CS292F StatRL Lecture 15 Uniform OPE and Near-Optimal Offline Learning

Instructor: Yu-Xiang Wang

Spring 2021

UC Santa Barbara

Logistics

- Homework 3 is released.
  - Due end of the quarter / June 2
  - 5 questions, but you only need to do either Q4 or Q5.

- Two more lectures on offline RL.
  - including this one.

- I will send schedules for the project presentations this week

Recap: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)

Offline Trajectory data D Collected by running $\boldsymbol{\mu}$

Task: design OPE methods

Evaluate fixed Target Policy $\boldsymbol{\pi}$

Via Uniform OPE

- Task 2: Offline Policy Learning. (OPL)

Offline Trajectory data D Collected by running $\boldsymbol{\mu}$

Task: design OPO methods

Find near optimal Policy $\widehat{\boldsymbol{\pi}}^*$

OPE: the tabular MDP case, they are all equivalent.

- TMIS

$$\widehat{v}^{\pi}_{\mathrm{MIS}} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\widehat{d}^{\pi}_t(s_t^{(i)})}{\widehat{d}^{\mu}_t(s_t^{(i)})}\widehat{r}^{\pi}_t(s^{(i)}).$$

- Model-based Plugin

$$\hat{v}^{\pi}_{\mathrm{DM}} = \sum_{h=1}^{H}\sum_{s\in\mathcal{S}}\hat{d}^{\pi}_h(s)\hat{r}^{\pi}_h(s)$$

- Fitted Q Iteration

$$\hat{v}^{\pi}_{\mathrm{FQI}} = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\hat{d}_1(s)\pi(a|s)\hat{Q}_1(s,a)$$

They are also information-theoretically optimal

$d_m = \min_{s,a,t} d^{\mu}(s,a)$

$s \cdot h, \exists \pi, d^{\pi}(s,a) > 0$

Offline Policy Evaluation

Simulation lemma (Kearns and Singh, 1998)	IS / DR (Jiang and Li, 2016)	MIS (Xie, Ma, W.,2019)	TMIS (Yin & W. 2020)	Fitted Q-Iteration (Duan and Wang, 2020)
$\sqrt{\frac{H^4 S^2}{n d_m}}$	$\sqrt{\frac{e^H poly(S,A)}{n}}$	$\sqrt{\frac{H^3}{n\, d_m}}$	$\sqrt{\frac{H^2}{n\, d_m}}$	$\sqrt{\frac{H^2}{n\, d_m}}$

Per-instance optimal.

$d_m = \frac{1}{SA}$

$\sqrt{\frac{H^2 SA}{n}}$

$n = \frac{T}{H}$

$$\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2]$$

$$\leq \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h, a_h} \frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \mathrm{Var}\left[(V_{h+1}^{\pi}(s_{h+1}^{(1)}) + r_h^{(1)}) \middle| s_h^{(1)} = s_h, a_h^{(1)} = a_h \right]$$

$$+ O(n^{-1.5})$$

Matching Cramer-Rao lower bound up to low-order terms.

Recap: From OPE to offline learning

- Empirical Risk Minimization (ERM)?

$$\hat{\pi} = \arg\max_{\pi \in \Pi} \hat{v}^{\pi}$$ (For some OPE estimator $\hat{v}^{\pi}$)

$$\pi^* = \arg\max_{\pi \in \Pi} v^{\pi}$$

- A uniform convergence argument

$$\sup_{\pi \in \Pi} |\hat{v}^{\pi} - v^{\pi}| \leq \epsilon \quad \textit{w.h.p.}$$

$$\Rightarrow$$

$$v^{\pi^*} - v^{\hat{\pi}} \leq 2\epsilon \quad \textit{w.h.p}$$

$$v^{\pi^*} - \hat{v}^{\pi^*} + \underbrace{\hat{v}^{\pi^*} - \hat{v}^{\hat{\pi}}}_{\leq 0} + \hat{v}^{\hat{\pi}} - v^{\hat{\pi}}$$

$$\mathbb{E}\left[\sup_{\pi \in \Pi} |\hat{v}^{\pi} - v^{\pi}|^2\right] \leq \epsilon^2$$

