Strategic Exploration in Large Scale MDPs

Recap on Bellman Error and Bellman Operator

Bellman error of $f(s, a)$

Recap on Bellman Error and Bellman Operator

Bellman error of $f(s, a)$

$$BE(s,a) = f(s,a) - \left(r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} f(s',a') \right)$$

Recap on Bellman Error and Bellman Operator

Bellman error of $f(s, a)$

$$BE(s,a) = f(s,a) - \left(r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} f(s', a') \right)$$

$= 0, \forall s, a$

$f = Q^*$

If $BE(s,a) = 0, \forall s, a$, then $f(s,a) = Q^\star(s,a), \forall s, a$

Recap on Bellman Error and Bellman Operator

Bellman error of $f(s, a)$

$$BE(s, a) = f(s, a) - \left(r(s, a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} f(s', a') \right)$$

If $BE(s, a) = 0, \forall s, a$, then $f(s, a) = Q^{\star}(s, a), \forall s, a$

Bellman Operator $\mathcal{T}$ **of** f

Recap on Bellman Error and Bellman Operator

Bellman error of $f(s, a)$

$$BE(s, a) = f(s, a) - \left(r(s, a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} f(s', a') \right)$$

If $BE(s, a) = 0, \forall s, a$, then $f(s, a) = Q^\star(s, a), \forall s, a$

Bellman Operator $\mathcal{T}$ of f

$$\mathcal{T} f : S \times A \to \mathbb{R}$$

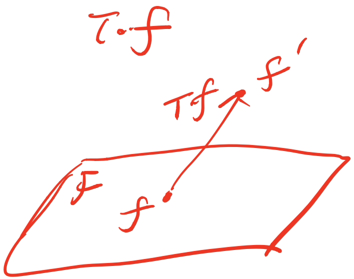
Recap on Bellman Error and Bellman Operator

Bellman error of $f(s, a)$

$$BE(s,a) = f(s,a) - \left(r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a'} f(s',a') \right)$$

If $BE(s,a) = 0, \forall s, a$, then $f(s,a) = Q^\star(s,a), \forall s, a$

Bellman Operator $\mathcal{T}$ of f



$$\mathcal{T}f : S \times A \to \mathbb{R}$$

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

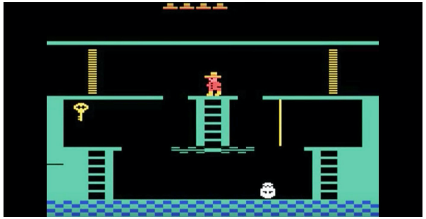
Notations

Probability of π visiting (s, a) at time step h: $d_h^\pi(s, a)$

Setting

Finite horizon episodic MDP $\left\{\{X_h\}_{h=0}^{H}, \{A_h\}_{h=0}^{H-1}, H, x_0, r, P\right\}$

State space X_h is extremely large:

$x =$ 

Not acceptable: poly $(|X|)$

Q: can we generalize using function approximation

Let's set up function class in RL setting

Mode-free & Value-based

We will consider **Q function class** for now (and model class later)

$$\mathcal{F} \subset X \times A \mapsto [0,1]$$

$F = \{f_1, f_2 \cdots f_{|F|}\}$ $\approx Q^*$

Let's set up function class in RL setting

We will consider **Q function class** for now (and model class later)

$$\mathscr{F} \subset X \times A \mapsto [0,1]$$

Realizability assumption:

$$Q^\star \in \mathscr{F}$$

Let's set up function class in RL setting

We will consider **Q function class** for now (and model class later)

$$\mathcal{F} \subset X \times A \mapsto [0,1]$$

Realizability assumption:

$$Q^\star \in \mathcal{F}$$

$$f \approx Q^*$$

Define **policy class:** $\Pi = \{\pi : \pi(x) = \arg\max_{a \in A} f(x, a), \forall x \in X \,|\, f \in \mathcal{F}\}$

Let's set up function class in RL setting

We will consider **Q function class** for now (and model class later)

$$\mathcal{F} \subset X \times A \mapsto [0,1]$$

Realizability assumption:

$$Q^\star \in \mathcal{F}$$

Define **policy class:** $\Pi = \{\pi : \pi(x) = \arg\max_{a \in A} f(x, a), \forall x \in X \,|\, f \in \mathcal{F}\}$

i.e., each Q-approximator f induces a policy (greedy w.r.t f)

$Q^* \in \mathcal{F} \Rightarrow \pi^* = \arg\max_a Q^* \in \Pi$

Learning Goal:

We will do PAC in this lecture rather than regret.

Given approximation error ϵ and failure prob δ,
can we learn ϵ *near optimal policy* (i.e., $V^{\hat{\pi}} \geq V^\star - \epsilon$) in # of samples scaling *poly* with all relevant parameters *(here, we need poly in* $\ln(|\mathscr{F}|)$*)*

poly $|X|$ ×
poly $|A|$

How to check if a Q-approximator is good?

We define **average** Bellman error below:

$$\mathscr{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

How to check if a Q-approximator is good?

We define **average** Bellman error below:

$$\mathcal{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \right]$$

$\bar{f}$: defines roll-in distribution over x_h, a_h at stage h.

How to check if a Q-approximator is good?

$f, \bar{f} \in \mathcal{F}$

We define **average** Bellman error below:

$$\mathcal{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \right]$$

$\bar{f}$: defines roll-in distribution over x_h, a_h at stage h.

Bellman error measures consistency in one-step Bellman backup, e.g., $\mathcal{E}(Q^\star; \bar{f}, h) = 0$

$BE_{Q^\star}(x, a) = 0$

How to check if a Q-approximator is good?

We define **average** Bellman error below:

$$\mathscr{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

$\bar{f}$: defines roll-in distribution over x_h, a_h at stage h.

Bellman error measures consistency in one-step Bellman backup, e.g., $\mathscr{E}(Q^\star; \bar{f}, h) = 0$

Hence, any f such that $\mathscr{E}(f; \pi, h) \neq 0$, is an incorrect $Q^\star$ approximator

Optimism Led Iterative Value Function Elimination (OLIVE)

Optimism Led Iterative Value Function Elimination (OLIVE)

pre- defined

Initialize $\mathscr{F}_0 = \mathscr{F}$

Optimism Led Iterative Value Function Elimination (OLIVE)

Initialize $\mathscr{F}_0 = \mathscr{F}$

For $t = 0, \ldots$

Optimism Led Iterative Value Function Elimination (OLIVE)

Initialize $\mathscr{F}_0 = \mathscr{F}$

For $t = 0, \ldots$

$$f_t = \arg\max_{f \in \mathscr{F}_t} \left(\max_a f(x_0, a) \right)$$

← optimism

$f \approx Q^*$

$\approx V^*(x_0)$

Optimism Led Iterative Value Function Elimination (OLIVE)

Initialize $\mathscr{F}_0 = \mathscr{F}$

For $t = 0, \ldots$

$$f_t = \arg\max_{f \in \mathscr{F}_t} \left(\max_a f(x_0, a) \right)$$

If $\left| \widetilde{V}^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

π_{f_t}

$\mathbb{E}\left[\sum_{h=0}^{H-1} r_h \mid \pi_{f_t}\right]$

$\approx V^*$

Optimism Led Iterative Value Function Elimination (OLIVE)

Initialize $\mathscr{F}_0 = \mathscr{F}$

For $t = 0, \ldots$

$$f_t = \arg\max_{f \in \mathscr{F}_t} \left(\max_a f(x_0, a) \right)$$

If $\left| \widetilde{V}^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Version space update:

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \widetilde{\mathscr{E}}(f; \pi_{f_t}, h) \leq \delta, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

Eliminate f: $\widetilde{\mathcal{E}}(f; \pi_{f_t}, h) > \delta$

$\widetilde{\mathcal{E}}(Q^\star; \pi_f, h) \approx 0$

Estimating Bellman Error under a fixed Roll-in Policy:

Given a **fixed** $\pi_{\bar{f}}$, we can evaluate all f efficiently **statistically (not computationally)**:

$$\forall f: \mathcal{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \right]$$

Estimating Bellman Error under a fixed Roll-in Policy:

Given a **fixed** $\pi_{\bar{f}}$, we can evaluate all f efficiently **statistically (not computationally)**:

$$\forall f: \mathscr{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}, x_{h+1} \sim P(\cdot | x_h, a_h)} \left[f(x_h, a_h) - r(x_h, a_h) - \max_{a \in \mathscr{A}} f(x_{h+1}, a) \right]$$

Estimating Bellman Error under a fixed Roll-in Policy:

Given a **fixed** $\pi_{\bar{f}}$, we can evaluate all f efficiently **statistically (not computationally)**:

$$\forall f: \mathscr{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}, x_{h+1} \sim P(\cdot | x_h, a_h)} \left[f(x_h, a_h) - r(x_h, a_h) - \max_{a \in \mathscr{A}} f(x_{h+1}, a) \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \left[f(x_h^i, a_h^i) - r(x_h^i, a_h^i) - \max_{a \in \mathscr{A}} f(x_{h+1}^i, a) \right]$$

