CSE 542: Statistical Reinforcement Learning

Lecture 15: Bellman Rank

Kevin Jamieson

Paul G. Allen School of Computer Science & Engineering
University of Washington

Outline

Today's Plan

What's the Most General Setting We Can Learn In?

Where we are.

- Tabular MDPs: $\widetilde{O}(\mathrm{poly}(S, A, H))$.
- Linear MDPs: $\widetilde{O}(\mathrm{poly}(d, H))$, no $|\mathcal{S}|$.
- Linear Bellman completeness: similar, with a regression-based analysis.

Each of these is a specific structural assumption. Question: is there a unifying complexity measure that captures *why* sample-efficient RL is possible in all these cases?

What's the Most General Setting We Can Learn In?

Where we are.

- Tabular MDPs: $\widetilde{O}(\mathrm{poly}(S, A, H))$.
- Linear MDPs: $\widetilde{O}(\mathrm{poly}(d, H))$, no $|\mathcal{S}|$.
- Linear Bellman completeness: similar, with a regression-based analysis.

Each of these is a specific structural assumption. Question: is there a unifying complexity measure that captures *why* sample-efficient RL is possible in all these cases?

The framework: bounded Bellman rank

A general PAC-RL framework due to Jiang et al. (2017):

- Given an arbitrary hypothesis class $\mathcal{H}$ of candidate Q-functions.
- Realizability ($Q^\star \in \mathcal{H}$) + a low-rank condition on the "Bellman error matrix."
- Polynomial sample complexity in d, H, $\log|\mathcal{H}|$, $1/\varepsilon$, with no $|\mathcal{S}|$ dependence.

Today's reference: AJKS Ch. 8
(Jiang–Krishnamurthy–Agarwal–Langford–Schapire 2017).

Setting

Finite-horizon MDP with horizon H, fixed start state s_0, rewards in $[0, 1]$.

Hypothesis class. $\mathcal{H}$ is a (possibly large but finite for now) class of Q-function candidates:

$$\mathcal{H} = \mathcal{H}_0 \times \mathcal{H}_1 \times \cdots \times \mathcal{H}_{H-1}.$$

Each $f \in \mathcal{H}$ gives functions $Q_{h,f}$, $V_{h,f}(s) := \max_a Q_{h,f}(s, a)$, and induces a *greedy* policy $\pi_{h,f}(s) := \arg\max_a Q_{h,f}(s, a)$.

Realizability. Assume $f^\star \in \mathcal{H}$ with $Q_{h,f^\star} = Q^\star_h$ for every h.

Goal (PAC-RL). Output $\widehat{\pi}$ with $V^\star_0(s_0) - V^{\widehat{\pi}}_0(s_0) \leq \varepsilon$ with prob. $\geq 1 - \delta$, using polynomially many *trajectories* from the MDP.

Setting

Finite-horizon MDP with horizon H, fixed start state s_0, rewards in $[0, 1]$.

Hypothesis class. $\mathcal{H}$ is a (possibly large but finite for now) class of Q-function candidates:

$$\mathcal{H} = \mathcal{H}_0 \times \mathcal{H}_1 \times \cdots \times \mathcal{H}_{H-1}.$$

Each $f \in \mathcal{H}$ gives functions $Q_{h,f}$, $V_{h,f}(s) := \max_a Q_{h,f}(s, a)$, and induces a *greedy* policy $\pi_{h,f}(s) := \arg\max_a Q_{h,f}(s, a)$.

Realizability. Assume $f^\star \in \mathcal{H}$ with $Q_{h,f^\star} = Q^\star_h$ for every h.

Goal (PAC-RL). Output $\widehat{\pi}$ with $V^\star_0(s_0) - V^{\widehat{\pi}}_0(s_0) \leq \varepsilon$ with prob. $\geq 1 - \delta$, using polynomially many *trajectories* from the MDP.

Why this matters

We've ditched all structural assumptions on P, r. The complexity will now depend only on:

- the size of $\mathcal{H}$ (statistical),
- a new "rank" quantity (algorithmic / informational).

Today's Plan

The Per-Hypothesis Bellman Residual

For any hypothesis $g \in \mathcal{H}$, define the one-step Bellman residual integrand:

$$\ell_h(s, a, s'; g) \;:=\; Q_{h,g}(s, a) \;-\; r_h(s, a) \;-\; V_{h+1,g}(s').$$

This is a *deterministic* function of (s, a, s') and the hypothesis g.

The Per-Hypothesis Bellman Residual

For any hypothesis $g \in \mathcal{H}$, define the one-step Bellman residual integrand:

$$\ell_h(s, a, s'; g) \;:=\; Q_{h,g}(s, a) \;-\; r_h(s, a) \;-\; V_{h+1,g}(s').$$

This is a *deterministic* function of (s, a, s') and the hypothesis g.

Observation. If $g = f^\star$ (the true $Q^\star$), then by the Bellman equation,

$$\mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[\ell_h(s, a, s'; f^\star)] \;=\; Q_h^\star(s, a) - r_h(s, a) - \mathbb{E}[V_{h+1}^\star(s')] \;=\; 0.$$

The Per-Hypothesis Bellman Residual

For any hypothesis $g \in \mathcal{H}$, define the one-step Bellman residual integrand:

$$\ell_h(s, a, s'; g) \;:=\; Q_{h,g}(s, a) \;-\; r_h(s, a) \;-\; V_{h+1,g}(s').$$

This is a *deterministic* function of (s, a, s') and the hypothesis g.

Observation. If $g = f^\star$ (the true $Q^\star$), then by the Bellman equation,

$$\mathbb{E}_{s' \sim P_h^\star(\cdot|s,a)}[\ell_h(s, a, s'; f^\star)] \;=\; Q_h^\star(s, a) - r_h(s, a) - \mathbb{E}[V_{h+1}^\star(s')] \;=\; 0.$$

For any candidate g, average ℓ_h over states/actions visited by some roll-in policy. Two flavors:

V-Bellman error (roll-in π_f, action π_g)

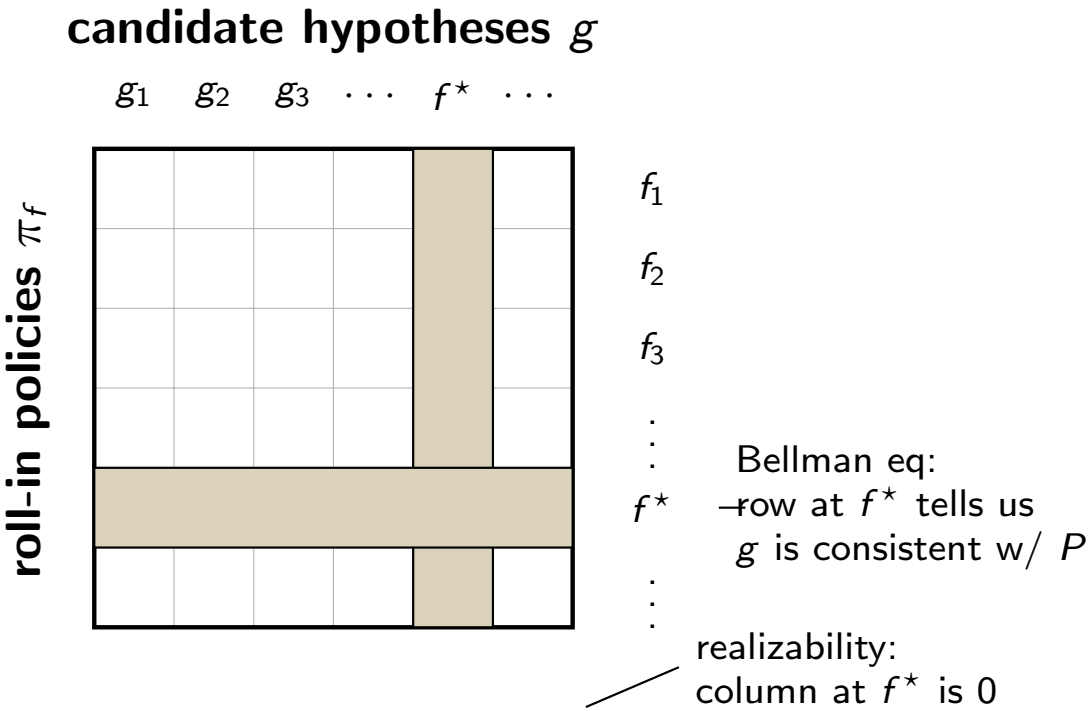
$$\mathcal{E}_h^V(f, g) \;:=\; \mathbb{E}_{s \sim d_h^{\pi_f},\; a = \pi_{h,g}(s),\; s' \sim P_h^\star}\big[\ell_h(s, a, s'; g)\big].$$

Q-Bellman error (roll-in π_f, action π_f)

$$\mathcal{E}_h^Q(f, g) \;:=\; \mathbb{E}_{(s,a) \sim d_h^{\pi_f},\; s' \sim P_h^\star}\big[\ell_h(s, a, s'; g)\big].$$

Think of It as a Matrix

The Bellman error is a function of two arguments: a roll-in policy (via f) and a candidate hypothesis (g). Picture it as a matrix:



Entry (f, g)**:** $\mathcal{E}_h(f, g)$ = how much does g's Bellman equation fail, on average, when we roll in with π_f?

