CSE 542: Statistical Reinforcement Learning

Lecture 12: Linear Function Approximation – Generative Model & Offline

Kevin Jamieson

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

Outline

Today's Plan

The Trouble With Tabular RL

Last time: UCB-VI achieves

$$R(K) \;=\; \widetilde{O}\left(H^{3/2}\sqrt{SAK}\right).$$

- Sample complexity scales linearly in S and A.
- In real applications S is astronomical:
  - Atari: $\sim 10^{60,000}$ frames
  - Robot manipulation: continuous $\mathbb{R}^d$
  - Language: vocabularies of $10^4 - 10^5$ tokens
- No hope of visiting every (s, a), let alone $\sqrt{K}$ times each.

The Trouble With Tabular RL

Last time: UCB-VI achieves

$$R(K) = \widetilde{O}\left(H^{3/2}\sqrt{SAK}\right).$$

- Sample complexity scales linearly in S and A.
- In real applications S is astronomical:
  - Atari: $\sim 10^{60,000}$ frames
  - Robot manipulation: continuous $\mathbb{R}^d$
  - Language: vocabularies of $10^4 - 10^5$ tokens
- No hope of visiting every (s, a), let alone $\sqrt{K}$ times each.

The fundamental question

Can we learn a near-optimal policy in time *independent of* $|\mathcal{S}|$, by exploiting structure?

The Trouble With Tabular RL

Last time: UCB-VI achieves

$$R(K) = \widetilde{O}\left(H^{3/2}\sqrt{SAK}\right).$$

- Sample complexity scales linearly in S and A.
- In real applications S is astronomical:
  - Atari: $\sim 10^{60,000}$ frames
  - Robot manipulation: continuous $\mathbb{R}^d$
  - Language: vocabularies of $10^4 - 10^5$ tokens
- No hope of visiting every (s, a), let alone $\sqrt{K}$ times each.

The fundamental question

Can we learn a near-optimal policy in time *independent of* $|\mathcal{S}|$, by exploiting structure?

Answer (today): yes — if value functions are linear in known features.

Function Approximation: Two Generic Approaches

Value-based (model-free)

Approximate the *value function* Q from a class $\mathcal{F}$. Solve an empirical Bellman equation in $\mathcal{F}$.

Examples: linear Q, kernel Q, deep Q-networks.

Model-based

Approximate the *model* (P, r) from a class $\mathcal{M}$. Plan in the learned model.

Examples: linear models, kernel models, neural-net dynamics.

Function Approximation: Two Generic Approaches

Value-based (model-free)

Approximate the *value function* Q from a class $\mathcal{F}$. Solve an empirical Bellman equation in $\mathcal{F}$.

Examples: linear Q, kernel Q, deep Q-networks.

Model-based

Approximate the *model* (P, r) from a class $\mathcal{M}$. Plan in the learned model.

Examples: linear models, kernel models, neural-net dynamics.

Plan for today

1. Start *model-free*: assume Q linear in features ϕ.
2. In the generative-model setting this already gives $\mathrm{poly}(d, H)/\varepsilon^2$ samples, under a closure condition (**Bellman completeness**).
3. For online RL with exploration, model-free turns out to be insufficient — we need a *model-based* structural assumption (the **Linear MDP** of Jin et al. 2020).

Today's Plan

Setup: A Linear Value-Function Class

A finite-horizon MDP $(\mathcal{S}, \mathcal{A}, \{P_h\}, \{r_h\}, H, \nu)$ where $\mathcal{S}$ may be *infinite or continuous*.

We are given a known feature map

$$\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d, \qquad \|\phi(s,a)\|_2 \leq 1 \text{ for all } (s,a),$$

and consider the linear value-function class

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

Think of ϕ as: a fixed neural-net embedding, polynomial / Fourier / RBF features, or one-hot indicators (recovers tabular with $d = SA$).

Setup: A Linear Value-Function Class

A finite-horizon MDP $(\mathcal{S}, \mathcal{A}, \{P_h\}, \{r_h\}, H, \nu)$ where $\mathcal{S}$ may be *infinite or continuous*.

We are given a known feature map

$$\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d, \qquad \|\phi(s,a)\|_2 \leq 1 \text{ for all } (s,a),$$

and consider the linear value-function class

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

Think of ϕ as: a fixed neural-net embedding, polynomial / Fourier / RBF features, or one-hot indicators (recovers tabular with $d = SA$).

This is a *model-free* starting point

We assume nothing about the dynamics P_h or rewards r_h. We only commit to a d-dimensional class of candidate Q-functions.

Two Natural Model-Free Assumptions

Bellman optimality operator: $(\mathcal{T}_h Q)(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim P_h}\big[\max_{a'} Q(s',a')\big]$.

Two assumptions on the class $\mathcal{F}$

(R) Linear $Q^\star$-realizability: $Q_h^\star \in \mathcal{F}$ for every h.

Two Natural Model-Free Assumptions

Bellman optimality operator: $(\mathcal{T}_h Q)(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim P_h}\big[\max_{a'} Q(s',a')\big]$.

Two assumptions on the class $\mathcal{F}$

(R) Linear $Q^\star$-realizability: $Q_h^\star \in \mathcal{F}$ for every h.

(C) Bellman completeness: for every h and every $f \in \mathcal{F}$, $\mathcal{T}_h f \in \mathcal{F}$.

Two Natural Model-Free Assumptions

Bellman optimality operator: $(\mathcal{T}_h Q)(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim P_h}\big[\max_{a'} Q(s',a')\big]$.

Two assumptions on the class $\mathcal{F}$

(R) Linear $Q^\star$-realizability: $Q_h^\star \in \mathcal{F}$ for every h.

(C) Bellman completeness: for every h and every $f \in \mathcal{F}$, $\mathcal{T}_h f \in \mathcal{F}$.

Note. (C) $\Rightarrow$ (R): apply $\mathcal{T}_h$ iteratively from $f = 0 \in \mathcal{F}$ to obtain $Q^\star \in \mathcal{F}$.

Two Natural Model-Free Assumptions

Bellman optimality operator: $(\mathcal{T}_h Q)(s,a) := r_h(s,a) + \mathbb{E}_{s' \sim P_h}\big[\max_{a'} Q(s',a')\big]$.

Two assumptions on the class $\mathcal{F}$

(R) Linear $Q^\star$-realizability: $Q_h^\star \in \mathcal{F}$ for every h.

(C) Bellman completeness: for every h and every $f \in \mathcal{F}$, $\mathcal{T}_h f \in \mathcal{F}$.

Note. (C) $\Rightarrow$ (R): apply $\mathcal{T}_h$ iteratively from $f = 0 \in \mathcal{F}$ to obtain $Q^\star \in \mathcal{F}$.

What each assumption says

(R) says only the *single* target $Q^\star$ lives in $\mathcal{F}$.

(C) says the class $\mathcal{F}$ is *stable* under one Bellman backup — algorithms can iterate within $\mathcal{F}$ without leaving it.

Why (C) Is the "Right" Algorithmic Assumption

Most algorithms estimate $Q^\star$ by iterating Bellman backups:

$$\widehat{Q}_{H+1} \equiv 0, \qquad \widehat{Q}_h \approx \mathcal{T}_h \widehat{Q}_{h+1}.$$

If $\mathcal{F}$ is closed under $\mathcal{T}_h$, the iterates stay in $\mathcal{F}$ and least squares recovers them at the parametric rate.

If $\mathcal{F}$ is *not* closed under $\mathcal{T}_h$, then $\mathcal{T}_h \widehat{Q}_{h+1}$ has no representation in $\mathcal{F}$ — least squares introduces bias that compounds over the H stages.

