JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -3 EXAMINATION- 2024

M.Tech- (CSE/IT)

COURSE CODE (CREDITS): 22M1WCI235 MAX. MARKS: 35

COURSE NAME: REINFORCEMENT LEARNING

COURSE INSTRUCTORS:DHA MAX. TIME: 2 Hours

***Note: (a)** All questions are compulsory.*

(b)Marks are indicated against each question in square brackets.

(c) The candidate is allowed to make Suitable numeric assumptions wherever required for solving problems

---

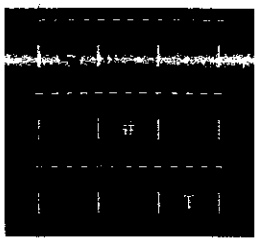
Q1. a) How does Temporal Difference (TD) learning estimate the value function?
b) In TD prediction, what is the role of the temporal difference error?
c) Explain TD(λ). **[CO-3, Marks:3+3+3]**

Q2. a) How would you define SARSA and Q-learning in the context of reinforcement learning?
b) What is the primary objective of Q-learning algorithm? **[CO-3, Marks: 3+3+3]**
c) Explain how SARSA algorithm aims to improve the policy of an agent.

Q3. a) Provide a real-world application where TD learning can be successfully applied?Justify
b) Explain Off Policy and On Policy Learning?
c) How does TD learning differ from SARSA algorithm?

[CO- 3, Marks: 3+3+3]

Q4. Suppose we have a 3x3 grid world where the agent can move in four directions: up, down, left, and right. The goal of the agent is to reach the terminal state (denoted by 'T') from the start state (denoted by 'S') while avoiding obstacles. The grid world looks like this: **[CO- 4, Marks: 8]**



'S' represents the start state, '#' represents an obstacle, and 'T' represents the terminal state. The agent receives a reward of +10 when reaching the terminal state and a reward of -1 for each step taken. The discount factor, γ, is 0.9. use Q-learning to find the optimal policy for the agent to navigate through this grid world. Initialize the Q-values arbitrarily. Use the following Q-learning update rule:

$Q(s, a) = Q(s, a) + \alpha * [r + \gamma * \max(Q(s', a')) - Q(s, a)]$

Where: α (alpha) is the learning rate, r is the reward received for taking action 'a' from state 's', s' is the next state after taking action 'a' from state 's', γ (gamma) is the discount factor, max(Q(s', a')) is the maximum Q-value over all possible actions 'a' in state 's'.

Assume a learning rate (α) of 0.1 and perform Q-learning till 3 iterations. Make assumptions where ever necessary.