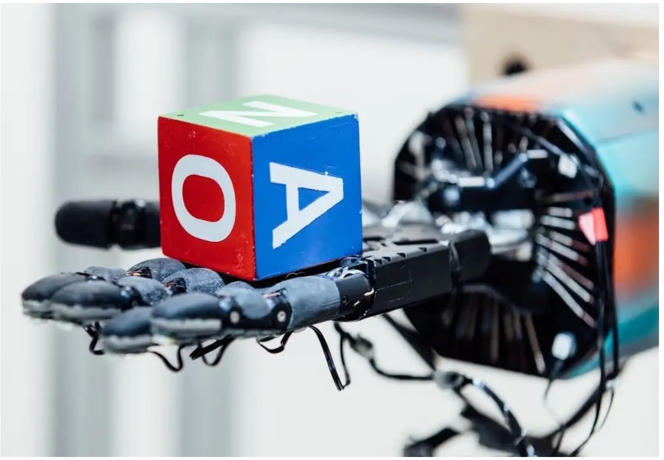
Computationally Efficient RL under Linear Bellman Completeness for Deterministic Dynamics

Runzhe Wu

Cornell University

Joint work with **Ayush Sekhari**, **Akshay Krishnamurthy**, and **Wen Sun**

RL + Rich Function Approximation



(OpenAI, 2018)

(Baker et al., 2022)

(OpenAI, 2023)

Can we design provably efficient RL algorithm under

Rich Function Approximation **?**

Can we design provably efficient RL algorithm under

Linear Function Approximation **?**

"Linear Bellman Completeness"

