RL with Linear Features: When Does It Work & When Doesn't It Work?

Part 2: Linear BC & The Offline RL Case

CS 2284: Foundations of Reinforcement Learning

Kianté Brantley & Sham Kakade

Hi!

Agenda

Announcements

- HW1 due Monday.

Recap

- Regression + D-opt design.

Today

- Finish linear BC analysis.
- Offline RL

Recap

Least-Squares Value Iteration (LSVI)

Setting: Finite Horizon H, Generative Model (Simulator).

Algorithm: Backward Induction via Regression.

1. Initialize $\widehat{V}_H(s) = 0$.
2. For $h = H-1, \ldots, 0$:
   - **Collect Data:** Generate dataset $D_h = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$.
   - **Form Targets:** Compute regression targets using the *next* value function:

$$y_i = r_h(s_i, a_i) + \widehat{V}_{h+1}(s_i')$$

   - **Regression:** Solve for parameters $\widehat{\theta}_h$:

$$\widehat{\theta}_h \in \operatorname*{argmin}_{\theta \in \mathbb{R}^d} \sum_{i=1}^N \left(\theta^\top \phi(s_i, a_i) - y_i\right)^2$$

   - **Update:** Set $\widehat{Q}_h(s,a) = \widehat{\theta}_h^\top \phi(s,a)$ and $\widehat{V}_h(s) = \max_a \widehat{Q}_h(s,a)$.

$E[y_i \mid s, a]$

$= r[\theta_{h+1}]$

$= \operatorname*{argmax}_{a'} \hat{\theta}_{h+1} \phi(s', a')$

The Assumption Ladder

We can organize linear RL assumptions from weakest (hardest) to strongest (easiest).

(A) **Agnostic Approximation**
No realizability. $Q^\star$ is "close" to linear.
Status: Hard. Requires strong distribution assumptions.

(B) **Linear $Q^\star$ Realizability**
$Q^\star_h(s,a) = (\theta^\star_h)^\top \phi(s,a)$.
Status: **Insufficient.** *Exponential lower bounds exist.*

(C) **All-Policies Realizability**
$Q^\pi_h(s,a) = (\theta^\pi_h)^\top \phi(s,a)$ for **all** π.
Status: **Subtle.** *Fails offline even with perfect coverage.*

(D) **Linear Bellman Completeness**
$\mathcal{T}_h f \in \mathcal{F}$ for all $f \in \mathcal{F}$.
Status: **Sufficient!** *(This Lecture)*

Fixed Design OLS (The Tool)

$\|\vec{x}\|_M^2 = x^\top M x$

Consider the standard linear regression setting:

$$y_i = x_i^\top \theta^\star + \xi_i, \quad \text{with } \mathbb{E}[\xi_i | x_i] = 0 \text{ (sub-Gaussian).}$$

The OLS estimator is $\widehat{\theta} = \Lambda^{-1}\left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right)$, where $\Lambda = \frac{1}{N}\sum_i x_i x_i^\top$.

Fixed-design OLS Bound

With probability at least $1-\delta$, the prediction error is bounded in the Λ-norm:

$$\|\widehat{\theta} - \theta^\star\|_\Lambda^{2} \lesssim \sigma\sqrt{\frac{d\log(1/\delta)}{N}}.$$

This translates to a *pointwise* bound using leverage scores:

$$|(\widehat{\theta}-\theta^\star)^\top \phi(s,a)| \;\le\; \|\widehat{\theta}-\theta^\star\|_\Lambda \sqrt{\phi(s,a)^\top \Lambda^{-1} \phi(s,a)}$$

OLS review

Consider the standard linear regression setting:

$\xi_i \sim N(0,\sigma^2)$ $\quad X \in R^{n\times d}$

$\vec{Y} = X\theta^* + \vec{\epsilon}$

$$y_i = x_i^\top \theta^\star + \xi_i, \quad \text{with } \mathbb{E}[\xi_i | x_i] = 0 \text{ (sub-Gaussian).}$$

$\vec{Y}, \vec{\epsilon} \in R^n$

The OLS estimator is $\widehat{\theta} = \Lambda^{-1}\left(\underbrace{\frac{1}{N}\sum_{i=1}^N x_i y_i}_{v}\right)$, where $\Lambda = \frac{1}{N}\sum_i x_i x_i^\top$.

