Reinforcement Learning Lab Report

Assignment 3

Jayanth S (201081003)
Praveen Kumar N (201082001)
Rishabh Roy (201082002)

1 Algorithms analysed

- Value-Iteration
- Policy-Iteration

2 Classical Maze Problem

We considered the classical maze problem (Finite Markov Decision Process problem) and use value iteration and policy iteration to solve them. The maze environment consisted of *free cells* (including the start and terminal state) and *wall*. The agent/robot should find the shortest path from the start state to the terminal state such that it's total discounted reward is maximized. i.e.,

$$G = \sum_{t=0}^{\infty} \gamma^t R_t, \tag{1}$$

where $\gamma \in [0, 1)$ is the discount factor, R_t is the reward obtained at time instant t and G_T is the total discounted reward over the entire path from starting state to the terminal state.

The detailed problem setting is given as follows: In each state the agent can take one of the following actions :

1. Left ($<$)
2. Up ($\wedge$)
3. Right ($>$)
4. Down ($\vee$)

The reward structure is as follows:

- Movement to an adjacent state will cost = 0.1.
- Attempt to move towards a wall will cost = 0.75 and resulting state remains same as the starting state.
- Attempt to move towards boundary will cost = 0.8 and resulting state remains same as the starting state.
- Reaching the terminal state will result in a reward = 10.

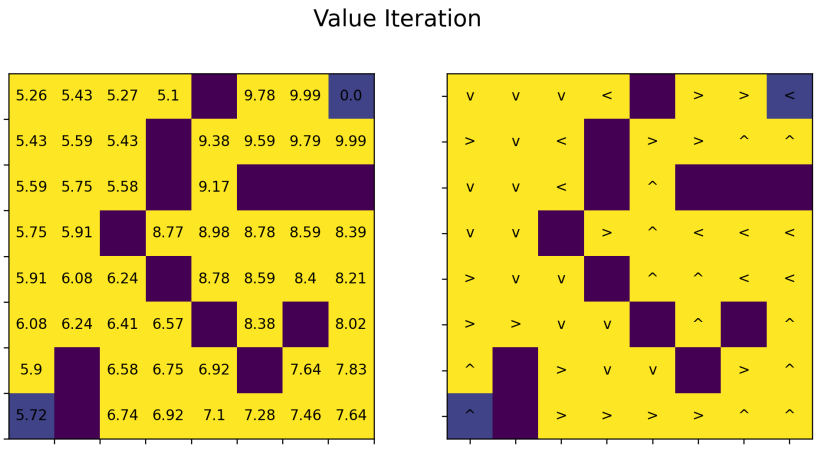

Figure 1: The value in each cell in left figure corresponds to the optimal state value for that cell. The right figure is the plot of optimal deterministic actions corresponding to each cell obtained using value iteration. (Ignore the action of the terminal state)

3 Observations

- It is observed that value iteration takes relatively lesser time when compared with the policy iteration. This might be because for each iteration, value iteration performs operations of $O(|\mathcal{A}||\mathcal{S}|^2)$ and policy iteration performs operations of $O(|S|^3+|A||S|^2)$, where $\mathcal{A}$ is the action space and $\mathcal{S}$ is the state space.
- However policy iteration requires relatively very less iterations to converge when compared with value iteration.

4 References

- Deep Reinforcement Learning for Maze Solving
- Markov Decision Processes
- Markov Decision Processes and Exact Solution Methods

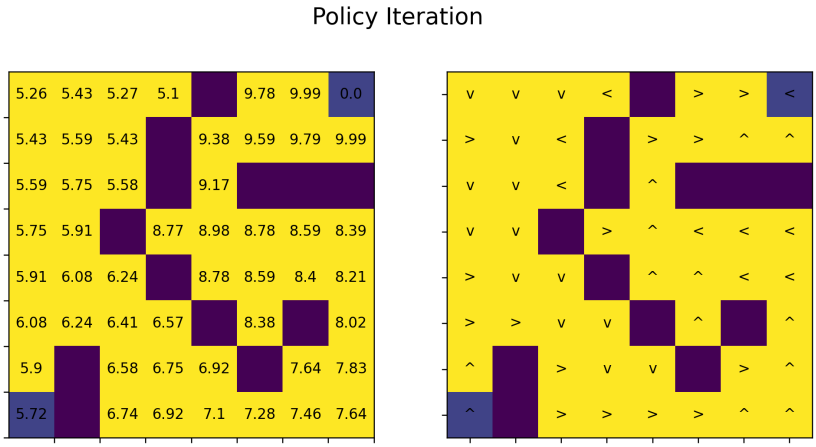

Figure 2: The value in each cell in left figure corresponds to the optimal state value for that cell. The right figure is the plot of optimal deterministic actions corresponding to each cell obtained using policy iteration. (Ignore the action of the terminal state)

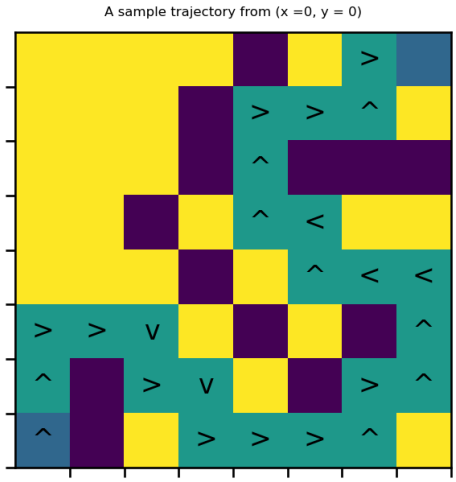

Figure 3: A sample trajectory obtained by following the optimal deterministic policy that's obtained using policy iteration