Policy Iteration

Local planning / Online planning

$\boxed{\text{P.I}}$

$M = (S, \mathcal{A}, P, r, \gamma)$, $0 \le \gamma < 1$

π_0 ML

π_1

π_2 π_{k+1} greedy w.r.t v^{π_k}

$\vdots$

$$T_{\pi_{k+1}} v^{\pi_k} = T v^{\pi_k}$$

$$\pi_{k+1}(s) = \arg\max_a \underbrace{r_a(s) + \gamma \langle P_a(s), v^{\pi_k} \rangle}$$

$\forall s$ syst. tie resolution

$$\boxed{\textbf{Lemma 1: } \|v^{\pi_k} - v^*\|_\infty \le \gamma^k \|v^{\pi_0} - v^*\|_\infty}$$

Proof: Claim: $v^* \ge v^{\pi_k} \ge T^k v^{\pi_0}$

?

$$\|v^* - v^{\pi_k}\|_\infty \underset{v^* = T v^*}{\le} \|T^k v^* - T^k v^{\pi_0}\|_\infty \le \gamma^k \|v^* - v^{\pi_0}\|_\infty$$

Need: $\boxed{v^{\pi_k} \ge T^k v^{\pi_0}}$

$k-1 \Rightarrow k$

$$T_{\pi_k} v^{\pi_{k-1}} = \overset{\max}{T} v^{\pi_{k-1}} \ge T_{\pi_{k-1}} v^{\pi_{k-1}} = v^{\pi_{k-1}}$$

$$T_{\pi_k}^2 v^{\pi_{k-1}} \ge T_{\pi_k} v^{\pi_{k-1}} = T v^{\pi_{k-1}} \ge v^{\pi_{k-1}}$$

$$\boxed{v^{\pi_k}} \xleftarrow[i \to \infty]{} T_{\pi_k}^i v^{\pi_{k-1}} \boxed{\ge T v^{\pi_{k-1}}} \ge T^2 v^{\pi_{k-2}}$$

// Q.e.d.

$v^{\pi'} = (I-\gamma P_{\pi'})^{-1} r_{\pi'}$

Value-difference identity

π, π' ML

$v^{\pi'} = r_{\pi'} + \gamma P_{\pi'} v^{\pi'}$

$v^{\pi'} = T_{\pi'} v^{\pi'}$

$$v^{\pi'} - v^{\pi} = (I-\gamma P_{\pi'})^{-1}\left[r_{\pi'} - (I-\gamma P_{\pi'})v^{\pi}\right]$$

$$= (I-\gamma P_{\pi'})^{-1}\Big[\underbrace{T_{\pi'} v^{\pi} - v^{\pi}}_{g(\pi',\pi)}\Big]$$

[advantage of π' relative to π]

π_0 $\quad v^*$

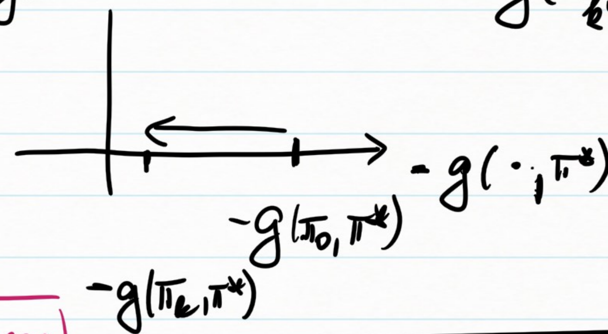
$$\boxed{v^{\pi'} - v^{\pi} = (I-\gamma P_{\pi'})^{-1} g(\pi', \pi)}$$

