CSE 542: Statistical Reinforcement Learning

Lecture 13: Linear MDPs and Online RL

Kevin Jamieson

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

Outline

Where We Left Off

Linear function approximation: feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, $\|\phi\|_2 \leq 1$, class $\mathcal{F} = \{\langle \phi, w \rangle : w \in \mathbb{R}^d\}$.

Two model-free assumptions:

(R) $Q^\star \in \mathcal{F}$ (linear realizability).

(C) $\mathcal{T}_h \mathcal{F} \subseteq \mathcal{F}$ (Bellman completeness): the algorithmic assumption.

What we proved:

- Generative model + (C) $\Rightarrow \widetilde{O}(d^2 H^5/\varepsilon^2)$ samples via LSVI with G-optimal design.
- (R) alone $\Rightarrow \Omega(2^{\min(d,H)})$ lower bound.
- Offline + (C) $\Rightarrow$ vanilla LSVI under all-policy concentrability. Pessimism (PEVI) would buy single-policy concentrability but already needs more than (C) — the same obstruction we're about to face online.

Where We Left Off

Linear function approximation: feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, $\|\phi\|_2 \leq 1$, class $\mathcal{F} = \{\langle \phi, w \rangle : w \in \mathbb{R}^d\}$.

Two model-free assumptions:

(R) $Q^\star \in \mathcal{F}$ (linear realizability).

(C) $\mathcal{T}_h \mathcal{F} \subseteq \mathcal{F}$ (Bellman completeness): the algorithmic assumption.

What we proved:

- Generative model + (C) $\Rightarrow \widetilde{O}(d^2 H^5/\varepsilon^2)$ samples via LSVI with G-optimal design.
- (R) alone $\Rightarrow \Omega(2^{\min(d,H)})$ lower bound.
- Offline + (C) $\Rightarrow$ vanilla LSVI under all-policy concentrability. Pessimism (PEVI) would buy single-policy concentrability but already needs more than (C) — the same obstruction we're about to face online.

Today: online RL

No generative model. Must explore via on-policy trajectories. We'll see that (C) is *not* enough — need a stronger model-based assumption.

Today's Plan

What Changes Without a Generative Model

Generative model

- Pick *any* (s, a) to query.
- G-optimal design covers feature space.
- Estimation only, no exploration.

Online RL

- Roll out a policy from $s_1 \sim \nu$.
- Visited (s, a) depend on past behavior.
- Must explore actively.

What Changes Without a Generative Model

Generative model

- Pick *any* (s, a) to query.
- G-optimal design covers feature space.
- Estimation only, no exploration.

Online RL

- Roll out a policy from $s_1 \sim \nu$.
- Visited (s, a) depend on past behavior.
- Must explore actively.

Tabular UCB-VI (Lecture 9): added a count-based bonus $H\sqrt{\log(\cdot)/n_h^k(s,a)}$ to the empirical Bellman backup.

What Changes Without a Generative Model

Generative model

- Pick *any* (s, a) to query.
- G-optimal design covers feature space.
- Estimation only, no exploration.

Online RL

- Roll out a policy from $s_1 \sim \nu$.
- Visited (s, a) depend on past behavior.
- Must explore actively.

Tabular UCB-VI (Lecture 9): added a count-based bonus $H\sqrt{\log(\cdot)/n_h^k(s,a)}$ to the empirical Bellman backup.

Natural attempt for linear FA: compute $\widehat{w}_h^k$ by ridge regression on past trajectory data, then use

$$\widehat{Q}_h^k(s,a) \;=\; \left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + \beta \left\| \phi(s,a) \right\|_{(\Lambda_h^k)^{-1}}.$$

What Changes Without a Generative Model

Generative model

- Pick *any* (s, a) to query.
- G-optimal design covers feature space.
- Estimation only, no exploration.

Online RL

- Roll out a policy from $s_1 \sim \nu$.
- Visited (s, a) depend on past behavior.
- Must explore actively.

Tabular UCB-VI (Lecture 9): added a count-based bonus $H\sqrt{\log(\cdot)/n_h^k(s,a)}$ to the empirical Bellman backup.

Natural attempt for linear FA: compute $\widehat{w}_h^k$ by ridge regression on past trajectory data, then use

$$\widehat{Q}_h^k(s,a) \;=\; \left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + \beta \left\| \phi(s,a) \right\|_{(\Lambda_h^k)^{-1}}.$$

Question: is Bellman completeness enough to make this work?

What Optimism Demands

For the optimism argument from Lecture 9 to go through, we need: for any value function V that the algorithm might compute,

$$\left|((P_h - \widehat{P}_h^k)V)(s,a)\right| \lesssim \beta \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

This is the linear analog of the tabular concentration $|((P_h - \widehat{P}_h^k)V)| \leq H\sqrt{\log/n}$.

What Optimism Demands

For the optimism argument from Lecture 9 to go through, we need: for any value function V that the algorithm might compute,

$$\left|((P_h - \widehat{P}_h^k)V)(s,a)\right| \lesssim \beta \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

This is the linear analog of the tabular concentration $|((P_h - \widehat{P}_h^k)V)| \leq H\sqrt{\log/n}$.

Crucial requirement. The function $(s,a) \mapsto (P_h V)(s,a)$ must lie in $\mathrm{span}(\phi)$ for the value functions V that LSVI-UCB actually produces.

What Optimism Demands

For the optimism argument from Lecture 9 to go through, we need: for any value function V that the algorithm might compute,

$$\left|((P_h - \widehat{P}_h^k)V)(s,a)\right| \lesssim \beta \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}.$$

This is the linear analog of the tabular concentration $|((P_h - \widehat{P}_h^k)V)| \leq H\sqrt{\log/n}$.

Crucial requirement. The function $(s,a) \mapsto (P_h V)(s,a)$ must lie in $\mathrm{span}(\phi)$ for the value functions V that LSVI-UCB actually produces.

What does LSVI-UCB produce?

$$\widehat{V}_h^k(s) = \max_a \left[\left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + \beta \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \right].$$

The max over a and the bonus mean $\widehat{V}_h^k$ is not an element of $\mathcal{F}$.

Bellman Completeness Doesn't Suffice for Online Optimism

What (C) gives

Closure of $\mathcal{F}$ under $\mathcal{T}_h$: $f \in \mathcal{F} \implies \mathcal{T}_h f \in \mathcal{F}$.

Bellman Completeness Doesn't Suffice for Online Optimism

What (C) gives

Closure of $\mathcal{F}$ under $\mathcal{T}_h$: $f \in \mathcal{F} \implies \mathcal{T}_h f \in \mathcal{F}$.

Problem. $\widehat{V}_h^k \notin \mathcal{F}$ (it has a max and a bonus term). So (C) tells us *nothing* about $(P_h \widehat{V}_h^k)(s, a)$.

Bellman Completeness Doesn't Suffice for Online Optimism

What (C) gives

Closure of $\mathcal{F}$ under $\mathcal{T}_h$: $f \in \mathcal{F} \implies \mathcal{T}_h f \in \mathcal{F}$.

Problem. $\widehat{V}_h^k \notin \mathcal{F}$ (it has a max and a bonus term). So (C) tells us *nothing* about $(P_h \widehat{V}_h^k)(s, a)$.

What we actually need

For *every bounded* $V : \mathcal{S} \to [0, H]$ — not just $V \in \mathcal{F}$ — the function $(s, a) \mapsto (P_h V)(s, a)$ should be linear in $\phi(s, a)$.

Bellman completeness only controls the linear class; we need control over a much larger class of V's.

Bellman Completeness Doesn't Suffice for Online Optimism

What (C) gives

Closure of $\mathcal{F}$ under $\mathcal{T}_h$: $f \in \mathcal{F} \implies \mathcal{T}_h f \in \mathcal{F}$.

Problem. $\widehat{V}_h^k \notin \mathcal{F}$ (it has a max and a bonus term). So (C) tells us *nothing* about $(P_h \widehat{V}_h^k)(s, a)$.

What we actually need

For *every bounded* $V : \mathcal{S} \to [0, H]$ — not just $V \in \mathcal{F}$ — the function $(s, a) \mapsto (P_h V)(s, a)$ should be linear in $\phi(s, a)$.

Bellman completeness only controls the linear class; we need control over a much larger class of V's.

This is exactly what Linear MDPs guarantee.

