| Roll No. | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|

# PRESIDENCY UNIVERSITY

## BENGALURU

**Make up Examinations – December 2025**

**Date:** 31 – 12- 2025 **Time:** 09:30am – 12:30pm

| | | |
|---|---|---|
| **School:** SOCSE | **Program:** B.Tech | |
| **Course Code :**CSE3011 | **Course Name :**REINFORCEMENT LEARNING | |
| **Semester**: MK | **Max Marks**:100 | **Weightage**:50% |

| **CO - Levels** | **CO1** | **CO2** | **CO3** | **CO4** | **CO5** |
|---|---|---|---|---|---|
| **Marks** | **16** | **26** | **34** | **24** | |

**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 x2M=20M**

| | | | | |
|---|---|---|---|---|
| **1** | Define the return of a trajectory for a continuous task | **2 Marks** | **L1** | **CO1** |
| **2** | Illustrate the equation to find Q(S) in the Monte Carlo method | **2 Marks** | **L2** | **CO2** |
| **3** | Interpret SARSA update rule | **2 Marks** | **L2** | **CO3** |
| **4** | Define the following<br>a) Agent b) Environment | **2 Marks** | **L1** | **CO1** |
| **5** | Illustrate the equation to find V(S) in the Monte Carlo method | **2 Marks** | **L2** | **CO2** |
| **6** | State optimal policy | **2 Marks** | **L1** | **CO2** |
| **7** | X is a random variable with the outcomes of throwing a dice. Find the expectation E[f(x)] where f(x)=X3 | **2 Marks** | **L1** | **CO1** |
| **8** | Distinguish between SARSA and Q-Learning algorithms | **2 Marks** | **L2** | **CO3** |
| **9** | Interpret upper confidence bound of an arm in MAB Problem | **2 Marks** | **L1** | **CO4** |
| **10** | Describe Thompson sampling | **2 Marks** | **L2** | **CO4** |

## Part B

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

| | | | | | |
|---|---|---|---|---|---|
| **11.** | **a.** | Discuss stochastic environment and deterministic environment in RL with an example | **10 Marks** | **L2** | **CO1** |
| | **b** | Explain Markov chain and Markov property in Markov decision process | **10 Marks** | **L3** | **CO2** |

**Or**

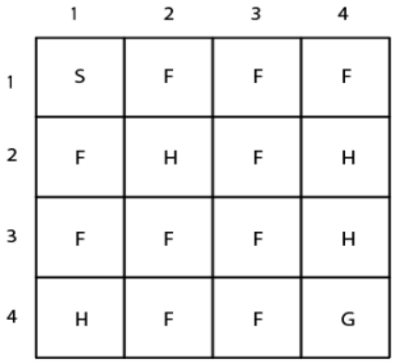
| | | | | | |
|---|---|---|---|---|---|
| **12.** | **a.** | Illustrate Bellman Equation for the Value function in deterministic environment | **10 Marks** | **L2** | **CO1** |
| | **b** | Using the value iteration algorithm and the model dynamics of state A given in the table below, Compute the optimal value of state A, after the first iteration<br><br>State $(s)$ / Action $(a)$ / Next State $(s')$ / Transition Probability $P(s'\|s,a)$ or $P^a_{ss'}$ / Reward Function $R(s,a,s')$ or $R^a_{ss'}$<br>A, 0, A, 0.1, 0<br>A, 0, B, 0.8, -1<br>A, 0, C, 0.1, 1<br>A, 1, A, 0.1, 0<br>A, 1, B, 0.0, -1<br>A, 1, C, 0.9, 0 | **10 Marks** | **L3** | **CO2** |

| | | | | | |
|---|---|---|---|---|---|
| **13.** | **a.** | Explain different types of RL environments with an example each | **10 Marks** | **L2** | **CO2** |
| | **b.** | Articulate TD Prediction algorithm in FZLE environment | **10 Marks** | **L3** | **CO3** |

Or

| | | | | | |
|---|---|---|---|---|---|
| **14.** | **a.** | Discuss Markov Decision Process in Detail | **10 Marks** | **L2** | **CO2** |
| | **b.** | Articulate TD Control algorithm in FZLE environment | **10 Marks** | **L3** | **CO3** |

| | | | | | |
|---|---|---|---|---|---|
| **15.** | **a.** | Determine the Algorithmic steps of the off policy TD Control-Q Learning technique .Also find the updated Q Value of state (3,2) using the following data , apply epsilon greedy technique (choose exploitation) to choose current and next actions. Assume alpha as 0.1 and gamma as 1 | **10 Marks** | **L3** | **CO3** |




| 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 |
| ⋮ | ⋮ | ⋮ |
| (4,4) | Right | 0.5 |

Figure 5.13: The Frozen Lake environment with a randomly initialized Q table

| 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 |
| ⋮ | ⋮ | ⋮ |
| (4,4) | Right | 0.5 |

| | | | | | |
|---|---|---|---|---|---|
| | **b** | Illustrate upper confidence bound method to overcome the exploration –exploitation dilemma with the algorithm | **10 Marks** | **L3** | **CO4** |

**Or**

| | | | | | |
|---|---|---|---|---|---|
| **16.** | **a.** | Articulate MAB problem and list out the applications of MAB | **10 Marks** | **L3** | **CO3** |
| | **b** | Apply softmax exploration method of exploration-exploitation strategy for the following 4 arm bandit<br><br>arm 1: Q = 1<br>arm 2: Q = 0<br>arm 3: Q = 0<br>arm 4: Q = 0 | **10 Marks** | **L3** | **CO4** |

| | | | | | |
|---|---|---|---|---|---|
| **17.** | **a.** | Explain SARSA in FZLE | **10 Marks** | **L2** | **CO3** |
| | **b.** | Implement the reinforcement learning environment namely, Frozen Lake Environment using a random policy and show the output of the following: | **10 Marks** | **L3** | **CO4** |

| | | | | | |
|---|---|---|---|---|---|
| | | a. Create and render the environment b. State Space<br>c. Action Space d. State transition probabilities | | | |

**Or**

| | | | | | |
|---|---|---|---|---|---|
| **18.** | **a.** | Explain TD control types in brief | **10 Marks** | **L2** | **CO3** |
| | **b.** | Write a python program to find an optimal policy using Q-learning for the frozen lake environment with alpha=0.85 and gamma=0.90 and epsilon =0.8<br>Create and render the environment<br>Generate policy using 30 episodes with 60 timesteps<br>Print the optimal policy | **10 Marks** | **L3** | **CO4** |

# ***** BEST WISHES *****