Why (C) Is the "Right" Algorithmic Assumption

Most algorithms estimate $Q^\star$ by iterating Bellman backups:

$$\widehat{Q}_{H+1} \equiv 0, \qquad \widehat{Q}_h \approx \mathcal{T}_h \widehat{Q}_{h+1}.$$

If $\mathcal{F}$ is closed under $\mathcal{T}_h$, the iterates stay in $\mathcal{F}$ and least squares recovers them at the parametric rate.

If $\mathcal{F}$ is *not* closed under $\mathcal{T}_h$, then $\mathcal{T}_h \widehat{Q}_{h+1}$ has no representation in $\mathcal{F}$ — least squares introduces bias that compounds over the H stages.

Roadmap for today

1. **(C) + generative model:** $\widetilde{O}(\mathrm{poly}(d, H)/\varepsilon^2)$ samples, no $|\mathcal{S}|, |\mathcal{A}|$.
2. **(R) alone:** provably hopeless — $2^{\min(d,H)}$ lower bound.
3. **Online RL (no generative model):** even (C) is not enough — need an additional *model-based* structural assumption (the Linear MDP).

Today's Plan

Generative Model: Recap and Question

Generative model: a simulator that, given any (s, a, h), returns

$$r \sim r_h(s, a), \qquad s' \sim P_h(\cdot \mid s, a).$$

Lectures 4–5 (tabular): $\widetilde{O}(SAH^3/\varepsilon^2)$ samples suffice for an ε-optimal policy.

Generative Model: Recap and Question

Generative model: a simulator that, given any (s, a, h), returns

$$r \sim r_h(s, a), \qquad s' \sim P_h(\cdot \mid s, a).$$

Lectures 4–5 (tabular): $\widetilde{O}(SAH^3/\varepsilon^2)$ samples suffice for an ε-optimal policy.

Question

With linear function approximation, can we replace SA by d?
Sample complexity $= \mathrm{poly}(d, H, 1/\varepsilon)$, independent of $|\mathcal{S}|, |\mathcal{A}|$?

Generative Model: Recap and Question

Generative model: a simulator that, given any (s, a, h), returns

$$r \sim r_h(s, a), \qquad s' \sim P_h(\cdot \mid s, a).$$

Lectures 4–5 (tabular): $\widetilde{O}(SAH^3/\varepsilon^2)$ samples suffice for an ε-optimal policy.

Question

With linear function approximation, can we replace SA by d?
Sample complexity $= \mathrm{poly}(d, H, 1/\varepsilon)$, independent of $|\mathcal{S}|, |\mathcal{A}|$?

Answer. Yes, under Bellman completeness (C). The generative model lets us choose query points freely — exploration is trivial, only estimation matters.

LSVI with a Generative Model

Setup. Use Kiefer–Wolfowitz to obtain a distribution $\rho^\star$ on $\mathcal{S} \times \mathcal{A}$ supported on $m \le d(d+1)/2$ points $\{(s_i, a_i)\}_{i=1}^m$.

Proportional allocation. Given a per-stage budget n, query the generative model at (s_i, a_i) exactly $n_i := \lceil n\, \rho^\star(s_i, a_i) \rceil$ times. Total queries per stage: $\sum_i n_i \le n + m \le n + d^2$.

LSVI (generative model, deterministic allocation)

$\widehat{Q}_{H+1}(\cdot) \equiv 0$. For $h = H, H-1, \ldots, 1$, run least-squares on all collected tuples:

$$\widehat{w}_h \;=\; \arg\min_{w} \sum_{i=1}^{m} \sum_{j=1}^{n_i} \Big(\langle \phi(s_i, a_i), w \rangle \;-\; y_h^{i,j} \Big)^2,$$

$$\text{where} \quad y_h^{i,j} \;:=\; r_h^{i,j} \;+\; \max_{a'} \Big\langle \phi(s_h'^{,i,j}, a'), \widehat{w}_{h+1} \Big\rangle,$$

and $r_h^{i,j}, s_h'^{,i,j}$ are the j-th sample at (s_i, a_i). Output greedy policy $\widehat{\pi}_h(s) = \arg\max_a \langle \phi(s, a), \widehat{w}_h \rangle$.

Note: the $+m \le +d^2$ rounding overhead is negligible whenever $n \gg d^2$.

Why Bellman Completeness Is Exactly What LSVI Needs

Each individual regression target satisfies

$$\mathbb{E}[y_h^{i,j} \mid s_i, a_i] \;=\; (\mathcal{T}_h \widehat{Q}_{h+1})(s_i, a_i).$$

Why Bellman Completeness Is Exactly What LSVI Needs

Each individual regression target satisfies

$$\mathbb{E}[y_h^{i,j} \mid s_i, a_i] \;=\; (\mathcal{T}_h \widehat{Q}_{h+1})(s_i, a_i).$$

Under **(C)**, $\mathcal{T}_h \widehat{Q}_{h+1} \in \mathcal{F}$, so there exists $w_h^\star \in \mathbb{R}^d$ with

$$\mathbb{E}[y_h^{i,j}] \;=\; \langle \phi(s_i, a_i),\, w_h^\star \rangle \,.$$

Why Bellman Completeness Is Exactly What LSVI Needs

Each individual regression target satisfies

$$\mathbb{E}[y_h^{i,j} \mid s_i, a_i] \;=\; (\mathcal{T}_h \widehat{Q}_{h+1})(s_i, a_i).$$

Under **(C)**, $\mathcal{T}_h \widehat{Q}_{h+1} \in \mathcal{F}$, so there exists $w_h^\star \in \mathbb{R}^d$ with

$$\mathbb{E}[y_h^{i,j}] \;=\; \langle \phi(s_i, a_i),\, w_h^\star \rangle \,.$$

$\Rightarrow$ standard linear-regression analysis applies: $\widehat{w}_h - w_h^\star$ concentrates at the parametric rate.

Why Bellman Completeness Is Exactly What LSVI Needs

Each individual regression target satisfies

$$\mathbb{E}[y_h^{i,j} \mid s_i, a_i] \;=\; (\mathcal{T}_h \widehat{Q}_{h+1})(s_i, a_i).$$

Under **(C)**, $\mathcal{T}_h \widehat{Q}_{h+1} \in \mathcal{F}$, so there exists $w_h^\star \in \mathbb{R}^d$ with

$$\mathbb{E}[y_h^{i,j}] \;=\; \langle \phi(s_i, a_i),\, w_h^\star \rangle \,.$$

$\Rightarrow$ standard linear-regression analysis applies: $\widehat{w}_h - w_h^\star$ concentrates at the parametric rate.

Without (C), even fixing h this fails

Under realizability (R) alone, $\mathcal{T}_h \widehat{Q}_{h+1}$ might *not* lie in $\mathcal{F}$, so the regression target has irreducible misspecification.

The misspecification compounds across H stages, leading to vacuous bounds.

