JAYPEE UNIVERSITY OF INFORMATION TECHNOLOGY, WAKNAGHAT

TEST -2 EXAMINATION- APRIL-2024

COURSE CODE(CREDITS): 22M1WCI235 MAX. MARKS: 25

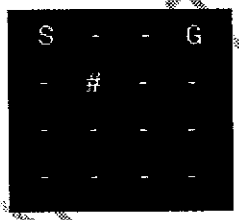
COURSE NAME: REINFORCEMENT LEARNING

COURSE INSTRUCTORS: DHA MAX. TIME: 1 Hour 30 Minutes

Note: *All questions are compulsory. Marks are indicated against each question in square brackets. Make assumptions wherever necessary.*

Q1. There is a grid-world environment with a goal state, obstacles, and a terminal state, where an agent needs to navigate to the goal while avoiding obstacles. The agent can move in four directions: up, down, left, and right. The goal is to find the optimal policy that leads the agent from the starting state (S) to the goal state (G) while avoiding obstacles.

Consider a 4x4 grid-world with the following layout: **[CO-3, Marks: 3+3]**



S: Starting state, G: Goal state, #: Obstacle

Assume the given policy to be a uniform random policy.

a) Give an algorithm to obtain the optimal policy using policy iteration method.

b) Trace the steps of the algorithm until the policy converges. Then, give the final optimal policy for the grid world problem.

Q2. a) Dynamic Programming can be used to solve the Planning problem in Reinforcement Learning. Justify. **[CO-1, Marks: 3+2+1]**

b) Explain utility of Bellman Optimality equations.

b) Define Monte Carlo Reinforcement Learning

Q3. a) Describe how the choice of the discount factor affects the performance of the first-visit Monte Carlo method. **[CO-3, Marks: 3+4]**

b) Given a game: A player competes against a dealer in a simplified version of the Blackjack card game, where the goal is to obtain a hand with a total value as close to 21 as possible without exceeding it. Each state in this problem represents the player's current hand value and the dealer's

face-up card. The possible actions are hitting (requesting an additional card) or standing (ending the turn). The player receives positive rewards for winning (say +1), negative rewards(say -5) for losing, and zero rewards for draws. By applying the first-visit Monte Carlo method, give the strategy for maximizing their expected winnings over time.

Q4. a) Model-free prediction methods are used in real-world applications where it's challenging to model the dynamics of the environment. Justify the above statement in case of autonomous driving. **[CO-3, Marks: 3+3]**

b) Compare First Visit and Every Visit Monte Carlo Policy Evaluation.