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

Part 3: Lower Bounds

CS 2284: Foundations of Reinforcement Learning

Kianté Brantley & Sham Kakade

Agenda

Announcements

- more project dates coming soon
- reading3 posted

Recap

- Offline RL +linear BC

Today

- Lower bounds!

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).

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.

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.

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:

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

$\mathcal{T}^\pi Q_{h+1} = r_h(s, \pi(s)) + \mathbb{E}_{s' \sim P_{sa}}[Q_{h+1}(s', \pi(s'))]$

Weaker Assumption:

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

Result: Basically the same as LSVI (with slighly better H dependence).

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.

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

(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).

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:

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

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 $\approx (\sqrt{d})^H$.

Offline Lower Bounds with Realizatiblity+Coverage

The Offline Setting: Strong Assumptions

Let's try to "break" the lower bound by making extremely strong assumptions.

The Setup:

- **Offline Data:** Fixed datasets $D_0, \ldots, D_{H-1}$.

The Offline Setting: Strong Assumptions

Let's try to "break" the lower bound by making extremely strong assumptions.

The Setup:

- **Offline Data:** Fixed datasets $D_0, \ldots, D_{H-1}$.
- **Assumption 1 (All-Policies Realizability):** For *every* policy π, Q_h^π is linear in ϕ.

The Offline Setting: Strong Assumptions

Let's try to "break" the lower bound by making extremely strong assumptions.

The Setup:

- **Offline Data:** Fixed datasets $D_0, \ldots, D_{H-1}$.
- **Assumption 1 (All-Policies Realizability):** For *every* policy π, Q_h^π is linear in ϕ.
- **Assumption 2 (Perfect Coverage):** The offline data covariance is "isotropic" (perfectly conditioned):

$$\Sigma_{D_h} = \frac{1}{d} I \quad \text{for all } h.$$

$\|\phi\| \leq 1$

The Offline Setting: Strong Assumptions

Let's try to "break" the lower bound by making extremely strong assumptions.

The Setup:

- **Offline Data:** Fixed datasets $D_0, \ldots, D_{H-1}$.
- **Assumption 1 (All-Policies Realizability):** For *every* policy π, Q_h^π is linear in ϕ.
- **Assumption 2 (Perfect Coverage):** The offline data covariance is "isotropic" (perfectly conditioned):

$$\Sigma_{D_h} = \frac{1}{d} I \quad \text{for all } h.$$

The Question: Under these ideal conditions (Realizability + Perfect Coverage), can we learn V^π with $\mathrm{poly}(d, H)$ samples?

Theorem: Offline Hardness

$T \circ f \in \mathcal{F}$ $\quad f \in \mathcal{F} = \{ w \cdot \phi \mid w \in \mathbb{R}^d \}$

Theorem 3.2 (Offline Policy Evaluation Hardness)

There exists a class of MDPs with dimension d and horizon H satisfying **All-Policies Realizability** and **Perfect Coverage**, such that any algorithm requires

$$N \geq \Omega\left((d/2)^H\right)$$

samples to estimate $Q^\pi(s_0, a)$ to constant accuracy.

Interpretation: Even with infinite data in the "covered" directions, variance accumulates exponentially in the "hidden" directions.

The High-Level Idea: Hiding a Secret

The Goal: Construct an MDP where the optimal value depends on a tiny parameter r_∞, but we can only learn it by solving a "hard" regression problem.

