第二课 bellman equation

1.state value

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

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

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

where

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

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

- $v_\pi = [v_\pi(s_1),\ldots,v_\pi(s_n)]^T \in \mathbb{R}^n$
- $r_\pi = [r_\pi(s_1),\ldots,r_\pi(s_n)]^T \in \mathbb{R}^n$
- $P_\pi \in \mathbb{R}^{n\times n}$, 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.

v_k $\quad v_{k+1}$

$\to v_0 \to v_1$

$v_1 \to v_2$

$v_2 \to v_3$

- *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

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

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

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

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

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

当然，也可由action value算出state value