$\Sigma = \frac{1}{n} X^T X$

$v = \frac{1}{n} X^T Y$

$v = \frac{1}{n} X^T (X\theta^* + \vec{\epsilon}) = \Lambda\theta^* + \frac{1}{n} X^T \vec{\epsilon}$

$\hat{\theta} - \theta^* = \Lambda^{-1} v - \theta^* = \frac{1}{n}\Lambda^{-1} X^T \vec{\epsilon}$

$\sigma^2 I$

$E[\|\hat{\theta} - \theta^*\|_\Lambda^2] = \frac{1}{n^2} E[\vec{\epsilon}^T X \Lambda^{-1} X^T \vec{\epsilon}] = \frac{1}{n^2} Tr(X^T E[\epsilon\epsilon^T] X \Lambda^{-1})$

$= \frac{\sigma^2}{n} Tr(\Lambda^{-1}\Lambda) = \frac{\sigma^2}{n} Tr(I) = \frac{d\sigma^2}{n}$

D-optimal design: the leverage-minimizing geometry

To guarantee uniform bounds, we must choose our training data carefully.

The feature set is:

$$\Phi := \{\phi(s,a) : (s,a) \in \mathcal{S} \times \mathcal{A}\} \subset \mathbb{R}^d.$$

D-optimal design (lemma; geometric fact)

Suppose Φ is compact. There exists a distribution ρ supported on at most $d(d+1)/2$ state–action pairs s.t. with

$$\Sigma := \mathbb{E}_{(s,a)\sim\rho}\left[\phi(s,a)\phi(s,a)^\top\right],$$

we have $\Sigma \succ 0$ and

$$\sup_{(s,a)} \; \phi(s,a)^\top \Sigma^{-1} \phi(s,a) \; = \; d.$$

Furthermore, no distribution ρ can achieve a lower (worst-case) leverage score.

Today: Analysis of LSVI

Roadmap: Proving LSVI Works

Recall our strategy:

$$\text{Regression Error} \xrightarrow{\text{D-Opt}} \text{Pointwise Error} \xrightarrow{\text{Completeness}} \text{Bellman Residual} \xrightarrow{\text{Sim. Lemma}} \text{Policy Loss}$$

We will formalize this in three steps:

1. **Regression Analysis:** Bound the error $\|\widehat{\theta}_h - \theta^\star_h\|$ using OLS.
2. **Bellman Residuals:** D-Optimal Design lets us translate parameter err into uniform value function error Res_h.
3. **The Simulation Lemma:** Show how residuals propagate to the final value/policy.

Step 1: The Regression at Stage h

Consider a fixed stage h. We have the "next" value function $\widehat{V}_{h+1}$.

The Data:

- Dataset $D_h = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ collected via D-optimal design ρ.

Step 1: The Regression at Stage h

Consider a fixed stage h. We have the "next" value function $\widehat{V}_{h+1}$.

The Data:

- Dataset $D_h = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ collected via D-optimal design ρ.
- Design Matrix $\Lambda_h = \frac{1}{N}\sum_{i=1}^N \phi(s_i, a_i)\phi(s_i, a_i)^\top \approx \Sigma_\rho$.

Step 1: The Regression at Stage h

Consider a fixed stage h. We have the "next" value function $\widehat{V}_{h+1}$.

The Data:

- Dataset $D_h = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ collected via D-optimal design ρ.
- Design Matrix $\Lambda_h = \frac{1}{N}\sum_{i=1}^N \phi(s_i, a_i)\phi(s_i, a_i)^\top \approx \Sigma_\rho$.

The Target:

$$y_i = r_h(s_i, a_i) + \widehat{V}_{h+1}(s_i')$$

Step 1: The Regression at Stage h

H-1

fit ✓

H-2

$r + \hat{V}_{H-1}$

Consider a fixed stage h. We have the "next" value function $\widehat{V}_{h+1}$.

