Reinforcement Learning : Homework

Realized by :

VARGANYI Kylian

Table des matières

1 Question 1 : Enumerate all the possible policies

The policies for each state are defined as follows :

— $\pi(S0) = a1$, $\pi(S0) = a2$
— $\pi(S1) = a0$
— $\pi(S2) = a0$
— $\pi(S3) = a0$

For state $S0$, two actions are possible : $a1$ and $a2$. For the other states, only one action is possible.

2 Question 2 : Write the equation for each optimal value function for each state

The optimal value function equations $V^*(s)$ for each state are :

1. $V^*(S3)$: $S3$ has one action $a0$ that leads to $S0$ with a reward of 10.
The equation is :

$$V^*(S3) = 10 + \gamma V^*(S0)$$

2. $V^*(S2)$: The reward for $S2$ is 1, and the only available action $a0$ leads to either $S0$ (with probability $1 - y$) or $S3$ (with probability y).

$$V^*(S2) = 1 + \gamma \left[(1 - y) \cdot V^*(S0) + y \cdot V^*(S3)\right]$$

3. $V^*(S1)$: $S1$ follows the same idea applied for S2.

$$V^*(S1) = 0 + \gamma \left[(1 - x) \cdot V^*(S1) + x \cdot V^*(S3)\right]$$

4. $V^*(S0)$: For $S0$ there is two possible actions ($a1$ and $a2$). The optimal value function for $S0$ is therefore determined by the maximum between the two actions :
- With $a1$, we go to $S1$:

$$V(S0, a1) = 0 + \gamma V^*(S1)$$

- With $a2$, we go to $S2$:

$$V(S0, a2) = 0 + \gamma V^*(S2)$$

Combining these, we get :

$$V^*(S0) = \max\left(\gamma V^*(S1), \gamma V^*(S2)\right)$$

3 Question 3 : is there exist a value for x, that for all $\gamma \in [0,1)$ and $y \in [0,1]$, $\pi^*(S0) = a2$.

the optimal policy for a state s is defined by :

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

To ensure that $\pi^*(S0) = a2$, it is necessary that :

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

The values of $V^*(S1)$ and $V^*(S2)$ are given by :

—

$$V^*(S1) = \gamma\left[(1-x) \cdot V^*(S1) + x \cdot V^*(S3)\right]$$
$$\iff V^*(S1) = \frac{\gamma x V^*(S3)}{1-\gamma(1-x)}$$

—

$$V^*(S2) = 1 + \gamma\left[(1-y)V^*(S0) + yV^*(S3)\right]$$

By choosing a value of x close to 0, the value of $V^*(S1)$ becomes small. It depends directly to the discount factor γ and y. This maximizes that $V^*(S2) > V^*(S1)$, ensuring that the optimal action for $S0$ is $a2$.

4 Question 4 : is there exist a value for y, that for all $\gamma \in [0,1)$ and $x > 1$, $\pi^*(S0) = a1$.

To ensure that $\pi^*(S0) = a1$, we need to process like the previous question, but with a value of y.

To ensure that $\pi^*(S0) = a1$, it is necessary that :

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

By choosing a value of y close to 0, the value of $V^*(S2)$ becomes small. It depends directly to the discount factor γ and x. This maximizes that $V^*(S1) > V^*(S2)$, ensuring that the optimal action for $S0$ is $a1$.

5 Question 5 : Python Code

Using $x = y = 0.25$ and $\gamma = 0.9$, we need to calculate π^* and V^* for all states using value iteration. The termination rule is $|V_k(S) - V_{k-1}(S)| < 0.0001$.

Below the result that we obtained :

Optimal values V* for each state :

$$V*(S0) = 14.1854$$
$$V*(S1) = 15.7617$$
$$V*(S2) = 15.6977$$
$$V*(S3) = 22.7669$$

Optimal policy for each state :

$$pi*(S0) = a1$$
$$pi*(S1) = a0$$
$$pi*(S2) = a0$$
$$pi*(S3) = a0$$