Lec-16, IT567, 24-25

$$\pi'(s) \overset{\Delta}{=} \arg\max_{a} \; q_\pi(s,a)$$

$$= \arg\max_{a} \; E\left[R_{t+1} + \gamma v_\pi(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right]$$

$$= \arg\max_{a} \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_\pi(s')\right]$$

$$v_\pi(s) = v_{\pi'}(s) \quad \forall s \in \mathcal{S} \text{ or } \mathcal{S}^+ \quad \Rightarrow$$

$$\pi'(s) = \arg\max_{a} \sum_{s',r} p(s',r \mid s,a)\left[r + \gamma v_{\pi'}(s')\right]$$

$$= \arg\max_{a} \; q_{\pi'}(s,a)$$

$$v_{\pi'}(s) = q_{\pi'}(s,\pi'(s)) = \max_{a} q_{\pi'}(s,a)$$

$\Rightarrow$ $V_{\pi'}(s) = \max_a E[R_{t+1} + \gamma V_{\pi'}(S_{t+1}) | S_t = s, A_t = a]$

this is BOE, implying $\pi'(\cdot)$ as one of the optimal policies & $V_{\pi'}(s)$ as the optimal state-value function

$\hookrightarrow V_*(s)$

$\Pi_* = \{\pi_1, \pi_2\}$

$\hookrightarrow$ set of optimal policies

<u>Value iteration</u>:- (PE → PI → PItr)

$\hookrightarrow$ prediction

Policy itr involves policy evaluation which itself is iterative & requires multiple sweeps through the state space.

Is it possible to truncate PE?

without losing the convergence guarantee of Policy itr

What about if PE is stopped after just one sweep?

(one update of each state)

This is called as Value iteration (V Its) → combination of PI & truncated PE steps

$$V_{k+1}(s) \triangleq \max_{a} E\left[R_{t+1} + \gamma V_k(S_{t+1}) \,\middle|\, S_t = s, A_t = a\right]$$

$$\forall s \in \mathcal{S}$$

$$\begin{tikzcd}
\pi_0 \arrow[r, "E"] & V_{\pi_0} \arrow[r, "I"] \arrow[r, bend right, "\text{V Its}"'] & \pi_1 \arrow[r, "E"] \arrow[r, bend right] & V_{\pi_1} & \cdots & \pi_* \arrow[r] & V_{\pi_*}
\end{tikzcd}$$

BE

BOE

for any policy π, $V_\pi(s)$ is called the state-value func. if it satisfies BE.

optimal policy π_*