The Data:

- Dataset $D_h = \{(s_i, a_i, r_i, s_i')\}_{i=1}^N$ collected via D-optimal design ρ.
- Design Matrix $\Lambda_h = \frac{1}{N}\sum_{i=1}^N \phi(s_i, a_i)\phi(s_i, a_i)^\top \approx \Sigma_\rho$.

The Target:

$$y_i = r_h(s_i, a_i) + \widehat{V}_{h+1}(s_i')$$

$\tilde{\theta}$ is better notation (best not to confuse $\hat{\theta}$ with opt θ^*)

By **Completeness**, the true expected target is linear:

$$\mathbb{E}[y_i|s_i, a_i] = (\mathcal{T}_h \widehat{Q}_{h+1})(s_i, a_i) = (\tilde{\theta}_h^{\star})^\top \phi(s_i, a_i)$$

Thus, $y_i = (\tilde{\theta}_h^\star)^\top \phi(s_i, a_i) + \xi_i$, where ξ_i is sub-Gaussian noise.

$\hat{\theta}$ is from BC.

[fixed in pdf slides & chapter]

Step 1: The Fixed Design Bound

We apply the standard Fixed-Design OLS bound.

OLS Generalization Bound

With probability $1 - \delta$, for all h

$$\|\widehat{\theta}_h - \widetilde{\theta}_h^\star\|_{\Lambda_h} \leq c \cdot H\sqrt{\frac{d\log(H/\delta)}{N}}$$

(Here the noise scale is H because values are bounded by H).

$$\sup_{s,a} \left| \hat{\theta}_h \cdot \phi(s,a) - \tilde{\theta}_h \phi(s,a) \right| \leq ??$$

Step 2: From Parameters to Pointwise Error

$\sqrt{d}$

We want to bound the prediction error at *any* (s, a).

$$|(\widehat{\theta}_h - \theta_h^\star)^\top \phi(s,a)| \;\leq\; \|\widehat{\theta}_h - \theta_h^\star\|_{\Lambda_h} \sqrt{\phi(s,a)^\top \Lambda_h^{-1} \phi(s,a)}$$

T

$H\sqrt{d/N}$

Step 2: From Parameters to Pointwise Error

We want to bound the prediction error at *any* (s, a).

$$|(\widehat{\theta}_h - \theta_h^\star)^\top \phi(s,a)| \;\leq\; \|\widehat{\theta}_h - \theta_h^\star\|_{\Lambda_h} \sqrt{\phi(s,a)^\top \Lambda_h^{-1} \phi(s,a)}$$

Plug in our bounds (using $\Lambda_h \approx \Sigma_\rho$):

- Parameter Error (OLS): $\approx H\sqrt{\frac{d}{N}}$
- Leverage Score (D-Opt): $\sqrt{\phi^\top \Lambda_h^{-1} \phi} \approx \sqrt{\phi^\top \Sigma_\rho^{-1} \phi} \leq \sqrt{d}$

Step 2: From Parameters to Pointwise Error

We want to bound the prediction error at *any* (s, a).

$$|(\widehat{\theta}_h - \theta^\star_h)^\top \phi(s, a)| \;\leq\; \|\widehat{\theta}_h - \theta^\star_h\|_{\Lambda_h} \sqrt{\phi(s, a)^\top \Lambda_h^{-1} \phi(s, a)}$$

Plug in our bounds (using $\Lambda_h \approx \Sigma_\rho$):

- Parameter Error (OLS): $\approx H\sqrt{\frac{d}{N}}$
- Leverage Score (D-Opt): $\sqrt{\phi^\top \Lambda_h^{-1} \phi} \approx \sqrt{\phi^\top \Sigma_\rho^{-1} \phi} \leq \sqrt{d}$

Result (Uniform Error): Multiply them together:

$$\sup_{(s,a)} |\widehat{Q}_h(s, a) - (\mathcal{T}_h \widehat{Q}_{h+1})(s, a)| \;\lesssim\; H\sqrt{\frac{d}{N}} \cdot \sqrt{d} \;=\; \frac{Hd}{\sqrt{N}}$$

Step 3: The Simulation Lemma

Bellman residual
of g w.r.t. f
$\|g - \mathcal{T} f\|_\infty$

Definitions:

