Computationally Efficient RL Under Linear Bellman Completeness For Deterministic Dynamics

Runzhe Wu

Cornell University

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

Can we design provably efficient RL algorithm under

Linear (Value) Function Approximation ?

(Computationally + Statistically)

Can we design provably efficient RL algorithm under

Linear (Value) Function Approximation ?

(Computationally + Statistically)

Can we design provably efficient RL algorithm under

Linear Bellman Completeness ?

Linear Bellman Completeness

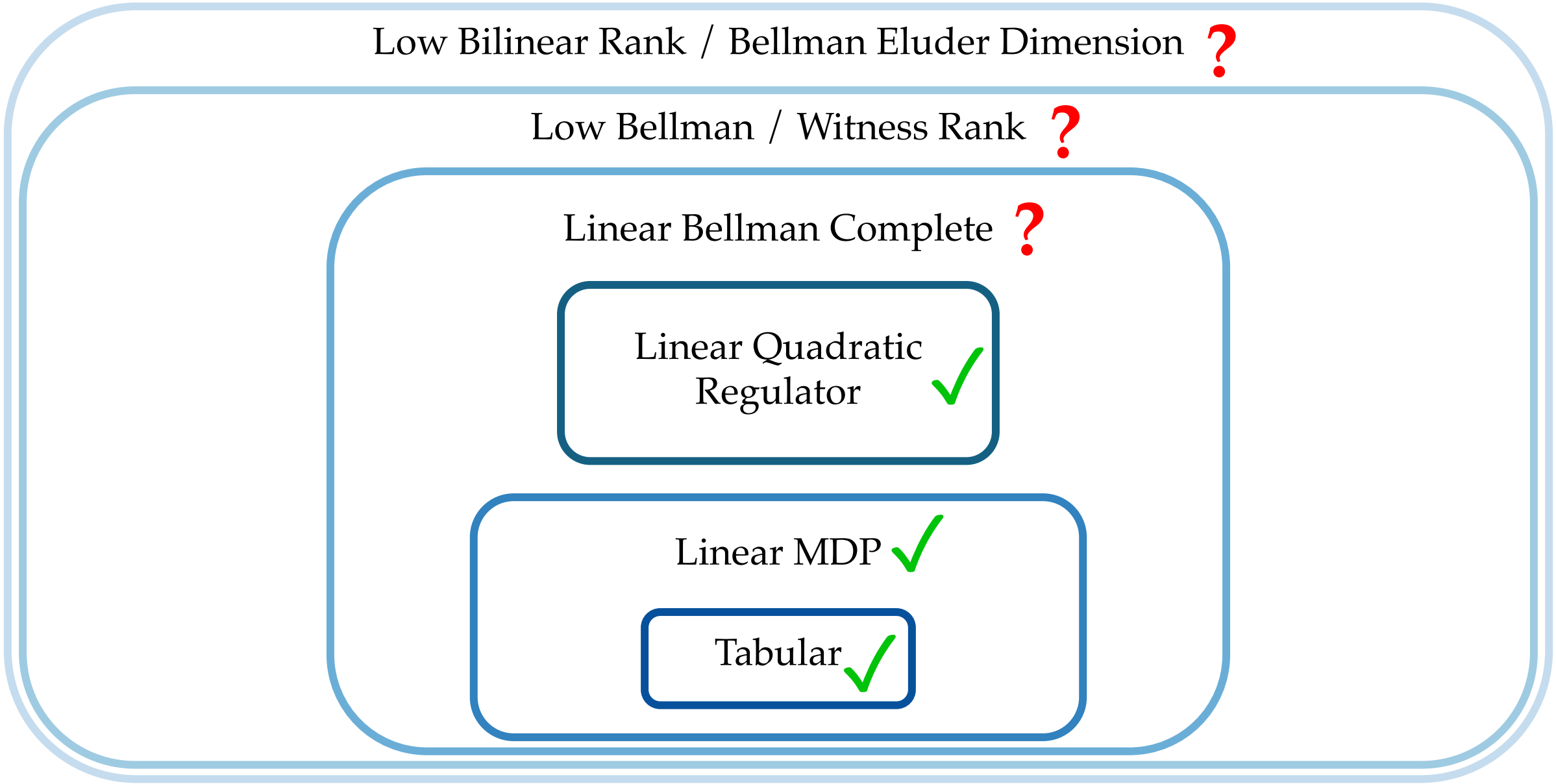
$$\forall f : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$$

