# 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)

There are many ways to prove the result. Please find some help in the links below

- https://runzhe-yang.science/2017-10-04-contraction/.
- https://towardsdatascience.com/mathematical-analysis-of-reinforcement-learning-bellman-equation-ac9f0954e19f

(b) Prove that the iterative policy evalution algorithm converges geometrically (3 Points)

We can prove that,

$$\|V_{k+1} - V\|_\infty \leq \gamma \|V_k - V\|_\infty$$

(See the answer to question below) Applying this inequality recursively, we get the desired result

(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)

The last iterate of the algorithm is $V_{k+1}$ and we know that $\|V_{k+1} - V_k\|_\infty \leq \varepsilon$. By using the triangular inequality (of norms) and by using the fact $BV_k = V_{k+1}$ where $B$ is the Bellman evaluaiton backup, we have,

$$\begin{aligned} \|V_k - V\|_\infty &\leq \|V_k - V_{k+1}\|_\infty + \|V_{k+1} - V\|_\infty = \|V_k - V_{k+1}\|_\infty + \|BV_k - BV\|_\infty \\ &\leq \|V_k - V_{k+1}\|_\infty + \gamma\|V_k - V\|_\infty = \varepsilon + \gamma\|V_k - V\|_\infty \end{aligned}$$

Therefore, $\|V_k - V\|_\infty \leq \frac{\varepsilon}{1-\gamma}$. This allows us to conclude that,

$$\|V_{k+1} - V\|_\infty = \|BV_k - BV\|_\infty \leq \gamma\|V_k - V\|_\infty \leq \frac{\gamma\varepsilon}{1-\gamma}$$

## Problem 2 : Programming Value Iteration

(a) Are there any stochastic optimal policies ? If so, does any of the algorithm find any stochastic optimal policy ? If not, why not ? (2 Points)

The value and policy iteration algorithm will not find any stochastic policies as they both look out for deterministic policies.

(b) 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.

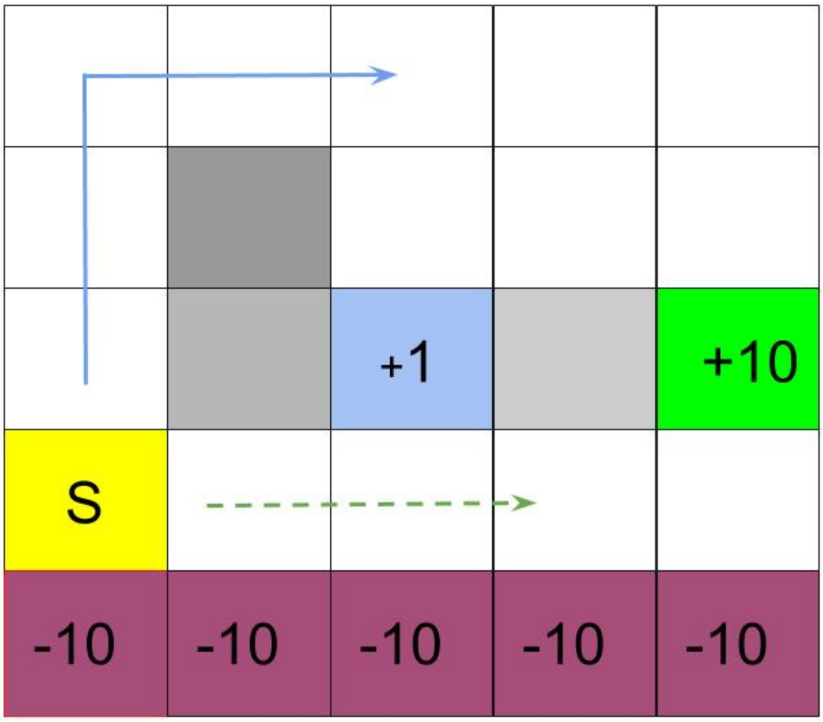

Figure 1: Modified Grid World

- 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 $\gamma$ and the other is the noise factor ($\eta$) 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 $\gamma$ and $\eta$. Identify what values of $\gamma$ and $\eta$ that could lead the agent to each of the optimal paths listed and explain the reasoning for the answer obtained. (4 Points)

   i. When $\gamma$ is low, RL agent is 'short sighted' and better rewards available in the distant future is not given importance. Further, when noise is zero in the environment, there is no danger of tripping to the cliff. Therefore, for low $\gamma$ and low $\eta$, the agent would prefer the close exit and risk the cliff.

   ii. When $\gamma$ is low, RL agent is 'short sighted' and better rewards available in the distant future is not given importance. Further, when noise is high or moderate in the environment, there is danger of tripping to the cliff. Therefore, for low $\gamma$ and low $\eta$, the agent would prefer the close exit and not risk the cliff.

   iii. When $\gamma$ is high, RL agent is 'far sighted' and better rewards available in the distant future is given importance. Further, when noise is low or zero in the environment, there is less or no danger of tripping to the cliff. Therefore, for high $\gamma$ and low $\eta$, the agent would prefer the distant exit and risk the cliff.

   iv. When $\gamma$ is high, RL agent is 'far sighted' and better rewards available in the distant future is given importance. Further, when noise is high or medium in the environment, there is danger of tripping to the cliff. Therefore, for high $\gamma$ and high $\eta$, the agent would prefer the distant exit and not risk the cliff.

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

The example above illustrates that scaling the discount factor does affect the choice of optimal policy.

## ALL THE BEST