Example: Bellman Equation for Q-Function

Consider a simple grid world with three states (S1, S2, and the goal state G) and two actions (left and right). The agent receives a reward of -1 for each step and a reward of $+10$ for reaching the goal state. The discount factor γ is set to 0.9.

The Bellman equation for the Q-function $Q(s, a)$ of a state s and action a is given by:

$$Q(s,a) = R(s,a) + \gamma \sum_{s'} P(s'|s,a) \max_{a'} Q(s',a')$$

where:

- $R(s, a)$ is the immediate reward received after taking action a in state s,
- γ is the discount factor,
- $P(s'|s, a)$ is the probability of transitioning to state s' from state s after taking action a.

Let's calculate the Q-values for each state-action pair:

1. **Initialization**: Start with initial Q-values of zero for all state-action pairs.

State	Action (left)	Action (right)
$S1$	0	0
$S2$	0	0
G	0	0

2. **Update Q-Value for State S1 and Action Left**:

$$\begin{aligned} R(S1, \text{left}) &= -1 \quad \text{(immediate reward for moving left)} \\ \max_{a'} Q(S2, a') &= \max(0, 0) = 0 \quad \text{(maximum Q-value for the next state S2)} \\ Q(S1, \text{left}) &= -1 + 0.9 \times 0 = -1 \end{aligned}$$

3. **Update Q-Value for State S1 and Action Right**:

$$\begin{aligned} R(S1, \text{right}) &= -1 \quad \text{(immediate reward for moving right)} \\ \max_{a'} Q(G, a') &= \max(0, 0) = 0 \quad \text{(maximum Q-value for the next state G)} \\ Q(S1, \text{right}) &= -1 + 0.9 \times 0 = -1 \end{aligned}$$

4. **Update Q-Value for State S2 and Action Left**:

$$\begin{aligned} R(S2, \text{left}) &= -1 \quad \text{(immediate reward for moving left)} \\ \max_{a'} Q(G, a') &= \max(0, 0) = 0 \quad \text{(maximum Q-value for the next state G)} \\ Q(S2, \text{left}) &= -1 + 0.9 \times 0 = -1 \end{aligned}$$

5. **Update Q-Value for State S2 and Action Right**:

$$R(S2, \text{right}) = -1 \quad \text{(immediate reward for moving right)}$$
$$\max_{a'} Q(G, a') = \max(0, 0) = 0 \quad \text{(maximum Q-value for the next state G)}$$
$$Q(S2, \text{right}) = -1 + 0.9 \times 0 = -1$$

6. **Update Q-Value for State G and Actions**:

$$\begin{aligned} R(G, \text{left}) &= 10 \quad \text{(reward for reaching the goal state)} \\ R(G, \text{right}) &= 10 \quad \text{(reward for reaching the goal state)} \\ Q(G, \text{left}) &= 10 + 0.9 \times 0 = 10 \\ Q(G, \text{right}) &= 10 + 0.9 \times 0 = 10 \end{aligned}$$

$$V(s) = \frac{1}{N(s)} \sum_{i=1}^{N(s)} G_i$$

After updating all state-action pairs, the Q-values are as follows:

State	Action (left)	Action (right)
$S1$	-1	-1
$S2$	-1	-1
G	10	10

The final Q-values represent the expected cumulative rewards the agent can achieve from each state-action pair following an optimal policy.