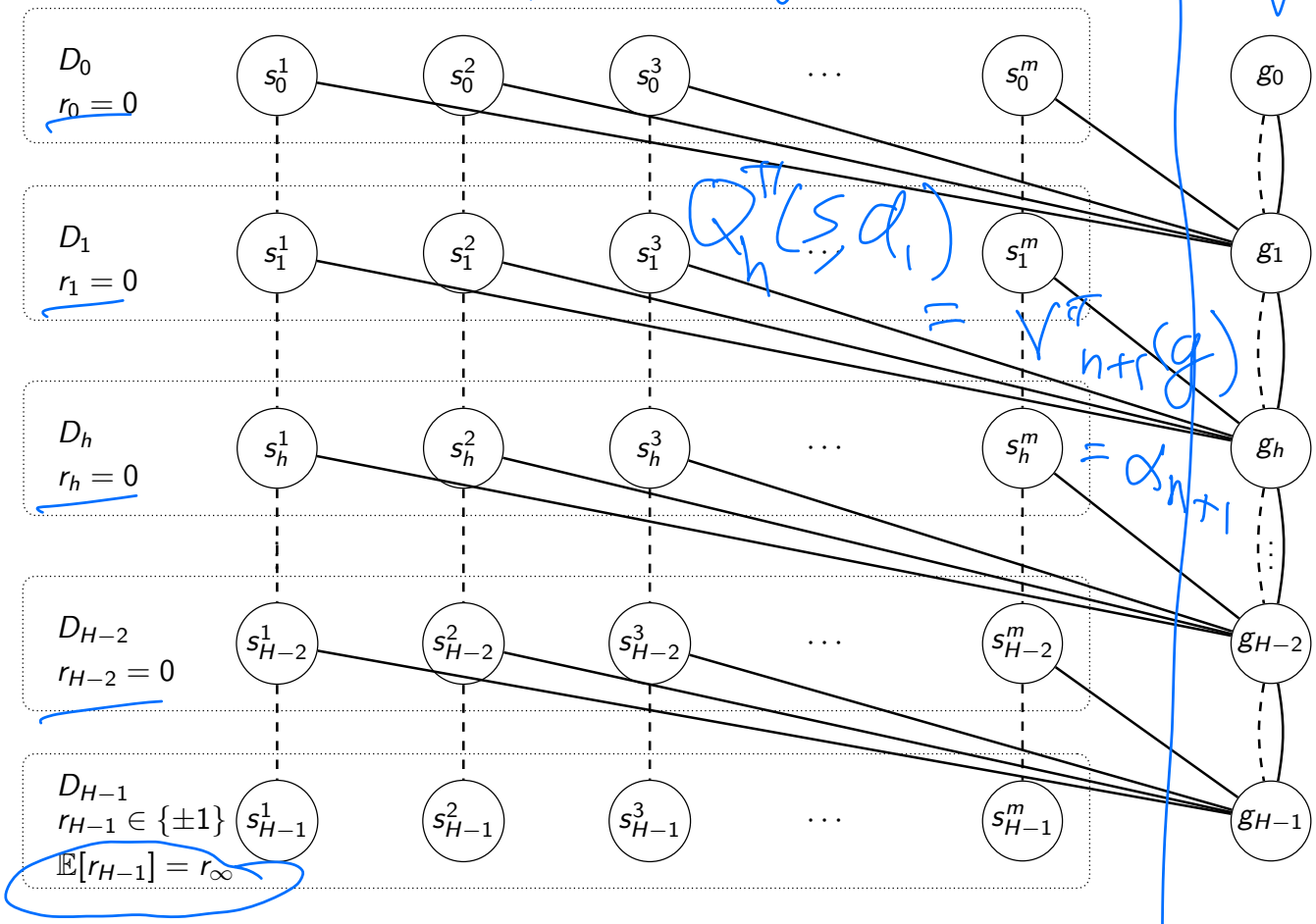
The Mechanism (Two Chains):

1. **The Hidden Chain (g):** Carries the true "signal" ($V \approx$ Large).
2. **The Observed Chain (s):** Carries a "noisy version" of the signal ($V \approx$ Small).

The "Trap": Linearly Realizable implies that $V(g)$ is coupled to $V(s)$ by a factor of $\sqrt{d}$.

- To know the large value at g, we must estimate the small value at s.
- Estimating the small value requires cancelling out huge noise.
- This requires exponentially many samples.

Offline PE hard instance (schematic)



a_1

a_2

Reward parameters.

