Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
Second Semester 2022-2023
Comprehensive Test (EC-3 Regular)

Course No. : AIMLCZG512 **Course Title** :Deep Reinforcement Learning
Nature of Exam : Open Book **Weightage** : 40% **No. of Pages** = 3 ; **No. Of Questions =** 4;
Duration : 2 Hours; **Date of Exam** : 15-10-2023 (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 all the answers neatly in A4 papers, scan and upload them.**
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) **[5+5=10 M]**

(a) In a 4-arm bandit problem, after executing 100 iterations of the UCB algorithm, the estimates of Q values are- $Q_{50}(1) = 13.3$, $Q_{50}(2) = 13.8$, $Q_{50}(3) = 12.98$, $Q_{50}(4) = 12.85$ and the number of times each of them are sampled are $n_1 = 15$, $n_2 = 10$, $n_3 = 20$, $n_4 = 5$. How do you decide on the arm to be sampled in the subsequent trial**[1 M]**? Find out the next arm that will get sampled next with the necessary details **[1 M]**. Compute the next three arms sampled If the rewards received in the next three trials for each arm are as below. Show all the details**[3 M]**.

Trials	**Arm #1**	**Arm #2**	**Arm #3**	**Arm #4**
i	0	0	20	0
i+1	10	20	10	10
i+2	0	0	0	10

(b) Consider the following grid world Markov decision problem, along with the transition probabilities.

Reward 0	Reward 0	Reward +5 (Terminal State)
Start	Reward 0	Reward -5 (Non-Terminal State)

Action	Transition Probability (x = last digit of your student id)
Up (↑)	(x+2) / 10 { let this be a}
Right(→)	1 - (a /2)
Left(←)	1 - (a /2)

Note that if the agent tries to move past the wall, the agent will remain in the same state. What could be an ideal policy **[1 M]**.? Use the bellman optimality equation for to compute values for two iterations assuming the values are initizlized to 0's to begin with and the discount factor is 0.9 **[3.0 M]**. How to you go about learning the optimal actions if the transition probabilities are not given **[1 M]**.?

Q2) **[3.5+0.5+4+2=10M]**

a) Consider the following grid world:

S_0	S_1	S_2	$S_3 = 10$

The states are S_0, S_1,S_2,S_3 with S_3 being the terminal state with utility value of 10. Move Left and Move right are the two actions possible with a transition probability of 0.2 for left and 0.8 for right respectively. You get a reward of -1 for states and actions except for the terminal which has a utility value of 10. Let the discount factor be 0.25

Using value iteration of dynamic programming, determine the values of all states. Let the values of all states be initialised to 0. Show upto 2 Iterations.**[3.5 M]**

(b) What would be the optimal policy as the result of value iteration for this given MDP? **[0.5 M]**

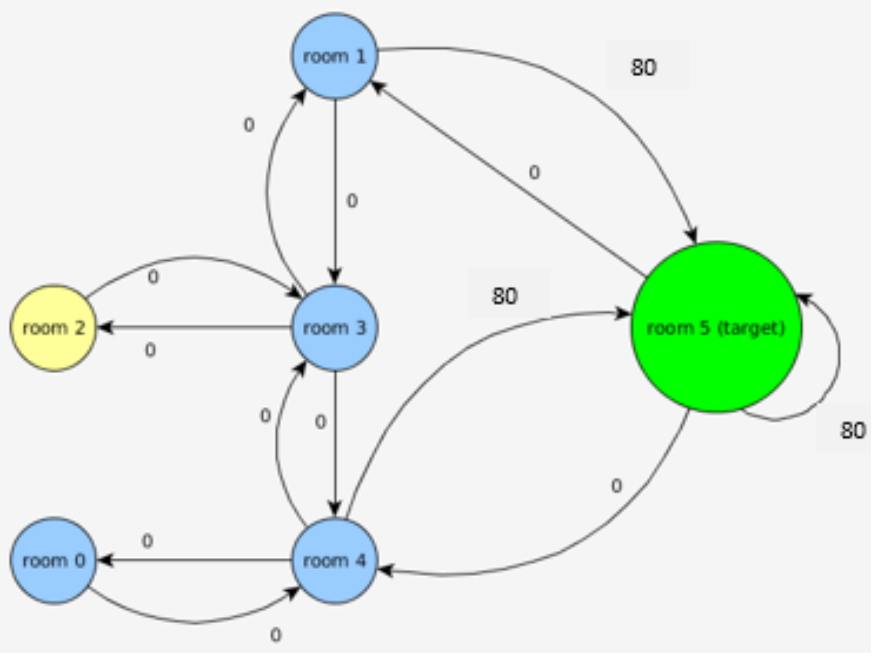
(c) Let b(s|a) = {left(0.46), right(0.54)} and π(s|a) = {right(0.97), left(0.03)}; τ = A,up,0; B,up,2; A,down,1; B,up,3; A,up,0; B,down,2

