0.5V^* &= \max_{a \in [0,1]} \left[2a^2 - a + 1 \right] \\ \implies V^* &= \max_{a \in [0,1]} \left[4a^2 - 2a + 2 \right] \end{aligned}$$

For $a \in [0, 1]$, the RHS maximizes for $a = 1$. So, the optimal policy is $\pi^*(s) = 1$ for all states $s \in S$. Substituting for $a = 1$, the optimal value function is given by, $V^* = 4 \times 1^2 - 2 \times 1 + 1 = 4$.

(c) State whether the following statement is *True* or *False* with a brief justification. **(2)**

For every finite MDP, $M = \langle S, A, R, \gamma \rangle$, with bounded rewards ($R$) and $\gamma \in [0, 1)$, for all policies π, we have: $\max_{a \in A} q_\pi(s,a) \geq v_\pi(s), \quad \forall s \in S.$

Answer: True

Because, $v_\pi(s) = \sum_{a \in A} \pi(a|s) q_\pi(s,a)$ combined with the fact that, *maximum* is always better than *expectation*.

(d) State whether the following statement is *True* or *False* with proper justification. **(3)**

If the only difference between two MDPs is the value of the discount factor then they must have the same optimal policy.

Answer: False

A counterexample suffices to show the statement is *false*. Consider an MDP with two sink states. Transitioning into sink state A gives a reward of 1, transitioning into sink state B gives a reward of 10. All other transitions have 0 (zero) rewards. Let A be one step North from the start state. Let B be two steps South from the start state. Assume actions always succeed. Then if the discount factor $\gamma < 0.1$ the optimal policy takes the agent one step North from the start state into A, if the discount factor $\gamma > 0.1$ the optimal policy takes the agent two steps South from the start state into B.

(e) Consider an infinite-horizon, discounted MDP $M = \langle S, A, R, P, \gamma \rangle$. Define the maximal reward, $R_{MAX} = \max_{(s,a) \in S \times A} R_s^a$ and for any policy $\pi : S \to A$, show that $V^\pi(s) \leq \frac{R_{MAX}}{1-\gamma}$ $(\forall s \in S)$. **(3)**

Answer:

$$
\begin{aligned}
V^\pi(s) &= \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_{s_t}^{a_t} \;\middle|\; s_0 = s\right] \\
&\leq \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t R_{MAX} \;\middle|\; s_0 = s\right] \\
&= R_{MAX} \cdot \mathbb{E}_\pi\left[\sum_{t=0}^{\infty} \gamma^t \;\middle|\; s_0 = s\right] \\
&= \frac{R_{MAX}}{1-\gamma}
\end{aligned}
$$

Q2. [Value Iteration] **15 marks**

Consider a simple MDP with 3 states s_1, s_2, s_3 and 2 actions a_1, a_2. The transition probabilities and expected rewards are given in the following table. Assume discount factor $\gamma = 1$.

(from) State	Action	Reward	Transition Probability (to State)		
			s_1	s_2	s_3
s_1	a_1	8	0.2	0.6	0.2
	a_2	10	0.1	0.2	0.7
s_2	a_1	1	0.3	0.3	0.4
	a_2	-1	0.5	0.3	0.2
s_3	a_1	0	0	0	1.0
	a_2	0	0	0	1.0

Your task is to determine an optimal deterministic policy by manually working out simply the first two iterations of *value iteration* algorithm.

Initialize the value function for each state to be it's max (over actions) reward, i.e., we initialize the value function to be $v_0(s_1) = 10,\ v_0(s_2) = 1,\ v_0(s_3) = 0$. Then answer the following:

(a) Considering all states and actions, calculate $q_k(\cdot,\cdot)$ and $v_k(\cdot)$ from $v_{k-1}(\cdot)$ using the value iteration update, and then calculate the greedy policy $\pi_k(\cdot)$ from $q_k(\cdot,\cdot)$ for 2 iterations (i.e. for $k = 1$ and $k = 2$). **(10)**

Answer:

Value iteration algorithm follows Bellman's optimality equation for iterative updates:

$$q_k(s,a) = R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_{k-1}(s')$$

$$v_k(s) = \max_{a \in A} q_k(s,a) = \max_{a \in A} \left[R_s^a + \gamma \sum_{s' \in S} P_{ss'}^a v_{k-1}(s') \right]$$

Following this, we present the updates for two value iterations as follows:

For $k = 1$,

$$q_1(s_1,a_1) = 8 + 0.2 \times 10.0 + 0.6 \times 1.0 + 0.2 \times 0 = 10.6$$
$$q_1(s_1,a_2) = 10 + 0.1 \times 10.0 + 0.2 \times 1.0 + 0.7 \times 0 = 11.2$$
$$\therefore\ v_1(s_1) = \max\left[10.6, 11.2\right] = 11.2 \quad \text{and} \quad \pi_1(s_1) = a_2$$

$$q_1(s_2,a_1) = 1 + 0.3 \times 10.0 + 0.3 \times 1.0 + 0.4 \times 0 = 4.3$$
$$q_1(s_2,a_2) = -1 + 0.5 \times 10.0 + 0.3 \times 1.0 + 0.2 \times 0 = 4.3$$
$$\therefore\ v_1(s_2) = \max\left[4.3, 4.3\right] = 4.3 \quad \text{and} \quad \pi_1(s_2) = a_1 \text{ or } a_2$$

For $k = 2$,

$$\begin{aligned} q_2(s_1,a_1) &= 8 + 0.2 \times 11.2 + 0.6 \times 4.3 + 0.2 \times 0 = 12.82 \\ q_2(s_1,a_2) &= 10 + 0.1 \times 11.2 + 0.2 \times 4.3 + 0.7 \times 0 = 11.98 \\ \therefore\ v_2(s_1) &= \max\left[12.82, 11.98\right] = 12.82 \quad \text{and} \quad \pi_2(s_1) = a_1 \end{aligned}$$

$$\begin{aligned} q_2(s_2,a_1) &= 1 + 0.3 \times 11.2 + 0.3 \times 4.3 + 0.4 \times 0 = 5.65 \\ q_2(s_2,a_2) &= -1 + 0.5 \times 11.2 + 0.3 \times 4.3 + 0.2 \times 0 = 5.89 \\ \therefore\ v_2(s_2) &= \max\left[5.65, 5.89\right] = 5.89 \quad \text{and} \quad \pi_2(s_2) = a_2 \end{aligned}$$

(b) Mathematically argue / show that, $\pi_k(\cdot)$ for $k > 2$ will be the same as $\pi_2(\cdot)$.
(**Hint:** You can make the argument by examining the structure of how you get $q_k(\cdot,\cdot)$ from $v_{k-1}(\cdot)$. With such argument, there is *no need to go beyond the two iterations* you performed above, and so you can proclaim $\pi_2(\cdot)$ as an optimal deterministic policy for this MDP.) **(5)**
Answer:

$$\begin{aligned} q_k(s_1,a_1) - q_k(s_1,a_2) &= (8-10) + (0.2-0.1)v_{k-1}(s_1) + (0.6-0.2)v_{k-1}(s_2) + (0.2-0.7)v_{k-1}(s_3) \\ &= -2 + 0.1v_{k-1}(s_1) + 0.4v_{k-1}(s_2) + 0, \quad \text{for all } k \geq 1 \\ q_k(s_2,a_1) - q_k(s_2,a_2) &= (1-(-1)) + (0.3-0.5)v_{k-1}(s_1) + (0.3-0.3)v_{k-1}(s_2) + (0.4-0.2)v_{k-1}(s_3) \\ &= 2 - 0.2v_{k-1}(s_1) + 0 + 0, \quad \text{for all } k \geq 1 \end{aligned}$$

Since $v_{k-1}(s_1) \geq 12.82$ and $v_{k-1}(s_2) \geq 5.89$ for all $k \geq 3$, we see that:

$$\begin{aligned} q_k(s_1,a_1) - q_k(s_1,a_2) &\geq -2.0 + 0.1 \times 12.82 + 0.4 \times 5.89 > 0, \quad \text{for all } k \geq 3 \\ q_k(s_2,a_1) - q_k(s_2,a_2) &\geq 2.0 - 0.2 \times 12.82 < 0, \quad \text{for all } k \geq 3 \end{aligned}$$

So, we find that, $q_k(s_1,a_1) > q_k(s_1,a_2)$ and $q_k(s_2,a_1) < q_k(s_2,a_1)$, for all $k \geq 3$.
Hence, $\pi_k(s_1) = a_1$ and $\pi_k(s_2) = a_2$ for all $k \geq 3$.