Outline

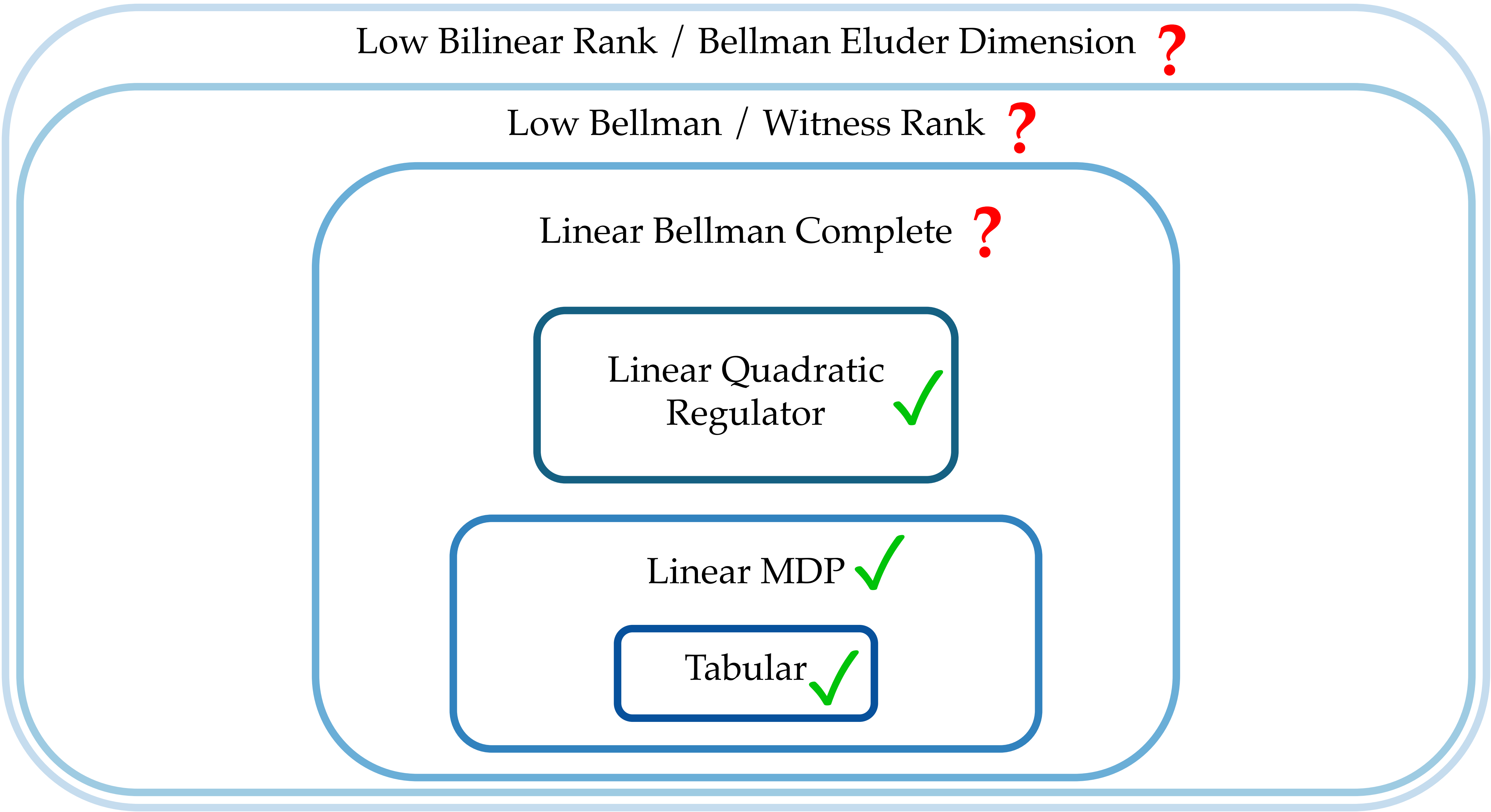
Part I Background

Part II Algorithm : the Trick of ***Span*** vs ***Null Space***

Part III The Norm Issue

Outline

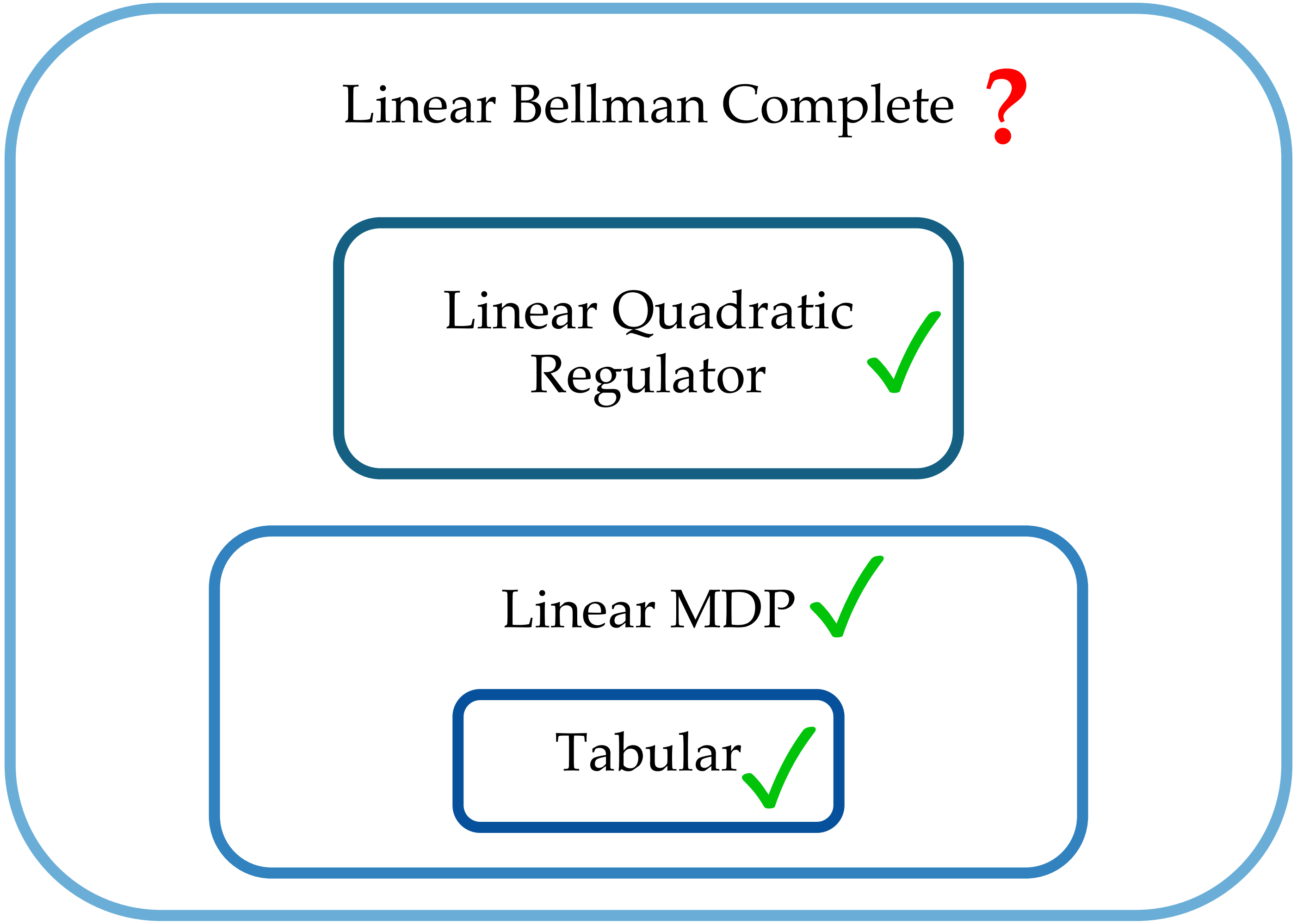
✓ : Computational and sample efficiency ? : Only sample efficiency Credit: Akshay Krishnamurthy



✓ : Computational and sample efficiency   ? : Only sample efficiency   Credit: Akshay Krishnamurthy

Open problem: does computationally efficient algorithm exist under linear BC?

Answer: yes, when the transition is deterministic!



We allow…

Adversarial Initial States ✓

Random Rewards ✓

Large Action Spaces ✓

(Assume known reward for simplicity)

Episodic Finite-Horizon MDP

$$s_1 \xrightarrow[r_1]{a_1} s_2 \xrightarrow[r_2]{a_2} s_3 \quad \cdots \quad s_{H-1} \xrightarrow[r_{H-1}]{a_{H-1}} s_H \xrightarrow[r_H]{a_H} \Big|$$

$$Q_h^\pi(s, a) = \mathbb{E}_\pi \left[\sum_{i=h}^{H} r_i \,\middle|\, s_h = s, a_h = a \right]$$

$$V_h^\pi(s) = Q(s, \pi(s))$$

$$V_h^\star(s) = \max_\pi V_h^\pi(s)$$

Regret Minimization

$$\mathrm{Reg}_T = \mathbb{E}\left[\sum_{t=1}^{T} \Big(V_1^\star(s_1) - V_1^{\pi_t}(s_1) \Big)\right]$$

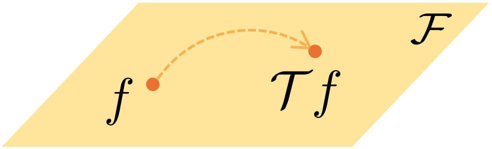
Linear Bellman Completeness

An MDP is **Bellman complete** w.r.t. a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ if

$$\forall f \in \mathcal{F} : \boxed{(s,a) \mapsto r(s,a) + \mathop{\mathbb{E}}_{s' \sim P(s,a)} \max_{a'} f(s',a')} \in \mathcal{F}$$

$=: \mathcal{T}f$ where $\mathcal{T}$ is the **Bellman operator**

In other words, $\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$



It is **Linear Bellman Complete** if $\mathcal{F}$ is linear

$$\mathcal{F} = \left\{ (s,a) \mapsto \langle \phi(s,a),\, \theta \rangle : \theta \in \mathbb{R}^d \right\}$$

Known

Outline

RLSVI

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a),\, \theta \rangle - r - V_{h+1}(s') \Big)^2 + \lambda \|\theta\|_2^2$$

$$\xi_h \sim \mathcal{N}(0, \sigma^2 \Sigma_h^{-1}) \text{ where } \Sigma_h = \sum_{(s,a) \in \mathcal{D}_h} \phi(s,a)\phi(s,a)^\top + \lambda I$$

$$Q_h(\cdot,\cdot) \leftarrow \min \Big\{ \langle \theta_h + \xi_h,\, \phi(\cdot,\cdot) \rangle,\, H \Big\}, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

Bayes optimal

Key Idea : ξ_h can cancel out estimation error $(\theta_h - \theta_h^\star)$ to achieve optimism (w/ constant prob)

*The clipping above is modified from the original version to facilitate presentation.

