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

Course No. : AIMLCZG512 **Course Title** :Deep Reinforcement Learning
Nature of Exam : Open Book **Weightage** : 40% **No. of Pages** = 3 ; **No. Of Questions =** 5;
Duration : 2.5 Hours / 150 Mins **Date of Exam** : 13-04-2024 (EN)

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. Assumptions made, if any, should be stated clearly at the beginning of your answer.

Q1) [Answer the following questions in the same order. Long and vague answers, penalized]

(a) Consider that you design an RL agent that learns only by playing against itself (i.e., using legal moves). Will the agent learn to excel at playing the game? [One/Two Statements] [1 M]. What is one most critical aspect in the design of such an agent? [Answer just in one precise statement] [1 M]. Why do you think so? [Answer just in one precise statement] [0.5 M].

(b) Consider learning the values of states in a tic-tac-toe game. Two students (Ram & Raghav) from the DRL course are working together on a solution. Ram argues that there are sets of symmetric positions, and it's unnecessary to learn separate values for all of them. Learning value for any one of the states covers all the symmetric positions. For example, symmetric positions are:



Raghav says it is better to learn values of each state independently. If you have to support Ram, what will be your explanation?[One/Two Statements] [1 M]. If you have to support Raghav, what will be your explanation?[One/Two Statements] [1 M]. Propose a better solution with a reason can be a better solution than Ram's and Raghav's if you are asked to provide one? Explain if your solution is better than theirs [One/Two Statements] [0.5 M].

(c) Consider an agent who learns to trade. The agent only considers two buying shares of companies, X and Y. The agent understands two kinds of days (A and B). On days of A-kind, the true values of A and B are 66.66 and 200, with probabilities 0.75 and 0.25, respectively. On days of B-kind, the true values of A and B are 200 and 87.5, with probabilities 0.2 and 0.8, respectively. How would you design your agent if the kind of day is inferred at the beginning of the day and not inferred? Show the steps. [3 M].

[2.5+2.5 + 3 = 8 M]

crackbitswilp.in

Q2) [Answer the following questions in the same order. Long and vague answers, penalized]

(a) Consider an agent that attempts to continuously inspect a closed environment safely. Ram and Raghav work to arrive at a reward model for this MDP. Ram proposes -20 and +20, respectively, for bumping against the wall and safely exploring. Raghav proposes a change in the rewards such that subtracts 10 from both positive and negative rewards (i.e., -30 and 10 instead of -20 and +20). They have their arguments for the choice of reward suggestions. Explain mathematically the difference between Ram's and Raghav's proposals [3 M]. Will there be differences in their proposals if the task is episodic with γ=1? [1 M].

(b) Consider the following four episodes of agent taking actions (left,right) until it reaches the terminal states:

Episode #1: x, left, 16, x, right, 12, x,left, 16, T, 100
Episode #2: x,left, 16, x, left, 16, x, right, 12,T, 200
Episode #3: x, right, 15, x, right, 11, x,left, 13, T, 150
Episode #4: x, right, 12, x,left, 16, x,left, 16, x, left, 16, x, 110

Use first visit MC to estimate values of x and y [2 M]. Suggest two ways to encourage exploration [No innovation here, only two from the classroom discussions accepted.] and explain [One/Two Statements] how they encourage exploration. [2 M].

[The question asks to estimate x and y, but the episodes have only x, not y. So accept the answers only for x]

[4+4= 8 Marks]

Q3) [Answer the following questions in the same order. Long and vague answers, penalized]

(a) Consider an environment with states x and T (where T is terminal state) , two actions 'left' and 'right'. TD algorithm generates the following episode:

x, left, 16, x, right, 12, x,left, 16, x, left, 16, T, 100

[**Note: the format is Format - state, action, reward, next state, action, reward.....**]

Policy π is given by: π(left|x) = 0.9, π(right|x) = 0.1;
Current q- values of q are: q(x, left) = 1 and q(x, right) = 2.
Discount factor, γ, is 1 2 .
Step size, α, is 0.1

Compute using 1-step SARSA [2 M], 2-step SARSA [1 M] and 2-step TD [2 M] the values of q(x,left) and q(x,right) after the first update. Present your answers neatly with all the computations.

