Roll No.												

PRESIDENCY UNIVERSITY

BENGALURU

End - Term Examinations – May/June 2026

Date: 14 – 05- 2026 **Time:** 01:00pm – 04:00pm

School: SOCSE	**Program:** B.Tech - CAI	
Course Code: CAI2507	**Course Name:** REINFORCEMENT LEARNING	
Semester: VI	**Max Marks**: 100	**Weightage**: 50%

CO - Levels	CO1	CO2	CO3	CO4
Marks	24	24	26	26

Instructions:

(i) Read all questions carefully and answer accordingly.

(ii) Do not write anything on the question paper other than roll number.

Part A

Answer ALL the Questions. Each question carries 2marks. **10Q x 2M=20M**

1.	How do we derive the value function from the Q function?	**2 Marks**	**L2**	**CO1**
2.	What are the steps involved in policy iteration?	**2 Marks**	**L1**	**CO1**
3.	What is the difference between first-visit MC and every-visit MC?	**2 Marks**	**L2**	**CO2**
4.	How does on-policy control differ from off-policy control?	**2 Marks**	**L2**	**CO2**
5.	What is the update rule of TD learning?	**2 Marks**	**L1**	**CO3**
6.	How does the TD prediction method work?	**2 Marks**	**L1**	**CO3**
7.	How does Q learning differ from SARSA?	**2 Marks**	**L2**	**CO3**
8.	How do we compute the upper confidence bound?	**2 Marks**	**L1**	**CO4**
9.	What are the steps involved in Thompson sampling?	**2 Marks**	**L1**	**CO4**
10.	What are contextual bandits?	**2 Marks**	**L1**	**CO4**

Part B

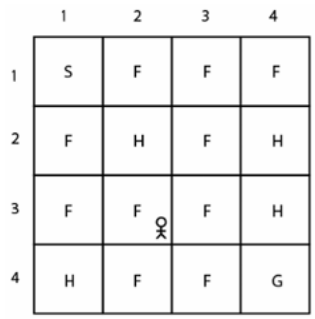
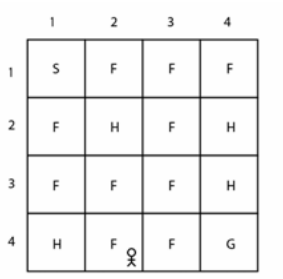
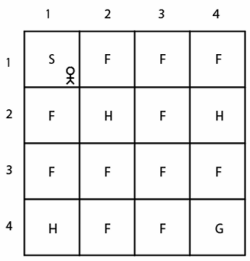
Answer the Questions. **Total Marks 80M**

<table>
<tr><td>11.</td><td>a.</td><td>Explain the concepts of value function in Reinforcement Learning. Discuss their mathematical foundations of value function with suitable example.</td><td>10 Marks</td><td>L2</td><td>CO1</td></tr>
<tr><td></td><td>b.</td><td>Explain the various elements of RL with the grid world environment as an example.</td><td>10 Marks</td><td>L2</td><td>CO1</td></tr>
<tr><td colspan="6">Or</td></tr>
<tr><td>12.</td><td>a.</td><td>Discuss Bellman optimality equation and explore how it is useful for finding the optimal Bellman value and Q functions.</td><td>10 Marks</td><td>L2</td><td>CO1</td></tr>
<tr><td></td><td>b.</td><td>Outline the working steps of the Value Iteration algorithm. Using the model dynamics of State A, compute the optimal value of State A after the first iteration of Value Iteration. (γ = 1)

<table>
<tr><th>State
(s)</th><th>Action
(a)</th><th>Next State
(s')</th><th>Transition Probability
$P(s'\mid s,a)$ or $P^a_{ss'}$</th><th>Reward Function
$R(s,a,s')$ or $R^a_{ss'}$</th></tr>
<tr><td>A</td><td>0</td><td>A</td><td>0.1</td><td>0</td></tr>
<tr><td>A</td><td>0</td><td>B</td><td>0.8</td><td>-1</td></tr>
<tr><td>A</td><td>0</td><td>C</td><td>0.1</td><td>1</td></tr>
<tr><td>A</td><td>1</td><td>A</td><td>0.1</td><td>0</td></tr>
<tr><td>A</td><td>1</td><td>B</td><td>0.0</td><td>-1</td></tr>
<tr><td>A</td><td>1</td><td>C</td><td>0.9</td><td>1</td></tr>
</table>
Model dynamics of state A</td><td>10 Marks</td><td>L3</td><td>CO1</td></tr>
</table>

