Paper / Subject Code: 50014 / Reinforcement Learning

15/05/2025 BE CSE-AIM SEM-VIII C-SCHEME REINFORCEMENT LEARNING QP CODE: 10082885

Duration 3 hours **Total marks 80**

N.B: (1) Question No. 1 is compulsory.
(2) Attempt any three questions out of the remaining five questions

Q 1. Attempt any four question **20 marks**

a. Define reinforcement learning and explain the key components involved in the RL framework. 5

b. Explain exploration approach and exploitation approach in multi armed bandit problem? 5

c. Enlist components of MDP model and explain in detail?5

d. What is the Bellman equation, and how does it relate to value iteration and policy iteration?5

e. Define Temporal Difference and explain parameters of TD in detail? 5

Q 2. A. **20 marks**

i. Discuss the difference between on-policy and off-policy learning. Provide examples of algorithms that fall into each category. 6

ii. What is optimal policies and explain optimal value function (q*)? 4

B.

i. Compare between value iteration and Policy iteration? 5

ii. Write gradient bandit algorithm and explain its steps? 5

Q. 3 **20 marks**

a. Define Offpolicy algorithm and onpolicy algorithm and identify SARSA is which type of algorithm and why? Write SARSA algorithm in detail? 10

b. Write Epsilon Greedy algorithm in detail with any one example? 10

Q. 4 **20 marks**

a. Explain the concept of Monte Carlo Prediction in reinforcement learning and describe the main steps involved in a Monte Carlo prediction algorithm. 10

b. Explain the concept of Deep Q-Networks (DQN) and discuss how deep learning can be integrated with Q-learning to solve complex problems. 10

Q. 5 **20 marks**

a. Write and explain off policy TD control using Q-learning? 5

b. Explain Generalised policy iteration of policy evaluation and policy improvement? 5

c. Define Agent and Environment and explain Agent Environment interface with diagram? 5

d. After 12 iterations of the UCB 1 algorithm applied on a 4-arm bandit problem, we have n1 = 3, n2 = 4, n3 = 3, n4 = 2 and Q12(1) = 0.55, Q12(2) = 0.63, Q12(3) = 0.61, Q12(4) = 0.40. Which arm should be played next? 5

Q. 6 **20 marks**

a. Explain the differences between TD learning and Monte Carlo methods. Also, describe the main components and key steps involved in TD prediction algorithms. 10

b. Explain the concept of Elevator Dispatching in a multi-floor building with diagram. Discuss the objectives and challenges of an elevator dispatching system. 10

****