Q3. [Monte-Carlo and Temporal Difference Learning] **12 marks**

Assume you have data in the form of just the following 5 complete episodes for an MRP.

- *Episode 1:* `A 2 A 6 B 1 B 0 T`
- *Episode 2:* `A 3 B 2 A 4 B 2 B 0 T`
- *Episode 3:* `B 3 B 6 A 1 B 0 T`
- *Episode 4:* `A 0 B 2 A 4 B 4 B 2 B 0 T`
- *Episode 5:* `B 8 B 0 T`

In the episodic data presented above, the non-terminal states are labeled `A` and `B`, the numbers denote *rewards*, and all states end in a terminal state `T`. Assume discount factor $\gamma = 1$. Answer the following:

(a) Given only this data and experience replay (repeatedly and endlessly drawing an episode at random from this pool of 5 episodes), calculate the value function estimates, i.e. $V(\texttt{A})$ and $V(\texttt{B})$, that (a) First-Visit Monte-Carlo and (b) Every-Visit Monte-Carlo converge to. **(4)**

Answer:

First-Visit Monte Carlo averages the returns starting from the first occurrence of each of the states across all episodes. Therefore, the First-Visit Monte Carlo Value Function estimate (with experience replay) would converge to:

$$V(\texttt{A}) = \frac{(2+6+1+0)+(3+2+4+2+0)+(1+0)+(0+2+4+4+2+0)}{4} = \frac{33}{4} = 8.25$$

$$V(\texttt{B}) = \frac{(1+0)+(2+4+2+0)+(3+6+1+0)+(2+4+4+2+0)+(8+0)}{5} = \frac{39}{5} = 7.8$$

Every-Visit Monte Carlo averages the returns starting from each occurrence of each of the states across all episodes. Therefore, the Every-Visit Monte Carlo Value Function estimate (with experience replay) would converge to:

$$V(\texttt{A}) = \frac{[(2+6+1+0)+(6+1+0)]+[(3+2+4+2+0)+(4+2+0)]+[(1+0)]+[(0+2+4+4+2+0)+(4+4+2+0)]}{2+2+1+2}$$

$$= \frac{(9+7)+(11+6)+(1)+(12+10)}{7} = \frac{56}{7} = 8$$

$$V(\texttt{B}) = \frac{[(1+0)+(0)]+[(2+4+2+0)+(2+0)+(0)]+[(3+6+1+0)+(6+1+0)+(0)]+[(2+4+4+2+0)+(4+2+0)+(2+0)+(0)]+[(8+0)+(0)]}{2+3+3+4+2}$$

$$= \frac{(1+0)+(8+2+0)+(10+7+0)+(12+6+2+0)+(8+0)}{14} = \frac{56}{14} = 4$$

(b) Construct an MRP that TD(0) (i.e., one-step TD) essentially builds with transition probabilities and reward function estimated, from the one-step transitions and sample rewards seen in the episodic data presented above. **(4)**

Answer:

TD(0) essentially constructs an MDP with transition probabilities and reward function estimated from the one-step transitions and sample rewards seen in the data, and its Value Function estimate is the Value Function of that estimated MDP.

