U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

June 2025 Semester End Main Examinations

Programme: B.E. **Semester: VI**

Branch: Artificial Intelligence and Machine Learning **Duration: 3 hrs.**

Course Code: 24AM6PCREL **Max Marks: 100**

Course: Reinforcement Learning

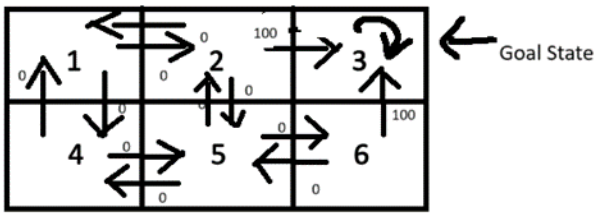
Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

			UNIT - I	*CO*	*PO*	Marks
1	a)		Explain the concept of evaluative feedback in REL (Reinforcement Learning) and highlight how it contrasts with instructive feedback. Also, mention an example of an algorithm that relies on instructive feedback.	*CO1*	*PO1*	**06**
	b)		Analyze how learning occurs in the Tic-Tac-Toe example using value functions.	*CO2*	*PO1*	**06**
	c)	i.	Illustrate the working of the ε-greedy algorithm and its approach to balancing exploration and exploitation in reinforcement learning with an example.	*CO2*	*PO1*	**08**
		ii.	Interpret the significance of decay in the ε-greedy algorithm and its impact on the learning process.			
			OR			
2	a)		Outline the main components of a reinforcement learning system and describe the role of each component with a diagram.	*CO1*	*PO1*	**06**
	b)		Derive the incremental implementation equation from the Action-Value Method.	*CO2*	*PO1*	**06**
	c)	i.	Formulate the Softmax equation used in reinforcement learning and identify its key components.	*CO2*	*PO1*	**08**
		ii.	Consider initial preferences H(A) = 2, H(B) = 1, and H(C) = 0, compute the action probabilities using the Softmax method.			
		iii.	Based on the computed probabilities, identify the most likely action to be chosen.			
			UNIT - II			
3	a)	i.	Explain the Markov Property in the context of reinforcement learning.	*CO1*	*PO2*	**06**

			ii. How does the state transition matrix represent the dynamics of a Markov process?			
		b)	Identify the differences between the State Value Function and the Action Value Function, and present their mathematical definitions.	*CO2*	*PO1*	**06**
		c)	In reinforcement learning, agents may interact with their environment through tasks that are either episodic or continuing in nature. i. Describe how episodic and continuing tasks differ in terms of agent-environment interaction. ii. Explain the importance of discount factor $\gamma \in [0,1]$ in continuing tasks. iii. Provide mathematical expression for the term G_t. iv. An agent receives a sequence of rewards over 5-time steps as follows: R = [3, 2, 4, 1, 5]. Compute the total return G using discount factor: i. $\gamma = 0.9$ ii. $\gamma = 0.1$ v. What value of γ reflects a far-sighted agent and which reflects a near-sighted agent and compare both?	*CO3*	*PO1*	**08**
			OR			
	4	a)	Differentiate between Dynamic programming and Monte Carlo method.	*CO1*	*PO1*	**06**
		b)	The value function $v\pi(s)$ under a policy π is defined as the expected return when starting from state s and following π thereafter. Derive the recursive form of the value function $v\pi(s)$ in terms of immediate reward, discount factor γ, and the next state's value. Clearly show how the Bellman equation is obtained.	*CO2*	*PO1*	**06**
		c)	A Markov Decision Process has the states S= {S1, S2, S3} and actions A = {A1, A2}. The transition and reward probabilities are given in the table. Compute the following: i. p(S3,7|S1,A1) ii. r (S1, A1). iii. p(S3|S1,A1) iv. r(S1,A1,S3)	*CO3*	*PO1*	**08**

States	Action a	Next States	Reward r	Probability p(s`, r|s, a)
S1	A1	S2	3	0.5
S1	A1	S3	7	0.5
S2	A2	S3	4	0.6
S2	A2S	S1	2	0.4

			UNIT - III			
	5	a)	Consider a 2x3 Grid world with immediate rewards given below.  Let the state space be S = {1,2,3,4,5,6} with grid 3 being the Goal State. The agent starts in any grid (0 to 6) and must learn to reach grid 3 using Q-learning. Moving into grid 3 (from rooms 2,6) gives a reward of +100 and all other moves give a reward of 0. i. Represent the environment using a reward matrix R. ii. Implement the Q-learning algorithm to update the Q-values in the Q Table (Consider Discount factor γ= 0.9).	*CO3*	*PO1*	**10**
		b)	Imagine you are designing a reinforcement learning agent to play a complex strategy game. The environment is initially unknown to the agent, but you have two options for your learning approach: • Option A: The agent tries to learn a model of the game's rules, including how the game state changes based on its actions, and then uses this model to plan future moves. • Option B: The agent learns a policy directly by interacting with the game and adjusting its actions based on the rewards it receives, without explicitly modelling the game's rules. i**.** Identify which option corresponds to model-based learning and which corresponds to model-free learning. ii**.** Discuss the differences between the two.	*CO2*	*PO1*	**06**
		c)	Describe the $TD(0)$ algorithm used for policy evaluation, including its update rule and key steps.	*CO2*	*PO1*	**04**
			OR			
	6	a)	A delivery drone is flying through different checkpoints (P, Q, R, S) to deliver packages. Each checkpoint grants the drone a reward based on delivery success or time saved. The drone's flight path and rewards are recorded over 5 episodes. The discount factor γ=1 (no discounting).	*CO3*	*PO1*	**10**

Episode	Time Step	State	Reward
1	0	P	0
	1	Q	1
	2	R	2
	3	S	3
	4	Terminal	-
2	0	Q	1
	1	R	0
	2	Q	2
	3	S	2

					4	Terminal	-	
				3	0	R	2	
					1	Q	1	
					2	S	3	
					3	Terminal	-	
				4	0	P	1	
					1	S	3	
					2	R	2	
					3	Terminal	-	
				5	0	Q	3	
					1	R	1	
					2	P	2	
					3	S	0	
					4	Terminal	-	
			i. Calculate the return Gt (sum of future rewards) for each visit to each checkpoint in all episodes (with γ=1). ii. Compute the estimates for each checkpoint by averaging the returns from their first visits in each episode. iii. Compute the every-visit Monte Carlo estimates for each checkpoint by averaging the returns from all visits across episodes.					
		b)	Differentiate between SARSA and Q-Learning Algorithms			*CO1*	*PO1*	**06**
		c)	List the advantages of Monte Carlo over Dynamic Programming.			*CO1*	*PO1*	**04**
			UNIT - IV					
	7	a)	i. Analyze how the use of the max operator in Q-learning can lead to instability during training. ii. Suggest any one technique that addresses the instability caused by the max operator in Deep Q-Learning. iii. Elaborate on the use of Experience Reply buffer in Deep Q – Learning.			*CO1*	*PO1*	**10**
		b)	An agent interacts with the environment and observes the following transition: State A → Action a → Reward = 2 → State B using a Deep Q-Network (DQN), where the Q- function is approximated using a second-degree polynomial over state features as follows: $Q(s,a:w) = w_1\phi_1(S) + w_2[\phi_1(S)]^2 + w_3\phi_2(S) + w_4[(\phi_2(S)]^2$ ➔ Feature vectors: ● $\phi(A) = [1,2]$ ● $\phi(B) = [2,0]$ ➔ Network Weights: ● Online network: $w_{online} = [0.5,0.1,0.0,0.2]$ ● Target network: $w_{target} = [0.2,0.1,0.1,0.1]$ ● Target network is updated once in 2 iterations ➔ Parameters: ● Discount factor: $\gamma = 0.9$ ● Learning rate: $\alpha = 0.5$			*CO3*	*PO2*	**10**

			OR			
	8	a)	You're implementing reinforcement learning agents to solve a complex navigation task in a simulated environment. You are considering two policy-gradient methods: A2C and A3C. i. Outline the key architectural difference between A2C and A3C ii. "A3C typically perform better in highly dynamic or large environments". Justify iii. Illustrate how the update mechanism differ between A2C and A3C.	*CO2*	*PO1*	**10**
		b)	Derive the mathematical formulation underlying direct policy optimization that maximizes expected return without relying on value function estimation or knowledge of environment dynamics.	*CO2*	*PO1*	**10**
			UNIT - V			
	9	a)	A delivery robot in a warehouse starts at point A, moves to checkpoint B, and then to its destination C, completing one delivery episode. The transitions and rewards are: ● From A to B: reward = +1 ● From B to C: reward = +2 ● C is a terminal state. The robot wants to estimate the value of being at each location using TD(λ) with Discount factor $\gamma = 1$, Trace decay parameter $\lambda = 0.5$, Initial values: V(A) = V(B) = V(C) = 0, Learning rate $\alpha = 1$. i. Using the forward view (λ-return), compute the value update for V(A) after this episode. ii. Using the backward view (eligibility traces), compute the value update for V(A) after this episode. iii. Explain why the results are the same, illustrating the equivalence of the two views.	*CO2*	*PO1*	**10**
		b)	Watkins' Q(λ) is one of the earliest attempts to combine Q-learning with eligibility traces in an off-policy setting, yet it exhibits two fundamental shortcomings. Identify these two key limitations of Watkins' Q(λ) and briefly explain why each impedes in off policy learning with eligibility traces.	*CO2*	*PO1*	**10**
			OR			
	10	a)	Consider a robot navigating a linear environment with four states: A, B, C, and D (where D is the Terminal state). The robot has only one available action at each state: Move Forward. The state transitions and corresponding rewards are as follows: • moving from A to B yields a reward of 1, • from B to C yields a reward of 2, and • from C to D yields a reward of 3. The robot always follows a fixed policy of moving forward at every state. Initially, the Q-values for all state-action pairs are set	*CO2*	*PO1*	**10**

			to 0. Use the SARSA(λ) algorithm with the following parameters: a learning rate (α) of 0.1, a discount factor (γ) of 1.0, and a trace decay parameter (λ) of 0.8. i. Perform one complete episode starting from state A and ending at the terminal state D. After each step, update the Q-values using replacing eligibility traces. ii. Show the eligibility trace values after each step, update the Q-values accordingly, and finally report the Q-values after the episode ends.			
		b)	Eligibility traces enhance learning in reinforcement learning by supporting both forward and backward perspectives. Elaborate on the following points: i. The role of eligibility traces in unifying TD and Monte Carlo methods using a spectrum of n-step updates. ii. The function of eligibility traces as temporary memory traces for assigning credit during learning updates. iii. The significance of combining forward and backward views for improved temporal credit assignment and learning performance.	*CO2*	*PO1*	**10**

****