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

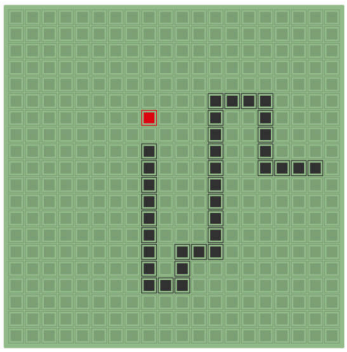
Course No. : AIMLCZG512 **Course Title** :Deep Reinforcement Learning
Nature of Exam : Open Book **Weightage** : 30% **No. of Pages** = 3 ; **No. Of Questions =** 4;
Duration : 2 Hours; **Date of Exam** : 06-08-2023 (FN)

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 ; 06-Aug-2023]

Consider the famous Snake game. The player here is an autonomous agent. The player controls the game with four controls - up, down,left, right. There will be food anywhere on the screen.The food disappears in 10 secs. The player has to control the snake in a way it eats food and he has to avoid the frame's borders otherwise the snake dies.Each time the snake eats food, the body length of the snake is increased. During the game the player has also to avoid the snake's body otherwise the game terminates, as well. The rewards are Eat food = +10, Game over = -10, else = 0. Make the required assumptions and state them clearly. [DRL Exam Question 3 ; 06-Aug-2023]



(i) Write down an MDP formulation for the given scenario with an explanation for all design choices. [3 M]
(ii) After some time you find the game not progressing with the rewards. What can go wrong? [1 M]
(iii) Your initial reward is 0 followed by infinite 10's. What is the initial and the next expected return ? Let the discount rate be 0.4 [2 M]
(iv) Suppose you treated this as an episodic task but also used discounting, with all rewards zero except for 1 upon failure. What then would the return be at each time? How does this return differ from that in the discounted, continuing formulation of this task? [1.5 M]

Q2. Consider the stock investment scenario, where individual has to invest in various stocks ($\in$ {p,q,r,s}) and observes the reward ($\in$ [-5, 5]). The following sequence of investments and rewards were observed over a period of time: **1**: (p,-2), **2**: (q,3), **3**: (r,-1), **4**: (r,-4), **5:** (s,3), **6**: (s,4), **7**: (p,2), **8**: (q,-1), **9**: (r,3), **10**: (r,3), **11**: (s,0), **12**: (q,-1). *Assuming the initial value of all the*

actions are the same as the last digit of your student id divided by 2, provide answers to the following questions in this specific context. [DRL Exam Question 2 ; 06-Aug-2023]

(i) (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 $\alpha = 0.5$? Why? [1.5 M]

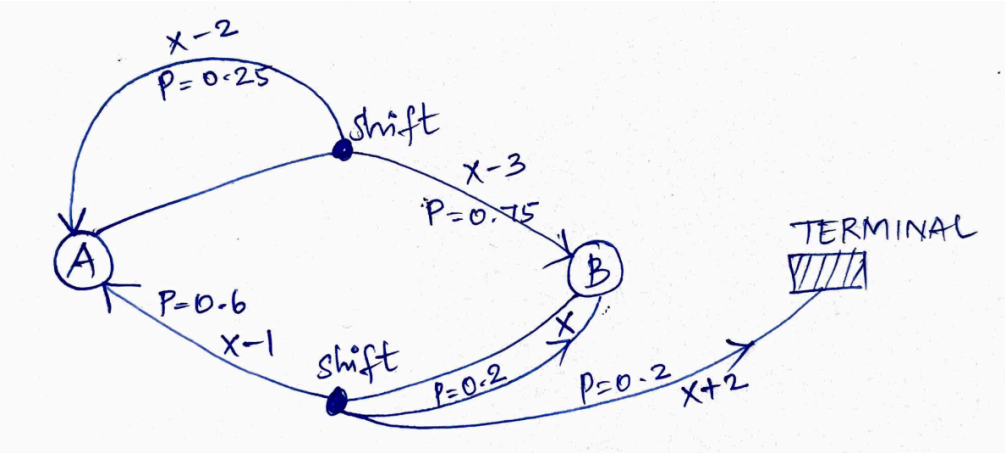
(ii) How do you make an action-selection among the non-greedy actions but still try to maximize the returns ? *Let the value of degree of exploration be the last digit of your student ID. If it is 0, let the value be 1.* [2.5 M]

(iii) Say the individual gets a preference for each stock in the order {2,1,5,6}. How would you create a differentiable policy towards the action choice? [2.5 M]

(iv) How would you map the scenario to an associative setting ? [1 M]

Q3. Answer the following questions:

Consider the MDP given below containing 2 states A and B with action *Shift* that may result in A,B or terminal state.. The rewards obtained are as indicated along the edges in the figure (X-2,X-3,X-1,X,X+2). *X is the last digit of your student ID.* The transition probabilities are as given along the edges. Let the discount factor be 0.4. [DRL Exam Question 3 ; 06-Aug-2023]



(i) Evaluate the given deterministic policy π_1(A)=Right ,π_1(B)=Right and improve it upto 1 iterations. Use Dynamic Programming solution to MDP [3 M]

(ii) Using value iteration of dynamic programming, determine the values of states A and B. Let the values of A and B be initialized to 1. Show 1 Iterations. [3 M]

(iii) With the final improved policy obtained in (i), find the values of state A and B. [1 M]

(iv) With the final improved policy obtained in (i), find the action value of A,*Right* [0.5 M]

Q4. Consider a board game with two states A and B, where two actions are possible from each state : Up and Down. Given below are some sample episodes from their game.

E1 = A,up,0; B,up,2; A,down,1; B,up,3; A,up,0; B,down,2
E2 = A,down,1; A,up,2; B,down,3; A,up,2; B,down,1; A,up,2

Let b(s|a) = {up(0.46), down(0.54)} and π(s|a) = {up(0.97), down(0.03)}; $\gamma = 0.3$ [DRL Exam Question 4 ; 06-Aug-2023]

(i) Estimate the values of states A and B using first-visit and every visit. [1.5 M]
(ii) Determine the estimates : $V_{\pi}(B)$, $V_{\pi}(A)$ using ordinary importance sampling (and first-visit) [2 M]
(iii) Improve and estimate the policy π such that $\pi \cong \pi^{*}$. Use the given b(s|a) and E1. Let the initial Q(s, a) for all (s,a) are 0's; and the initial V(s) for all s to be 0's. [4 M]