The transition probabilities would be estimated as:

$$\mathbb{P}[\mathtt{A} \to \mathtt{A}] = \frac{1}{7} \qquad \mathbb{P}[\mathtt{A} \to \mathtt{B}] = \frac{6}{7} \qquad \mathbb{P}[\mathtt{A} \to \mathtt{T}] = 0$$

$$\mathbb{P}[\mathtt{B} \to \mathtt{A}] = \frac{3}{14} \qquad \mathbb{P}[\mathtt{B} \to \mathtt{B}] = \frac{6}{14} \qquad \mathbb{P}[\mathtt{B} \to \mathtt{T}] = \frac{5}{14}$$

The reward function would be estimated as:

$$R(\mathtt{A}) = \frac{(2+6)+(3+4)+(1)+(0+4)}{7} = \frac{20}{7} = 2.857$$

$$R(\mathtt{B}) = \frac{(1+0)+(2+2+0)+(3+6+0)+(2+4+2+0)+(8+0)}{14} = \frac{15}{7} = 2.143$$

(c) From the constructed MRP in Part (b), what will be the value function estimates, i.e. $V(\mathtt{A})$ and $V(\mathtt{B})$, that TD(0) converge to? Show your calculations. **(4)**

Answer:

The MRP with these transition probabilities and reward function leads to the following Bellman Equations:

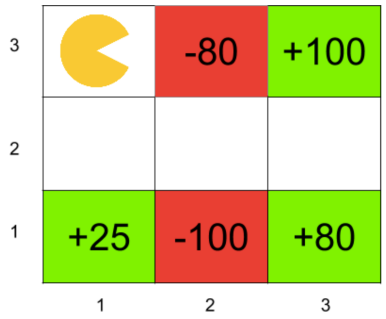
$$V(\mathtt{A}) = \frac{20}{7} + \frac{1}{7}V(\mathtt{A}) + \frac{6}{7}V(\mathtt{B}) \implies 3V(\mathtt{A}) - 3V(\mathtt{B}) = 10$$

$$V(\mathtt{B}) = \frac{15}{7} + \frac{3}{14}V(\mathtt{A}) + \frac{6}{14}V(\mathtt{B}) \implies -3V(\mathtt{A}) + 8V(\mathtt{B}) = 30$$

This yields: $V(\mathtt{A}) = \frac{34}{3} = 11.33$ and $V(\mathtt{B}) = 8.$

Q4. [Q-Learning with Function Approximation] **15 marks**

Consider the grid-world given (right) and Pacman who is trying to learn the optimal policy. If an action results in landing into one of the shaded states the corresponding reward is awarded during that transition. All shaded states are terminal states, i.e., the MDP terminates once arrived in a shaded state. The other states have the *North, East, South, West* actions available, which deterministically move Pacman to the corresponding neighboring state (or have Pacman stay in place if the action tries to move out of the grid). Assume the discount factor $\gamma = 0.5$ and the Q-learning rate $\alpha = 0.5$ for all calculations. Pacman starts in state $(1,3)$ as shown in the figure. Answer the following:



(a) Calculate the optimal value function V^* for the following states: $V^*(2,2)$ and $V^*(1,3)$. **(3)**

Answer:

$V^*(2,2) = 50$ and $V^*(1,3) = 12.5$

The optimal values for the states can be found by computing the expected reward for the agent acting optimally from that state onwards. Note that you get a reward when you transition into the shaded states and not out of them. So for example the optimal path starting from $(2,2)$ is to go to the $+100$ square which has a discounted reward of $0+\gamma\times 100 = 50$. For $(1,3)$, going to either of $+25$ or $+100$ has the same discounted reward of $0+\gamma\times 25 = 12.5$ or $0+\gamma\times 0+\gamma^2\times 0+\gamma^3\times 25 = 12.5$.

(b) The agent starts from the top left corner and you are given the following episodes from runs of the agent through this grid-world. Each line in an Episode is a tuple containing (s,a,s',r).

