Mathematical Foundations of Reinforcement Learning

Chapter 2: State Values and Bellman Equation

Minseok Seo
Artificial Intelligence Graduate School
Gwangju Institute of Science and Technology (GIST)

July 17, 2025

Overview

Motivating example 1: Why are returns important?

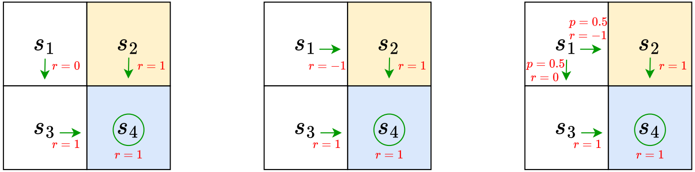

Figure: Examples for demonstrating the importance of returns.

- Following the first policy, discounted return is

$$\begin{aligned} R_1 &= 0 + \gamma 1 + \gamma^2 1 + \cdots \\ &= \gamma(1 + \gamma + \gamma^2 + \cdots) \\ &= \frac{\gamma}{1-\gamma} \end{aligned}$$

where $\gamma \in (0,1)$ is the discount rate.

Motivating example 1: Why are returns important?

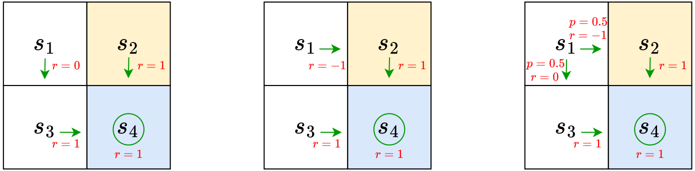

Figure: Examples for demonstrating the importance of returns.

- Following the second policy, discounted return is

$$\begin{aligned} R_2 &= -1 + \gamma 1 + \gamma^2 1 + \cdots \\ &= -1 + \gamma(1 + \gamma + \gamma^2 + \cdots) \\ &= -1 + \frac{\gamma}{1-\gamma} \end{aligned}$$

Motivating example 1: Why are returns important?

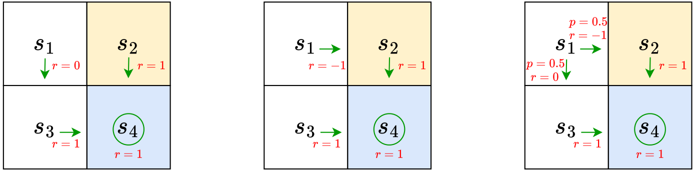

Figure: Examples for demonstrating the importance of returns.

- Following the third policy, discounted return is

$$R_3 = 0.5\left(-1+\frac{\gamma}{1-\gamma}\right)+0.5\left(\frac{\gamma}{1-\gamma}\right)$$
$$= -0.5+\frac{\gamma}{1-\gamma}$$

By Comparing the returns of the three policies, we notice that $R_1 > R_2 > R_3$ for any γ. It is notable that R_3 does not strictly comply with the definition of returns because it is more like an expected value.

How to calculate returns?

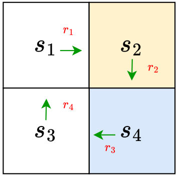

Figure: An example for demonstrating how to calculate returns.

There are two ways to calculate returns: The first is simply by definition: a return equals the discounted sum of all rewards collected along a trajectory.

$$\begin{aligned}
v_1 &= r_1 + \gamma r_2 + \gamma^2 r_3 + \cdots \\
v_2 &= r_2 + \gamma r_3 + \gamma^2 r_4 + \cdots \\
v_3 &= r_3 + \gamma r_4 + \gamma^2 r_1 + \cdots \\
v_4 &= r_4 + \gamma r_1 + \gamma^2 r_2 + \cdots
\end{aligned}$$

How to calculate returns?

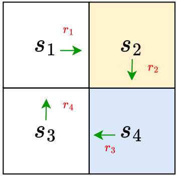

Figure: An example for demonstrating how to calculate returns.

The second way is based on the idea of bootstrapping.

$$\begin{aligned}
v_1 &= r_1 + \gamma(r_2 + \gamma r_3 + \cdots) = r_1 + \gamma v_2 \\
v_2 &= r_2 + \gamma(r_3 + \gamma r_4 + \cdots) = r_2 + \gamma v_3 \\
v_3 &= r_3 + \gamma(r_4 + \gamma r_1 + \cdots) = r_3 + \gamma v_4 \\
v_4 &= r_4 + \gamma(r_1 + \gamma r_2 + \cdots) = r_4 + \gamma v_1
\end{aligned} \tag{1}$$

How to calculate returns?

