TP2 Reinforcement Learning

Yan CHEN & Dajing GU

September 2020

In this TP, the algorithm of Value Iteration in Reinforcement Learning is implemented in python. The code associated can be found in the GitHub repository: TP2. All $V^*(*)$ is supposed to be > 0 in this report.

1 Question 1

	s_0	s_1	s_2	s_3
π_1	a_1	a_0	a_0	a_0
π_2	a_2	a_0	a_0	a_0

Table 1.1: All possible policies

2 Question 2

According to the formula:

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

and

$$R(S) = \begin{cases} 10, \text{ for state } S3 \\ 1, \text{ for state } S2 \\ 0, \text{ otherwise} \end{cases}$$

$$T(S,a0,S') = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1-x & 0 & x \\ 1-y & 0 & 0 & y \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$T(S,a1,S') = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$T(S, a2, S') = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

We can calculate easily:

$$\begin{aligned} V^*(S_0) &= R(S_0) + \max_a \gamma \sum_{S'} T(S_0, a, S')V^*(S')) \\ &= \gamma \cdot \max(B_0, B_1, B_2) \end{aligned}$$

With [1]

$$\begin{aligned} B_0 &= 0 * V^*(S_0) + 0 * V^*(S_1) + 0 * V^*(S_2) + 0 * V^*(S_3) \\ &= 0 \end{aligned}$$

$$\begin{aligned} B_1 &= 0 * V^*(S_0) + 1 * V^*(S_1) + 0 * V^*(S_2) + 0 * V^*(S_3) \\ &= V^*(S_1) \end{aligned}$$

$$\begin{aligned} B_2 &= 0 * V^*(S_0) + 0 * V^*(S_1) + 1 * V^*(S_2) + 0 * V^*(S_3) \\ &= V^*(S_2) \end{aligned}$$

So we obtain:

$$V^*(S_0) = \gamma \cdot \max[0, V^*(S_1), V^*(S_2)] \tag{1}$$

In the same way, we have:

$$V^*(S_1) = \gamma \cdot [(1 - x)V^*(S_1) + xV^*(S_3)] \tag{2}$$

$$V^*(S_2) = 1 + \gamma \cdot [(1 - y)V^*(S_0) + yV^*(S_3)] \tag{3}$$

$$V^*(S_3) = 10 + \gamma \cdot V^*(S_0) \tag{4}$$

3 Question 3

(1) Analysis:

Knowing that

$$\begin{aligned} \pi^*(S_0) &= \arg\max_a \sum_{S'} T\left(S_0, a, S'\right) V^*\left(S'\right) \\ &= \arg\max_a (0, V^*(S_1), V^*(S_2)) \end{aligned}$$

With respectively:

$$\arg(0) = a_0, \arg(V^*(S_1)) = a_1, \arg(V^*(S_2)) = a_2.$$

In order to make $\pi^*(S_0) = a_2$, we need to find a value for x which allows:

[1] $T(S_0, a, S_0)V^*(S_0) + T(S_0, a, S_1)V^*(S_1) + T(S_0, a, S_2)V^*(S_2) + T(S_0, a, S_3)V^*(S_3))$

$$V^*(S_2) > V^*(S_1) and V^*(S_2) > 0$$

(2) Solution:

Supposing that the original value of $V^*(s_0), V^*(s_1), V^*(s_2), V^*(s_3)$ are 0. Since the reward is always greater than or equal to 0, so every new Value of $V^*(s_0), V^*(s_1), V^*(s_2), V^*(s_3)$ after each iteration is always greater than or equal to 0.

1) if $\gamma = 0$:

   We have always $V^*(S_2) = 1 > V^*(S_1) = 0$ no matter the value of x.

2) if $\gamma \in (0, 1)$:

   When $x = 0$:

$$V^*(S_1) = \gamma \cdot V^*(S_1) \Leftrightarrow (1 - \gamma) \cdot V^*(S_1) = 0$$

   Due to $0 < 1 - \gamma < 1$, $V^*(S_1) = 0$.
   When $y \in [0, 1]$:

$$V^*(S_2) \geq 1$$

$$=> V^*(S_1) = 0 < 1 < V^*(S_2)$$

(3) Conclusion:

$x = 0$ can cater to our need.

4 Question 4

(1) Analysis:

With the same analysis in Question 3, we need to find a value for y which allows:

$$V^*(S_2) < V^*(S_1)$$

According to Question 2, we have:

$$\begin{aligned} V^*(S_2) &= 1 + \gamma \cdot [(1 - y)V^*(S_0) + yV^*(S_3)] \\ &= 1 + \gamma \cdot [\gamma \cdot (1 - y)V^*(S_1) + (10\gamma + \gamma^2) \cdot y \cdot V^*(S_1)] \\ &> 1 \end{aligned}$$

While

$$\begin{aligned} V^*(S_1) &= \gamma \cdot [(1 - x)V^*(S_1) + xV^*(S_3)] \\ &< 1 \ (\ if \ \gamma < \frac{1}{(1 - x)V^*(S_1) + xV^*(S_3)} \ ) \end{aligned}$$

(2) Conclusion:

The y doesn't exist.

5 Question 5

Please read the code in the jupyter notebook: TP2.