Policy Iteration

Init $\pi_0 : S \to A$.

For $k = 1, 2, 3, \ldots$

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

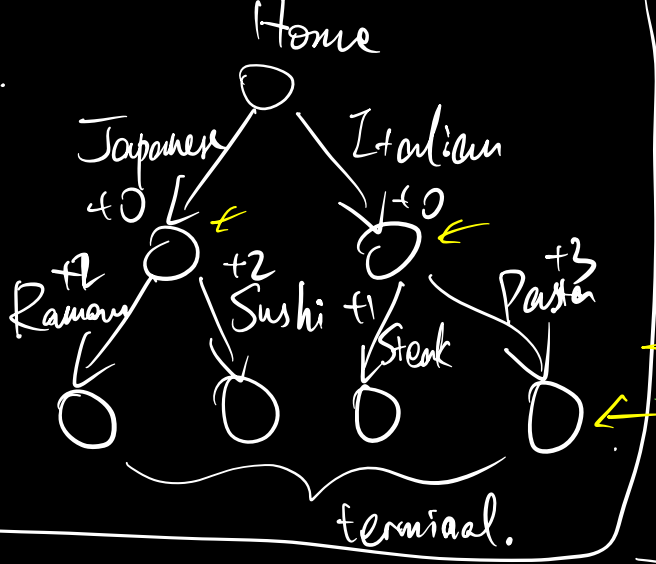
Policy evaluation: Compute $Q^{\pi_{k-1}} \in \mathbb{R}^{S \times A}$

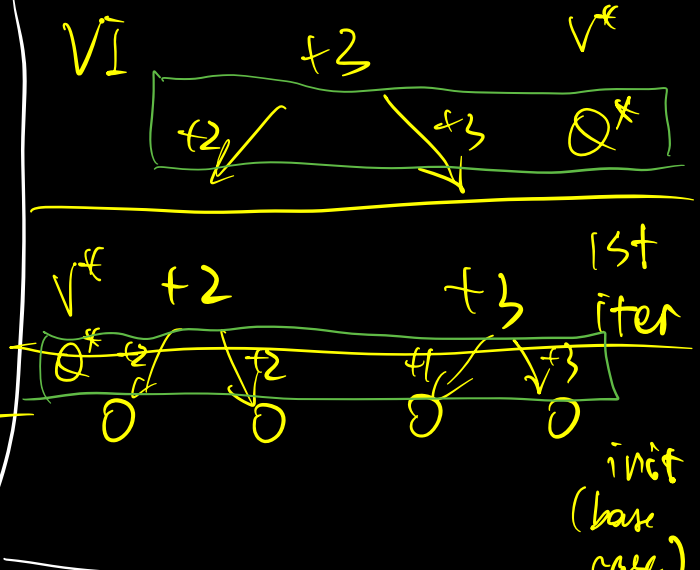
Policy Improvement: $\pi_k \leftarrow \pi_{Q^{\pi_{k-1}}}$

