Q-Learning Homework

Q-Learning in a Deterministic 4-State World

Environment Setup

- **States:** There are four states arranged linearly:

State	Description
S1	Leftmost state (absorbing)
S2	Second state from the left
S3	Third state from the left
S4	Rightmost state

- **Actions:**
  - **S1:** No actions (absorbing state).
  - **S2:**
    - **Move Left:** Transitions to S1.
    - **Move Right:** Transitions to S3.
  - **S3:**
    - **Move Left:** Transitions to S2.
    - **Move Right:** Transitions to S4.
  - **S4:**
    - **Move Left:** Transitions to S3.
- **Rewards:**
  - Entering **S1:** Reward = 10 (absorbing state).
  - Entering **S2, S3, S4:** Reward = 0.
- **Discount Factor:** $\gamma = 0.8$
- **Initial Q-Values:** All Q-values are initialized to 0.

Q-Learning Update Rule

The Q-Learning update rule for each state-action pair (s, a) is given by:

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

Where:

- $R(s,a)$ is the immediate reward after taking action a in state s.
- s' is the next state after taking action a.
- α is the learning rate (not specified here, assuming convergence).

Since the environment is deterministic and we aim for the optimal Q-values, we can directly compute them using the Bellman optimality equations.

Bellman Optimality Equations

For each non-absorbing state, the optimal Q-values are:

- **S2:**

$$Q^*(S2, \text{Left}) = R(S2, \text{Left}) + \gamma V(S1) = 10 + 0.8 \times 10 = 18$$

$$Q^*(S2, \text{Right}) = R(S2, \text{Right}) + \gamma V(S3) = 0 + 0.8 \times V(S3)$$

- **S3:**

$$Q^*(S3, \text{Left}) = R(S3, \text{Left}) + \gamma V(S2) = 0 + 0.8 \times V(S2)$$

$$Q^*(S3, \text{Right}) = R(S3, \text{Right}) + \gamma V(S4) = 0 + 0.8 \times V(S4)$$

- **S4:**

$$Q^*(S4, \text{Left}) = R(S4, \text{Left}) + \gamma V(S3) = 0 + 0.8 \times V(S3)$$

The value function $V(s)$ for each state is defined as:

$$V(s) = \max_a Q^*(s,a)$$

Solving for Optimal Q-Values

1. **From State S2:**

$$V(S2) = \max(Q^*(S2, \text{Left}), Q^*(S2, \text{Right})) = \max(18, 0.8 \times V(S3))$$

2. **From State S3:**

$$V(S3) = \max(Q^*(S3, \text{Left}), Q^*(S3, \text{Right})) = \max(0.8 \times V(S2), 0.8 \times V(S4))$$

3. **From State S4:**

$$V(S4) = Q^*(S4, \text{Left}) = 0.8 \times V(S3)$$

Assuming:

- $V(S1) = 10$ (absorbing state).

Solving the Equations:

- From **S4**:

$$V(S4) = 0.8 \times V(S3)$$

- From **S3**:

$$V(S3) = \max(0.8 \times V(S2), 0.8 \times V(S4)) = \max(0.8 \times V(S2), 0.8 \times 0.8 \times V(S3))$$

  Let's denote $V(S3) = \max(0.8V(S2), 0.64V(S3))$. Since $V(S3) \geq 0$, this simplifies to:

$$V(S3) = 0.8V(S2)$$

- From **S2**:

$$V(S2) = \max(18, 0.8 \times V(S3)) = \max(18, 0.8 \times 0.8V(S2)) = \max(18, 0.64V(S2))$$

  To satisfy the equation:

$$18 \geq 0.64V(S2) \Rightarrow V(S2) \leq \frac{18}{0.64} \approx 28.125$$

  Since $V(S2) = \max(18, 0.64V(S2))$, and assuming $V(S2) = 18$, we get:

$$V(S3) = 0.8 \times 18 = 14.4$$

$$V(S4) = 0.8 \times 14.4 = 11.52$$

Final Q-Values:

State	Action	Q*(State, Action)
S2	Left	18
S2	Right	$0.8 \times 14.4 = 11.52$
S3	Left	$0.8 \times 18 = 14.4$
S3	Right	$0.8 \times 11.52 = 9.216$
S4	Left	$0.8 \times 14.4 = 11.52$

Optimal Q-Values Summary

State	Action	Q*(State, Action)
S2	Left	18
S2	Right	11.52
S3	Left	14.4
S3	Right	9.216
S4	Left	11.52

Optimal Policy

The optimal policy selects the action with the highest Q* value in each state.

State	Optimal Action
S1	*No Action (Terminal)*
S2	**Move Left**
S3	**Move Left**
S4	**Move Left**

Description:

- **S2:** Moving left leads directly to the absorbing state **S1** with a high reward.
- **S3:** Moving left transitions to **S2**, which can then lead to **S1**.
- **S4:** Only action available is to move left towards **S3**.

Final Optimal Q-Values and Policy

The final optimal Q-values and the corresponding optimal policy are as follows:

Q-Values Table

State	Action	Q*(State, Action)
S2	Left	18
S2	Right	11.52
S3	Left	14.4
S3	Right	9.216
S4	Left	11.52

Optimal Policy Table

State	Optimal Action
S1	*No Action (Terminal)*
S2	**Move Left**
S3	**Move Left**
S4	**Move Left**

Interpretation:

- The optimal policy directs the agent to always move left from states **S2**, **S3**, and **S4**, effectively guiding it towards the absorbing state **S1** to receive the maximum possible reward of 10.