The equation 1 can be reformed into a linear matrix-vector equation:

$$\underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{v} = \begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix} + \begin{bmatrix} \gamma v_2 \\ \gamma v_3 \\ \gamma v_4 \\ \gamma v_1 \end{bmatrix} = \underbrace{\begin{bmatrix} r_1 \\ r_2 \\ r_3 \\ r_4 \end{bmatrix}}_{r} + \gamma \underbrace{\begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \end{bmatrix}}_{P} \underbrace{\begin{bmatrix} v_1 \\ v_2 \\ v_3 \\ v_4 \end{bmatrix}}_{v}, \tag{2}$$

which can be written compactly as

$$v = r + \gamma P v.$$

State values

Starting from t, we can obtain a state-action-reward trajectory:

$$S_t \xrightarrow{A_t} S_{t+1}, R_{t+1} \xrightarrow{A_{t+1}} S_{t+2}, R_{t+2} \xrightarrow{A_{t+2}} S_{t+3}, R_{t+3} \xrightarrow{A_{t+3}} \cdots$$

By definition, the discounted return along the trajectory is

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots$$

where $\gamma \in (0,1)$ is the discount rate.

State values

Note that G_t is a random variable since $R_{t+1}, R_{t+2}, \cdots$ are all random variables.
Since G_t is a random variable, we can calculate its expected value (also called the expectation or mean):

$$v_\pi(s) = \mathbb{E}[G_t | S_t = s]$$

Bellman equation

We can be rewritten the G_t as follows:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \\ &= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots) \\ &= R_{t+1} + \gamma G_{t+1} \end{aligned}$$

where $G_{t+1} = R_{t+2} + \gamma R_{t+3} + \cdots$.
This equation establishes the relationship between G_t and G_{t+1}.

Bellman equation

Then, the state value can written as

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}[G_t | S_t = s] \\
&= \mathbb{E}[R_{t+1} + \gamma G_{t+1} | S_t = s] \\
&= \mathbb{E}[R_{t+1} | S_t = s] + \gamma \mathbb{E}[G_{t+1} | S_t = s]
\end{aligned}
\tag{3}
$$

The two terms in Eq. 3 are analzed next slide.

Bellman equation

The first term $\mathbb{E}[R_{t+1}|S_t = s]$, is the expectation of the immediate rewards.
It can be calculated as follows:

$$
\begin{aligned}
\mathbb{E}[R_{t+1}|S_t = s] &= \sum_{a \in \mathcal{A}} \pi(a|s) \mathbb{E}[R_{t+1}|S_t = s, A_t = a] \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}} p(r|s, a) r
\end{aligned}
\tag{4}
$$

Bellman equation

The second term $\mathbb{E}[G_{t+1}|S_t = s]$, is the expectation of the future rewards. It can be calculated as follows:

$$
\begin{aligned}
\mathbb{E}[G_{t+1}|S_t = s] &= \sum_{s \in \mathcal{S}} \mathbb{E}[G_{t+1}|S_t = s, S_{t+1} = s']p(s'|s) \\
&= \sum_{s' \in \mathcal{S}} \mathbb{E}[G_{t+1}|S_{t+1} = s']p(s'|s) \quad \text{(due to the Markov property)} \\
&= \sum_{s' \in \mathcal{S}} v_\pi(s')p(s'|s) \\
&= \sum_{s' \in \mathcal{S}} v_\pi(s') \sum_{a \in \mathcal{A}} p(s'|s, a)\pi(a|s)
\end{aligned}
\tag{5}
$$

Bellman equation

Substituting Eq. 4 and Eq. 5 into Eq. 3, we obtain the Bellman equation:

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[G_{t+1}|S_t = s], \\
&= \underbrace{\sum_{a\in\mathcal{A}} \pi(a|s) \sum_{r\in\mathcal{R}} p(r|s,a) r}_{\text{mean of immediate rewards}} + \underbrace{\gamma \sum_{a\in\mathcal{A}} \pi(a|s) \sum_{s'\in\mathcal{S}} p(s'|s,a) v_\pi(s')}_{\text{mean of future rewards}} \\
&= \sum_{a\in\mathcal{A}} \pi(a|s) \left[\sum_{r\in\mathcal{R}} p(r|s,a) r + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a) v_\pi(s') \right], \quad \text{for all } s \in \mathcal{S}.
\end{aligned} \tag{6}
$$

This equation is in an elementwise form.
Since it is valid for every state, we can combine all these equations and write them concisely in a matrix-vector form.

Matrix-vector form of the Bellman equation

$$
\begin{aligned}
v_\pi(s) &= \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[G_{t+1}|S_t = s], \\
&= \underbrace{\sum_{a\in\mathcal{A}} \pi(a|s) \sum_{r\in\mathcal{R}} p(r|s,a) r}_{r_\pi(s)} + \gamma \sum_{s'\in\mathcal{S}} v_\pi(s') \underbrace{\sum_{a\in\mathcal{A}} p(s'|s,a)\pi(a|s)}_{p_\pi(s'|s)} \\
&= r_\pi(s) + \gamma \sum_{s'\in\mathcal{S}} p_\pi(s'|s) v_\pi(s')
\end{aligned}
\tag{7}
$$

Here, $r_\pi(s)$ denotes the mean of the immediate rewards, and $p_\pi(s'|s)$ is the probability of transitioning from s to s' under policy π.