Today's Plan

Definition: Linear MDP (Jin, Yang, Wang, Jordan 2020)

Linear MDP assumption (L)

An MDP is *linear* with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ (with $\|\phi(s,a)\|_2 \leq 1$) if for each $h \in [H]$ there exist

- an unknown signed measure $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ on $\mathcal{S}$,
- an unknown vector $\theta_h \in \mathbb{R}^d$,

such that

$$P_h(s' \mid s,a) = \langle \phi(s,a), \mu_h(s') \rangle, \qquad r_h(s,a) = \langle \phi(s,a), \theta_h \rangle .$$

Normalization: $\|\mu_h(\mathcal{S})\|_2 \leq \sqrt{d}$ and $\|\theta_h\|_2 \leq \sqrt{d}$.

Definition: Linear MDP (Jin, Yang, Wang, Jordan 2020)

Linear MDP assumption (L)

An MDP is *linear* with feature map $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$ (with $\|\phi(s,a)\|_2 \leq 1$) if for each $h \in [H]$ there exist

- an unknown signed measure $\mu_h = (\mu_h^{(1)}, \ldots, \mu_h^{(d)})$ on $\mathcal{S}$,
- an unknown vector $\theta_h \in \mathbb{R}^d$,

such that

$$P_h(s' \mid s,a) = \langle \phi(s,a), \mu_h(s') \rangle, \qquad r_h(s,a) = \langle \phi(s,a), \theta_h \rangle .$$

Normalization: $\|\mu_h(\mathcal{S})\|_2 \leq \sqrt{d}$ and $\|\theta_h\|_2 \leq \sqrt{d}$.

This is a model-based assumption: we commit to a specific form of P_h and r_h.

Why $\sqrt{d}$ Is the "Natural" Scaling

The bounds $\|\mu_h(\mathcal{S})\|_2 \leq \sqrt{d}$ and $\|\theta_h\|_2 \leq \sqrt{d}$ may look arbitrary. They are calibrated so that the tabular case sits exactly at the boundary.

Tabular case. $\phi(s,a) = e_{(s,a)} \in \mathbb{R}^d$ with $d = SA$.

- Signed measures: $\mu_h^{(s,a)}(s') = P_h(s'|s,a) \in [0,1]$, so $\mu_h^{(s,a)}(\mathcal{S}) = 1$ for each component.

$$\|\mu_h(\mathcal{S})\|_2 = \sqrt{\sum_{(s,a)} 1^2} = \sqrt{SA} = \sqrt{d}.$$

- Reward weights: $\theta_h^{(s,a)} = r_h(s,a) \in [0,1]$, so

$$\|\theta_h\|_2 \leq \sqrt{SA} = \sqrt{d}.$$

Why $\sqrt{d}$ Is the "Natural" Scaling

The bounds $\|\mu_h(\mathcal{S})\|_2 \le \sqrt{d}$ and $\|\theta_h\|_2 \le \sqrt{d}$ may look arbitrary. They are calibrated so that the tabular case sits exactly at the boundary.

Tabular case. $\phi(s,a) = e_{(s,a)} \in \mathbb{R}^d$ with $d = SA$.

- Signed measures: $\mu_h^{(s,a)}(s') = P_h(s'|s,a) \in [0,1]$, so $\mu_h^{(s,a)}(\mathcal{S}) = 1$ for each component.

$$\|\mu_h(\mathcal{S})\|_2 \;=\; \sqrt{\textstyle\sum_{(s,a)} 1^2} \;=\; \sqrt{SA} \;=\; \sqrt{d}.$$

- Reward weights: $\theta_h^{(s,a)} = r_h(s,a) \in [0,1]$, so

$$\|\theta_h\|_2 \;\le\; \sqrt{SA} \;=\; \sqrt{d}.$$

Interpretation

Each "feature direction" $j \in [d]$ contributes $O(1)$ to the relevant operator norms; with d such directions the ℓ_2 aggregate scales as $\sqrt{d}$. This is what allows the bounds $\|w_h^\pi\|_2 \le 2H\sqrt{d}$ that drive all subsequent regret analyses.

What (L) Buys: Universal Linearity of $P_h V$

Lemma (Universal linearity)

Under (L), for *every* bounded function $V : \mathcal{S} \to \mathbb{R}$,

$$(P_h V)(s, a) = \left\langle \phi(s, a), \int V(s')\, \mu_h(s')\, ds' \right\rangle.$$

That is, P_h maps any V into $\mathrm{span}(\phi)$ — not just functions in $\mathcal{F}$.

What (L) Buys: Universal Linearity of $P_h V$

Lemma (Universal linearity)

Under (L), for *every* bounded function $V : \mathcal{S} \to \mathbb{R}$,

$$(P_h V)(s,a) \;=\; \Big\langle \phi(s,a),\; \int V(s')\,\mu_h(s')\,ds' \Big\rangle.$$

That is, P_h maps any V into $\mathrm{span}(\phi)$ — not just functions in $\mathcal{F}$.

Proof. $(P_h V)(s,a) = \int V(s')\,\langle \phi(s,a), \mu_h(s')\rangle\, ds' = \left\langle \phi(s,a),\, \int V \mu_h\, ds' \right\rangle.$ □

What (L) Buys: Universal Linearity of $P_h V$

Lemma (Universal linearity)

Under (L), for *every* bounded function $V : \mathcal{S} \to \mathbb{R}$,

$$(P_h V)(s,a) \;=\; \Big\langle \phi(s,a),\; \int V(s')\,\mu_h(s')\,ds' \Big\rangle.$$

That is, P_h maps any V into $\mathrm{span}(\phi)$ — not just functions in $\mathcal{F}$.

Proof. $(P_h V)(s,a) = \int V(s')\,\langle \phi(s,a), \mu_h(s')\rangle\, ds' = \left\langle \phi(s,a),\, \int V \mu_h\, ds' \right\rangle.$ □

Exactly what online optimism needed

For any $\widehat{V}_h^k$ produced by LSVI-UCB (with bonus, max, clipping), $(P_h \widehat{V}_h^k)(s,a) = \langle \phi(s,a), w\rangle$ for some $w \in \mathbb{R}^d$.

So $((P_h - \widehat{P}_h^k)\widehat{V}_h^k)$ becomes a self-normalized linear-regression error, controlled at rate $\|\phi\|_{(\Lambda_h^k)^{-1}}$.

Consequence: Q^π Is Linear for Every Policy

Lemma (Linearity of Q^π)

In a linear MDP, for *every* policy π and every $h \in [H]$, there exists $w_h^\pi \in \mathbb{R}^d$ with

$$Q_h^\pi(s,a) = \langle \phi(s,a), w_h^\pi \rangle \qquad \forall (s,a).$$

Moreover $\|w_h^\pi\|_2 \leq 2H\sqrt{d}$.

Consequence: Q^π Is Linear for Every Policy

Lemma (Linearity of Q^π)

In a linear MDP, for *every* policy π and every $h \in [H]$, there exists $w_h^\pi \in \mathbb{R}^d$ with

$$Q_h^\pi(s,a) \;=\; \langle \phi(s,a),\, w_h^\pi \rangle \qquad \forall (s,a).$$

Moreover $\|w_h^\pi\|_2 \le 2H\sqrt{d}$.

Proof. Backward induction. $h = H+1$: $Q_{H+1}^\pi \equiv 0$. ✓

Step. Apply universal linearity to V_{h+1}^π:

$$Q_h^\pi(s,a) = \langle \phi(s,a), \theta_h \rangle + (P_h V_{h+1}^\pi)(s,a) = \left\langle \phi(s,a),\; \theta_h + \int V_{h+1}^\pi \, d\mu_h \right\rangle.$$

So $w_h^\pi = \theta_h + \int V_{h+1}^\pi \, d\mu_h$, with $\|w_h^\pi\| \le \sqrt{d} + H\sqrt{d} \le 2H\sqrt{d}$. □

Consequence: Q^π Is Linear for Every Policy

Lemma (Linearity of Q^π)

In a linear MDP, for *every* policy π and every $h \in [H]$, there exists $w_h^\pi \in \mathbb{R}^d$ with

$$Q_h^\pi(s,a) \;=\; \langle \phi(s,a),\, w_h^\pi \rangle \qquad \forall (s,a).$$

Moreover $\|w_h^\pi\|_2 \le 2H\sqrt{d}$.