Choosing the Core Set: G-Optimal Design

G-optimal design (Kiefer–Wolfowitz)

There exists a distribution $\rho^\star$ supported on at most $d(d+1)/2$ points such that

$$\max_{(s,a)} \|\phi(s,a)\|^2_{\Lambda(\rho^\star)^{-1}} \;\leq\; d, \qquad \Lambda(\rho) := \mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top].$$

Choosing the Core Set: G-Optimal Design

G-optimal design (Kiefer–Wolfowitz)

There exists a distribution $\rho^\star$ supported on at most $d(d+1)/2$ points such that

$$\max_{(s,a)} \|\phi(s,a)\|^2_{\Lambda(\rho^\star)^{-1}} \;\le\; d, \qquad \Lambda(\rho) := \mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top].$$

Algorithmic version. Allocate $n_i = \lceil n\,\rho^\star(s_i,a_i)\rceil$ deterministic queries to each support point (*not* i.i.d. from $\rho^\star$). Then

$$\Lambda_{\mathcal{D}} \;=\; \sum_i n_i\,\phi_i\phi_i^\top \;\succeq\; n\,\Lambda(\rho^\star) \quad\Rightarrow\quad \max_{(s,a)} \|\phi(s,a)\|^2_{\Lambda_{\mathcal{D}}^{-1}} \;\le\; \tfrac{d}{n}.$$

Choosing the Core Set: G-Optimal Design

G-optimal design (Kiefer–Wolfowitz)

There exists a distribution $\rho^\star$ supported on at most $d(d+1)/2$ points such that

$$\max_{(s,a)} \|\phi(s,a)\|^2_{\Lambda(\rho^\star)^{-1}} \;\le\; d, \qquad \Lambda(\rho) := \mathbb{E}_{(s,a)\sim\rho}[\phi(s,a)\phi(s,a)^\top].$$

Algorithmic version. Allocate $n_i = \lceil n\,\rho^\star(s_i,a_i)\rceil$ deterministic queries to each support point (*not* i.i.d. from $\rho^\star$). Then

$$\Lambda_{\mathcal{D}} \;=\; \sum_i n_i\,\phi_i\phi_i^\top \;\succeq\; n\,\Lambda(\rho^\star) \quad\Rightarrow\quad \max_{(s,a)} \|\phi(s,a)\|^2_{\Lambda_{\mathcal{D}}^{-1}} \;\le\; \tfrac{d}{n}.$$

Key consequence

Prediction error of least squares at *any* $\phi(s,a)$ is bounded by the in-sample regression error scaled by $\sqrt{d/n}$ — eliminating the $|\mathcal{S}|\cdot|\mathcal{A}|$ dependence. Rounding $\lceil\cdot\rceil$ adds at most d^2 extra samples per stage, negligible whenever $n \gg d^2$.

The Per-Step Bellman Error

Key quantity. Define the per-step Bellman error of stage h:

$$\zeta_h := \left\| \widehat{Q}_h - \mathcal{T}_h \widehat{Q}_{h+1} \right\|_\infty = \sup_{(s,a)} \left| \widehat{Q}_h(s,a) - (\mathcal{T}_h \widehat{Q}_{h+1})(s,a) \right|.$$

The Per-Step Bellman Error

Key quantity. Define the per-step Bellman error of stage h:

$$\zeta_h \;:=\; \left\| \widehat{Q}_h \;-\; \mathcal{T}_h \widehat{Q}_{h+1} \right\|_\infty \;=\; \sup_{(s,a)} \left| \widehat{Q}_h(s,a) - (\mathcal{T}_h \widehat{Q}_{h+1})(s,a) \right|.$$

Sanity check: where did the reward go?

Both terms in $\Delta_h(s,a) := \widehat{Q}_h(s,a) - (\mathcal{T}_h \widehat{Q}_{h+1})(s,a)$ contain the reward:

- $\mathcal{T}_h \widehat{Q}_{h+1}(s,a) = r_h(s,a) + \mathbb{E}_{s'}[\max_{a'} \widehat{Q}_{h+1}(s',a')]$ (reward explicit).
- $\widehat{Q}_h(s,a) = \langle \phi(s,a), \widehat{w}_h \rangle$ where $\widehat{w}_h$ was fit to targets $y_h^{i,j} = r_h^{i,j} + \max_{a'} \widehat{Q}_{h+1}(s'^{,i,j}, a')$ (reward absorbed into $\widehat{w}_h$).

Under (C), $\mathcal{T}_h \widehat{Q}_{h+1} = \langle \phi, w_h^\star \rangle$ for the unique $w_h^\star$, so

$$\Delta_h(s,a) \;=\; \langle \phi(s,a),\, \widehat{w}_h - w_h^\star \rangle \,.$$

The reward cancels — Δ_h is the linear-regression error in the coefficient $\widehat{w}_h$. $\Delta_h \to 0$ as $\widehat{w}_h \to w_h^\star$.

The Per-Step Bellman Error

Key quantity. Define the per-step Bellman error of stage h:

$$\zeta_h \;:=\; \left\| \widehat{Q}_h \;-\; \mathcal{T}_h \widehat{Q}_{h+1} \right\|_\infty \;=\; \sup_{(s,a)} \left| \widehat{Q}_h(s,a) - (\mathcal{T}_h \widehat{Q}_{h+1})(s,a) \right|.$$

Sanity check: where did the reward go?

Both terms in $\Delta_h(s,a) := \widehat{Q}_h(s,a) - (\mathcal{T}_h \widehat{Q}_{h+1})(s,a)$ contain the reward:

- $\mathcal{T}_h \widehat{Q}_{h+1}(s,a) = r_h(s,a) + \mathbb{E}_{s'}[\max_{a'} \widehat{Q}_{h+1}(s',a')]$ (reward explicit).
- $\widehat{Q}_h(s,a) = \langle \phi(s,a), \widehat{w}_h \rangle$ where $\widehat{w}_h$ was fit to targets $y_h^{i,j} = r_h^{i,j} + \max_{a'} \widehat{Q}_{h+1}(s'^{,i,j}, a')$ (reward absorbed into $\widehat{w}_h$).

Under (C), $\mathcal{T}_h \widehat{Q}_{h+1} = \langle \phi, w_h^\star \rangle$ for the unique $w_h^\star$, so

$$\Delta_h(s,a) \;=\; \langle \phi(s,a),\, \widehat{w}_h - w_h^\star \rangle \,.$$

The reward cancels — Δ_h is the linear-regression error in the coefficient $\widehat{w}_h$. $\Delta_h \to 0$ as $\widehat{w}_h \to w_h^\star$.

Strategy. (1) Bound $V_0^\star - V_0^{\widehat{\pi}}$ by $\sum_h \zeta_h$ (Lemma 1, model-free). (2) Bound $\left\| \widehat{w}_h - w_h^\star \right\|_{\Lambda_\mathcal{D}}$ via a cover argument under (C) (Lemma 2). (3) Convert to ℓ_∞ via Cauchy–Schwarz + Kiefer–Wolfowitz.

Lemma 1 (Error Propagation): Statement

Lemma (Approximate VI bound)

Let $\widehat{Q}_1, \ldots, \widehat{Q}_H$ be *any* sequence of functions, $\widehat{V}_h(s) := \max_a \widehat{Q}_h(s,a)$, $\widehat{\pi}_h(s) := \arg\max_a \widehat{Q}_h(s,a)$. Define the (deterministic) Bellman residual

$$\Delta_h(s,a) \;:=\; \widehat{Q}_h(s,a) \;-\; (\mathcal{T}_h \widehat{Q}_{h+1})(s,a), \qquad \zeta_h := \|\Delta_h\|_\infty .$$

Then for any initial state s_0,

$$V_0^\star(s_0) - V_0^{\widehat{\pi}}(s_0) \;\leq\; 2\sum_{h=1}^{H} \zeta_h.$$

Lemma 1 (Error Propagation): Statement

Lemma (Approximate VI bound)

Let $\widehat{Q}_1, \ldots, \widehat{Q}_H$ be *any* sequence of functions, $\widehat{V}_h(s) := \max_a \widehat{Q}_h(s,a)$, $\widehat{\pi}_h(s) := \arg\max_a \widehat{Q}_h(s,a)$. Define the (deterministic) Bellman residual

$$\Delta_h(s,a) \;:=\; \widehat{Q}_h(s,a) \;-\; (\mathcal{T}_h \widehat{Q}_{h+1})(s,a), \qquad \zeta_h := \|\Delta_h\|_\infty .$$

Then for any initial state s_0,

$$V_0^\star(s_0) - V_0^{\widehat{\pi}}(s_0) \;\le\; 2\sum_{h=1}^{H} \zeta_h.$$

Important: Δ_h is a deterministic function

The operator $\mathcal{T}_h$ uses the *true* dynamics P_h. Once $\widehat{Q}_h, \widehat{Q}_{h+1}$ are fixed, $\Delta_h(s,a)$ is just a number for each (s,a) — no expectation over data.

