

RollNo.

--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--	--

## ANNA UNIVERSITY (UNIVERSITY DEPARTMENTS)

B.E. /B.Tech / B. Arch (Full Time) - END SEMESTER EXAMINATIONS, NOV / DEC 2024

INFORMATION TECHNOLOGY  
VII Semester  
ITM504 REINFORCEMENT LEARNING  
(Regulation 2019)

Max.Marks: 100

Time: 3hrs

CO 1	Understand the different terminologies of RL and concepts of Probability
CO 2	Illustrate the Markov decision Process and Bellman Equation for learning.
CO 3	Apply dynamic programming techniques to Markov Decision Process and Monte Carlo Methods
CO 4	Implement Time Differencing Learning for Real World Problems
CO 5	Apply the approximation methods for Learning and Q-Learning Techniques.

**BL – Bloom's Taxonomy Levels**

(L1-Remembering, L2-Understanding, L3-Applying, L4-Analysing, L5-Evaluating, L6-Creating)

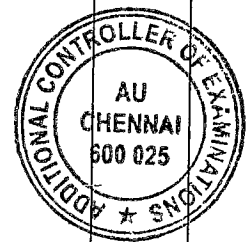
**PART- A(10x2=20Marks)**  
(Answer all Questions)

Q.No.	Questions	Marks	CO	BL
1	Explain the trade-off between exploration and exploitation in reinforcement learning. Why is it important, and how do algorithms like $\epsilon$ -greedy address this trade-off?	2	1	1
2	List out the components of Reinforcement learning.	2	1	1
3	What is a discount factor? Why is it required in reward calculation?	2	2	2
4	State the Markovian assumption. Why is it important?	2	2	2
5	State the characteristics of dynamic programming. State the Bellman equation.	2	3	2
6	What is the Monte Carlo algorithm for random walk?	2	3	2
7	State the differences between Q-learning and SARSA learning.	2	4	1
8	State the differences between online and offline policy.	2	4	1
9	What is the universal approximation theorem? How is it helpful for linear function approximation?	2	5	1
10	What is a policy gradient theorem?	2	5	1

**PART- B(5x 13=65Marks)**  
(Restrict to a maximum of 2 subdivisions)

Q.No.	Questions	Marks	CO	BL
11 (a)	A two-state Markov chain consists of two states, let's call them S1 and S2, with a Transition Probability Matrix (TPM) that describes the probabilities of moving from one state to another.  Suppose the TPM is:	13	1	L3

	$\begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$ <p>Assuming the initial probability distribution as (0.5,0.5), show the prediction after 3 years.</p> <p>If the initial probability distribution is not given, how will you find it and repeat the prediction for 3 years?</p>																			
OR																				
11 (b)	Explain the multiarmed bandit problem. Explain the approaches of epsilon-greedy, softmax, UCB and Thompson approach for solving multiarmed bandit problem with respect to a company advertisement policy or three advertisements.	13	1	L3																
12 (a)	<p>Explain in detail the value and policy iteration algorithms. Apply it to the following scenario:</p> <p>Assume a 3x3 grid. The goal is to reach the top-right corner (3,3) from the bottom-left corner (1,1). The states are each cell in the grid is a state, represented by its coordinates (x, y). Let the actions be <b>Up</b>, <b>Down</b>, <b>Left</b>, or <b>Right</b> (unless at the grid boundary where some actions are not allowed). The reward is +10 when you reach (3,3) and -1 for each step taken otherwise. Let the discount factor be 0.9. Let the transition probabilities be - 80% chance the intended action occurs and 20% chance you move randomly to one of the other adjacent cells.</p> <p>Explain how the value and policy iteration algorithms work.</p>	13	2	L4																
OR																				
12 (b)	What is dynamic programming? List out its characteristics and explain how it is useful in solving the RL problems with the value and policy iteration.	13	2	L4																
13 (a)	<p>Consider a 3x3 grid where you start at the top-left corner (0, 0) and aim to reach the bottom-right corner (2, 2). You can only move <b>right</b> or <b>down</b>. Each cell (i,j) has a cost associated with entering it, as given as</p> <table><tr><td></td><td>(0)</td><td>(1)</td><td>(2)</td></tr><tr><td>(0)</td><td>1</td><td>4</td><td>5</td></tr><tr><td>(1)</td><td>5</td><td>6</td><td>10</td></tr><tr><td>(2)</td><td>11</td><td>8</td><td>9</td></tr></table> <p>Find the shortest path using a dynamic programming approach.</p>		(0)	(1)	(2)	(0)	1	4	5	(1)	5	6	10	(2)	11	8	9	13	3	L3
	(0)	(1)	(2)																	
(0)	1	4	5																	
(1)	5	6	10																	
(2)	11	8	9																	
OR																				
13 (b)	Explain in detail the first visit and every visit MC algorithm. Outline the differences between them and highlight the differences with a simple numerical example.	13	3	L3																
14 (a)	Explain the SARSA algorithm in detail. Take a grid problem and show the application of it.	13	4	L4																



OR				
14 (b)	Explain the Q-Learning algorithm in detail. Take a maze problem and show its application to find the solution.	13	4	<u>L4</u>
15 (a)	Explain the concept of policy gradient and explain the REINFORCE algorithm with its variants.	13	5	<u>L5</u>
OR				
15 (b)	Explain in detail the working of an actor-critic method and its implementation in A2C and A3C algorithms.	13	5	<u>L5</u>

**PART- C(1x 15=15Marks)**  
(Q.No.16 is compulsory)

Q.No.	Questions	Marks	CO	BL
16.	<p>1. Find the optimal path from (3,1) to the goal (1,3) using dynamic programming. Assume the given minimum costs as given below:</p> <pre> 4  6  10 3  X  7 2  1  X </pre> <p>The X's are obstacles. So, it is necessary to avoid that. What value would one assign then (high or low), and do accordingly and find the shortest path possibility.</p> <p>2. How would you approach the above problem using deep neural networks? Outline the approach of DQN networks with its implementation details to solve the above problem. Show layer-by-layer details of DQN.</p>	7 + 8	<u>5</u>	<u>6</u>