$x_h^i, a_h^i \sim \pi_{\bar{f}}$

$x_{h+1}^i \sim P(\cdot | x_h^i, a_h^i)$

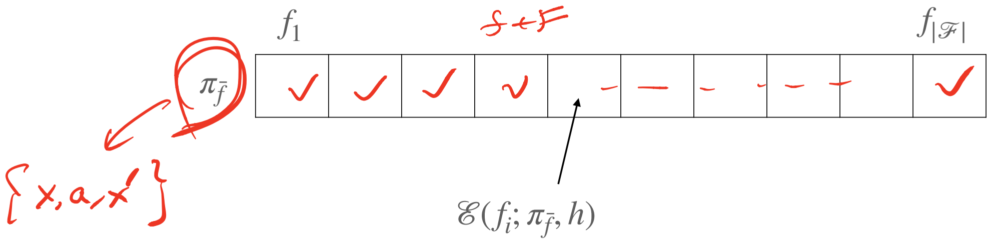
Estimating Bellman Error under a fixed Roll-in Policy:

Given a **fixed** $\pi_{\bar{f}}$, we can evaluate all f efficiently **statistically (not computationally)**:

$$\forall f : \mathscr{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}, x_{h+1} \sim P(\cdot | x_h, a_h)} \left[f(x_h, a_h) - r(x_h, a_h) - \max_{a \in \mathscr{A}} f(x_{h+1}, a) \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \left[f(x_h^i, a_h^i) - r(x_h^i, a_h^i) - \max_{a \in \mathscr{A}} f(x_{h+1}^i, a) \right]$$



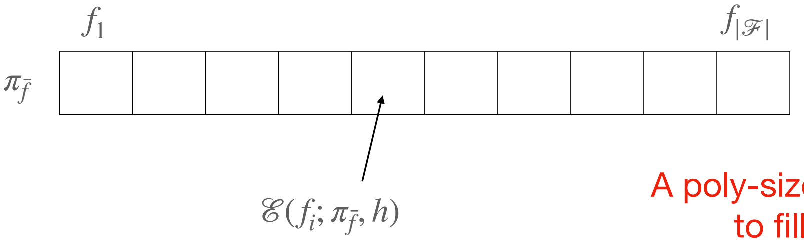
Estimating Bellman Error under a fixed Roll-in Policy:

Given a **fixed** $\pi_{\bar{f}}$, we can evaluate all f efficiently **statistically (not computationally)**:

$$\forall f : \mathscr{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}, x_{h+1} \sim P(\cdot | x_h, a_h)} \left[f(x_h, a_h) - r(x_h, a_h) - \max_{a \in \mathscr{A}} f(x_{h+1}, a) \right]$$

$$\approx \frac{1}{N} \sum_{i=1}^{N} \left[f(x_h^i, a_h^i) - r(x_h^i, a_h^i) - \max_{a \in \mathscr{A}} f(x_{h+1}^i, a) \right]$$



A poly-size dataset allows us to fill up all entries

OLIVE Revisit (ignoring statistical error for simplicity)

E ✓
Δ

Initialize $\mathscr{F}_0 = \mathscr{F}$

For $t = 0, \ldots$

$f_t = \arg\max_{f \in \mathscr{F}_t} \max_a f(x_0, a)$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Version space update:

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

Δ

OLIVE Revisit (ignoring statistical error for simplicity)

Initialize $\mathscr{F}_0 = \mathscr{F}$

For $t = 0, \ldots$

$f_t = \arg\max_{f \in \mathscr{F}_t} \max_a f(x_0, a)$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Version space update:

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

1. Upon termination we succeed (due to optimism)

OLIVE Revisit (ignoring statistical error for simplicity)

Initialize $\mathcal{F}_0 = \mathcal{F}$

For $t = 0, \ldots$

$$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(x_0, a)$$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Version space update:

$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

1. Upon termination we succeed (due to optimism)

2. If not terminate, we make non-trivial progress

OLIVE Revisit (ignoring statistical error for simplicity)

Initialize $\mathcal{F}_0 = \mathcal{F}$

For $t = 0, \ldots$

$$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(x_0, a)$$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Version space update:

$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

1. Upon termination we succeed (due to optimism)

2. If not terminate, we make non-trivial progress

3. Total # of such non-trivial progress is bounded

Quality of Returned Policy upon Termination:

$$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(s_0, a)$$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon,$ return π_{f_t} ✓

$$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(s_0, a)$$

Quality of Returned Policy upon Termination:

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Note $Q^\star \in \mathcal{F}_t, \forall t,$

(we only eliminate things that are obviously wrong & $Q^\star$ **has zero bellman error everywhere**)

$BE_{Q^\star}(x,a) = 0$

$$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(s_0, a)$$

Quality of Returned Policy upon Termination:

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon,$ return π_{f_t}

Note $Q^\star \in \mathcal{F}_t, \forall t,$

(we only eliminate things that are obviously wrong & $Q^\star$ **has zero bellman error everywhere**)

$$\max_a f_t(x_0, a) \geq \max_a Q^\star(x_0, a) = V^\star(x_0)$$

$Q^\star \in \mathcal{F}_t$

$f_t = \arg\max_{f \in \mathcal{F}_t} \left(\max_a f(x_0, a) \right)$

Quality of Returned Policy upon Termination:

$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(s_0, a)$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon,$ return π_{f_t}

Note $Q^\star \in \mathcal{F}_t, \forall t,$

(we only eliminate things that are obviously wrong & $Q^\star$ **has zero bellman error everywhere**)

$$\max_a f_t(x_0, a) \geq \max_a Q^\star(x_0, a) = V^\star(x_0)$$

$$| V^{\pi_{f_t}} - \max_a f_t(x_0, a) | \leq \epsilon \Rightarrow V^{\pi_{f_t}} \geq \max_a f_t(x_0, a) - \epsilon \geq V^\star(x_0) - \epsilon$$

Quality of Returned Policy upon Termination:

$$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(s_0, a)$$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Note $Q^\star \in \mathcal{F}_t, \forall t,$

(we only eliminate things that are obviously wrong & $Q^\star$ **has zero bellman error everywhere**)

$$\max_a f_t(x_0, a) \geq \max_a Q^\star(x_0, a) = V^\star(x_0)$$

$$| V^{\pi_{f_t}} - \max_a f_t(x_0, a) | \leq \epsilon \Rightarrow V^{\pi_{f_t}} \geq \max_a f_t(x_0, a) - \epsilon \geq V^\star(x_0) - \epsilon$$

Optimism ensures that once termination happens, we are done!

No termination means we found a bad $Q^\star$-approximator:

Claim [performance difference lemma]:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

$\mathcal{E}(f_t; \pi_{f_t}, h)$

Proof: a straight telescoping sum

No termination means we found a bad $Q^\star$-approximator:

Claim [performance difference lemma]:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

Proof: a straight telescoping sum

1

$$f_t(x_0, \pi_{f_t}(x_0)) - r(x_0, \pi_{f_t}(x_0)) - \mathbb{E}_{x_1 \sim P(\cdot | x_0, \pi_{f_t}(x_0))} \max_a f_t(x_1, a)$$

No termination means we found a bad $Q^\star$-approximator:

Claim [performance difference lemma]:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

x_0 is fixed

Proof: a straight telescoping sum

$$f_t(x_0, \pi_{f_t}(x_0)) - r(x_0, \pi_{f_t}(x_0)) - \mathbb{E}_{x_1 \sim P(\cdot | x_0, \pi_{f_t}(x_0))} \max_a f_t(x_1, a)$$

$$+ \mathbb{E}_{x_1 \sim d_1^{\pi_{f_t}}} \left[f_t(x_1, \pi_{f_t}(x_1)) - r(x_1, \pi_{f_t}(x_1)) - \mathbb{E}_{x_2 \sim P(\cdot | x_1, \pi_{f_t}(x_1))} \max_a f_t(x_2, a) \right]$$

← h = 1

No termination means we found a bad $Q^\star$-approximator:

Claim [performance difference lemma]:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

Proof: a straight telescoping sum

$$f_t(x_0, \pi_{f_t}(x_0)) - r(x_0, \pi_{f_t}(x_0)) - \mathbb{E}_{x_1 \sim P(\cdot | x_0, \pi_{f_t}(x_0))} \max_a f_t(x_1, a)$$

$$+ \mathbb{E}_{x_1 \sim d_1^{\pi_{f_t}}} \left[f_t(x_1, \pi_{f_t}(x_1)) - r(x_1, \pi_{f_t}(x_1)) - \mathbb{E}_{x_2 \sim P(\cdot | x_1, \pi_{f_t}(x_1))} \max_a f_t(x_2, a) \right]$$

$$+ \mathbb{E}_{x_2 \sim d_2^{\pi_{f_t}}} \left[f_t(x_2, \pi_{f_t}(x_2)) - r(x_2, \pi_{f_t}(x_2)) - \mathbb{E}_{x_3 \sim P(\cdot | x_2, \pi_{f_t}(x_2))} \max_a f_t(x_3, a) \right]$$

h=2

3

No termination means we found a bad $Q^\star$-approximator:

Claim [performance difference lemma]:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

$$= \sum_{h=0}^{H-1} \mathcal{E}(f_t; \pi_{f_t}, h)$$

Proof: a straight telescoping sum

$$f_t(x_0, \pi_{f_t}(x_0)) - r(x_0, \pi_{f_t}(x_0)) - \mathbb{E}_{x_1 \sim P(\cdot | x_0, \pi_{f_t}(x_0))} \max_a f_t(x_1, a)$$

$$+ \mathbb{E}_{x_1 \sim d_1^{\pi_{f_t}}} \left[f_t(x_1, \pi_{f_t}(x_1)) - r(x_1, \pi_{f_t}(x_1)) - \mathbb{E}_{x_2 \sim P(\cdot | x_1, \pi_{f_t}(x_1))} \max_a f_t(x_2, a) \right]$$

$$+ \mathbb{E}_{x_2 \sim d_2^{\pi_{f_t}}} \left[f_t(x_2, \pi_{f_t}(x_2)) - r(x_2, \pi_{f_t}(x_2)) - \mathbb{E}_{x_3 \sim P(\cdot | x_2, \pi_{f_t}(x_2))} \max_a f_t(x_3, a) \right]$$

