Policy iteration

init π_0.

for $k = 1, 2, \cdots$,

policy eval. $Q^{\pi_{k-1}}$

policy improve $\pi_k \leftarrow \pi_{Q^{\pi_{k-1}}}$

$\pi_f(s) = \arg\max_a f(s,a)$

$\pi_{k-1} = \pi^*$. $\rightarrow$ $Q^{\pi_{k-1}} = Q^*$ $\rightarrow$ $\pi_k = \pi_{Q^*} = \pi^*$.

Monotone improvement: (1) $\forall k$, $V^{\pi_k} \geq V^{\pi_{k-1}}$

(2) as long as $\pi_{k-1} \neq \pi^*$. $\exists s$.

$V^{\pi_k}(s) > V^{\pi_{k-1}}(s)$.

Corollary: $\pi_k = \pi^*$ $\forall k \geq |A|^{|S|}$.

Performance difference lemma: $\forall \pi, \pi'$.

$$V^{\pi'}(s) - V^{\pi}(s) = \frac{1}{1-\gamma} \mathbb{E}_{\tilde{s} \sim d^{\pi', s}} \underbrace{\left[Q^{\pi}(\tilde{s}, \pi') - V^{\pi}(\tilde{s}) \right]}_{A^{\pi}(\tilde{s}, \pi')}.$$

"advantage"

$$(1-\gamma) \sum_{t=1}^{\infty} \gamma^{t-1} \underbrace{d_t^{\pi', s}}_{\mathbb{P}[s_t = \cdot \mid s_1 = s, \pi']}.$$

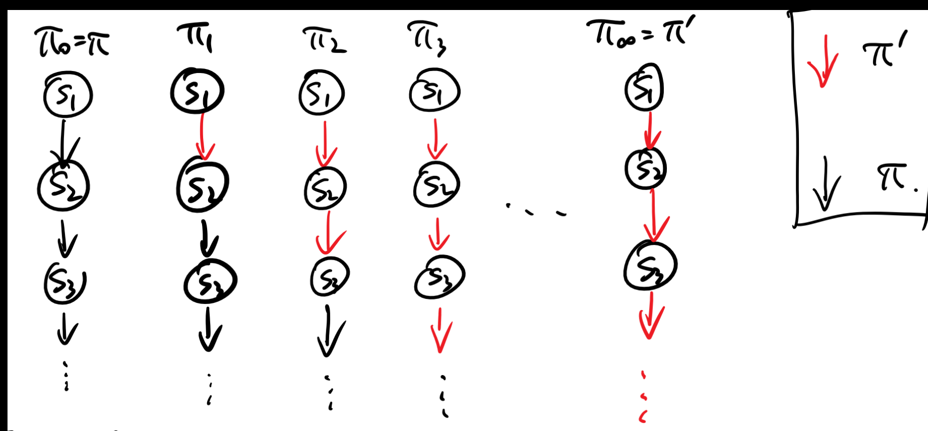
Given P-D Lemma:

$$V^{\pi_k}(s) - V^{\boxed{\pi_{k-1}}}(s) = \frac{1}{1-\gamma}\,\mathbb{E}_{\boxed{\tilde{s}}}\Big[\underline{Q^{\pi_{k-1}}(\tilde{s}, \pi_k)} - Q^{\pi_{k-1}}(\tilde{s}, \pi_{k-1})\Big].$$

$$\pi_k \leftarrow \pi_{Q^{\pi_{k-1}}}. \qquad \boxed{\geq 0.} \qquad V^{\pi_{k-1}}(\tilde{s}).$$

$$V^{\pi}(s) - V^{\pi'}(s).$$

$$= \sum_{i=0}^{\infty} V^{\pi_{i+1}}(s) - V^{\pi_i}(s)$$



$$V^{\pi_2}(s) = \cancel{\mathbb{E}[r_1 + \gamma r_2 \mid s_1 = s, \pi']}$$

$$+ \mathbb{E}\Big[\sum_{t=3}^{\infty} \gamma^{t-1} r_t \,\Big|\, s_1 = s,\ a_{1:2} \sim \pi',\ a_{3:\infty} \sim \pi\Big].$$

