Name:

Student ID:

Deep RL Course
Instructor: S. A. Emami
TA: M. H. Narimani
Fall 2024, SUT

Problem statement:

Consider a 2×2 grid-world problem as described below. We aim to solve the control problem using the SARSA algorithm with a greedy policy at each step. Consider $\gamma = 1$, $\alpha = 1$.

1 (Starting point)	2
3	4 (Target point)

- **Generate Q-tables** for 10 iterations. If, after an iteration, you reach the target, return to the starting point to begin a new episode. Report the best trajectory at the end of the training.

Action \ State	a (up)	b (down)	c (left)	d (right)
1	12	12	12	12
2	12	12	12	12
3	12	12	12	12

- Initialize the Q-table with $Q(x, u) = 12$ for all cells. This approach is known as "**Optimistic Initial Values**."
- **Rewards**: If your new state is inside the grid but not the target, the reward is -2. If the new state is outside the grid, the reward is -4. If you reach the target, the reward is 10.
- If an action takes you outside the grid, you will return to the previous cell (not the starting cell).
- If two actions have the same value, you may choose one randomly.
- The values for the target point and all cells outside the grid are assumed to be zero.