The point of Bellman completeness is to make ζ_h *small*. Lemma 2 will show this.

Lemma 1: Proof

Decompose $V_0^\star - V_0^{\widehat{\pi}} = \underbrace{(V_0^\star - \widehat{V}_0)}_{\text{(I)}} + \underbrace{(\widehat{V}_0 - V_0^{\widehat{\pi}})}_{\text{(II)}}$. Bound each by $\sum_h \zeta_h$.

(I) Greediness $\Rightarrow \widehat{V}_h(s) \geq \widehat{Q}_h(s, \pi_h^\star(s))$. So

$$\begin{aligned} V_h^\star(s) - \widehat{V}_h(s) &\leq Q_h^\star(s, \pi_h^\star(s)) - \widehat{Q}_h(s, \pi_h^\star(s)) \\ &= \left[r_h + P_h V_{h+1}^\star\right](s, \pi_h^\star(s)) - \left[\mathcal{T}_h \widehat{Q}_{h+1} + \Delta_h\right](s, \pi_h^\star(s)) \\ &= P_h\big(V_{h+1}^\star - \widehat{V}_{h+1}\big)(s, \pi_h^\star(s)) - \Delta_h(s, \pi_h^\star(s)). \end{aligned}$$

Take expectation under $\pi^\star$ and unroll: $V_0^\star - \widehat{V}_0 \leq \sum_h \mathbb{E}_{\pi^\star}[-\Delta_h] \leq \sum_h \zeta_h$.

Lemma 1: Proof

Decompose $V_0^\star - V_0^{\widehat{\pi}} = \underbrace{(V_0^\star - \widehat{V}_0)}_{\text{(I)}} + \underbrace{(\widehat{V}_0 - V_0^{\widehat{\pi}})}_{\text{(II)}}$. Bound each by $\sum_h \zeta_h$.

(I) Greediness $\Rightarrow \widehat{V}_h(s) \geq \widehat{Q}_h(s, \pi_h^\star(s))$. So

$$\begin{aligned} V_h^\star(s) - \widehat{V}_h(s) &\leq Q_h^\star(s, \pi_h^\star(s)) - \widehat{Q}_h(s, \pi_h^\star(s)) \\ &= \left[r_h + P_h V_{h+1}^\star\right](s, \pi_h^\star(s)) - \left[\mathcal{T}_h \widehat{Q}_{h+1} + \Delta_h\right](s, \pi_h^\star(s)) \\ &= P_h\big(V_{h+1}^\star - \widehat{V}_{h+1}\big)(s, \pi_h^\star(s)) - \Delta_h(s, \pi_h^\star(s)). \end{aligned}$$

Take expectation under $\pi^\star$ and unroll: $V_0^\star - \widehat{V}_0 \leq \sum_h \mathbb{E}_{\pi^\star}[-\Delta_h] \leq \sum_h \zeta_h$.

(II) By definition $\widehat{V}_h(s) = \widehat{Q}_h(s, \widehat{\pi}_h(s))$, so

$$\begin{aligned} \widehat{V}_h(s) - V_h^{\widehat{\pi}}(s) &= \widehat{Q}_h(s, \widehat{\pi}_h(s)) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \\ &= \left[\mathcal{T}_h \widehat{Q}_{h+1} + \Delta_h\right](s, \widehat{\pi}_h(s)) - \left[r_h + P_h V_{h+1}^{\widehat{\pi}}\right](s, \widehat{\pi}_h(s)) \\ &= \Delta_h(s, \widehat{\pi}_h(s)) + P_h\big(\widehat{V}_{h+1} - V_{h+1}^{\widehat{\pi}}\big)(s, \widehat{\pi}_h(s)). \end{aligned}$$

Take expectation under $\widehat{\pi}$ and unroll: $\widehat{V}_0 - V_0^{\widehat{\pi}} = \sum_h \mathbb{E}_{\widehat{\pi}}[\Delta_h] \leq \sum_h \zeta_h$.

Lemma 1: Proof

Decompose $V_0^\star - V_0^{\widehat{\pi}} = \underbrace{(V_0^\star - \widehat{V}_0)}_{\text{(I)}} + \underbrace{(\widehat{V}_0 - V_0^{\widehat{\pi}})}_{\text{(II)}}$. Bound each by $\sum_h \zeta_h$.

(I) Greediness $\Rightarrow \widehat{V}_h(s) \geq \widehat{Q}_h(s, \pi_h^\star(s))$. So

$$\begin{aligned} V_h^\star(s) - \widehat{V}_h(s) &\leq Q_h^\star(s, \pi_h^\star(s)) - \widehat{Q}_h(s, \pi_h^\star(s)) \\ &= \left[r_h + P_h V_{h+1}^\star\right](s, \pi_h^\star(s)) - \left[\mathcal{T}_h \widehat{Q}_{h+1} + \Delta_h\right](s, \pi_h^\star(s)) \\ &= P_h\big(V_{h+1}^\star - \widehat{V}_{h+1}\big)(s, \pi_h^\star(s)) - \Delta_h(s, \pi_h^\star(s)). \end{aligned}$$

Take expectation under $\pi^\star$ and unroll: $V_0^\star - \widehat{V}_0 \leq \sum_h \mathbb{E}_{\pi^\star}[-\Delta_h] \leq \sum_h \zeta_h$.

(II) By definition $\widehat{V}_h(s) = \widehat{Q}_h(s, \widehat{\pi}_h(s))$, so

$$\begin{aligned} \widehat{V}_h(s) - V_h^{\widehat{\pi}}(s) &= \widehat{Q}_h(s, \widehat{\pi}_h(s)) - Q_h^{\widehat{\pi}}(s, \widehat{\pi}_h(s)) \\ &= \left[\mathcal{T}_h \widehat{Q}_{h+1} + \Delta_h\right](s, \widehat{\pi}_h(s)) - \left[r_h + P_h V_{h+1}^{\widehat{\pi}}\right](s, \widehat{\pi}_h(s)) \\ &= \Delta_h(s, \widehat{\pi}_h(s)) + P_h\big(\widehat{V}_{h+1} - V_{h+1}^{\widehat{\pi}}\big)(s, \widehat{\pi}_h(s)). \end{aligned}$$

Take expectation under $\widehat{\pi}$ and unroll: $\widehat{V}_0 - V_0^{\widehat{\pi}} = \sum_h \mathbb{E}_{\widehat{\pi}}[\Delta_h] \leq \sum_h \zeta_h$.

Combining (I) + (II): $V_0^\star - V_0^{\widehat{\pi}} \leq 2 \sum_h \zeta_h$.