$f^\star$-column is zero ($\mathcal{E}_h(f, f^\star) = 0$ for any f) by the Bellman equation.

Two Special Entries of the Matrix

Observation 1: the $f^\star$-column is zero. The true $Q^\star$ has zero Bellman residual. So for any roll-in f, $\mathcal{E}_h(f, f^\star) = \mathbb{E}_{(s,a)\sim d_h^{\pi_f}}[\, Q_h^\star(s,a) - r_h - \mathbb{E}[V_{h+1}^\star(s')]\,] = 0$.

Meaning: the true $Q^\star$ is consistent with every possible roll-in — realizability gives a free constraint we can use to test other g's.

Two Special Entries of the Matrix

Observation 1: the $f^\star$-column is zero. The true $Q^\star$ has zero Bellman residual. So for any roll-in f, $\mathcal{E}_h(f, f^\star) = \mathbb{E}_{(s,a)\sim d_h^{\pi_f}}[\, Q_h^\star(s,a) - r_h - \mathbb{E}[V_{h+1}^\star(s')]\,] = 0$.

Meaning: the true $Q^\star$ is consistent with every possible roll-in — realizability gives a free constraint we can use to test other g's.

Observation 2: a diagonal entry $\mathcal{E}_h(f, f)$ measures f's "self-prediction" error.

By "diagonal" we mean $\mathcal{E}_h(f, f)$ — roll in with π_f *and* evaluate the hypothesis f (same f for both row and column).

Telescoping identity (*not* the PDL)

$$V_{0,f}(s_0) \;-\; V_0^{\pi_f}(s_0) \;=\; \sum_{h=0}^{H-1} \mathcal{E}_h(f, f).$$

LHS: the value f *predicts* for the start state, minus the value π_f *actually achieves* in the true MDP.
RHS: sum of the diagonal entries.

This is not the performance-difference lemma (which compares two *true* value functions via advantages). It's the same telescoping argument we did in Lecture 7 Lemma A, now applied to "hypothesis-predicted" vs. "actually-realized" value of π_f. Proof on the next slide.

Special case: $\mathcal{E}_h(f, f) = 0\ \forall h \implies V_{0,f}(s_0) = V_0^{\pi_f}(s_0)$ (self-prediction matches reality).

Aside: Proof of the Telescoping Identity

Let $a_h := \pi_{h,f}(s_h)$, so $V_{h,f}(s_h) = Q_{h,f}(s_h, a_h)$ (greedy). Let $\Delta_h(s,a) := Q_{h,f}(s,a) - r_h(s,a) - \mathbb{E}_{s' \sim P_h^\star}[V_{h+1,f}(s')]$. Note $\mathbb{E}[\Delta_h(s,a)]$ averaged under $(s,a) \sim d_h^{\pi_f}$ is exactly $\mathcal{E}_h(f,f)$.

One-step decomposition. $Q_h^{\pi_f}$ satisfies $Q_h^{\pi_f}(s,a) = r_h(s,a) + \mathbb{E}_{s'}[V_{h+1}^{\pi_f}(s')]$. So

$$\begin{aligned} V_{h,f}(s_h) - V_h^{\pi_f}(s_h) &= Q_{h,f}(s_h,a_h) - Q_h^{\pi_f}(s_h,a_h) \\ &= \underbrace{Q_{h,f}(s_h,a_h) - r_h(s_h,a_h) - \mathbb{E}[V_{h+1,f}(s_{h+1})]}_{=\Delta_h(s_h,a_h)} \\ &\quad + \mathbb{E}_{s_{h+1} \sim P_h^\star(\cdot|s_h,a_h)}\big[V_{h+1,f}(s_{h+1}) - V_{h+1}^{\pi_f}(s_{h+1})\big]. \end{aligned}$$

Aside: Proof of the Telescoping Identity

Let $a_h := \pi_{h,f}(s_h)$, so $V_{h,f}(s_h) = Q_{h,f}(s_h, a_h)$ (greedy). Let $\Delta_h(s,a) := Q_{h,f}(s,a) - r_h(s,a) - \mathbb{E}_{s' \sim P_h^\star}[V_{h+1,f}(s')]$. Note $\mathbb{E}[\Delta_h(s,a)]$ averaged under $(s,a) \sim d_h^{\pi_f}$ is exactly $\mathcal{E}_h(f,f)$.

One-step decomposition. $Q_h^{\pi_f}$ satisfies $Q_h^{\pi_f}(s,a) = r_h(s,a) + \mathbb{E}_{s'}[V_{h+1}^{\pi_f}(s')]$. So

$$\begin{aligned} V_{h,f}(s_h) - V_h^{\pi_f}(s_h) &= Q_{h,f}(s_h,a_h) - Q_h^{\pi_f}(s_h,a_h) \\ &= \underbrace{Q_{h,f}(s_h,a_h) - r_h(s_h,a_h) - \mathbb{E}[V_{h+1,f}(s_{h+1})]}_{=\Delta_h(s_h,a_h)} \\ &\quad + \mathbb{E}_{s_{h+1} \sim P_h^\star(\cdot|s_h,a_h)}\big[V_{h+1,f}(s_{h+1}) - V_{h+1}^{\pi_f}(s_{h+1})\big]. \end{aligned}$$

Telescope. Iterate the recursion from $h = 0$ down to $h = H$ ($V_{H,f} \equiv V_H^{\pi_f} \equiv 0$), taking expectation along the trajectory generated by π_f in the true MDP:

$$V_{0,f}(s_0) - V_0^{\pi_f}(s_0) \;=\; \mathbb{E}_{\pi_f}\left[\sum_{h=0}^{H-1} \Delta_h(s_h,a_h)\right] \;=\; \sum_h \mathbb{E}_{(s,a)\sim d_h^{\pi_f}}[\Delta_h(s,a)] \;=\; \sum_h \mathcal{E}_h(f,f). \qquad \square$$

Aside: Proof of the Telescoping Identity

Let $a_h := \pi_{h,f}(s_h)$, so $V_{h,f}(s_h) = Q_{h,f}(s_h, a_h)$ (greedy). Let $\Delta_h(s,a) := Q_{h,f}(s,a) - r_h(s,a) - \mathbb{E}_{s' \sim P_h^\star}[V_{h+1,f}(s')]$. Note $\mathbb{E}[\Delta_h(s,a)]$ averaged under $(s,a) \sim d_h^{\pi_f}$ is exactly $\mathcal{E}_h(f,f)$.

One-step decomposition. $Q_h^{\pi_f}$ satisfies $Q_h^{\pi_f}(s,a) = r_h(s,a) + \mathbb{E}_{s'}[V_{h+1}^{\pi_f}(s')]$. So

$$\begin{aligned} V_{h,f}(s_h) - V_h^{\pi_f}(s_h) &= Q_{h,f}(s_h,a_h) - Q_h^{\pi_f}(s_h,a_h) \\ &= \underbrace{Q_{h,f}(s_h,a_h) - r_h(s_h,a_h) - \mathbb{E}[V_{h+1,f}(s_{h+1})]}_{=\Delta_h(s_h,a_h)} \\ &\quad + \mathbb{E}_{s_{h+1} \sim P_h^\star(\cdot|s_h,a_h)}\big[V_{h+1,f}(s_{h+1}) - V_{h+1}^{\pi_f}(s_{h+1})\big]. \end{aligned}$$

Telescope. Iterate the recursion from $h = 0$ down to $h = H$ ($V_{H,f} \equiv V_H^{\pi_f} \equiv 0$), taking expectation along the trajectory generated by π_f in the true MDP:

$$V_{0,f}(s_0) - V_0^{\pi_f}(s_0) \;=\; \mathbb{E}_{\pi_f}\left[\sum_{h=0}^{H-1} \Delta_h(s_h,a_h)\right] \;=\; \sum_h \mathbb{E}_{(s,a)\sim d_h^{\pi_f}}[\Delta_h(s,a)] \;=\; \sum_h \mathcal{E}_h(f,f). \qquad \square$$

What's Δ_h?

$\Delta_h(s,a)$ is the Bellman residual of f at (s,a) — "the inconsistency of using f to predict its own next-stage value." Averaging Δ_h over states/actions actually visited by π_f gives the diagonal entry $\mathcal{E}_h(f,f)$.