Proof. Backward induction. $h = H+1$: $Q_{H+1}^\pi \equiv 0$. ✓

Step. Apply universal linearity to V_{h+1}^π:

$$Q_h^\pi(s,a) = \langle \phi(s,a), \theta_h \rangle + (P_h V_{h+1}^\pi)(s,a) = \left\langle \phi(s,a),\; \theta_h + \textstyle\int V_{h+1}^\pi \, d\mu_h \right\rangle.$$

So $w_h^\pi = \theta_h + \int V_{h+1}^\pi \, d\mu_h$, with $\|w_h^\pi\| \le \sqrt{d} + H\sqrt{d} \le 2H\sqrt{d}$. □

Hierarchy of linear assumptions

(R) linear $Q^\star$ $\subsetneq$ (C) Bellman completeness $\subsetneq$ (L) Linear MDP.

(L) gives linearity of Q^π for *all* π — the strongest of the three.

Today's Plan

From UCB-VI to LSVI-UCB

Previous section: linear FA with a generative model — queries chosen via G-optimal design.
Now: no generative model; we see only on-policy trajectories and must explore to discover relevant (s, a) pairs.

Recall UCB-VI in the tabular case:

$$\widehat{Q}_h^k(s,a) := r_h(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + \underbrace{H\sqrt{\frac{\log(\cdot)}{n_h^k(s,a)}}}_{\text{count-based bonus}} .$$

Linear MDP analogue.

- Replace the empirical Bellman backup with a ridge regression of $r_h + \widehat{V}_{h+1}^k$ onto ϕ.
- Replace counts $n_h^k(s,a)$ by the elliptical norm $\|\phi(s,a)\|^2_{(\Lambda_h^k)^{-1}}$.

From UCB-VI to LSVI-UCB

Previous section: linear FA with a generative model — queries chosen via G-optimal design.
Now: no generative model; we see only on-policy trajectories and must explore to discover relevant (s, a) pairs.

Recall UCB-VI in the tabular case:

$$\widehat{Q}_h^k(s,a) := r_h(s,a) + (\widehat{P}_h^k \widehat{V}_{h+1}^k)(s,a) + \underbrace{H\sqrt{\frac{\log(\cdot)}{n_h^k(s,a)}}}_{\text{count-based bonus}} .$$

Linear MDP analogue.

- Replace the empirical Bellman backup with a ridge regression of $r_h + \widehat{V}_{h+1}^k$ onto ϕ.
- Replace counts $n_h^k(s,a)$ by the elliptical norm $\|\phi(s,a)\|^2_{(\Lambda_h^k)^{-1}}$.

Slogan

Counts $\longleftrightarrow$ *Elliptical norm* $\qquad \sqrt{\frac{1}{n_h^k(s,a)}} \longleftrightarrow \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$

Ridge Regression Recap

Given data $\{(\phi_i, y_i)\}_{i=1}^n$ with $\phi_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, ridge regression with regularizer $\lambda > 0$:

$$\widehat{w} \;=\; \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^n (\langle \phi_i, w \rangle - y_i)^2 + \lambda \|w\|_2^2 .$$

Closed form:

$$\widehat{w} \;=\; \Lambda^{-1} \sum_{i=1}^n \phi_i y_i, \qquad \Lambda := \lambda I + \sum_{i=1}^n \phi_i \phi_i^\top .$$

Ridge Regression Recap

Given data $\{(\phi_i, y_i)\}_{i=1}^n$ with $\phi_i \in \mathbb{R}^d$, $y_i \in \mathbb{R}$, ridge regression with regularizer $\lambda > 0$:

$$\widehat{w} = \arg\min_{w \in \mathbb{R}^d} \sum_{i=1}^n (\langle \phi_i, w \rangle - y_i)^2 + \lambda \|w\|_2^2 .$$

Closed form:

$$\widehat{w} = \Lambda^{-1} \sum_{i=1}^n \phi_i y_i, \qquad \Lambda := \lambda I + \sum_{i=1}^n \phi_i \phi_i^\top .$$

Elliptical norm

$$\|x\|_\Lambda^2 := x^\top \Lambda x, \qquad \|x\|_{\Lambda^{-1}}^2 := x^\top \Lambda^{-1} x.$$

Interpretation. $\|\phi(s,a)\|_{\Lambda^{-1}}^2$ measures how "new" the direction $\phi(s,a)$ is, given past features:

- Tabular case ($\phi(s,a) = e_{(s,a)}$): $\Lambda = \mathrm{diag}(\lambda + n(s,a))$, so $\|\phi(s,a)\|_{\Lambda^{-1}}^2 = 1/(\lambda + n(s,a))$.
- Hence $\|\phi(s,a)\|_{\Lambda^{-1}}^2$ *is* the linear-algebraic analog of the inverse count.

LSVI-UCB (Jin, Yang, Wang, Jordan 2020)

Algorithm

Input: feature map ϕ, regularizer $\lambda > 0$, bonus scale $\beta > 0$.

For episode $k = 1, 2, \ldots, K$:

1. Set $\widehat{V}^k_{H+1}(\cdot) \equiv 0$. For $h = H, H-1, \ldots, 1$:

$$\Lambda^k_h = \lambda I + \sum_{\tau=1}^{k-1} \phi(s^\tau_h, a^\tau_h)\phi(s^\tau_h, a^\tau_h)^\top,$$

$$\widehat{w}^k_h = (\Lambda^k_h)^{-1} \sum_{\tau=1}^{k-1} \phi(s^\tau_h, a^\tau_h)\left[r^\tau_h + \widehat{V}^k_{h+1}(s^\tau_{h+1})\right],$$

$$b^k_h(s,a) = \beta \, \|\phi(s,a)\|_{(\Lambda^k_h)^{-1}},$$

$$\widehat{Q}^k_h(s,a) = \min\left\{H, \left\langle \phi(s,a), \widehat{w}^k_h \right\rangle + b^k_h(s,a)\right\},$$

$$\widehat{V}^k_h(s) = \max_a \widehat{Q}^k_h(s,a), \quad \pi^k_h(s) = \arg\max_a \widehat{Q}^k_h(s,a).$$

2. Execute π^k, collect trajectory $(s^k_1, a^k_1, r^k_1, \ldots, s^k_H, a^k_H, r^k_H)$.

Why the Closed Form Looks Right

If we knew $V^\star_{h+1}$ exactly, the linearity lemma says

$$r_h(s,a) + \mathbb{E}_{s'} V^\star_{h+1}(s') \;=\; \langle \phi(s,a),\, w^\star_h \rangle\,,$$

with $w^\star_h = \theta_h + \int V^\star_{h+1}\, d\mu_h$.

Why the Closed Form Looks Right

If we knew $V^\star_{h+1}$ exactly, the linearity lemma says

$$r_h(s,a) + \mathbb{E}_{s'} V^\star_{h+1}(s') \;=\; \langle \phi(s,a),\, w^\star_h \rangle\,,$$

with $w^\star_h = \theta_h + \int V^\star_{h+1}\, d\mu_h$.

With $\widehat{V}^k_{h+1}$ as a stand-in for $V^\star_{h+1}$, the regression target

$$y^\tau_h \;=\; r^\tau_h + \widehat{V}^k_{h+1}(s^\tau_{h+1})$$

satisfies

$$\mathbb{E}[y^\tau_h \mid s^\tau_h, a^\tau_h] \;=\; \left\langle \phi(s^\tau_h, a^\tau_h),\, w^k_h \right\rangle, \qquad w^k_h \triangleq \theta_h + \int \widehat{V}^k_{h+1}\, d\mu_h.$$

Ridge regression $\Rightarrow$ $\widehat{w}^k_h$ approximates w^k_h.

Why the Closed Form Looks Right

If we knew $V^\star_{h+1}$ exactly, the linearity lemma says

$$r_h(s,a) + \mathbb{E}_{s'} V^\star_{h+1}(s') \;=\; \langle \phi(s,a),\, w^\star_h \rangle\,,$$

with $w^\star_h = \theta_h + \int V^\star_{h+1}\, d\mu_h$.

With $\widehat{V}^k_{h+1}$ as a stand-in for $V^\star_{h+1}$, the regression target

$$y^\tau_h \;=\; r^\tau_h + \widehat{V}^k_{h+1}(s^\tau_{h+1})$$

satisfies

$$\mathbb{E}[y^\tau_h \mid s^\tau_h, a^\tau_h] \;=\; \left\langle \phi(s^\tau_h, a^\tau_h),\, w^k_h \right\rangle, \qquad w^k_h \triangleq \theta_h + \int \widehat{V}^k_{h+1}\, d\mu_h.$$

Ridge regression $\Rightarrow$ $\widehat{w}^k_h$ approximates w^k_h.

