Given MDP. S. A. R(s.a.s')

Policy Evaluation: until V-values converge.

$$V_{k+1}^{\pi_i} \leftarrow \sum_{s'} P(s.\pi_i(s).s')\left[R(s.\pi_i(s).s') + \gamma V_k^{\pi_i}(s')\right]$$

$\pi_i(s)$

Policy Improvement: $\pi_{i+1}(s) \in \arg\max_a Q^{\pi_i}(s.a)$

$$\pi_{i+1}(s) \in \arg\max_a \sum_{s'} P(s.a.s')\left[R(s.a.s') + \gamma V^{\pi_i}(s')\right]$$

argmax : a

Proof: ① $V^{\pi_{i+1}}(s) \geq V^{\pi_i}(s)$

$\pi_i(s) = a = \pi_{i+1}(s)$

$\Updownarrow$

$V^{\pi_i}(s) = \max Q(s.a)$

② Policy Iteration converges to an optimal policy.

① First, prove the convergency of policy evaluation:

$$V_{k+1}^{\pi_i} - V_k^{\pi_i} = \sum_{s'} P(s.\pi_i(s), s')\left[R(s.\pi_i(s).s') + \gamma V_k^{\pi_i}(s')\right]$$

$$- \sum_{s'} P(s.\pi_i(s), s')\left[R(s.\pi_i(s).s') + \gamma V_{k-1}^{\pi_i}(s')\right]$$

$$= \gamma \sum_{s'} P(s.\pi_i(s).s')\left(V_k^{\pi_i}(s') - V_{k-1}^{\pi_i}(s')\right) \leq \gamma \| V_k^{\pi_i} - V_{k-1}^{\pi_i} \|_\infty$$

$$\leq \cdots \leq \gamma^k \| V_1^{\pi_i} - V^{\pi_i} \|_\infty \Rightarrow \text{converge}$$

Then. prove $V^{\pi_{i+1}}(s) \geq V^{\pi_i}(s)$.

$\because a = \arg\max Q^{\pi_i}(s.a) \quad \therefore V^{\pi_i}(s) \leq Q^{\pi_i}(s.a) = Q^{\pi_i}(s.\pi_{i+1}(s))$

$$V^{\pi_i}(s) \leq Q^{\pi_i}(s.\pi_{i+1}(s)) = \mathbb{E}\left[R_{t+1} + \gamma V^{\pi_i}(S_{t+1}) \mid S_t = s. A_t = \pi_{i+1}(s)\right]$$

$$= \mathbb{E}_{\pi_{i+1}}\left[R_{t+1} + \gamma V^{\pi_i}(S_{t+1}) \mid S_t = s\right]$$ (中间不断展开的步骤实则就是 policy evaluation 在做的事

$$\leq \mathbb{E}_{\pi_{i+1}}\left[R_{t+1} + \gamma (R_{t+2}) + \gamma^2 V^{\pi_i}(S_{t+2}) \mid S_t = s.\right]$$

$$\leq \cdots \leq \mathbb{E}_{\pi_{i+1}}\left[R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots \mid S_t = s\right]$$

$$= V^{\pi_{i+1}}(s)$$

$\arg\max Q(s.a) = \pi_{i+1}(s)$

$\downarrow$

$= \pi_i(s)$.

② From ① $V^{\pi_i}(s) \leq Q^{\pi_i}(s.\pi_{i+1}(s)) \leq V^{\pi_{i+1}}(s)$

policy converge $\Leftrightarrow \pi_i = \pi_{i+1} \Leftrightarrow V^{\pi_i}(s) = Q^{\pi_i}(s.\pi_{i+1}(s)) = V^{\pi_{i+1}}(s)$

Because state and action space are finite and discrete, the number of feasible policy is finite, which assures the existence of optimal policy. Therefore, after finite policy iterations the policy must converge to an policy π^* for $V^{\pi_i}(s) \leq V^{\pi_{i+1}}(s)$. If there are two policy π_a and π_b on a "policy oscillation", $\pi_a \to V_a \to \pi_b$. $\pi_b \to V_b \to \pi_a$. then we must have $V_a \leq V_b \leq V_a \Rightarrow \pi_a$ and π_b are same policy.

$\therefore$ Policy iteration can converge to an optimal policy