<table>
<tr><td>13.</td><td>a.</td><td>Describe the step-by-step algorithm of Monte Carlo prediction (Q function), and illustrate with a suitable example.</td><td>10 Marks</td><td>L2</td><td>CO2</td></tr>
<tr><td></td><td>b.</td><td>Explain First-Visit Monte Carlo prediction. Compare it with Every-Visit MC prediction using an illustrative example.</td><td>10 Marks</td><td>L2</td><td>CO2</td></tr>
<tr><td colspan="6">Or</td></tr>
<tr><td>14.</td><td>a.</td><td>Discuss the On-policy Monte Carlo Control with Exploring Starts (MC-ES) algorithm used in reinforcement learning.</td><td>10 Marks</td><td>L2</td><td>CO2</td></tr>
<tr><td></td><td>b.</td><td>Explain the On-policy Monte Carlo Control with epsilon-greedy policy algorithm used in reinforcement learning.</td><td>10 Marks</td><td>L2</td><td>CO2</td></tr>
</table>

<table>
<tr><td>15.</td><td>a.</td><td>Explain the steps involved in the off-policy TD control algorithm and compute Q-learning algorithm for the Frozen Lake environment under the given policy. Using the provided transition data, determine the updated value estimates for the state (3,2). Assume the learning rate $\boldsymbol{\alpha = 0.1}$ and the discount factor $\boldsymbol{\gamma = 1}$.</td><td>10 Marks</td><td>L3</td><td>CO3</td></tr>
</table>

		 **State** | **Action** | **Value** (1,1) Up 0.5 (3,2) Up 0.1 (3,2) Down 0.8 (3,2) Left 0.5 (3,2) Right 0.6 (4,2) Up 0.2 (4,2) Down 0.5 (4,2) Left 0.1 (4,2) Right 0.8 (4,4) Right 0.5			
	b.	Discuss the steps involved in the on-policy TD control algorithm and compute SARSA algorithm for the Frozen Lake environment under the given policy. Using the provided transition data, determine the updated value estimates for the state (4,2). Assume the learning rate $\alpha = 0.1$ and the discount factor $\gamma = 1$.  **State** | **Action** | **Value** (1,1) Up 0.5 (4,2) Up 0.3 (4,2) Down 0.5 (4,2) Left 0.1 (4,2) Right 0.8 (4,3) Up 0.1 (4,3) Down 0.3 (4,3) Left 1.0 (4,3) Right 0.9 (4,4) Right 0.5	**10 Marks**	**L3**	**CO3**
		Or			
16.	**a.**	Describe the steps involved in the Temporal Difference (TD) prediction algorithm and compute the Temporal Difference (TD) prediction for the Frozen Lake environment under the specified policy. Using the provided transition data, determine the updated value estimates for the states (1,1), (1,2), and (1,3). Assume the learning rate $\alpha = 0.1$and the discount factor $\gamma = 1$. **State** | **Action** (1,1) Right (1,2) Right (1,3) Left ⋮ ⋮ (4,4) Down  **State** | **Value** (1,1) 0.9 (1,2) 0.6 (1,3) 0.8 ⋮ ⋮ (4,4) 0.7	**10 Marks**	**L3**	**CO3**
	b.	Describe and compare DP, MC, and TD learning techniques using an illustrative example.	**10 Marks**	**L2**	**CO3**

17.	**a.**	Illustrate with example, how to find the best arm with the epsilon-greedy method in Multi-Armed Bandit (MAB) problem.	**10 Marks**	**L2**	**CO4**
	b.	Explain in detail Softmax exploration strategies for finding the optimal arm in Multi-Armed Bandit (MAB) problem with example.	**10 Marks**	**L2**	**CO4**
		Or			

18.	**a.**	Discuss the steps involved in Thompson sampling method and how Thompson sampling exploration strategy to overcome the exploration-exploitation dilemma based on a beta distribution.	**10 Marks**	**L2**	**CO4**
	b.	Describe the Upper confidence bound for handling the exploration-exploitation dilemma in MAB problem with example.	**10 Marks**	**L2**	**CO4**