Learning with Linear Bellman Completion & Generative Model

Wen Sun

CS 6789: Foundations of Reinforcement Learning
Given feature $\phi$, take any linear function $w^T \phi(s, a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, \text{ s.t. }, \theta^T \phi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P_{h(s,a)}} \max_{a'} w^T \phi(s', a'), \forall s, a$$
Recap: Linear Bellman Completion

Given feature $\phi$, take any linear function $w^T \phi(s,a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, s.t., \theta^T \phi(s,a) = r(s,a) + \mathbb{E}_{s' \sim P_{h(s,a)}} \max_{a'} w^T \phi(s',a'), \forall s, a$$

It implies that $Q_h^*$ is linear in $\phi$: $Q_h^* = (\theta_h^*)^T \phi, \forall h$
Recap: Linear Bellman Completion

Given feature $\phi$, take any linear function $w^T \phi(s, a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, \text{ s.t., } \theta^T \phi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P_{h(s, a)}} \max_{a'} w^T \phi(s', a'), \forall s, a$$

It implies that $Q^*_h$ is linear in $\phi$: $Q^*_h = (\theta^*_h)^T \phi, \forall h$

Captures Tabular MDPs, and Linear Quadratic Regulators
Recap: Linear Bellman Completion

Given feature $\phi$, take any linear function $w^T \phi(s, a)$:

$$\forall h, \exists \theta \in \mathbb{R}^d, \text{s.t., } \theta^T \phi(s, a) = r(s, a) + \mathbb{E}_{s' \sim P_h(s, a)} \max_{a'} w^T \phi(s', a'), \forall s, a$$

It implies that $Q^*_h$ is linear in $\phi$: $Q^*_h = (\theta^*_h)^T \phi, \forall h$

Captures Tabular MDPs, and Linear Quadratic Regulators

But adding additional elements may just break the condition
Recap: Least-Square Value Iteration
Recap: Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ 
$\mathcal{D}_h = \{s, a, r, s'\}$, $r = r(s, a)$, $s' \sim P_h(\cdot | s, a)$
Recap: Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ 
$\mathcal{D}_h = \{s, a, r, s'\}, \ r = r(s, a), \ s' \sim P_h(\cdot | s, a)$

Set $V_H(s) = 0, \forall s$
Recap: Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ 
$\mathcal{D}_h = \{s, a, r, s'\}$, $r = r(s, a)$, $s' \sim P_h(\cdot | s, a)$

Set $V_H(s) = 0, \forall s$

For $h = H - 1$ to 0:

$$\theta_h = \arg \min_{\theta} \sum_{\mathcal{D}_h} \left( \theta^T \phi(s, a) - \left( r + V_{h+1}(s') \right) \right)^2$$
Recap: Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ 
$\mathcal{D}_h = \{s, a, r, s'\}$, $r = r(s, a)$, $s' \sim P_h(\cdot | s, a)$

Set $V_H(s) = 0, \forall s$

For $h = H-1$ to 0:

\[
\theta_h = \arg \min_{\theta} \sum_{\mathcal{D}_h} \left( \theta^T \phi(s, a) - (r + V_{h+1}(s')) \right)^2
\]

Set $V_h(s) := \max_a \theta_h^T \phi(s, a), \forall s$
Recap: Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ $\mathcal{D}_h = \{s, a, r, s'\}, r = r(s, a), s' \sim P_h(\cdot | s, a)$

Set $V_H(s) = 0, \forall s$

For $h = H-1$ to 0:

- $\theta_h = \arg \min_\theta \sum_{\mathcal{D}_h} \left( \theta^T \phi(s, a) - (r + V_{h+1}(s')) \right)^2$
- Set $V_h(s) := \max_a \theta_h^T \phi(s, a), \forall s$

Return $\hat{\pi}_h(s) = \arg \max_a \theta_h^T \phi(s, a), \forall h$
Recap: Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ $\mathcal{D}_h = \{s, a, r, s'\}, r = r(s, a), s' \sim P_h(\cdot | s, a)$

Set $V_H(s) = 0, \forall s$

For $h = H-1$ to 0:

$$\theta_h = \arg \min_{\theta} \sum_{\mathcal{D}_h} \left( \theta^T \phi(s, a) - (r + V_{h+1}(s')) \right)^2$$

Set $V_h(s) := \max_a \theta_h^T \phi(s, a), \forall s$

Return $\hat{\pi}_h(s) = \arg \max_a \theta_h^T \phi(s, a), \forall h$
Outline for Today

