

# AI 3000 / CS 5500 : REINFORCEMENT LEARNING

## ASSIGNMENT No 2

DUE DATE : 25/09/2023

---

Couse Instructor : Easwar Subramanian

10/09/2023

### Problem 1 : Value Iteration

- (a) Prove that the Bellman optimality operator is a contraction under the max-norm (5 Points)
- (b) Prove that the iterative policy evaluation algorithm converges geometrically (3 Points)
- (c) Let  $M$  be an infinite horizon MDP and  $V^*$  be its optimal value function. Suppose if the value iteration algorithm is terminated after  $k + 1$  iterations as  $\|V_{k+1} - V_k\|_\infty < \epsilon$  for some chosen  $\epsilon > 0$ , how far is the estimate  $V_{k+1}$  from the optimal value function  $V^*$  ? Provide details of your derivation. (5 Points)

### Problem 2 : Programming Value Iteration

- (a) Implement value iteration and policy iteration algorithm. Modularize the implementation to contain separate functions for policy evaluation, value iteration and policy improvement. Test the implementation on the  $4 \times 4$  **Frozen Lake** environment available in Gymnasium (formerly known as Open AI Gym). The stochastic version of the environment is default which can be changed by modifying the `is-slippy` flag. (8 Points)
- (b) Document your findings with number of iterations needed for both algorithms to converge (or nearly converge) to optimal policy on the **Frozen Lake** environment. Further, provide a snapshot of the optimal policy obtained via the two algorithms. (4 Points)
- (c) Are there any stochastic optimal policies ? If so, does any of the algorithm find any stochastic optimal policy ? If not, why not ? (2 Points)
- (d) Consider the grid world problem similar to **Frozen Lake** shown in Figure 1. The grid has two terminal states with positive payoff (+1 and +10). The bottom row is a cliff where each state is a terminal state with negative payoff (-10). The greyed squares in the grid are walls. The agent starts from the yellow state  $S$ . As usual, the agent has four actions  $\mathcal{A} = (\text{Left}, \text{Right}, \text{Up}, \text{Down})$  to choose from any non-terminal state and the actions that take the agent off the grid leaves the state unchanged. Notice that, if agent follows the dashed path, it needs to be careful not to step into any terminal state at the bottom row that has negative payoff. There are four possible (optimal) paths that an agent can take.

