How to solve for V^*/Q^* | Value Iteration (VI)

Definition: Bellman optimality operator $\mathcal{T}: \mathbb{R}^{S\times A} \to \mathbb{R}^{S\times A}$
(function → function)

$\forall f \in \mathbb{R}^{S\times A}$, $(\mathcal{T}f) \in \mathbb{R}^{S\times A}$.

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

Bellman opt. eq. for Q^*: $\boxed{Q^* = \mathcal{T}Q^*}$

(Q^* is the fixed point of operator $\mathcal{T}$).

VI for solving Q^*

- Initialize f_0 arbitrarily (e.g. $f_0 = \vec{0}$)
- for $k = 1, 2, 3, \cdots$, $f_k = \mathcal{T}f_{k-1}$.

Intuition: when $f_{k-1} = Q^*$, $f_k = \mathcal{T}Q^* = Q^*$.

Convergence of VI:

$$\|f\|_\infty = \max_{s,a} |f(s,a)|.$$

Proposition: $\mathcal{T}$ is a γ-contraction under $\|\cdot\|_\infty$, That is,

$\forall f, f' \in \mathbb{R}^{S\times A}$, $\|\mathcal{T}f - \mathcal{T}f'\|_\infty \le \gamma \cdot \|f - f'\|_\infty$

Use contraction of $\mathcal{T}$ to show convergence of VI:

$$\|f_k - Q^*\|_\infty = \|\mathcal{T}f_{k-1} - \mathcal{T}Q^*\|_\infty$$

$$\le \gamma \cdot \|f_{k-1} - Q^*\|_\infty \le \gamma^2 \cdot \|f_{k-2} - Q^*\|_\infty \le \cdots \le \gamma^k \|f_0 - Q^*\|_\infty$$

$$\le \gamma^k \cdot \frac{R_{max}}{1-\gamma} \quad (\text{for } f_0 = \vec{0}).$$

Proof of contraction: $\forall s, a$.

$$|(\mathcal{T}f)(s,a) - (\mathcal{T}f')(s,a)|$$

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

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

$$\le \gamma \cdot \max_{s' \in S} \Big| \max_{a'} f(s',a') - \max_{a'} f'(s',a') \Big| .$$

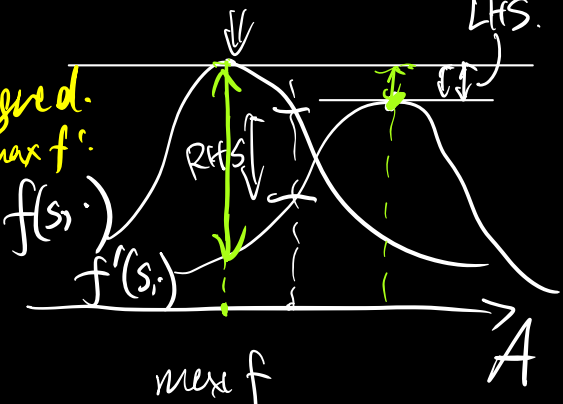
$$= \gamma \cdot \max_{s \in S} \Big| \max_{a} f(s,a) - \max_{a} f'(s,a) \Big| . \quad \searrow$$

(want to show) $\overset{?}{\le} \gamma \max_{s \in S} \max_{a \in A} |f(s,a) - f'(s,a)|$.

Suffice to show: $\forall s$.

$$\Big|\max_a f(s,a) - \max_a f'(s,a)\Big| \le \max_a |f(s,a) - f'(s,a)| .$$

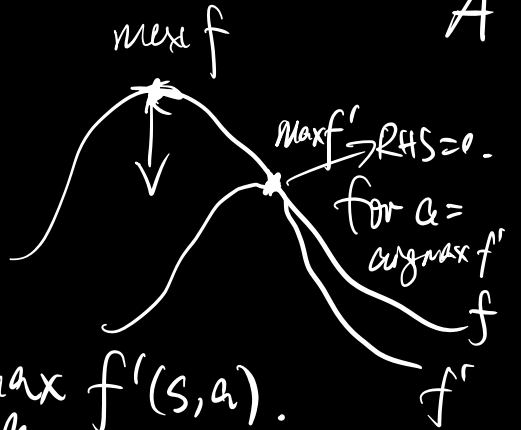
actions not aligned. argmax f vs. argmax f'.



aligned.

w.l.o.g. $\max_a f(s,a) \ge \max_a f'(s,a)$.

