

第二课 bellman equation

1.state value

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

2.bellman equation

Recall that:

$$v_{\pi}(s) = \sum_a \pi(a|s) \left[\sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_{\pi}(s') \right]$$

Rewrite the Bellman equation as

$$v_{\pi}(s) = r_{\pi}(s) + \gamma \sum_{s'} p_{\pi}(s'|s)v_{\pi}(s') \quad (1)$$

where

$$r_{\pi}(s) \triangleq \sum_a \pi(a|s) \sum_r p(r|s, a)r, \quad p_{\pi}(s'|s) \triangleq \sum_a \pi(a|s)p(s'|s, a)$$

Suppose the states could be indexed as s_i ($i = 1, \dots, n$).

For state s_i , the Bellman equation is

$$v_{\pi}(s_i) = r_{\pi}(s_i) + \gamma \sum_{s_j} p_{\pi}(s_j|s_i)v_{\pi}(s_j)$$

Put all these equations for all the states together and rewrite to a matrix-vector form

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

where

- $v_{\pi} = [v_{\pi}(s_1), \dots, v_{\pi}(s_n)]^T \in \mathbb{R}^n$
- $r_{\pi} = [r_{\pi}(s_1), \dots, r_{\pi}(s_n)]^T \in \mathbb{R}^n$
- $P_{\pi} \in \mathbb{R}^{n \times n}$, where $[P_{\pi}]_{ij} = p_{\pi}(s_j|s_i)$, is the *state transition matrix*

example

If there are four states, $v_\pi = r_\pi + \gamma P_\pi v_\pi$ can be written out as

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

3. policy evaluation

给定策略求对应的state value叫做policy evaluation

The Bellman equation in matrix-vector form is

$$v_\pi = r_\pi + \gamma P_\pi v_\pi$$

- The *closed-form solution* is:

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

In practice, we still need to use numerical tools to calculate the matrix inverse.

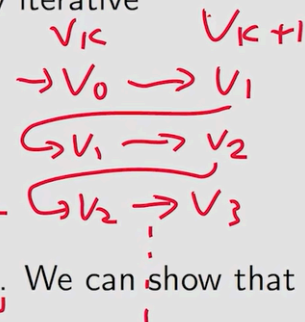
Can we avoid the matrix inverse operation? Yes, by iterative algorithms.

- An *iterative solution* is:

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

This algorithm leads to a sequence $\{v_0, v_1, v_2, \dots\}$. We can show that

$$\underbrace{v_k}_{\text{red circle}} \rightarrow \underbrace{v_\pi}_{\text{red circle}} = (I - \gamma P_\pi)^{-1} r_\pi, \quad \underbrace{k \rightarrow \infty}_{\text{red underline}}$$



4. action value

根据下面等式可以由state value算出action value

Recall that the state value is given by

$$v_\pi(s) = \sum_a \pi(a|s) \left[\underbrace{\sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_\pi(s')}_{q_\pi(s, a)} \right] \quad (3)$$

By comparing (2) and (3), we have the **action-value function** as

$$q_\pi(s, a) = \sum_r p(r|s, a)r + \gamma \sum_{s'} p(s'|s, a)v_\pi(s') \quad (4)$$

当然, 也可由action value算出state value