1. Proof Sketch of LSVI

2. LSVI in Offline RL
Theorem: There exists a way to construct datasets $\{\mathcal{D}_h\}_{h=0}^{H-1}$, such that with probability at least $1 - \delta$, we have:

$$V^{\hat{\pi}} - V^* \leq \epsilon$$

w/ total number of samples in these datasets scaling $\tilde{O}\left(d^2 + H^6d^2/\epsilon^2\right)$
Theorem

**Theorem**: There exists a way to construct datasets \( \{ \mathcal{D}_h \}_{h=0}^{H-1} \), such that with probability at least \( 1 - \delta \), we have:

\[
V_{\hat{\pi}} - V^* \leq \epsilon
\]

w/ total number of samples in these datasets scaling \( \tilde{O}\left(d^2 + H^6d^2/\epsilon^2\right) \)

1. How to actively design / construct datasets \( \mathcal{D}_h \) via the Generative Model property
Theorem

**Theorem**: There exists a way to construct datasets $\{D_h\}_{h=0}^{H-1}$, such that with probability at least $1 - \delta$, we have:

$$V^{\hat{\pi}} - V^* \leq \epsilon$$

w/ total number of samples in these datasets scaling $\widetilde{O} \left( d^2 + H^6 d^2 / \epsilon^2 \right)$

1. How to actively design / construct datasets $D_h$ via the Generative Model property

2. Show that our estimators are near-bellman consistent: $\|\theta_h^T \phi - \mathcal{T}_h(\theta_{h+1}^T \phi)\|_\infty$ is small
Theorem

**Theorem:** There exists a way to construct datasets \( \{ \mathcal{D}_h \}_{h=0}^{H-1} \), such that with probability at least \( 1 - \delta \), we have:

\[
V_{\hat{\pi}} - V^* \leq \epsilon
\]

w/ total number of samples in these datasets scaling \( \widetilde{O}\left( d^2 + H^6 d^2 / \epsilon^2 \right) \)

1. How to actively design / construct datasets \( \mathcal{D}_h \) via the Generative Model property
2. Show that our estimators are near-bellman consistent: \( \| \theta_h^T \phi - \mathcal{T}_h(\theta_{h+1}^T \phi) \|_\infty \) is small
3. Near-Bellman consistency implies near optimal performance (s.t. \( H \) error amplification)
Detour: D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume $\text{span}(\mathcal{X}) = \mathbb{R}^d$).
Detour: D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume $\text{span}(\mathcal{X}) = \mathbb{R}^d$)

**D-optimal Design** $\rho^* \in \Delta(\mathcal{X})$: $\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} \left[ xx^T \right] \right)$
Detour: D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume $\text{span}(\mathcal{X}) = \mathbb{R}^d$)

**D-optimal Design** $\rho^* \in \Delta(\mathcal{X})$: $\rho^* = \arg\max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} \left[ xx^T \right] \right)$

Properties of the D-optimal Design:

$\text{support}(\rho^*) \leq d(d + 1)/2$
Detour: D-optimal Design

Consider a compact space \( \mathcal{X} \subset \mathbb{R}^d \) (without loss of generality, assume \( \text{span}(\mathcal{X}) = \mathbb{R}^d \))

**D-optimal Design** \( \rho^* \in \Delta(\mathcal{X}) \):

\[
\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} [xx^T] \right)
\]

**Properties of the D-optimal Design:**

\[
\text{support}(\rho^*) \leq d(d + 1)/2
\]

\[
\max_{y \in \mathcal{X}} y^T \left[ \mathbb{E}_{x \sim \rho^*} xx^T \right]^{-1} y \leq d
\]
Detour: OLS w/ D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume $\text{span}(\mathcal{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^* \in \Delta(\mathcal{X})$: $\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} [xx^T] \right)$
Detour: OLS w/ D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume $\text{span}(\mathcal{X}) = \mathbb{R}^d$)

**D-optimal Design** $\rho^* \in \Delta(\mathcal{X})$: $\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} [xx^T] \right)$

We actively construct a dataset $\mathcal{D}$, which contains $\lceil \rho^*(x)N \rceil$ many copies of $x$
Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume $\text{span}(\mathcal{X}) = \mathbb{R}^d$)

D-optimal Design $\rho^* \in \Delta(\mathcal{X})$:  
$$\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} [xx^T] \right)$$

We actively construct a dataset $\mathcal{D}$, which contains $[\rho^*(x)N]$ many copies of $x$.

