AI 3000 / CS 5500 : Reinforcement Learning

Assignment № 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 evalution algorithm converges geometrically (3 Points)

(c) Let M be an infinite horizon MDP and V^* be its optmal 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 evalution, value iteration and policy improvement. Test the implementation on the 4×4 **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-slippery 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** envrionment. 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 $\mathcal{A} =$ (Left, Right, Up, 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.

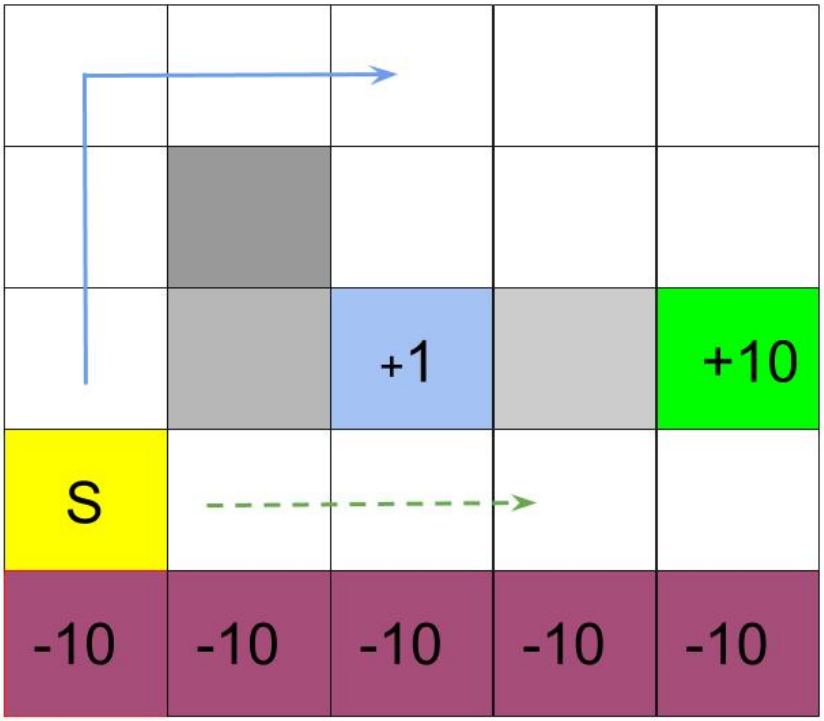

Figure 1: Modified Grid World

- Prefer the close exit (state with reward +1) but risk the cliff (dashed path to +1)
- Prefer the distant exit (state with reward +10) but risk the cliff (dashed path to +10)
- Prefer the close exit (state with reward +1) by avoiding the cliff (solid path to +1)
- Prefer the distant exit (state with reward +10) by avoiding the cliff (solid path to +10)

There are two free parameters to this problem. One is the discount factor γ and the other is the noise factor (η) in the environment. Noise makes the environment stochastic. For example, a noise of 0.2 would mean the action of the agent is successful only 80 % of the times. The rest 20 % of the time, the agent may end up in an unintended state after having chosen an action.

(i) Implement the above environment in Python 3.8+. (8 Points)

(ii) Use any of the DP algorithms implemented above on this environment and observe the optimal paths for various choices of γ and η. Identify what values of γ and η that could lead the agent to each of the optimal paths listed and explain the reasoning for the answer obtained. (4 Points)

(iii) After solving this grid world example, please re-visit your answer to question 2(h) of Assignment 1 (1 Point)

ALL THE BEST