Russo, Daniel. "Worst-case regret bounds for exploration via randomized value functions." NeurIPS, 2019. Zanette, Andrea, et al. "Frequentist regret bounds for randomized least-squares value iteration." AISTATS, 2020.

RLSVI

Linear Bellman Complete

$$\mathcal{F} = \left\{ (s,a) \mapsto \langle \phi(s,a), \theta \rangle : \theta \in \mathbb{R}^d \right\}$$

$$\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$$

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

Bayes optimal must be linear

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2 + \lambda \|\theta\|_2^2$$

$$\xi_h \sim \mathcal{N}(0, \sigma^2 \Sigma_h^{-1}) \text{ where } \Sigma_h = \sum_{(s,a) \in \mathcal{D}_h} \phi(s,a)\phi(s,a)^\top + \lambda I$$

Not linear ✕

$$Q_h(\cdot,\cdot) \leftarrow \min \left\{ \langle \theta_h + \xi_h, \phi(\cdot,\cdot) \rangle, H \right\}, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Linear

Collect data w/ π_t

Apply to Linear BC ? ✕

What if we don't clip?

Non-linear Bayes optimal ⇒ linear regression fails

*The clipping above is modified from the original version to facilitate presentation.

Russo, Daniel. "Worst-case regret bounds for exploration via randomized value functions." NeurIPS, 2019. Zanette, Andrea, et al. "Frequentist regret bounds for randomized least-squares value iteration." AISTATS, 2020.

Observation: $\|\xi_h\| \approx \|\theta_h - \theta_h^\star\| =: \|\theta_h\| \cdot \epsilon$

Bayes optimal

RLSVI (w/o clipping)

Linear Bellman Complete

$$\mathcal{F} = \left\{(s,a) \mapsto \langle \phi(s,a), \theta \rangle : \theta \in \mathbb{R}^d\right\}$$

$$\mathcal{T}\mathcal{F} \subseteq \mathcal{F}$$

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_\theta \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a),\, \theta \rangle - r - V_{h+1}(s')\Big)^2 + \lambda\|\theta\|_2^2$$

$$\xi_h \sim \mathcal{N}(0, \sigma^2 \Sigma_h^{-1}) \text{ where } \Sigma_h = \sum_{(s,a) \in \mathcal{D}_h} \phi(s,a)\phi(s,a)^\top + \lambda I$$

$$Q_h(\cdot,\cdot) \leftarrow \min\left\{\langle \theta_h + \xi_h,\, \phi(\cdot,\cdot)\rangle,\, H\right\}, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

$\forall h$ Assume $\|\theta_h\| = L$

$\|\theta_h + \xi_h\| \approx (1+\epsilon)L$

$|Q_h| \approx (1+\epsilon)L$

$\|\theta_{h-1}\| \approx (1+\epsilon)L$

$\|\theta_{h-1} + \xi_{h-1}\| \approx (1+\epsilon)^2 L$

$|Q_{h-1}| \approx (1+\epsilon)^2 L$

$\|\theta_{h-2}\| \approx (1+\epsilon)^2 L$

$\|\theta_{h-2} + \xi_{h-2}\| \approx (1+\epsilon)^3 L$

$|Q_{h-2}| \approx (1+\epsilon)^3 L$

......

Without clipping, $\|\theta_h\|$ grows **exponentially**

On exponentially large $\|\theta_h\|$

$\|\theta_h\|$ ↗

Happens in **null space** only

① **Regression**

$$\|\theta_h - \theta_h^\star\|_{\Sigma_t}^2 \lesssim \mathrm{Poly}(d, H)$$

$$\|\phi\|_{\Sigma_t^{-1}}^2 \lesssim \frac{\mathrm{Poly}(d)}{N}$$

② **Bellman residual**

$$|Q_h - \mathcal{T} Q_h| \lesssim \frac{\mathrm{Poly}(d, H)}{\sqrt{N}}$$

③ **Performance gap**

$$|V^{\hat{\pi}} - V^\star| \lesssim \frac{\mathrm{Poly}(d, H)}{\sqrt{N}}$$

Solution:

When transition is deterministic, add noise in <u>the null space of data</u> only.

Key Observation

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_\theta \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a),\, \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\xi_h \sim \mathcal{N}(0, \sigma^2 \Sigma_t^{-1}) \text{ where } \Sigma_t = \sum_{(s,a) \in \mathcal{D}_h} \phi(s,a)\phi(s,a)^\top + \lambda I$$

$$Q_h(\cdot,\cdot) \leftarrow \min \Big\{ \langle \theta_h + \xi_h,\, \phi(\cdot,\cdot) \rangle,\, H \Big\}, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