$$V^{\pi_3}(s) = \cancel{\mathbb{E}[r_1 + \gamma r_2 \mid s_1 = s, \pi']}$$

$$+ \mathbb{E}\Big[\sum_{t=3}^{\infty} \gamma^{t-1} r_t \,\Big|\, s_1 = s,\ a_{1:3} \sim \pi',\ a_{4:\infty} \sim \pi\Big].$$

$$\gamma^2 \mathbb{E}_{s_3, a_3} \mathbb{E}\Big[\sum_{t=3}^{\infty} \gamma^{t-3} r_t \,\Big|\, \cancel{s_1 = s,\ a_{1:2} \sim \pi'},\ s_3 = s',\ a_3 = a',\ a_{4:\infty} \sim \pi\Big].$$

$$- \gamma^2 \mathbb{E}\left[Q^{\pi}(s_3, \underline{\pi}) \;\middle|\; \underline{s_1 = s,\ a_{1:2} \sim \pi'} \right] -$$

$$+ \gamma^2 \mathbb{E}\left[Q^{\pi}(s_3, \underline{\pi'}) \;\middle|\; s_1 = s,\ a_{1:2} \sim \pi' \right]$$

$$= \gamma^2 \mathbb{E}_{s_3 \sim d_3^{\pi', s}} \left[Q^{\pi}(s_3, \pi') - Q^{\pi}(s_3, \pi) \right].$$

$$\sum_{i=0}^{\infty} V^{\pi_{i+1}}(s) - V^{\pi_i}(s)$$

$$= \sum_{i=0}^{\infty} \gamma^i \mathbb{E}_{\tilde{s} \sim d_{i+1}^{\pi', s}} \left[Q^{\pi}(\tilde{s}, \pi') - Q^{\pi}(\tilde{s}, \pi) \right] -$$

$$\mathbb{E}_p[f] + \mathbb{E}_q[f].$$

$$= \langle p, f \rangle + \langle q, f \rangle.$$

$$= 2 \cdot \left\langle \frac{p+q}{2}, f \right\rangle. = 2\, \mathbb{E}_{\frac{p+q}{2}}[f].$$

$$\boxed{\begin{aligned}&\text{If } \pi_{k-1} \neq \pi^*.\quad \exists \bar{s},\ Q^{\pi_{k-1}}(\bar{s}, \pi_{k-1}) < \max_a Q^{\pi_{k-1}}(\bar{s}, a)\\ &\exists \bar{s}\ \ Q^{\pi_{k-1}}(\bar{s}, \pi_k) > Q^{\pi_{k-1}}(\bar{s}, \pi_{k-1})\end{aligned}}$$

$$\left(\text{o.w. } \forall s,\ Q^{\pi_{k-1}}(s, \pi_k) = Q^{\pi_{k-1}}(s, \pi_{k-1})\right).$$

$$\to V^{\pi^*}(s) - V^{\pi_{k-1}}(s) = \frac{1}{1-\gamma} \mathbb{E}_{\boxed{\cdot}\ \pi^*} \left[Q^{\pi_{k-1}}(s, \pi^*) - Q^{\pi_{k-1}}(s, \pi_{k-1}) \right]$$

$$\to V^{\pi_k}(s) - V^{\pi_{k-1}}(s) = \frac{1}{1-\gamma} \cdot \mathbb{E}_{\tilde{s} \sim d^{\pi_k, s}} \underbrace{\left[Q^{\pi_{k-1}}(\tilde{s}, \pi_k) - Q^{\pi_{k-1}}(\tilde{s}, \pi_{k-1}) \right]}_{\geq 0}.$$

$$s = \bar{s}\ \checkmark.$$

$$\Rightarrow d^{\pi_k, \bar{s}}(\bar{s}) > 0$$

$$(\geq (1-\gamma)).$$