ROB311-TD2

Iad Abdul-Raouf and Madeleine Becker

September 2020, 28th

Preliminary observations

It is said in the questions of this TP2 that $\gamma \in [0, 1]$ (for instance see question 4). However the environment does not contain any terminal states and does contain only positive rewards with some rewards being strictly positive. Thus $\gamma = 1$ would lead V^* to take infinite values in the iterative value algorithm. Therefore, we assumed it was a misprint and that γ actually lies in the interval $[0, 1[$ (french notation to exclude 1 from the interval).

Moreover, in the formulas of V^* and π^*, it is our understanding that max_a and $argmax_a$ implicitly implies that only possible actions a are taken into account. In order to make everything explicit, let us introduce the following notation :

$$A(S) = \{a \,|\, a \text{ is a possible action from state } S\}$$

This notation will be used in the demonstration of the two lemmas before questions 3 and 4.

Question 1

All states only have the action a_0 associated with, except for the state S_0 which can be associated with the actions a_1 and a_2 (see figure 4).

Therefore, there are only two possible policies :

$$\pi_1 : (S_0, S_1, S_2, S_3) \longmapsto (a_1, a_0, a_0, a_0)$$
$$\text{and } \pi_2 : (S_0, S_1, S_2, S_3) \longmapsto (a_2, a_0, a_0, a_0)$$

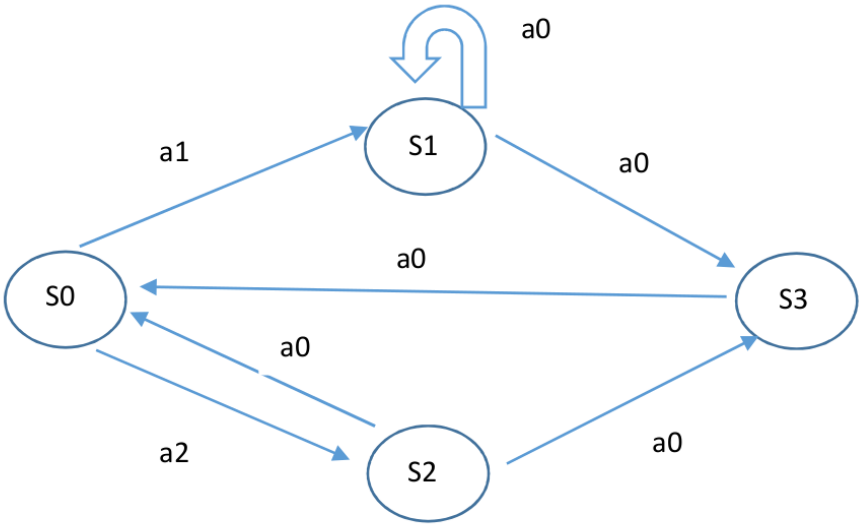

Figure 1: Graph of the links between states and associated actions

Question 2

The TD statement gives the following formula :

$$V^*(S) = R(S) + max_a \left(\gamma \sum_{S'} T(S, a, S') V^*(S') \right) \tag{1}$$

From there we can deduce for each state :

$$\begin{aligned} V^*(S_0) &= R(S_0) + max_a \left(\gamma T(S_0, a_1, S_1) V^*(S_0), \gamma T(S_0, a_2, S_2) V^*(S_0) \right) \\ &= 0 + max_a \left(\gamma * 1 * V^*(S_0), \gamma * 1 * V^*(S_0) \right) \\ &= max_a \left(\gamma V^*(S_0), \gamma V^*(S_0) \right) \end{aligned}$$

$$\begin{aligned} V^*(S_1) &= R(S_1) + max_a \left(\gamma T(S_1, a_0, S_1) V^*(S_1) + \gamma T(S_1, a_0, S_3) V^*(S_3) \right) \\ &= \gamma(1-x) V^*(S_1) + \gamma x V^*(S_3) \end{aligned}$$

$$\begin{aligned} V^*(S_2) &= R(S_2) + max_a \left(\gamma T(S_2, a_0, S_0) V^*(S_0) + \gamma T(S_2, a_0, S_3) V^*(S_3) \right) \\ &= 1 + \gamma(1-y) V^*(S_0) + \gamma y V^*(S_3) \end{aligned}$$

$$\begin{aligned} V^*(S_3) &= R(S_3) + max_a \left(\gamma T(S_3, a_0, S_0) V^*(S_0) \right) \\ &= 10 + \gamma V^*(S_0) \end{aligned}$$

Lemmas

To answer the two following questions, we are first going to demonstrate two lemmas :

Lemma. *If all rewards are positive then* $\forall S, V^*(S) \geqslant 0$

and

Lemma. *For a given* $\gamma \in [0, 1[$*,* V^* *is upper-bounded.*

Demonstration of the first lemma

We have the formula (1) :

$$V^*(S) = R(S) + max_{a \in A(S)} \left(\gamma \sum_{S'} T(S, a, S') V^*(S') \right) \tag{2}$$

where $A(S)$ is the set of possible actions from S (see Preliminary observations).