Episode 1	*Episode 2*	*Episode 3*
(1,3), S, (1,2), 0	(1,3), S, (1,2), 0	(1,3), S, (1,2), 0
(1,2), E, (2,2), 0	(1,2), E, (2,2), 0	(1,2), E, (2,2), 0
(2,2), S, (2,1), -100	(2,2), E, (3,2), 0	(2,2), E, (3,2), 0
	(3,2), N, (3,3), +100	(3,2), S, (3,1), +80

Using Q-Learning updates, calculate the following Q-values for $Q((3,2),N)$, $Q((3,2),S)$ and $Q((2,2),E)$ after the above three episodes. **(4.5)**

Answer:

Q-values obtained by Q-learning updates: $Q(s,a) \leftarrow (1-\alpha)Q(s,a) + \alpha\left(R(s,a,s') + \gamma \max_{a' \in A} Q(s',a')\right)$

So, we have,

$$Q((3,2),N) = 0.5 \times \underbrace{Q((3,2),N)}_{\text{since init val = 0}}^{0} + 0.5 \times \left(100 + 0.5 \times \underbrace{\max_{a'} Q((3,3),a')}_{\text{since no trans. from (3,3)}}^{0}\right) = 50$$

$$Q((3,2),S) = 0.5 \times \underbrace{Q((3,2),S)}_{\text{since init val = 0}}^{0} + 0.5 \times \left(80 + 0.5 \times \underbrace{\max_{a'} Q((3,1),a')}_{\text{since no trans. from (3,1)}}^{0}\right) = 40$$

$$Q((2,2),E) = 0.5 \times \underbrace{Q((2,2),E)}_{\text{since init val = 0}}^{0} + 0.5 \times \left(0 + 0.5 \times \max[\underbrace{Q((3,2),N)}_{=50}, \underbrace{Q((3,2),S)}_{=40}]\right) = 12.5$$

(c) Consider a feature based representation of the Q-value function:

$$Q_f(s,a) = w_1 f_1(s) + w_2 f_2(s) + w_3 f_3(a)$$

$f_1(s)$: The x coordinate of the state $\quad$ $f_2(s)$: The y coordinate of the state

$f_3(N) = 1, \quad f_3(S) = 2, \quad f_3(E) = 3, \quad f_3(W) = 4$

(i) Given that all w_i are initially 0, calculate the values of these weights after *Episode 1*. **(4.5)**

Answer:

Using the approximate Q-learning weight updates:

$$w_i \leftarrow w_i + \alpha\Big[\big(R(s,a,s') + \gamma \max_{a' \in A} Q_f(s',a')\big) - Q_f(s,a)\Big]\nabla_{w_i} Q_f(s,a)$$

The only time the reward is non-zero in the first episode is when it transitions into -100 state.

$$\begin{aligned}
\therefore\ w_1 &= \cancel{w_1}^{0} + 0.5 \times \Big[\ \big(-100 + 0.5 \times \max_{a' \in A} \cancel{Q_f(s',a')}^{0}\big) - \cancel{Q_f(s,a)}^{0}\Big] \cdot f_1((2,2)) \\
&= 0.5 \times -100 \times 2 = -100 \\
w_2 &= \cancel{w_2}^{0} + 0.5 \times \Big[\ \big(-100 + 0.5 \times \max_{a' \in A} \cancel{Q_f(s',a')}^{0}\big) - \cancel{Q_f(s,a)}^{0}\Big] \cdot f_2((2,2)) \\
&= 0.5 \times -100 \times 2 = -100 \\
w_3 &= \cancel{w_3}^{0} + 0.5 \times \Big[\ \big(-100 + 0.5 \times \max_{a' \in A} \cancel{Q_f(s',a')}^{0}\big) - \cancel{Q_f(s,a)}^{0}\Big] \cdot f_3(S) \\
&= 0.5 \times -100 \times 2 = -100
\end{aligned}$$

(ii) Assume the weight vector **w** is equal to $(1,1,1)$. What is the action prescribed by the Q-function in state $(2,2)$? State the calculations to reason the same. **(3)**

Answer: West

The action prescribed at $(2,2)$ is $\arg\max_{a \in A} Q_f((2,2),a)$ where,

$$Q_f((2,2),a) = w_1 f_1((2,2)) + w_2 f_2((2,2)) + w_3 f_3(a) = 2 + 2 + f_3(a)$$

is computed using the feature representation. In this case, the Q-value for *West* gives maximum value with $Q_f((2,2),West) = 2+2+4 = 8$.

— The question paper ends here. —