INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

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)**
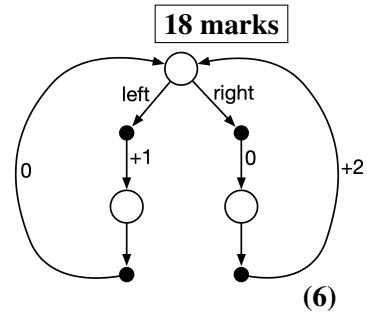
Answer:

(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:

(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: $\quad \max\limits_{a \in A} q_\pi(s,a) \geq v_\pi(s), \quad \forall s \in S.$

Answer:

(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:

(e) Consider an infinite-horizon, discounted MDP $M = \langle S, A, R, P, \gamma \rangle$. Define the maximal reward, $R_{MAX} = \max\limits_{(s,a) \in S \times A} R_s^a$ 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:

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:

(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:

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. $V(A)$ and $V(B)$, that (a) First-Visit Monte-Carlo and (b) Every-Visit Monte-Carlo converge to.　　**(4)**

Answer:

(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)**
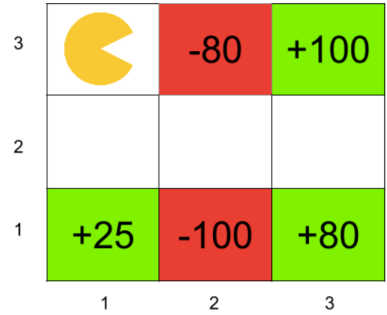
Answer:

(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:

Q4. **[Q-Learning with Function Approximation]** $\boxed{\text{15 marks}}$

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:

	1	2	3
3	◖	-80	+100
2			
1	+25	-100	+80

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

Answer:

(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:

(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 $f_2(s)$: 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:

(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: