

The Note of Reinforcement Learning

Aoxiang

xuyuan

Oct 2025

1 Bellman equation

1.1 basic concept

The agent in time t is in state S_t , takes action A_t , receives reward R_{t+1} , the next state is S_{t+1} , it can be represented as a state-action-reward trajectory:

$$S_t \xrightarrow{A_t} S_{t+1}, R_{t+1} \xrightarrow{A_{t+1}} S_{t+2}, R_{t+2} \xrightarrow{A_{t+2}} S_{t+3}, R_{t+3} \dots \quad (1.1)$$

and the discounted return can be defined as:

$$\begin{aligned} G_t &= R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \dots \\ &= R_{t+1} + \gamma(R_{(t+1)+1} + \gamma R_{(t+1)+2} + \dots) = R_{t+1} + \gamma G_{t+1} \end{aligned} \quad (1.2)$$

where $\gamma \in (0, 1)$ is the discount rate, and we also anoted the R_{t+1} as imediate reward¹.

Cause R_t, A_t is random variable (even for a fixed π , the A_t is also random²), so is G_t , we can define the value function as the expectation of G_t :

$$\begin{aligned} & \text{\textcolor{red}{ s is a typical state}} \\ & \text{\textcolor{red}{ $v_{\pi}(s)$ }} = \mathbb{E}[G_t | S_t = s] = \mathbb{E}[\text{\textcolor{blue}{ $G_t|s$ }}] \\ & \text{\textcolor{red}{same as $v(\pi, s)$ }} \quad \text{\textcolor{blue}{简写为}} \end{aligned} \quad (1.3)$$

Notice that when $|s$ occurs in $\mathbb{E}[G_t|s]$, it equals to $|S_t = s$. (And $\mathbb{E}[G_{t+1}|S_{t+1} = 1] \leftrightarrow \mathbb{E}[G_{t+1}|s]$)

And $v_{\pi(s)}$ is time-independent, it only releates to the state s and policy π (for diifferent policies, the action space may be different).

$$\text{when } P(S_{a_i}|S_t) = p_i \quad \& \quad \sum p_i = 1 \quad \text{then} \quad v_{\pi}(s) = \sum p_i G_{a_i} \quad (1.4)$$

1.2 simply $v_{\pi(s)}$

From the definition of G_t , we have:

¹when agent receives reward, the agent is in time $t + 1$

²for example, $P(S_a|S_t) = 0.5, P(S_b|S_t) = 0.5 \quad a \neq b$

$$\begin{aligned}
v_{\pi(s)} &= \mathbb{E}[G_t|s] = \mathbb{E}[(R_{t+1} + \gamma G_{t+1})|s] \\
&= \mathbb{E}[R_{t+1}|S_t = s] + \gamma \mathbb{E}[G_{t+1}|S_t = s]
\end{aligned} \tag{1.5}$$

Notice there $\mathbb{E}[G_{t+1}|S_t = s]$ can,t be simplified to $\mathbb{E}[G_{t+1}|s]$.

When agent in s at time t , it will be lots of possible $S_{t+1} = s_i$ when take action a_i .

We first consider $\mathbb{E}[R_{t+1}|s]$:

$$\begin{aligned}
\mathbb{E}[R_{t+1}|S_t = s] &= \sum_i^n p(a_i|s, \pi) \mathbb{E}[R_{t+1}|S_t = s, A_t = a_i] \\
&= \sum_i^n \pi(a_i|s) \mathbb{E}[R_{t+1}|S_t = s, A_t = a_i] \\
&= \sum_i^n \pi(a_i|s) \sum_j^m p(r_j|s, a_i) r_j
\end{aligned} \tag{1.6}$$

where n is number of possible actions in \mathcal{A}_s , m is the number of possible rewards in $\mathcal{R}_{s,a}$