Why This Matrix Is the Right Object

Small diagonal + optimism $\Rightarrow$ near-optimal policy.

Suppose the algorithm guarantees optimism: $V_{0,f_t}(s_0) \geq V_0^\star(s_0)$. Then by Observation 2:

$$V_0^\star(s_0) - V_0^{\pi_{f_t}}(s_0) \;\leq\; V_{0,f_t}(s_0) - V_0^{\pi_{f_t}}(s_0) \;=\; \sum_h \mathcal{E}_h(f_t, f_t).$$

So if we can force the diagonal entries to be small for some f_t, π_{f_t} is near-optimal.

Why This Matrix Is the Right Object

Small diagonal + optimism $\Rightarrow$ near-optimal policy.

Suppose the algorithm guarantees optimism: $V_{0,f_t}(s_0) \geq V_0^\star(s_0)$. Then by Observation 2:

$$V_0^\star(s_0) - V_0^{\pi_{f_t}}(s_0) \;\leq\; V_{0,f_t}(s_0) - V_0^{\pi_{f_t}}(s_0) \;=\; \sum_h \mathcal{E}_h(f_t, f_t).$$

So if we can force the diagonal entries to be small for some f_t, π_{f_t} is near-optimal.

Where will optimism come from? Observation 1 (the $f^\star$-column is zero) lets the algorithm build constraints that $f^\star$ always satisfies, hence $f^\star$ is always feasible. Maximizing the predicted value over the feasible set then yields $V_{0,f_t}(s_0) \geq V_{0,f^\star}(s_0) = V_0^\star(s_0)$.

Why This Matrix Is the Right Object

Small diagonal + optimism $\Rightarrow$ near-optimal policy.

Suppose the algorithm guarantees optimism: $V_{0,f_t}(s_0) \geq V_0^\star(s_0)$. Then by Observation 2:

$$V_0^\star(s_0) - V_0^{\pi_{f_t}}(s_0) \;\leq\; V_{0,f_t}(s_0) - V_0^{\pi_{f_t}}(s_0) \;=\; \sum_h \mathcal{E}_h(f_t, f_t).$$

So if we can force the diagonal entries to be small for some f_t, π_{f_t} is near-optimal.

Where will optimism come from? Observation 1 (the $f^\star$-column is zero) lets the algorithm build constraints that $f^\star$ always satisfies, hence $f^\star$ is always feasible. Maximizing the predicted value over the feasible set then yields $V_{0,f_t}(s_0) \geq V_{0,f^\star}(s_0) = V_0^\star(s_0)$.

The PAC-RL game

We don't see the matrix. At iteration t, the algorithm:

- picks a row f_t (the roll-in),
- estimates the row entries $\mathcal{E}_h(f_t, g)$ for every column g from a batch of trajectories,
- uses these to eliminate g's whose row entries are too far from zero.

Goal: after few iterations, force some f_t to have small *diagonal* $\mathcal{E}_h(f_t, f_t)$.

Today's Plan

Definition: Bellman Rank

Definition (Bellman rank, Jiang et al. 2017)

$(\mathcal{M}, \mathcal{H})$ has **Bellman rank at most** d (in V- or Q-mode) if, for each stage h, there exist maps

$$X_h : \mathcal{H} \to \mathbb{R}^d, \qquad W_h : \mathcal{H} \to \mathbb{R}^d,$$

such that for every pair $f, g \in \mathcal{H}$,

$$\mathcal{E}_h(f, g) \;=\; \langle X_h(f),\; W_h(g) \rangle \,.$$

(With bounded norms $\|X_h(f)\|_2 \leq B_X$, $\|W_h(g)\|_2 \leq B_W$.)

Definition: Bellman Rank

Definition (Bellman rank, Jiang et al. 2017)

$(\mathcal{M}, \mathcal{H})$ has **Bellman rank at most** d (in V- or Q-mode) if, for each stage h, there exist maps

$$X_h : \mathcal{H} \to \mathbb{R}^d, \qquad W_h : \mathcal{H} \to \mathbb{R}^d,$$

such that for every pair $f, g \in \mathcal{H}$,

$$\mathcal{E}_h(f, g) \;=\; \langle X_h(f),\; W_h(g) \rangle \,.$$

(With bounded norms $\|X_h(f)\|_2 \leq B_X$, $\|W_h(g)\|_2 \leq B_W$.)

What this says: the giant $|\mathcal{H}| \times |\mathcal{H}|$ matrix $[\mathcal{E}_h(f, g)]_{f,g}$ has rank at most d.

- $X_h(f)$: a d-dimensional "signature" of the *roll-in policy* π_f.
- $W_h(g)$: a d-dimensional "signature" of the *hypothesis* g.

Definition: Bellman Rank

Definition (Bellman rank, Jiang et al. 2017)

$(\mathcal{M}, \mathcal{H})$ has **Bellman rank at most** d (in V- or Q-mode) if, for each stage h, there exist maps

$$X_h : \mathcal{H} \to \mathbb{R}^d, \qquad W_h : \mathcal{H} \to \mathbb{R}^d,$$

such that for every pair $f, g \in \mathcal{H}$,

$$\mathcal{E}_h(f, g) \;=\; \langle X_h(f),\; W_h(g) \rangle \,.$$

(With bounded norms $\|X_h(f)\|_2 \le B_X$, $\|W_h(g)\|_2 \le B_W$.)

What this says: the giant $|\mathcal{H}| \times |\mathcal{H}|$ matrix $[\mathcal{E}_h(f, g)]_{f,g}$ has rank at most d.

- $X_h(f)$: a d-dimensional "signature" of the *roll-in policy* π_f.
- $W_h(g)$: a d-dimensional "signature" of the *hypothesis* g.

Convention. Under realizability, $\mathcal{E}_h(f, f^\star) = 0$ for all f, so we can take $W_h(f^\star) = 0$ without loss of generality. "How far is g from being correct" is then measured by $\|W_h(g)\|$.

Low Rank = Few Directions to Search

The matrix $[\mathcal{E}_h(f, g)]$ has rank $\leq d$. Equivalently, the set of "roll-in signatures" $\{X_h(f) : f \in \mathcal{H}\} \subset \mathbb{R}^d$ spans an at-most-d-dimensional subspace.

Low Rank = Few Directions to Search

The matrix $[\mathcal{E}_h(f, g)]$ has rank $\leq d$. Equivalently, the set of "roll-in signatures" $\{X_h(f) : f \in \mathcal{H}\} \subset \mathbb{R}^d$ spans an at-most-d-dimensional subspace.

Why this changes the game

Suppose we pick a roll-in f_t and observe $\langle X_h(f_t), W_h(g) \rangle = \mathcal{E}_h(f_t, g)$ for every $g \in \mathcal{H}$. This is one linear equation for each g, with unknown "parameter" $W_h(g)$.

After at most d *linearly independent* roll-in directions $X_h(f_0), X_h(f_1), \ldots,$ $W_h(g)$ is fully determined — regardless of $|\mathcal{H}|$.

$\Rightarrow$ at most $\widetilde{O}(d)$ rounds are needed, each costing $\mathrm{poly}(H, \log|\mathcal{H}|/\varepsilon)$ trajectories.

Low Rank = Few Directions to Search

The matrix $[\mathcal{E}_h(f, g)]$ has rank $\leq d$. Equivalently, the set of "roll-in signatures" $\{X_h(f) : f \in \mathcal{H}\} \subset \mathbb{R}^d$ spans an at-most-d-dimensional subspace.

Why this changes the game

Suppose we pick a roll-in f_t and observe $\langle X_h(f_t), W_h(g) \rangle = \mathcal{E}_h(f_t, g)$ for every $g \in \mathcal{H}$. This is one linear equation for each g, with unknown "parameter" $W_h(g)$.

After at most d *linearly independent* roll-in directions $X_h(f_0), X_h(f_1), \ldots,$ $W_h(g)$ is fully determined — regardless of $|\mathcal{H}|$.

$\Rightarrow$ at most $\widetilde{O}(d)$ rounds are needed, each costing $\mathrm{poly}(H, \log|\mathcal{H}|/\varepsilon)$ trajectories.

Where statistics enters. We don't see $\mathcal{E}_h$ exactly — we estimate it from finite samples. So the bound is more like "at most $\widetilde{O}(dH)$ rounds, each batch size $m = \mathrm{poly}(H, \log|\mathcal{H}|, 1/\varepsilon)$."

Q vs. V Bellman Rank

The two versions differ only in how actions are chosen at the test stage h.

	Q-Bellman rank	**V-Bellman rank**
Roll-in	π_f to stage h	π_f to stage h
Action at h	$a_h = \pi_{h,f}(s_h)$	$a_h = \pi_{h,g}(s_h)$
Estimation	directly on-policy	importance weight (uniform a_h)
$\lvert\mathcal{A}\rvert$ factor	none	yes ($\varepsilon_{\mathrm{gen}} \propto A$)

Q vs. V Bellman Rank

The two versions differ only in how actions are chosen at the test stage h.