Matrix-vector form of the Bellman equation

Let $v_\pi = [v_\pi(s_1), v_\pi(s_2), \cdots, v_\pi(s_n)]^T \in \mathbb{R}^n$, $r_\pi = [r_\pi(s_1), r_\pi(s_2), \cdots, r_\pi(s_n)]^T \in \mathbb{R}^n$, and $P_\pi \in \mathbb{R}^{n\times n}$ with $[P_\pi]_{ij} = p_\pi(s_j|s_i)$.
Then, the Bellman equation can be written in a matrix-vector form:

$$v_\pi = r_\pi + \gamma P_\pi v_\pi. \tag{8}$$

where v_π is the unknown to be solved, and r_π, P_π are known.

$$\underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi} = \underbrace{\begin{bmatrix} r_\pi(s_1) \\ r_\pi(s_2) \\ r_\pi(s_3) \\ r_\pi(s_4) \end{bmatrix}}_{r_\pi} + \gamma \underbrace{\begin{bmatrix} p_\pi(s_1|s_1) & p_\pi(s_2|s_1) & p_\pi(s_3|s_1) & p_\pi(s_4|s_1) \\ p_\pi(s_1|s_2) & p_\pi(s_2|s_2) & p_\pi(s_3|s_2) & p_\pi(s_4|s_2) \\ p_\pi(s_1|s_3) & p_\pi(s_2|s_3) & p_\pi(s_3|s_3) & p_\pi(s_4|s_3) \\ p_\pi(s_1|s_4) & p_\pi(s_2|s_4) & p_\pi(s_3|s_4) & p_\pi(s_4|s_4) \end{bmatrix}}_{P_\pi} \underbrace{\begin{bmatrix} v_\pi(s_1) \\ v_\pi(s_2) \\ v_\pi(s_3) \\ v_\pi(s_4) \end{bmatrix}}_{v_\pi}.$$

Closed-from solution

Since $v_\pi = r_\pi + \gamma P_\pi v_\pi$ is a simple linear equation, its closed-form solution can be easily obtained as follows:

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi$$

Iterative solution

Although the closed-form solution is useful for theoretical analysis purposes, it is not applicable in practice because it involves a matrix inversion operation, which still needs to be calculated by other numerical algorithms.
In fact, we can directly solve the Bellman equation using the following iterative algorithm:

$$v_{k+1} = r_\pi + \gamma P_\pi v_k, \quad k = 0, 1, 2, \cdots$$

This algorithm generates a sequence of vectors $\{v_0, v_1, v_2, \cdots\}$, where $v_0 \in \mathbb{R}^n$ is an initial guess of v_π.
It holds that

$$v_k \to v_\pi = (I - \gamma P_\pi)^{-1} r_\pi, \quad \text{as } k \to \infty$$

Action values

The action value indicates the value of taking an action at a state.

$$q_\pi(s, a) = \mathbb{E}[G_t | S_t = s, A_t = a]$$

What is the relationship between action values and state values?

From action value to state value

First, it follows from the properties of conditional expectation that

$$\underbrace{\mathbb{E}[G_t|S_t = s]}_{v_\pi(s)} = \sum_{a\in\mathcal{A}} \underbrace{\mathbb{E}[G_t|S_t = s, A_t = a]}_{q_\pi(s,a)} \pi(a|s)$$

It then follows that

$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) q_\pi(s, a) \tag{9}$$

Eq. 9 shows how to obtain state values from action values.

From state value to action value

Second, since the state value is given by Eq. 6

$$v_\pi(s) = \sum_{a\in\mathcal{A}} \pi(a|s) \underbrace{\left[\sum_{r\in\mathcal{R}} p(r|s,a)r + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)v_\pi(s')\right]}_{q_\pi(s,a)}$$

comparing it with Eq. 9, leads to:

$$q_\pi(s,a) = \sum_{r\in\mathcal{R}} p(r|s,a)r + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a)v_\pi(s') \tag{10}$$

Above Eq. 10 shows how to obtain action values from state values.

The Bellman equation in terms of action values

Substituting Eq. 9 into Eq. 10, we can rewrite the Bellman equation in terms of action values:

$$q_\pi(s,a) = \sum_{r\in\mathcal{R}} p(r|s,a)r + \gamma \sum_{s'\in\mathcal{S}} p(s'|s,a) \sum_{a'\in\mathcal{A}(s')} \pi(a'|s')q_\pi(s',a')$$