Mahindra University Hyderabad
École Centrale School of Engineering
Minor-II
Subject: Reinforcement Learning and Autonomous Systems (AI 4102)

Date: 11/11/2023 **Start Time: 2:00 pm**
Time Duration: 1 hour 30 minutes **Max. Marks: 30**

Instructions:

1) The submitted answer sheet should only contain your final answer. Use separate sheets for rough work.
2) You must show your solution procedure in the submitted answer sheet.
3) The paper has a sum total of **40 marks**. However, you can score a maximum of **30 marks**.

Q1: Multi-Armed Bandits (15 marks, 25 minutes)

Consider a Multi-Armed Bandit (MAB) setup with three actions indexed 1 to 3. The agents actions and rewards for the first ten time slots are as follows:

Time slot	1	2	3	4	5	6	7	8	9	10
Action	2	3	2	2	1	1	1	2	3	3
Reward	4.5	2.0	1.0	4.0	2.5	2.6	7.0	5.0	1.6	3.2

Answer the following questions:

(a) Suppose that the agent is using UCB policy. What action will it choose in time slot 11? **(5 marks)**

(b) Suppose that the agent is using ε-greedy policy with $\varepsilon = 0.1$. What is the probability of choosing action 2 in time slot 11? **(5 marks)**

(c) Suppose that the agent is using ε-greedy policy with $\varepsilon = 0.1$. Suppose that the true reward distributions of the three actions are normal distributions. The variance of the normal distribution is 1 for all the actions. The mean of the normal distribution for actions 1, 2, and 3 are 3.5, 4, and 2 respectively. What is the probability distribution of the agent's reward in time slot 11? Instructions for part (c): **(5 marks)**

- This is a **slightly tough** question. So, spend your time judiciously.
- You do not have to calculate the final answer.
- In case you forgot, the formula for normal distribution is $f(x) = \frac{1}{\sigma\sqrt{2\pi}} exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Q2: Markov Decision Process (15 marks, 40 minutes)

Let's design a policy to play "Kaun Banega Crorepati"! To keep things simple, we assume that there are no "lifelines" and "checkpoints" *(If you don't know what lifelines and/or checkpoints are, then don't worry. It is not required to solve this question)*.

There are N questions indexed n. We have $1 \leq n \leq N$. As n increases, the difficulty of the questions increases. In time slot t, the contestant is presented "question number n" where $n = t$. After being presented the question, the contestant has to decide the following:

1. *Quit:* If the contestant quits, the game terminates and the contestant gets to **keep the money** that he/she has won till now.

2. *Answer question:* If the contestant answers the question incorrectly, the game terminates and the contestant **loses all the money**. If the contestant answers correctly, he/she wins $f(n)$ INR and one of the two things happen:
   i. If $t < N$, the contestant moves to "question number $n + 1$" in the next time slot.
   ii. If $t = N$, then the game terminates and the contestant **keeps all the money** he/she has earned.

   The probability of answering "question number n" correctly is θ_n. As n increases, $f(n)$ increases and θ_n decreases.

We want design a policy to maximize the net expected money that the contestant wins. This is equivalent to maximizing the β-discounted reward for this problem with $\beta = 1$ *(you do not have to worry how these two problems are equivalent)*. Answer the following questions:

(a) What is the state and state space for this problem? **(2 marks)**

(b) What is the action and action space for this problem? **(2 marks)**

(c) Define the reward for all state-action pair for this problem? **(4 marks)**
CAUTION: Put emphasis on defining the reward for the case when the contestant answers and loses all the money.

(d) What is the Bellman optimality equation for this problem? **(7 marks)**

(PTO)

Q3: Reinforcement Learning (10 marks, 25 minutes)

(a) Briefly explain Temporal Difference (TD) Learning? Solve the problem shown in Figure 1 using TD Method to calculate state transition and find utility policy $U^{\pi}(s)$? (Assume: Discount factor $\gamma = 1$ and learning rate $\alpha = 0.5$.) **(5 Marks)**

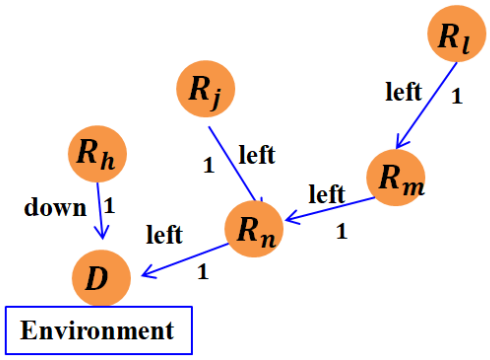

Figure 1

(b) List the differences between Q-learning and SARSA? **(5 Marks)**