	Q-Bellman rank	**V-Bellman rank**
Roll-in	π_f to stage h	π_f to stage h
Action at h	$a_h = \pi_{h,f}(s_h)$	$a_h = \pi_{h,g}(s_h)$
Estimation	directly on-policy	importance weight (uniform a_h)
$\lvert\mathcal{A}\rvert$ factor	none	yes ($\varepsilon_{\mathrm{gen}} \propto A$)

Q-mode. At test time we use the action π_f would take — a *single* roll-out trajectory under π_f gives unbiased estimates of $\mathcal{E}_h^Q(f, g)$ for all g.

V-mode. At test time we use π_g's action — but g varies across the comparison. Inject a uniform action at h and importance-weight. Single roll-out under π_f still works for all g, but at cost factor A.

Q vs. V Bellman Rank

The two versions differ only in how actions are chosen at the test stage h.

	Q-Bellman rank	**V-Bellman rank**
Roll-in	π_f to stage h	π_f to stage h
Action at h	$a_h = \pi_{h,f}(s_h)$	$a_h = \pi_{h,g}(s_h)$
Estimation	directly on-policy	importance weight (uniform a_h)
$\lvert\mathcal{A}\rvert$ factor	none	yes ($\varepsilon_{\mathrm{gen}} \propto A$)

Q-mode. At test time we use the action π_f would take — a *single* roll-out trajectory under π_f gives unbiased estimates of $\mathcal{E}_h^Q(f, g)$ for all g.

V-mode. At test time we use π_g's action — but g varies across the comparison. Inject a uniform action at h and importance-weight. Single roll-out under π_f still works for all g, but at cost factor A.

Convention for the rest of the lecture. We'll mostly talk about Q-mode for cleanliness; V-mode is similar with an A overhead.

Examples

Many familiar models have bounded Bellman rank.

- **Tabular MDP.** $d \leq SA$ (just take $X_h(f) = d_h^{\pi_f}$ as a vector indexed by (s, a), $W_h(g) =$ Bellman residual at each (s, a)).
- **Linear MDP.** $\mathcal{E}_h^Q(f, g) = \mathbb{E}_{d_h^{\pi_f}}[\langle \phi(s, a), w_h^g - w_h^\star \rangle]$.
  Set $X_h(f) = \mathbb{E}[\phi(s, a)]_{(s,a)\sim d_h^{\pi_f}}$ and $W_h(g) = w_h^g - w_h^\star$. Rank $= d$.
- **Linear Bellman completeness.** $W_h(g) = w_{h,g} - \mathcal{T}_h(w_{h+1,g})$, $X_h(f) = \mathbb{E}_{d_h^{\pi_f}}[\phi]$.
  Rank $\leq d$.
- **Contextual linear bandits ($H = 1$).** Trivially Bellman rank $\leq d$.
- **Low-rank transitions** $P_h(s'|s, a) = \langle \phi^\star(s, a), \mu^\star(s') \rangle$: V-Bellman rank $\leq d$, even if $\phi^\star$ is unknown.

Examples

Many familiar models have bounded Bellman rank.

- **Tabular MDP.** $d \leq SA$ (just take $X_h(f) = d_h^{\pi_f}$ as a vector indexed by (s, a), $W_h(g) =$ Bellman residual at each (s, a)).
- **Linear MDP.** $\mathcal{E}_h^Q(f, g) = \mathbb{E}_{d_h^{\pi_f}}[\langle \phi(s, a), w_h^g - w_h^\star \rangle]$.
  Set $X_h(f) = \mathbb{E}[\phi(s, a)]_{(s,a)\sim d_h^{\pi_f}}$ and $W_h(g) = w_h^g - w_h^\star$. Rank $= d$.
- **Linear Bellman completeness.** $W_h(g) = w_{h,g} - \mathcal{T}_h(w_{h+1,g})$, $X_h(f) = \mathbb{E}_{d_h^{\pi_f}}[\phi]$.
  Rank $\leq d$.
- **Contextual linear bandits ($H = 1$).** Trivially Bellman rank $\leq d$.
- **Low-rank transitions** $P_h(s'|s, a) = \langle \phi^\star(s, a), \mu^\star(s') \rangle$: V-Bellman rank $\leq d$, even if $\phi^\star$ is unknown.

Takeaway

Bellman rank is a *unifying* parameter. Each of the structural assumptions we've seen so far (linear MDP, linear completeness, low-rank transitions) implies bounded Bellman rank — so the same algorithm gives PAC bounds for all of them, paying only in $\log |\mathcal{H}|$ on top of $\mathrm{poly}(d, H)$.

Today's Plan

The Algorithm in One Picture

Idea. Iterate: pick an *optimistic* hypothesis f_t that hasn't yet been ruled out, roll in with π_{f_t}, use the resulting batch to eliminate any g whose Bellman error along this roll-in is too large.

Algorithm (Bellman-rank optimistic elimination, AJKS Alg. 8)

For $t = 0, 1, \ldots, T-1$:

1. Pick $f_t \in \arg\max_{g \in \mathcal{H}} V_{0,g}(s_0)$ subject to

$$\sum_{i<t} \widehat{\mathcal{E}}_{i,h}(g)^2 \leq R^2 \quad \forall h.$$

2. Roll in with π_{f_t} to collect a batch $D_{t,h}$ at each stage.
3. Compute, for each $g \in \mathcal{H}$ and stage h, the estimate $\widehat{\mathcal{E}}_{t,h}(g)$.

Output: $\widehat{\pi} := \pi_{f_{\hat{t}}}$ where $\hat{t} = \arg\min_t \sum_h \widehat{\mathcal{E}}_{t,h}(f_t)$.

The Algorithm in One Picture

Idea. Iterate: pick an *optimistic* hypothesis f_t that hasn't yet been ruled out, roll in with π_{f_t}, use the resulting batch to eliminate any g whose Bellman error along this roll-in is too large.

Algorithm (Bellman-rank optimistic elimination, AJKS Alg. 8)

For $t = 0, 1, \ldots, T-1$:

1. Pick $f_t \in \arg\max_{g \in \mathcal{H}} V_{0,g}(s_0)$ subject to

$$\sum_{i<t} \widehat{\mathcal{E}}_{i,h}(g)^2 \;\leq\; R^2 \quad \forall h.$$

2. Roll in with π_{f_t} to collect a batch $D_{t,h}$ at each stage.
3. Compute, for each $g \in \mathcal{H}$ and stage h, the estimate $\widehat{\mathcal{E}}_{t,h}(g)$.

Output: $\widehat{\pi} := \pi_{f_{\hat{t}}}$ where $\hat{t} = \arg\min_t \sum_h \widehat{\mathcal{E}}_{t,h}(f_t)$.

Reading the constraint. "g is currently feasible" means: in every past roll-in f_i, the Bellman residual of g averaged out to ≈ 0. By the matrix picture, this is a set of (noisy) linear equations on $W_h(g)$.

Uniform Estimation

What the algorithm actually sees. Each iteration t produces a batch of m trajectories rolled in by π_{f_t}, from which we form, for every $g \in \mathcal{H}$ and stage h, the empirical Bellman-error estimate

$$\widehat{\mathcal{E}}_{t,h}(g) \;:=\; \tfrac{1}{m} \sum_{j=1}^{m} \tilde{\ell}_h(s^j, a^j, s'^{,j}; g),$$

where $\tilde{\ell}_h = \ell_h$ in Q-mode, $\tilde{\ell}_h$ is the importance-weighted version in V-mode.

Uniform Estimation

What the algorithm actually sees. Each iteration t produces a batch of m trajectories rolled in by π_{f_t}, from which we form, for every $g \in \mathcal{H}$ and stage h, the empirical Bellman-error estimate

$$\widehat{\mathcal{E}}_{t,h}(g) \;:=\; \tfrac{1}{m}\sum_{j=1}^{m} \tilde{\ell}_h(s^j, a^j, s'^{,j}; g),$$

where $\tilde{\ell}_h = \ell_h$ in Q-mode, $\tilde{\ell}_h$ is the importance-weighted version in V-mode.

Lemma (Uniform estimation event $\mathcal{E}$)

There exists $\varepsilon_{\mathrm{gen}} = \varepsilon_{\mathrm{gen}}(m, H, \delta/(TH))$ such that with probability $\geq 1-\delta$, simultaneously for all $t < T$, $h < H$, $g \in \mathcal{H}$:

$$\left|\widehat{\mathcal{E}}_{t,h}(g) \;-\; \mathcal{E}_h(f_t, g)\right| \;\leq\; \varepsilon_{\mathrm{gen}}.$$

For finite $\mathcal{H}$ in Q-mode: $\varepsilon_{\mathrm{gen}} = O\big(H\sqrt{\log(|\mathcal{H}|TH/\delta)/m}\big)$.
For V-mode: extra factor of $|\mathcal{A}|$ from importance weighting.