The key issue

$\widehat{V}^k_{h+1}$ is itself a function of past data, so the regression targets are *not* i.i.d. around a fixed w^k_h.

We need a uniform concentration bound over a function class containing all candidate $\widehat{V}^k_{h+1}$.

Today's Plan

Self-Normalized Concentration: Warm-up

Setup: features $\phi_1, \ldots, \phi_n \in \mathbb{R}^d$ (possibly chosen adaptively) and noises $\eta_1, \ldots, \eta_n$ with $\mathbb{E}[\eta_i \mid \mathcal{F}_{i-1}] = 0$ and $|\eta_i| \leq R$.

Let $\Lambda_n = \lambda I + \sum_{i \leq n} \phi_i \phi_i^\top$ and $S_n = \sum_{i \leq n} \phi_i \eta_i$.

Self-Normalized Concentration: Warm-up

Setup: features $\phi_1, \ldots, \phi_n \in \mathbb{R}^d$ (possibly chosen adaptively) and noises $\eta_1, \ldots, \eta_n$ with $\mathbb{E}[\eta_i \mid \mathcal{F}_{i-1}] = 0$ and $|\eta_i| \leq R$.

Let $\Lambda_n = \lambda I + \sum_{i \leq n} \phi_i \phi_i^\top$ and $S_n = \sum_{i \leq n} \phi_i \eta_i$.

Theorem (Abbasi-Yadkori, Pál, Szepesvári 2011)

With probability at least $1 - \delta$, for all $n \geq 1$ simultaneously:

$$\|S_n\|_{\Lambda_n^{-1}} \leq R\sqrt{2\log\left(\frac{\det(\Lambda_n)^{1/2}}{\delta \det(\lambda I)^{1/2}}\right)} \leq R\sqrt{d\log\big(1 + n/(\lambda d)\big) + 2\log(1/\delta)}.$$

Self-Normalized Concentration: Warm-up

Setup: features $\phi_1, \ldots, \phi_n \in \mathbb{R}^d$ (possibly chosen adaptively) and noises $\eta_1, \ldots, \eta_n$ with $\mathbb{E}[\eta_i \mid \mathcal{F}_{i-1}] = 0$ and $|\eta_i| \le R$.

Let $\Lambda_n = \lambda I + \sum_{i \le n} \phi_i \phi_i^\top$ and $S_n = \sum_{i \le n} \phi_i \eta_i$.

Theorem (Abbasi-Yadkori, Pál, Szepesvári 2011)

With probability at least $1 - \delta$, for all $n \ge 1$ simultaneously:

$$\|S_n\|_{\Lambda_n^{-1}} \le R\sqrt{2\log\left(\frac{\det(\Lambda_n)^{1/2}}{\delta \det(\lambda I)^{1/2}}\right)} \le R\sqrt{d\log\big(1 + n/(\lambda d)\big) + 2\log(1/\delta)}.$$

Interpretation. This is the linear analogue of Hoeffding's bound: $\frac{1}{\sqrt{n}}$ becomes $\|\phi\|_{\Lambda_n^{-1}}$, and $\log(1/\delta)$ becomes $\log(\det \Lambda_n / \det(\lambda I)) \le d \log(1 + n/(\lambda d))$.

Confidence Set for w_h^k

Lemma (Confidence ellipsoid)

Suppose the regression target satisfies $y_h^\tau = \langle \phi(s_h^\tau, a_h^\tau), w_h^k \rangle + \eta_h^\tau$ with $\mathbb{E}[\eta_h^\tau \mid \mathcal{F}_{\tau-1}] = 0$ and $|\eta_h^\tau| \leq H$. Then with probability $\geq 1 - \delta$, for all k, h:

$$\left\| \widehat{w}_h^k - w_h^k \right\|_{\Lambda_h^k} \leq \beta_*, \qquad \beta_* := H\sqrt{d \log(1 + K/(\lambda d))} + \sqrt{\lambda} \left\| w_h^k \right\|_2 .$$

Confidence Set for w_h^k

Lemma (Confidence ellipsoid)

Suppose the regression target satisfies $y_h^\tau = \langle \phi(s_h^\tau, a_h^\tau), w_h^k \rangle + \eta_h^\tau$ with $\mathbb{E}[\eta_h^\tau \mid \mathcal{F}_{\tau-1}] = 0$ and $|\eta_h^\tau| \le H$. Then with probability $\ge 1-\delta$, for all k, h:

$$\left\| \widehat{w}_h^k - w_h^k \right\|_{\Lambda_h^k} \le \beta_*, \qquad \beta_* := H\sqrt{d \log(1 + K/(\lambda d))} + \sqrt{\lambda} \left\| w_h^k \right\|_2 .$$

Consequence. By Cauchy–Schwarz:

$$\left| \left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle \right| \le \left\| \widehat{w}_h^k - w_h^k \right\|_{\Lambda_h^k} \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \le \beta_* \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} .$$

Confidence Set for w_h^k

Lemma (Confidence ellipsoid)

Suppose the regression target satisfies $y_h^\tau = \langle \phi(s_h^\tau, a_h^\tau), w_h^k \rangle + \eta_h^\tau$ with $\mathbb{E}[\eta_h^\tau \mid \mathcal{F}_{\tau-1}] = 0$ and $|\eta_h^\tau| \le H$. Then with probability $\ge 1-\delta$, for all k, h:

$$\left\| \widehat{w}_h^k - w_h^k \right\|_{\Lambda_h^k} \le \beta_*, \qquad \beta_* := H\sqrt{d \log(1 + K/(\lambda d))} + \sqrt{\lambda} \left\| w_h^k \right\|_2 .$$

Consequence. By Cauchy–Schwarz:

$$\left| \left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle \right| \le \left\| \widehat{w}_h^k - w_h^k \right\|_{\Lambda_h^k} \cdot \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \le \beta_* \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} .$$

The bonus has the right form

Choosing $\beta = \widetilde{O}(H\sqrt{d})$, we get $\left| \left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle \right| \le b_h^k(s,a)$ on the good event.

The Problem: Adaptive Targets

The confidence-ellipsoid lemma assumed a *fixed* target w_h^k. But $w_h^k = \theta_h + \int \widehat{V}_{h+1}^k \, d\mu_h$ depends on $\widehat{V}_{h+1}^k$, which depends on the data through $\widehat{w}_{h+1}^k, \widehat{w}_{h+2}^k, \ldots$

The Problem: Adaptive Targets

The confidence-ellipsoid lemma assumed a *fixed* target w_h^k. But $w_h^k = \theta_h + \int \widehat{V}_{h+1}^k \, d\mu_h$ depends on $\widehat{V}_{h+1}^k$, which depends on the data through $\widehat{w}_{h+1}^k, \widehat{w}_{h+2}^k, \ldots$

Naive worry. If we apply the self-normalized bound *conditional* on $\widehat{V}_{h+1}^k$, the conditioning destroys the martingale structure — $\widehat{V}_{h+1}^k$ was built from the very samples η_h^τ that enter the noise.

The Problem: Adaptive Targets

The confidence-ellipsoid lemma assumed a *fixed* target w_h^k. But $w_h^k = \theta_h + \int \widehat{V}_{h+1}^k \, d\mu_h$ depends on $\widehat{V}_{h+1}^k$, which depends on the data through $\widehat{w}_{h+1}^k, \widehat{w}_{h+2}^k, \ldots$

Naive worry. If we apply the self-normalized bound *conditional* on $\widehat{V}_{h+1}^k$, the conditioning destroys the martingale structure — $\widehat{V}_{h+1}^k$ was built from the very samples η_h^τ that enter the noise.

Fix: union-bound over a cover of all possible V's. Define

$$\mathcal{V} := \Big\{ V(s) = \min\big\{H,\ \max_a \big[\langle \phi(s,a), w\rangle + \beta \, \|\phi(s,a)\|_{\Lambda^{-1}}\big]\big\} \ :\ \|w\| \le L,\ \Lambda \succeq \lambda I \Big\}.$$

By construction, $\widehat{V}_{h+1}^k \in \mathcal{V}$ for every k.