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

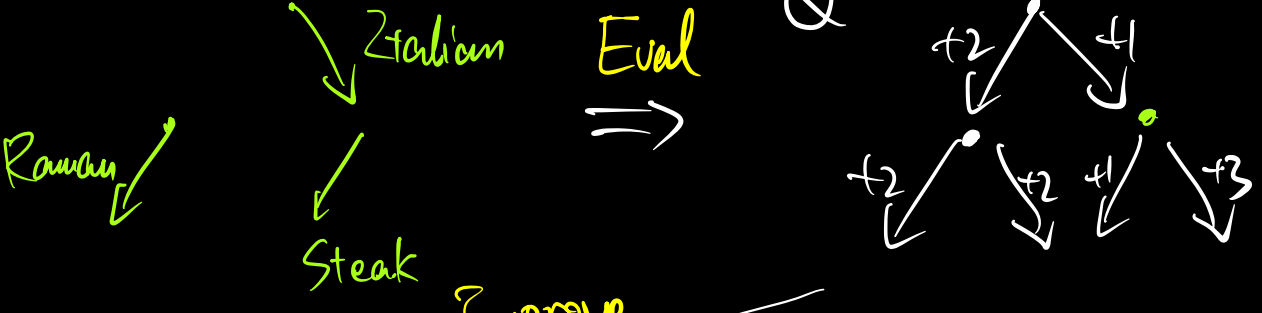
V_1

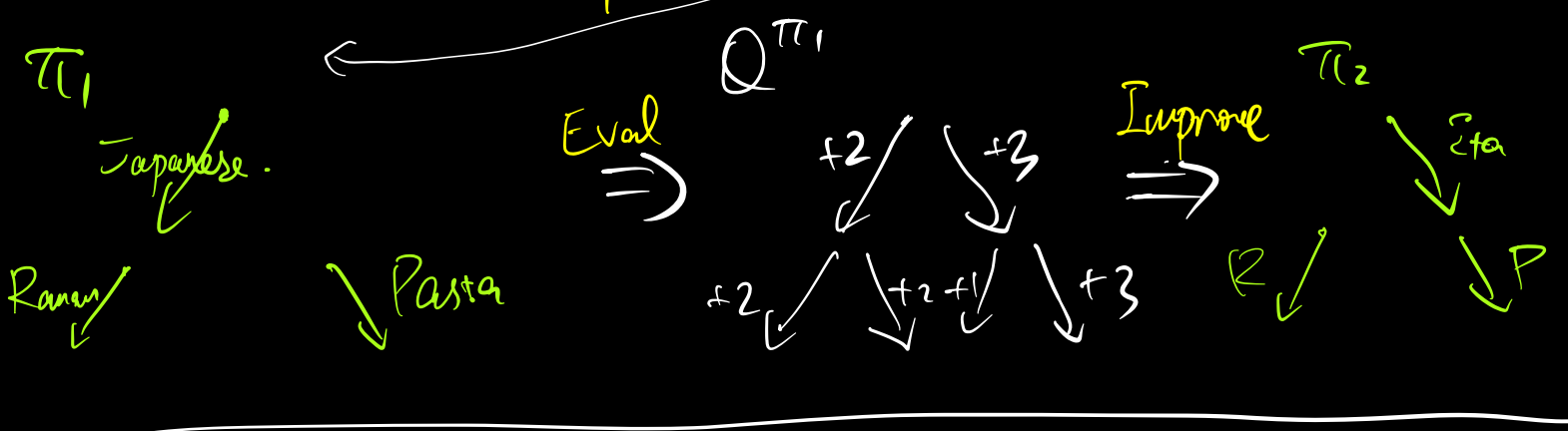
Example:





PI: π_0:





$\gamma = 0.9$.

$f_0 = 0, \quad f_1 = 1, \quad f_2 = 1 + \gamma \cdot 1 = 1.9.$

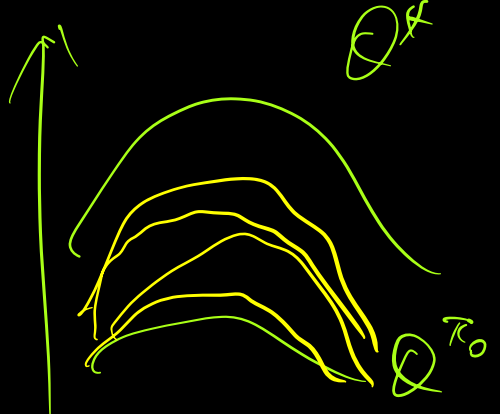
$f_3 = 1 + \gamma + \gamma^2, \ \ldots$

Monotone Policy Improvement: $\forall k$.

$$V^{\pi_k}(s) \geq V^{\pi_{k-1}}(s) \quad \forall s.$$

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

$\Rightarrow$ #iter $\leq |A|^{|S|}$.



Lemma: T^π is monotone. that is,

$\forall f \le f'$. $T^\pi f \le T^\pi f'$.

$(T^\pi f)(s,a) - (T^\pi f')(s,a)$.

$= \big(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[f(s', \pi)]\big)$

$- \big(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[f'(s', \pi)]\big)$

$= \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[f(s', \pi) - f'(s', \pi)]$.

≤ 0.

Want to show: $Q^{\pi_k} \le Q^{\pi_{k+1}}$.

$Q^{\pi_k} = T^{\pi_k} Q^{\pi_k} \le T Q^{\pi_k} = T^{\pi_{k+1}} Q^{\pi_k}$ (alg.)

$(Tf)(s,a) = R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)}[\max_{a'} f(s', a')]$

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

$\forall f, Tf = T^{\pi_f} f$.

$$Q^{\pi_k} \le \mathcal{T}^{\pi_{k+1}} Q^{\pi_k}$$

$$\Downarrow$$

$$\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \le \mathcal{T}^{\pi_{k+1}} \left(\mathcal{T}^{\pi_{k+1}} Q^{\pi_k} \right)$$

$$\Downarrow \cdots$$

$$Q^{\pi_k} \le \underbrace{\left(\mathcal{T}^{\pi_{k+1}} \right)^{\infty} Q^{\pi_k}}_{\text{"VI"}} = Q^{\pi_{k+1}}.$$

Recall: VI for Q^π: $\forall f_0 \in \mathbb{R}^{S \times A}$

$$f_k \leftarrow \mathcal{T}^\pi f_{k-1}.$$

$$f_k = \left(\mathcal{T}^\pi \right)^k f_0$$

as $k \to \infty$, $f_k \to Q^\pi$.