For each $x \in \mathcal{D}$, query $y$ (noisy measure);
Detour: OLS w/ D-optimal Design

Consider a compact space \( \mathcal{X} \subset \mathbb{R}^d \) (without loss of generality, assume \( \text{span}(\mathcal{X}) = \mathbb{R}^d \))

**D-optimal Design** \( \rho^* \in \Delta(\mathcal{X}) \): \( \rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} \left[ xx^T \right] \right) \)

We **actively** construct a dataset \( \mathcal{D} \), which contains \( \lceil \rho^*(x)N \rceil \) many copies of \( x \)

For each \( x \in \mathcal{D} \), query \( y \) (noisy measure);

The OLS solution \( \hat{\theta} \) on \( \mathcal{D} \) has the following point-wise guarantee: w/ prob \( 1 - \delta \)

\[
\max_{x \in \mathcal{X}} \left| \langle \hat{\theta} - \theta^*, x \rangle \right| \leq \frac{\sigma d \ln(1/\delta)}{\sqrt{N}}
\]
Detour: OLS w/ D-optimal Design

Consider a compact space $\mathcal{X} \subset \mathbb{R}^d$ (without loss of generality, assume $\text{span}(\mathcal{X}) = \mathbb{R}^d$)

**D-optimal Design** $\rho^* \in \Delta(\mathcal{X})$: $\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} \left[ xx^T \right] \right)$

We **actively** construct a dataset $\mathcal{D}$, which contains $\lceil \rho^*(x)N \rceil$ many copies of $x$.

For each $x \in \mathcal{D}$, query $y$ (noisy measure);

The OLS solution $\hat{\theta}$ on $\mathcal{D}$ has the following point-wise guarantee: w/ prob $1 - \delta$

$$\max_{x \in \mathcal{X}} \left| \langle \hat{\theta} - \theta^*, x \rangle \right| \leq \frac{\sigma d \ln(1/\delta)}{\sqrt{N}}$$

$$\left| (\hat{\theta} - \theta^*)^T x \right| \leq \left\| \Lambda^{1/2}(\hat{\theta} - \theta^*) \right\|_2 \left\| \Lambda^{-1/2} x \right\|_2$$
Summary so far on OLS & D-optimal Design

D-optimal Design $\rho^* \in \Delta(\mathcal{X})$: $\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \text{det} \left( \mathbb{E}_{x \sim \rho} [xx^T] \right)$
Summary so far on OLS & D-optimal Design

D-optimal Design $\rho^* \in \Delta(\mathcal{X})$: $\rho^* = \arg \max_{\rho \in \Delta(\mathcal{X})} \ln \det \left( \mathbb{E}_{x \sim \rho} [xx^T] \right)$

D-optimal design allows us to actively construct a dataset $\mathcal{D} = \{x, y\}$, such that OLS solution is POINT-WISE accurate:

$$\max_{x \in \mathcal{X}} \left| \langle \hat{\theta} - \theta^*, x \rangle \right| \leq \frac{\sigma d \ln(1/\delta)}{\sqrt{N}}$$
Using D-optimal design to construct $\mathcal{D}_h$ in LSVI

Consider the space $\Phi = \{\phi(s, a) : s, a \in S \times A\}$
Using D-optimal design to construct $\mathcal{D}_h$ in LSVI

Consider the space $\Phi = \{\phi(s, a) : s, a \in S \times A\}$

**D-optimal Design** $\rho^* \in \Delta(\Phi)$: $\rho^* = \arg \max_{\rho \in \Delta(\Phi)} \ln \det \left( \mathbb{E}_{s,a \sim \rho} \left[ \phi(s,a) \phi(s,a)^T \right] \right)$
Using D-optimal design to construct $\mathcal{D}_h$ in LSVI

Consider the space $\Phi = \{\phi(s, a) : s, a \in S \times A\}$

**D-optimal Design** $\rho^* \in \Delta(\Phi)$: $\rho^* = \arg \max_{\rho \in \Delta(\Phi)} \ln \det \left( \mathbb{E}_{s, a \sim \rho} [\phi(s, a)\phi(s, a)^\top] \right)$

Construct $\mathcal{D}_h$ that contains $[\rho(s, a)N]$ many copies of $\phi(s, a)$, for each $\phi(s, a)$, \textbf{query} $y := r(s, a) + V_{h+1}(s'), s' \sim P_h(. | s, a)$
Using D-optimal design to construct $\mathcal{D}_h$ in LSVI