$$v^{\pi^*} - \mathbb{E}[v^{\hat{\pi}}] \leq 2\epsilon$$

Recap: exploration assumption

- The logging policy μ is out of our control

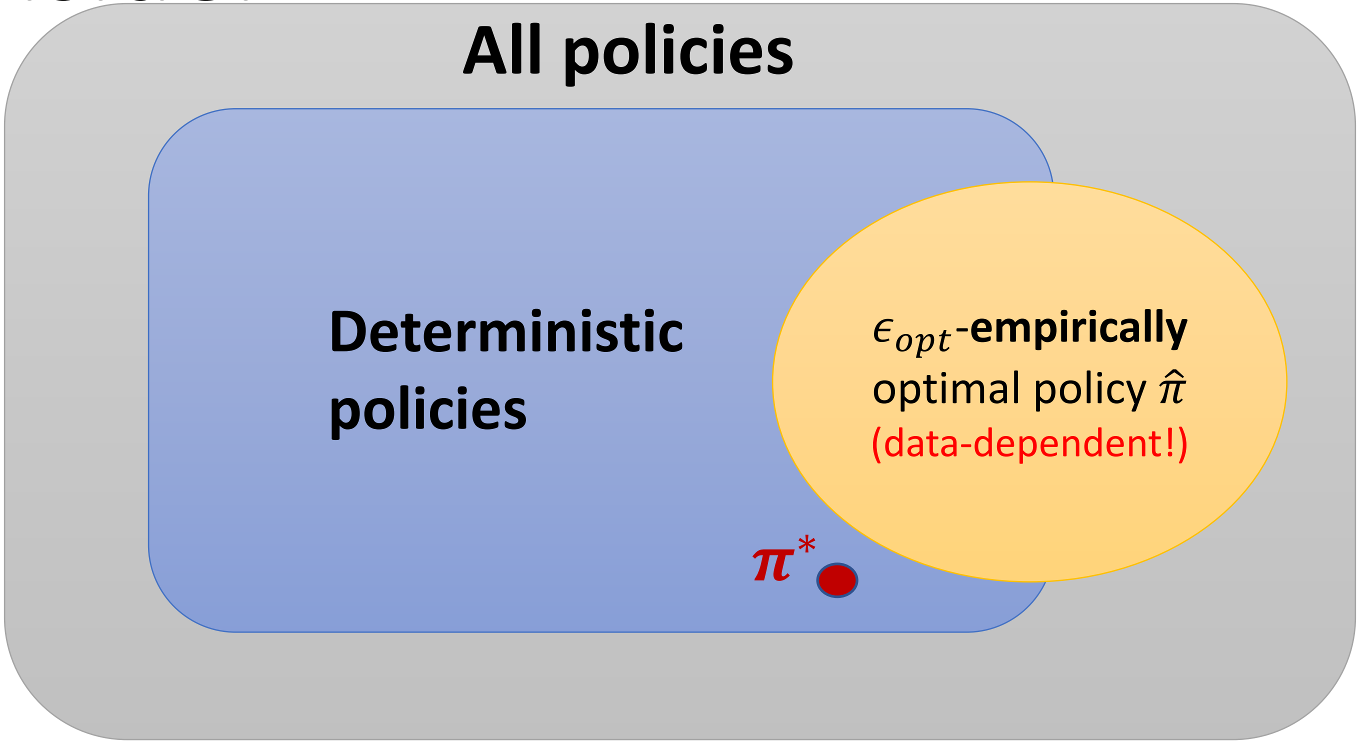
- Need to make assumptions about it

$$d_m := \min_{t,s,a} d_t^{\mu}(s,a) > 0 \text{ for all } t, s, a$$

$$\text{s.t. } d_t^{\pi}(s,a) > 0 \text{ for some } \pi \in \Pi$$

  - Assumed to simplify the discussion on optimality
  - Sometimes appear only in low-order terms.

Recap: The policy classes we consider



For ERM, it suffices to consider the smaller policy class.
But we also want to cover other planning algorithms.

The remainder of the lecture is based on:
Yin, Bai and W. (2020) https://arxiv.org/pdf/2007.03760.pdf

This lecture

- Characterize the uniform OPE on deterministic policy class
- Optimal offline learning via a local uniform OPE

Recap: counting the number of deterministic policies

- Setting: Tabular MDP with S states, A actions and H steps.

$P_1 \dots P_H$ differen

$r_t(s,a)$ ... differ at t

- How many deterministic policies are there?