Uniform Estimation

What the algorithm actually sees. Each iteration t produces a batch of m trajectories rolled in by π_{f_t}, from which we form, for every $g \in \mathcal{H}$ and stage h, the empirical Bellman-error estimate

$$\widehat{\mathcal{E}}_{t,h}(g) \;:=\; \tfrac{1}{m}\sum_{j=1}^{m} \tilde{\ell}_h(s^j, a^j, s'^{,j}; g),$$

where $\tilde{\ell}_h = \ell_h$ in Q-mode, $\tilde{\ell}_h$ is the importance-weighted version in V-mode.

Lemma (Uniform estimation event $\mathcal{E}$)

There exists $\varepsilon_{\mathrm{gen}} = \varepsilon_{\mathrm{gen}}(m, H, \delta/(TH))$ such that with probability $\geq 1-\delta$, simultaneously for all $t < T$, $h < H$, $g \in \mathcal{H}$:

$$\left|\widehat{\mathcal{E}}_{t,h}(g) \;-\; \mathcal{E}_h(f_t, g)\right| \;\leq\; \varepsilon_{\mathrm{gen}}.$$

For finite $\mathcal{H}$ in Q-mode: $\varepsilon_{\mathrm{gen}} = O\big(H\sqrt{\log(|\mathcal{H}|TH/\delta)/m}\big)$.
For V-mode: extra factor of $|\mathcal{A}|$ from importance weighting.

What this gives us. A single "good event" on which every empirical Bellman-error estimate (across iterations t, stages h, and hypotheses g) tracks the population quantity to within $\varepsilon_{\mathrm{gen}}$. We condition on $\mathcal{E}$ for the rest of the analysis.

Why the Algorithm Works: Choosing R and Optimism

Condition on the uniform estimation event $\mathcal{E}$ throughout.

Step 1: Choice of R. Set

$$\boxed{R^2 \;:=\; T\,\varepsilon_{\mathrm{gen}}^2.}$$

Rationale. The feasibility constraint accumulates one squared error of size at most $\varepsilon_{\mathrm{gen}}^2$ per past iteration (under $\mathcal{E}$). Over T iterations, $T\varepsilon_{\mathrm{gen}}^2$ is the smallest budget that lets $f^\star$ stay feasible no matter what roll-ins $f_0, \ldots, f_{T-1}$ the algorithm has played.

Why the Algorithm Works: Choosing R and Optimism

Condition on the uniform estimation event $\mathcal{E}$ throughout.

Step 1: Choice of R. Set

$$\boxed{R^2 \;:=\; T\,\varepsilon_{\mathrm{gen}}^2.}$$

Rationale. The feasibility constraint accumulates one squared error of size at most $\varepsilon_{\mathrm{gen}}^2$ per past iteration (under $\mathcal{E}$). Over T iterations, $T\varepsilon_{\mathrm{gen}}^2$ is the smallest budget that lets $f^\star$ stay feasible no matter what roll-ins $f_0, \ldots, f_{T-1}$ the algorithm has played.

Step 2: $f^\star$ is always feasible. By realizability, $\mathcal{E}_h(f, f^\star) = 0$ for every f, h. On $\mathcal{E}$, $|\widehat{\mathcal{E}}_{i,h}(f^\star)| = |\widehat{\mathcal{E}}_{i,h}(f^\star) - \mathcal{E}_h(f_i, f^\star)| \le \varepsilon_{\mathrm{gen}}$ for every i, h. Hence

$$\sum_{i<t} \widehat{\mathcal{E}}_{i,h}(f^\star)^2 \;\le\; t\,\varepsilon_{\mathrm{gen}}^2 \;\le\; T\,\varepsilon_{\mathrm{gen}}^2 \;=\; R^2.$$

$f^\star$ is feasible at every (t, h). ✓

Why the Algorithm Works: Choosing R and Optimism

Condition on the uniform estimation event $\mathcal{E}$ throughout.

Step 1: Choice of R. Set

$$\boxed{R^2 \;:=\; T\,\varepsilon_{\mathrm{gen}}^2.}$$

Rationale. The feasibility constraint accumulates one squared error of size at most $\varepsilon_{\mathrm{gen}}^2$ per past iteration (under $\mathcal{E}$). Over T iterations, $T\varepsilon_{\mathrm{gen}}^2$ is the smallest budget that lets $f^\star$ stay feasible no matter what roll-ins $f_0, \ldots, f_{T-1}$ the algorithm has played.

Step 2: $f^\star$ is always feasible. By realizability, $\mathcal{E}_h(f, f^\star) = 0$ for every f, h. On $\mathcal{E}$, $|\widehat{\mathcal{E}}_{i,h}(f^\star)| = |\widehat{\mathcal{E}}_{i,h}(f^\star) - \mathcal{E}_h(f_i, f^\star)| \le \varepsilon_{\mathrm{gen}}$ for every i, h. Hence

$$\sum_{i<t} \widehat{\mathcal{E}}_{i,h}(f^\star)^2 \;\le\; t\,\varepsilon_{\mathrm{gen}}^2 \;\le\; T\,\varepsilon_{\mathrm{gen}}^2 \;=\; R^2.$$

$f^\star$ is feasible at every (t, h). ✓

Step 3: Optimism. Since f_t maximizes $V_{0,g}(s_0)$ subject to feasibility,

$$V_{0,f_t}(s_0) \;\ge\; V_{0,f^\star}(s_0) \;=\; V_0^\star(s_0).$$

Why the Algorithm Works: Self-Error Gives the Gap

Still on $\mathcal{E}$.

Step 4: Self-error. By the telescoping identity (derived earlier) applied to $f = f_t$: the residual of Q_{h,f_t} averaged along the trajectory π_{f_t} generates in the true MDP gives

$$V_{0,f_t}(s_0) \; - \; V_0^{\pi_{f_t}}(s_0) \; = \; \mathbb{E}_{\pi_{f_t}}\left[\sum_{h=0}^{H-1} \ell_h(s_h, a_h, s_{h+1}; f_t)\right] \; = \; \sum_{h=0}^{H-1} \mathcal{E}_h(f_t, f_t).$$

Why the Algorithm Works: Self-Error Gives the Gap

Still on $\mathcal{E}$.

Step 4: Self-error. By the telescoping identity (derived earlier) applied to $f = f_t$: the residual of Q_{h,f_t} averaged along the trajectory π_{f_t} generates in the true MDP gives

$$V_{0,f_t}(s_0) \; - \; V_0^{\pi_{f_t}}(s_0) \; = \; \mathbb{E}_{\pi_{f_t}}\left[\sum_{h=0}^{H-1} \ell_h(s_h, a_h, s_{h+1}; f_t)\right] \; = \; \sum_{h=0}^{H-1} \mathcal{E}_h(f_t, f_t).$$

Combining with optimism (Step 3):

$$\boxed{V_0^\star(s_0) - V_0^{\pi_{f_t}}(s_0) \; \leq \; \sum_{h=0}^{H-1} \mathcal{E}_h(f_t, f_t).}$$

Why the Algorithm Works: Self-Error Gives the Gap

Still on $\mathcal{E}$.

Step 4: Self-error. By the telescoping identity (derived earlier) applied to $f = f_t$: the residual of Q_{h,f_t} averaged along the trajectory π_{f_t} generates in the true MDP gives

$$V_{0,f_t}(s_0) \; - \; V_0^{\pi_{f_t}}(s_0) \; = \; \mathbb{E}_{\pi_{f_t}}\left[\sum_{h=0}^{H-1} \ell_h(s_h, a_h, s_{h+1}; f_t)\right] \; = \; \sum_{h=0}^{H-1} \mathcal{E}_h(f_t, f_t).$$

Combining with optimism (Step 3):

$$\boxed{V_0^\star(s_0) - V_0^{\pi_{f_t}}(s_0) \; \leq \; \sum_{h=0}^{H-1} \mathcal{E}_h(f_t, f_t).}$$

Remaining question

Why must some $t \in \{0, \ldots, T-1\}$ have small $\sum_h \mathcal{E}_h(f_t, f_t)$, with $T = \mathrm{poly}(d, H)$?
Bounded Bellman rank — next slide.

From Feasibility to a Bound on $W_h(f_t)$, Part 1

Goal. Translate feasibility (an *empirical* condition on $\widehat{\mathcal{E}}$) into a bound on the *signature* $W_h(f_t) \in \mathbb{R}^d$ given by the Bellman-rank factorization $\mathcal{E}_h(f, g) = \langle X_h(f), W_h(g) \rangle$.