Aside: Linear Regression Concentration (Setup)

Generic problem: data $\{(\phi_i, y_i)\}_{i=1}^n$ with $\phi_i \in \mathbb{R}^d$ *deterministic* (or fixed), and

$$y_i \;=\; \langle \phi_i, w^\star \rangle + \eta_i, \qquad \eta_i \text{ mean-zero, } \sigma\text{-sub-Gaussian, indep. across } i.$$

Assume the design $\Lambda := \sum_i \phi_i \phi_i^\top$ is invertible. Plain least squares:

$$\widehat{w} \;=\; \arg\min_w \sum_{i=1}^n (\langle \phi_i, w \rangle - y_i)^2 \;=\; \Lambda^{-1} \sum_{i=1}^n \phi_i y_i.$$

Aside: Linear Regression Concentration (Setup)

Generic problem: data $\{(\phi_i, y_i)\}_{i=1}^n$ with $\phi_i \in \mathbb{R}^d$ *deterministic* (or fixed), and

$$y_i \;=\; \langle \phi_i, w^\star \rangle + \eta_i, \qquad \eta_i \text{ mean-zero, } \sigma\text{-sub-Gaussian, indep. across } i.$$

Assume the design $\Lambda := \sum_i \phi_i \phi_i^\top$ is invertible. Plain least squares:

$$\widehat{w} \;=\; \arg\min_w \sum_{i=1}^n (\langle \phi_i, w \rangle - y_i)^2 \;=\; \Lambda^{-1} \sum_{i=1}^n \phi_i y_i.$$

Decomposition. Plugging in $y_i = \langle \phi_i, w^\star \rangle + \eta_i$:

$$\widehat{w} - w^\star \;=\; \Lambda^{-1} \sum_i \phi_i \eta_i.$$

Aside: Linear Regression Concentration (Setup)

Generic problem: data $\{(\phi_i, y_i)\}_{i=1}^n$ with $\phi_i \in \mathbb{R}^d$ *deterministic* (or fixed), and

$$y_i \;=\; \langle \phi_i, w^\star \rangle + \eta_i, \qquad \eta_i \text{ mean-zero, } \sigma\text{-sub-Gaussian, indep. across } i.$$

Assume the design $\Lambda := \sum_i \phi_i \phi_i^\top$ is invertible. Plain least squares:

$$\widehat{w} \;=\; \arg\min_w \sum_{i=1}^n (\langle \phi_i, w \rangle - y_i)^2 \;=\; \Lambda^{-1} \sum_{i=1}^n \phi_i y_i.$$

Decomposition. Plugging in $y_i = \langle \phi_i, w^\star \rangle + \eta_i$:

$$\widehat{w} - w^\star \;=\; \Lambda^{-1} \sum_i \phi_i \eta_i.$$

Goal. Bound the elliptical norm $\|\widehat{w} - w^\star\|_\Lambda$.

Aside: Direction-by-Direction Sub-Gaussian Bound

Key trick. For any unit vector $u \in \mathbb{R}^d$,

$$\left\langle u, \Lambda^{1/2}(\widehat{w} - w^\star) \right\rangle = \left\langle u, \Lambda^{-1/2} \sum_i \phi_i \eta_i \right\rangle = \sum_{i=1}^n \underbrace{(u^\top \Lambda^{-1/2} \phi_i)}_{\text{coefficient}} \cdot \eta_i .$$

This is a sum of independent sub-Gaussians $\Rightarrow$ sub-Gaussian with proxy

$$\sigma^2 \sum_i (u^\top \Lambda^{-1/2} \phi_i)^2 = \sigma^2 \, u^\top \Lambda^{-1/2} \Big(\sum_i \phi_i \phi_i^\top \Big) \Lambda^{-1/2} u = \sigma^2 \, u^\top u = \sigma^2,$$

since $\sum_i \phi_i \phi_i^\top = \Lambda$.

Aside: Direction-by-Direction Sub-Gaussian Bound

Key trick. For any unit vector $u \in \mathbb{R}^d$,

$$\left\langle u, \Lambda^{1/2}(\widehat{w} - w^\star) \right\rangle = \left\langle u, \Lambda^{-1/2} \sum_i \phi_i \eta_i \right\rangle = \sum_{i=1}^n \underbrace{(u^\top \Lambda^{-1/2} \phi_i)}_{\text{coefficient}} \cdot \eta_i.$$

This is a sum of independent sub-Gaussians $\Rightarrow$ sub-Gaussian with proxy

$$\sigma^2 \sum_i (u^\top \Lambda^{-1/2} \phi_i)^2 = \sigma^2 \, u^\top \Lambda^{-1/2} \Big(\sum_i \phi_i \phi_i^\top \Big) \Lambda^{-1/2} u = \sigma^2 \, u^\top u = \sigma^2,$$

since $\sum_i \phi_i \phi_i^\top = \Lambda$.

Hoeffding bound. For any *fixed* unit u, with prob. $\geq 1 - \delta$,

$$\left| \left\langle u, \Lambda^{1/2}(\widehat{w} - w^\star) \right\rangle \right| \leq \sigma \sqrt{2 \log(2/\delta)}.$$

Aside: Cover Argument $\Rightarrow$ Uniform Bound

We want $\|\widehat{w} - w^\star\|_\Lambda = \sup_{\|u\|=1} \left\langle u, \Lambda^{1/2}(\widehat{w} - w^\star) \right\rangle$, i.e., the sup over *all* directions.

ε-net. The unit sphere in $\mathbb{R}^d$ admits an ε-net $\mathcal{N}_\varepsilon$ with $|\mathcal{N}_\varepsilon| \le (3/\varepsilon)^d$ (covering by balls of radius ε).

Discretization. For any unit u, pick $u_0 \in \mathcal{N}_\varepsilon$ with $\|u - u_0\| \le \varepsilon$. Then

$$\langle u, x \rangle \;=\; \langle u_0, x \rangle + \langle u - u_0, x \rangle \;\le\; \langle u_0, x \rangle + \varepsilon \|x\| .$$

Taking sup and rearranging: $\sup_{\|u\|=1} \langle u, x \rangle \le \frac{1}{1-\varepsilon} \sup_{u_0 \in \mathcal{N}_\varepsilon} \langle u_0, x \rangle$.

Aside: Cover Argument $\Rightarrow$ Uniform Bound

We want $\|\widehat{w} - w^\star\|_\Lambda = \sup_{\|u\|=1} \left\langle u, \Lambda^{1/2}(\widehat{w} - w^\star) \right\rangle$, i.e., the sup over *all* directions.

ε-net. The unit sphere in $\mathbb{R}^d$ admits an ε-net $\mathcal{N}_\varepsilon$ with $|\mathcal{N}_\varepsilon| \le (3/\varepsilon)^d$ (covering by balls of radius ε).

Discretization. For any unit u, pick $u_0 \in \mathcal{N}_\varepsilon$ with $\|u - u_0\| \le \varepsilon$. Then

$$\langle u, x \rangle \;=\; \langle u_0, x \rangle + \langle u - u_0, x \rangle \;\le\; \langle u_0, x \rangle + \varepsilon \|x\| .$$

Taking sup and rearranging: $\sup_{\|u\|=1} \langle u, x \rangle \le \frac{1}{1-\varepsilon} \sup_{u_0 \in \mathcal{N}_\varepsilon} \langle u_0, x \rangle$.

Union bound. Apply the previous slide's bound at each $u_0 \in \mathcal{N}_\varepsilon$, with confidence $\delta/|\mathcal{N}_\varepsilon|$. With $\varepsilon = 1/2$ (so $|\mathcal{N}_{1/2}| \le 6^d$):

Lemma (Linear regression concentration)

With probability at least $1 - \delta$,

$$\boxed{\|\widehat{w} - w^\star\|_\Lambda \;\le\; 2\sigma\sqrt{2\big(d \log 6 + \log(2/\delta)\big)} \;=\; \widetilde{O}(\sigma\sqrt{d}).}$$

Aside: Cover Argument $\Rightarrow$ Uniform Bound

We want $\|\widehat{w} - w^\star\|_\Lambda = \sup_{\|u\|=1} \left\langle u, \Lambda^{1/2}(\widehat{w} - w^\star) \right\rangle$, i.e., the sup over *all* directions.