Let $a^* \overset{=}{=} \operatorname{argmax}_a f(s,a)$.



$$\max_a f(s,a) - \max_a f'(s,a) = f(s,a^*) - \max_a f'(s,a).$$

$$\le f(s,a^*) - f'(s,a^*) \le \max_a |f(s,a) - f'(s,a)| . \quad \blacksquare$$

Remarks: $\|Q^* - f_k\|_\infty \le \gamma^k \cdot \|Q^* - f_0\|_\infty$. b/c $\mathcal{T}$ is contraction.

There are other "VI" algorithms.

- $V^*(s) = \max_a \left(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [V^*(s')] \right)$.

$$:= (\mathcal{T} V^*)(s).$$

$\rightarrow$ abuse of notation.

VI for V^*: init $f_0 \in \mathbb{R}^S$. $f_k \leftarrow \mathcal{T} f_{k-1}$.

- $V^\pi = R^\pi + \gamma P^\pi V^\pi$

Define $\mathcal{T}^\pi : \mathbb{R}^S \to \mathbb{R}^S$. $\forall f \in \mathbb{R}^S$.

Bellman operator for policy π.

$$\mathcal{T}^\pi f = R^\pi + \gamma P^\pi f. \Rightarrow V^\pi = \mathcal{T}^\pi V^\pi.$$

Also for Q^π: $Q^\pi = \mathcal{T}^\pi Q^\pi$.

↖ abuse of notation

VI for Q^*: as $k \uparrow$, $\|f_k - Q^*\|_\infty \downarrow$.

$\pi_f(s) = \operatorname{argmax}_a f(s,a)$ "greedy policy"

We know: $\pi_{Q^*} = \pi^*$.

But we only have $f_k \approx Q^*$.

How to induce a good policy? Naive: π_{f_k}.

Q: Is this near-optimal?

A: YES!

Lemma: $\forall f \in \mathbb{R}^{S\times A}$. $\|V^* - V^{\pi_f}\|_\infty \le \frac{2\|f - Q^*\|_\infty}{1-\gamma}$.

Proof: $\forall s \in S$.

"real": V^π, Q^π, V^*, Q^*

"arbitrary": f, f'

$$V^*(s) - V^{\pi_f}(s)$$

$$= \underbrace{Q^*(s,\pi^*) - Q^*(s,\pi_f)}_{(i)} + \underbrace{Q^*(s,\pi_f) - Q^{\pi_f}(s,\pi_f)}_{(ii)}.$$

$Q^*(s,\pi_f) \ge 0$

$$\le \underbrace{Q^*(s,\pi^*) - f(s,\pi^*) + f(s,\pi_f) - Q^*(s,\pi_f)}_{(i)}.$$

actions aligned

$$\underbrace{\left(R(s,\pi_f) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,\pi_f)}[V^*(s')]\right) - \left(R(s,\pi_f) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,\pi_f)}[V^{\pi_f}(s')]\right)}_{(ii)}$$

$$\le \underbrace{2\|f - Q^*\|_\infty}_{(i)} + \underbrace{\gamma \mathbb{E}_{s'\sim P(\cdot|s,\pi_f)}[V^*(s') - V^{\pi_f}(s')]}_{(ii)}.$$

$V^*(s') - V^{\pi_f}(s')$: non-negative.