- Let $\widehat{V}_h(s) := \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) := \operatorname{argmax}_a \widehat{Q}_h(s,a)$.
- Define the stage-h **Bellman Residual** as the prediction error:

$$\mathrm{Res}_h \;:=\; \left\| \widehat{Q}_h - \mathcal{T}_h \widehat{Q}_{h+1} \right\|_\infty$$

Step 3: The Simulation Lemma

Definitions:

- Let $\widehat{V}_h(s) := \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) := \text{argmax}_a \widehat{Q}_h(s,a)$.
- Define the stage-h **Bellman Residual** as the prediction error:

$\eta \approx \frac{Hd}{\sqrt{N}}$

$$\mathrm{Res}_h \; := \; \left\| \widehat{Q}_h - \mathcal{T}_h \widehat{Q}_{h+1} \right\|_\infty$$

Lemma 3.3 (Small residual $\Rightarrow$ good policy)

Assume $\widehat{Q}_H \equiv 0$ and $\mathrm{Res}_h \leq \eta$ for all h. Then:

1. **Value Accuracy:** For all $h \in [H]$,

$$\|\widehat{Q}_h - Q^\star_h\|_\infty \; \leq \; (H-h)\eta.$$

Step 3: The Simulation Lemma

Definitions:

- Let $\widehat{V}_h(s) := \max_a \widehat{Q}_h(s,a)$ and $\widehat{\pi}_h(s) := \operatorname{argmax}_a \widehat{Q}_h(s,a)$.
- Define the stage-h **Bellman Residual** as the prediction error:

$$\mathrm{Res}_h \;:=\; \left\| \widehat{Q}_h - \mathcal{T}_h \widehat{Q}_{h+1} \right\|_\infty$$

Lemma 3.3 (Small residual $\Rightarrow$ good policy)

Assume $\widehat{Q}_H \equiv 0$ and $\mathrm{Res}_h \le \eta$ for all h. Then:

1. **Value Accuracy:** For all $h \in [H]$,

$$\|\widehat{Q}_h - Q_h^\star\|_\infty \;\le\; (H-h)\eta.$$

2. **Policy Loss:** For the greedy policy $\widehat{\pi}$,

$$V_0^\star(s) - V_0^{\widehat{\pi}}(s) \;\le\; 2H^2\eta.$$

Proof of Claim 1: Value Accuracy

We prove $\|\widehat{Q}_h - Q_h^\star\|_\infty \leq (H-h)\eta$ by backward induction.

Base Case ($h = H$): $Q_H^\star = \widehat{Q}_H = 0$, so error is 0.

$\pm \tilde{\mathcal{T}}_h \hat{Q}_{h+1}$

Inductive Step: Consider the error at stage h:

$$\|\hat{Q}_h - Q^*_h\|_\infty = \|\hat{Q}_h - \tilde{\mathcal{T}}_h Q^*_{h+1}\|_\infty$$

$$\leq \|\hat{Q}_h - \tilde{\mathcal{T}}_h \hat{Q}_{h+1}\|_\infty + \|\tilde{\mathcal{T}} \hat{Q}_{h+1} - \tilde{\mathcal{T}} Q^*_{h+1}\|_\infty$$

Proof of Claim 1: Value Accuracy

We prove $\|\widehat{Q}_h - Q_h^\star\|_\infty \le (H-h)\eta$ by backward induction.

Base Case ($h = H$): $Q_H^\star = \widehat{Q}_H = 0$, so error is 0.
Inductive Step: Consider the error at stage h:

$$\begin{aligned}\|\widehat{Q}_h - Q_h^\star\|_\infty &= \|\widehat{Q}_h - \mathcal{T}_h Q_{h+1}^\star\|_\infty \\ &\le \underbrace{\|\widehat{Q}_h - \mathcal{T}_h \widehat{Q}_{h+1}\|_\infty}_{\text{Residual } \le \eta} + \underbrace{\|\mathcal{T}_h \widehat{Q}_{h+1} - \mathcal{T}_h Q_{h+1}^\star\|_\infty}_{\text{Recursive Error}}\end{aligned}$$

Proof of Claim 1: Value Accuracy

We prove $\|\widehat{Q}_h - Q_h^\star\|_\infty \le (H-h)\eta$ by backward induction.