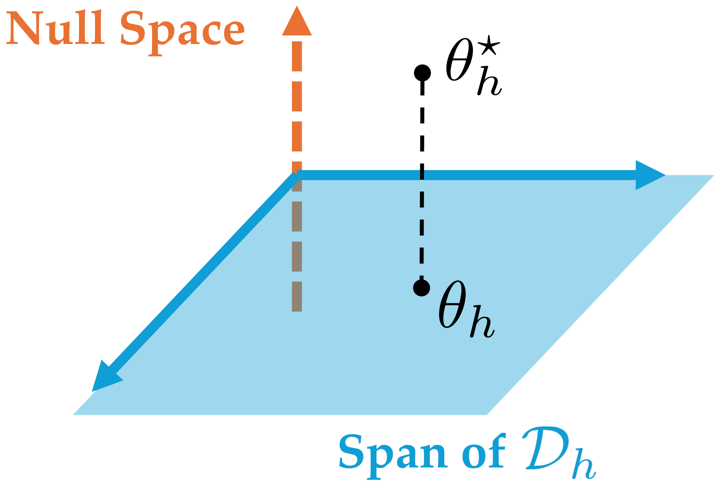
$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

Key Observation

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s')\in\mathcal{D}_h} \Big(\langle \phi(s,a),\, \theta\rangle - r - V_{h+1}(s')\Big)^2 = 0$$

Deterministic Transition $\Rightarrow$ $V_{h+1}(s')$ is deterministic $\Rightarrow$ θ_h zeros the empirical risk



P_h : orthogonal projection matrix onto the span

$\theta_h^\star - \theta_h \perp$ Span $\qquad P_h(\theta_h^\star - \theta_h) = 0$

$\theta_h^\star - \theta_h \in$ Null $\qquad (I - P_h)(\theta_h^\star - \theta_h) = \theta_h^\star - \theta_h$

Key Observation

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_\theta \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\boxed{\xi_h \sim \mathcal{N}(0, \sigma^2 \Sigma_t^{-1}) \text{ where } \Sigma_t = \sum_{(s,a) \in \mathcal{D}_h} \phi(s,a)\phi(s,a)^\top + \lambda I}$$

No need to explore in the **span**

$$Q_h(\cdot,\cdot) \leftarrow \big\langle \theta_h + \xi_h, \phi(\cdot,\cdot) \big\rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

P_h : orthogonal projection onto span

Key Observation

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\xi_h \sim \mathcal{N}(0, \sigma^2 \Sigma_t^{-1}) \text{ where } \Sigma_t = \sum_{(s,a) \in \mathcal{D}_h} \phi(s,a)\phi(s,a)^\top + \lambda I$$

$+$ $$\widetilde{\xi}_h \leftarrow (I - P_h)\xi_h$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

Only explore in the **null space**

Simplify to : $\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma^2 (I - P_h))$

P_h : orthogonal projection onto span

Key Observation

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

⚠️ ***Caveat**: Setting σ is **challenging** (covered in next section).*

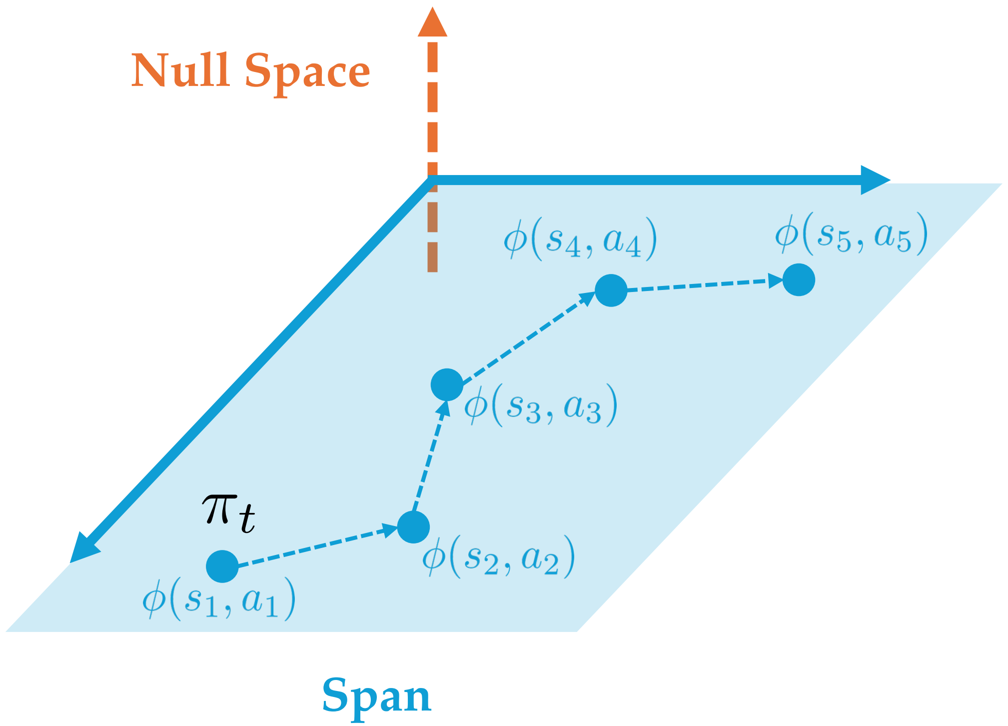
For now, assume we know how to set it.