ε-net. The unit sphere in $\mathbb{R}^d$ admits an ε-net $\mathcal{N}_\varepsilon$ with $|\mathcal{N}_\varepsilon| \le (3/\varepsilon)^d$ (covering by balls of radius ε).

Discretization. For any unit u, pick $u_0 \in \mathcal{N}_\varepsilon$ with $\|u - u_0\| \le \varepsilon$. Then

$$\langle u, x \rangle \;=\; \langle u_0, x \rangle + \langle u - u_0, x \rangle \;\le\; \langle u_0, x \rangle + \varepsilon \|x\| .$$

Taking sup and rearranging: $\sup_{\|u\|=1} \langle u, x \rangle \le \frac{1}{1-\varepsilon} \sup_{u_0 \in \mathcal{N}_\varepsilon} \langle u_0, x \rangle$.

Union bound. Apply the previous slide's bound at each $u_0 \in \mathcal{N}_\varepsilon$, with confidence $\delta/|\mathcal{N}_\varepsilon|$. With $\varepsilon = 1/2$ (so $|\mathcal{N}_{1/2}| \le 6^d$):

Lemma (Linear regression concentration)

With probability at least $1 - \delta$,

$$\boxed{\|\widehat{w} - w^\star\|_\Lambda \;\le\; 2\sigma\sqrt{2\big(d \log 6 + \log(2/\delta)\big)} \;=\; \widetilde{O}(\sigma\sqrt{d}).}$$

$\Rightarrow$ least-squares error is $\widetilde{O}(\sigma\sqrt{d})$ in elliptical norm.

Lemma 2: Regression Error for Our LSVI Setting

Lemma (Regression error under (C))

Fix stage h, assume $\widehat{Q}_{h+1}$ is independent of the stage-h samples, and use the proportional KW allocation (so $\Lambda_{\mathcal{D}} \succeq n\Lambda(\rho^\star)$). Under (C), there exists $w_h^\star \in \mathbb{R}^d$ with $\langle \phi, w_h^\star \rangle = \mathcal{T}_h \widehat{Q}_{h+1}$, and with probability $\geq 1-\delta$,

$$\|\widehat{w}_h - w_h^\star\|_{\Lambda_{\mathcal{D}}} \;\leq\; c\, H \sqrt{d + \log(1/\delta)},$$

for an absolute constant $c > 0$.

Lemma 2: Regression Error for Our LSVI Setting

Lemma (Regression error under (C))

Fix stage h, assume $\widehat{Q}_{h+1}$ is independent of the stage-h samples, and use the proportional KW allocation (so $\Lambda_{\mathcal{D}} \succeq n\Lambda(\rho^\star)$). Under (C), there exists $w_h^\star \in \mathbb{R}^d$ with $\langle \phi, w_h^\star \rangle = \mathcal{T}_h \widehat{Q}_{h+1}$, and with probability $\geq 1-\delta$,

$$\|\widehat{w}_h - w_h^\star\|_{\Lambda_{\mathcal{D}}} \;\leq\; c\, H \sqrt{d + \log(1/\delta)},$$

for an absolute constant $c > 0$.

Proof.

- By (C), $\mathcal{T}_h \widehat{Q}_{h+1} \in \mathcal{F} \Rightarrow w_h^\star$ exists with $\langle \phi, w_h^\star \rangle = \mathcal{T}_h \widehat{Q}_{h+1}$.
- Each individual target satisfies $\mathbb{E}[y_h^{i,j}] = \langle \phi(s_i, a_i), w_h^\star \rangle$ *exactly*.
- Noise $\eta_h^{i,j} := y_h^{i,j} - \langle \phi(s_i, a_i), w_h^\star \rangle$ takes values in $[-H, H]$, hence is sub-Gaussian with proxy $\sigma = H$.
- Apply the regression concentration lemma with $\Lambda = \Lambda_{\mathcal{D}}$. □

Lemma 2: Regression Error for Our LSVI Setting

Lemma (Regression error under (C))

Fix stage h, assume $\widehat{Q}_{h+1}$ is independent of the stage-h samples, and use the proportional KW allocation (so $\Lambda_{\mathcal{D}} \succeq n\Lambda(\rho^\star)$). Under (C), there exists $w_h^\star \in \mathbb{R}^d$ with $\langle \phi, w_h^\star \rangle = \mathcal{T}_h \widehat{Q}_{h+1}$, and with probability $\geq 1-\delta$,

$$\|\widehat{w}_h - w_h^\star\|_{\Lambda_{\mathcal{D}}} \;\leq\; c\,H\sqrt{d + \log(1/\delta)},$$

for an absolute constant $c > 0$.

Proof.

- By (C), $\mathcal{T}_h \widehat{Q}_{h+1} \in \mathcal{F} \Rightarrow w_h^\star$ exists with $\langle \phi, w_h^\star \rangle = \mathcal{T}_h \widehat{Q}_{h+1}$.
- Each individual target satisfies $\mathbb{E}[y_h^{i,j}] = \langle \phi(s_i, a_i), w_h^\star \rangle$ *exactly*.
- Noise $\eta_h^{i,j} := y_h^{i,j} - \langle \phi(s_i, a_i), w_h^\star \rangle$ takes values in $[-H, H]$, hence is sub-Gaussian with proxy $\sigma = H$.
- Apply the regression concentration lemma with $\Lambda = \Lambda_{\mathcal{D}}$. □

Note: the bound is independent of n in the $\Lambda_{\mathcal{D}}$-norm; the n dependence enters when we convert to a pointwise ℓ_∞ bound via Kiefer–Wolfowitz.

Putting the Lemmas Together

Per stage, allocate n samples proportionally to $\rho^\star$ (rounding overhead $\leq d^2$). By Kiefer–Wolfowitz, $\|\phi(s,a)\|^2_{\Lambda_{\mathcal{D}}^{-1}} \leq d/n$ uniformly.

Lemma 2 + Cauchy–Schwarz. For any (s,a),

$$\left| \langle \phi(s,a), \widehat{w}_h - w_h^\star \rangle \right| \leq \|\phi(s,a)\|_{\Lambda_{\mathcal{D}}^{-1}} \cdot \|\widehat{w}_h - w_h^\star\|_{\Lambda_{\mathcal{D}}} \leq \sqrt{\tfrac{d}{n}} \cdot \widetilde{O}(H\sqrt{d}).$$

So $\zeta_h \leq \widetilde{O}(Hd/\sqrt{n})$.

Lemma 1. $V_0^\star - V_0^{\widehat{\pi}} \leq 2\sum_h \zeta_h \leq \widetilde{O}(H^2 d/\sqrt{n})$.

Putting the Lemmas Together

Per stage, allocate n samples proportionally to $\rho^\star$ (rounding overhead $\le d^2$). By Kiefer–Wolfowitz, $\|\phi(s,a)\|^2_{\Lambda_{\mathcal{D}}^{-1}} \le d/n$ uniformly.