We define S_{min}, the state which minimizes $V^*(S)$. Then, for all states S :

$$\begin{aligned} V^*(S) &\geqslant R(S) + max_{a \in A(S)} \left(\gamma \sum_{S'} T(S, a, S') V^*(S_{min}) \right) \\ &\geqslant R(S) + max_{a \in A(S)} \left(\gamma V^*(S_{min}) \sum_{S'} T(S, a, S') \right) \\ &\geqslant R(S) + max_{a \in A(S)} \left(\gamma V^*(S_{min}) \times 1 \right) \\ &\geqslant R(S) + \gamma V^*(S_{min}) \end{aligned}$$

Thus, applied in particular to $S = S_{min}$:

$$\begin{aligned} V^*(S_{min}) &\geqslant R(S_{min}) + \gamma V^*(S_{min}) \\ \iff (1 - \gamma) V^*(S_{min}) &\geqslant R(S_{min}) \geqslant min_S(R(S)) \\ \iff V^*(S_{min}) &\geqslant \frac{min_S(R(S))}{1 - \gamma} \end{aligned}$$

Here are all rewards positive, therefore for all states S :

$$\boxed{V^*(S) \geqslant V^*(S_{min}) \geqslant 0}.$$

Demonstration of the second lemma

In the same way, we define S_{max}, the state which maximizes $V^*(S)$. Then, for all states S :

$$\begin{aligned} V^*(S) &\leqslant R(S) + max_{a \in A(S)} \left(\gamma \sum_{S'} T(S, a, S') V^*(S_{max}) \right) \\ &\leqslant R(S) + \gamma V^*(S_{max}) \end{aligned}$$

Thus, applied in particular to $S = S_{max}$,

$$\begin{aligned} & V^*(S_{max}) \leqslant R(S_{max}) + \gamma V^*(S_{max}) \\ \iff & (1-\gamma)V^*(S_{max}) \leqslant R(S_{max}) \leqslant max_S(R(S)) \\ \iff & V^*(S_{max}) \leqslant \frac{max_S(R(S))}{1-\gamma} \end{aligned}$$

Yet we have $\forall S, V^*(S) \leqslant V^*(S_{max})$.
Therefore, for a given $\gamma \in [0, 1[$, $\boxed{\text{the value function is upper-bounded}}$.

Question 3

We are looking for a value of x for which $\forall \gamma \in [0, 1[, y \in [0, 1], \pi^*(S_0) = a_2$, where :

$$\pi^*(S) = argmax_a \left(\sum_{S'} T(S, a, S') V^*(S') \right). \tag{3}$$

For $S = S_0$:

$$\pi^*(S) = argmax_a \left(V^*(S_1), V^*(S_2) \right)$$

Yet we have on first hand :

$$V^*(S_1) = \gamma(1-x)V^*(S_1) + \gamma x V^*(S_3)$$

With $x = 0$:

$$\begin{aligned} & V^*(S_1) = \gamma V^*(S_1) \\ \iff & (1-\gamma)V^*(S_1) = 0 \end{aligned}$$

With $\gamma < 1$, we can deduce that $V^*(S_1) = 0$.
On the other hand,

$$V^*(S_2) = 1 + \gamma(1-y)V^*(S_0) + \gamma y V^*(S_3)$$

Thanks to the first lemma, we can state that $\gamma(1-y)V^*(S_0) + \gamma y V^*(S_3) \geqslant 0$, therefore $V^*(S_2) \geqslant 1 > 0 = V^*(S_1)$. We can deduce from that :

$$\boxed{x = 0 \implies \pi^*(S_0) = a_2}$$

Question 4

We want to prove that the TD statement is wrong, therefore that $\forall \gamma \in [0,1[, \forall y \in [0,1], \exists x > 0 \,|\, \pi^*(S_0) = a_2$.

As before,

$$\pi^*(S) = argmax_a \left(V^*(S_1), V^*(S_2)\right)$$

We already have proven at the previous question that

$$\forall x \in [0,1], V^*(S_2) = 1 + \gamma(1-y)V^*(S_0) + \gamma y V^*(S_3) \geqslant 1.$$

We want to prove that there exists an x value such as $V^*(S_1) < 1$, because then it would be lower than $V^*(S_2)$, and then $\pi^*(S_0) = a_2$.

$$V^*(S_1) = \gamma(1-x)V^*(S_1) + \gamma x V^*(S_3) \iff V^*(S_1) = \frac{\gamma x}{1-\gamma(1-x)} V^*(S_3)$$

Yet, since the highest reward is 10, the second lemma states that :

$$V^*(S_1) = \frac{\gamma x}{1-\gamma(1-x)} V^*(S_3) \leqslant \frac{\gamma x}{1-\gamma(1-x)} \frac{10}{1-\gamma} \xrightarrow[x \to 0]{} 0$$

Therefore, for a given $0 \leqslant \gamma < 1$, there exists an x value close enough to zero such as $V^*(S_1) < 1 = V^*(S_2)$, and for this value of x, $\pi^*(S_0) = a_2$.