Fix round t

Span Argument

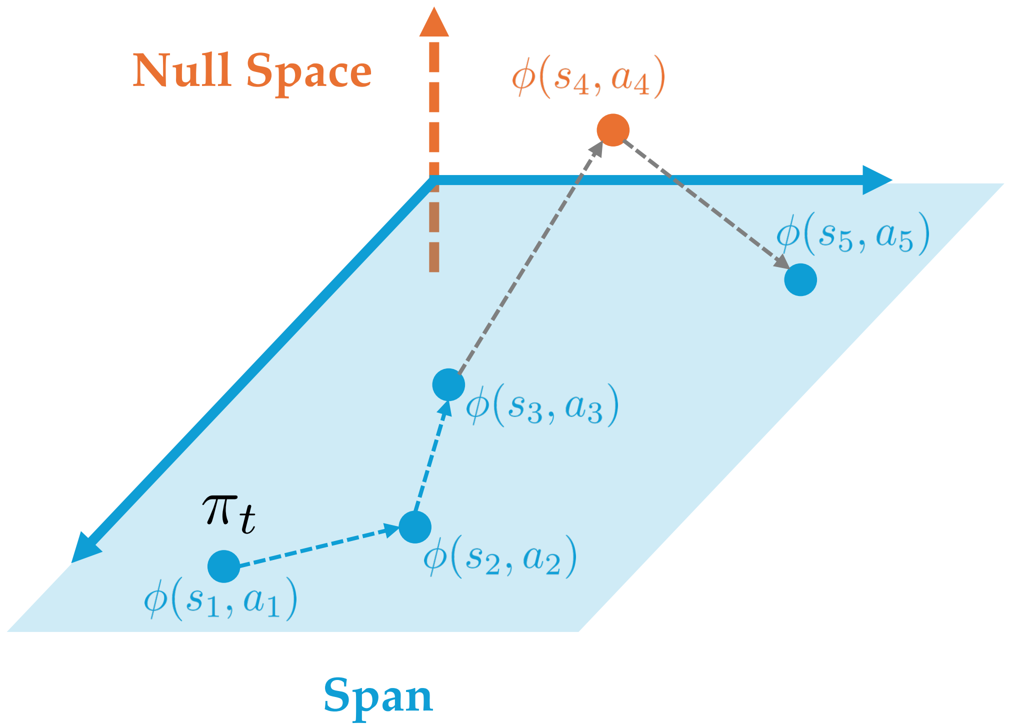
(1) All in Span

$\forall h : \phi(s_h, a_h) \in \text{Span}$



(2) Some in Null Space

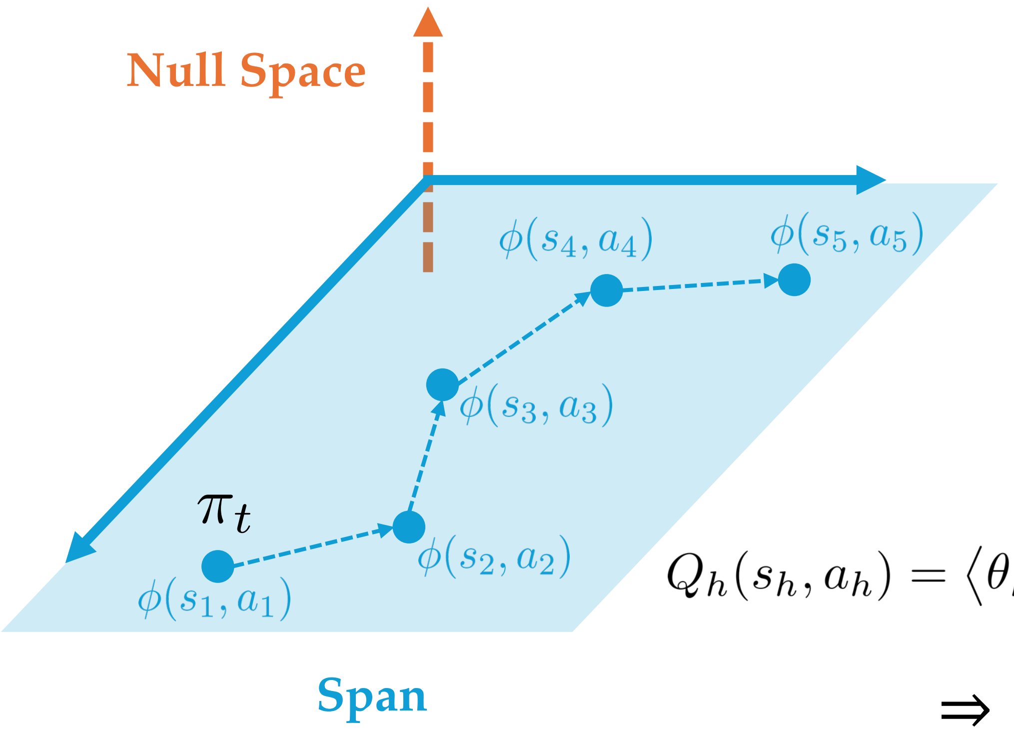
$\exists h : \phi(s_h, a_h) \notin \text{Span}$



Span Argument

(1) All in Span

$$\forall h : \phi(s_h, a_h) \in \text{Span}$$



Algorithm

For $t = 1, \ldots, T$

  For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a),\, \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h,\, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

  $\pi_t \leftarrow$ greedy policy w.r.t. Q_h

  Collect data w/ π_t

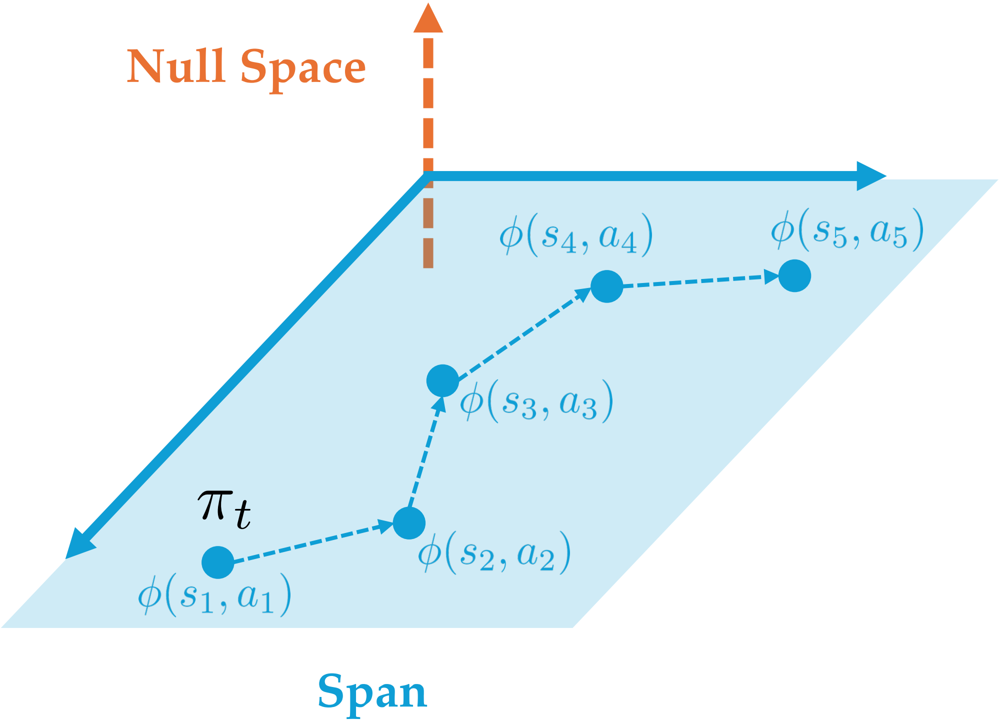
$$Q_h(s_h, a_h) = \langle \theta_h + \widetilde{\xi}_h,\, \phi(s_h, a_h) \rangle = \langle \theta_h,\, \phi(s_h, a_h) \rangle = \langle \theta_h^\star,\, \phi(s_h, a_h) \rangle = Q_h^{\pi_t}(s_h, a_h)$$