Improve and estimate the policy π such that $\pi \cong \pi^*$. Use the given b(s|a) and τ. 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]**

(d) In the given episode τ what are the estimates of A and B through every-visit and fist-visit methods? **[2M]**

Q3) **[4+0.5+1+1.5+3=10M]**

Consider the given scenario : Consider a robot that needs to learn how to leave a house in the best path possible. We have a house with 5 rooms, and one "exit" room. A graph representing it is given below. On this graph all rooms are nodes, and the arrows the actions that can be taken on each node. The arrow values are the immediate rewards that the agent receives by taking some action on a specific room. We choose our reinforcement learning environment to give 0 reward for all rooms that are not the exit room. In our target room we give a 100 reward. Let the discount factor be 0.7 and the learning rate be 0.4. An episode starts with a random start node and ends upon reaching the target room.



(a) Consider you are starting from room 2. Using Q learning , find out the best path to exit the house.**[4 M]**.

(b) Also show the updated Q-table.**[0.5 M]** .

(c) Is there any other alternative path other than your answer when starting from room 2? If so, what makes your path a better one ? **[1 M]**

(d) In the above scenario, let's assume there are four actions - up,down,left,right. Consider the following observations in the order (s, a, s' , R(s, a, s')) - (Room 4, up,Room 3, +2), (Room 3, left,Room 2, +4), (Room 3, up,Room 1, +2) , (Room 1, right, Room 5, +4). What is the learned value of Room 3 after these set of observations ? Let the discount factor be 0.7 and the learning rate be 0.4. **[1.5M]**

(e) Consider that you are training a DQN agent to play a simple video game. During training, the following (refer to table below) Q-values are observed while training for 5 different episodes. Do you find the Q values converging **[1 M]**? How do you explain the agent's behaviour for the epsilon 0.1 **[1 M]**? What problems with behaviour cloning that experience replay technique attempt to solve **[1 M]** ?

Episode	Steps	Q-value (State-Action Pair)
1	100	0.5
2	200	0.7
3	300	1.2
4	400	1.6
5	500	1.8

Q4) Answer the following questions: **[2+1.5+2+3+1.5=10M]**

(a) Your state space is high-dimensional and you will run out of memory during computation. You are in need of an algorithm that will definitely reach a local optimum. Would you choose DQN or REINFORCE ? Justify your answer. **[2M]**

(b) Reason for the statement - " Introducing baseline in REINFORCE does not affect the validity of the algorithm ". **[1.5M]**

(c) Can you categorise policy iteration as an actor-critic reinforcement algorithm ? Justify your answer. **[2 M]**

(d) Consider the grid world MDP given in the question paper or use any version of grid world MDP that you are comfortable with. Provide a formulation of this problem in a way that we can perform imitation learning using DAGGER and demonstrate how DAGGER solves this problem using 2 iterations. **[3 M]**

(e) How is exploration and exploitation tradeoff balanced in Monte Carlo Tree search if action with best Q value is chosen at each branch? **[1.5 M]**