Then we consider $\mathbb{E}[G_{t+1}|S_t = s]$

$$\begin{aligned}
\mathbb{E}[G_{t+1}|S_t = s] &= \sum_i^l P(s_i|s, \pi) \mathbb{E}[G_{t+1}|s_i] = \sum_i^l p(s_i|s, \pi) v_{\pi}(s_i) \\
&= \sum_i^l p(a_i|s, \pi) p(s_i|a_i, s) v_{\pi}(s_i) = \sum_i^l \pi(a_i|s) p(s_i|a_i, s) v_{\pi}(s_i)
\end{aligned} \tag{1.7}$$

so finally we have:

$$\begin{aligned}
v_{\pi(s)} &= \sum_i^n \pi(a_i|s) \left[\sum_j^m p(r_j|s, a_i) r_j + \gamma \sum_k^l p(s_k|s, a_i) v_{\pi}(s_k) \right] \\
&= \sum_{a \in \mathcal{A}} \pi(a|s) \left[\sum_{r \in \mathcal{R}_s} p(r|s, a) r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a) v_{\pi}(s') \right] \quad \text{for all } s \in \mathcal{S}
\end{aligned} \tag{1.8}$$

where l is the number of possible states in \mathcal{S}_{t+1} when $S_t = s$.

And it is noteds below:

- The equation Gleichung (1.8) called the Bellman equation is a **set** of linear equations for all $s \in \mathcal{S}$.
- $\pi(s), \pi(s')$ is unknown and need to be solved.
- what is $\pi(a|s)$? $\pi(a_i|s) \equiv p(a_i|s, \pi)$
- $p(r|s, a) \neq 1, p(s'|s, a)$ represented the **system model** which can capture the strong randomness of the environment—meaning that the agent cannot know the exact subsequent state and reward even if it takes fixed action a in state s .

1.3 Matrix-vector form of the Bellman equation

For the bellman equation mentioned above, we can rewrite in another form (use some different notations).

$$\sum_{a \in \mathcal{A}} \pi(a|s) \left[\sum_{r \in \mathcal{R}_s} p(r|s, a)r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v_\pi(s') \right] \quad (1.9)$$

Firstly, we consider $\sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}_s} p(r|s, a)r$

$$\begin{aligned} \sum_{a \in \mathcal{A}} \pi(a|s) \sum_{r \in \mathcal{R}_s} p(r|s, a)r &= \sum_{a \in \mathcal{A}} \sum_{r \in \mathcal{R}_s} p(a|\pi, s)p(r|s, a)r \\ &= \sum_{r \in \mathcal{R}_s} p(r|s, \pi)r \\ &= \sum_{r \in \mathcal{R}_s} p_\pi(r|s)r \\ &\equiv \mathbb{E}[R|s, \pi] \equiv r_{\pi(s)} \end{aligned} \quad (1.10)$$

It means the expected immediate reward when agent in state s following policy π .

Secondly, we consider $\sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} p(s'|s, a)v_\pi(s')$

$$\sum_{a \in \mathcal{A}} \pi(a|s) \sum_{s' \in \mathcal{S}} p(s'|s, a)v_\pi(s') = \sum_{s' \in \mathcal{S}} p(s'|s, \pi)v_\pi(s') \quad (1.11)$$

And we notation $p(p'|s, \pi)$ as

$$p(s'|s, \pi) \equiv p_\pi(s'|s) \quad (1.12)$$

so the second part of bellman equation can be rewritten as:

$$\sum_{s' \in \mathcal{S}} p(s'|s, \pi)v_\pi(s') = \sum_{s' \in \mathcal{S}} p_\pi(s'|s)v_\pi(s') \quad (1.13)$$

The bellman equation can be rewritten as:

$$\begin{aligned} v_{\pi(s)} &= \sum_{a \in \mathcal{A}} \pi(a|s) \left[\sum_{r \in \mathcal{R}_s} p(r|s, a)r + \gamma \sum_{s' \in \mathcal{S}} p(s'|s, a)v_\pi(s') \right] \\ &= r_\pi(s) + \sum_{s' \in \mathcal{S}} p_\pi(s'|s)v_\pi(s') \end{aligned} \quad (1.14)$$