From Feasibility to a Bound on $W_h(f_t)$, Part 1

Goal. Translate feasibility (an *empirical* condition on $\widehat{\mathcal{E}}$) into a bound on the *signature* $W_h(f_t) \in \mathbb{R}^d$ given by the Bellman-rank factorization $\mathcal{E}_h(f, g) = \langle X_h(f), W_h(g) \rangle$.

Step 1: population bound from empirical. Feasibility: $\sum_{i<t} \widehat{\mathcal{E}}_{i,h}(f_t)^2 \leq R^2 = T\varepsilon_{\mathrm{gen}}^2$. With $|\widehat{\mathcal{E}}_{i,h}(f_t) - \mathcal{E}_h(f_i, f_t)| \leq \varepsilon_{\mathrm{gen}}$ and $(a+b)^2 \leq 2a^2 + 2b^2$:

$$\sum_{i<t} \mathcal{E}_h(f_i, f_t)^2 \;\leq\; 2R^2 + 2T\varepsilon_{\mathrm{gen}}^2 \;=\; 4T\varepsilon_{\mathrm{gen}}^2.$$

From Feasibility to a Bound on $W_h(f_t)$, Part 1

Goal. Translate feasibility (an *empirical* condition on $\widehat{\mathcal{E}}$) into a bound on the *signature* $W_h(f_t) \in \mathbb{R}^d$ given by the Bellman-rank factorization $\mathcal{E}_h(f, g) = \langle X_h(f), W_h(g) \rangle$.

Step 1: population bound from empirical. Feasibility: $\sum_{i<t} \widehat{\mathcal{E}}_{i,h}(f_t)^2 \le R^2 = T\varepsilon_{\mathrm{gen}}^2$. With $|\widehat{\mathcal{E}}_{i,h}(f_t) - \mathcal{E}_h(f_i, f_t)| \le \varepsilon_{\mathrm{gen}}$ and $(a+b)^2 \le 2a^2 + 2b^2$:

$$\sum_{i<t} \mathcal{E}_h(f_i, f_t)^2 \;\le\; 2R^2 + 2T\varepsilon_{\mathrm{gen}}^2 \;=\; 4T\varepsilon_{\mathrm{gen}}^2.$$

Step 2: rewrite as a quadratic form in $W_h(f_t)$. Substituting $\mathcal{E}_h(f_i, f_t) = \langle X_h(f_i), W_h(f_t) \rangle$:

$$\sum_{i<t} \langle X_h(f_i),\, W_h(f_t) \rangle^2 \;=\; W_h(f_t)^\top \Big(\sum_{i<t} X_h(f_i) X_h(f_i)^\top \Big) W_h(f_t) \;=\; \|W_h(f_t)\|_{\Sigma_{t,h}}^2 ,$$

where $\Sigma_{t,h} := \sum_{i<t} X_h(f_i) X_h(f_i)^\top \in \mathbb{R}^{d \times d}$ is the Gram matrix of past X-signatures.

From Feasibility to a Bound on $W_h(f_t)$, Part 1

Goal. Translate feasibility (an *empirical* condition on $\widehat{\mathcal{E}}$) into a bound on the *signature* $W_h(f_t) \in \mathbb{R}^d$ given by the Bellman-rank factorization $\mathcal{E}_h(f, g) = \langle X_h(f), W_h(g) \rangle$.

Step 1: population bound from empirical. Feasibility: $\sum_{i<t} \widehat{\mathcal{E}}_{i,h}(f_t)^2 \le R^2 = T\varepsilon_{\mathrm{gen}}^2$. With $|\widehat{\mathcal{E}}_{i,h}(f_t) - \mathcal{E}_h(f_i, f_t)| \le \varepsilon_{\mathrm{gen}}$ and $(a+b)^2 \le 2a^2 + 2b^2$:

$$\sum_{i<t} \mathcal{E}_h(f_i, f_t)^2 \;\le\; 2R^2 + 2T\varepsilon_{\mathrm{gen}}^2 \;=\; 4T\varepsilon_{\mathrm{gen}}^2.$$

Step 2: rewrite as a quadratic form in $W_h(f_t)$. Substituting $\mathcal{E}_h(f_i, f_t) = \langle X_h(f_i), W_h(f_t) \rangle$:

$$\sum_{i<t} \langle X_h(f_i),\, W_h(f_t) \rangle^2 \;=\; W_h(f_t)^\top \Big(\sum_{i<t} X_h(f_i) X_h(f_i)^\top \Big) W_h(f_t) \;=\; \|W_h(f_t)\|_{\Sigma_{t,h}}^2 ,$$

where $\Sigma_{t,h} := \sum_{i<t} X_h(f_i) X_h(f_i)^\top \in \mathbb{R}^{d \times d}$ is the Gram matrix of past X-signatures.

Combining Steps 1 & 2:

$$\boxed{\|W_h(f_t)\|_{\Sigma_{t,h}}^2 \;\le\; 4T\varepsilon_{\mathrm{gen}}^2.}$$

In English: $W_h(f_t)$ has small "$\Sigma_{t,h}$-norm," i.e., is small *in every direction we've already probed* via past roll-ins.

From Feasibility to a Bound on $W_h(f_t)$, Part 2

The trouble. $\Sigma_{t,h} = \sum_{i<t} X_h(f_i) X_h(f_i)^\top$ has rank at most t, which is less than d in early rounds. So $\Sigma_{t,h}$ is generally *singular* — $\Sigma_{t,h}^{-1}$ doesn't exist, and the $\Sigma_{t,h}$-norm is only a *seminorm* (it's zero on directions not yet explored).

The fix: regularize. For some $\lambda > 0$ to be chosen later, define

$$\Lambda_{t,h} \;:=\; \lambda I + \Sigma_{t,h}.$$

Now $\Lambda_{t,h} \succeq \lambda I \succ 0$ is invertible. And $\|x\|^2_{\Lambda_{t,h}} = \lambda \|x\|_2^2 + \|x\|^2_{\Sigma_{t,h}}$.

From Feasibility to a Bound on $W_h(f_t)$, Part 2

The trouble. $\Sigma_{t,h} = \sum_{i<t} X_h(f_i) X_h(f_i)^\top$ has rank at most t, which is less than d in early rounds. So $\Sigma_{t,h}$ is generally *singular* — $\Sigma_{t,h}^{-1}$ doesn't exist, and the $\Sigma_{t,h}$-norm is only a *seminorm* (it's zero on directions not yet explored).

The fix: regularize. For some $\lambda > 0$ to be chosen later, define

$$\Lambda_{t,h} \;:=\; \lambda I + \Sigma_{t,h}.$$

Now $\Lambda_{t,h} \succeq \lambda I \succ 0$ is invertible. And $\|x\|^2_{\Lambda_{t,h}} = \lambda \|x\|_2^2 + \|x\|^2_{\Sigma_{t,h}}$.

Step 3: bound on $W_h(f_t)$. Lift Step 2 to $\Lambda_{t,h}$:

$$\|W_h(f_t)\|^2_{\Lambda_{t,h}} \;=\; \lambda \|W_h(f_t)\|_2^2 + \|W_h(f_t)\|^2_{\Sigma_{t,h}} \;\le\; \lambda B_W^2 + 4T\varepsilon^2_{\mathrm{gen}}.$$

(Used the Bellman-rank bound $\|W_h(g)\|_2 \le B_W$.)

From Feasibility to a Bound on $W_h(f_t)$, Part 2

The trouble. $\Sigma_{t,h} = \sum_{i<t} X_h(f_i) X_h(f_i)^\top$ has rank at most t, which is less than d in early rounds. So $\Sigma_{t,h}$ is generally *singular* — $\Sigma_{t,h}^{-1}$ doesn't exist, and the $\Sigma_{t,h}$-norm is only a *seminorm* (it's zero on directions not yet explored).

The fix: regularize. For some $\lambda > 0$ to be chosen later, define

$$\Lambda_{t,h} \;:=\; \lambda I + \Sigma_{t,h}.$$

Now $\Lambda_{t,h} \succeq \lambda I \succ 0$ is invertible. And $\|x\|_{\Lambda_{t,h}}^2 = \lambda \|x\|_2^2 + \|x\|_{\Sigma_{t,h}}^2$.

Step 3: bound on $W_h(f_t)$. Lift Step 2 to $\Lambda_{t,h}$:

$$\|W_h(f_t)\|_{\Lambda_{t,h}}^2 \;=\; \lambda \|W_h(f_t)\|_2^2 + \|W_h(f_t)\|_{\Sigma_{t,h}}^2 \;\le\; \lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2.$$

(Used the Bellman-rank bound $\|W_h(g)\|_2 \le B_W$.)