π_k $\quad \pi^*$

$\pi = \pi^*$ an opt. ML policy

$\forall \pi' : \ g(\pi', \pi^*) \le 0$

Progress measure: $-g(\pi_k, \pi^*)$



Progress Lemma: $\forall \pi_0 \quad v^{\pi_0} \neq v^* \qquad \{\pi_k\}$ PI

$\exists s_0 \in S$ s.t. $\forall k \ge k^*(\gamma)$

$\pi_k(s_0) \neq \pi_0(s_0)$

Proof: $-g(\pi_k, \pi^*) = (I-\gamma P_{\pi_k})(v^* - v^{\pi_k})$

$$= v^* - v^{\pi_k} - \underbrace{\gamma P_{\pi_k} \underbrace{(v^* - v^{\pi_k})}_{\ge 0}}_{\ge 0}$$

$$\le v^* - v^{\pi_k}$$

$$\|g(\pi_k, \pi^*)\|_\infty \le \|v^* - v^{\pi_k}\|_\infty \le \gamma^k \|v^* - v^{\pi_0}\|_\infty$$

$$\|g(\pi_k,\pi_*)\|_\infty \le \gamma^k \|v^* - v^{\pi_0}\|_\infty =$$

$$(I-\gamma P_{\pi_0})^{-1} g(\pi_0,\pi_*) = v^{\pi_0} - v^*$$

$$= \gamma^k \|(I-\gamma P_{\pi_0})^{-1} g(\pi_0,\pi_*)\|_\infty$$

$$\le \gamma^k \|\underbrace{(I-\gamma P_{\pi_0})^{-1}}_{\|\sum \gamma^i P_{\pi_0}^i\|_\infty \le \sum_{i\ge 0}\gamma^i = \frac{1}{1-\gamma}}\|_\infty \|g(\pi_0,\pi_*)\|_\infty$$

$$\le \frac{\gamma^k}{1-\gamma}\|g(\pi_0,\pi_*)\|_\infty = \frac{\gamma^k}{1-\gamma}(-g(\pi_0,\pi_*)(s_0)) > 0$$

$$\underbrace{(I-A)}^{-1} = \sum_{i=0}^{\infty} A^i$$

$$\frac{1}{1-a} = \sum_{i\ge 0} a^i$$

$$-g(\pi_0,\pi_*)(s_0) = \|g(\pi_0,\pi_*)\|_\infty > 0 \quad \pi_0 \ne \pi^*$$

Claim: s_0 suits the bill!

$$\frac{\gamma^k}{1-\gamma}(-g(\pi_0,\pi_*)(s_0)) \ge \|g(\pi_k,\pi_*)\|_\infty$$

$$\ge -g(\pi_k,\pi_*)(s_0)$$

g expand

$$A[\;\;] > \frac{\gamma^k}{1-\gamma}[v^* - T_{\pi_0}v^*](s_0)$$

$$\ge [v^* - T_{\pi_k}v^*](s_0)$$

$$k \ge k^*(\gamma) \to \boxed{\frac{\gamma^k}{1-\gamma} < 1} \Rightarrow [v^* - T_{\pi_0}v^*](s_0) > [v^* - T_{\pi_k}v^*](s_0)$$

$k \geq k^*(\gamma)$

$$(T_{\pi_k} v^*)(s_0) > (T_{\pi_0} v^*)(s_0)$$

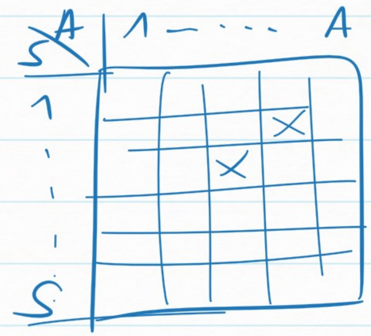
$$\underbrace{r_{\pi_k(s_0)}}(s_0) + \gamma \langle P_{\pi_k(s_0)}(s_0), v^* \rangle$$

$$> \underline{r_{\pi_0(s_0)}(s_0) + \gamma \langle P_{\pi_0(s_0)}(s_0), v^* \rangle}$$

$$\Rightarrow \pi_k(s_0) \neq \pi_0(s_0)$$

$k^*(\gamma)$ $\qquad \cancel{\frac{\gamma^k}{1-\gamma} < 1}$

$$k^*(\gamma) := \lceil H_{\gamma,1} \rceil = \boxed{\frac{\log\left(\frac{1}{1-\gamma}\right)}{\log(1/\gamma)} \lesssim \frac{\log(1/(1-\gamma))}{1-\gamma}}$$



k^*

$$\underbrace{SA}_{\text{Total \# SA pairs}} - S$$

Thm: $k \geq O\left(\frac{SA-S}{1-\gamma} \log\left(\frac{1}{1-\gamma}\right)\right)$