(Answer: A^{SH})

- Together with a $\boxed{\text{high-probability OPE bound}}$

fixed π_1 $\quad \left|\hat{V}^{\pi_1} - V^{\pi_1}\right| < O\left(\sqrt{\frac{H^2 \log\frac{1}{\delta}}{n \cdot d_m}}\right)$ w.p. $1-\delta$

then

union bound $\quad \delta = \frac{\delta'}{A^{SH}}$

w.p. $1-\delta'$

$$\sup_{\pi \in \Pi} \left|\hat{V}^{\pi} - V^{\pi}\right| \le O\left(\sqrt{\frac{H^3 S \log\frac{A}{\delta'}}{n \cdot d_m}}\right)$$

Uniform convergence theorem for all deterministic policies

Theorem 3.5: with probability $\geq 1-\delta$

$$\sup_{\pi\in\Pi_{\text{deterministic}}} |\hat{v}^{\pi}-v^{\pi}| \lesssim \sqrt{\frac{H^3 S}{n d_m}\log(\frac{HSA}{\delta})} + O(1/n)$$

- Optimal in H.
- Suboptimal in S?

- Proof: Union bound with a high-probability pointwise OPE bound.

How do we obtain a high-probability pointwise OPE bound?

- Steps in the analysis

  $\hat{v}$, $\hat{Q}$, $\hat{V}$, $\hat{P}_t$, $\hat{r}_t$, $\hat{d}_t^{\pi}$, $\hat{d}_t^{\mu}$

  1. Fictitious estimator technique and multiplicative Chernoff bounds.
     - Nice event: $E_t := \{n_{s_t,a_t} \geq n d_t^{\mu}(s_t, a_t)/2\}$

  2. Error decomposition (reducing to the case with known reward function)

  3. Further decomposition of the occupancy measure into a Martingale.

     $\hat{d}_t^{\pi}(s) - d_t^{\pi}(s)$

  4. Apply Freedman's inequality --- a Bernstein-style Martingale Concentration.

Step 1 Recap: The fictitious estimator is easier to analyze, because:

- Always unbiased.
- Has an *epistemical* Bellman-equation of variance
- Has nice martingale decompositions
- Moreover: Lemma C.3

$$\sup_{\pi\in\Pi}\left|\tilde{v}^{\pi}-\hat{v}^{\pi}\right|=0 \qquad \text{w.h.p.}$$

Under mild condition: $n \gtrsim \frac{1}{d_m}\log\frac{HSA}{\delta}$

Step 2 Recap: The noise in the reward is straightforward to handle.

$$\begin{aligned}
\sup_{\pi\in\Pi} |\widetilde{v}^\pi - v^\pi| &= \sup_{\pi\in\Pi} \Big|\sum_{t=1}^{H}\langle \widetilde{d}_t^\pi, \widetilde{r}_t\rangle - \sum_{t=1}^{H}\langle d_t^\pi, r_t\rangle\Big| \\
&= \sup_{\pi\in\Pi} \Big|\sum_{t=1}^{H}\langle \widetilde{d}_t^\pi, \widetilde{r}_t\rangle - \sum_{t=1}^{H}\langle \widetilde{d}_t^\pi, r_t\rangle + \sum_{t=1}^{H}\langle \widetilde{d}_t^\pi, r_t\rangle - \sum_{t=1}^{H}\langle d_t^\pi, r_t\rangle\Big| \\
&\le \underbrace{\sup_{\pi\in\Pi} \Big|\sum_{t=1}^{H}\langle \widetilde{d}_t^\pi - d_t^\pi, r_t\rangle\Big|}_{(*)} + \underbrace{\sup_{\pi\in\Pi} \Big|\sum_{t=1}^{H}\langle \widetilde{d}_t^\pi, \widetilde{r}_t - r_t\rangle\Big|}_{(**)}
\end{aligned}$$

Lemma C.4: $(**) \lesssim \sqrt{H^2/(nd_m)}$

Therefore, it suffices to consider the case with **deterministic rewards**.

