

Prompt:

I have a toy problem involving reinforcement learning. In particular, q-learning.

My environment is very simple - a linear line with 5 stops. from left to right, 1 is a hole, 2 is empty, 3 is the starting point for the agent, 4 is empty and 5 is the destination. -1 point if you hit the hole and +1 point if you arrive at the destination.

Give me a sequence of q-table updates so that i can trace the agent's progress and understand the q-learning.

This is a classic introductory problem for Q-learning! To trace the agent's progress, let's establish the environment and then walk through a few episodes, showing the **Q-table updates** step-by-step.

Environment Setup

State	Description	Reward
1	Hole (Terminal)	-1
2	Empty	0
3	Start	0
4	Empty	0
5	Destination (Terminal)	+1

- **States (S):** {1,2,3,4,5}
- **Actions (A):** {Left (-1),Right (+1)}. An agent at state s moves to s+a.
- **Initial State:** s0=3
- **Learning Rate (α):** 0.5 (A common value for initial examples)
- **Discount Factor (γ):** 0.9 (A common value)

- **Exploration Strategy:** ϵ -greedy. We'll assume the agent explores a few times initially.
- Q-Learning Update Rule:

$$Q(s,a) \leftarrow Q(s,a) + \alpha [R_t + 1 + \gamma \max_{a'} Q(s',a') - Q(s,a)]$$

- **Initial Q-Table:** All values are zero.

Q(s,a)	Left (-1)	Right (+1)
1 (Hole)	0	0
2 (Empty)	0	0
3 (Start)	0	0
4 (Empty)	0	0
5 (Dest.)	0	0

Q-Table Update Sequence (Trace)

We'll trace **three episodes** to demonstrate how the values converge.

Episode 1: Initial Exploration (Going Left)

The agent starts at $s=3$. Let's assume the agent **explores** by choosing a random action, $a=Left$.

Step 1: $s=3$ Left $s'=2$

- **Action Taken:** Left ($a=-1$).
- **New State:** $s'=2$.
- **Reward:** $R=0$ (non-terminal state).

- **Max Q for new state (s'=2):** $\max_a Q(2,a) = \max(0,0) = 0$.
- Update:

$$Q(3, \text{Left}) \leftarrow 0 + 0.5[0 + 0.9 \cdot 0 - 0] = 0$$

(No change because all initial Q-values are 0)

Step 2: s=2 Left s'=1 (Hole)

The agent is now at s=2. Let's assume it **explores** again and chooses a=Left.

- **Action Taken:** Left (a=-1).
- **New State:** s'=1 (**Terminal: Hole**).
- **Reward:** R=-1.
- **Max Q for new state (s'=1): 0** (Since s'=1 is terminal, the future discounted reward is 0).
- Update:

$$Q(2, \text{Left}) \leftarrow 0 + 0.5[-1 + 0.9 \cdot 0 - 0] = -0.5$$

Resulting Q-Table after Episode 1:

Q(s,a)	Left (-1)	Right (+1)
1 (Hole)	0	0
2 (Empty)	-0.5	0
3 (Start)	0	0
4 (Empty)	0	0
5 (Dest.)	0	0

Episode 2: Reaching the Destination

The agent starts at $s=3$. This time, let's assume the agent **explores** by choosing $a=Right$ (perhaps a better initial exploration).

Step 1: $s=3$ Right $s'=4$

- **Action Taken:** Right ($a=+1$).
- **New State:** $s'=4$.
- **Reward:** $R=0$.
- **Max Q for new state ($s'=4$):** $\max_a Q(4,a)=\max(0,0)=0$.
- **Update:**

$$Q(3,Right) \leftarrow 0 + 0.5[0 + 0.9 \cdot 0 - 0] = 0$$

Step 2: $s=4$ Right $s'=5$ (Destination)

The agent is at $s=4$. Let's assume it **explores** again and chooses $a=Right$.

- **Action Taken:** Right ($a=+1$).
- **New State:** $s'=5$ (**Terminal: Destination**).
- **Reward:** $R=+1$.
- **Max Q for new state ($s'=5$):** **0** (Since $s'=5$ is terminal, the future discounted reward is 0).
- **Update:**

$$Q(4,Right) \leftarrow 0 + 0.5[+1 + 0.9 \cdot 0 - 0] = +0.5$$

Resulting Q-Table after Episode 2:

$Q(s,a)$	Left (-1)	Right (+1)
1 (Hole)	0	0
2 (Empty)	-0.5	0
3 (Start)	0	0

4 (Empty)	0	+0.5
5 (Dest.)	0	0

Episode 3: Propagating the Reward (Exploitation)

The agent starts at $s=3$. Now the Q-table is not all zeros, so the agent can **exploit** (choose the action with the max Q-value).

Step 1: $s=3$ Right $s'=4$

- **Action Choice:** $\max_a Q(3,a) = \max(0,0) = 0$. We'll assume a tie-breaker favors **Right**.
- **Action Taken:** Right ($a=+1$).
- **New State:** $s'=4$.
- **Reward:** $R=0$.
- **Max Q for new state ($s'=4$):** The agent looks at the current Q-table for state 4:
 $\max_{a'} Q(4,a') = \max(0,0.5) = 0.5$.
- **Update:**

$$Q(3, \text{Right}) \leftarrow 0 + 0.5[0 + 0.9 \cdot 0.5 - 0] = 0.5 \cdot 0.45 = 0.225$$

Step 2: $s=4$ Right $s'=5$ (Destination)

The agent is at $s=4$.

- **Action Choice:** $\max_a Q(4,a) = \max(0,0.5) = 0.5$. The agent chooses **Right**.
- **Action Taken:** Right ($a=+1$).
- **New State:** $s'=5$ (Terminal).
- **Reward:** $R=+1$.
- **Max Q for new state ($s'=5$):** 0.
- **Update:**

$$Q(4, \text{Right}) \leftarrow 0.5 + 0.5[+1 + 0.9 \cdot 0 - 0.5] = 0.5 + 0.5 \cdot 0.5 = 0.5 + 0.25 = 0.75$$

Resulting Q-Table after Episode 3:

Q(s,a)	Left (-1)	Right (+1)
1 (Hole)	0	0
2 (Empty)	-0.5	0
3 (Start)	0	0.225
4 (Empty)	0	0.75
5 (Dest.)	0	0

Conclusion of Trace

After just three episodes, the Q-learning algorithm has started to learn the path:

- The negative reward from the hole has made $Q(2, \text{Left})$ negative.
- The positive reward from the destination has **propagated backward** from state 5, to $Q(4, \text{Right})$, and then to $Q(3, \text{Right})$.
- The **optimal policy** for states 3 and 4 is now to go **Right** (the action with the higher Q-value).

If you continued the training, the values would increase and converge toward their optimal discounted return:

- $Q(4, \text{Right})$ would approach 1.0 (immediate reward of +1).
- $Q(3, \text{Right})$ would approach $\gamma \cdot 1.0 = 0.9$ (discounted reward of +1 one step away).
- $Q(2, \text{Right})$ would approach $\gamma^2 \cdot 1.0 = 0.81$, and so on.