$\pi_k = \pi^*$

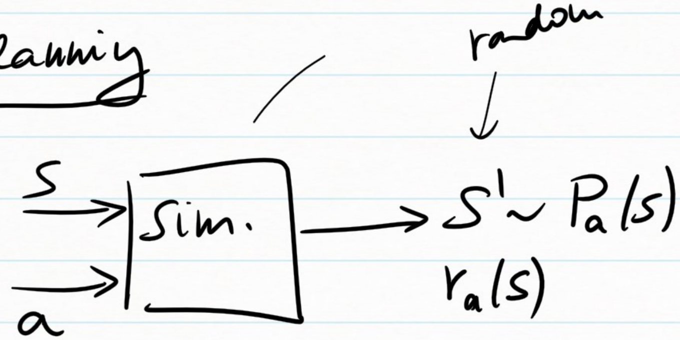
Bruno Scherrer / . . . Hansen

Yinyu Ye

2013

2016 (MOR)

Local planning



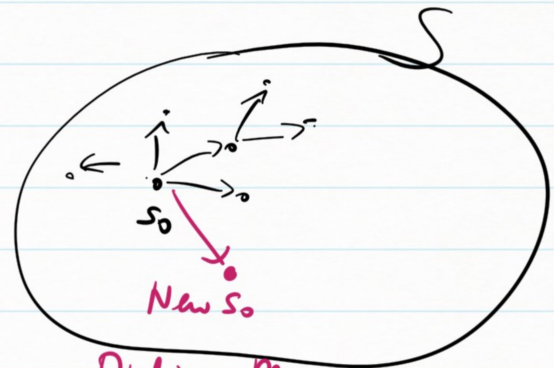
Local planner

Input:	$s_0 \in S$
Output:	$A \in \mathcal{A}$

$$\mathbb{P}_{s_0}(A = a) = \pi(a|s)$$

$$\boxed{v^\pi \geq v^* - \delta 1}$$

$\boxed{\delta > 0 \text{ planner}}$

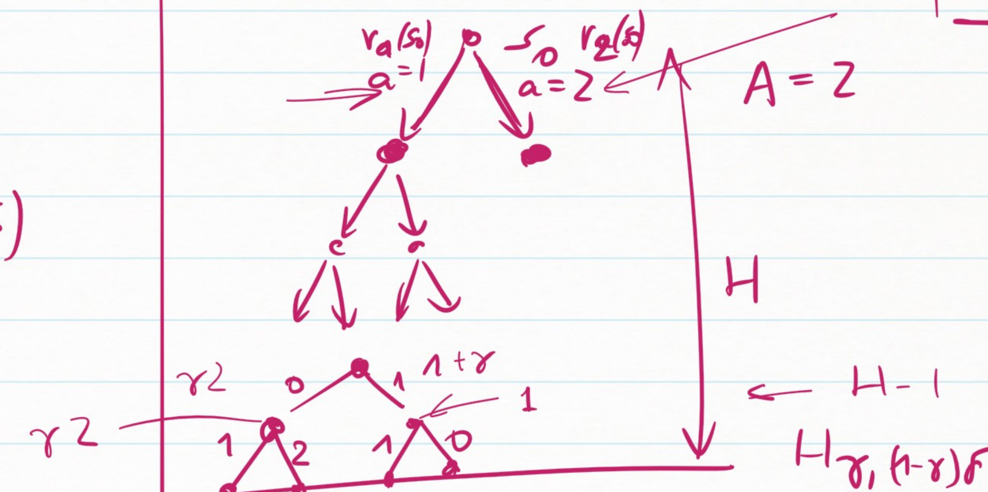


Online Planning

$\boxed{\text{Independent of } |S|}$

Deterministic MDP

$\boxed{A^H}$



$\boxed{\text{Stochastic Case?}}$