$$\le 2\cdot\|f - Q^*\|_\infty + \gamma \cdot \max_{s'}(V^*(s') - V^{\pi_f}(s'))$$

$$= 2\cdot\|f - Q^*\|_\infty + \gamma \cdot \|V^* - V^{\pi_f}\|_\infty.$$

Therefore:

$$\|V^* - V^{\pi_f}\|_\infty \le 2\cdot\|f - Q^*\|_\infty + \gamma\cdot\|V^* - V^{\pi_f}\|_\infty.$$

$$\Rightarrow \|V^* - V^{\pi_f}\|_\infty \le \frac{2}{1-\gamma}\cdot\|f - Q^*\|_\infty. \quad \blacksquare$$

Remarks:

1. $\|f_k - Q^*\|_\infty \le \gamma^k \|f_0 - Q^*\|_\infty \le \gamma^k \cdot \frac{R_{max}}{1-\gamma}$!! V_{max}.

$\Rightarrow$ if we want $\|f_k - Q^*\|_\infty \le \varepsilon$. how large should k be?

$$\gamma^k V_{max} \le \varepsilon \Rightarrow \gamma^k \le \frac{\varepsilon}{V_{max}}.$$

$$k \ge \log_{1/\gamma} \frac{V_{max}}{\varepsilon} \Rightarrow k \ge \frac{\ln \frac{V_{max}}{\varepsilon}}{\ln 1/\gamma} \approx \frac{\ln \frac{V_{max}}{\varepsilon}}{1-\gamma} = O\left(\frac{1}{1-\gamma}\right).$$

"effective horizon"

2. Stopping criterion: $\|f_k - f_{k+1}\|_\infty \le \varepsilon'$.

Why? $\|f_k - f_{k+1}\|_\infty = \|f_k - \mathcal{T} f_k\|_\infty$.

Bellman error / residual of f_k.

$Q^* = \mathcal{T} Q^*$.

$$\|f_k - Q^*\|_\infty$$
$$= \|f_k - \mathcal{T} f_k + \mathcal{T} f_k - \mathcal{T} Q^*\|_\infty$$
$$\le \|f_k - \mathcal{T} f_k\|_\infty + \|\mathcal{T} f_k - \mathcal{T} Q^*\|_\infty$$
$$\le \varepsilon' + \gamma \cdot \|f_k - Q^*\|_\infty$$

$$\Rightarrow \|f_k - Q^*\|_\infty \le \frac{\varepsilon'}{1-\gamma}.$$

3. Similar algorithm / analyses for V^*, V^π, Q^π.

$Q^\pi = \mathcal{T}^\pi Q^\pi$ $\rightarrow$ "VI" alg: $f_k \leftarrow \mathcal{T}^\pi f_{k-1}$.

Alternative Proof for convergence of VI

"truncated value-func" $V^{\pi,H}(s) = \mathbb{E}\left[\sum_{t=1}^{H} \gamma^{t-1} r_t \mid s_1 = s, \pi\right]$.

(in fact, $V^{\pi}(s) = V^{\pi,\infty}(s)$).

Consider VI for solving V^*.

Init $f_0 = \vec{0} \in \mathbb{R}^S$. $f_k \leftarrow T f_{k-1}$.

Claim: $f_k(s) = \max_{\pi} V^{\pi,k}(s) =: V^{*,k}(s)$.

(π) ← "all policies, possibly non-stationary"

Examples: $f_k(s) = \max_a \left(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [f_{k-1}(s')] \right)$

$k=0$. $f_0 = V^{*,0} = \vec{0}$.

$k=1$. $f_1(s) = \max_a R(s,a) = V^{*,1}(s)$

$k=2$. $f_2(s) = V^{*,2}(s) =$

$\max_a \left(R(s,a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} [\max_{a'} R(s',a')] \right)$.

depends on 1st time step.

$k=3, \ldots$

Remark: optimal policy for $V^{*,H}$ is non-stationary.