Base Case ($h = H$): $Q_H^\star = \widehat{Q}_H = 0$, so error is 0.
Inductive Step: Consider the error at stage h:

$$\begin{aligned}\|\widehat{Q}_h - Q_h^\star\|_\infty &= \|\widehat{Q}_h - \mathcal{T}_h Q_{h+1}^\star\|_\infty \\ &\le \underbrace{\|\widehat{Q}_h - \mathcal{T}_h \widehat{Q}_{h+1}\|_\infty}_{\text{Residual } \le \eta} + \underbrace{\|\mathcal{T}_h \widehat{Q}_{h+1} - \mathcal{T}_h Q_{h+1}^\star\|_\infty}_{\text{Recursive Error}}\end{aligned}$$

Using the contraction property (non-expansion) of $\mathcal{T}_h$:

$$\|\mathcal{T}_h \widehat{Q}_{h+1} - \mathcal{T}_h Q_{h+1}^\star\|_\infty \le \|\widehat{Q}_{h+1} - Q_{h+1}^\star\|_\infty$$

Thus, $\text{Error}_h \le \eta + \text{Error}_{h+1}$. Unrolling gives $\sum \eta = (H-h)\eta$. □

Proof of Claim 2: Policy Loss

This follows the standard "Performance Difference" logic.

The Argument:

- The loss of a greedy policy is bounded by the sum of suboptimality gaps at each step.
- Since $\widehat{\pi}$ is greedy with respect to $\widehat{Q}$, the single-step loss is bounded by $2\times$ estimation error:

$$Q_h^\star(s, \pi^\star) - Q_h^\star(s, \widehat{\pi}) \;\leq\; 2\|\widehat{Q}_h - Q_h^\star\|_\infty$$

Total Loss: Summing over H steps:

$$V_0^\star - V_0^{\widehat{\pi}} \;\leq\; \sum_{h=0}^{H-1} 2\underbrace{\|\widehat{Q}_h - Q_h^\star\|_\infty}_{\leq (H-h)\eta} \leq \sum_{h=0}^{H-1} 2(H-h)\eta \approx H^2\eta$$

Final Result: LSVI Sample Complexity

$$\|\hat{Q}_h - Q^*_h\| \leq H^2 \eta$$

Putting it all together:

1. We want final error $V^\star - V^{\widehat{\pi}} \leq \epsilon$.
2. By Simulation Lemma, we need $\text{Res}_h \leq \epsilon/(2H^2)$. $= \eta$
3. By Regression Analysis, we need $Hd/\sqrt{N} \approx \epsilon/H^2$.

η

Final Result: LSVI Sample Complexity

Putting it all together:

1. We want final error $V^\star - V^{\widehat{\pi}} \leq \epsilon$.
2. By Simulation Lemma, we need $\mathrm{Res}_h \leq \epsilon/(2H^2)$.
3. By Regression Analysis, we need $Hd/\sqrt{N} \approx \epsilon/H^2$.

Solve for N:

$$\frac{Hd}{\sqrt{N}} \approx \frac{\epsilon}{H^2} \implies \sqrt{N} \approx \frac{H^3 d}{\epsilon} \implies N \approx \frac{H^6 d^2}{\epsilon^2}$$

Theorem (LSVI Generative)

LSVI with D-optimal design yields an ϵ-optimal policy with $\tilde{O}(H^6 d^2/\epsilon^2)$ samples.

Offline RL ~~Recap~~

Switching to Offline RL

The Setting:

- We can no longer query the simulator.
- We are given static datasets $D_0, \ldots, D_{H-1}$.
- **Key Question:** When does LSVI still work?

The Challenge:

- In Generative Mode, we used *D-Optimal Design* to ensure:

$$\Lambda_h \approx \Sigma_{\rho^\star} \implies \text{Good Coverage Everywhere}$$

- In Offline RL, we are stuck with the behavior policy's distribution.

The Coverage Assumption

To guarantee success, the offline data must "cover" the feature space at least as well as the optimal design (up to a constant).

The Coverage Assumption

To guarantee success, the offline data must "cover" the feature space at least as well as the optimal design (up to a constant).