Step 3 Recap: Martingale decomposition of the error $\tilde{v}^{\pi} - v^{\pi}$

Primal representation (Marginal distribution style):

$$\sum_{t=1}^{H} \langle \widetilde{d_t^{\pi}} - d_t^{\pi}, r_t \rangle$$

$\hat{d}_t^{\tilde{\pi}_t}(s) \cdot \tilde{\pi}_t(a|s)$

$\hat{d}_t^{\tilde{\pi}_t} \in \mathbb{R}^{S \times A}$

|| (Lemma C.5)

Dual representation (Value function style):

$\in \mathbb{R}^{S \times SA}$

$$\langle V_1^{\pi}(s), (\widetilde{d_1^{\pi}} - d_1^{\pi})(s) \rangle + \sum_{h=2}^{H} \langle V_h^{\pi}(s), ((\widetilde{T}_h - T_h)\widetilde{d_{h-1}^{\pi}})(s) \rangle$$

$\hat{P}_h(s_h'|s_{h-1},a_{h-1})$ $P_h(s_h'|s_{h-1},a_{h-1})$

(You can prove this by simulation lemma, see HW3 Q5.)

Let's check that this is a Martingale (w.r.t. the parallel data sequence)

$D_0 = \phi \quad D_t : D_{t-1} \cup \{ s_t^{(i)}, a_t^{(i)}, r_t^{(i)} \}_{i=1}^{n}$

Recall definition of Martingale:

a. E[X_t | D_{t-1}] = X_{t-1}

b. E[|X_t|] is bounded

$$E\left[X_t \,\middle|\, D_{t-1} \right] := \sum_{h=2}^{t} \langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi \rangle + \langle V_1^\pi, \widetilde{d}_1^\pi - d_1^\pi \rangle .$$

fixed

$$= E\left[\langle V_t^\pi, (\widetilde{T}_t^n - T_t) \widetilde{d}_{t-1}^{n,\pi} \rangle \,\middle|\, D_{t-1} \right] + \sum_{h=2}^{t-1} \langle V_h^\pi, (\widetilde{T}_h^n - T_h) \widetilde{d}_{h-1}^{n,\pi} \rangle$$

Random fixed

$= X_{t-1}$

$$E(|X_t|) \le H^2$$

Step 4: Freedman's inequality

Lemma A.6 (Freedman's inequality Tropp et al. (2011)). *Let X be the martingale associated with a filter $\mathcal{F}$ (i.e. $X_i = \mathbb{E}[X|\mathcal{F}_i]$) satisfying $|X_i - X_{i-1}| \le M$ for $i = 1, ..., n$. Denote $W := \sum_{i=1}^{n} \mathrm{Var}(X_i|\mathcal{F}_{i-1})$ then we have*

$$\mathbb{P}(|X - \mathbb{E}[X]| \ge \epsilon, W \le \sigma^2) \le 2e^{-\frac{\epsilon^2}{2(\sigma^2 + M\epsilon/3)}}.$$

Or in other words, with probability $1 - \delta$,

$$|X - \mathbb{E}[X]| \le \sqrt{8\sigma^2 \cdot \log(1/\delta)} + \frac{2M}{3} \cdot \log(1/\delta), \quad Or \quad W \ge \sigma^2.$$

Bernstein $\left|\frac{X}{n} - \frac{\mathbb{E}X}{n}\right| \le O\left(\frac{\sqrt{\mathrm{Var}(X_i - X_{i-1}) \log\frac{1}{\delta}}}{n} + \frac{M \log\frac{1}{\delta}}{n}\right)$

- To apply this inequality, we need to
  - bound M
  - Need to work out the variance.

*In fact we will use a more flexible version of Freedman's inequality due to (Chung and Liu 2006) that allows the bound M to hold w.h.p rather than with prob 1.

Bound M in high probability

- Lemma E.2 With prob at least 1-δ

$$\sup_t |X_t - X_{t-1}| \le O\left(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_m}}\right).$$

Proof:

$$|X_t - X_{t-1}| = \langle V_t^\pi, (\widetilde{T}_t - T_t)\widetilde{d}^\pi_{t-1} \rangle$$

$$\le ||(\widetilde{T}_t - T_t)^T V_t^\pi||_\infty ||\widetilde{d}^\pi_{t-1}||_1 = ||(\widetilde{T}_t - T_t)^T V_t^\pi||_\infty .$$

$$= \mathbb{1}(E_{t-1}) \max_{s,a} \left| \sum_{s_t} \left(\widetilde{T}_t(s_t|s,a) - T_t(s_t|s,a)\right) \cdot V_t^\pi(s_t) \right| \qquad \left| \sum_{s_t} \widetilde{T}_t(s_t|s,a) V_t^\pi(s_t) \right|$$

