Learning with Linear Bellman Completion & Generative Model

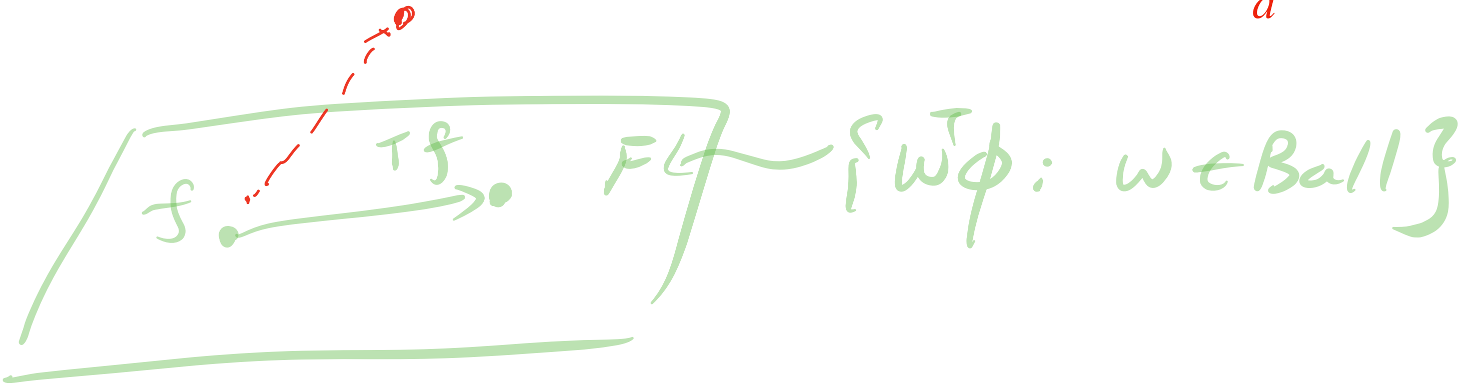
Wen Sun

CS 6789: Foundations of Reinforcement Learning

Recap: Linear Bellman Completion

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

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



Recap: Linear Bellman Completion

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

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

It implies that $Q_h^\star$ is linear in ϕ: $Q_h^\star = (\theta_h^\star)^\top \phi, \forall h$

Recap: Linear Bellman Completion

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

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

It implies that $Q_h^\star$ is linear in ϕ: $Q_h^\star = (\theta_h^\star)^\top \phi, \forall h$

Captures Tabular MDPs, and Linear Quadratic Regulators

Recap: Linear Bellman Completion

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

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

It implies that $Q_h^\star$ is linear in ϕ: $Q_h^\star = (\theta_h^\star)^\top \phi, \forall h$

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/

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

Recap: Least-Square Value Iteration

Datasets $\mathscr{D}_0, \ldots, \mathscr{D}_{H-1}$, w/

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

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

Recap: Least-Square Value Iteration

Datasets $\mathscr{D}_0, \ldots, \mathscr{D}_{H-1}$, w/

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

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

For h = H-1 to 0:

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

$\approx V^\star_{h+1}(s')$

$\theta_h^T \phi \approx Q^\star_h$

Recap: Least-Square Value Iteration

Datasets $\mathscr{D}_0, \ldots, \mathscr{D}_{H-1}$, w/

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

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

For h = H-1 to 0:

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

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

Recap: Least-Square Value Iteration

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

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

For h = H-1 to 0:

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

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

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

Recap: Least-Square Value Iteration

Datasets $\mathscr{D}_0, \ldots, \mathscr{D}_{H-1}$, w/

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

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

For h = H-1 to 0:

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

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

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

BC always ensures linear regression is realizable:

i.e., our regression target

$$r(s,a) + \mathbb{E}_{s' \sim P_h(s,a)} \max_{a'} \theta_{h+1}^\top \phi(s', a')$$

is always linear:

