State-action value function (Q-Function)

State-action value function definition

The Q function, denoted as $Q(s,a)$, computes the expected return of **taking action** a in **state** s and then **behaving optimally thereafter**.

> The definition may seem **circular**, as knowing the **optimal policy** would eliminate the need to compute $Q(s,a)$, but this will be resolved in later discussions.

When creating the policy:

- The best possible return from state s is $\max_a Q(s,a)$.
- The best possible action in state s is the action a that gives $\max_a Q(s,a)$.

$$\pi(s) = \arg\max_a Q(s,a)$$

> Sometimes, instead of Q, we can see Q^* - this is the optimal Q function.

Example:

- For state 2, taking the action to go right results in a Q value of 12.5, while going left yields a higher return of 50.
- In state 4, going left also results in a Q value of 12.5, indicating that this action is **optimal**.

Bellman Equation

The **Bellman equation** helps compute $Q(s,a)$, which represents the expected return from taking action a in state s and then acting optimally thereafter.

If $R(s)$ is the reward for the current state, γ is the discount factor, and

- s is the current state.
- a is the current action
- s' is the new state after taking action a,
- a' is the action taking in state s'.

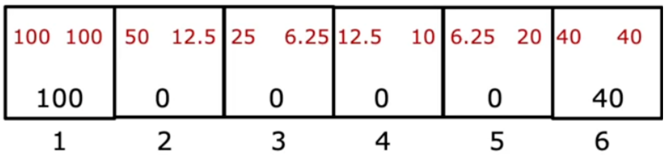
The **Bellman equation** is defined as:

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

In other words,

$$Q(s,a) = \text{Reward you get right away}[R(s)] + \text{Return from behaving optimally starting from state } s'$$

Examples of Applying the Bellman Equation



- For $Q(2, \rightarrow)$,
  - $s = 2, a = \rightarrow, s' = 3$
  - The calculation involves the reward of state 2 (which is 0) and the maximum Q value from state 3, resulting in $Q(2, \rightarrow) = 12.5$.

$$Q(2, \rightarrow) = R(2) + 0.5 \max_a Q(3, a') = 0 + 0.5 \times 25 = 12.5$$

- Similarly, for $Q(4, \leftarrow)$, the calculation also results in $Q(4, \leftarrow) = 12.5$.
  - $s = 4, a = \leftarrow, s' = 3$

$$Q(4, \leftarrow) = R(4) + 0.5 \max_a Q(3, a') = 0 + 0.5 \times 25 = 12.5$$

Explanation of Bellman equation:

$$\begin{aligned} Q(4, \leftarrow) &= 0 + (0.5) \cdot 0 + (0.5)^2 \cdot 0 + (0.5)^3 \cdot 100 \\ &= R(4) + (0.5) \left[0 + (0.5) \cdot 0 + (0.5)^2 \cdot 100 \right] \\ &= R(4) + (0.5) \max_a Q(3, a') \end{aligned}$$

Random (stochastic) environment

In stochastic environments, **actions may NOT always lead to the expected outcomes** due to **random factors**, such as *terrain conditions* affecting the rover's movement.

- For instance, if the rover is commanded to go left, there is a 90% chance it will succeed, but a 10% chance it may slip and go right instead.

Therefore, we need calculate the average reward of total actions to avoid expected outcomes due to random factors.

$$\text{Expected Return} = \text{Average}(R_1 + \gamma R_2 + \gamma^2 R_3 + \gamma^3 R_4 + \dots)$$

The word Expected is somehow related the word Average. Then, finally, the equation is described as:

$$Q(s,a) = R(s) + \gamma \mathbb{E}\left[\max_{a'} Q(s', a') \right]$$

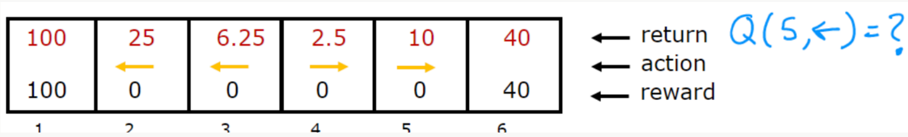
Which of the following accurately describes the state-action value function $Q(s,a)$?

- It is the return if you start from state s, take action a (once), then behave optimally after that. 👍
- It is the return if you start from state s and repeatedly take action a.
- It is the return if you start from state s and behave optimally.
- It is the immediate reward if you start from state s and take action a (once).

You are controlling a robot that has 3 actions: ← (left), → (right) and STOP. From a given state s, you have computed $Q(s, \leftarrow) = -10, Q(s, \rightarrow) = -20, Q(s, STOP) = 0$.
What is the optimal action to take in state s?

- STOP 👍👍
- ← (left)
- → (right)
- Impossible to tell

For this problem, $\gamma = 0.25$ The diagram below shows the return and the optimal action from each state. Please compute $Q(5, \leftarrow)$.



- 0.625 👍👍
- 0.391
- 1.25
- 2.5

<u>*Explain:*</u> Yes, we get 0 reward in state 5. Then $0 * 0.25$ discounted reward in state 4, since we moved left for our action. Now we behave optimally starting from state 4 onwards. So, we move right to state 5 from state 4 and receive $0 * 0.25^2$ discounted reward. Finally, we move right in state 5 to state 6 to receive a discounted reward of $40 * 0.25^3$. Adding these together we get 0.625