Roll No:.....

JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION- June 2023

M.Tech. CSE/IT 2nd Semester

COURSE CODE: 22M1WCI235 MAX. MARKS: 35

COURSE NAME: REINFORCEMENT LEARNING

COURSE CREDITS: 03 MAX. TIME: 2Hr

COURSE COORRDINATOR: Prof. (Dr.) Vivek Kumar Sehgal

Note: *All questions are compulsory. Carrying of mobile phone during examinations will be treated as case of unfair means.*

1. Consider two MDPs that are identical, except for their initial state distributions, d_0. Let π^* and μ^* be optimal policies for the first and second MDP, respectively. Let $s^* \in \mathcal{S}$ be a state that has a non-zero probability of occurring when using π^* on the first MDP and a non-zero probability of occurring when using μ^* on the second MDP. Consider a new policy, π' such that $\pi'(s,a) = \pi^*(s,a)$ for all $s \in \mathcal{S} \setminus \{s^*\}$ and $a \in \mathcal{A}$ and $\pi'(s^*,a) = \mu^*(s^*,a)$ for all $a \in \mathcal{A}$. Is π' an optimal policy for the first $MDPQ$ **[CO-1, 2 -5]**

2. Consider an MDP with one state, $S = \{1\}$ and $\mathcal{A} = \mathbb{R}$. Let

$$R_t = \begin{cases} A_t & \text{if } A_t < 1 \\ 0 & \text{otherwise.} \end{cases}$$

Let $\gamma < 1$. In this case, what is the optimal policy? **[CO-2, - 5]**

3. Can you derive the Bellman equation for $q^\pi(s,a)$ from the definition of q^π ? **[CO-3, - 5]**

4. You are given an environment with 1 state, x, and 2 actions, b and c. T is the terminal state. Your TD algorithm generates the following episode using the policy π when interacting with its environment:

Timestep	Reward	State	Action
0		x	b
1	16	x	c
2	12	x	b
3	16	T	

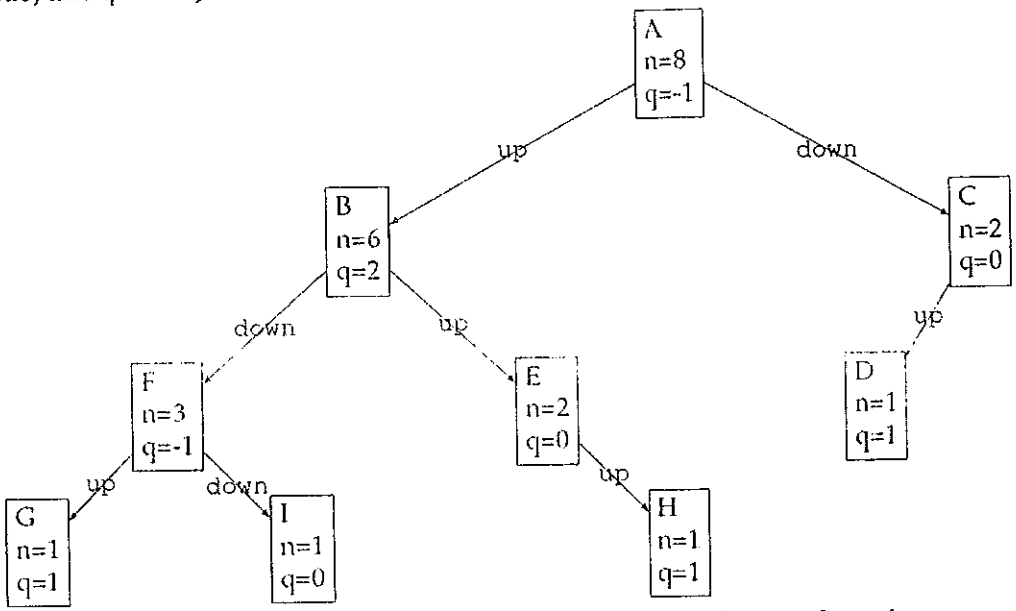
- The policy π is given by: $\pi(b \mid x) = 0.9, \pi(c \mid x) = 0.1$
- The current values of q are: $q(x,b) = 1$ and $q(x,c) = 2$.
- the discount factor, γ, is $\frac{1}{2}$.
- the step size, α, is 0.1

Show the values of $q(x, b)$ and $q(x, c)$ after their first update using 1-step Sarsa, 2step Sarsa, 2-step Expected Sarsa, and 2-step Tree Backup. Note: you should update $q(x, b)$ and $q(x, c)$ only once per learning algorithm. Show your work and carry out your calculations to two decimal places.

Learning Algorithm	$q(x, b)$ after its first update	$q(x, c)$ after its first update
1-step Sarsa	2.6	3.13
2-step Sarsa	3.13	3.8
2-step Expected Sarsa	3.13	3.8
2-step Tree Backup	2.61	3.73

[CO-4.5, - 12]

5. You are using Monte Carlo Tree Search to decide on the next action for a two-person competitive game with 2 actions at each state (up and down). It is player 1 's turn to play in state A. The state of the tree so far is as follows (each node consists of state identifier, n value, and q value):



Remember that the formula for the UCT value for a node, v, is:

$$UCT(v) = \frac{q(v)}{n(v)} + c\sqrt{\frac{\ln n(v \cdot \text{parent})}{n(v)}}$$

Assume the constant c in the UCT formula is 0.5 .

i. What is the node that is next selected (show your work)?

ii. Assuming that the simulation (rollout) from the expanded node gives a value of 1 (that is, player 1 wins), backup that value to all of the affected nodes.

[CO-5, - 8]