Consider the space $\Phi = \{\phi(s, a) : s, a \in S \times A\}$

**D-optimal Design** $\rho^* \in \Delta(\Phi)$: $\rho^* = \arg \max_{\rho \in \Delta(\Phi)} \ln \det \left( \mathbb{E}_{s, a \sim \rho} [\phi(s, a)\phi(s, a)^T] \right)$

Construct $\mathcal{D}_h$ that contains $[\rho(s, a)N]$ many copies of $\phi(s, a)$, for each $\phi(s, a)$, query $y := r(s, a) + V_{h+1}(s'), s' \sim P_h(., | s, a)$

What’s the Bayes optimal $\mathbb{E}[y | s, a]$?
Using D-optimal design to construct $\mathcal{D}_h$ in LSVI

Consider the space $\Phi = \{\phi(s,a) : s, a \in S \times A\}$

**D-optimal Design** $\rho^* \in \Delta(\Phi):$ $\rho^* = \arg \max_{\rho \in \Delta(\Phi)} \ln \det \left( \mathbb{E}_{s,a \sim \rho} [\phi(s,a)\phi(s,a)^\top] \right)$

Construct $\mathcal{D}_h$ that contains $\lfloor \rho(s,a)N \rfloor$ many copies of $\phi(s,a)$, for each $\phi(s,a)$, query $y := r(s,a) + V_{h+1}(s'), s' \sim P_h(\cdot | s,a)$

What's the Bayes optimal $\mathbb{E}[y | s,a]$?

OLS /w D-optimal design implies that $\theta_h$ is point-wise accurate:

$$\max_{s,a} \left| \theta_h^\top \phi(s,a) - \mathcal{T}_h(\theta_{h+1})^\top \phi(s,a) \right| \leq \tilde{O}\left( H d / \sqrt{N} \right).$$
Concluding the proof of LSVI

1. OLS /w D-optimal design implies that $\theta_h$ is point-wise accurate:

$$\max_{s,a} \left| \theta_h \phi(s,a) - \mathcal{T}_h(\theta_{h+1})^\top \phi(s,a) \right| \leq O\left(\frac{hd}{\sqrt{N}}\right).$$
Concluding the proof of LSVI

1. OLS /w D-optimal design implies that $\theta_h$ is point-wise accurate:

$$\max_{s,a} \left| \theta_h^\top \phi(s, a) - T_h(\theta_{h+1})^\top \phi(s, a) \right| \leq O\left(\frac{Hd}{\sqrt{N}}\right).$$

2. This implies that our estimator $Q_h := \theta_h^\top \phi$ is nearly Bellman-consistent, i.e.,

$$\| Q_h - T_h Q_{h+1} \|_{\infty} \leq O\left(\frac{Hd}{\sqrt{N}}\right)$$
Concluding the proof of LSVI

1. OLS /w D-optimal design implies that $\theta_h$ is point-wise accurate:

$$\max_{s,a} \left| \theta_h^\top \phi(s, a) - \mathcal{T}_h(\theta_{h+1})^\top \phi(s, a) \right| \leq O\left( \frac{Hd}{\sqrt{N}} \right).$$

2. This implies that our estimator $Q_h := \theta_h^\top \phi$ is nearly Bellman-consistent, i.e.,

$$\| Q_h - \mathcal{T}_h Q_{h+1} \|_\infty \leq O\left( \frac{Hd}{\sqrt{N}} \right)$$

3. Nearly-Bellman consistency implies $Q_h$ is close to $Q_h^*$ (this holds in general)

$$\| Q_h - Q_h^* \|_\infty \leq O\left( \frac{H^2d}{\sqrt{N}} \right)$$
Concluding the proof of LSVI

1. OLS /w D-optimal design implies that $\theta_h$ is point-wise accurate:

$$\max_{s,a} \left| \theta_h^\top \phi(s, a) - \mathcal{T}_h(\theta_{h+1})^\top \phi(s, a) \right| \leq O\left(\frac{Hd}{\sqrt{N}}\right).$$

2. This implies that our estimator $Q_h := \theta_h^\top \phi$ is nearly Bellman-consistent, i.e.,

$$\| Q_h - \mathcal{T}_h Q_{h+1} \|_\infty \leq O\left(\frac{Hd}{\sqrt{N}}\right)$$

