Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2022-2023
Mid-Semester Test (EC-2 Regular)

Course No. : AIMLCZG512 **Course Title** :Deep Reinforcement Learning
Nature of Exam : Open Book **Weightage** : 30% **No. of Pages** = 2 ; **No. Of Questions =** 4;
Duration : 2 Hours; **Date of Exam** : 23-07-2023 (FN)

Note to Students:

1. Answer all the questions. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
2. Write answers neatly in A4 papers, scan and upload.
3. Write your name and sign at the end of all the pages.
4. Assumptions made if any, should be stated clearly at the beginning of your answer.
5. Refer to the Honor Code published at the course page.

Q1) Answer the following questions. [DRL Exam Question 1 ; 22-July-2023]

(a) Consider the problem of learning optimal values that control traffic signals at a zebra crossing. The amount of time pedestrians wait for the signal (red), and the amount of time they get while crossing the road (green) are adequate to control the traffic. Provide a MDP formulation to this scenario with all the necessary details. Explain **all** the design choices. [3.5 M]

(b) Write the bellman state-value equation for your MDP in (a). Show the corresponding back-up diagram. [2.0 M]

(c) For the given formulation, which of the approaches to solving MDP is appropriate. Explain. [1.0 M]

(d) Explain what happens if all the rewards in (a) are multiplied by 10? [1.0 M]

Q2) Considering Netflix's personalised video recommendation for an individual, where Netflix suggests a genre from the available list of genres ($\in$ {a, b, c, d}) and observes the reward (satisfactory rating) $\in$ {0, 1, 2, 3} from the individual, let's analyse the following sequence of tried genres and rewards: **1**: (a,2), **2**: (b,3), **3**: (c,1), **4**: (a,1), **5:** (b,1), **6**: (d,2), **7**: (b,1), **8**: (d,1), **9**: (c,3), **10**: (c,3), **11**: (c,1), **12**: (d,1). *Assuming the initial value of all the actions are the same as the last digit of your student id divided by 2*, please provide answers to the following questions in this specific context. [DRL Exam Question 2 ; 22-July-2023]

(a) (Stationary case) Which of those suggestions are exploratory in nature assuming ε-greedy action selection and action values are obtained by sample averages ? Why? [1.5 M]

(b) (Non-Stationary case)Which of those suggestions are exploratory in nature assuming ε-greedy action selection and action values are obtained by weighted averages with α = 0.5? Why? [1.5 M]

(c) Compare your answers to (a) and (b) and explain the difference, ignoring the assumption on stationarity. [1.5 M]

(d) If we initialise all the action values to be 5, will the exploratory moves due to ε-greedy behave differently? Explain. [1.5 M]

(e) Assuming there are **n**-individuals whose likings on movies are different. How would you extend the solution based on armed bandits in this case? [1.5 M]

Q3) Consider a robot navigating in an environment consisting of six positions,numbered from 0 to 5.

0	1	2	3	4	5

At each location, the robot has two actions available: moving left (⇦) and moving right (⇒). These actions are selected with equal probability. States 0 and 5 are considered terminal states, meaning no further actions are permitted from there. When the agent transitions into state 5, it receives a reward of +x and when the agent transitions into state 0, it receives a reward of -x. *x is determined by the last digit of your student ID plus 5*. For all other transitions, the agent receives a reward of 0.

Answer the following questions in this context. Use in-place version (asynchronous dynamic programming) of the algorithms. The discount factor, γ is 0.5. [DRL Exam Question 3 ; 22-July-2023]

(a) Consider the values initialised to 0's. Evaluate the following policy. Use your answer to improve the policy (i.e. one iteration). Show all the steps. [3 M]

	⇦	⇦	⇦	⇒	

(b) Consider the values initialised to 0's. Show two iterations of value iteration. Show all the steps. [3 M]

(c) Comment on the applicability of both the algorithms (a) & (b) for problems with large state and action spaces in no more than 3 well articulated statements. [1.5 M]

Q4) Considering Netflix's personalised video recommendation for an individual, where Netflix suggests a genre from the available list of genres (∈ {a, b, c }) and observes the reward (satisfactory rating) ∈ {0, 1, 2, 3} from three individuals { p, q, r }. It is desired to learn the most suitable recommendations for the individuals from the following episodes {E1 … E3} of the trial run:

E1: p, a, 0 ; q, c, 2 ; p, b, 0 ; q, c, 3 ; r, c, 2; r, b, 2 ; q, c, 1;
E2: q, b, 3 ; r, a, 2 ; q, c, 1 ; p, a, 0 ; q, a, 2; p, c, 2 ; r, b, 1;
E2: r, a, 2 ; p, b, 1 ; p, a, 2 ; q, c, 3 ; q, b, 3; r, a, 3 ; q, c, 0;

Assume the initial policy to be π (p) = a ; π (q) = a; π (r) = c; the initial Q(s, a) for all (s,a) are 0's; and the initial V(s) for all s to be 0's [DRL Exam Question 4 ; 22-July-2023]

(a) Compute first visit and every visit prediction for all the states. Which of these values converge faster to $V^*_\pi(s)$? Explain [3 M]

(b) Generate a random episode and find out the important sampling ratio for E2. Explain how the importance sampling ratio and model dynamics (when available) are related. [2 M]

(c) Explain two approaches discussed in the class that help an agent learning policy using monte-carlo methods continue exploring. [1.5 M]

(d) Is it mandatory for the off-policy learning method to have non-zero probabilities for the choice of all actions ? Explain. [1.0 M]