ÉCOLE NATIONALE SUPÉRIEURE DE TECHNIQUES AVANCÉES

CSC_5RO11

TP Reinforcement Learning

Tiago LOPES REZENDE

Guilherme TROFINO

Palaiseau, France 2024

1 Question 1

The possible policies can be seen in the Table 1.

π	Current state	Input	Next State
π_1	S_0	a_1	S_1
π_2	S_0	a_2	S_2
π_3	S_1	a_0	S_1
π_4	S_1	a_0	S_3
π_5	S_3	a_0	S_0
π_6	S_2	a_0	S_0
π_7	S_2	a_0	S_3

TABLE 1 – State Transitions

2 Question 2

The transitions functions are as following :

$$T(S, a_0, S') = \begin{pmatrix} 0 & 0 & 0 & 0 \\ 0 & 1-x & 0 & x \\ 1-y & 0 & 0 & y \\ 1 & 0 & 0 & 0 \end{pmatrix}$$

$$T(S, a_1, S') = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

$$T(S, a_2, S') = \begin{pmatrix} 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$

The reward for every state is given by :

$$R(s) = \begin{cases} 10, & \text{for state } S_3 \\ 1, & \text{for state } S_2 \\ 0, & \text{otherwise} \end{cases}$$

The equation for each optimal state is given by :

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

So we can find the equation for the diferent states doing :

$$V^*(S0) = R(0) + \max_a \gamma T(S0, a1, S1)V^*(S1) + T(S0, a2, S2)V^*(S2)$$

$$V^*(S0) = \max_a \gamma \left[V^*(S_1) + V^*(S_2)\right]$$

$$V^*(S1) = R(1) + \max_a \gamma T(S1, a0, S1)V^*(S1) + T(S1, a0, S3)V^*(S3)$$

$$V^*(S1) = \max_a \gamma \left[(1-x)V^*(S_1) + xV^*(S_3)\right]$$

$$V^*(S2) = R(2) + \max_a \gamma T(S2, a0, S3)V^*(S3) + T(S2, a0, S0)V^*(S0)$$

$$V^*(S2) = 1 + \max_a \gamma \left[yV^*(S_3) + (1-y)V^*(S_0)\right]$$

$$V^*(S3) = R(3) + \max_a \gamma T(S3, a0, S0)V^*(S0)$$

$$V^*(S3) = 10 + \max_a \gamma \left[V^*(S_0)\right]$$

3 Question 3

With :

$$\pi^*(S) = \arg\max_a \sum_{S'} T(S, a, S')V^*(S')$$

It is possible to do :

$$\pi^*(S0) = \arg\max_a T(S0, a1, S1)V^*(S1) + T(S0, a2, S2)V^*(S2)$$

$$\pi^*(S0) = \arg\max_a V^*(S1) + V^*(S2)$$

So, for $\pi^*(S0) = a2$:

$$V^*(S1) < V^*(S2)$$

$$\max_a \gamma \Big[(1-x)V^*(S_1) + xV^*(S_3)] < 1 + \max_a \gamma \left[yV^*(S_3) + (1-y)V^*(S_0)\right]$$

If we chose x=0, we have :

$$\max_a \gamma \Big[V^*(S_1) < 1 + \max_a \gamma \left[yV^*(S_3) + (1-y)V^*(S_0)\right]$$

After multiples interaction we would have :

$$V^*(S_1) = \max_a \gamma^n \left[V^*(S_1)\right]$$

Since the factor γ is greater than zero and less than 1, it converges to zero as n tends to infinity. Since $V^*(S_2)$ has a reward of 1, it is always greater than zero and, therefore, greater than $V^*(S_1)$, fulfilling the requirements for an x that makes $\pi^*(S_0) = a_2$.

4 Question 4

With :

$$\pi^*(S) = \arg\max_a \sum_{S'} T(S, a, S')V^*(S')$$

It is possible to do :

$$\pi^*(S0) = \arg\max_a \left[T(S0, a1, S1)V^*(S1) + T(S0, a2, S2)V^*(S2)\right]$$

$$\pi^*(S0) = \arg\max_a \left[V^*(S1) + V^*(S2)\right]$$

So, for $\pi^*(S0) = a1 the \mathrm{V}^*(S1)$ should be :

$$V^*(S1) \geqslant V^*(S2)$$

$$\max_a \gamma \left[(1-x)V^*(S_1) + xV^*(S_3)\right] \geqslant 1 + \max_a \gamma \left[yV^*(S_3) + (1-y)V^*(S_0)\right]$$

5 Question 5

Algorithm was implemented in Python and can be found in this GitHub Repository under the TP5 folder.