Assumption: Uniform Coverage

There exists a constant $\kappa \geq 1$ such that for all h, the empirical covariance Λ_h satisfies:

$$\Lambda_h \succeq \frac{1}{\kappa}\Sigma_{\rho^\star}$$

$$\Lambda_h^{-1} \lesssim \kappa \Sigma_{\rho^*}^{-1}$$

where $\Sigma_{\rho^\star}$ is the covariance of the D-optimal design.

The Coverage Assumption

To guarantee success, the offline data must "cover" the feature space at least as well as the optimal design (up to a constant).

Assumption: Uniform Coverage

There exists a constant $\kappa \geq 1$ such that for all h, the empirical covariance Λ_h satisfies:

$$\Lambda_h \succeq \frac{1}{\kappa}\Sigma_{\rho^\star}$$

where $\Sigma_{\rho^\star}$ is the covariance of the D-optimal design.

Interpretation:

- κ is the "relative condition number."
- If $\kappa = 1$, our data is perfect (D-optimal).
- If κ is huge, we have missing directions (poor coverage).

Analysis of Offline LSVI

The analysis remains almost identical! We just swap the leverage bound.

Analysis of Offline LSVI

The analysis remains almost identical! We just swap the leverage bound.

1. Generative (D-Optimal):

$$\phi^\top \Lambda^{-1} \phi \approx \phi^\top \Sigma_{\rho^\star}^{-1} \phi \;\leq\; d$$

Analysis of Offline LSVI

The analysis remains almost identical! We just swap the leverage bound.

1. **Generative (D-Optimal):**

$$\phi^\top \Lambda^{-1} \phi \approx \phi^\top \Sigma_{\rho^\star}^{-1} \phi \; \leq \; d$$

2. **Offline (Coverage κ):**

$$\Lambda \succeq \frac{1}{\kappa} \Sigma_{\rho^\star} \implies \Lambda^{-1} \preceq \kappa \Sigma_{\rho^\star}^{-1}$$

$$\implies \phi^\top \Lambda^{-1} \phi \; \leq \; \kappa \left(\phi^\top \Sigma_{\rho^\star}^{-1} \phi \right) \; \leq \; \kappa d$$

Analysis of Offline LSVI

The analysis remains almost identical! We just swap the leverage bound.

1. Generative (D-Optimal):

$$\phi^\top \Lambda^{-1} \phi \approx \phi^\top \Sigma_{\rho^\star}^{-1} \phi \ \leq \ d$$

2. Offline (Coverage κ):

$$\Lambda \succeq \frac{1}{\kappa} \Sigma_{\rho^\star} \implies \Lambda^{-1} \preceq \kappa \Sigma_{\rho^\star}^{-1}$$

$$\implies \phi^\top \Lambda^{-1} \phi \ \leq \ \kappa \left(\phi^\top \Sigma_{\rho^\star}^{-1} \phi \right) \ \leq \ \kappa d$$

Result: The error scales by $\sqrt{\kappa}$. Sample complexity scales linearly with κ:

$$\text{Error} \approx \frac{Hd\sqrt{\kappa}}{\sqrt{N}} \implies N \approx \frac{\kappa H^6 d^2}{\epsilon^2}$$

Task 2: Policy Evaluation

Sometimes we don't want to find $\pi^\star$, but just evaluate a fixed policy π.

Task 2: Policy Evaluation

Sometimes we don't want to find $\pi^\star$, but just evaluate a fixed policy π.

Algorithm: Least-Squares Policy Evaluation (LSPE)

- Same backward regression structure.
- **Target Change:** Instead of $\max_{a'} \widehat{Q}_{h+1}(s', a')$, we use the value of our specific policy:

$$y_i = r_i + \widehat{Q}_{h+1}(s_i', \pi(s_i'))$$

Task 2: Policy Evaluation

Sometimes we don't want to find $\pi^\star$, but just evaluate a fixed policy π.

Algorithm: Least-Squares Policy Evaluation (LSPE)

- Same backward regression structure.
- **Target Change:** Instead of $\max_{a'} \widehat{Q}_{h+1}(s', a')$, we use the value of our specific policy:

$$y_i = r_i + \widehat{Q}_{h+1}(s_i', \pi(s_i'))$$

Weaker Assumption:

- We don't need closure under $\mathcal{T}$ (Optimality).
- We only need closure under $\mathcal{T}^\pi$ (Policy Operator).