Outline for Today

1. Proof Sketch of LSVI

2. LSVI in Offline RL

Theorem

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

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

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

Theorem

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

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

$V^{\pi} = V_0^{\pi}(s_0)$

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 $\mathscr{D}_h$ via the Generative Model property

Theorem

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

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

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 $\mathscr{D}_h$ via the Generative Model property
2. Show that our estimators are near-bellman consistent: $\|\theta_h^\top \phi - \mathscr{T}_h(\theta_{h+1}^\top \phi)\|_\infty$ is small

Theorem

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

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

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 $\mathscr{D}_h$ via the Generative Model property
2. Show that our estimators are near-bellman consistent: $\|\theta_h^\top \phi - \mathscr{T}_h(\theta_{h+1}^\top \phi)\|_\infty$ 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 span$(\mathcal{X}) = \mathbb{R}^d$)

Detour: D-optimal Design

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

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

Detour: D-optimal Design

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

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

Properties of the D-optimal Design:

support$(\rho^\star) \leq d(d+1)/2$

Detour: D-optimal Design

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

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

Properties of the D-optimal Design:

support$(\rho^\star) \le d(d+1)/2$

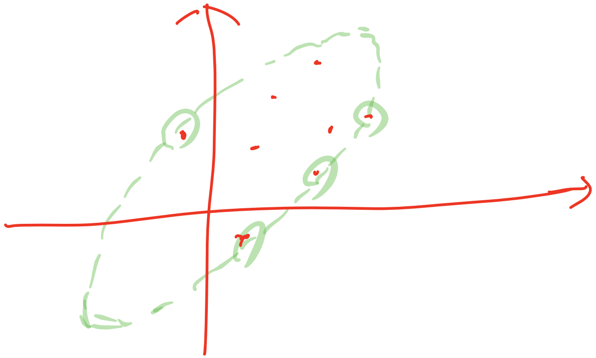
$$\max_{y\in\mathcal{X}} y^\top \left[\mathbb{E}_{x\sim\rho^\star} xx^\top\right]^{-1} y \le d$$

$$\sum_{i=1}^{d} \frac{1}{\sigma_i}(y^\top u_i)^2 \le d$$

$$\Sigma^* = \mathbb{E}_{x\sim\rho^*} xx^\top$$

$$V = \{x : x^\top (\Sigma^*)^{-1} x \le d\}$$

poly(d, ln



Detour: OLS w/ D-optimal Design

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

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

Detour: OLS w/ D-optimal Design

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

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

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

Detour: OLS w/ D-optimal Design

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

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

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

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

Detour: OLS w/ D-optimal Design

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

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

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

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

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

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

Detour: OLS w/ D-optimal Design

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

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

We **actively** construct a dataset $\mathcal{D}$, which contains $\lceil \rho^\star(x) N \rceil$ 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 $1-\delta$

$\Lambda = \frac{1}{N}\sum_{i=1}^{N} x_i x_i^\top$

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

$$\left|(\hat\theta - \theta^\star)^\top x\right| \le \left\| \Lambda^{1/2}(\hat\theta - \theta^\star)\right\|_2 \left\| \Lambda^{-1/2} x \right\|_2$$

$(\hat\theta - \theta^\star)^\top \Lambda^{\frac{1}{2}} \Lambda^{-\frac{1}{2}} x$

$\le \sqrt{d}$ ← second property of $\rho^\star$

Summary so far on OLS & D-optimal Design

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

Summary so far on OLS & D-optimal Design

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

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

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

Using D-optimal design to construct $\mathscr{D}_h$ in LSVI

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

Using D-optimal design to construct $\mathscr{D}_h$ in LSVI

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

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

Using D-optimal design to construct $\mathscr{D}_h$ in LSVI

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

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

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

⇒ Generative Access

Using D-optimal design to construct $\mathscr{D}_h$ in LSVI

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

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

Construct $\mathscr{D}_h$ that contains $\lceil \rho(s,a) N \rceil$ 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 $\mathbb{E}[y\,|\,s,a]$?

