ROB311

TP-5 Reinforcement Learning

XU Neng

3 November 2024

Table des matières

1 Question 1 **3**

2 Question 2 **3**

2.1 For s_0 3

2.2 For s_1 3

2.3 For s_2 4

2.4 For s_3 4

3 Question 3 **4**

4 Question 4 **5**

5 Question 5 **5**

5.1 Method 5

5.2 Code 6

5.3 Results 6

5.4 Optimal Policy 6

1 Question 1

Each state S can be assigned an action a from the available actions for that state. Since there are 4 states (s_0, s_1, s_2, s_3) and each state has multiple actions, the policies are the mappings of each state to one of its actions.

s_0 have possible actions a_1, a_2. s_1 have possible action a_0. s_2 have possible actions a_0. s_3 have possible actions a_0.

Thus, all possible policies are combinations of actions chosen for each state :

$$\pi(s_0) = \{a_1, a_2\}$$
$$\pi(s_1) = \{a_0\}$$
$$\pi(s_2) = \{a_0\}$$
$$\pi(s_3) = \{a_0\}$$

2 Question 2

The optimal value function for each state S is given by :

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

2.1 For s_0

As the agent can apply action a_1, a_2 at state s_0, we can write the optimal value function of s_0 as follow :

$$V^*(s_0) = R(s_0) + \max_a \left(\gamma \sum_{S'} T(s_0, a_1, S') V^*(S'),\ \gamma \sum_{S'} T(s_0, a_2, S') V^*(S') \right)$$

According to the reward and the transition function,

$$V^*(s_0) = \max_a \left(\gamma V^*(s_1),\ \gamma V^*(s_2) \right)$$

2.2 For s_1

As the agent can only apply action a_0 at state s_1, we can write the optimal value function of s_1 as follow :

$$V^*(s_1) = R(s_1) + \max_a \left(\gamma \sum_{S'} T(s_1, a_0, S') V^*(S') \right)$$

According to the reward and the transition function,

$$V^*(s_1) = \gamma \left[(1-x) V^*(s_1) + x \cdot V^*(s_3) \right]$$

2.3 For s_2

As the agent can only apply action a_0 at state s_2, we can write the optimal value function of s_2 as follow :

$$V^*(s_2) = R(s_2) + \max_a \left(\gamma \sum_{S'} T(s_2, a_0, S') V^*(S') \right)$$

According to the reward and the transition function,

$$V^*(s_2) = 1 + \gamma \left[(1-y) V^*(s_0) + y \cdot V^*(s_3) \right]$$

2.4 For s_3

As the agent can only apply action a_0 at state s_3, we can write the optimal value function of s_3 as follow :

$$V^*(s_3) = R(s_3) + \max_a \left(\gamma \sum_{S'} T(s_3, a_0, S') V^*(S') \right)$$

According to the reward and the transition function,

$$V^*(s_3) = 10 + \gamma \cdot V^*(s_0)$$

3 Question 3

Given :

$$\pi^*(s_0) = \max_a \left(\gamma V^*(s_1), \ \gamma V^*(s_2) \right) \quad \text{and} \quad \pi^*(s_0) = a_2$$

This means that the policy π^* at s_0 will choose action a_2 if, for all values of γ and y, the expected return from a_2 is greater than or equal to the expected return from a_1 at s_0. So we have $V^*(s_2) > V^*(s_1)$.

From the expressions caculated in Q2,

$$V^*(s_1) = \gamma \left[(1-x) V^*(s_1) + x \cdot V^*(s_3) \right]$$

$$V^*(s_2) = 1 + \gamma \left[(1-y) V^*(s_0) + y \cdot V^*(s_3) \right]$$

the value of x must be chosen such that $V^*(s_1)$ does not grow larger than $V^*(s_2)$, regardless of γ and y values.

When $x = 0$, the expression for $V^*(s_1)$ is

$$V^*(s_1) = \gamma \left[(1-0) V^*(s_1) + 0 \cdot V^*(s_3) \right] = \gamma V^*(s_1)$$

This implies that $V^*(s_1) = 0$.

For $V^*(s_2)$:

$$V^*(s_2) = 1 + \gamma \left[(1-y) V^*(s_0) + y \cdot V^*(s_3) \right]$$

Since $V^*(s_1) = 0$, the value $V^*(s_2)$ will generally be positive due to the constant term 1 and the potential contributions from $V^*(s_0)$ and $V^*(s_3)$.