LSPE Analysis: Why it's better

Evaluating a fixed policy is easier than finding the optimal one.

1. **No "Greedy" Error Propagation**
   - In LSVI, we incur error from the Bellman residual AND the greedy step.
   - In LSPE, we only track the estimation error of a fixed operator.
2. **Tighter Simulation Lemma**
   - **Control (LSVI):** $V^\star - V^{\widehat{\pi}} \leq 2H^2\eta$.
   - **Evaluation (LSPE):** $|V^\pi - \widehat{V}^\pi| \leq H\eta$.

LSPE Analysis: Why it's better

Evaluating a fixed policy is easier than finding the optimal one.

1. **No "Greedy" Error Propagation**
   - In LSVI, we incur error from the Bellman residual AND the greedy step.
   - In LSPE, we only track the estimation error of a fixed operator.
2. **Tighter Simulation Lemma**
   - **Control (LSVI):** $V^\star - V^{\widehat{\pi}} \leq 2H^2\eta$.
   - **Evaluation (LSPE):** $|V^\pi - \widehat{V}^\pi| \leq H\eta$.

Result: Better dependence on horizon H (typically H^4 instead of H^6).

1. The Critical Decomposition

To analyze error propagation, we introduce the **Infinite-Sample Target**.

Definition ($f_h^\star$): The function LSVI would learn with infinite data.

$$f_h^\star := \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_\rho \left[\left(f(s,a) - \mathcal{T}_h \widehat{Q}_{h+1}(s,a) \right)^2 \right] = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1})$$

$= \mathcal{T}\hat{Q}$ if B.C.

(Ideally, we want $f_h^\star$ to track the optimal value $Q_h^\star$).

1. The Critical Decomposition

To analyze error propagation, we introduce the **Infinite-Sample Target**.

Definition ($f_h^\star$): The function LSVI would learn with infinite data.

$$f_h^\star := \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_\rho \left[\left(f(s,a) - \mathcal{T}_h \widehat{Q}_{h+1}(s,a) \right)^2 \right] = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1})$$

(Ideally, we want $f_h^\star$ to track the optimal value $Q_h^\star$).

The Error Triangle: We split the total error into Statistical Variance and Recursive Stability.

$$\|\widehat{Q}_h - Q_h^\star\|_\infty \leq \underbrace{\|\widehat{Q}_h - f_h^\star\|_\infty}_{\text{Statistical Error}} + \underbrace{\|f_h^\star - Q_h^\star\|_\infty}_{\text{Recursive Stability}}$$

1. The Critical Decomposition

To analyze error propagation, we introduce the **Infinite-Sample Target**.

Definition ($f_h^\star$): The function LSVI would learn with infinite data.

$$f_h^\star := \operatorname*{argmin}_{f \in \mathcal{F}} \mathbb{E}_\rho \left[\left(f(s,a) - \mathcal{T}_h \widehat{Q}_{h+1}(s,a) \right)^2 \right] = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1})$$

(Ideally, we want $f_h^\star$ to track the optimal value $Q_h^\star$).

The Error Triangle: We split the total error into Statistical Variance and Recursive Stability.

$$\|\widehat{Q}_h - Q_h^\star\|_\infty \leq \underbrace{\|\widehat{Q}_h - f_h^\star\|_\infty}_{\text{Statistical Error}} + \underbrace{\|f_h^\star - Q_h^\star\|_\infty}_{\text{Recursive Stability}}$$

- **Statistical Error:** Controlled by N and D-Optimal Design ($\approx \sqrt{d/N}$).
- **Recursive Stability:** This is where the universe splits.

2. The Split Universe

How does the Recursive Error $\|f_h^\star - Q_h^\star\|_\infty$ behave?

Recall $f_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1})$ and $Q_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h Q_{h+1}^\star)$ (by realizability).

2. The Split Universe

How does the Recursive Error $\|f_h^\star - Q_h^\star\|_\infty$ behave?
Recall $f_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1})$ and $Q_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h Q_{h+1}^\star)$ (by realizability).

Universe A: Completeness

Assumption: $\mathcal{T}_h$ preserves linearity.

Since $\widehat{Q}_{h+1} \in \mathcal{F}$, we have:
$f_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1}) = \mathcal{T}_h \widehat{Q}_{h+1}$, and thus:

$$\begin{aligned} \|f_h^\star - Q_h^\star\|_\infty &= \|\mathcal{T}_h \widehat{Q}_{h+1} - \mathcal{T}_h Q_{h+1}^\star\|_\infty \\ &\le \|\widehat{Q}_{h+1} - Q_{h+1}^\star\|_\infty \end{aligned}$$

Result: Error is stable (contraction).

2. The Split Universe

How does the Recursive Error $\|f_h^\star - Q_h^\star\|_\infty$ behave?
Recall $f_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1})$ and $Q_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h Q_{h+1}^\star)$ (by realizability).

Universe A: Completeness

Assumption: $\mathcal{T}_h$ preserves linearity.

Since $\widehat{Q}_{h+1} \in \mathcal{F}$, we have:
$f_h^\star = \Pi_{\mathcal{F},\rho}(\mathcal{T}_h \widehat{Q}_{h+1}) = \mathcal{T}_h \widehat{Q}_{h+1}$, and thus:

$$\begin{aligned} \|f_h^\star - Q_h^\star\|_\infty &= \|\mathcal{T}_h \widehat{Q}_{h+1} - \mathcal{T}_h Q_{h+1}^\star\|_\infty \\ &\leq \|\widehat{Q}_{h+1} - Q_{h+1}^\star\|_\infty \end{aligned}$$

Result: Error is stable (contraction).

Universe B: Realizability Only

Assumption: Only $Q^\star \in \mathcal{F}$.

The Bellman backup $\mathcal{T}_h \widehat{Q}_{h+1}$ may be **non-linear** (off-manifold).

We must project it back to $\mathcal{F}$:

$$f_h^\star = \Pi_{\mathcal{F},\rho}(\text{Non-Linear Target})$$

Result: We pay for the stability of the projection operator $\Pi_{\mathcal{F},\rho}$.

3. The Amplification Mechanism (Universe B)

Without Completeness, we must bound the stability of the projection $\Pi_{\mathcal{F},\rho}$. Let $\Delta = \mathcal{T}_h \widehat{Q}_{h+1} - \mathcal{T}_h Q^\star_{h+1}$.

The Chain of Inequalities:

$$\begin{aligned}
\|f^\star_h - Q^\star_h\|_\infty &= \|\Pi_{\mathcal{F},\rho}\Delta\|_\infty \\
&\leq \sqrt{d} \cdot \|\Pi_{\mathcal{F},\rho}\Delta\|_{L_2(\rho)} \quad && \text{(\textbf{Step 1:} Norm Equivalence)} \\
&\leq \sqrt{d} \cdot \|\Delta\|_{L_2(\rho)} && \text{(\textbf{Step 2:} } L_2 \text{ Stability of LS)} \\
&\leq \sqrt{d} \cdot \|\Delta\|_\infty && \text{(\textbf{Step 3:} Norm Monotonicity)} \\
&\leq \sqrt{d} \cdot \|\widehat{Q}_{h+1} - Q^\star_{h+1}\|_\infty
\end{aligned}$$

The Verdict: The "price" of converting the L_2 guarantee (regression) to the L_∞ guarantee (DP) is exactly $\sqrt{d}$.

Total Amplification over H steps $\approx (\sqrt{d})^H$.

Summary & What's Next

Summary of Lecture 2:

- **Generative Model:** LSVI + D-Optimal Design achieves $\mathrm{poly}(d, H)$ sample complexity.
- **Offline RL:** The same algorithm works if we assume coverage (κ).
- **The Key Ingredient: Linear Bellman Completeness** ensures the regression targets remain realizable, preventing bias propagation.

Next Lecture (The Negative Results):

- What if we *don't* have Completeness?
- What if we only know $Q^\star$ is linear?
- We will show that without Completeness, sample complexity can become **Exponential in H or d.**