Want to upper bound: $\|f_K - V^*\|_\infty$

$f_K = V^{*,k}$. $\quad V^{*,k} - V^*$.

Proof: (1) $V^{*,k} \le V^*$

$$V^{*,k} = \max_\pi \mathbb{E}\left[\sum_{t=1}^{k} \gamma^{t-1} \cdot r_t \,\middle|\, s_1 = s, \pi\right].$$

$$V^* = V^{*,\infty} = \max_\pi \mathbb{E}\left[\sum_{t=1}^{\infty} \cdots\right]$$

$V^{*,k} \le V^*$ follows directly from $r_t \ge 0$.

(2) $V^{*,k} \ge V^* - \gamma^k V_{max}$ $\quad = \frac{R_{max}}{1-\gamma}$.

$$V^{*,k} = \max_\pi \mathbb{E}\left[\sum_{t=1}^{k} \gamma^{t-1} \cdot r_t \,\middle|\, s_1 = s, \pi\right]$$

$$\ge \mathbb{E}\left[\sum_{t=1}^{k} \gamma^{t-1} r_t \,\middle|\, s_1 = s, \pi^*\right]$$ ← optimal for infinite horizon

$$= \mathbb{E}\left[\sum_{t=1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s, \pi^*\right]$$

$$- \mathbb{E}\left[\sum_{t=k+1}^{\infty} \gamma^{t-1} r_t \,\middle|\, s_1 = s, \pi^*\right].$$

$$= V^*(s) - \gamma^k \mathbb{E}\left[\sum_{t=k+1}^{\infty} \gamma^{t-k-1} r_t \,\middle|\, s_1 = s, \pi^*\right]$$

$t' := t - k$.

$$\sum_{t'=1}^{\infty} \gamma^{t'-1} r_{t'+k} \,\middle|\, s_{1-k} = s, \pi^*$$ ignore.

$$\ge V^*(s) - \gamma^k \cdot V_{max}.$$

$$V^*(s) - \gamma^k V_{max} \le V^{*,k}(s) \le V^*(s).$$

$$f_k''(s)$$

$$\Rightarrow \|V^* - f_k\|_\infty \le \gamma^k V_{max}$$

(same guarantee as before)

$V_{max} \le \sum_{t'=1}^{\infty} \gamma^{t'-1} R_{max} = \frac{1}{1-\gamma} R_{max} \le \frac{R_{max}}{1-\gamma}$.

Remark: for Q^*, we have.

$$\|V^* - V^{\pi_{f_k}}\|_\infty \le \frac{2 \cdot \|f_k - Q^*\|}{1-\gamma} \le \frac{2 \cdot \gamma^k V_{max}}{1-\gamma}.$$

Turns out you can save $\frac{2}{1-\gamma}$ factor!

Trick: instead of outputting π_{f_k}.

output instead $\pi_{NS} = \pi_{f_k} \circ \pi_{f_{k-1}} \circ \pi_{f_{k-2}} \circ \cdots \circ \pi_{f_1}$ ("non-stat")

$\Leftrightarrow$ $a_1 \sim \pi_{f_k}$, $a_2 \sim \pi_{f_{k-1}}$, $a_3 \sim \pi_{f_{k-2}}$, ...

$a_k \sim \pi_{f_1}$, $a_{k+1:\infty} \sim$ arbitrary.

Claim: this policy optimizes $V^{*,k}$.

$$V^{\pi_{NS}} \ge V^{\pi_{NS},k} = V^{*,k} \ge V^* - \gamma^k R_{max}.$$

$$\Rightarrow \|V^* - V^{\pi_{NS}}\|_\infty \le \gamma^k R_{max}. \quad \triangleleft$$

In practice, γ is often introduced for convenience.

Given finite-horizon undiscounted problem (w/ H) can turn into inf-horizon discounted, $H \approx O\left(\frac{1}{1-\gamma}\right)$.