For all $s \in \mathcal{S}$, we notation s as s_i

$$\begin{aligned} v_\pi(s_i) &= r_\pi(s_i) + \gamma \sum_{s' \in \mathcal{S}} p_\pi(s'|s_i)v_\pi(s') \\ &= r_\pi(s_i) + \gamma \sum_j^n p_\pi(s_j|s_i)v_\pi(s_j) \end{aligned} \quad (1.15)$$

where n is the number of states in \mathcal{S} . Then, we define some vector notation:

$$\begin{aligned}
v_\pi &= [v_\pi(s_1), v_\pi(s_2), \dots, v_\pi(s_n)]^T \\
r_\pi &= [r_\pi(s_1), r_\pi(s_2), \dots, r_\pi(s_n)]^T \\
P_\pi[i, j] &= p_\pi(s_j | s_i) \quad \{P_\pi[i, j] > 0, \sum (P_\pi[i, :]) = 1\}
\end{aligned} \tag{1.16}$$

simply Gleichung (1.15) in matrix-vector form:

$$v_\pi(s_i) = r_\pi(s_i) + \gamma P_{\pi[i, :]} v_\pi \tag{1.17}$$

Take $n = 1, 2, 3 \dots n$ as example

$$\begin{aligned}
v_\pi(s_1) &= r_\pi(s_1) + \gamma P_{\pi[1, :]} v_\pi \\
v_\pi(s_2) &= r_\pi(s_2) + \gamma P_{\pi[2, :]} v_\pi \\
&\dots \\
v_\pi(s_n) &= r_\pi(s_n) + \gamma P_{\pi[n, :]} v_\pi
\end{aligned} \tag{1.18}$$

Obviously, we can rewrite above equations in matrix-vector form:

$$v_\pi = r_\pi + \gamma P_\pi v_\pi \tag{1.19}$$

1.4 Solving state values from the Bellman equation

1.4.1 close form solution

not applicable in practice because it involves a matrix inversion operation, which still needs to be calculated by other numerical algorithms

$$v_\pi = (I - \gamma P_\pi)^{-1} r_\pi \tag{1.20}$$

1.4.2 Iterative solution

In fact, we can directly solve the Bellman equation using the following iterative algorithm

$$v_{k+1} = r_\pi + \gamma P_\pi v_k \tag{1.21}$$

where v_0 is a initial guess of v_π , and when $k \rightarrow \infty$, v_k will converge to v_π .

Proof is below :

First define $\delta_k = v_k - v_\pi$, we need to prove when $k \rightarrow \infty, \delta_k \rightarrow 0$

$$v_k = v_\pi + \delta_k \quad v_{k+1} = v_\pi + \delta_{k+1} \quad \dots \tag{1.22}$$

Take Gleichung (1.22) into Gleichung (1.21), we have:

$$v_\pi + \delta_{k+1} = r_\pi + \gamma P_\pi (v_\pi + \delta_k) \tag{1.23}$$

Then simply the notation of δ_{k+1} and δ_k :

$$\delta_{k+1} = r_\pi + \gamma P_\pi \delta_k + \gamma P_\pi v_\pi - v_\pi \quad (1.24)$$

Use Gleichung (1.24) $v_\pi = r_\pi + \gamma P_\pi v_\pi$, we have:

$$\delta_{k+1} = \gamma P_\pi \delta_k = \gamma^k P_\pi \delta_0 \quad (1.25)$$

Since $\gamma \in (0, 1)$, when $k \rightarrow \infty$, $\gamma^k \rightarrow 0$, $\delta_{k+1} \rightarrow 0$.

