Reg No.:_____ Name:_____

APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY

B.Tech Degree S8 (R,S) Examinations April 2026 (2019 Scheme)

Course Code: AMT402

Course Name: CONCEPTS IN REINFORCEMENT LEARNING

Max. Marks: 100 **Duration: 3 Hours**

PART A

Answer all questions, each carries 3 marks. Marks

1	Let A and B be independent events, where P(A)=0.5 and P(B)=0.6. Find: 1. P(A∩B) 2. P(A∪B) 3. P(A∩B′)	(3)
2	A biased coin has P(heads)=0.6. The coin is tossed repeatedly until two successive heads (HH) occur for the first time. What is the probability that the experiment is completed on the 5th toss?	(3)
3	List the main components and flow of a reinforcement learning cycle.	(3)
4	What are some real-world problems where reinforcement learning is commonly used?	(3)
5	Why do we use a discount factor in reinforcement learning, and what does it represent?	(3)
6	What does an optimal policy mean in reinforcement learning, and what is its significance?	(3)
7	Explain the main strengths and limitations of using Monte Carlo methods for policy evaluation in reinforcement learning.	(3)
8	How is the depth parameter utilized in Monte Carlo Tree Search, and what are three important roles it plays?	(3)
9	In what situations is importance sampling used, and why is it necessary in reinforcement learning?	(3)
10	What are two techniques to encourage exploration in a Q-learning agent?	(3)

PART B

Answer any one full question from each module, each carries 14 marks.

Module I

11	a)	An experiment consists of tossing a fair coin three times. Find the Probability Mass Function (PMF) and the Cumulative Distribution Function (CDF) for the following random variables: i. The number of tails before the first head ii. The number of heads after the first tail iii. The difference between number of heads and tails iv. The sum of positions (indices) where heads appear	(7)
	b)	A fair six-sided die is rolled **three times**. Let the random variable X be the **maximum value** observed among the three rolls. Find the **Probability Mass Function (PMF)**, the **expected value** E[X], and the **standard deviation** σX.	(7)
		OR	
12	a)	The first two digits of a vehicle number are **KA**. The remaining **four digits** (from 0 to 9) are randomly assigned. What is the probability that the complete number contains **six distinct characters**?	(5)
	b)	Suppose that two components have independent exponentially distributed lifetimes, T1and T2, with parameters αand β, respectively. Find (a) P (T1> T2) and (b) P (T1> 2 T2).	(9)
		Module II	
13	a)	How do model-based and model-free reinforcement learning methods differ? Provide one example of each.	(5)
	b)	How is reinforcement learning different from supervised and unsupervised learning? Explain with examples, focusing on how RL handles sequential decision-making and feedback.	(9)
		OR	
14	a)	In what ways does reinforcement learning differ from traditional machine learning paradigms such as supervised and unsupervised learning? Discuss with reference to learning objectives, feedback mechanisms, and real-world applications.	(7)
	b)	What are the strengths and limitations of reinforcement learning? In what kinds of problems does it excel, and what challenges arise when deploying it in real-world applications? Suggest ways to overcome these challenges.	(7)
		Module III	

15 a) How do episodic and continuing reinforcement learning tasks differ? Provide the reward calculation formula for each type. (4)

b) Consider designing a recycling robot whose job is to collect empty bottles around the building. The robot has a sensor to detect when a bottle is in front of it, and a gripper to pick up the bottle. It also senses the level of its battery. The robot can navigate, as well as pick up a bottle and throw a bottle it is holding in the trash. There is a battery charger in the building, and the robot should not run out of battery. Describe this problem as an MDP. What are the states and actions? (10)

Suppose that you want the robot to collect as many bottles as possible, while not running out of battery. Describe what rewards would enable it to achieve this task.

OR

16 a) What is the difference between a state value function and an action-value function in reinforcement learning? (4)

b) Based on the given transition matrix, answer the following: (10)

i) State Space: Define the set of all possible states in the given Markov Decision Process (MDP).

ii) Transition Graph: Draw the state transition probability graph based on the given transition matrix.

iii) Episodes: Explain the concept of episodes in reinforcement learning and describe how an episode would progress through the given state transitions.

Transition Matrix

$P =$

	Facebook	Class 1	Class 2	Class 3	Pub	Sleep	Pass
Facebook	0.9	0.1	0	0	0	0	0
Class 1	0.5	0	0.5	0	0	0	0
Class 2	0	0	0	0.8	0.2	0.2	0
Class 3	0	0	0	0	0.4	0	0.6
Pub	0	0.2	0	0.4	0	0.4	0
Sleep	0	0	0	0	0	1.0	0
Pass	0	0	0	0	0	0	1.0

Module IV

17 a) In a given MDP, assume we aim to evaluate a policy using Monte Carlo methods based solely on actual interaction with the environment. We choose the first-visit method and randomly select the start state from the set of all possible states for each (7)

episode. Will this strategy allow us to fully evaluate the action-value function for the policy?

b) In order for Q-learning to converge to the optimal policy, what condition must the exploration strategy satisfy regarding the exploration-exploitation trade-off? (7)

OR

18 a) Can value updates in the Value Iteration algorithm be performed in parallel across states? Why or why not? (5)

b) Consider an undiscounted Markov Reward Process with two states A and B. The transition matrix and reward function are unknown, but you have observed two sample episodes: (9)

A +3 --> A +2 --> B -4 --> A +4 --> B -3

B -2 --> A +3 --> B -3

i. Using first-visit Monte-Carlo evaluation, estimate the state-value function V(A), V(B).

ii. Using every-visit Monte-Carlo evaluation, estimate the state-value function V(A), V(B).

iii. Draw a diagram of the Markov Reward Process that best explains these two episodes. Show rewards and transition probabilities on your diagram.

Module V

19 a) How do on-policy and off-policy learning differ in reinforcement learning, and why is SARSA considered on-policy while Q-learning is off-policy? Provide an example for each. (7)

b) What is the n-step Tree Backup algorithm, and how does it support off-policy learning without relying on importance sampling? (7)

OR

20 a) What are the key benefits of using Temporal-Difference (TD) prediction in reinforcement learning compared to other methods like Monte Carlo or Dynamic Programming? (7)

b) How does the SARSA algorithm work as an on-policy Temporal Difference (TD) control method? Illustrate with an example (7)

****