Summary of this slide

$$\|W_h(f_t)\|_{\Lambda_{t,h}}^2 \;\le\; \lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2.$$

The λB_W^2 is the regularization cost. λ is still a free parameter; we'll pick it at the end.

Self-Error via Cauchy–Schwarz

Step 4: self-error via Cauchy–Schwarz. Plug in $\mathcal{E}_h(f_t, f_t) = \langle X_h(f_t), W_h(f_t) \rangle$:

$$|\mathcal{E}_h(f_t, f_t)| \;\le\; \|X_h(f_t)\|_{\Lambda_{t,h}^{-1}} \, \|W_h(f_t)\|_{\Lambda_{t,h}} .$$

By Step 3 ($\|W_h(f_t)\|_{\Lambda_{t,h}}^2 \le \lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2$):

$$\boxed{\sum_h |\mathcal{E}_h(f_t, f_t)| \;\le\; \sqrt{\lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2} \sum_h \|X_h(f_t)\|_{\Lambda_{t,h}^{-1}} .}$$

Goal: bound $\sum_h \|X_h(f_t)\|_{\Lambda_{t,h}^{-1}}$ *for some round* t. Strategy: determinant lemma + pigeonhole.

Step 5: Determinant Lemma

One-step update. By the matrix determinant lemma,

$$\det(\Lambda_{t+1,h}) \;=\; \det(\Lambda_{t,h})\left(1 + \|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\right).$$

Telescope: $\sum_{t=0}^{T-1} \log\left(1 + \|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\right) \;=\; \log \frac{\det(\Lambda_{T,h})}{\det(\lambda I)}$.

Step 5: Determinant Lemma

One-step update. By the matrix determinant lemma,

$$\det(\Lambda_{t+1,h}) \;=\; \det(\Lambda_{t,h})\,\big(1 + \|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\big).$$

Telescope: $\sum_{t=0}^{T-1} \log\big(1 + \|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\big) \;=\; \log \frac{\det(\Lambda_{T,h})}{\det(\lambda I)}$.

Bound the log-determinant. $\mathrm{tr}(\Lambda_{T,h}) \le \lambda d + TB_X^2$, and AM–GM for PSD matrices gives $\det(M) \le (\mathrm{tr}(M)/d)^d$. So

$$\log \frac{\det(\Lambda_{T,h})}{\det(\lambda I)} \;\le\; d \log\Big(1 + \frac{TB_X^2}{d\lambda}\Big).$$

Step 5: Determinant Lemma

One-step update. By the matrix determinant lemma,

$$\det(\Lambda_{t+1,h}) \;=\; \det(\Lambda_{t,h})\,\big(1 + \|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\big).$$

Telescope: $\sum_{t=0}^{T-1} \log\big(1 + \|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\big) \;=\; \log \frac{\det(\Lambda_{T,h})}{\det(\lambda I)}$.

Bound the log-determinant. $\mathrm{tr}(\Lambda_{T,h}) \le \lambda d + TB_X^2$, and AM–GM for PSD matrices gives $\det(M) \le (\mathrm{tr}(M)/d)^d$. So

$$\log \frac{\det(\Lambda_{T,h})}{\det(\lambda I)} \;\le\; d \log\Big(1 + \frac{TB_X^2}{d\lambda}\Big).$$

Step 5 conclusion

For every h: $\sum_{t=0}^{T-1} \log\big(1 + \|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\big) \;\le\; d \log(1 + TB_X^2/(d\lambda))$.

Note. This is the same identity that drives Lecture 13's elliptical potential lemma. Here we keep it in the $\log(1+\cdot)$ form rather than passing to $\min\{1, \cdot\}$.

Step 6: Pigeonhole + Exp/Log

Sum over h, average over t. From Step 5:

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\log\left(1+\|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\right) \;\leq\; \frac{Hd}{T}\log\left(1+\frac{TB_X^2}{d\lambda}\right).$$

By pigeonhole, there exists $t^\star \in \{0,\ldots,T-1\}$ achieving (at most) the average:

$$\sum_{h}\log\left(1+\|X_h(f_{t^\star})\|^2_{\Lambda_{t^\star,h}^{-1}}\right) \;\leq\; \frac{Hd}{T}\log\left(1+\frac{TB_X^2}{d\lambda}\right).$$

Step 6: Pigeonhole + Exp/Log

Sum over h, average over t. From Step 5:

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\log\big(1+\|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\big) \;\le\; \frac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big).$$

By pigeonhole, there exists $t^\star \in \{0,\dots,T-1\}$ achieving (at most) the average:

$$\sum_{h}\log\big(1+\|X_h(f_{t^\star})\|^2_{\Lambda_{t^\star,h}^{-1}}\big) \;\le\; \frac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big).$$

Each summand is non-negative. Hence for every h individually:

$$\log\big(1+\|X_h(f_{t^\star})\|^2_{\Lambda_{t^\star,h}^{-1}}\big) \;\le\; \frac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big).$$

Exponentiating:

$$\boxed{\|X_h(f_{t^\star})\|^2_{\Lambda_{t^\star,h}^{-1}} \;\le\; \exp\Big(\tfrac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big)\Big)-1.}$$

Step 6: Pigeonhole + Exp/Log

Sum over h, average over t. From Step 5:

$$\frac{1}{T}\sum_{t=0}^{T-1}\sum_{h=0}^{H-1}\log\big(1+\|X_h(f_t)\|^2_{\Lambda_{t,h}^{-1}}\big) \;\le\; \frac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big).$$

By pigeonhole, there exists $t^\star \in \{0,\dots,T-1\}$ achieving (at most) the average:

$$\sum_{h}\log\big(1+\|X_h(f_{t^\star})\|^2_{\Lambda_{t^\star,h}^{-1}}\big) \;\le\; \frac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big).$$

Each summand is non-negative. Hence for every h individually:

$$\log\big(1+\|X_h(f_{t^\star})\|^2_{\Lambda_{t^\star,h}^{-1}}\big) \;\le\; \frac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big).$$

Exponentiating:

$$\boxed{\|X_h(f_{t^\star})\|^2_{\Lambda_{t^\star,h}^{-1}} \;\le\; \exp\Big(\tfrac{Hd}{T}\log\big(1+\tfrac{TB_X^2}{d\lambda}\big)\Big)-1.}$$

Interpretation. At round $t^\star$, *simultaneously* every stage h has small elliptical norm — $X_h(f_{t^\star})$ lies in the span of past directions.

Step 7: Plug In and Choose Parameters

Combining Step 4 with Step 6 at $t = t^\star$, applied to each h and summed:

$$\sum_h |\mathcal{E}_h(f_{t^\star}, f_{t^\star})| \;\le\; H\sqrt{\lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2}\,\sqrt{\exp\Big(\tfrac{Hd}{T}\log(1+\tfrac{TB_X^2}{d\lambda})\Big) - 1}.$$

Step 7: Plug In and Choose Parameters

Combining Step 4 with Step 6 at $t = t^\star$, applied to each h and summed:

$$\sum_h |\mathcal{E}_h(f_{t^\star}, f_{t^\star})| \;\le\; H\sqrt{\lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2}\,\sqrt{\exp\Big(\tfrac{Hd}{T}\log(1 + \tfrac{TB_X^2}{d\lambda})\Big) - 1}.$$

Choose parameters (AJKS Thm 8.6). Let $L := \log(eHd(B_X^2 B_W^2/\varepsilon_{\mathrm{gen}}^2 + 1))$, $\lambda := B_X^2/(B_X^2 B_W^2/\varepsilon_{\mathrm{gen}}^2 + 1)$, $T := \lceil c\, Hd\, L \rceil$. Then $\lambda B_W^2 \le \varepsilon_{\mathrm{gen}}^2$ and $\frac{Hd}{T}\log(1 + TB_X^2/(d\lambda)) \le 1$.

Using $e^x - 1 \le 2x$ for $x \in [0, 1]$ and $\sqrt{\lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2} \le 3\sqrt{T}\varepsilon_{\mathrm{gen}}$:

$$\sum_h |\mathcal{E}_h(f_{t^\star}, f_{t^\star})| \;\le\; H \cdot 3\sqrt{T}\varepsilon_{\mathrm{gen}} \cdot \sqrt{2HdL/T} \;=\; O\big(\varepsilon_{\mathrm{gen}}\sqrt{dH^3 L}\big).$$

Step 7: Plug In and Choose Parameters

Combining Step 4 with Step 6 at $t = t^\star$, applied to each h and summed:

$$\sum_h |\mathcal{E}_h(f_{t^\star}, f_{t^\star})| \;\le\; H\sqrt{\lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2}\,\sqrt{\exp\Big(\tfrac{Hd}{T}\log(1+\tfrac{TB_X^2}{d\lambda})\Big) - 1}.$$

Choose parameters (AJKS Thm 8.6). Let $L := \log(eHd(B_X^2 B_W^2/\varepsilon_{\mathrm{gen}}^2 + 1))$, $\lambda := B_X^2/(B_X^2 B_W^2/\varepsilon_{\mathrm{gen}}^2 + 1)$, $T := \lceil c\, Hd\, L \rceil$. Then $\lambda B_W^2 \le \varepsilon_{\mathrm{gen}}^2$ and $\frac{Hd}{T}\log(1 + TB_X^2/(d\lambda)) \le 1$.