Covering number bound (Jin et al. 2020)

$\mathcal{V}$ admits an ε-cover $\mathcal{N}_\varepsilon$ (in $\|\cdot\|_\infty$) of size
$\log |\mathcal{N}_\varepsilon| \ \lesssim\ d \log(1 + L/\varepsilon) + d^2 \log(1 + \beta^2/(\lambda \varepsilon^2))$.

Cover Argument, in Detail

Step 1: pointwise bound for fixed V. For each *fixed* $V \in \mathcal{V}$, let $w(V) := \theta_h + \int V \, d\mu_h$. Targets $r_h^\tau + V(s_{h+1}^\tau)$ are i.i.d. around $\langle \phi_h^\tau, w(V) \rangle$; self-normalized lemma at confidence δ': $\left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \leq H\sqrt{2\log(1/\delta') + d\log(1 + K/(\lambda d))}$.

Cover Argument, in Detail

Step 1: pointwise bound for fixed V. For each *fixed* $V \in \mathcal{V}$, let $w(V) := \theta_h + \int V \, d\mu_h$. Targets $r_h^\tau + V(s_{h+1}^\tau)$ are i.i.d. around $\langle \phi_h^\tau, w(V) \rangle$; self-normalized lemma at confidence δ':
$\left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \leq H\sqrt{2\log(1/\delta') + d\log(1 + K/(\lambda d))}$.

Step 2: union bound. Apply Step 1 to all $V \in \mathcal{N}_\varepsilon$ at $\delta' = \delta/|\mathcal{N}_\varepsilon|$:
$\forall V \in \mathcal{N}_\varepsilon : \ \left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \leq H\sqrt{2\log(|\mathcal{N}_\varepsilon|/\delta) + d\log(1 + K/(\lambda d))}$.

Cover Argument, in Detail

Step 1: pointwise bound for fixed V. For each *fixed* $V \in \mathcal{V}$, let $w(V) := \theta_h + \int V \, d\mu_h$. Targets $r_h^\tau + V(s_{h+1}^\tau)$ are i.i.d. around $\langle \phi_h^\tau, w(V) \rangle$; self-normalized lemma at confidence δ':
$\left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \leq H\sqrt{2\log(1/\delta') + d\log(1 + K/(\lambda d))}$.

Step 2: union bound. Apply Step 1 to all $V \in \mathcal{N}_\varepsilon$ at $\delta' = \delta/|\mathcal{N}_\varepsilon|$:
$\forall V \in \mathcal{N}_\varepsilon : \ \left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \leq H\sqrt{2\log(|\mathcal{N}_\varepsilon|/\delta) + d\log(1 + K/(\lambda d))}$.

Step 3: discretization. For $\widehat{V}_{h+1}^k \in \mathcal{V}$, pick $\tilde{V} \in \mathcal{N}_\varepsilon$ with $\|\widehat{V}_{h+1}^k - \tilde{V}\|_\infty \leq \varepsilon$. Triangle inequality gives $\left\| \widehat{w}_h^k(\widehat{V}_{h+1}^k) - w(\widehat{V}_{h+1}^k) \right\|_{\Lambda_h^k} \leq \left\| \widehat{w}_h^k(\tilde{V}) - w(\tilde{V}) \right\|_{\Lambda_h^k} + O(\varepsilon\sqrt{Kd})$.

Cover Argument, in Detail

Step 1: pointwise bound for fixed V. For each *fixed* $V \in \mathcal{V}$, let $w(V) := \theta_h + \int V \, d\mu_h$. Targets $r_h^\tau + V(s_{h+1}^\tau)$ are i.i.d. around $\langle \phi_h^\tau, w(V) \rangle$; self-normalized lemma at confidence δ':
$\left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \leq H\sqrt{2\log(1/\delta') + d\log(1 + K/(\lambda d))}$.

Step 2: union bound. Apply Step 1 to all $V \in \mathcal{N}_\varepsilon$ at $\delta' = \delta/|\mathcal{N}_\varepsilon|$:
$\forall V \in \mathcal{N}_\varepsilon : \ \left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \leq H\sqrt{2\log(|\mathcal{N}_\varepsilon|/\delta) + d\log(1 + K/(\lambda d))}$.

Step 3: discretization. For $\widehat{V}_{h+1}^k \in \mathcal{V}$, pick $\tilde{V} \in \mathcal{N}_\varepsilon$ with $\|\widehat{V}_{h+1}^k - \tilde{V}\|_\infty \leq \varepsilon$. Triangle inequality gives $\left\| \widehat{w}_h^k(\widehat{V}_{h+1}^k) - w(\widehat{V}_{h+1}^k) \right\|_{\Lambda_h^k} \leq \left\| \widehat{w}_h^k(\tilde{V}) - w(\tilde{V}) \right\|_{\Lambda_h^k} + O(\varepsilon\sqrt{Kd})$.

Choose $\varepsilon = 1/K$. Then $\log|\mathcal{N}_\varepsilon| = \widetilde{O}(d^2)$, discretization $= \widetilde{O}(1)$. Net: replace $\log(1/\delta)$ in the self-normalized bound by $\widetilde{O}(d^2 + \log(1/\delta))$.

Cover Argument, in Detail

Step 1: pointwise bound for fixed V. For each *fixed* $V \in \mathcal{V}$, let $w(V) := \theta_h + \int V \, d\mu_h$. Targets $r_h^\tau + V(s_{h+1}^\tau)$ are i.i.d. around $\langle \phi_h^\tau, w(V) \rangle$; self-normalized lemma at confidence δ':
$\left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \le H\sqrt{2\log(1/\delta') + d\log(1 + K/(\lambda d))}$.

Step 2: union bound. Apply Step 1 to all $V \in \mathcal{N}_\varepsilon$ at $\delta' = \delta/|\mathcal{N}_\varepsilon|$:
$\forall V \in \mathcal{N}_\varepsilon : \ \left\| \widehat{w}_h^k(V) - w(V) \right\|_{\Lambda_h^k} \le H\sqrt{2\log(|\mathcal{N}_\varepsilon|/\delta) + d\log(1 + K/(\lambda d))}$.

Step 3: discretization. For $\widehat{V}_{h+1}^k \in \mathcal{V}$, pick $\tilde{V} \in \mathcal{N}_\varepsilon$ with $\|\widehat{V}_{h+1}^k - \tilde{V}\|_\infty \le \varepsilon$. Triangle inequality gives $\left\| \widehat{w}_h^k(\widehat{V}_{h+1}^k) - w(\widehat{V}_{h+1}^k) \right\|_{\Lambda_h^k} \le \left\| \widehat{w}_h^k(\tilde{V}) - w(\tilde{V}) \right\|_{\Lambda_h^k} + O(\varepsilon\sqrt{Kd})$.

Choose $\varepsilon = 1/K$. Then $\log|\mathcal{N}_\varepsilon| = \widetilde{O}(d^2)$, discretization $= \widetilde{O}(1)$. Net: replace $\log(1/\delta)$ in the self-normalized bound by $\widetilde{O}(d^2 + \log(1/\delta))$.

End-to-end consequence

Choosing $\beta = c\,Hd\sqrt{\log(dKH/\delta)}$, on the good event $\mathcal{E}$:
$|\langle \phi(s,a), \widehat{w}_h^k - w_h^k \rangle| \le \beta \, \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} = b_h^k(s,a)$ holds simultaneously for all k, h, s, a.

Today's Plan

Lemma: Optimism

Lemma (Optimism)

On the good event $\mathcal{E}$ (the uniform self-normalized bound holds), for all k, h, s, a:

$$\widehat{Q}_h^k(s,a) \geq Q_h^\star(s,a), \qquad \widehat{V}_h^k(s) \geq V_h^\star(s).$$

Lemma: Optimism

Lemma (Optimism)

On the good event $\mathcal{E}$ (the uniform self-normalized bound holds), for all k, h, s, a:

$$\widehat{Q}_h^k(s,a) \geq Q_h^\star(s,a), \qquad \widehat{V}_h^k(s) \geq V_h^\star(s).$$

Proof (backward induction on h):
Base ($h = H+1$): both sides are zero. ✓

Step. Suppose $\widehat{V}_{h+1}^k \geq V_{h+1}^\star$. Recall $w_h^k = \theta_h + \int \widehat{V}_{h+1}^k d\mu_h$, so

$$\left\langle \phi(s,a), w_h^k \right\rangle = r_h(s,a) + (P_h \widehat{V}_{h+1}^k)(s,a) \geq r_h(s,a) + (P_h V_{h+1}^\star)(s,a) = Q_h^\star(s,a).$$

On $\mathcal{E}$, $|\left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle| \leq b_h^k(s,a)$. Therefore

$$\widehat{Q}_h^k(s,a) \geq \left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + b_h^k(s,a) \geq \left\langle \phi(s,a), w_h^k \right\rangle \geq Q_h^\star(s,a).$$

(If the min with H is active, $\widehat{Q}_h^k = H \geq Q_h^\star$ trivially.) Take max over a. □

Regret-to-Surplus Certificate (Linear Version)

Lemma (Certificate)

Define the surplus

$$\varepsilon_h^k(s,a) \;:=\; \widehat{Q}_h^k(s,a) - r_h(s,a) - (P_h \widehat{V}_{h+1}^k)(s,a).$$

On the good event, for every episode k:

$$V_0^\star - V_0^{\pi^k} \;\le\; \sum_{h=1}^{H} \mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k) \,\middle|\, \pi^k\right].$$

Regret-to-Surplus Certificate (Linear Version)

Lemma (Certificate)

Define the surplus

$$\varepsilon_h^k(s,a) \;:=\; \widehat{Q}_h^k(s,a) - r_h(s,a) - (P_h \widehat{V}_{h+1}^k)(s,a).$$

On the good event, for every episode k:

$$V_0^\star - V_0^{\pi^k} \;\le\; \sum_{h=1}^{H} \mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k) \,\middle|\, \pi^k\right].$$

Proof. Identical to the tabular UCB-VI case (Lecture 9): define $\Gamma_h^k := \widehat{V}_h^k - V_h^{\pi^k}$ and unroll

$$\Gamma_h^k(s_h^k) \;=\; \varepsilon_h^k(s_h^k, a_h^k) + (P_h \Gamma_{h+1}^k)(s_h^k, a_h^k).$$

Optimism gives $V_0^\star \le \widehat{V}_0^k$, so $V_0^\star - V_0^{\pi^k} \le \mathbb{E}[\Gamma_1^k(s_1^k)] = \sum_h \mathbb{E}[\varepsilon_h^k]$. □

Regret-to-Surplus Certificate (Linear Version)

Lemma (Certificate)

Define the surplus

$$\varepsilon_h^k(s,a) \;:=\; \widehat{Q}_h^k(s,a) - r_h(s,a) - (P_h \widehat{V}_{h+1}^k)(s,a).$$

On the good event, for every episode k:

$$V_0^\star - V_0^{\pi^k} \;\le\; \sum_{h=1}^{H} \mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k) \,\middle|\, \pi^k\right].$$

Proof. Identical to the tabular UCB-VI case (Lecture 9): define $\Gamma_h^k := \widehat{V}_h^k - V_h^{\pi^k}$ and unroll

$$\Gamma_h^k(s_h^k) \;=\; \varepsilon_h^k(s_h^k, a_h^k) + (P_h \Gamma_{h+1}^k)(s_h^k, a_h^k).$$

Optimism gives $V_0^\star \le \widehat{V}_0^k$, so $V_0^\star - V_0^{\pi^k} \le \mathbb{E}[\Gamma_1^k(s_1^k)] = \sum_h \mathbb{E}[\varepsilon_h^k]$. □

Same structure as tabular UCB-VI. All the work goes into bounding the surplus.

Bounding the Surplus

Recall the good event:
$\mathcal{E} := \left\{ \forall k, h, s, a : \ |\left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle| \le \beta \ \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \right\}, \ \mathbb{P}(\mathcal{E}) \ge 1 - \delta.$

Unpack the surplus. When $\min\{H, \cdot\}$ is inactive, $\widehat{Q}_h^k(s,a) = \left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + b_h^k(s,a)$:

$$\begin{aligned}
\varepsilon_h^k(s,a) &= \widehat{Q}_h^k(s,a) - r_h(s,a) - (P_h \widehat{V}_{h+1}^k)(s,a) \\
&= \left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + b_h^k(s,a) - \underbrace{[r_h(s,a) + (P_h \widehat{V}_{h+1}^k)(s,a)]}_{= \left\langle \phi(s,a), w_h^k \right\rangle \text{ by (L)}} \\
&= \underbrace{\left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle}_{\text{regression error}} + \underbrace{b_h^k(s,a)}_{\text{bonus}}.
\end{aligned}$$

(The reward r_h is part of $w_h^k = \theta_h + \int \widehat{V}_{h+1}^k \, d\mu_h$ and cancels.)

Bounding the Surplus

Recall the good event:
$\mathcal{E} := \left\{ \forall k, h, s, a : \ |\left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle| \le \beta \ \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}} \right\}, \ \mathbb{P}(\mathcal{E}) \ge 1 - \delta.$

Unpack the surplus. When $\min\{H, \cdot\}$ is inactive, $\widehat{Q}_h^k(s,a) = \left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + b_h^k(s,a)$:

$$\begin{aligned}
\varepsilon_h^k(s,a) &= \widehat{Q}_h^k(s,a) - r_h(s,a) - (P_h \widehat{V}_{h+1}^k)(s,a) \\
&= \left\langle \phi(s,a), \widehat{w}_h^k \right\rangle + b_h^k(s,a) - \underbrace{[r_h(s,a) + (P_h \widehat{V}_{h+1}^k)(s,a)]}_{= \left\langle \phi(s,a), w_h^k \right\rangle \text{ by (L)}} \\
&= \underbrace{\left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle}_{\text{regression error}} + \underbrace{b_h^k(s,a)}_{\text{bonus}}.
\end{aligned}$$

(The reward r_h is part of $w_h^k = \theta_h + \int \widehat{V}_{h+1}^k \, d\mu_h$ and cancels.)

Apply $\mathcal{E}$. $|\left\langle \phi(s,a), \widehat{w}_h^k - w_h^k \right\rangle| \le b_h^k(s,a)$, so $0 \le \varepsilon_h^k(s,a) \le 2\, b_h^k(s,a) = 2\beta \ \|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$.

Today's Plan

Putting It Together

Step 1: regret to sum of surpluses. By the certificate (Lemma), on $\mathcal{E}$,

$$R(K) \;=\; \sum_{k=1}^{K} \left(V_0^\star - V_0^{\pi^k}\right) \;\le\; \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k) \,\middle|\, \pi^k\right].$$

Step 2: surplus to elliptical norm. Apply the surplus bound $\varepsilon_h^k \le 2\beta \, \|\phi\|_{(\Lambda_h^k)^{-1}}$:

$$R(K) \;\le\; 2\beta \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\left[\left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}}\right].$$

Putting It Together

Step 1: regret to sum of surpluses. By the certificate (Lemma), on $\mathcal{E}$,

$$R(K) \;=\; \sum_{k=1}^{K}\left(V_0^\star - V_0^{\pi^k}\right) \;\le\; \sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k)\,\middle|\,\pi^k\right].$$

Step 2: surplus to elliptical norm. Apply the surplus bound $\varepsilon_h^k \le 2\beta\,\|\phi\|_{(\Lambda_h^k)^{-1}}$:

$$R(K) \;\le\; 2\beta\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{E}\left[\left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}}\right].$$

Step 3: bound the sum of elliptical norms. For each fixed h, we need

$$\sum_{k=1}^{K}\left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}}.$$

Putting It Together

Step 1: regret to sum of surpluses. By the certificate (Lemma), on $\mathcal{E}$,

$$R(K) = \sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \leq \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\left[\varepsilon_h^k(s_h^k, a_h^k) \,\middle|\, \pi^k\right].$$

Step 2: surplus to elliptical norm. Apply the surplus bound $\varepsilon_h^k \leq 2\beta \, \|\phi\|_{(\Lambda_h^k)^{-1}}$:

$$R(K) \leq 2\beta \sum_{k=1}^{K} \sum_{h=1}^{H} \mathbb{E}\left[\left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}}\right].$$

Step 3: bound the sum of elliptical norms. For each fixed h, we need

$$\sum_{k=1}^{K} \left\|\phi(s_h^k, a_h^k)\right\|_{(\Lambda_h^k)^{-1}}.$$

Linear analogue of the counting argument

Tabular: $\sum_k 1/\sqrt{n_h^k(s_h^k, a_h^k)} \leq 2\sqrt{SAK}$.

