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' across all actions a'.

Initial Q-Table

The Q-table is initialized with all zeros, and we will update it based on the agent's actions.

State	Up	Down	Left	Right
(0,0)	0.0	0.0	0.0	0.0
(0,1)	0.0	0.0	0.0	0.0
(0,2)	0.0	0.0	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	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)

Iterations

Iteration 1: - State: (0,0) - Action: Right - Next state: (0,1) - Reward: -1

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

Substituting values:

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

$$Q(0,0,\text{Right}) = -0.5$$

Iteration 2: - State: (0,1) - Action: Right - Next state: (0,2) - Reward: -1

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

Substituting values:

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

$$Q(0,1,\text{Right}) = -0.5$$

Iteration 3: - State: (0,2) - Action: Down - Next state: (1,2) - Reward: -1

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

Substituting values:

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

$$Q(0,2,\text{Down}) = -0.5$$

Iteration 4: - State: (1,2) - Action: Down - Next state: (2,2) (Goal state) - Reward: $+10$

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

Since reaching the goal yields no future rewards, substitute:

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

$$Q(1,2,\text{Down}) = 5.0$$

Iteration 5 (Modified): - State: (0,0) - Action: Down - Next state: (1,0) - Reward: -1

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

Substituting values:

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

$$Q(0,0,\text{Down}) = -0.5$$

Iteration 6: - State: (1,0) - Action: Right - Next state: (1,1) - Reward: -1

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

Substituting values:

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

$$Q(1,0,\text{Right}) = -0.5$$

Iteration 7: - State: (1,1) - Action: Right - Next state: (1,2) - Reward: -1

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

Substituting values:

$$Q(1,1,\text{Right}) \leftarrow 0 + 0.5 \left[-1 + 0.9 \times 5.0 - 0 \right]$$

$$Q(1,1,\text{Right}) = 1.7$$

Iteration 8: - State: (1,1) - Action: Down - Next state: (2,1) - Reward: -1

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

Substituting values:

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

$$Q(1,1,\text{Down}) = -0.5$$

Updated Q-Table

After completing these iterations, the Q-table becomes:

State	Up	Down	Left	Right
(0,0)	0.0	-0.5	0.0	-0.5
(0,1)	0.0	0.0	0.0	-0.5
(0,2)	0.0	-0.5	0.0	0.0
(1,0)	0.0	0.0	0.0	-0.5
(1,1)	0.0	-0.5	0.0	1.7
(1,2)	0.0	5.0	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)

Conclusion

Through these iterations, you can observe how the Q-learning update formula is applied step-by-step and how the agent learns the optimal policy by maximizing rewards and learning from the environment. Continue experimenting with different paths to understand how Q-learning converges.