$$\Rightarrow \quad \forall h : Q_h(s_h, a_h) = Q_h^{\pi_t}(s_h, a_h), \quad V_h(s_h) = V_h^{\pi_t}(s_h)$$

Span Argument

(1) All in Span

$\forall h : \phi(s_h, a_h) \in \text{Span}$



$$V_1(s_1) = V_1^{\pi_t}(s_1)$$

$$V_1(s_1) \geq V_1^{\star}(s_1)$$

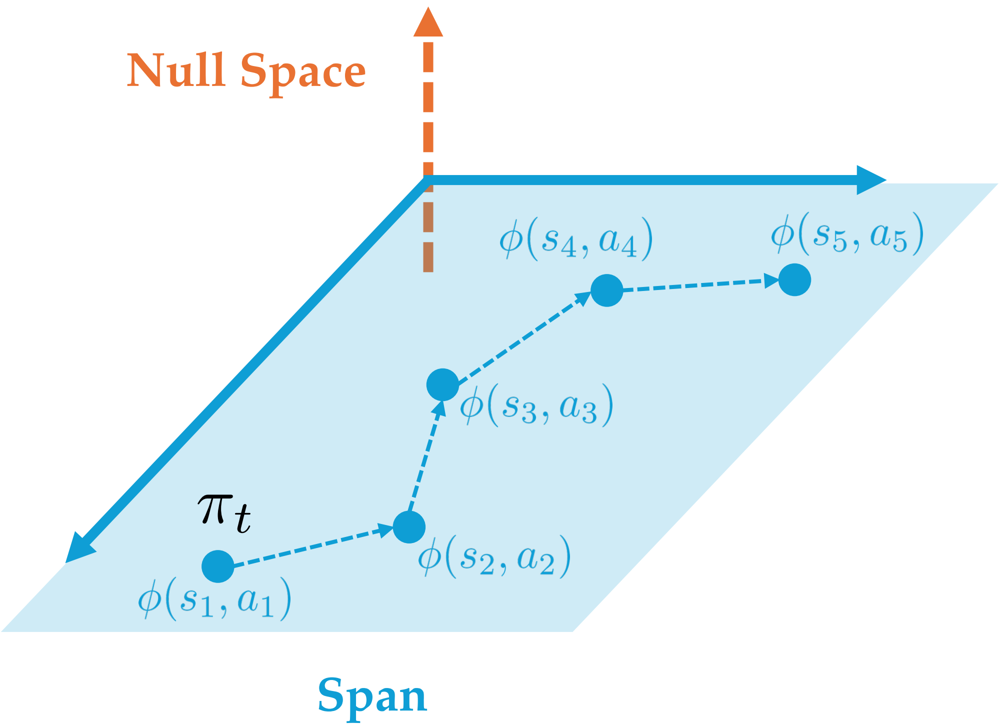
(optimism holds with constant probability)

Then, $$\begin{aligned} & V^{\star}(s_1) - V^{\pi_t}(s_1) \\ & \leq V_1(s_1) - V^{\pi_t}(s_1) \\ & = V_1(s_1) - V_1(s_1) \\ & = 0 \end{aligned}$$