(b) Can we use 'tiling' to construct features for three-dimensional representational space where the importance of dimensions is not equal? Why / Why not? [One/Two Statements] [1 M] Suggest two other feature construction techniques that work here [0.5 + 0.5 M].

(c) In the class, we have discussed that Q-Learning is an off-policy method. Explain in 2-3 statements for your reasons for the same. You can begin your answer by reproducing the expression from the book. [1 M]

[5+2+1= 8 Marks]

crackbitswilp.in

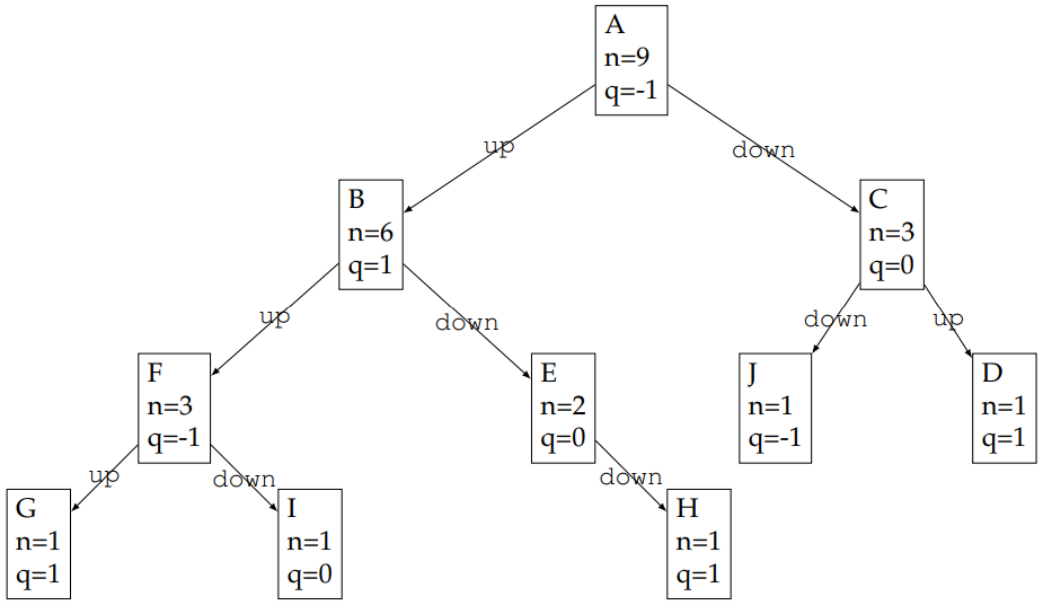
Q4) [Answer the following questions in the same order. Long and vague answers, penalized]

(a) In training DQN, the collected episodes are not trained from start to end. What is the reasoning behind this [One/Two Statements] [1.0 M]. Will the replay buffer used for training DQN ever overflow? Explain the rationale for the design of experience replay buffer [One/Two Statements] [1.0 M].

(b) Explain DDQN's proposed change over DQN [One/Two Statements] [2 M].

(c) Why should the action chosen by Monte Carlo Tree Search tend to be better than the action the underlying rollout policy would choose? [One/Two Statements] [2 M].

(d) Assume you are using MCTS to decide the next move for the two-player game, with possible actions being up or down. Your present state is A, and you must decide between making an up or down move. Based on the iterations, understand the tree as a 2-player game tree. Let q be the estimated value and n be the number of visits to that state so far. Assume the moves are equi-probable if required. Explain how MCTS selects the next state for expansion [4.0 M].



[2 + 2+2+ 4 = 10 Marks]

Q5) [Answer the following questions in the same order. Long and vague answers, penalized]

(a) How do we compute policy gradients by finite difference [One/Two Statements] [1.0 M]? How are policy gradients computed for larger problems [One/Two Statements]? Write an example from classroom coverage on this. [2.0 M].

(b) Select any classic machine learning problems (classification, regression, similarity learning, etc.) and attempt modeling the solution using inverse reinforcement learning. Show the steps. [3.0 M].

[3+3= 6 Marks]

crackbitswilp.in