Thus, setting $x = 0$ satisfies the condition for $\pi^*(s_0) = a_2$ for all $\gamma \in [0, 1)$ and $y \in [0, 1]$.

4 Question 4

Given :

$$\pi^*(s_0) = \max_a \big(\gamma V^*(s_1),\ \gamma V^*(s_2)\big) \quad \text{and} \quad \pi^*(s_0) = a_1$$

This means that the policy π^* at s_0 will choose action a_1 if, for all values of γ and y, the expected return from a_1 is greater than or equal to the expected return from a_2 at s_0. So we have $V^*(s_1) > V^*(s_2)$.

As in the Q3, when x approaches zero, $V^*(s_1)$ will converge to zero.

For s_2, the value function is :

$$V^*(s_2) = 1 + \gamma\left[(1-y)V^*(s_0) + y \cdot V^*(s_3)\right]$$

Therefore, regardless of the value of x, $V^*(s_2)$ will be positive and > 1.

To sum up, when $x \to 0$, we observe $V^*(s_1) \to 0$ and $V^*(s_2) > 0$. This implies that $V^*(s_2)$ will always be greater than $V^*(s_1)$. We cannot find a y to satisfy $\pi^*(s_0) = a_1$.

5 Question 5

5.1 Method

To calculate the optimal policy π^* and value function V^* for each state s_0, s_1, s_2, and s_3, we follow these steps :

1. **Initialization** : We start by assigning an initial value of 0 to each state V_0, V_1, V_2, and V_3.
2. **Iteration** : The code then enters a loop, repeating the following updates to each state's value :
   — For s_0, we calculate :
   $$V^*(s_0) = \max(\gamma V^*(s_1), \gamma V^*(s_2))$$
   — For s_1, we calculate :
   $$V^*(s_1) = \gamma\left[(1-x)V^*(s_1) + x \cdot V^*(s_3)\right]$$
   — For s_2, we calculate :
   $$V^*(s_2) = 1 + \gamma\left[(1-y)V^*(s_0) + y \cdot V^*(s_3)\right]$$
   — For s_3, we calculate :
   $$V^*(s_3) = 10 + \gamma \cdot V^*(s_0)$$
3. **Updating Values** : After calculating the new values, we update V_0, V_1, V_2, and V_3 with the new results.
4. **Convergence** : The loop iterates until the values stabilize (The difference between the two approaching iterations is less than 0.0001), providing the final values for $V^*(s_0), V^*(s_1), V^*(s_2)$, and $V^*(s_3)$.

5.2 Code

The codes are as follows.

```
# Initial guesses for V* values for each state
V0, V1, V2, V3 = 0, 0, 0, 0

# Iterative approach
for i in range(1000):
    V0_new = max(gamma * V1, gamma * V2)
    V1_new = gamma * ((1 - x) * V1 + x * V3)
    V2_new = 1 + gamma * ((1 - y) * V0 + y * V3)
    V3_new = 10 + gamma * V0

    if abs(V0_new-V0)<0.0001 and abs(V1_new-V1)<0.0001
    and abs(V2_new-V2)<0.0001 and abs(V3_new-V3)<0.0001:
        print("Finish in iteration =", i)
        break

    V0, V1, V2, V3 = V0_new, V1_new, V2_new, V3_new
```

5.3 Results

The calculated optimal value function V^* for each state is :

$$V^*(s_0) \approx 14.18$$
$$V^*(s_1) \approx 15.76$$
$$V^*(s_2) \approx 15.70$$
$$V^*(s_3) \approx 22.77$$

5.4 Optimal Policy

To determine the optimal policy π^* for s_0, we choose the action that maximizes the value.

$$\pi^*(s_0) = \max_a \left(\gamma V^*(s_1),\ \gamma V^*(s_2)\right)$$

Since $V^*(s_1) \approx 15.76$ and $V^*(s_2) \approx 15.70$, the optimal action at s_0 would correspond to the action that leads to $V^*(s_1)$, as it provides a higher value. For s_1, s_2, and s_3, the only available action is a_0. The policy for these states is fixed.

Thus, the optimal policy π^* is :

— $\pi^*(s_0) = a_1$

— $\pi^*(s_1) = a_0$

— $\pi^*(s_2) = a_0$

— $\pi^*(s_3) = a_0$