Using $e^x - 1 \le 2x$ for $x \in [0,1]$ and $\sqrt{\lambda B_W^2 + 4T\varepsilon_{\mathrm{gen}}^2} \le 3\sqrt{T}\varepsilon_{\mathrm{gen}}$:

$$\sum_h |\mathcal{E}_h(f_{t^\star}, f_{t^\star})| \;\le\; H \cdot 3\sqrt{T}\varepsilon_{\mathrm{gen}} \cdot \sqrt{2HdL/T} \;=\; O\big(\varepsilon_{\mathrm{gen}}\sqrt{dH^3 L}\big).$$

Per-round bound (AJKS Theorem 8.6)

$$V_0^\star - V_0^{\pi_{f_{t^\star}}} \le C\,\varepsilon_{\mathrm{gen}}\sqrt{dH^3 \log(eHd(B_X^2 B_W^2/\varepsilon_{\mathrm{gen}}^2 + 1))}.$$

The algorithm picks $\hat{t}$ minimizing empirical self-error; selection costs an extra $O(H\varepsilon_{\mathrm{gen}})$, absorbed.

Main Theorem

Theorem 8.6 (PAC RL under bounded Bellman rank)

Assume realizability and bounded Bellman rank d. For finite $\mathcal{H}$ with values in $[0, H]$, the optimistic-elimination algorithm with $T = \widetilde{O}(Hd)$ rounds and batch size $m = \widetilde{O}(dH^5 \log|\mathcal{H}|/\varepsilon^2)$ outputs $\widehat{\pi}$ with $V_0^\star - V_0^{\widehat{\pi}} \leq \varepsilon$ with probability $\geq 1 - \delta$, using

$$m \cdot T \;=\; \widetilde{O}\left(\frac{d^2 H^6 \log|\mathcal{H}|}{\varepsilon^2}\right) \text{ trajectories (Q-mode).}$$

V-mode: extra factor of $|\mathcal{A}|$ from importance weighting.

Main Theorem

Theorem 8.6 (PAC RL under bounded Bellman rank)

Assume realizability and bounded Bellman rank d. For finite $\mathcal{H}$ with values in $[0, H]$, the optimistic-elimination algorithm with $T = \widetilde{O}(Hd)$ rounds and batch size $m = \widetilde{O}(dH^5 \log|\mathcal{H}|/\varepsilon^2)$ outputs $\widehat{\pi}$ with $V_0^\star - V_0^{\widehat{\pi}} \le \varepsilon$ with probability $\ge 1-\delta$, using

$$m \cdot T \;=\; \widetilde{O}\left(\frac{d^2 H^6 \log|\mathcal{H}|}{\varepsilon^2}\right) \text{ trajectories (Q-mode).}$$

V-mode: extra factor of $|\mathcal{A}|$ from importance weighting.

What you pay for.

- H^6: from $\varepsilon_{\text{gen}}^2 \propto H^2/m$ combined with the per-round $\sqrt{dH^3}$ in Step 7.
- $\log|\mathcal{H}|$: uniform convergence over hypotheses — no $|\mathcal{S}|$ anywhere.
- Independent of $|\mathcal{S}|$ entirely. $|\mathcal{A}|$ enters only in V-mode.

Putting the Pieces Together

The three-line summary of the algorithm:

1. Optimism: $V_{0,f_t} \geq V_0^\star$ because $f^\star$ is feasible.
2. Self-error bound: $V_0^\star - V_0^{\pi_{f_t}} \leq \sum_h \mathcal{E}_h(f_t, f_t)$ (telescoping identity).
3. Rank-d exploration: self-error must be small for some $t \in \{0, \ldots, T-1\}$ with $T = \widetilde{O}(dH)$ (elliptical potential on X-signatures).

Putting the Pieces Together

The three-line summary of the algorithm:

1. Optimism: $V_{0,f_t} \geq V_0^\star$ because $f^\star$ is feasible.
2. Self-error bound: $V_0^\star - V_0^{\pi_{f_t}} \leq \sum_h \mathcal{E}_h(f_t, f_t)$ (telescoping identity).
3. Rank-d exploration: self-error must be small for some $t \in \{0, \ldots, T-1\}$ with $T = \widetilde{O}(dH)$ (elliptical potential on X-signatures).

The matrix perspective, restated

The algorithm probes the Bellman-error matrix one row at a time (rolling in with π_{f_t} reveals the f_t-th row). Since the matrix has rank d, after $\widetilde{O}(d)$ rows we've "seen" enough to determine W_h on the subspace where it matters.

The optimism principle steers us into rows that are *actually informative* — not just any d rolls of $\mathcal{H}$, but rolls whose X-signatures span new directions where the optimistic hypothesis could disagree with $f^\star$.

Today's Plan

What Bellman Rank Unifies

Single algorithm, single analysis gives PAC bounds for:

- Tabular MDPs ($d \leq SA$).
- Linear MDPs ($d =$ feature dim).
- Linear Bellman completeness.
- Low-rank transition models.
- Sparse-feature linear MDPs (representation learning): d governed by latent dimension; only $\log |\Phi|$ extra cost for feature selection.
- Various "contextual" RL settings.

What Bellman Rank Unifies

Single algorithm, single analysis gives PAC bounds for:

- Tabular MDPs ($d \leq SA$).
- Linear MDPs ($d =$ feature dim).
- Linear Bellman completeness.
- Low-rank transition models.
- Sparse-feature linear MDPs (representation learning): d governed by latent dimension; only $\log |\Phi|$ extra cost for feature selection.
- Various "contextual" RL settings.

Important caveat: computation

The optimistic-elimination algorithm assumes an *optimization oracle*: at each step we must find the most optimistic hypothesis subject to many quadratic constraints. This is generally NP-hard.

The Bellman-rank framework gives *statistical* efficiency. Making it computationally efficient typically requires extra structure (e.g. the linear MDP case, where the oracle is tractable via LSVI-UCB).

Beyond Bellman Rank

Generalizations. The Bellman-rank framework has been extended in several directions:

- **Bellman–Eluder dimension** (Jin, Liu, Miryoosefi 2021): combines Bellman rank with eluder dimension; captures non-linear function classes.
- **Bilinear classes** (Du, Kakade, Lee, Lovett, Mahajan, Sun, Wang 2021): a yet more general low-rank factorization viewpoint.
- **Decision-Estimation Coefficient** (Foster, Kakade, Qian, Rakhlin 2021): an information-theoretic complexity measure that subsumes the above.

Beyond Bellman Rank

Generalizations. The Bellman-rank framework has been extended in several directions:

- **Bellman–Eluder dimension** (Jin, Liu, Miryoosefi 2021): combines Bellman rank with eluder dimension; captures non-linear function classes.
- **Bilinear classes** (Du, Kakade, Lee, Lovett, Mahajan, Sun, Wang 2021): a yet more general low-rank factorization viewpoint.
- **Decision-Estimation Coefficient** (Foster, Kakade, Qian, Rakhlin 2021): an information-theoretic complexity measure that subsumes the above.

Take-aways from this lecture

- The *Bellman error matrix* $[\mathcal{E}_h(f, g)]$ is the right object: rows indexed by roll-in policies, columns by candidate hypotheses.
- Realizability puts a zero column at $f^\star$.
- Bounded Bellman rank = this matrix is low-rank, i.e., few independent "search directions."
- The algorithm uses optimism to pick informative rows; elliptical potential shows $\widetilde{O}(d)$ rows suffice.

Reading & References

Primary reference:

AJKS Agarwal, Jiang, Kakade, Sun, *Reinforcement Learning: Theory and Algorithms*, Chapter 8.

Foundational paper:

- Jiang, Krishnamurthy, Agarwal, Langford, Schapire, *Contextual decision processes with low Bellman rank are PAC-learnable*, ICML 2017. (**OLIVE algorithm**)

Generalizations:

- Jin, Liu, Miryoosefi, *Bellman Eluder dimension: New rich classes of RL problems, and sample-efficient algorithms*, NeurIPS 2021.
- Du, Kakade, Lee, Lovett, Mahajan, Sun, Wang, *Bilinear classes: A structural framework for provable generalization in RL*, ICML 2021.
- Foster, Kakade, Qian, Rakhlin, *The statistical complexity of interactive decision making*, 2021.

Questions?