…

$$= f_t(x_0, \pi_{f_t}(x_0)) - \mathbb{E}\left[\sum_{h=0}^{H-1} r_h \right]$$

$$= V^{\pi_{f_t}}$$

No termination means we found a bad $Q^\star$-approximator:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, \pi_{f_t}(x_h)) - r(x_h, \pi_{f_t}(x_h)) - \mathbb{E}_{x_{h+1} \sim P(\cdot \mid x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

If we do not terminate, i.e.,

$$\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \geq \epsilon,$$

then:

No termination means we found a bad $Q^\star$-approximator:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, \pi_{f_t}(x_h)) - r(x_h, \pi_{f_t}(x_h)) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

If we do not terminate, i.e.,

$$\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \geq \epsilon,$$

$= \sum_{h=0}^{H-1} \mathcal{E}(f_t; \pi_{f_t, h})$

then:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, \pi_{f_t}(x_h)) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right] \geq \epsilon$$

No termination means we found a bad $Q^\star$-approximator:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, \pi_{f_t}(x_h)) - r(x_h, \pi_{f_t}(x_h)) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right]$$

If we do not terminate, i.e.,

$$\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \geq \epsilon,$$

then:

$$\max_a f_t(x_0, a) - V^{\pi_{f_t}}(x_0) = \sum_{h=0}^{H-1} \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, \pi_{f_t}(x_h)) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right] \geq \epsilon$$

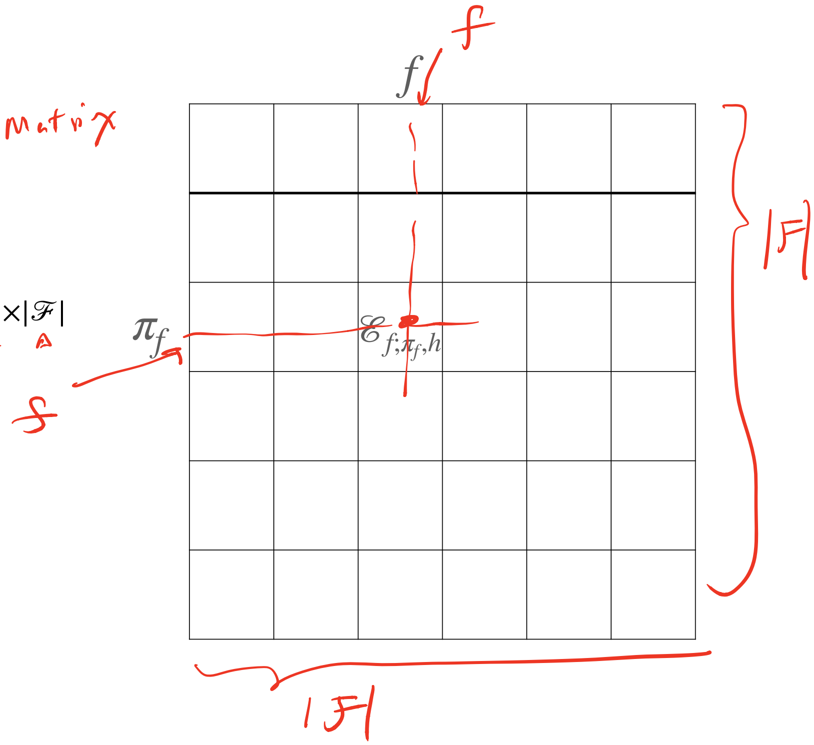
$a + b \geq 1$

$$\Rightarrow \exists h, \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{f_t}}} \left[f_t(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_a f_t(x_{h+1}, a) \right] \right] \geq \epsilon / H$$

We need to argue how many episodes we have before termination

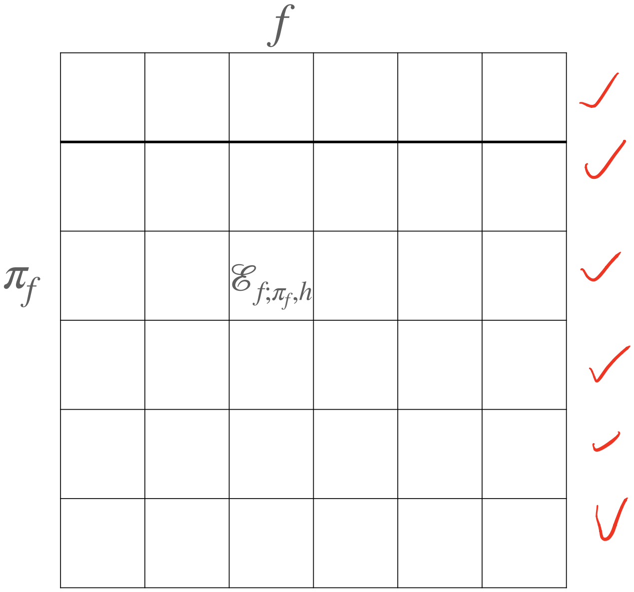
Bellman Error Matrix

$$\forall h : \mathscr{E}_h \in \mathbb{R}^{|\mathscr{F}| \times |\mathscr{F}|}$$



We need to argue how many episodes we have before termination

$\forall h : \mathscr{E}_h \in \mathbb{R}^{|\mathscr{F}|\times|\mathscr{F}|}$



Rank of this Matrix is defined as Bellman Rank

Small-Rank

Progress on Value Function Elimination

← # of Iter = Bellman Rank

$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

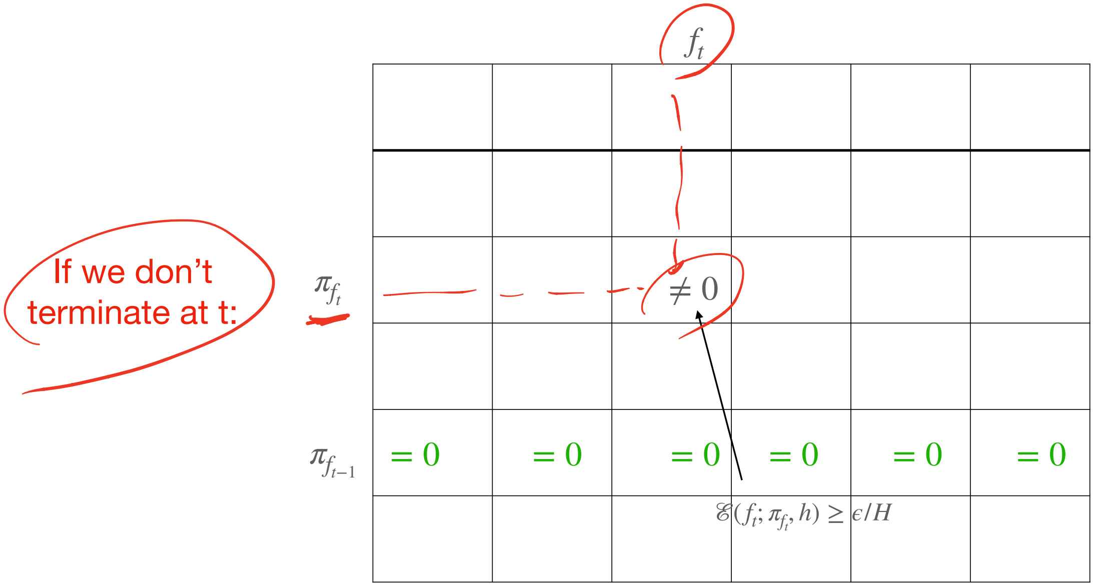
Progress on Value Function Elimination

$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

$\pi_{f_{t-1}}$	$= 0$	$= 0$	$= 0$	$= 0$	$= 0$	$= 0$

Progress on Value Function Elimination

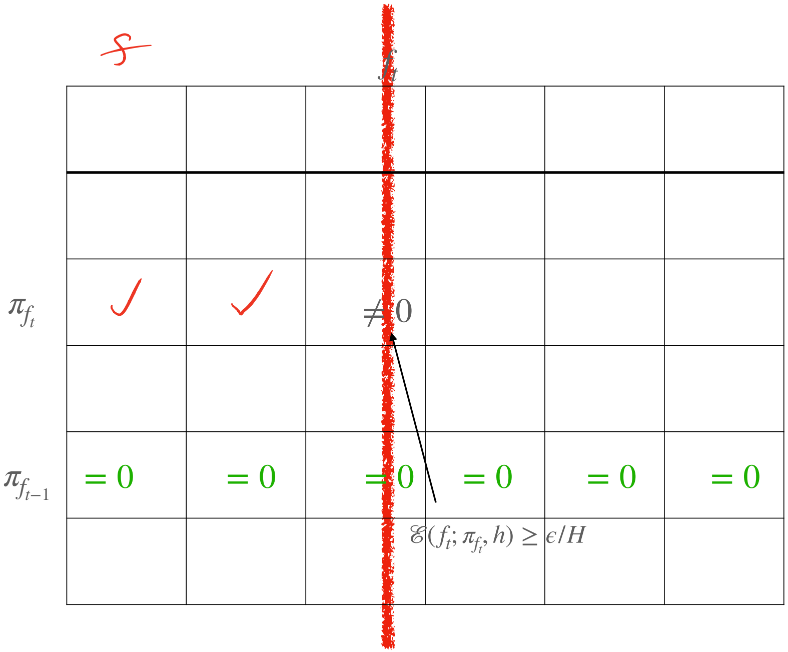
$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$



