Q-Learning Example: Solved by Hand

Instructor: Ahmed Métwalli

Problem Setup: A Simple Gridworld

We have a simple 3×3 grid where the agent starts in the top-left corner and needs to reach the goal at the bottom-right corner. The agent can move in four directions: up, down, left, or right. Every step incurs a penalty of -1, except for reaching the goal, which provides a reward of +10.

$(0,0)$	$(0,1)$	$(0,2)$
$(1,0)$	$(1,1)$	$(1,2)$
$(2,0)$	$(2,1)$	$(2,2)$ [Goal]

Rewards and Setup

- Reward for each step: -1
- Reward for reaching the goal: $+10$
- Start state: $(0,0)$
- Goal state: $(2,2)$
- Learning rate (α): 0.5
- Discount factor (γ): 0.9
- Q-table initialization: All entries are initialized to zero.

Q-Learning Update Formula

The Q-learning update formula is as follows:

$$Q(s,a) \leftarrow Q(s,a) + \alpha \left[r + \gamma \max_{a'} Q(s',a') - Q(s,a) \right]$$

Where:

- $Q(s,a)$ is the Q-value for state s and action a.
- α is the learning rate.
- r is the reward received after taking action a in state s.
- γ is the discount factor, which determines how much future rewards are valued.
- $\max_{a'} Q(s',a')$ is the maximum Q-value for the next state s'.

Q-Table (Before the Last 3 Iterations)

The Q-table is partially filled and given below. Your task is to complete the last three iterations.

State	Up	Down	Left	Right
(0,0)	0.0	0.0	0.0	-0.5
(0,1)	0.0	0.0	0.0	-0.5
(0,2)	0.0	**-0.75**	0.0	0.0
(1,0)	0.0	0.0	0.0	0.0
(1,1)	0.0	0.0	0.0	0.0
(1,2)	0.0	**??**	0.0	**??**
(2,0)	0.0	0.0	0.0	0.0
(2,1)	0.0	0.0	0.0	0.0
(2,2)	-	-	-	- (Goal)

Your Task

You are asked to complete the next 3 iterations of the Q-learning process by applying the update formula.

Iteration 1: - State: (0,2) - Action: Down - Next state: (1,2) - Reward: -1
Update the Q-value for $Q(0, 2, \text{Down})$ using the Q-learning formula.

$$Q(0, 2, \text{Down}) \leftarrow Q(0, 2, \text{Down}) + \alpha \left[r + \gamma \max_{a'} Q(1, 2, a') - Q(0, 2, \text{Down}) \right]$$

$$Q(0, 2, \text{Down}) \leftarrow -0.5 + 0.5 \left[-1 + 0.9 \times 0 - (-0.5) \right]$$

$$Q(0, 2, \text{Down}) \leftarrow -0.5 + 0.5 \times (-0.5) = -0.75$$

Iteration 2: - State: (1,2) - Action: Down - Next state: (2,2) (Goal state) - Reward: $+10$
Update the Q-value for $Q(1, 2, \text{Down})$.

$$Q(1, 2, \text{Down}) \leftarrow Q(1, 2, \text{Down}) + \alpha \left[r + \gamma \max_{a'} Q(2, 2, a') - Q(1, 2, \text{Down}) \right]$$

$$Q(1, 2, \text{Down}) \leftarrow 0 + 0.5 \left[10 + 0.9 \times 0 - 0 \right]$$

$$Q(1, 2, \text{Down}) = 0 + 0.5 \times 10 = 5.0$$

Iteration 3: - State: (1,2) - Action: Right - Next state: (2,2) - Reward: $+10$
Update the Q-value for $Q(1, 2, \text{Right})$.

$$Q(1, 2, \text{Right}) \leftarrow Q(1, 2, \text{Right}) + \alpha \left[r + \gamma \max_{a'} Q(2, 2, a') - Q(1, 2, \text{Right}) \right]$$

$$Q(1, 2, \text{Right}) \leftarrow 0 + 0.5 \left[10 + 0.9 \times 0 - 0 \right]$$

$$Q(1, 2, \text{Right}) = 0 + 0.5 \times 10 = 5.0$$

Final Updated Q-Table

After completing these iterations, the Q-table becomes:

State	Up	Down	Left	Right
(0,0)	0.0	0.0	0.0	-0.5
(0,1)	0.0	0.0	0.0	-0.5
(0,2)	0.0	-0.75	0.0	0.0
(1,0)	0.0	0.0	0.0	0.0
(1,1)	0.0	0.0	0.0	0.0
(1,2)	0.0	5.0	0.0	5.0
(2,0)	0.0	0.0	0.0	0.0
(2,1)	0.0	0.0	0.0	0.0
(2,2)	-	-	-	- (Goal)

Conclusion

Through this exercise, you have seen how the Q-learning update formula is applied step-by-step. Continue practicing by filling in Q-tables for different iterations to fully grasp how the agent learns from its environment.