RL debts

28 октября 2023 г. 14:37

I. $V^*(s) \overset{def}{=} \max_\pi V_\pi(s) = \max_\pi \sum_{a \in A(s)} \pi(a|s) Q_\pi(s,a)$

<u>Prove</u>: $V^*(s) = \max_\pi \max_{a \in A(s)} Q_\pi(s,a)$

$$= \max_\pi \max_a \mathbb{E}_\pi [R_t \mid s_t = s, a_t = a]$$

$$= \max_\pi \max_a \mathbb{E}_\pi \left[\sum_{k=0}^{+\infty} \gamma^k r_{t+k+1} \mid s_t = s, a_t = a \right]$$

$$= \max_\pi \max_a \mathbb{E}_\pi \left[r_{t+1} + \gamma \sum_{k=0}^{+\infty} \gamma^k r_{t+k+2} \mid s_t = s, a_t = a \right]$$

$$= \max_{a \in A(s)} \max_\pi \mathbb{E}_\pi [\ldots]$$

$$= \max_a \left[\mathbb{E}_{*} \{ r_{t+1} \mid s_t = s, a_t = a \} + \gamma \max_\pi \mathbb{E}_\pi \{ R_{t+1} \mid s_{t+1} = s, a_{t+1} = a \} \right]$$

$$= \max_a \mathbb{E} \{ r_{t+1} + \gamma V^*(s_{t+1}) \mid s_t = s, a_t = a \}$$

The same logic applies to derivation of Q^*:

$$Q^*(s,a) = \mathbb{E} \{ r_{t+1} + \gamma \max_{a'} Q^*(s_{t+1}, a') \mid s_t = s, a_t = a \}$$

II. Let π and π' be two policies such that

$$\boxed{\forall s \in S \quad Q_\pi(s, \pi'(s)) \geqslant V_\pi(s)}$$

Then $V_{\pi'}(s) \geqslant V_\pi(s)$ $(\Leftrightarrow \pi' \geqslant \pi)$

$\triangleright$ $V_\pi(s) \leqslant Q_\pi(s, \pi'(s)) = \mathbb{E}_{\pi'} [r_{t+1} + \gamma V_\pi(s_{t+1}) \mid s_t = s]$

$$\leqslant \mathbb{E}_{\pi'} [r_{t+1} + \gamma Q_\pi(s_{t+1}, \pi'(s_{t+1})) \mid s_t = s]$$

$$= \mathbb{E}_{\pi'} [r_{t+1} + \gamma \mathbb{E}_{\pi'} \langle r_{t+2} + \gamma V_\pi(s_{t+2}) \mid s_{t+1} = s_{t+1} \rangle \mid s_t = s]$$

$$= \mathbb{E}_{\pi} [r_{t+1} + \gamma r_{t+2} + \gamma^2 V_\pi(s_{t+2}) \mid s_t = s]$$

$$\leqslant \mathbb{E}_{\pi} [r_{t+1} + \gamma r_{t+2} + \gamma^2 r_{t+3} + \gamma^3 V_\pi(s_{t+3}) \mid s_t = s]$$

$$\vdots$$

$$\leqslant V_{\pi'}(s) \qquad \square$$

In Policy Improvement:

$$\pi'(s) = \arg\max_a Q_\pi(s,a)$$

$$Q_\pi(s, \pi'(s)) \geqslant Q_\pi(s,a) \quad \forall a \in A$$

$$V_\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)} Q_\pi(s,a) \leqslant Q_\pi(s, \pi'(s)) \Rightarrow \pi' \geqslant \pi$$

When $\pi' = \pi \Rightarrow V_{\pi'} = V_\pi$

then it's optimal, since it satisfies:

$$V_{\pi'}(s) = \max_a \sum_{r,s'} p(r,s' \mid s,a) [r + \gamma V_\pi(s')]$$