$$= \mathbb{1}(E_{t-1}) \max_{s,a} \left| \sum_{s_t} \frac{1}{n_{s,a,t-1}} \sum_{i=1}^{n} \mathbb{1}(s_{t-1}^{(i)}=s, a_{t-1}^{(i)}=a, s_t^{(i)}=s_t) \cdot V_t^\pi(s_t) \right|$$

$$= \mathbb{1}(E_{t-1}) \max_{s,a} \left| \frac{1}{n_{s,a,t-1}} \sum_{i: s_{t-1}^{(i)}=s,\, a_{t-1}^{(i)}=a} V_t^\pi(s_t^{(i)}) - E[V_t^\pi | s,a] \right|,$$

Hoeffding's inequality, union bound

$W \geq \sum_{t=2}^{H} \mathrm{Var}(X_{t+1} | \mathcal{D}_t)$

Bound W: Sum of conditional variance

Lemma E.3. *We have the following decomposition of conditional variance:*

$$\mathrm{Var}[X_{t+1}|\mathcal{D}_t] = \sum_{s_t, a_t} \frac{\widetilde{d}_t^\pi(s_t, a_t)^2 \cdot \mathbb{1}(E_t)}{n_{s_t,a_t}} \cdot \mathrm{Var}[V_{t+1}^\pi(s_{t+1}^{(1)})|s_t^{(1)} = s_t, a_t^{(1)} = a_t]$$

Proof: $\mathrm{Var}\left[\sum_{s_t,a_t} \sum_{s_{t+1}} V_{t+1}^{\pi}(s_{t+1}) (\hat{T}_{t+1} - T_{t+1})(s_{t+1}|s_t,a_t) \cdot \hat{d}_t^{\pi}(s_t,a_t) \,\middle|\, \mathcal{D}_t\right]$ fixed

$= \mathbb{1}(E_t) \sum_{s_t a_t} \mathrm{Var}\left(\sum_{s_{t+1}} V_{t+1}^{\pi}(s_{t+1}) (\hat{T}_{t+1} - T_{t+1})(s_{t+1}|s_t a_t) \,\middle|\, \mathcal{D}_t\right) (\hat{d}_t^{\pi}(s_t a_t))^2$

$= \mathbb{1}(E_t) \sum_{s_t a_t} \mathrm{Var}\left(\frac{1}{n_{s_t a_t t}} \sum_{\substack{i\, s.t.\\ s_t^{(i)}=s_t\\ a_t^{(i)}=a_t}} V_{t+1}^{\pi}(s_{t+1}^{(i)}) \,\middle|\, \mathcal{D}_t\right) \cdot (\hat{d}_t^{\pi}(s_t a_t))^2$

$V_{t+1}^{\pi}(s_{t+1}^{(i)})$ iid samples

$= \mathbb{1}(E_t) \sum_{s_t a_t} \frac{\hat{d}_t^{\pi}(s_t a_t)^2}{n_{s_t a_t t}} \mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)}) | s_t^{(1)} = s_t, a_t^{(1)} = a_t]$

$\mathrm{Var}(X+Y) = \mathrm{Var}(X) + \mathrm{Var}(Y)$

if $X \perp Y$

s_t, a_t s_t', a_t'

$s_t^{(i)}$ $s_t'^{(i)}$

$\mathrm{Var}(x) = \mathrm{Var}(X - EX)$

$\mathrm{Var}(cX) = c^2 \mathrm{Var}(X)$

$\mathrm{Var}(\frac{1}{n}\sum X_i) = \frac{1}{n}\mathrm{Var}(X_1)$

Bound W: Sum of conditional variance

$\mathcal{E}_t = \{ n_{s_t,a_t,t} \geq \frac{1}{2} d_t^{\mu}(s_t,a_t) \ \forall s_t,a_t,t \}$

- If we can bound $\widetilde{d}_t^{\pi}(s_t, a_t) \le C \cdot d_t^{\pi}(s_t,a_t)$