Span Argument

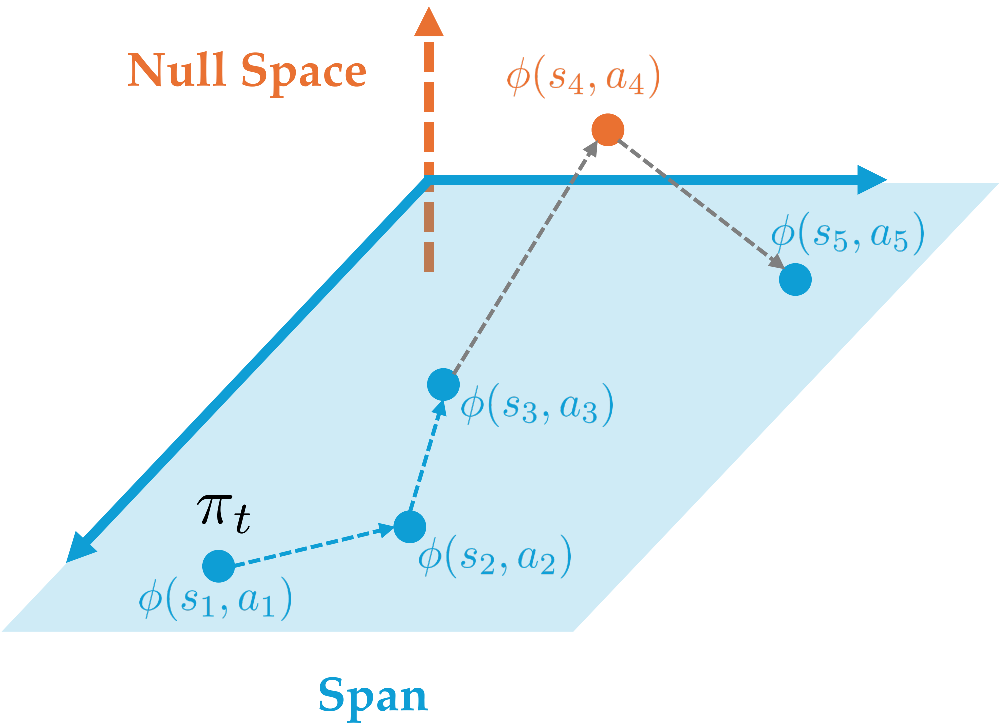
(1) All in Span

$\forall h : \phi(s_h, a_h) \in \text{Span}$



(2) Some in Null Space

$\exists h : \phi(s_h, a_h) \notin \text{Span}$



Span Argument

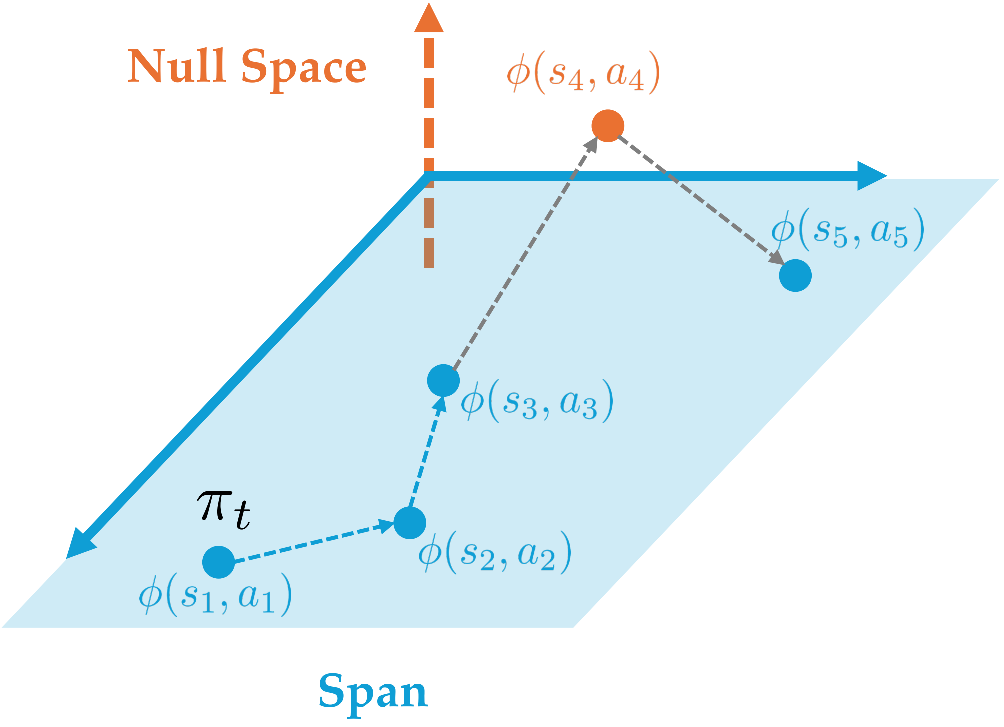
Next round, $\dim(\text{Span})$ increases by 1

But $\forall h : \dim(\text{Span}) \leq d$

Can happen at most dH times

(2) Some in Null Space

$\exists h : \phi(s_h, a_h) \notin \text{Span}$



Span Argument

(1) All in Span

$$\mathrm{Reg}_T = 0^{*}$$

(2) Some in Null Space

$$\mathrm{Reg}_T \leq dH \cdot H$$

Theorem. *If reward is known, we have*

$$\mathrm{Reg}_T \leq dH^2$$

Reward learning is standard

Theorem. *If reward is unknown, we have*

$$\mathrm{Reg}_T \leq \widetilde{O}\left(d^{5/2}H^{5/2} + d^2H^{3/2}\sqrt{T}\right)$$

* The derivation was intuitive; we actually incur some $\mathrm{Poly}(d, H)$ regret in case (1).

We are not done yet

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a),\, \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h,\, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

⚠️ ***Caveat**: Setting σ is **challenging (the norm issue)***

Outline

Part I Background

Part II Algorithm : the Trick of ***Span*** vs ***Null Space***

Part III The Norm Issue

The Norm Issue

Algorithm (w/ known reward)

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a),\, \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma_h^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h,\, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

We should have

$$\left| \widetilde{\xi}_h \right| \gtrsim \left| \theta_h - \theta_h^\star \right|$$

$$\sigma_h \gtrsim \left\| \theta_h - \theta_h^\star \right\|$$

How large can it be?

The Norm Issue

Algorithm (w/ known reward)

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma_h^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

Linear Bellman Complete

$$\mathcal{F} = \left\{ (s,a) \mapsto \langle \phi(s,a), \theta \rangle : \theta \in \mathbb{R}^d \right\} \qquad \mathcal{T}\mathcal{F} \subseteq \mathcal{F}$$

Assume

$$\|\theta_{h+1}\| \leq L \quad \|\theta^\star_{h+1}\| \leq L$$

$$\Downarrow$$

$$\|\widetilde{\xi}_{h+1}\| \approx \|\theta_{h+1} - \theta^\star_{h+1}\| \leq 2L$$

$$\Downarrow$$

$$\|\theta_{h+1} + \widetilde{\xi}_{h+1}\| \lesssim 3L$$

? $\quad \theta^\star_h = \mathcal{T}(\theta_{h+1} + \widetilde{\xi}_{h+1})$

$$\|\theta^\star_h\| \lesssim ?? \quad \|\theta_h - \theta^\star_h\| \lesssim ??$$

Don't know how to set σ_h

The Norm Issue

Algorithm (w/ known reward)

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma_h^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

Linear Bellman Complete

$$\mathcal{F} = \Big\{ (s,a) \mapsto \langle \phi(s,a), \theta \rangle : \theta \in \mathbb{R}^d \Big\} \qquad \mathcal{T}\mathcal{F} \subseteq \mathcal{F}$$