Lemma 2 + Cauchy–Schwarz. For any (s,a),

$$\left|\langle \phi(s,a), \widehat{w}_h - w_h^\star \rangle\right| \le \|\phi(s,a)\|_{\Lambda_{\mathcal{D}}^{-1}} \cdot \|\widehat{w}_h - w_h^\star\|_{\Lambda_{\mathcal{D}}} \le \sqrt{\tfrac{d}{n}} \cdot \widetilde{O}(H\sqrt{d}).$$

So $\zeta_h \le \widetilde{O}(Hd/\sqrt{n})$.

Lemma 1. $V_0^\star - V_0^{\widehat{\pi}} \le 2\sum_h \zeta_h \le \widetilde{O}(H^2 d/\sqrt{n})$.

Theorem (LSVI under Bellman completeness)

With $n = \widetilde{O}(d^2H^4/\varepsilon^2)$ samples per stage (rounding overhead $\le d^2$ negligible), with probability $\ge 1-\delta$, $V_0^\star - V_0^{\widehat{\pi}} \le \varepsilon$, using a total of

$$\boxed{n \cdot H \;=\; \widetilde{O}\left(\frac{d^2H^5}{\varepsilon^2}\log(1/\delta)\right)}$$

generative-model samples. No dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$.

Putting the Lemmas Together

Per stage, allocate n samples proportionally to $\rho^\star$ (rounding overhead $\leq d^2$). By Kiefer–Wolfowitz, $\|\phi(s,a)\|^2_{\Lambda_{\mathcal{D}}^{-1}} \leq d/n$ uniformly.

Lemma 2 + Cauchy–Schwarz. For any (s,a),

$$\left| \langle \phi(s,a), \widehat{w}_h - w_h^\star \rangle \right| \leq \|\phi(s,a)\|_{\Lambda_{\mathcal{D}}^{-1}} \cdot \|\widehat{w}_h - w_h^\star\|_{\Lambda_{\mathcal{D}}} \leq \sqrt{\tfrac{d}{n}} \cdot \widetilde{O}(H\sqrt{d}).$$

So $\zeta_h \leq \widetilde{O}(Hd/\sqrt{n})$.

Lemma 1. $V_0^\star - V_0^{\widehat{\pi}} \leq 2\sum_h \zeta_h \leq \widetilde{O}(H^2 d/\sqrt{n})$.

Theorem (LSVI under Bellman completeness)

With $n = \widetilde{O}(d^2H^4/\varepsilon^2)$ samples per stage (rounding overhead $\leq d^2$ negligible), with probability $\geq 1-\delta$, $V_0^\star - V_0^{\widehat{\pi}} \leq \varepsilon$, using a total of

$$\boxed{n \cdot H \;=\; \widetilde{O}\left(\frac{d^2H^5}{\varepsilon^2}\log(1/\delta)\right)}$$

generative-model samples. No dependence on $|\mathcal{S}|$ or $|\mathcal{A}|$.

Conditioning. Lemma 2 assumed $\widehat{Q}_{h+1}$ indep. of stage-h samples; achieved by using fresh samples per stage — hence the factor of H.

Realizability Alone Is Not Enough

What if we only assume (R), i.e., $Q^\star \in \mathcal{F}$ but $\mathcal{F}$ is *not* closed under $\mathcal{T}_h$?

Realizability Alone Is Not Enough

What if we only assume (R), i.e., $Q^\star \in \mathcal{F}$ but $\mathcal{F}$ is *not* closed under $\mathcal{T}_h$?

Theorem (Weisz, Amortila, Szepesvári 2021)

There exist MDPs with $Q_h^\star$ exactly linear in a d-dimensional feature map, such that any algorithm using a generative model requires at least $\Omega\left(2^{\min(d,H)}\right)$ samples to compute an ε-optimal policy, even for constant ε.

Realizability Alone Is Not Enough

What if we only assume (R), i.e., $Q^\star \in \mathcal{F}$ but $\mathcal{F}$ is *not* closed under $\mathcal{T}_h$?

Theorem (Weisz, Amortila, Szepesvári 2021)

There exist MDPs with $Q_h^\star$ exactly linear in a d-dimensional feature map, such that any algorithm using a generative model requires at least $\Omega\left(2^{\min(d,H)}\right)$ samples to compute an ε-optimal policy, even for constant ε.

Take-away. The gap between (R) and (C) is the difference between exponential and polynomial sample complexity.

Realizability Alone Is Not Enough

What if we only assume (R), i.e., $Q^\star \in \mathcal{F}$ but $\mathcal{F}$ is *not* closed under $\mathcal{T}_h$?

Theorem (Weisz, Amortila, Szepesvári 2021)

There exist MDPs with $Q^\star_h$ exactly linear in a d-dimensional feature map, such that any algorithm using a generative model requires at least $\Omega\left(2^{\min(d,H)}\right)$ samples to compute an ε-optimal policy, even for constant ε.

Take-away. The gap between (R) and (C) is the difference between exponential and polynomial sample complexity.

Moral

Bellman completeness — closure of $\mathcal{F}$ under the Bellman operator — is the right model-free abstraction for sample-efficient batch RL with linear function approximation.

Today's Plan

The Offline Setting

Generative model: we picked $\mathcal{D}$ via G-optimal design.
Offline: we are *given* a fixed dataset

$$\mathcal{D} = \left\{(s_h^i, a_h^i, r_h^i, s_{h+1}^i)\right\}_{i=1}^n, \quad (s_h^i, a_h^i) \sim \rho_h,$$

collected by some *unknown* behavior policy. No choice in queries.

The Offline Setting

Generative model: we picked $\mathcal{D}$ via G-optimal design.
Offline: we are *given* a fixed dataset

$$\mathcal{D} = \left\{(s_h^i, a_h^i, r_h^i, s_{h+1}^i)\right\}_{i=1}^n, \quad (s_h^i, a_h^i) \sim \rho_h,$$

collected by some *unknown* behavior policy. No choice in queries.

Question

How well can we estimate $Q^\star$ when the data distribution ρ_h is fixed and *not* chosen for our convenience?

The Offline Setting

Generative model: we picked $\mathcal{D}$ via G-optimal design.
Offline: we are *given* a fixed dataset

$$\mathcal{D} = \left\{(s_h^i, a_h^i, r_h^i, s_{h+1}^i)\right\}_{i=1}^n, \quad (s_h^i, a_h^i) \sim \rho_h,$$

collected by some *unknown* behavior policy. No choice in queries.

Question

How well can we estimate $Q^\star$ when the data distribution ρ_h is fixed and *not* chosen for our convenience?

Plan. Run vanilla LSVI on the offline data and bound the suboptimality in terms of how well ρ covers the feature directions visited by competing policies.

All-Policy Concentrability

Let $\Lambda_{\rho,h} := \mathbb{E}_{(s,a)\sim\rho_h}[\phi(s,a)\phi(s,a)^\top]$ be the offline data covariance at stage h.

All-policy concentrability C_{all}

$$C_{\text{all}} := \max_h \sup_\pi \mathbb{E}_{(s,a)\sim d_h^\pi}\left[\|\phi(s,a)\|^2_{\Lambda^{-1}_{\rho,h}} \right].$$

C_{all} measures the worst-case mismatch between the data covariance and the visitation covariance of any policy — not just $\pi^\star$.

All-Policy Concentrability

Let $\Lambda_{\rho,h} := \mathbb{E}_{(s,a)\sim\rho_h}[\phi(s,a)\phi(s,a)^\top]$ be the offline data covariance at stage h.

All-policy concentrability C_{all}

$$C_{\mathrm{all}} := \max_h \sup_\pi \mathbb{E}_{(s,a)\sim d_h^\pi}\left[\|\phi(s,a)\|^2_{\Lambda_{\rho,h}^{-1}}\right].$$

C_{all} measures the worst-case mismatch between the data covariance and the visitation covariance of any policy — not just $\pi^\star$.