- Then, we can apply Lemma 3.4 from the last lecture

$$\sum_{t=1}^{H} \mathrm{Var}[X_{t+1}|\mathcal{D}_t] \le O\left(\frac{1}{nd_m} \cdot \sum_{t=1}^{H} \mathbb{E}[\mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)}, a_t^{(1)}]]\right) \le \boxed{O\left(\frac{H^2}{nd_m}\right)}$$

$$\sum_{t=1}^{H} \sum_{s_t,a_t} \frac{\mathbb{1}(\mathcal{E}_t)\cdot d_t^{\mu\pi}(s_t,a_t)^2}{n_{s_t,a_t,t}} \mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)}=s_t, a_t^{(1)}=a_t]$$

$$\frac{c \cdot P_{max} H^2}{n}$$

$$\sum_{t=1}^{H} \sum_{s_t,a_t} \frac{c^2 d_t^{\pi}(s_t,a_t)^2}{\frac{1}{2} d_t^{\mu}(s_t,a_t)} \mathrm{Var}[V_{t+1}^{\pi}(s_{t+1}^{(1)})|s_t^{(1)}=s_t, a_t^{(1)}=a_t] \le \frac{2 P_{max} c^2 \mathrm{Var}[\sum_t r_t^{(1)}]}{n} =$$

Bounding $\widetilde{d}_t^{\pi}(s_t, a_t)$

- Martingale decomposition

$$\widetilde{d}_t^{\pi}(s_t, a_t) - d_t^{\pi}(s_t, a_t) = \sum_{h=2}^{t} (\Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^{\pi})(s_t, a_t) + (\Gamma_{1:t}(\widetilde{d}_1^{\pi} - d_1^{\pi}))(s_t, a_t),$$

- Bounded Martingale difference w.h.p

$$\sup_h ||\Gamma'_{h:t}(\widetilde{T}_h - T_h)||_\infty \le O(\sqrt{\frac{1}{n \cdot d_m} \log \frac{H^2 S^2 A^2}{\delta}}).$$

- (Chung and Liu, 06)-style Azuma-Hoeffding, and union bound

$$\sup_t ||\widetilde{d}_t^{\pi} - d_t^{\pi}||_\infty \le O(\sqrt{\frac{H}{n d_m} \log \frac{H^2 S^2 A^2}{\delta} \log \frac{HSA}{\delta}}).$$

Completing the analysis

$\|d_t^{u\pi} - d_t^{\pi}\|_\infty \le \sqrt{\frac{H}{n d_m}\log(\cdot)}$

- Bounding W

$(d_t^{u\pi})^2 = (d_t^{u\pi} - d_t^{\pi} + d_t^{\pi})^2 \le 2 \cdot (d_t^{\pi})^2 + \underbrace{2(d_t^{u\pi} - d_t^{\pi})^2}_{O\left(\frac{H}{n d_m}\log(\cdot)\right)}$

$\sum_{t=1}^{H} \mathrm{Var}(X_t | D_{t-1}) = \sigma^2 = W \le \underbrace{O\left(\frac{\rho_{max} H^2}{n}\right)}_{\frac{1}{n d_m}} + \underbrace{O\left(\frac{H}{n d_m} \cdot \frac{\rho_{max} H^2}{n}\right)}_{O\left(\frac{1}{n^2}\right)}$

- Bounding M

$$\sup_t |X_t - X_{t-1}| \le O\left(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_m}}\right).$$

Uniform over all π deterministic $\sqrt{\frac{H^3 S \log\frac{HSA}{\delta}}{n d_m}}$

- Apply (Chung and Liu's) Freedman's inequality

$$|X - \mathbb{E}[X]| \le \sqrt{8\sigma^2 \cdot \log(1/\delta)} + \frac{2M}{3} \cdot \log(1/\delta),$$

$O\left(\sqrt{\frac{1}{n d_m}}\right) + O\left(\frac{1}{n d_m}\right) + O\left(\frac{1}{n^{1.5}}\right)$

Uniform convergence theorem for all policies

$\epsilon \gtrsim \sqrt{\frac{H^3}{nd_m}}$ $\sqrt{\frac{H^3 S}{nd_m} \log \frac{HSA}{\delta}}\,(1+o(1))$

Theorem 3.3: with probability $\geq 1 - \delta$

$$\sup_{\pi \in \Pi} |\hat{v}^\pi - v^\pi| \lesssim \sqrt{\frac{H^4}{nd_m} \log(\frac{HSA}{\delta})} + \sqrt{\frac{H^4 S}{nd_m} \log(SA)}$$

- Optimal in S if $\delta < e^{-S}$, suboptimal in H.

- Proof idea: Martingale decomposition over H. Freedman's inequality. