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 : Closed Book **Weightage** : 30% **No. of Pages** = 2 ; **No. Of Questions =** 5;
Duration : 2 Hours; **Date of Exam**: 27-01-2024 (FN)

Note to Students:

1. Answer all the questions.
2. Write your name and sign at the end of all the pages.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Q1) [Write the answer in an A4 Sheet, Scan & Upload] Consider tic-tac-toe to answer this question. Assume that states are numbered from S_1 to S_n.

(a) List the four elements of reinforcement learning and write one well-articulated formal statement explaining the role of each element. [2 M]

(b) Write the temporal difference rule for learning each state's value. [0.5 M]. Explain various elements and the workings of this rule. [1.5 M]

(c) Let the value of the current state be 4.5, and all its possible successor/predecessor states have a value of 2.7. Use 0.9 to be the parameter value for any parameter you need to use to solve this problem. Given this, Revise the estimate of the value of the current state using your answer to (b). Explain your answer [2 M]

[2.0 + 2.0+ 2.0 = 6 Marks]

Q2) [Write the answer in A4 Sheet, Scan & Upload]

(a) Write any one use-case (application) which can be modeled as a multi-armed bandit for its solution. Your answer should have all the elements that identify the use case and can be modeled as a multi-armed bandit problem. [2 M]

(b) Write down a bandit **algorithm** that is ε-greedy. [Note: only a formal algorithm pseudo-code will be accepted as an answer.] [2 M]. Comment on the efficiency of this algorithm. [0.5 M].

(c) How does this algorithm balance exploration vs exploitation trade-off? Explain. [1.5 M].

[2.0 + 2.5+ 1.5 = 6 Marks]

Q3) [Write the answer in A4 Sheet, Scan & Upload]

(a) Devise two example tasks of your own that fit into the MDP framework, identifying for each its states, actions, and rewards. Make the examples as different from each other as possible. [2 M]

(b) In the following two cases you will find the reward sequences (Starting from R1) is given. You should find the Return (G_0 and G_1) in each case. Assume the $\gamma = 0.8$ in both cases.

(i) An episodic task with T=5. The rewards received are $R_1 = 1$, $R_2 = 2$, $R_3 = 6$, $R_4 = 3$, and $R_5 = 2$ [1.5 M]

(ii) A continuing task with $R_1 = 2$, followed by an infinite sequence of 0's. [1.5 M]

(c) Write down the complete expression for either $v_\pi(s)$ or $q_\pi(s,a)$. [1.0M]

[2.0 + 3.0+ 1.0 = 6 Marks]

Q4) [Type in the space provided]

(a) What do you perform in Policy Evaluation and Policy Improvement? How are they useful in estimating optimal policy? [Answer must be phrased using formal, unambiguous statement] [2 M]

(b) Provide a high-level algorithm for policy iteration. Use the computation of either action or state values. Make necessary assumptions. [2 M]

(c) Explain the characteristics of reinforcement learning problems for which a dynamic programming solution is appropriate. Provide any two examples of problems. [2 M]

[2.0 + 2.0+ 2.0 = 6 Marks]

Q5) [Write the answer in A4 Sheet, Scan & Upload]

(a) Assume you have data in the form of just the following 5 complete episodes. Non-terminal States are labeled A and B, the numbers in the episodes denote Rewards, and all states end in a terminal state T.

- A 2 A 6 B 1 B 0 T
- A 3 B 2 A 4 B 2 B 0 T
- A 0 B 2 A 4 B 4 B 2 B 0 T
- B 8 B 0 T

Estimate the value of states of A & B. Show the steps. [2 M]

(b) What techniques monte-carlo algorithms use to ensure the exploration vs. exploitation trade-off is balanced. Write one line for each of the techniques stating their use. [2 M]

(c) How are ε-soft and ε-greedy policies are related? A small example (a state with 4 actions, with one of the actions being greedy) illustrates the working of ε-greedy action selection. Assume suitable values for each action to help your explanation. [2 M]