$$\text{Define} \quad \mathcal{T} f(s,a) := r(s,a) + \mathop{\mathbb{E}}_{s' \sim P(s,a)} \max_{a'} f(s',a')$$

An MDP is **Linear Bellman Complete** w.r.t. a known feature map ϕ if

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

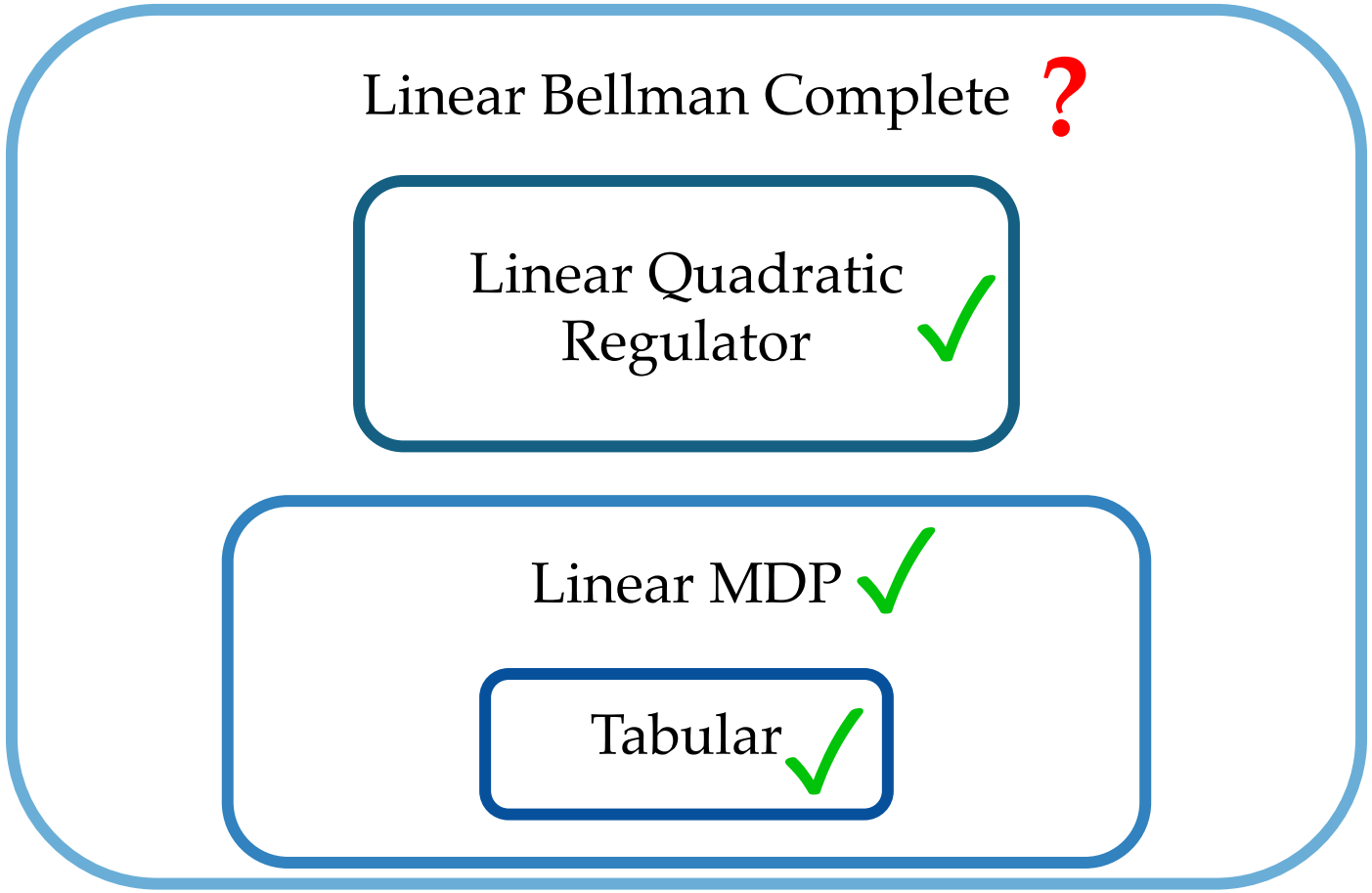
✓ : Computational and sample efficiency ? : Sample efficiency only Credit: Akshay Krishnamurthy