Linear: Elliptical Potential Lemma, replacing SA with d.

Elliptical Potential Lemma

Lemma (Elliptical Potential, Abbasi-Yadkori et al. 2011)

Let $\phi_1, \phi_2, \ldots, \phi_K \in \mathbb{R}^d$ with $\|\phi_k\| \leq 1$, and $\Lambda_k = \lambda I + \sum_{\tau<k} \phi_\tau \phi_\tau^\top$ with $\lambda \geq 1$. Then

$$\sum_{k=1}^{K} \min\left\{1,\ \|\phi_k\|^2_{\Lambda_k^{-1}}\right\} \ \leq\ 2\log\frac{\det(\Lambda_{K+1})}{\det(\lambda I)} \ \leq\ 2d\log\left(1+\frac{K}{\lambda d}\right).$$

By Cauchy–Schwarz:

$$\sum_{k=1}^{K} \min\left\{1,\ \|\phi_k\|_{\Lambda_k^{-1}}\right\} \ \leq\ \sqrt{2Kd\log(1+K/(\lambda d))} \ =\ \widetilde{O}(\sqrt{dK}).$$

Elliptical Potential Lemma

Lemma (Elliptical Potential, Abbasi-Yadkori et al. 2011)

Let $\phi_1, \phi_2, \ldots, \phi_K \in \mathbb{R}^d$ with $\|\phi_k\| \le 1$, and $\Lambda_k = \lambda I + \sum_{\tau<k} \phi_\tau \phi_\tau^\top$ with $\lambda \ge 1$. Then

$$\sum_{k=1}^{K} \min\left\{1,\ \|\phi_k\|^2_{\Lambda_k^{-1}}\right\} \ \le\ 2\log\frac{\det(\Lambda_{K+1})}{\det(\lambda I)} \ \le\ 2d\log\left(1+\frac{K}{\lambda d}\right).$$

By Cauchy–Schwarz:

$$\sum_{k=1}^{K} \min\left\{1,\ \|\phi_k\|_{\Lambda_k^{-1}}\right\} \ \le\ \sqrt{2Kd\log(1+K/(\lambda d))} \ =\ \widetilde{O}(\sqrt{dK}).$$

Intuition. $\mathbb{R}^d$ contains only $\widetilde{O}(d)$ "effectively new" directions to explore. Each new direction increases $\log\det\Lambda$, but $\log\det$ is bounded.

Tabular check. $\phi_k = e_{(s_k,a_k)}$, $d = SA$: gives $\sum_k (n_k(s_k,a_k))^{-1/2} \le \widetilde{O}(\sqrt{SAK})$. ✓

Proof of the Elliptical Potential Lemma

Let $u_k = \min\{1, \|\phi_k\|^2_{\Lambda_k^{-1}}\}$. Use $u \leq 2\log(1+u)$ for $u \in [0,1]$:

$$\sum_k u_k \;\leq\; 2\sum_k \log\big(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\big).$$

Proof of the Elliptical Potential Lemma

Let $u_k = \min\{1, \|\phi_k\|^2_{\Lambda_k^{-1}}\}$. Use $u \leq 2\log(1+u)$ for $u \in [0,1]$:

$$\sum_k u_k \;\leq\; 2\sum_k \log\big(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\big).$$

Matrix determinant lemma:

$$\det(\Lambda_{k+1}) = \det(\Lambda_k + \phi_k\phi_k^\top) = \det(\Lambda_k)\big(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\big).$$

Proof of the Elliptical Potential Lemma

Let $u_k = \min\{1, \|\phi_k\|^2_{\Lambda_k^{-1}}\}$. Use $u \leq 2\log(1+u)$ for $u \in [0,1]$:

$$\sum_k u_k \;\leq\; 2\sum_k \log\big(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\big).$$

Matrix determinant lemma:

$$\det(\Lambda_{k+1}) = \det(\Lambda_k + \phi_k\phi_k^\top) = \det(\Lambda_k)\big(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\big).$$

Telescoping:

$$\sum_{k=1}^{K} \log\big(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\big) \;=\; \log\frac{\det(\Lambda_{K+1})}{\det(\lambda I)}.$$

Proof of the Elliptical Potential Lemma

Let $u_k = \min\{1, \|\phi_k\|^2_{\Lambda_k^{-1}}\}$. Use $u \le 2\log(1+u)$ for $u \in [0,1]$:

$$\sum_k u_k \;\le\; 2\sum_k \log\bigl(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\bigr).$$

Matrix determinant lemma:

$$\det(\Lambda_{k+1}) = \det(\Lambda_k + \phi_k\phi_k^\top) = \det(\Lambda_k)\bigl(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\bigr).$$

Telescoping:

$$\sum_{k=1}^{K} \log\bigl(1 + \|\phi_k\|^2_{\Lambda_k^{-1}}\bigr) \;=\; \log\frac{\det(\Lambda_{K+1})}{\det(\lambda I)}.$$

Bound on $\det(\Lambda_{K+1})$**.** Eigenvalues of Λ_{K+1} sum to $\mathrm{tr}(\Lambda_{K+1}) = \lambda d + \sum_k \|\phi_k\|^2 \le \lambda d + K$.
By AM–GM:

$$\det(\Lambda_{K+1}) \le \left(\tfrac{\mathrm{tr}(\Lambda_{K+1})}{d}\right)^d \le (\lambda + K/d)^d.$$

Therefore $\log\det(\Lambda_{K+1})/\det(\lambda I) \le d\log(1 + K/(\lambda d))$.

Today's Plan

Main Theorem

Theorem (Jin, Yang, Wang, Jordan 2020)

Set $\lambda = 1$ and $\beta = c\, Hd\sqrt{\log(dKH/\delta)}$ for an appropriate constant c. With probability at least $1 - \delta$, the regret of LSVI-UCB satisfies

$$R(K) \;=\; \sum_{k=1}^{K} (V_0^\star - V_0^{\pi^k}) \;\leq\; \widetilde{O}\left(\sqrt{d^3 H^4 K}\right).$$

The bound has no dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$.

Main Theorem

Theorem (Jin, Yang, Wang, Jordan 2020)

Set $\lambda = 1$ and $\beta = c\,Hd\sqrt{\log(dKH/\delta)}$ for an appropriate constant c. With probability at least $1-\delta$, the regret of LSVI-UCB satisfies

$$R(K) \;=\; \sum_{k=1}^{K}\left(V_0^\star - V_0^{\pi^k}\right) \;\le\; \widetilde{O}\left(\sqrt{d^3H^4K}\right).$$

The bound has no dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$.

Proof. From the previous slide, $R(K) \le 2\beta \sum_{h,k} \mathbb{E}[\|\phi\|_{(\Lambda_h^k)^{-1}}]$.

Why $\min\{1,\cdot\}$**?** Values are in $[0,H]$, so $\varepsilon_h^k(s,a) \le H$. Hence $\varepsilon_h^k \le \min\{H, 2\beta\,\|\phi\|_{(\Lambda_h^k)^{-1}}\} = 2\beta\min\{H/(2\beta), \|\phi\|_{(\Lambda_h^k)^{-1}}\}$. Since $\beta \ge H/2$: $\varepsilon_h^k \le 2\beta\ \min\{1, \|\phi\|_{(\Lambda_h^k)^{-1}}\}$.

Elliptical potential (with min inside): $\sum_{k=1}^{K}\min\{1, \|\phi_h^k\|_{(\Lambda_h^k)^{-1}}\} \le \sqrt{2Kd\log(1+K/d)}$.

Main Theorem

Theorem (Jin, Yang, Wang, Jordan 2020)

Set $\lambda = 1$ and $\beta = c\,Hd\sqrt{\log(dKH/\delta)}$ for an appropriate constant c. With probability at least $1-\delta$, the regret of LSVI-UCB satisfies

$$R(K) \;=\; \sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}) \;\le\; \widetilde{O}\left(\sqrt{d^3H^4K}\right).$$

The bound has no dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$.

Proof. From the previous slide, $R(K) \le 2\beta \sum_{h,k} \mathbb{E}[\|\phi\|_{(\Lambda_h^k)^{-1}}]$.

