Example 2: The Bellman equation of the Q function

Consider a simple grid world where an agent can move left, right, up, or down. The grid has a reward of -1 for each step, and the agent receives 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-action pair (s, a) in this grid world is:

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

where:

- s is the current state,
- a is the action taken,
- $R(s, a)$ is the immediate reward for 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,
- $\max_{a'} Q(s', a')$ is the maximum Q-value for the next state s' over all possible actions a'.

Let's consider a specific example where the agent is in state S and can take actions to move left, right, up, or down. The rewards for each action are as follows:

- Moving left or right: -1
- Moving up or down: -1

The goal state (state G) has a reward of $+10$. Since the grid is deterministic, the agent moves to the desired state with probability 1.

We can calculate the Q-values for each state-action pair using the Bellman equation and the given rewards. Let's start with the initial Q-values:

$$\begin{aligned} Q(S, \text{left}) &= 0 \\ Q(S, \text{right}) &= 0 \\ Q(S, \text{up}) &= 0 \\ Q(S, \text{down}) &= 0 \end{aligned}$$

To update the Q-values, we apply the Bellman equation for each state-action pair. For example, to update $Q(S, \text{left})$:

$$\begin{aligned} Q(S, \text{left}) &= -1 + 0.9 \times \max(Q(\text{next state}, \text{all actions})) \\ &= -1 + 0.9 \times \max(Q(S, \text{left}), Q(S, \text{right}), Q(S, \text{up}), Q(S, \text{down})) \\ &= -1 + 0.9 \times \max(0, 0, 0, 0) \\ &= -1 \end{aligned}$$

Similarly, we can update $Q(S, \text{right})$, $Q(S, \text{up})$, and $Q(S, \text{down})$. After updating, the Q-values become:

$$\begin{aligned} Q(S, \text{left}) &= -1 \\ Q(S, \text{right}) &= -1 \\ Q(S, \text{up}) &= -1 \\ Q(S, \text{down}) &= -1 \end{aligned}$$

These updated Q-values reflect the expected cumulative rewards the agent can achieve from each state-action pair following an optimal policy in the grid world environment.