1711E1316

M.Sc. Sem.-3 Examination

504

D. S.

November-2025

Candidate's Seat No :_____

Time : 2-30 Hours] **[Max. Marks : 70**

Instructions: All questions are compulsory. Use of non-programmable scientific calculator is allowed.

Q.1 **(a)** Define Reinforcement Learning and describe its core components. Explain how these elements interact within the reinforcement learning framework **(07)**

(b) Explain exploitation and exploration dilemma in the context of bandit settings. How does this dilemma differ from that in full Markov Decision Processes (MDPs)? **(07)**

OR

(a) Define a Markov Decision Process (MDP). Explain the components of an MDP and the Markov property. How does the MDP framework support sequential decision-making under uncertainty? Illustrate with a real-world example. **(07)**

(b) How reinforcement learning differs from other machine learning approaches. **(07)**

Q.2 **(a)** Explain bandit algorithm and its role in decision making. **(07)**

(b) Explain Generalized Policy Iteration (GPI) framework in detail. **(07)**

OR

(a) Explain bellman optimality in reinforcement in detail. **(07)**

(b) An agent uses epsilon-greedy strategy to select among three actions. The action values after several iterations are $Q(a_1) = 4.2, Q(a_2) = 3.7$ and $Q(a_3) = 4.9$. If the epsilon value is 0.1, (i) Identify the greedy action, (ii) Compute the probability of selecting each action, and (iii) If the agent chooses a_2, what is the probability that this choice was due to exploration rather than exploitation? **(07)**

Q.3 **(a)** You are given an environment with 1 state, x, and 2 actions, b and c. T is the terminal state. Your TD algorithm generates following episode using the policy π. **(07)**

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

- The policy π is given by: $\pi(b|x) = 0.9,\ \pi(c|x) = 0.1$.
- The current values of q are: $q(x, b) = 1$ and $q(x, c) = 2$.
- Discount factor, $\gamma = 0.5$ and step size $\alpha = 0.1$.

Show the values of $q(x, b)$ and $q(x, c)$ after their first update using 1-step SARSA, 2-step SARSA, and 2-step Expected SARSA. Also, compute first update for $q(x, b)$ using Q learning.

(b) Compare Temporal-Difference (TD) learning with Monte Carlo (MC) methods and Dynamic Programming (DP) in the context of prediction tasks in Reinforcement Learning. What are the key advantages of TD methods over the other two approaches? **(07)**

OR

(a) Explain the roles of prediction and control in Reinforcement Learning. How do on-policy and off-policy methods differ in their approach to learning and decision-making? **(07)**

P.T.O

E 1316.2

(b) Consider a finite, episodic and undiscounted Markov Decision Process (MDP) with states P and Q apart from the terminal state. The following two episodes are observed during Monte Carlo evaluation: Episode 1: (P,3) → (P,2) → (Q, −4) →(P,4) →(Q,−3) and Episode 2: (Q,−2) → (P,3) → (Q,−3). Each tuple (s, r) denotes that the agent is in state s and receives reward r upon transitioning to the next state. The episodes terminate after the last state shown. (a) Using *First-Visit Monte Carlo* evaluation, and (b) Using *Every-Visit Monte Carlo* evaluation, estimate the value function V(s) for states P and Q based on the observed episodes. **(07)**

Q.4 (a) Explain the need for function approximation in Reinforcement Learning. Compare and contrast it with tabular methods. Also, provide examples of where function approximation is crucial in real-world RL problems. **(07)**

(b) Describe the *Tile Coding* technique used for function approximation. Explain how it handles continuous state spaces. **(07)**

OR

(a) Describe how function approximation is used in both value-based and policy-based Reinforcement Learning algorithms. **(07)**

(b) Explain the concept of *parallel learning* in reinforcement learning. How it is beneficial to train an agent? **(07)**

Q.5 Attempt any **SEVEN** out of **TWELVE:** **(14)**

(1) Define (i) reward, and (ii) return

(2) Suppose $\gamma = 0.5$ and the sequence of rewards is received $R_1 = 3, R_2 = 7$ and $R_3 = 4$ with $T = 3$. What are the values for G_0, G_1 and G_2?

(3) What are the key differences between a stochastic policy and a deterministic policy?

(4) What is the significance of the discount factor in reinforcement learning?

(5) Briefly describe what bootstrapping is in the context of Reinforcement Learning.

(6) Provide comparison between policy iteration and value iteration.

(7) What is the main principle behind Monte Carlo methods in reinforcement learning?

(8) Why Q learning is an off-policy algorithm?

(9) When using TD learning, why is it better to learn action values (Q-values) rather than state values (V -values)?

(10) In a Deep Q-Network (DQN), what are the input and output of the neural network?

(11) What does the actor learn in the actor-critic architecture?

(12) What is the advantage function?

**** ****