RO3002: Reinforcement Learning

Assignment-2

Question 1: **(30 marks)**

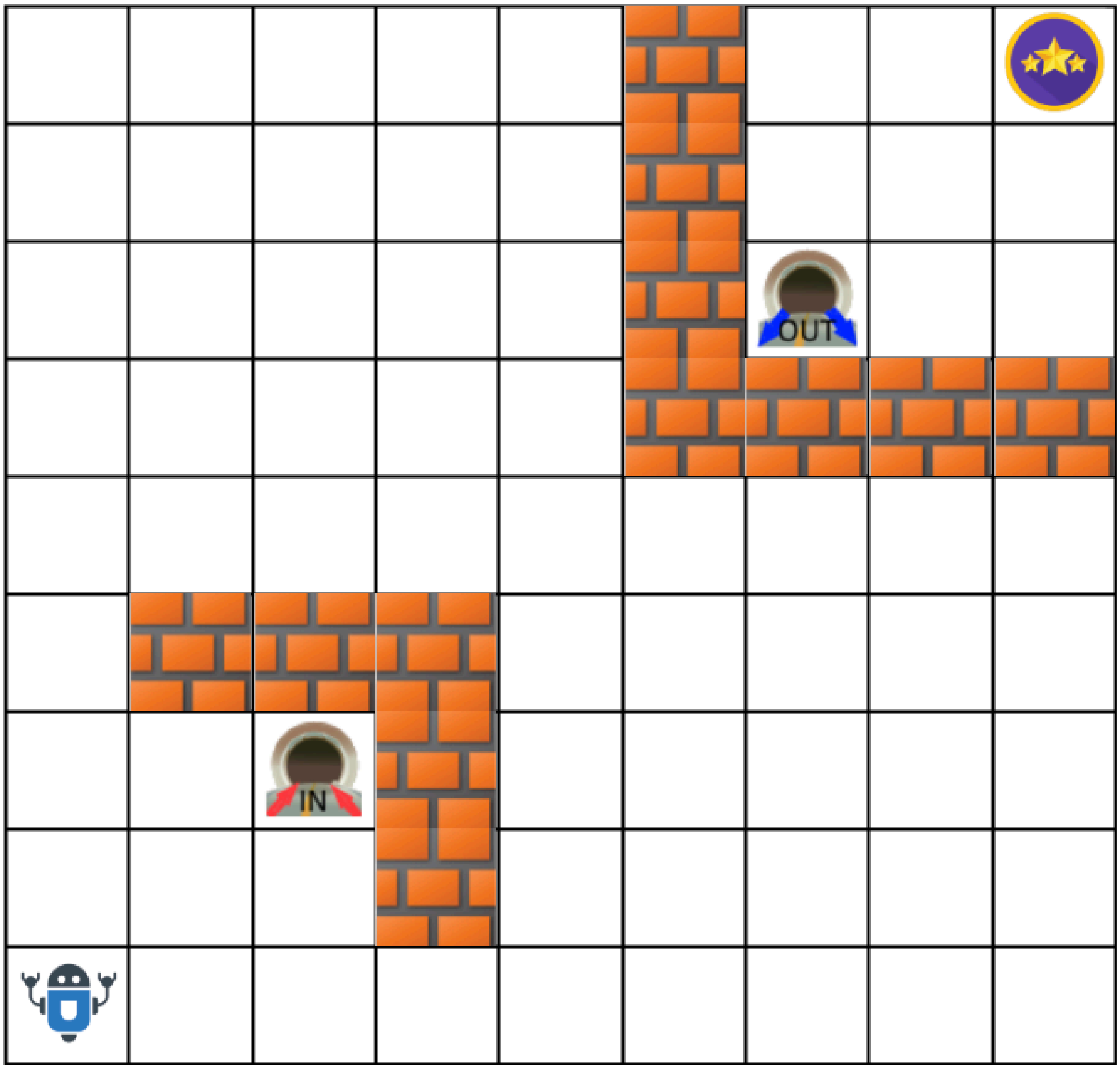
An undergraduate student at Plaksha University has the task of attending classes and eating food during their tenure in college. The student has access to three locations on campus: hostel, Bharti Airtel block (academic block), and mess (canteen). The student receives a reward of -1 for staying in the hostel, +3 for attending class at the Bharti Airtel block, and +1 for being in the mess. At any given time, the student can either eat food or attend class.

When the student is at the hostel, they attempt to attend classes by either going to the Bharti Airtel block with a 50% probability or staying in the hostel with a 50% probability. If the student is hungry, they always go to the mess from the hostel with a 100% probability. From the Bharti Airtel block, the student attends class by either staying in the Bharti Airtel block with a 70% probability or going to the mess with a 30% probability. If the student becomes hungry while in the Bharti Airtel block, they either go to the mess with an 80% probability or stay in the Bharti Airtel block with a 20% probability. At the mess, the student has a 60% chance of attending classes by going to the Bharti Airtel block, a 30% chance of attending class by going to the hostel, and a 10% chance of attending from the mess itself. If the student is hungry, they always stay in the mess with a 100% probability.

Using this information, design a finite MDP by writing down the possible combinations of states, actions, transition probability from one state to another for a given action, and rewards in a tabular form. Also, draw a diagram of the MDP from the information mentioning the probability and rewards.

- Based on the designed MDP, perform value iteration and show the optimal value for each state and the policy obtained.
- Based on the designed MDP, perform policy iteration and show the optimal policy.
- Discuss the results obtained from policy iteration and value iteration.

Question 2: **(30 marks)**



You are given a 9x9 grid-world environment where:

- The robot icon marks the agent's starting location.
- The star symbol represents the goal position.

- Two tunnels, labelled IN and OUT, serve as one-way portals. The agent can enter through IN and exit through OUT.
- The agent receives a reward of +1 upon reaching the goal; in all other states, the reward is 0.

Your task is to solve this problem using **Value Iteration**, **Policy Iteration, and Monte Carlo** techniques. Specifically, you are required to:

1. Implement Value Iteration to compute the optimal policies.
2. Implement Policy Iteration to compute the optimal policies.
3. Implement Monte Carlo with Exploring Starts to compute the optimal policies.
4. Visualize the optimal policy for each method by plotting a quiver plot, showing the direction of the agent's optimal movements at each grid cell for all three approaches.

Due Date:- 28th March 2025