1.5 From state value to action value

Finish the state value $v_{\pi(s)}$, we can easily get the action value $q_\pi(s, a)$ which means the agent expected reward when agent in state s and take action a :

$$\begin{aligned} q_\pi(s, a) &\equiv \mathbb{E}[G_t | S_t = s, A_t = a] \\ &\equiv \mathbb{E}[G_t | s, a] \end{aligned} \quad (1.26)$$

And use condition expectation, we have:

$$\begin{aligned} v_\pi(s) &= \mathbb{E}[G_t | s] = \sum_{a \in \mathcal{A}} \mathbb{E}[G_t | s, a] \pi(a | s) \\ &= \sum_{a \in \mathcal{A}} q_\pi(s, a) \pi(a | s) \end{aligned} \quad (1.27)$$

Then $v_\pi(s)$ can be represented by $q_\pi(s, a)$.

$$\begin{aligned} v_\pi(s) &= \sum_{a \in \mathcal{A}} \pi(a | s) \left[\sum_{r \in \mathcal{R}_s} p(r | s, a) r + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) v_\pi(s') \right] \\ &= \sum_{a \in \mathcal{A}} \pi(a | s) q_\pi(s, a) \longleftarrow \text{Gleichung (1.27)} \end{aligned} \quad (1.28)$$

So we can notation $q_\pi(s, a)$ as:

$$q_\pi(s, a) = \sum_{r \in \mathcal{R}_s} p(r | s, a) r + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) v_\pi(s') \quad (1.29)$$

Use q_π to replace $v_\pi(s')$ in Gleichung (1.29), we have:

$$\begin{aligned}
v_\pi(s') &= \sum_{a' \in \mathcal{A}} q_\pi(s', a') \pi(a' | s') \\
q_\pi(s, a) &= \sum_{r \in \mathcal{R}_s} p(r | s, a) r + \gamma \sum_{s' \in \mathcal{S}} p(s' | s, a) \sum_{a' \in \mathcal{A}} \pi(a' | s') q_\pi(s', a') \\
&= \sum_{r \in \mathcal{R}_s} p(r | s, a) r + \gamma \sum_k^{\text{len}(\mathcal{S})} p(s_k | s, a) \sum_l^{\text{len}(\mathcal{A})} \pi(a_l | s_k) q_\pi(s_k, a_l) \\
&= \sum_{r \in \mathcal{R}_s} p(r | s, a) r + \gamma \sum_k^{\text{len}(\mathcal{S})} \sum_l^{\text{len}(\mathcal{A})} p_\pi(s_k | s, a) \pi(a_l | s_k) q_\pi(s_k, a_l)
\end{aligned} \tag{1.30}$$

And like state value , we can also rewrite the equation by matrix-vector form: ($i = 1, 2, \dots, \text{len}(\mathcal{S}), j = 1, 2, \dots, \text{len}(\mathcal{A})$)

$$q_\pi(s_i, a_j) = \sum_{r \in \mathcal{R}_s} p(r | s, a) r + \gamma \sum_k^{\text{len}(\mathcal{S})} \sum_l^{\text{len}(\mathcal{A})} p_\pi(s_k | s_i, a_j) \pi(a_l | s_k) q_\pi(s_k, a_l) \tag{1.31}$$

Notation some useful notation:

1. Notation q_π as a $|\mathcal{A}| \times |\mathcal{S}|$ vector , and it can be accessed via the pair (i, j) , where $i = 1, 2, \dots, |\mathcal{A}|, j = 1, 2, \dots, |\mathcal{S}|$
2. $P\Pi$ is a $|\mathcal{A}| \times |\mathcal{A}| \times |\mathcal{S}|$ matrix, where the element at row pair (i, j) and column pair (k, l) is defined as:

$$p_\pi(s_k | s_i, a_j) \pi(a_l | s_k) \tag{1.32}$$

3. \tilde{r} is the immediate reward vector indexed by the state-action pairs (a, s) , is also a $|\mathcal{A}| \times |\mathcal{S}|$ vector. And the element at row pair (i, j) is defined as:

$$\sum_{r \in \mathcal{R}_s} p(r | s, a) r \tag{1.33}$$

So we have

$$\begin{aligned}
q_\pi[(i, j), 1] &= \tilde{r}[(i, j), 1] + \gamma P\Pi[(i, j), :] q_\pi \\
q_\pi &= \tilde{r} + \gamma P\Pi q_\pi
\end{aligned} \tag{1.34}$$

2 Optimal State Values and Bellman Optimality Equation

参考文献