Why $\min\{1,\cdot\}$**?** Values are in $[0,H]$, so $\varepsilon_h^k(s,a) \le H$. Hence $\varepsilon_h^k \le \min\{H, 2\beta\,\|\phi\|_{(\Lambda_h^k)^{-1}}\} = 2\beta\min\{H/(2\beta), \|\phi\|_{(\Lambda_h^k)^{-1}}\}$. Since $\beta \ge H/2$: $\varepsilon_h^k \le 2\beta\ \min\{1, \|\phi\|_{(\Lambda_h^k)^{-1}}\}$.

Elliptical potential (with min inside): $\sum_{k=1}^{K}\min\{1, \|\phi_h^k\|_{(\Lambda_h^k)^{-1}}\} \le \sqrt{2Kd\log(1+K/d)}$.

Assemble. $R(K) \le 2\beta \cdot H \cdot \sqrt{2Kd\log(1+K/d)} = \widetilde{O}(H \cdot Hd \cdot \sqrt{Kd}) = \widetilde{O}(\sqrt{d^3H^4K})$. □

Comparing the Bounds

Setting	**Algorithm**	**Regret**
Tabular MDP	UCB-VI (basic)	$\widetilde{O}(H^2 S\sqrt{SAK})$
Tabular MDP	UCB-VI (Bernstein)	$\widetilde{O}(H^{3/2}\sqrt{SAK})$
Linear MDP	LSVI-UCB	$\widetilde{O}(\sqrt{d^3H^4K})$
Linear MDP	lower bound	$\Omega(d\sqrt{H^3K})$

Comparing the Bounds

Setting	**Algorithm**	**Regret**
Tabular MDP	UCB-VI (basic)	$\widetilde{O}(H^2 S\sqrt{SAK})$
Tabular MDP	UCB-VI (Bernstein)	$\widetilde{O}(H^{3/2}\sqrt{SAK})$
Linear MDP	LSVI-UCB	$\widetilde{O}(\sqrt{d^3H^4K})$
Linear MDP	lower bound	$\Omega(d\sqrt{H^3K})$

Take-aways.

- Replacing S, A with d is a huge win when features are well-chosen.
- Polynomial gap to the lower bound: closing it has been an active research direction.
- Computation: each Bellman backup requires only inverting a $d \times d$ matrix and computing one ridge-regression solution per stage.

Where the d^3H^4 Comes From

Decomposition of d^3H^4K:

Source	**Contribution**
Bonus scale $\beta = \widetilde{O}(H\sqrt{d})$	$H^2 d$
Elliptical potential $(\sqrt{dK})$	dK
Sum over $h \in [H]$ stages	H^2
Covering number for $\mathcal{V}$	extra $\widetilde{O}(d)$ in β
Product	$\widetilde{O}(\sqrt{d^3H^4K})$

Where Can We Tighten?

- **d-dependence ($d^3 \to d^2$):** improved uniform concentration / randomized LSVI.

  Zanette, Lazaric, Kochenderfer, Brunskill, *Frequentist regret bounds for randomized least-squares value iteration*, AISTATS 2020.

- **H-dependence ($H^4 \to H^3$):** variance-aware (Bernstein) recursion, analogous to Lecture 10.

  Hu, Yang, Wang, *Nearly minimax optimal RL with linear function approximation*, ICML 2022;

  Agarwal, Jin, Zhang, *VOQL: Towards optimal regret in model-free RL with nonlinear function approximation*, COLT 2023.

- **Closing the gap to $\Omega(d\sqrt{H^3K})$:**

  He, Zhou, Gu, *Nearly minimax optimal RL for linear Markov decision processes*, ICML 2023.

- **Lower bound:**

  Zhou, Gu, Szepesvári, *Nearly minimax optimal RL for linear mixture Markov decision processes*, COLT 2021.

Today's Plan

Linear Realizability is Not Enough

Weaker assumption: only $Q_h^\star(s,a) = \langle \phi(s,a), w_h^\star \rangle$ for some $w_h^\star \in \mathbb{R}^d$.

Linear Realizability is Not Enough

Weaker assumption: only $Q_h^\star(s,a) = \langle \phi(s,a), w_h^\star \rangle$ for some $w_h^\star \in \mathbb{R}^d$.

Theorem (Du, Kakade, Lee, Lovett, Mahajan, Sun, Wang 2020)

There exist MDPs with linear $Q^\star$ (in d-dimensional features) for which any algorithm requires $\Omega(\min(2^d, 2^H))$ samples to find an ε-optimal policy.

Linear Realizability is Not Enough

Weaker assumption: only $Q_h^\star(s,a) = \langle \phi(s,a), w_h^\star \rangle$ for some $w_h^\star \in \mathbb{R}^d$.

Theorem (Du, Kakade, Lee, Lovett, Mahajan, Sun, Wang 2020)

There exist MDPs with linear $Q^\star$ (in d-dimensional features) for which any algorithm requires $\Omega(\min(2^d, 2^H))$ samples to find an ε-optimal policy.

Why is the linear MDP assumption stronger?
Linear MDP $\Rightarrow$ Q^π is linear for *every* policy π.
Linear $Q^\star$ $\Rightarrow$ only $Q^\star$ (and not other Q^π) is linear.

Linear Realizability is Not Enough

Weaker assumption: only $Q_h^\star(s,a) = \langle \phi(s,a), w_h^\star \rangle$ for some $w_h^\star \in \mathbb{R}^d$.

Theorem (Du, Kakade, Lee, Lovett, Mahajan, Sun, Wang 2020)

There exist MDPs with linear $Q^\star$ (in d-dimensional features) for which any algorithm requires $\Omega(\min(2^d, 2^H))$ samples to find an ε-optimal policy.

Why is the linear MDP assumption stronger?
Linear MDP $\Rightarrow$ Q^π is linear for *every* policy π.
Linear $Q^\star$ $\Rightarrow$ only $Q^\star$ (and not other Q^π) is linear.

Moral

Sample-efficient RL with function approximation requires *closure under Bellman operators*, not just realizability of the optimal value function.

Summary

What we did today

- Three nested assumptions: **(R)** $Q^\star \in \mathcal{F}$, **(C)** $\mathcal{T}_h \mathcal{F} \subseteq \mathcal{F}$, **(L)** linear MDP.
- **Generative model + (C):** LSVI with G-optimal design gives $\widetilde{O}(d^2 H^4 / \varepsilon^2)$ samples.
- **(R) alone is not enough:** $2^{\min(d,H)}$ lower bound.
- **Online + (L):** LSVI-UCB achieves $R(K) = \widetilde{O}(\sqrt{d^3 H^4 K})$, no $|\mathcal{S}|, |\mathcal{A}|$ dependence.
- Key tools: self-normalized concentration, elliptical potential lemma.

Summary

What we did today

- Three nested assumptions: **(R)** $Q^\star \in \mathcal{F}$, **(C)** $\mathcal{T}_h\mathcal{F} \subseteq \mathcal{F}$, **(L)** linear MDP.
- **Generative model + (C):** LSVI with G-optimal design gives $\widetilde{O}(d^2H^4/\varepsilon^2)$ samples.
- **(R) alone is not enough:** $2^{\min(d,H)}$ lower bound.
- **Online + (L):** LSVI-UCB achieves $R(K) = \widetilde{O}(\sqrt{d^3H^4K})$, no $|\mathcal{S}|, |\mathcal{A}|$ dependence.
- Key tools: self-normalized concentration, elliptical potential lemma.

Mental dictionary:

Tabular	**Linear**
$n_h^k(s,a)$	$\Lambda_h^k = \lambda I + \sum_\tau \phi_h^\tau {\phi_h^\tau}^\top$
$1/\sqrt{n_h^k(s,a)}$	$\|\phi(s,a)\|_{(\Lambda_h^k)^{-1}}$
counting bound $(\sqrt{SAK})$	elliptical potential $(\sqrt{dK})$
$S \cdot A$	d

Reading & References

Primary reference:

AJKS Agarwal, Jiang, Kakade, Sun, *Reinforcement Learning: Theory and Algorithms*, Chapter 3.

Foundational:

- Jin, Yang, Wang, Jordan, COLT 2020 (LSVI-UCB).
- Abbasi-Yadkori, Pál, Szepesvári, NeurIPS 2011 (self-normalized concentration).
- Weisz, Amortila, Szepesvári, ALT 2021 (realizability lower bound).

Improvements:

- Zanette, Lazaric, Kochenderfer, Brunskill, AISTATS 2020.
- Hu, Yang, Wang, ICML 2022.
- He, Zhou, Gu, ICML 2023.

Questions?