3. Nearly-Bellman consistency implies $Q_h$ is close to $Q^*_h$ (this holds in general)

$$\| Q_h - Q^*_h \|_\infty \leq O\left(\frac{H^2d}{\sqrt{N}}\right)$$

$$\Rightarrow V^* - \hat{V} \leq \tilde{O}\left(\frac{H^3d}{\sqrt{N}}\right)$$
Outline for Today

1. Proof Sketch of LSVI

2. LSVI in Offline RL
Offline Reinforcement Learning
Offline Reinforcement Learning

Learner cannot interact with the environment, instead, learner is given static datasets:

\[ \mathcal{D}_h = \{s, a, r, s'\}, \quad s, a \sim \nu, \quad r = r(s, a), \quad s' \sim P_h(\cdot \mid s, a) \]
Offline Reinforcement Learning

Learner **cannot interact** with the environment, instead, learner is given static datasets:

$$\mathcal{D}_h = \{s, a, r, s'\}, \quad s, a \sim \nu, r = r(s, a), s' \sim P_h(\cdot \mid s, a)$$

**Offline Distribution** (e.g., maybe is $d^{\pi_b}$ for some behavior policy $\pi_b$)
Offline Reinforcement Learning

Learner cannot interact with the environment, instead, learner is given static datasets:

$$\mathcal{D}_h = \{s, a, r, s'\}, \quad s, a \sim \nu, r = r(s, a), s' \sim P_h(\cdot | s, a)$$

Offline Distribution (e.g., maybe is $d^{\pi_b}$ for some behavior policy $\pi_b$)

Offline RL is promising for safety critical applications (i.e., learning from logged data for health applications...)

Recall Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ 
$\mathcal{D}_h = \{s, a, r, s'\}, r = r(s, a), s' \sim P_h(\cdot|s, a)$

Set $V_H(s) = 0, \forall s$

For $h = H-1$ to 0:

$$\theta_h = \arg\min_{\theta} \sum_{\mathcal{D}_h} \left( \theta^T \phi(s, a) - (r + V_{h+1}(s')) \right)^2$$

Set $V_h(s) := \max_a \theta_h^T \phi(s, a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^T \phi(s, a), \forall h$
Recall Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/ $\mathcal{D}_h = \{s, a, r, s'\}$, $r = r(s, a), s' \sim P_h(\cdot | s, a)$

Set $V_H(s) = 0, \forall s$

For $h = H-1$ to 0:

$$\theta_h = \arg\min_{\theta} \sum_{\mathcal{D}_h} \left( \theta^T \phi(s, a) - (r + V_{h+1}(s')) \right)^2$$

Set $V_h(s) := \max_a \theta_h^T \phi(s, a), \forall s$

Return $\hat{\pi}_h(s) = \arg\max_a \theta_h^T \phi(s, a), \forall h$

LSVI directly can directly operate in offline model!
Least-Square Value Iteration Guarantee

Recall $\mathcal{D}_h = \{s, a, r, s'\}, s, a \sim \nu, r = r(s, a), s' \sim P_h( \cdot | s, a)$
Least-Square Value Iteration Guarantee

Recall $\mathcal{D}_h = \{s, a, r, s'\}$, $s, a \sim \nu$, $r = r(s, a)$, $s' \sim P_h(\cdot | s, a)$

Assumptions
1. Full offline data coverage: $\sigma_{\min}(\mathbb{E}_{s,a\sim\nu}\phi(s,a)\phi(s,a)^\top) \geq \kappa$
2. Linear Bellman completion
Least-Square Value Iteration Guarantee

Recall $\mathcal{D}_h = \{s, a, r, s'\}, s, a \sim \nu, r = r(s, a), s' \sim P_h(\cdot | s, a)$

Assumptions
1. Full offline data coverage: $\sigma_{\min} \left( \mathbb{E}_{s, a \sim \nu} \phi(s, a) \phi(s, a)^T \right) \geq \kappa$
2. Linear Bellman completion

Then, with probability at least $1 - \delta$, LSVI return $\hat{\pi}$ with $V^* - V^{\hat{\pi}} \leq \epsilon$, using at most $\text{poly} \left( H, 1/\epsilon, 1/\kappa, d, \ln(1/\delta) \right)$
The proof for the offline set is almost identical

Key step:
Linear Bellman completion + Linear Regression w/ full data coverage

=> Near-Bellman consistency, i.e., $\|Q_h - \mathcal{T}_h Q_{h+1}\|_{\infty}$ is small
The proof for the offline set is almost identical