Progress on Value Function Elimination

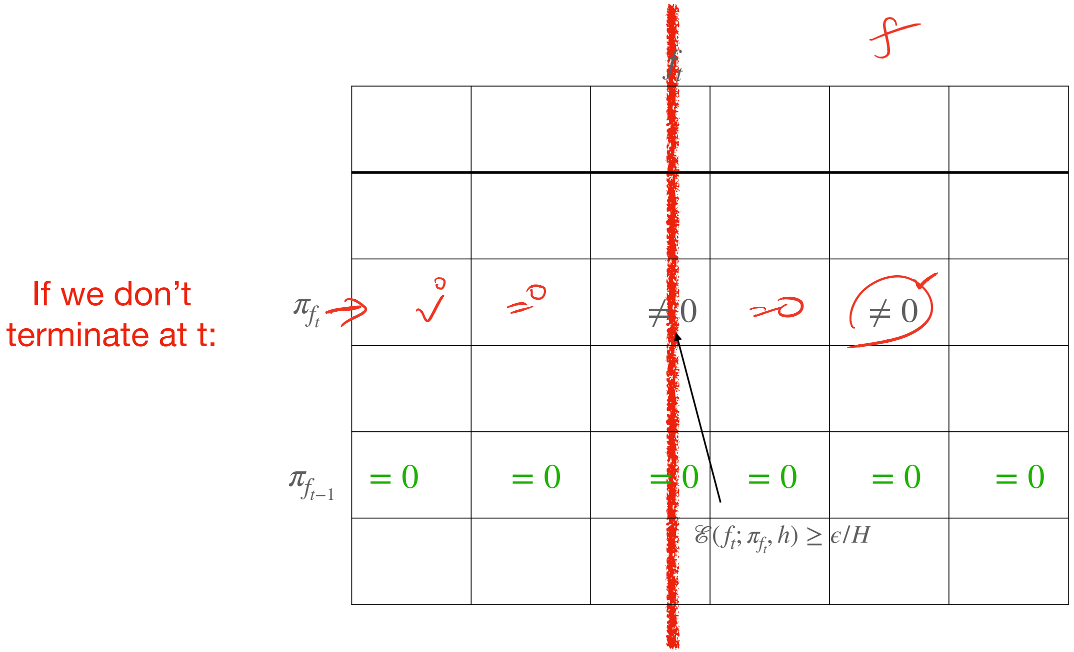
$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

If we don't terminate at t:



Progress on Value Function Elimination

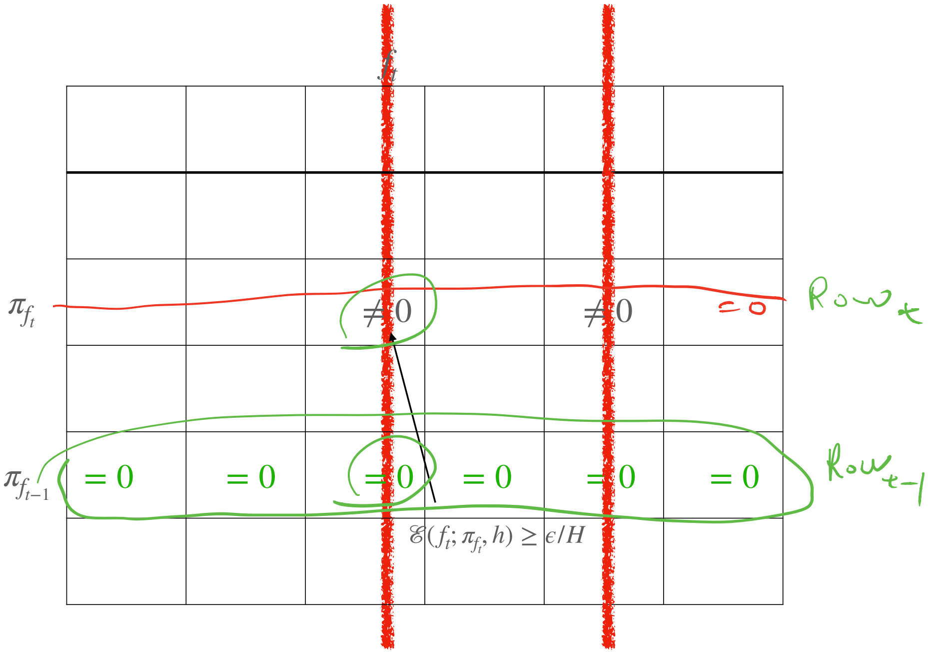
$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$



Progress on Value Function Elimination

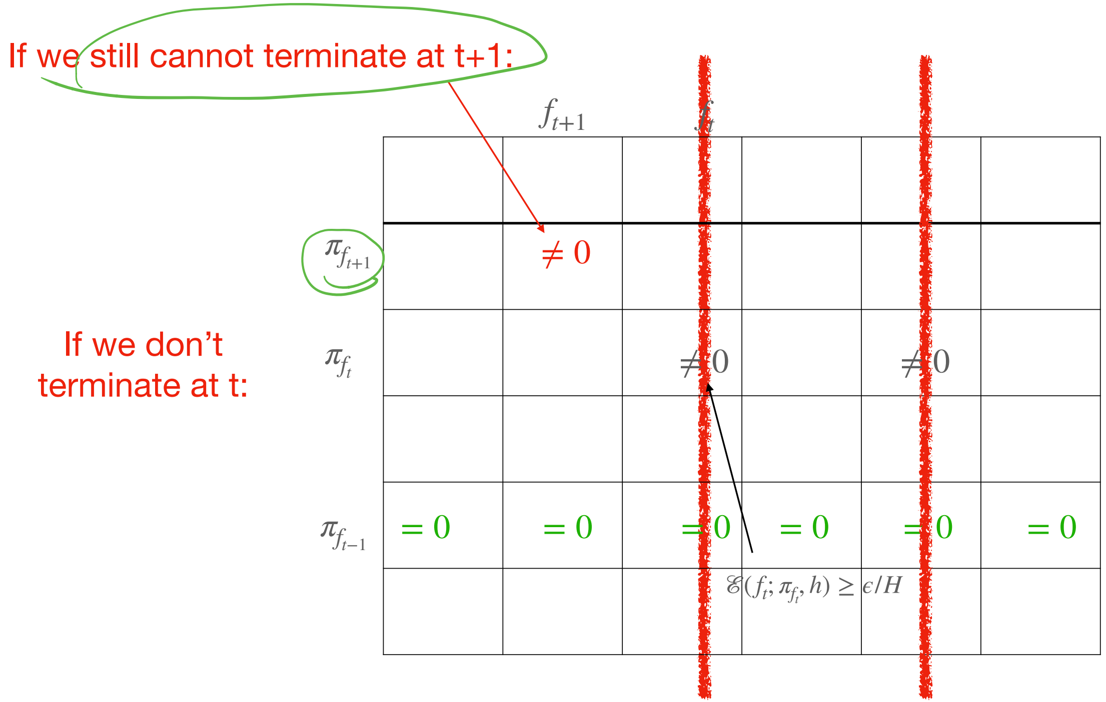
$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

If we don't terminate at t:



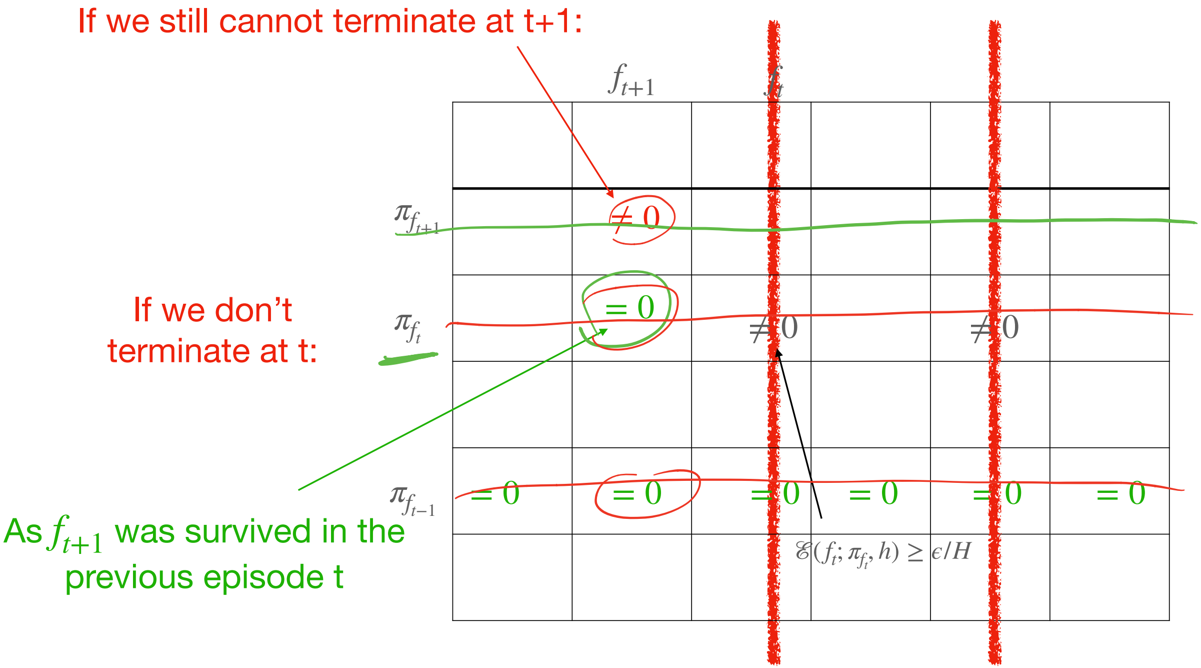
Progress on Value Function Elimination

$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$



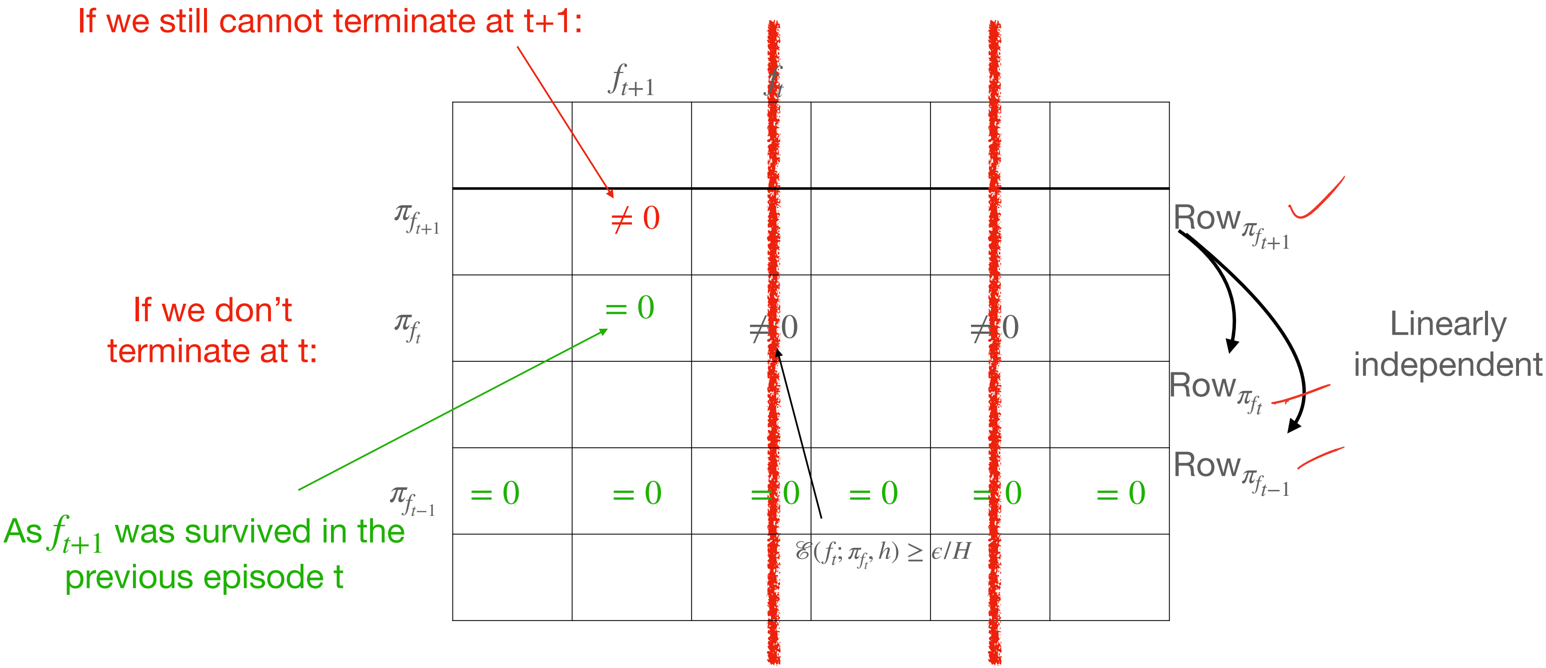
Progress on Value Function Elimination

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$



Progress on Value Function Elimination

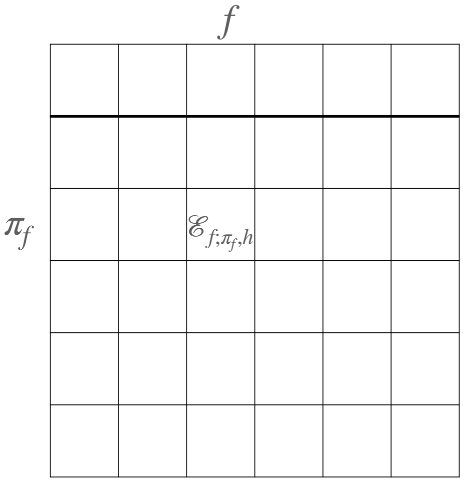
$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$



Progress on Value Function Elimination

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

$$\forall h : \mathscr{E}_h \in \mathbb{R}^{|\mathscr{F}| \times |\mathscr{F}|}$$

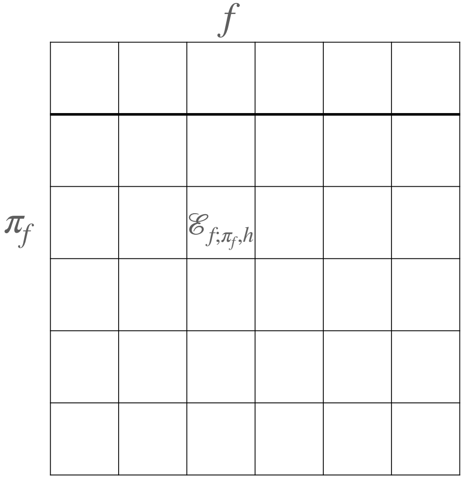


Rank

Progress on Value Function Elimination

$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

$\forall h : \mathcal{E}_h \in \mathbb{R}^{|\mathcal{F}| \times |\mathcal{F}|}$

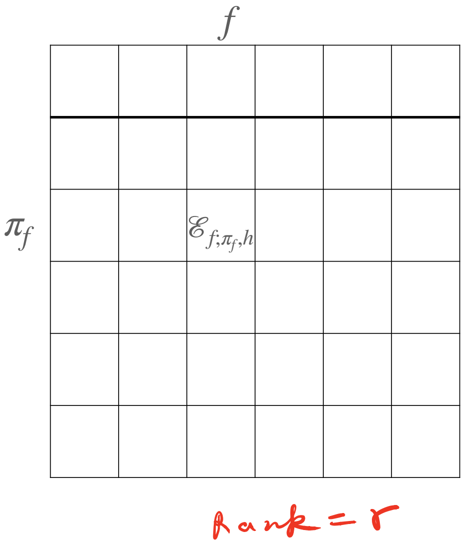


Every episode, we identify a row that is linearly independent of all previous rows we found!

Progress on Value Function Elimination

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

$$\forall h : \mathscr{E}_h \in \mathbb{R}^{|\mathscr{F}|\times|\mathscr{F}|}$$



Every episode, we identify a row that is linearly independent of all previous rows we found!

Then we must terminate in # of iterations at most (**Rank H**)

OLIVE Revisit (ignoring statistical error for simplicity)

Initialize $\mathcal{F}_0 = \mathcal{F}$

For $t = 0, \ldots$

$f_t = \arg\max_{f \in \mathcal{F}_t} \max_a f(x_0, a)$ ✓ optimism

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Version space update:

$$\mathcal{F}_{t+1} = \left\{ f \in \mathcal{F}_t : \mathcal{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

1. Upon termination we succeed ✓
2. If not terminate, we make non-trivial progress (f_t)
3. Total # of such non-trivial progress is bounded

Bellman Rank

row_t Linear independent wrt $row_1 \cdots row_{t-1}$

OLIVE Revisit (ignoring statistical error for simplicity)

Initialize $\mathscr{F}_0 = \mathscr{F}$

1. Requires samples (poly in $1/\epsilon$)

For $t = 0, \ldots$

$f_t = \arg\max_{f \in \mathscr{F}_t} \max_a f(x_0, a)$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon,$ return π_{f_t}

Version space update:

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

1. Upon termination we succeed

2. If not terminate, we make non-trivial progress

3. Total # of such non-trivial progress is bounded

OLIVE Revisit (ignoring statistical error for simplicity)

Initialize $\mathscr{F}_0 = \mathscr{F}$

1. Requires samples (poly in $1/\epsilon$)

For $t = 0, \ldots$

$$f_t = \arg\max_{f \in \mathscr{F}_t} \max_a f(x_0, a)$$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon,$ return π_{f_t}

Version space update:

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$$

✓ Union Bound

2. Requires samples (poly in $1/\epsilon$, $\ln(|\mathscr{F}|)$)
(needs to hold for all f)

1. Upon termination we succeed

2. If not terminate, we make non-trivial progress

3. Total # of such non-trivial progress is bounded

OLIVE Revisit (ignoring statistical error for simplicity)

Initialize $\mathscr{F}_0 = \mathscr{F}$

1. Requires samples (poly in $1/\epsilon$)

For $t = 0, \ldots$

$$f_t = \arg\max_{f \in \mathscr{F}_t} \max_a f(x_0, a)$$

If $\left| V^{\pi_{f_t}} - \max_a f_t(x_0, a) \right| \leq \epsilon$, return π_{f_t}

Version space update:

$$\mathscr{F}_{t+1} = \left\{ f \in \mathscr{F}_t : \mathscr{E}(f; \pi_{f_t}, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

2. Requires samples (poly in $1/\epsilon$, $\ln(|\mathscr{F}|)$)
(needs to hold for all f)

1. Upon termination we succeed

2. If not terminate, we make non-trivial progress

3. Total # of such non-trivial progress is bounded

Terminate at most Rank Iterations

$$\text{Poly}\left(H, \frac{1}{\epsilon}, \text{Rank}, \ln(|\mathscr{F}|)\right)$$

Low Bellman Rank Example

$$\mathcal{E}(f;\bar{f},h) = \mathbb{E}_{x_h,a_h \sim d_h^{\pi_{\bar{f}}}}\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathcal{A}} f(x_{h+1},a)\right]\right] = \langle \xi(f), \eta(\bar{f})\rangle$$

Low Bellman Rank Example

$$\mathcal{E}(f; \bar{f}, h) = \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \right] = \langle \xi(f), \eta(\bar{f}) \rangle$$

1. Tabular MDP: ✓

$\leftarrow |X| \times |A|$

Low Bellman Rank Example

$$\mathscr{E}(f;\bar{f},h) = \mathbb{E}_{x_h,a_h\sim d_h^{\pi_{\bar{f}}}}\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathscr{A}} f(x_{h+1},a)\right]\right] = \langle \xi(f), \eta(\bar{f})\rangle$$

1. Tabular MDP:

$$\mathbb{E}_{x_h,a_h\sim d_h^{\pi_{\bar{f}}}}\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathscr{A}} f(x_{h+1},a)\right]\right] = \sum_{x,a\in X_h\times A_h} d_h^{\pi_{\bar{f}}}(s,a)\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathscr{A}} f(x_{h+1},a)\right]\right]$$

Low Bellman Rank Example

$$\mathscr{E}(f;\bar{f},h) = \mathbb{E}_{x_h,a_h\sim d_h^{\pi_{\bar{f}}}}\left[\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathscr{A}} f(x_{h+1},a)\right]\right]\right] = \langle \xi(f), \eta(\bar{f})\rangle$$

1. Tabular MDP:

$$\mathbb{E}_{x_h,a_h\sim d_h^{\pi_{\bar{f}}}}\left[\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathscr{A}} f(x_{h+1},a)\right]\right]\right] = \sum_{x,a\in X_h\times A_h} d_h^{\pi_{\bar{f}}}(s,a)\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathscr{A}} f(x_{h+1},a)\right]\right]$$

$$= \langle d_h^{\pi_{\bar{f}}}, f(\cdot,\cdot) - r(\cdot,\cdot) - \mathbb{E}_{x_{h+1}\sim P(\cdot|\cdot,\cdot)} \max_a f(x_{h+1},a)\rangle$$

A

$$d_h^{\pi_f} \in R^{|X|\times|A|}$$

$$\text{Bellman Rank} \leq |X|\times|A|$$

Low Bellman Rank Example

$$\mathcal{E}(f;\bar{f},h) = \mathbb{E}_{x_h,a_h\sim d_h^{\pi_{\bar{f}}}}\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathcal{A}} f(x_{h+1},a)\right]\right] = \langle \xi(f), \eta(\bar{f})\rangle$$

1. Tabular MDP:

$$\mathbb{E}_{x_h,a_h\sim d_h^{\pi_{\bar{f}}}}\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathcal{A}} f(x_{h+1},a)\right]\right] = \sum_{x,a\in X_h\times A_h} d_h^{\pi_{\bar{f}}}(s,a)\left[f(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1}\sim P(\cdot|x_h,a_h)}\left[\max_{a\in\mathcal{A}} f(x_{h+1},a)\right]\right]$$

$$= \langle d_h^{\pi_{\bar{f}}}, f(\cdot,\cdot) - r(\cdot,\cdot) - \mathbb{E}_{x_{h+1}\sim P(\cdot|\cdot,\cdot)} \max_a f(x_{h+1},a)\rangle$$

Rank at most $|X||A|$

2. Linear MDPs:

$$P(\cdot \mid x, a) = \mu^\star \phi(x, a),\ r(x, a) = \theta^\star \phi(x, a)$$

$$Q^\star(x, a) = (w^\star)^\top \phi(x, a)$$

$\phi \in \mathbb{R}^d$

2. Linear MDPs:

$$P(\cdot \mid x, a) = \mu^\star \phi(x, a),\ r(x, a) = \theta^\star \phi(x, a)$$

$$Q^\star(x, a) = (w^\star)^\top \phi(x, a)$$

$$\mathscr{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \leq W\}$$

2. Linear MDPs:

$$P(\cdot \mid x, a) = \mu^\star \phi(x, a),\ r(x, a) = \theta^\star \phi(x, a)$$

$$Q^\star(x, a) = (w^\star)^\top \phi(x, a)$$

$$\mathcal{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \leq W\}$$

$f(s,a) = w^\top \cdot \phi(s,a)$

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar f}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot \mid x_h, a_h)} \left[\max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \right]$$

2. Linear MDPs:

$$P(\cdot \mid x, a) = \mu^\star \phi(x, a),\ r(x, a) = \theta^\star \phi(x, a)$$

$$Q^\star(x, a) = (w^\star)^\top \phi(x, a)$$

$$\mathscr{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \leq W\}$$

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot \mid x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[w^\top \phi(x_h, a_h) - (\theta^\star)^\top \phi(x_h, a_h) - \phi(x_h, a_h)^\top (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right]$$

2. Linear MDPs:

$$P(\cdot \mid x, a) = \mu^\star \phi(x, a),\ r(x, a) = \theta^\star \phi(x, a)$$

$$Q^\star(x, a) = (w^\star)^\top \phi(x, a)$$

$$\mathcal{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \le W\}$$

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar f}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar f}}} \left[w^\top \phi(x_h, a_h) - (\theta^\star)^\top \phi(x_h, a_h) - \phi(x_h, a_h)^\top (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar f}}} \left[\left(w + \theta^\star + (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right)^\top \phi(x_h, a_h) \right]$$

2. Linear MDPs:

$$P(\cdot \mid x, a) = \mu^\star \phi(x, a),\ r(x, a) = \theta^\star \phi(x, a)$$

$$Q^\star(x, a) = (w^\star)^\top \phi(x, a)$$

$$\mathscr{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \le W\}$$

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot \mid x_h, a_h)} \left[\max_{a \in \mathscr{A}} f(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[w^\top \phi(x_h, a_h) - (\theta^\star)^\top \phi(x_h, a_h) - \phi(x_h, a_h)^\top (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[\left(w + \theta^\star + (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right)^\top \phi(x_h, a_h) \right]$$

$$= \left\langle \left(w + \theta^\star + (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right), \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} [\phi(x_h, a_h)] \right\rangle$$

2. Linear MDPs:

$$P(\cdot \mid x, a) = \mu^\star \phi(x, a),\ r(x, a) = \theta^\star \phi(x, a)$$

$$Q^\star(x, a) = (w^\star)^\top \phi(x, a)$$

$$\mathcal{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \le W\}$$

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[f(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot \mid x_h, a_h)} \left[\max_{a \in \mathcal{A}} f(x_{h+1}, a) \right] \right]$$

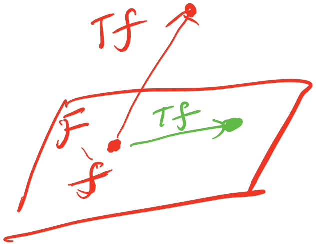
$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[w^\top \phi(x_h, a_h) - (\theta^\star)^\top \phi(x_h, a_h) - \phi(x_h, a_h)^\top (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[\left(w + \theta^\star + (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right)^\top \phi(x_h, a_h) \right]$$

$$= \left\langle \left(w + \theta^\star + (\mu^\star)^\top \left(\max_a f(\cdot, a) \right) \right), \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[\phi(x_h, a_h) \right] \right\rangle$$

Rank at most d

2. Linear Function with Bellman Completeness



$$\mathscr{F} = \{w^\top \cdot \phi(x,a) : w \in \mathbb{R}^d, \|w\|_2 \le W\}$$

$$r(x,a) = \theta^{\star}\phi(x,a),$$

and **for any linear function** $f(x,a) := w^\top \phi(x,a)$,

we have $\mathscr{T} f(x,a) = (w')^\top \phi(x,a)$

2. Linear Function with Bellman Completeness

$$\mathcal{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \leq W\}$$

$$r(x, a) = \theta^\star \phi(x, a),$$

and **for any linear function** $f(x, a) := w^\top \phi(x, a)$,

we have $\mathcal{T} f(x, a) = (w')^\top \phi(x, a)$

(Go and verify that linear MDP is a special instance of this setting)

2. Linear Function with Bellman Completeness

$$\mathscr{F} = \{w^\top \cdot \phi(x,a) : w \in \mathbb{R}^d, \|w\|_2 \le W\}$$

$$r(x,a) = \theta^\star \phi(x,a),$$

and **for any linear function** $f(x,a) := w^\top \phi(x,a)$,

we have $\mathcal{T} f(x,a) = (w')^\top \phi(x,a)$

(Go and verify that linear MDP is a special instance of this setting)

$f = w^\top \cdot \phi$ $\qquad$ $T \cdot f := T(w \cdot \phi) = w' \cdot \phi(s,a)$

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\tilde f}}}\left[w^\top \phi(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} w^\top \phi(x_{h+1}, a) \right] \right]$$

$w' \cdot \phi(x_h, a_h) \leftarrow$ By Bellman-completion

2. Linear Function with Bellman Completeness

$$\mathscr{F} = \{w^\top \cdot \phi(x, a) : w \in \mathbb{R}^d, \|w\|_2 \leq W\}$$

$$r(x, a) = \theta^\star \phi(x, a),$$

and **for any linear function** $f(x, a) := w^\top \phi(x, a)$,

we have $\mathcal{T} f(x, a) = (w')^\top \phi(x, a)$

(Go and verify that linear MDP is a special instance of this setting)

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[w^\top \phi(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} w^\top \phi(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\bar{f}}}} \left[w^\top \phi(x_h, a_h) - (w')^\top \phi(x_h, a_h) \right]$$

2. Linear Function with Bellman Completeness

$$\mathscr{F} = \{w^\top \cdot \phi(x,a) : w \in \mathbb{R}^d, \|w\|_2 \leq W\}$$

$$r(x,a) = \theta^\star \phi(x,a),$$

and **for any linear function** $f(x,a) := w^\top \phi(x,a)$,

we have $\mathscr{T} f(x,a) = (w')^\top \phi(x,a)$

(Go and verify that linear MDP is a special instance of this setting)

$$\mathbb{E}_{x_h,a_h \sim d_h^{\pi_{\bar f}}} \left[w^\top \phi(x_h,a_h) - r(x_h,a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot|x_h,a_h)} \left[\max_{a \in \mathscr{A}} w^\top \phi(x_{h+1},a) \right] \right]$$

$$= \mathbb{E}_{x_h,a_h \sim d_h^{\pi_{\bar f}}} \left[w^\top \phi(x_h,a_h) - (w')^\top \phi(x_h,a_h) \right]$$

$$= \left\langle \mathbb{E}_{x_h,a_h \sim d_h^{\pi_{\bar f}}} \left[\phi(x_h,a_h) \right], w - w' \right\rangle$$

2. Linear Function with Bellman Completeness

olive

poly$(H, \frac{1}{\epsilon}, d, \;)$

$$\mathscr{F} = \{w^\top \cdot \phi(x,a) : w \in \mathbb{R}^d, \|w\|_2 \le W\}$$

$$r(x,a) = \theta^\star \phi(x,a),$$

and **for any linear function** $f(x,a) := w^\top \phi(x,a)$,

we have $\mathcal{T} f(x,a) = (w')^\top \phi(x,a)$

(Go and verify that linear MDP is a special instance of this setting)

$$\mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\tilde f}}} \left[w^\top \phi(x_h, a_h) - r(x_h, a_h) - \mathbb{E}_{x_{h+1} \sim P(\cdot | x_h, a_h)} \left[\max_{a \in \mathscr{A}} w^\top \phi(x_{h+1}, a) \right] \right]$$

$$= \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\tilde f}}} \left[w^\top \phi(x_h, a_h) - (w')^\top \phi(x_h, a_h) \right]$$

Rank at most d

$$= \left\langle \mathbb{E}_{x_h, a_h \sim d_h^{\pi_{\tilde f}}} \left[\phi(x_h, a_h) \right], w - w' \right\rangle$$

Bellman Rank

↑

Tabular / Linear / Linear Complete

SA d d

Olive

- Computationally Inefficient

open problem.

① statical & computation efficient

② Linear-complete

Olive

③ Dependency Rank suboptimal

Other Examples it captures:

Bellman Error

Low rank MDP (requires a small modification in definition) ↙

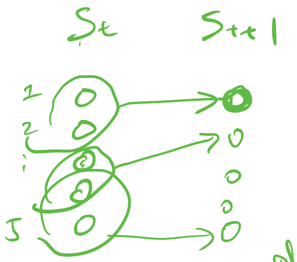
Reactive Predictive State Representation (PSRs) ↙

Reactive POMDP ↙

Examples it does not capture:

Factored MDPs ←

Linear $Q^\star$ (?)



$$P(s'|s) = \prod_{i=1}^{d} P(s'_i | s_{pa_i})$$

Model-based RL: Function Approximation in Model

Let's set up function class in Model-based RL Setting

We consider a model class $\mathscr{P}$

$$\mathscr{P} \subset X \times A \mapsto \Delta(X)$$

Let's set up function class in Model-based RL Setting

We consider a model class $\mathscr{P}$

$$\mathscr{P} \subset X \times A \mapsto \Delta(X)$$

1. **Realizability** assumption:

$$P^\star \in \mathscr{P}$$

Let's set up function class in Model-based RL Setting

We consider a model class $\mathscr{P}$

$$\mathscr{P} \subset X \times A \mapsto \Delta(X)$$

1. **Realizability** assumption:

$$P^\star \in \mathscr{P}$$

2. We also assume reward function r is known

Let's set up function class in Model-based RL Setting

We consider a model class $\mathscr{P}$

$$\mathscr{P} \subset X \times A \mapsto \Delta(X)$$

1. **Realizability** assumption:

$$P^\star \in \mathscr{P}$$

2. We also assume reward function r is known

3. Define **Optimal-Planner (OP)**: $\mathrm{OP}(P, r) => \left(\pi_P^\star, V_P^\star, Q_P^\star\right)$

Let's set up function class in Model-based RL Setting

We consider a model class $\mathscr{P}$

$$\mathscr{P} \subset X \times A \mapsto \Delta(X)$$

1. **Realizability** assumption:

$$P^\star \in \mathscr{P}$$

2. We also assume reward function r is known

Computational oracle, no real-world samples needed

3. Define **Optimal-Planner (OP)**: OP$(P, r) => \left(\pi_P^\star, V_P^\star, Q_P^\star\right)$

Let's set up function class in Model-based RL Setting

We consider a model class $\mathscr{P}$

$$\mathscr{P} \subset X \times A \mapsto \Delta(X)$$

1. **Realizability** assumption:

$$P^\star \in \mathscr{P}$$

2. We also assume reward function r is known

Computational oracle, no real-world samples needed

3. Define **Optimal-Planner (OP)**: $\mathrm{OP}(P, r) => \left(\pi_P^\star, V_P^\star, Q_P^\star\right)$

4. Induced Policy, Value function, and Q function class:

$$\Pi := \{\pi_P^\star : P \in \mathscr{P}\}, \mathscr{V} = \{V_P^\star : P \in \mathscr{P}\}, \mathscr{Q} = \{Q_P^\star : P \in \mathscr{P}\}$$

Let's set up function class in Model-based RL Setting

We consider a model class $\mathscr{P}$

$$\mathscr{P} \subset X \times A \mapsto \Delta(X)$$

1. **Realizability** assumption:

$$P^\star \in \mathscr{P}$$

2. We also assume reward function r is known

Computational oracle, no real-world samples needed

3. Define **Optimal-Planner (OP)**: $\mathrm{OP}(P, r) => \left(\pi_P^\star, V_P^\star, Q_P^\star\right)$

4. Induced Policy, Value function, and Q function class:

$$\Pi := \{\pi_P^\star : P \in \mathscr{P}\}, \mathscr{V} = \{V_P^\star : P \in \mathscr{P}\}, \mathscr{Q} = \{Q_P^\star : P \in \mathscr{P}\}$$

Indeed you can run OLIVE w/ $\mathscr{Q}$!

How to check if a model approximator is good?

Witness model error: $\mathcal{E}(P; \pi, h) = \max_{f \in \mathcal{F}} \mathbb{E}_{x_h, a_h \sim d_h^\pi} \left[\mathbb{E}_{x_{h+1} \sim P_{x_h, a_h}} f(x_h, a_h, x_{h+1}) - \mathbb{E}_{x_{h+1} \sim P^\star_{x_h, a_h}} f(x_h, a_h, x_{h+1}) \right]$

Witness function (or aka discriminators): $\mathcal{F} \subset X \times A \times X \mapsto \mathbb{R}$

How to check if a model approximator is good?

$\mathscr{E}(P^\star; \pi, h) = 0, \forall \pi, h$

Witness model error: $\mathscr{E}(P; \pi, h) = \max_{f \in \mathscr{F}} \mathbb{E}_{x_h, a_h \sim d_h^\pi} \left[\mathbb{E}_{x_{h+1} \sim P_{x_h, a_h}} f(x_h, a_h, x_{h+1}) - \mathbb{E}_{x_{h+1} \sim P^\star_{x_h, a_h}} f(x_h, a_h, x_{h+1}) \right]$

Witness function (or aka discriminators): $\mathscr{F} \subset X \times A \times X \mapsto \mathbb{R}$

How to check if a model approximator is good?

$\mathscr{E}(P^\star; \pi, h) = 0, \forall \pi, h$

Witness model error: $\mathscr{E}(P; \pi, h) = \max_{f \in \mathscr{F}} \mathbb{E}_{x_h, a_h \sim d_h^\pi} \left[\mathbb{E}_{x_{h+1} \sim P_{x_h, a_h}} f(x_h, a_h, x_{h+1}) - \mathbb{E}_{x_{h+1} \sim P^\star_{x_h, a_h}} f(x_h, a_h, x_{h+1}) \right]$

Witness function (or aka discriminators): $\mathscr{F} \subset X \times A \times X \mapsto \mathbb{R}$

Integral Probability Metric (IPM): given two distributions $P_1 \in \Delta(X), P_2 \in \Delta(X)$:

$$\text{IPM}_{\mathscr{F}} = \max_{f \in \mathscr{F}} \mathbb{E}_{x \sim P_1} f(x) - \mathbb{E}_{x \sim P_2} f(x)$$

How to check if a model approximator is good?

$\mathscr{E}(P^\star; \pi, h) = 0, \forall \pi, h$

Witness model error: $\mathscr{E}(P; \pi, h) = \max_{f \in \mathscr{F}} \mathbb{E}_{x_h, a_h \sim d_h^\pi} \left[\mathbb{E}_{x_{h+1} \sim P_{x_h, a_h}} f(x_h, a_h, x_{h+1}) - \mathbb{E}_{x_{h+1} \sim P^\star_{x_h, a_h}} f(x_h, a_h, x_{h+1}) \right]$

Witness function (or aka discriminators): $\mathscr{F} \subset X \times A \times X \mapsto \mathbb{R}$

Integral Probability Metric (IPM): given two distributions $P_1 \in \Delta(X), P_2 \in \Delta(X)$:

$$\text{IPM}_{\mathscr{F}} = \max_{f \in \mathscr{F}} \mathbb{E}_{x \sim P_1} f(x) - \mathbb{E}_{x \sim P_2} f(x)$$

$\mathscr{F} = \{f : \|f\|_\infty \leq 1\}$ => Total Variation Distance;

$\mathscr{F} = \{f : \text{f is 1-Lipschitz}\}$ => Wasserstein Distance;

How to check if a model approximator is good?

$\mathscr{E}(P^\star; \pi, h) = 0, \forall \pi, h$

Witness model error: $\mathscr{E}(P; \pi, h) = \max_{f \in \mathscr{F}} \mathbb{E}_{x_h, a_h \sim d_h^\pi} \left[\mathbb{E}_{x_{h+1} \sim P_{x_h, a_h}} f(x_h, a_h, x_{h+1}) - \mathbb{E}_{x_{h+1} \sim P^\star_{x_h, a_h}} f(x_h, a_h, x_{h+1}) \right]$

Witness function (or aka discriminators): $\mathscr{F} \subset X \times A \times X \mapsto \mathbb{R}$

Integral Probability Metric (IPM): given two distributions $P_1 \in \Delta(X), P_2 \in \Delta(X)$:

$$\text{IPM}_{\mathscr{F}} = \max_{f \in \mathscr{F}} \mathbb{E}_{x \sim P_1} f(x) - \mathbb{E}_{x \sim P_2} f(x)$$

$\mathscr{F} = \{f : \|f\|_\infty \leq 1\}$ => Total Variation Distance;

$\mathscr{F} = \{f : \text{f is 1-Lipschitz}\}$ => Wasserstein Distance;

Assumption (for analysis simplicity): $\mathscr{V} \subseteq \mathscr{F}$, where recall $\mathscr{V} = \{V_P^\star : P \in \mathscr{P}\}$

Optimism Led Iterative Model Elimination (OLIM)

Initialize $\mathscr{P}_0 = \mathscr{P}$

For $t = 0, \ldots$

$$P_t = \arg\max_{P \in \mathscr{P}_t} V^\star_P(x_0)$$

$$\pi_t := \pi^\star_{P_t}$$

If $\left| V^{\pi_t} - V^\star_{P_t} \right| \leq \epsilon$, return π_t

Version space update:

$$\mathscr{P}_{t+1} = \left\{ P \in \mathscr{P}_t : \mathscr{E}(P; \pi_t, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

Optimism Led Iterative Model Elimination (OLIM)

Initialize $\mathscr{P}_0 = \mathscr{P}$

For $t = 0, \ldots$

$$P_t = \arg\max_{P \in \mathscr{P}_t} V_P^\star(x_0)$$

$$\pi_t := \pi_{P_t}^\star$$

If $\left| V^{\pi_t} - V_{P_t}^\star \right| \leq \epsilon,$ return π_t

Version space update:

$$\mathscr{P}_{t+1} = \left\{ P \in \mathscr{P}_t : \mathscr{E}(P; \pi_t, h) = 0, \forall h \in \{0,1,\ldots,H-1\} \right\}$$

1. Terminate means we succeed

Optimism Led Iterative Model Elimination (OLIM)

Initialize $\mathscr{P}_0 = \mathscr{P}$

For $t = 0, \ldots$

$P_t = \arg\max_{P \in \mathscr{P}_t} V_P^\star(x_0)$

$\pi_t := \pi_{P_t}^\star$

If $\left| V^{\pi_t} - V_{P_t}^\star \right| \leq \epsilon,$ return π_t

Version space update:

$\mathscr{P}_{t+1} = \left\{ P \in \mathscr{P}_t : \mathscr{E}(P; \pi_t, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$

1. Terminate means we succeed

2. If not terminate, we make progress

Optimism Led Iterative Model Elimination (OLIM)

Initialize $\mathscr{P}_0 = \mathscr{P}$

For $t = 0, \ldots$

$P_t = \arg\max_{P \in \mathscr{P}_t} V_P^\star(x_0)$

$\pi_t := \pi_{P_t}^\star$

If $\left| V^{\pi_t} - V_{P_t}^\star \right| \leq \epsilon$, return π_t

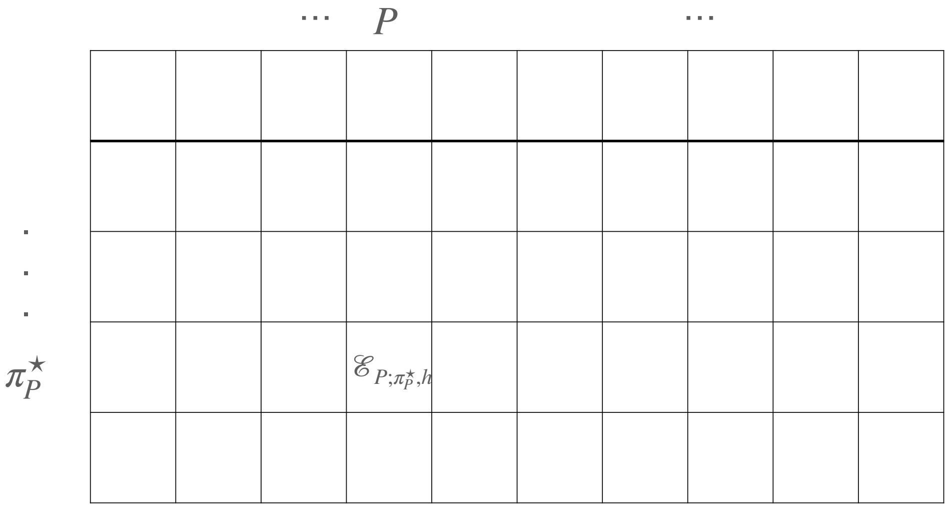
Version space update:

$\mathscr{P}_{t+1} = \left\{ P \in \mathscr{P}_t : \mathscr{E}(P; \pi_t, h) = 0, \forall h \in \{0, 1, \ldots, H-1\} \right\}$

1. Terminate means we succeed
2. If not terminate, we make progress
3. Total progress is upper bounded

Progress on Model Elimination

MDP + $\mathscr{P}$ + $\mathscr{F}$ determines the following matrix: $\mathscr{E} \in \mathbb{R}^{|\mathscr{P}| \times |\mathscr{P}|}$



Rank of this Matrix = Witness Rank

Sample complexity of Optimism Led Iterative Model Elimination:

Under some assumption of discriminators $\mathscr{F}$ (i.e., $\mathscr{V} \in \mathscr{F}$),

Witness rank <= Bellman Rank ($\mathscr{Q}$)

Sample complexity of Optimism Led Iterative Model Elimination:

$$\text{Poly}\left(H, \frac{1}{\epsilon}, \text{Witness-Rank}, \ln(|\mathscr{P}||\mathscr{F}|)\right)$$

Under some assumption of discriminators $\mathscr{F}$ (i.e., $\mathscr{V} \in \mathscr{F}$),

Witness rank <= Bellman Rank ($\mathscr{Q}$)

Comparison of Witness Rank and Bellman Rank

(More Broadly, Model-based Versus Model-free)

Comparison of Witness Rank and Bellman Rank
(More Broadly, Model-based Versus Model-free)

Let us compare under a specific setup:

We start with model class $\mathscr{P}$, we *convert models to Q functions* $\mathbb{Q} = \{Q_P^\star : P \in \mathscr{P}\}$

Comparison of Witness Rank and Bellman Rank
(More Broadly, Model-based Versus Model-free)

Let us compare under a specific setup:

We start with model class $\mathscr{P}$, we *convert models to Q functions* $\mathscr{Q} = \{Q_P^\star : P \in \mathscr{P}\}$

$\mathscr{P}$ => Witness Rank and OLIME

$\mathscr{Q}$ => Bellman Rank and OLIVE

Comparison of Witness Rank and Bellman Rank
(More Broadly, Model-based Versus Model-free)

Let us compare under a specific setup:

We start with model class $\mathscr{P}$, we *convert models to Q functions* $\mathscr{Q} = \{Q_P^\star : P \in \mathscr{P}\}$

$\mathscr{P}$ => Witness Rank and OLIME

$\mathscr{Q}$ => Bellman Rank and OLIVE

Theorem[Exponential Separation]:

$\exists$ MDPs (Factored MDP!) and realizable model class $\mathscr{P}$, s.t:

Witness-Rank($\mathscr{P}$) is *exponentially* smaller than Bellman-rank($\mathscr{Q}$) in horizon H