$r_\infty \in \{0, m^{-H/2}\}$

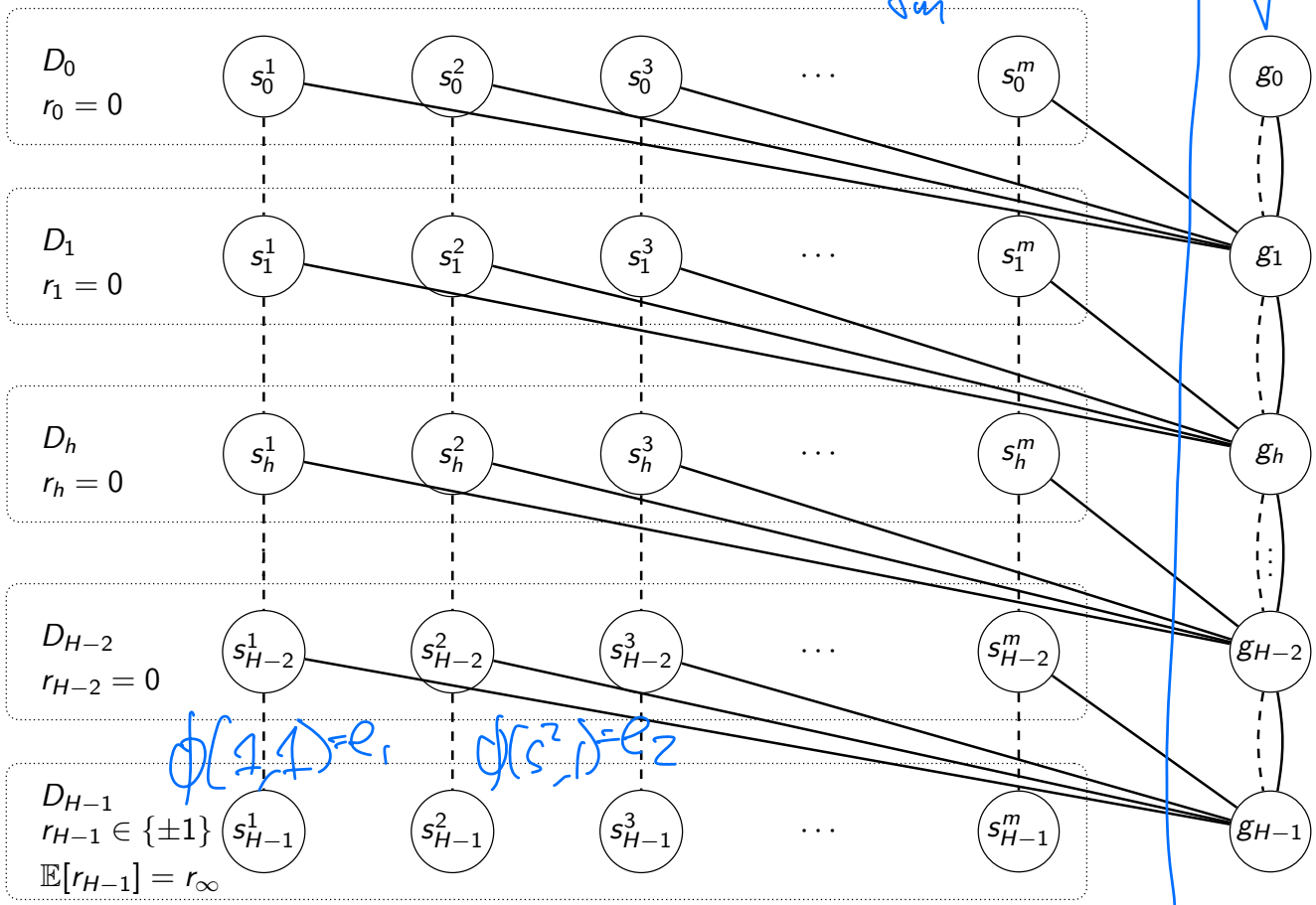
$\alpha_h := r_\infty m^{(H-h)/2}$

Aggregator rewards.

$r_h(g_h, \cdot) = \alpha_h - \alpha_{h+1}$.

Last layer: $r_{H-1}(g_{H-1}) = \alpha_{H-1}$.

Offline PE hard instance (schematic)



a_1 (solid)

a_2 (dashed)

Reward parameters.

$r_\infty \in \{0,\ m^{-H/2}\}$

$\alpha_h := r_\infty\, m^{(H-h)/2}$

Aggregator rewards.

$r_h(g_h, \cdot) = \alpha_h - \alpha_{h+1}$.

Last layer: $r_{H-1}(g_{H-1}) = \alpha_{H-1}$.

Features.

$\phi(s_h^c, a_1) = e_c$

$\phi(s_h^c, a_2) = e_{m+c}$

$\phi(g_h, a) = \frac{1}{\sqrt{m}}(e_1 + \cdots + e_m)$

State space and transitions (two chains)

Fix $m \geq 1$, set $d = 2m$, $|\mathcal{S}_h| = m + 1$, $\mathcal{A} = \{a_1, a_2\}$, horizon H.

State space (layered).

$$\mathcal{S}_h := \{s_h^1, \ldots, s_h^m\} \cup \{g_h\}, \qquad h = 0, \ldots, H-1, \quad \text{and } s_0 = g_0.$$

Deterministic transitions for $h \leq H - 2$:

From any state, action a_1 jumps to aggregators:	$P(g_{h+1} \mid s, a_1) = 1 \quad \forall s \in \mathcal{S}_h.$
Action a_2 keeps index on observed chain:	$P(s_{h+1}^i \mid s_h^i, a_2) = 1 \quad \forall i \in [m].$
Aggregator chain stays on aggregators:	$P(g_{h+1} \mid g_h, a_2) = 1.$

Interpretation: a_2 keeps you in the observed world; a_1 (or being at g_h) takes you off it.

