Policy Iteration

Init π_0 arbitrarily.

for $k = 1, 2, 3, \dots$, $\pi_k \leftarrow \pi_{Q^{\pi_{k-1}}}$ "Policy Eval" step

$\pi_f(s) := \operatorname{argmax}_{a \in A} f(s,a)$. ↳ [1. Compute $Q^{\pi_{k-1}}$ 2. Take its greedy policy to be π_k. "Policy Improvement step.

Remarks:

1. $\pi^* \xrightarrow{\text{eval}} Q^{\pi^*} = Q^* \xrightarrow{\text{improve}} \pi_{Q^*} = \pi^*$.
2. Assume Q^π is calculated exactly.

Convergence?

Policy Improvement Theorem.

In PI, $V^{\pi_k}(s) \geq V^{\pi_{k-1}}(s)$, $\forall k \geq 1$, $s \in S$.

(Further, if $\pi_{k-1} \neq \pi^*$, improvement is strict in at least 1 state)

Corollary: PI converges in $|A|^{|S|}$ iterations.

(b/c. polices in $\pi_0, \pi_1, \pi_2, \dots$, never repeats before π^*.)

See standard proof in notes (using "monotone property of T")

Lemma: (Performance-Difference Lemma) $\forall \pi, \pi'$. $s \in S$.

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

(s-th row of matrix. $(I - \gamma P^{\pi'})^{-1}$. ⟺ normalized discounted occupancy of π starting in s.

$$\preceq \mathbb{I} + \gamma P^{\pi'} + \gamma^2 (P^{\pi'})^2 + \gamma^3 (P^{\pi'})^3 + \cdots$$

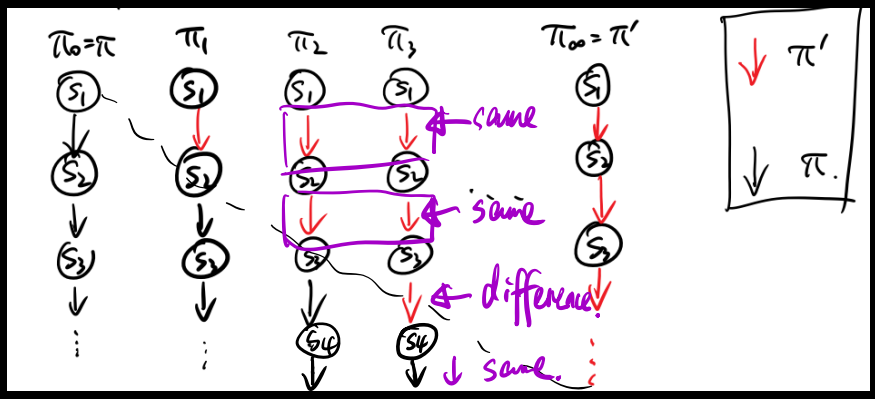
$\mathbb{I}$: s-th row: if $s_1 = s$, dist of s_t.

where $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$. "advantage function".

Construct π_i:

$\pi_0 = \pi$. $\pi_\infty = \pi'$.

For $1 \le i < \infty$.



π_i is non-stationary:

$$\begin{cases} a_t = \pi'(s_t) & \text{if } t \le i. \\ a_t = \pi(s_t) & \text{if } t > i. \end{cases} \qquad (s_1 = \ ).$$

$$V^{\pi'}(s) - V^\pi(s) = V^{\pi_\infty}(s) - V^{\pi_0}(s) = \sum_{i=0}^{\infty} \left(V^{\pi_{i+1}}(s) - V^{\pi_i}(s) \right).$$

$$= \sum_{i=0}^{\infty} \gamma^i \sum_{s' \in S} \underbrace{P[s_{i+1} = s' \mid s_1 = s, \pi']}_{\text{the distribution of } s_{i+1} \text{ for both } \pi_i \text{ \& } \pi_{i+1}} \cdot \underbrace{(Q^\pi(s', \pi') - Q^\pi(s, \pi))}$$

$$= \sum_{i=0}^{\infty} \gamma^i \sum_{s' \in S} P[s_{i+1} = s' \mid s_1 = s, \pi'] \cdot A^\pi(s', \pi')$$

$$= \sum_{s' \in S} \left(\sum_{i=0}^{\infty} \gamma^i P[s_{i+1} = s' \mid s_1 = s, \pi'] \right) A^\pi(s', \pi')$$

$$= \sum_{s' \in S} \cdot \frac{1}{1-\gamma} \cdot d^{\pi', s}(s') \cdot A^\pi(s', \pi').$$