Question 5

Python code

Please, find our code in the following github repository : `https://github.com/MalphaBecker/ROB311-TD2`.

It contains our main file applying the value iteration algorithm to the graph of the TP by calling our functions defined in the file `aux_function.py`.

You just have to run `main_iad_madeleine.py` in your IPython console. Parameters of the algorithm and parameters of the graph are at the beginning of the main file and can be changed at will.

Results

To measure convergence, we used the root mean square value (RMS method). Having $x = y = 0.25, \gamma = 0.9$, and the stop condition ϵ on the RMS value set to $\epsilon = 0.01$, leading to a convergence in 50 iterations, the optimal policy found by our algorithm is :

$$\pi^* : (S_0, S_1, S_2, S_3) \longmapsto (a_1, a_0, a_0, a_0)$$

You can also find the final values of V^*, the evolution of the RMS and the evolution of the policy on the figures below which clearly show convergence. For the policy, only $\pi^*(0)$ is shown as the others are constant equal to a_0.

```
In [1]: runfile('/home/iad/Documents/ENSTA/3A/rob_311/TP_rob311/TP2/ROB311-TD2/main_i
iad/Documents/ENSTA/3A/rob_311/TP_rob311/TP2/ROB311-TD2')
the estimated optimal policy is : [1. 0. 0. 0.]
number of iteration : 50
final V = [14.1025855  15.67876763 15.61484469 22.68402157]
final RMS : 0.009228192433676341
parameters : gamma = 0.9 , eps = 0.01
```

Figure 2: Final results of the algorithm

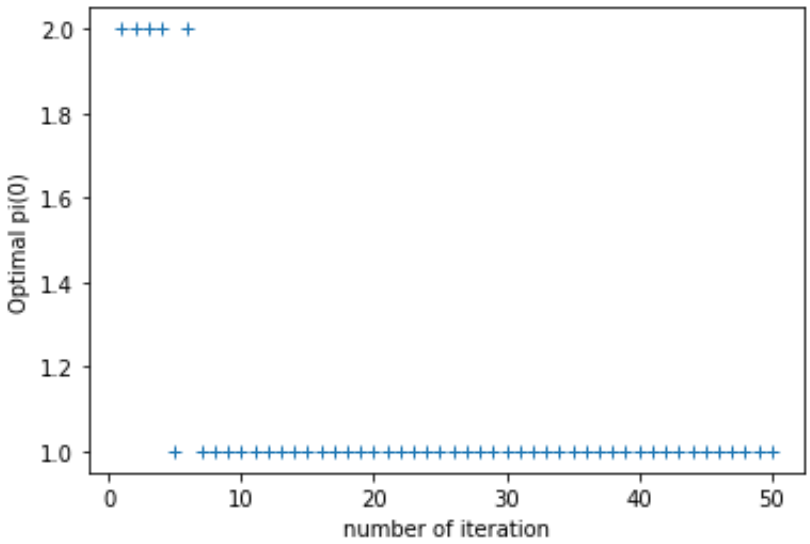

Figure 3: Optimal policy for S_0 at each iteration until convergence

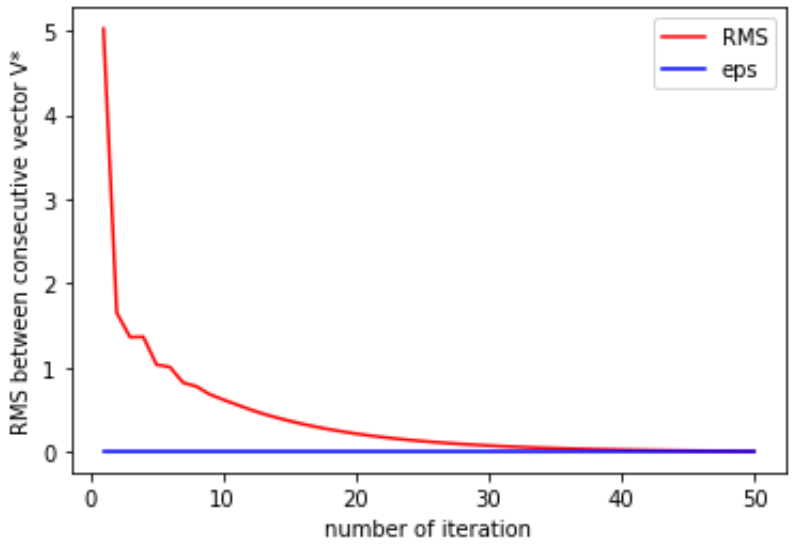

Figure 4: Root mean square error between two consecutive values of V*