$= r + \mathbb{E}_{s' \sim P(s,a)} V_{h+1}(s')$

$\because$ BC, linear in ϕ

Using D-optimal design to construct $\mathscr{D}_h$ in LSVI

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

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

noisy label

Construct $\mathscr{D}_h$ that contains $\lceil \rho(s,a)N \rceil$ 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 $\mathbb{E}[y\,|\,s,a]$?

OLS /w D-optimal design implies that θ_h is point-wise accurate:

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

$\approx Q_h^\star$ $\approx Q_{h+1}^\star$

Concluding the proof of LSVI

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

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

Concluding the proof of LSVI

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

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

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

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

Concluding the proof of LSVI

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

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

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

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

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

$$\|Q_h - Q_h^\star\|_\infty \leq O(H^2 d/\sqrt{N})$$

Concluding the proof of LSVI

$\mathcal{X} = \text{Ball} \subseteq \mathbb{R}^d$

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

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

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

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

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

$$\|Q_h - Q_h^\star\|_\infty \le O(H^2 d/\sqrt{N})$$

$$\Rightarrow V^\star - V^{\hat{\pi}} \le \widetilde{O}(H^3 d/\sqrt{N})$$

H $= \varepsilon \implies$ solve for N.

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:

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

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^{π_b} for some behavior policy π_b)

Offline Reinforcement Learning

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

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

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

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

Recall Least-Square Value Iteration

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

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

For h = H-1 to 0:

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

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

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

Recall Least-Square Value Iteration

Datasets $\mathcal{D}_0, \ldots, \mathcal{D}_{H-1}$, w/

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

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

For h = H-1 to 0:

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

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

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

LSVI directly can directly operate in offline model!

Least-Square Value Iteration Guarantee

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

Least-Square Value Iteration Guarantee

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

Assumptions

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

$\not\Rightarrow \sup_{s,a} \nu(s,a) \geq c$

Least-Square Value Iteration Guarantee

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

Assumptions

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

Then, with probability at least $1 - \delta$, LSVI return $\hat{\pi}$ with $V^\star - V^{\hat{\pi}} \leq \epsilon$, using at most 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 $r = r(s, a), s' \sim P_h(\cdot \mid s, a)$, 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 $r = r(s, a), s' \sim P_h(\cdot \mid s, a)$, we have

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

Then with Cauchy-Schwartz, we get

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

$\Sigma = \mathbb{E}_{s,a\sim\nu} \phi\phi^\top$

Training

$\leq \frac{1}{\kappa}$

$(\theta_h - \mathcal{T}_h(\theta_{h+1}))^\top \Sigma^{\frac{1}{2}} \Sigma^{-\frac{1}{2}} \phi(s,a)$

$\sigma_{min}(\Sigma) \geq \kappa$

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 $r = r(s, a), s' \sim P_h(\cdot \mid s, a)$, we have

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

Then with Cauchy-Schwartz, we get

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

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

Summary

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

$$\|\Lambda^{\frac{1}{2}}(\theta - \theta^*)\|_2 = (\theta - \theta^*)^T \Lambda (\theta - \theta^*)$$

$$\Lambda = \frac{1}{N}\sum_{i=1}^{n} x_i x_i^T$$

$$= \frac{1}{N}\sum_i (x_i^T\theta - x_i^T\theta^*)^2$$

Tese x:

$$(\theta - \theta^*)^T x \le \|\theta - \theta^*\|_\Lambda \|x\|_{\Lambda^{-1}}$$

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

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

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 θ_h is point-wise accurate, which ensures near Bellman consistent, i.e., $\left\| Q_h - \mathcal{T}_h Q_{h+1} \right\|_\infty$ 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 θ_h is point-wise accurate, which ensures near Bellman consistent, i.e., $\left\| Q_h - \mathcal{T}_h Q_{h+1} \right\|_\infty$ is small

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

$\| Q_h - Q_h^\star \|$ being small

$\Rightarrow \| V^\pi - V^\star \|$ being small

Next week

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