Linear BC allows *arbitrary norm explosion.*

⬇

We won't know how to set σ_h

An MDP is **Linear Bellman Complete** if

$$\forall f = \langle \phi, \theta \rangle, \quad \exists \tilde{f} = \langle \phi, \tilde{\theta} \rangle \quad \text{s.t.} \quad \tilde{f} = \mathcal{T} f$$

Prior Works

(1) Assume $\|\tilde{\theta}\|_2 \leq R$ (R : pre-fixed)

Looks not so natural…

(2) Assume $\|\tilde{\theta}\|_2 \leq \|\theta\|_2$

Not true in tabular MDPs

An MDP is **Linear Bellman Complete** if

$$\forall f = \langle \phi, \theta \rangle, \quad \exists \tilde{f} = \langle \phi, \tilde{\theta} \rangle \quad \text{s.t.} \quad \tilde{f} = \mathcal{T} f$$

Our Observation

$$\underbrace{\max_{s,a} \langle \phi(s,a), \tilde{\theta} \rangle}_{=: \|\tilde{\theta}\|_\infty^\phi} = \max_{s,a} \left(r(s,a) + \mathop{\mathbb{E}}_{s' \sim P(s,a)} \max_{a'} \langle \phi(s',a'), \theta \rangle \right) \le 1 + \underbrace{\max_{s,a} \langle \phi(s,a), \theta \rangle}_{=: \|\theta\|_\infty^\phi}$$

"ℓ_∞-functional-norm"

Linear BC controls *ℓ_∞-functional-norm !*

It is not an <u>assumption</u>; it is a <u>conclusion</u>.

A Second Visit to Norm

Algorithm (w/ known reward)

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \arg\min_{\theta} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma_h^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

Assume

$$\|\theta_{h+1}\|_\infty^\phi \le L \quad \|\theta^\star_{h+1}\|_\infty^\phi \le L$$

$$\|\widetilde{\xi}_{h+1}\|_\infty^\phi \approx \|\theta_{h+1} - \theta^\star_{h+1}\|_\infty^\phi \le 2L$$

$$\|\theta_{h+1} + \widetilde{\xi}_{h+1}\|_\infty^\phi \lesssim 3L$$

(was stuck here) $\theta^\star_h = \mathcal{T}(\theta_{h+1} + \widetilde{\xi}_{h+1})$

$$\|\theta^\star_h\|_\infty^\phi \lesssim 3L + 1 \checkmark$$

How about $\|\theta_h\|_\infty^\phi$?

A Second Visit to Norm

Algorithm (w/ known reward)

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

$$\theta_h \leftarrow \mathop{\arg\min}_{\theta: \|\theta\|_\infty^\phi \leq \|\theta^\star\|_\infty^\phi} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\langle \phi(s,a), \theta \rangle - r - V_{h+1}(s') \Big)^2$$

$$\widetilde{\xi}_h \sim \mathcal{N}(0, \sigma_h^2 (I - P_h))$$

$$Q_h(\cdot,\cdot) \leftarrow \langle \theta_h + \widetilde{\xi}_h, \phi(\cdot,\cdot) \rangle, \quad V_h(\cdot) \leftarrow \max_a Q_h(\cdot, a)$$

$\pi_t \leftarrow$ greedy policy w.r.t. Q_h

Collect data w/ π_t

Assume

$$\|\theta_{h+1}\|_\infty^\phi \leq L \quad \|\theta_{h+1}^\star\|_\infty^\phi \leq L$$

$$\|\widetilde{\xi}_{h+1}\|_\infty^\phi \approx \|\theta_{h+1} - \theta_{h+1}^\star\|_\infty^\phi \leq 2L$$

$$\|\theta_{h+1} + \widetilde{\xi}_{h+1}\|_\infty^\phi \lesssim 3L$$

(was stuck here) $\theta_h^\star = \mathcal{T}(\theta_{h+1} + \widetilde{\xi}_{h+1})$

$$\|\theta_h^\star\|_\infty^\phi \lesssim 3L + 1 \checkmark$$

Enforce $\|\theta_h\|_\infty^\phi \lesssim 3L + 1$

$$\|\theta_h^\star - \theta_h\|_\infty^\phi \lesssim 6L + 2$$

Set $\sigma_h \gtrsim 6L + 2 \checkmark$

Constrained Squared Loss Regression

$$\theta_h \leftarrow \underset{\theta : \|\theta\|_\infty^\phi \le \|\theta^\star\|_\infty^\phi}{\arg\min} \sum_{(s,a,r,s') \in \mathcal{D}_h} \Big(\big\langle \phi(s,a),\, \theta \big\rangle - r - V_{h+1}(s') \Big)^2$$

In general, need to solve…

$$\theta \leftarrow \arg\min_{\theta} \sum_{(x,y) \in \mathcal{D}} \Big(\big\langle x,\, \theta \big\rangle - y \Big)^2 \quad \text{s.t.} \quad \max_{s,a} \Big| \big\langle \phi(s,a),\, \theta \big\rangle \Big| \le W$$

Squared loss regression

ℓ_∞-functional constraints (exponential W)

Random Walks (Bertsimas & Vempala, 2004) ✓

needs a linear optimization oracle:

$$\max_{\phi(s,a)} \big\langle \phi(s,a),\, \theta \big\rangle$$

Outline

Part I Background

Part II Algorithm : the Trick of ***Span*** vs ***Null Space***

Part III The Norm Issue

Takeaways

An efficient RL algorithm under deterministic transition.

Key Ideas:

(1) ***Span*** vs ***Null Space***

(2) ℓ_∞-*functional-norm* instead of ℓ_2-*norm*

Future Work

1. Extending the "*Span* vs *Null Space*" technique to broader settings

2. Low-variance stochastic transitions

3. Can we ultimately resolve the linear BC problem?

Thank you! Any Questions?