Key step:
Linear Bellman completion + Linear Regression w/ full data coverage
=> Near-Bellman consistency, i.e., \( \| Q_h - \mathcal{T}_h Q_{h+1} \|_\infty \) is small

E.g., with \( N \) training examples where \((s, a) \sim \nu\), and \( r = r(s, a)\), \( s' \sim P_h(\cdot | s, a)\), we have

\[
\mathbb{E}_{s,a \sim \nu} \left( \theta_h^\top \phi(s, a) - \mathcal{T}_h(\theta_{h+1})^\top \phi(s, a) \right)^2 \leq \text{poly}(H, d, 1/N)
\]
The proof for the offline set is almost identical

**Key step:**
Linear Bellman completion + Linear Regression w/ full data coverage

=> Near-Bellman consistency, i.e., \( \| Q_h - \mathcal{T}_h Q_{h+1} \|_\infty \) is small

e.g., with N training examples where \((s, a) \sim \nu\), and \( r = r(s, a), s' \sim P_h(\cdot | s, a)\), we have

\[
\mathbb{E}_{s, a \sim \nu} \left( \theta_h^T \phi(s, a) - \mathcal{T}_h(\theta_{h+1})^T \phi(s, a) \right)^2 \leq \text{poly}(H, d, 1/N)
\]

Then with Cauchy-Schwartz, we get

\[
\forall s, a, \quad \left| (\theta_h - \mathcal{T}_h(\theta_{h+1}))^T \phi(s, a) \right| \leq \| \theta_h - \mathcal{T}_h(\theta_{h+1}) \|_\Sigma \| \phi(s, a) \|_{\Sigma^{-1}}
\]
The proof for the offline set is almost identical

**Key step:**
Linear Bellman completion + Linear Regression w/ full data coverage

=> Near-Bellman consistency, i.e., \( \|Q_h - \mathcal{T}_h Q_{h+1}\|_\infty \) is small

e.g., with N training examples where \((s, a) \sim \nu\), and \( r = r(s, a), s' \sim P_h(\cdot | s, a)\), we have

\[
\mathbb{E}_{s,a \sim \nu} \left( \theta_h^T \phi(s, a) - \mathcal{T}_h(\theta_{h+1})^T \phi(s, a) \right)^2 \leq \text{poly}(H, d, 1/N)
\]

Then with Cauchy-Schwartz, we get

\[
\forall s, a, \quad \left| (\theta_h - \mathcal{T}_h(\theta_{h+1}))^T \phi(s, a) \right| \leq \|\theta_h - \mathcal{T}_h(\theta_{h+1})\|_\Sigma \|\phi(s, a)\|_{\Sigma^{-1}}
\]

(we will give a HW question on a related topic)
Summary

1. Linear Bellman Completion definition (a strong assumption, though captures some models)
Summary

1. Linear Bellman Completion definition (a strong assumption, though captures some models)

2. Least square value iteration: integrate Linear regression into DP, i.e., $Q_h := \theta_h^\top \phi \approx Q_h^*$ via

$$
\phi(s, a) \mapsto r(s, a) + \max_{a'} \theta_{h+1}^\top \phi(s', a')
$$
Summary

1. Linear Bellman Completion definition (a strong assumption, though captures some models)

2. Least square value iteration: integrate Linear regression into DP, i.e., $Q_h := \theta_h^\top \phi \approx Q_h^*$ via

   $$\phi(s, a) \mapsto r(s, a) + \max_{a'} \theta_{h+1}^\top \phi(s', a')$$

3. Leverage D-optimal design, we make sure that $\theta_h$ is point-wise accurate, which ensures near Bellman consistent, i.e., $\| Q_h - \mathcal{T}_h Q_{h+1} \|_\infty$ is small
Summary

1. Linear Bellman Completion definition (a strong assumption, though captures some models)

2. Least square value iteration: integrate Linear regression into DP, i.e., $Q_h := \theta_h^\top \phi \approx Q_h^\star$ via

$$
\phi(s, a) \mapsto r(s, a) + \max_{a'} \theta_{h+1}^\top \phi(s', a')
$$

3. Leverage D-optimal design, we make sure that $\theta_h$ is point-wise accurate, which ensures near Bellman consistent, i.e.,

$$
\| Q_h - T_h Q_{h+1} \|_\infty
$$
is small

4. Near-Bellman consistency implies small approximation error of $Q_h$ (holds in general)
Next week

**Exploration**: Multi-armed Bandits and online learning in Tabular MDP