Proof of policy improvement:

$$V^{\pi_k}(s) - V^{\pi_{k-1}}(s) = \frac{1}{1-\gamma} \mathbb{E}_{s' \sim d^{\pi_k, s}} \left[A^{\pi_{k-1}}(s, \pi_k) \right]$$

$$A^{\pi_{k-1}}(s, \pi_k) = Q^{\pi_{k-1}}(s, \pi_k) - \underbrace{V^{\pi_{k-1}}(s)}_{Q^{\pi_{k-1}}(s, \pi_{k-1})}$$

Claim: ≥ 0.

recall: π_k is greedy w.r.t. $Q^{\pi_{k-1}}$.

$$\therefore \underbrace{Q^{\pi_{k-1}}(s, \pi_k)}_{\Delta} \geq Q^{\pi_{k-1}}(s, \pi_{k-1}) = V^{\pi_{k-1}}(s).$$

Strict improvement. $\pi_{k-1} \neq \pi^*$, then $\exists s, V^{\pi_k}(s) > V^{\pi_{k-1}}(s)$

$\Leftrightarrow \exists s, A^{\pi_{k-1}}(s, \pi_k) > 0$.

Proof: if $\pi_{k-1} \neq \pi^*$. $\exists s_0, V^{\pi^*}(s_0) - V^{\pi_{k-1}}(s_0) > 0$.

$$\Rightarrow \underbrace{\mathbb{E}_{s' \sim d^{\pi^*, s_0}} \left[A^{\pi_{k-1}}(s', \pi^*) \right]}_{\Delta} > 0.$$

$$\exists \tilde{s} : A^{\pi_{k-1}}(\tilde{s}, \pi^*) > 0.$$

$$\underbrace{A^{\pi_{k-1}}(\tilde{s}, \pi_k)}_{\frac{1}{1-\gamma}} \geq A^{\pi_{k-1}}(\tilde{s}, \pi^*) > 0.$$

$$V^{\pi_k}(\tilde{s}) - V^{\pi_{k-1}}(\tilde{s}) = \mathbb{E}_{s' \sim \underbrace{d^{\pi_k, \tilde{s}}}} \left[A^{\pi_{k-1}}(s', \pi_k) \right]$$

$$\geq A^{\pi_{k-1}}(\tilde{s}, \pi_k) > 0.$$

Exponential convergence to ε-optimality.

Thm: $\|Q^* - Q^{\pi_{k+1}}\|_\infty \le \gamma \cdot \|Q^* - Q^{\pi_k}\|_\infty$

Proof: Key facts. (to be established). "Pointwise"

(a) $T^{\pi_{k+1}} Q^{\pi_k} \ge T^{\pi} Q^{\pi_k}$ $\forall \pi$.

(b) $T^{\pi_{k+1}} Q^{\pi_k} \le Q^{\pi_{k+1}}$

$$\underbrace{Q^* - Q^{\pi_{k+1}}}_{\ge 0} = Q^* - T^{\pi_{k+1}} Q^{\pi_k} + \underbrace{T^{\pi_{k+1}} Q^{\pi_k} - Q^{\pi_{k+1}}}_{\le 0}$$

$$\le Q^* - T^{\pi^*} Q^{\pi_k} = T^{\pi^*} Q^* - T^{\pi^*} Q^{\pi_k}$$

$$\|Q^* - Q^{\pi_{k+1}}\|_\infty \le \|T^{\pi^*} Q^* - T^{\pi^*} Q^{\pi_k}\|_\infty \le \gamma \cdot \|Q^* - Q^{\pi_k}\|_\infty$$

To establish (a) & (b).

$$(T^{\pi_{k+1}} Q^{\pi_k})(s,a) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} \cdot r_t \,\middle|\, s_1 = s,\ a_1 = a,\ a_2 \sim \pi_{k+1},\ a_{3:\infty} \sim \pi_k\right].$$

$$T^{\pi} Q^{\pi_k} \le T^{\pi_{k+1}} Q^{\pi_k} \quad \left(Q^{\pi_k}(s, \pi_{k+1}) \ge Q^{\pi_k}(s,\pi)\ \forall \pi\right) \text{ better}$$

$$Q^{\pi_{k+1}}(s,a) = \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s,\ a_1 = a,\ a_2 \sim \pi_{k+1},\ a_{3:\infty} \sim \pi_{k+1}\right]$$

$$(T^{\pi} f)(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}\left[f(s', \pi)\right].$$