Why all-policy? Standard offline LSVI gives a uniform Q-function estimate over $\mathcal{F}$; the worst direction we might evaluate is the worst direction *any* policy could exploit.

All-Policy Concentrability

Let $\Lambda_{\rho,h} := \mathbb{E}_{(s,a)\sim\rho_h}[\phi(s,a)\phi(s,a)^\top]$ be the offline data covariance at stage h.

All-policy concentrability C_{all}

$$C_{\mathrm{all}} := \max_h \sup_\pi \mathbb{E}_{(s,a)\sim d_h^\pi}\left[\|\phi(s,a)\|^2_{\Lambda_{\rho,h}^{-1}}\right].$$

C_{all} measures the worst-case mismatch between the data covariance and the visitation covariance of any policy — not just $\pi^\star$.

Why all-policy? Standard offline LSVI gives a uniform Q-function estimate over $\mathcal{F}$; the worst direction we might evaluate is the worst direction *any* policy could exploit.

Tabular analog

$C_{\mathrm{all}} = \sup_h \sup_{\pi,(s,a)} d_h^\pi(s,a)/\rho_h(s,a)$. Fails the moment ρ misses any reachable (s,a).

This is the "weak" offline result — in Lecture 6 we got around it with pessimism + single-policy concentrability.

Offline LSVI Under (C)

Algorithm. Same LSVI as the generative-model case, run on the offline data $\mathcal{D}$ (no G-optimal design, no bonus).

$$\widehat{w}_h \;=\; \Lambda_h^{-1} \sum_i \phi(s_h^i, a_h^i) \left[r_h^i + \max_{a'} \left\langle \phi(s_{h+1}^i, a'), \widehat{w}_{h+1} \right\rangle \right].$$

Output $\widehat{\pi}_h(s) := \arg\max_a \langle \phi(s,a), \widehat{w}_h \rangle$.

Offline LSVI Under (C)

Algorithm. Same LSVI as the generative-model case, run on the offline data $\mathcal{D}$ (no G-optimal design, no bonus).

$$\widehat{w}_h \;=\; \Lambda_h^{-1} \sum_i \phi(s_h^i, a_h^i) \big[r_h^i + \max_{a'} \big\langle \phi(s_{h+1}^i, a'), \widehat{w}_{h+1} \big\rangle \big].$$

Output $\widehat{\pi}_h(s) := \arg\max_a \langle \phi(s,a), \widehat{w}_h \rangle$.

Theorem (Offline LSVI under (C), informal)

Under Bellman completeness (C) with all-policy concentrability C_{all}, with probability $\geq 1-\delta$:

$$V_0^\star - V_0^{\widehat{\pi}} \;\leq\; \widetilde{O}\left(\sqrt{\frac{dH^4\, C_{\mathrm{all}}}{n}} \right).$$

Offline LSVI Under (C)

Algorithm. Same LSVI as the generative-model case, run on the offline data $\mathcal{D}$ (no G-optimal design, no bonus).

$$\widehat{w}_h \;=\; \Lambda_h^{-1} \sum_i \phi(s_h^i, a_h^i) \big[r_h^i + \max_{a'} \big\langle \phi(s_{h+1}^i, a'), \widehat{w}_{h+1} \big\rangle \big].$$

Output $\widehat{\pi}_h(s) := \arg\max_a \langle \phi(s,a), \widehat{w}_h \rangle$.

Theorem (Offline LSVI under (C), informal)

Under Bellman completeness (C) with all-policy concentrability C_{all}, with probability $\geq 1-\delta$:

$$V_0^\star - V_0^{\widehat{\pi}} \;\leq\; \widetilde{O}\left(\sqrt{\frac{dH^4\, C_{\mathrm{all}}}{n}} \right).$$

Why this is what (C) supports. The proof goes through exactly as the generative-model proof: regression target $\mathcal{T}_h \widehat{Q}_{h+1}$ is realizable in $\mathcal{F}$ by (C); the in-sample regression error is $\widetilde{O}(H\sqrt{d/n})$; transferring to a uniform ℓ_∞ bound costs C_{all} instead of d/n.

Summary

What we did today

- Linear FA *model-free*: $\mathcal{F} = \{\langle \phi, w \rangle\}$. Two assumptions: **(R)** $Q^\star \in \mathcal{F}$ and **(C)** $\mathcal{T}_h \mathcal{F} \subseteq \mathcal{F}$.
- **Generative model + (C):** LSVI with G-optimal design (samples allocated proportionally to $\rho^\star$), $\widetilde{O}(d^2 H^5/\varepsilon^2)$ samples. Two lemmas:
  1. Error propagation: $V_0^\star - V_0^{\widehat{\pi}} \le 2 \sum_h \zeta_h$.
  2. Regression error under (C): $\|\widehat{w}_h - w_h^\star\|_{\Lambda_\mathcal{D}} \le \widetilde{O}(H\sqrt{d})$ (cover argument).

  Combined via Cauchy–Schwarz + Kiefer–Wolfowitz: $\zeta_h \le \widetilde{O}(Hd/\sqrt{n})$.
- **(R) alone is not enough:** $\Omega(2^{\min(d,H)})$ lower bound.
- **Offline + (C):** vanilla LSVI gives $\widetilde{O}(\sqrt{dH^4 C_{\mathrm{all}}/n})$, all-policy concentrability. Pessimism would buy single-policy concentrability but already needs (L) — foreshadowing online.

Summary

What we did today

- Linear FA *model-free*: $\mathcal{F} = \{\langle \phi, w \rangle\}$. Two assumptions: **(R)** $Q^\star \in \mathcal{F}$ and **(C)** $\mathcal{T}_h \mathcal{F} \subseteq \mathcal{F}$.
- **Generative model + (C):** LSVI with G-optimal design (samples allocated proportionally to $\rho^\star$), $\widetilde{O}(d^2 H^5/\varepsilon^2)$ samples. Two lemmas:
  1. Error propagation: $V_0^\star - V_0^{\widehat{\pi}} \le 2 \sum_h \zeta_h$.
  2. Regression error under (C): $\|\widehat{w}_h - w_h^\star\|_{\Lambda_\mathcal{D}} \le \widetilde{O}(H\sqrt{d})$ (cover argument).

  Combined via Cauchy–Schwarz + Kiefer–Wolfowitz: $\zeta_h \le \widetilde{O}(Hd/\sqrt{n})$.
- **(R) alone is not enough:** $\Omega(2^{\min(d,H)})$ lower bound.
- **Offline + (C):** vanilla LSVI gives $\widetilde{O}(\sqrt{dH^4 C_{\mathrm{all}}/n})$, all-policy concentrability. Pessimism would buy single-policy concentrability but already needs (L) — foreshadowing online.

Next lecture

Online RL with no generative model. We'll see why even Bellman completeness is not enough, and introduce the Linear MDP — a model-based assumption that makes optimistic exploration possible.

Reading & References

Primary reference for today:

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

Foundational papers:

- Weisz, Amortila, Szepesvári, *Exponential lower bounds for planning in MDPs with linearly-realizable optimal action-value functions*, ALT 2021.
- Jin, Yang, Wang, *Is pessimism provably efficient for offline RL?*, NeurIPS 2021.

For next lecture:

- Jin, Yang, Wang, Jordan, *Provably efficient reinforcement learning with linear function approximation*, COLT 2020.
- Abbasi-Yadkori, Pál, Szepesvári, *Improved algorithms for linear stochastic bandits*, NeurIPS 2011.

Questions?