Features and Verifying Perfect Coverage

The Features: We use the standard basis for the observed states, and a dense "mean" vector for the aggregators.

- $\phi(s^i, a_1) = e_i \quad$ and $\quad \phi(s^i, a_2) = e_{m+i}$
- $\phi(g, \cdot) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} e_i$

Features and Verifying Perfect Coverage

The Features: We use the standard basis for the observed states, and a dense "mean" vector for the aggregators.

- $\phi(s^i, a_1) = e_i \quad$ and $\quad \phi(s^i, a_2) = e_{m+i}$
- $\phi(g, \cdot) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} e_i$

Observed Data Distribution (μ_h): Uniform distribution over the m observed states $\{s_h^i\}_{i=1}^m$ and actions.

Features and Verifying Perfect Coverage

The Features: We use the standard basis for the observed states, and a dense "mean" vector for the aggregators.

- $\phi(s^i, a_1) = e_i \quad$ and $\quad \phi(s^i, a_2) = e_{m+i}$
- $\phi(g, \cdot) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} e_i$

Observed Data Distribution (μ_h): Uniform distribution over the m observed states $\{s_h^i\}_{i=1}^m$ and actions.

Coverage Check: Calculate the covariance matrix of the data:

$$\Sigma_D = \frac{1}{2m} \sum_{i=1}^{m} (e_i e_i^\top + e_{m+i} e_{m+i}^\top) = \frac{1}{2m} I = \frac{1}{d} I$$

Perfect isotropic coverage! The offline data explores every dimension of the feature space equally.

Verifying Realizability. Step 1: The Constraint

Linearity $Q^{\pi}(s,a) = \theta^{\top}\phi(s,a)$ imposes rigid structure. (Fix level h below. Let π be arbitrary.)

Verifying Realizability. Step 1: The Constraint

Linearity $Q(s,a) = \theta^\top \phi(s,a)$ imposes rigid structure. (Fix level h below. Let π be arbitrary.)

1. The Observed States are Tabular: Since $\phi(s^i, a_1) = e_i$, the weights θ just memorize the values:

$$\theta_i = Q(s^i, a_1) \quad \text{and} \quad \theta_{m+i} = Q(s^i, a_2)$$

We can represent *any* function on the observed states perfectly.

Verifying Realizability. Step 1: The Constraint

Linearity $Q(s,a) = \theta^\top \phi(s,a)$ imposes rigid structure. (Fix level h below. Let π be arbitrary.)

1. The Observed States are Tabular: Since $\phi(s^i, a_1) = e_i$, the weights θ just memorize the values:

$$\theta_i = Q(s^i, a_1) \quad \text{and} \quad \theta_{m+i} = Q(s^i, a_2)$$

We can represent *any* function on the observed states perfectly.

2. The Aggregator is Constrained: However, the value at g is fixed by the weights θ:

$$Q(g,\cdot) = \theta^\top \phi(g) = \theta^\top \left(\frac{1}{\sqrt{m}} \sum_{i=1}^{m} e_i \right)$$

Substituting $\theta_i = Q(s^i, a_1)$, we get the **Realizability Constraint**:

$$\boxed{Q(g,\cdot) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Q(s^i, a_1)}$$

To be realizable, the value at g must be the scaled sum of values at s.

Verifying Realizability. Step 2: Checking the Values

Does our specific MDP satisfy this constraint?

$$\text{Constraint: } Q_h(g_h, \cdot) \stackrel{?}{=} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Q_h(s_h^i, a_1)$$

Verifying Realizability. Step 2: Checking the Values

Does our specific MDP satisfy this constraint?

$$\text{Constraint: } Q_h(g_h, \cdot) \stackrel{?}{=} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Q_h(s_h^i, a_1)$$

First: By construction, $\alpha_h = \sqrt{m}\alpha_{h+1}$.

Verifying Realizability. Step 2: Checking the Values

Does our specific MDP satisfy this constraint?

$$\text{Constraint: } Q_h(g_h, \cdot) \stackrel{?}{=} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Q_h(s_h^i, a_1)$$

First: By construction, $\alpha_h = \sqrt{m}\alpha_{h+1}$.
LHS (Actual Value at g): By construction, rewards telescope so $V_h(g_h) = \alpha_h = \sqrt{m}V_{h+1}(g_{h+1})$.

Verifying Realizability. Step 2: Checking the Values

Does our specific MDP satisfy this constraint?

$$\text{Constraint: } Q_h(g_h, \cdot) \stackrel{?}{=} \frac{1}{\sqrt{m}} \sum_{i=1}^{m} Q_h(s_h^i, a_1)$$

First: By construction, $\alpha_h = \sqrt{m}\alpha_{h+1}$.
LHS (Actual Value at g): By construction, rewards telescope so $V_h(g_h) = \alpha_h = \sqrt{m}V_{h+1}(g_{h+1})$.

RHS (Sum of Values at s): Action a_1 at s_h^i jumps to g_{h+1}.

$$Q_h(s_h^i, a_1) = 0 + V_{h+1}(g_{h+1})$$

So the sum is:

$$\frac{1}{\sqrt{m}} \sum_{i=1}^{m} V_{h+1}(g_{h+1}) \;=\; \sqrt{m}V_{h+1}(g_{h+1})$$

Constraint Verified.

Step 3: The Impossibility Result

Putting it together:

- We verified Realizability + Opt. Coverage hold for all h.
- Problem choice: $r_\infty \in \{0,\ m^{-H/2}\}$
- Value at start: $V^\pi(g_0) = \alpha_0 = m^{H/2} r_\infty \in \{0, 1\}$.

Step 3: The Impossibility Result

Putting it together:

- We verified Realizability + Opt. Coverage hold for all h.
- Problem choice: $r_\infty \in \{0,\ m^{-H/2}\}$
- Value at start: $V^\pi(g_0) = \alpha_0 = m^{H/2} r_\infty \in \{0, 1\}$.

The Hypothesis Test: To decide if $V^\pi(g_0)$ is 0 or 1, we must distinguish:

$$r_\infty = 0 \quad \text{vs} \quad r_\infty = m^{-H/2}$$

Step 3: The Impossibility Result

Putting it together:

- We verified Realizability + Opt. Coverage hold for all h.
- Problem choice: $r_\infty \in \{0,\ m^{-H/2}\}$
- Value at start: $V^\pi(g_0) = \alpha_0 = m^{H/2} r_\infty \in \{0, 1\}$.

The Hypothesis Test: To decide if $V^\pi(g_0)$ is 0 or 1, we must distinguish:

$$r_\infty = 0 \quad \text{vs} \quad r_\infty = m^{-H/2}$$

The Bottleneck: We never see g_h. We only see the noisy leaves at $H - 1$.

- Signal Gap: $\Delta = m^{-H/2}$.
- Samples Needed: $N \approx 1/\Delta^2 = m^H = (d/2)^H$.

We need exponential samples to track the value of the unseen state.

Discussion 1: Why LSPE Fails (Variance Explosion)

The Paradox: LSPE is an unbiased estimator. Why does it fail?

Discussion 1: Why LSPE Fails (Variance Explosion)

The Paradox: LSPE is an unbiased estimator. Why does it fail?

1. Tabular Learning on Observed States (Base Case) On the observed states, the features are standard basis vectors e_i. Thus, LSPE estimates the parameters $\widehat{Q}_{H-1}(s^i, a_1)$ purely locally:

$$\widehat{Q}_{H-1}(s^i, a_1) \approx Q^{\star}_{H-1}(s^i, a_1) \pm \frac{\sigma}{\sqrt{N}}$$

Discussion 1: Why LSPE Fails (Variance Explosion)

The Paradox: LSPE is an unbiased estimator. Why does it fail?

1. Tabular Learning on Observed States (Base Case) On the observed states, the features are standard basis vectors e_i. Thus, LSPE estimates the parameters $\widehat{Q}_{H-1}(s^i, a_1)$ purely locally:

$$\widehat{Q}_{H-1}(s^i, a_1) \approx Q^{\star}_{H-1}(s^i, a_1) \pm \frac{\sigma}{\sqrt{N}}$$

2. The Recursive Update. At subsequent layers, the value of jumping $Q_h(s, a_1)$ is determined by the value at the (unseen) next state g_{h+1}. Working out the LSPE update shows:

$$\widehat{Q}_h(s, a_1) = \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \widehat{Q}_{h+1}(s^i, a_1) = \sqrt{m} \underbrace{\left(\frac{1}{m} \sum_{i=1}^{m} \widehat{Q}_{h+1}(s^i, a_1) \right)}_{\text{Average Estimate}}$$

Variance Explosion: The estimate at h is inflated by $\sqrt{m}$ times the average estimate at $h+1$.

We are extrapolating $\sqrt{m}$ further out than our data support, H times in a row.

Discussion 2: Can Online RL Save Us?

- The features are not **complete** (even though every policy is linear).
- **A Trivial Observation:** Our offline hard instance has very few states-action pairs. $|\mathcal{S}| \cdot |\mathcal{A}| = H \cdot (d + 2)$
- In an **Online** (episodic) or **Generative model** setting, we could simply visit every state and solve it tabularly!

Discussion 2: Can Online RL Save Us?

- The features are not **complete** (even though every policy is linear).
- **A Trivial Observation:** Our offline hard instance has very few states-action pairs. $|\mathcal{S}| \cdot |\mathcal{A}| = H \cdot (d + 2)$
- In an **Online** (episodic) or **Generative model** setting, we could simply visit every state and solve it tabularly!
- **The Question:** Is there a hard instance — where the state space is huge — where even active exploration fails because we can't "find" the hidden $\theta^\star$ direction?

Can we construct a "Needle in a Haystack" where we can't find the features?

Online Lower Bounds with Realizability

What if we can explore? (Generative Model)

The Question: We saw Offline RL fails even with coverage and all policies linear. What if we have a **Generative Model** (can sample any (s, a))?

What if we can explore? (Generative Model)

The Question: We saw Offline RL fails even with coverage and all policies linear. What if we have a **Generative Model** (can sample any (s, a))?

The Assumption (Weakening): Suppose **only** that $Q^\star$ is linear.

$$Q_h^\star(s, a) = \langle \theta_h^\star, \phi(s, a) \rangle$$

(So no linear BC, i.e. we do *not* assume $\mathcal{T}Q \in \mathcal{F}$ for all Q).

What if we can explore? (Generative Model)

The Question: We saw Offline RL fails even with coverage and all policies linear. What if we have a **Generative Model** (can sample any (s, a))?

The Assumption (Weakening): Suppose **only** that $Q^\star$ is linear.

$$Q_h^\star(s, a) = \langle \theta_h^\star, \phi(s, a) \rangle$$

(So no linear BC, i.e. we do *not* assume $\mathcal{T}Q \in \mathcal{F}$ for all Q).

Theorem (Hardness of Linear $Q^\star$)

Even with a Generative Model, any algorithm requires

$$N \geq \min\{2^{\Omega(d)}, 2^{\Omega(H)}\}$$

samples to find an ϵ-optimal policy (hard instance: $|\mathcal{A}| \approx \text{poly}(d)$, $|\mathcal{S}| \approx \exp(d)$).

Proof Comments

1. **The Construction:**
   - Construct a tree where only one specific path (the "needle") has high value.
   - Again there is this "amplfication" effect.
   - But the feature vector construction is very subtle.
2. **The Failure of Gradient Signal:**
   - Without **Completeness**, the Bellman updates on the "distractor" paths do not propagate useful gradient information about $\theta^\star$.
   - The value function looks "flat" everywhere except exactly on the optimal path.

Summary: The Linearity Ladder

We can classify RL problems based on their structural assumptions.

1. **Linear Bellman Completeness:**
   - **Result:** poly(d, H).
   - **Status:** Efficient (LSVI).

Summary: The Linearity Ladder

We can classify RL problems based on their structural assumptions.

1. **Linear Bellman Completeness:**
   - **Result:** poly(d, H).
   - **Status:** Efficient (LSVI).
2. **Linear $Q^\star$ Realizability:**
   - **Result:** Hard with a Generative Model.
   - **Status:** Exponential Lower Bounds.

Summary: The Linearity Ladder

We can classify RL problems based on their structural assumptions.

1. **Linear Bellman Completeness:**
   - **Result:** poly(d, H).
   - **Status:** Efficient (LSVI).
2. **Linear $Q^\star$ Realizability:**
   - **Result:** Hard with a Generative Model.
   - **Status:** Exponential Lower Bounds.
3. **All-Policy Realizability:**
   - **Offline (+ Coverage):** Hard (Variance Explosion).
   - **Generative Model:** Easy (Matrix Estimation).
   - **Online RL: Unknown/Open Question**.

Implication: To get polynomial sample complexity in RL, we need strong **Structural Assumptions** (like Completeness, Block MDPs, or Low Rank) that ensure values can be propagated globally.