✓ : Computational and sample efficiency ? : Sample efficiency only Credit: Akshay Krishnamurthy

Open problem: Do efficient algorithms exist under linear BC?

Our contribution: They do when transitions are deterministic.



*Golowich and Moitra (2024) solve the problem for a constant number of actions.

RLSVI

[Russo, 2019]
[Zanette et al., 2020]

For $t = 1, \ldots, T$

  For $h = H, \ldots, 1$

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

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

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

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

  Collect data w/ π_t

Key Idea : ξ_h cancels out estimation error to achieve optimism

*Algorithm modified from original for presentation clarity.

RLSVI

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

Non-linear Bayes optimal

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

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

Not linear

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

Linear

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

Collect data w/ π_t

Apply to Linear BC ? ✗

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

Key Observation

For $t = 1, \ldots, T$

  For $h = H, \ldots, 1$

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

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

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

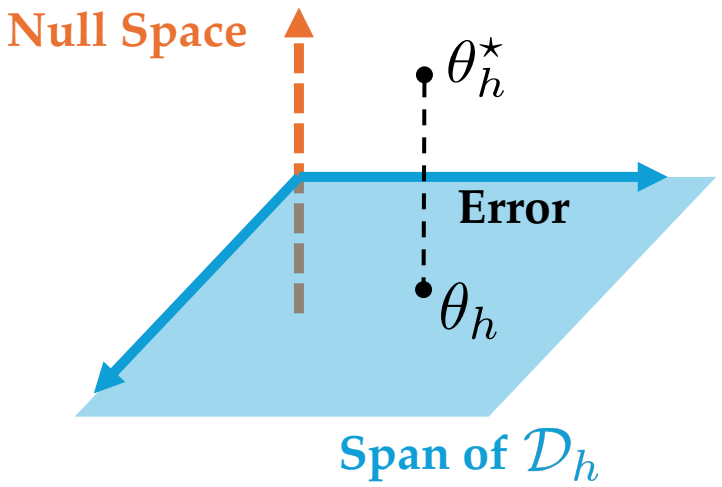
  $\pi_t \leftarrow$ greedy policy w.r.t. Q_h

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



P_h : orthogonal projection onto the span

$$P_h(\theta_h^\star - \theta_h) = 0$$

Key Observation

For $t = 1, \ldots, T$

For $h = H, \ldots, 1$

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

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

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

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

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

Collect data w/ π_t

+

1. Only explore in the **null space**
2. Remove clipping

Span Argument

Assume known reward

Fix $t \in [T]$

(1) All in Span

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

Optimism + Zero Empirical Risk

$\Downarrow$

Regret is **zero**

(2) Some in Null Space

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

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

Happens at most dH times

Regret $\leq dH^2$

Span Argument

Theorem (informal). *If reward is known, we have*

$$\mathbb{E}[\mathrm{Reg}_T] \leq dH^2$$

Standard reward learning

Theorem (informal). *If reward is unknown, we have*

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

*We assume deterministic transitions but allow adversarial initial states.

Takeaway

1. When transition is deterministic, consider the **span argument**
2. Regardless, efficient RL under **linear BC** remains an open problem

Thank you !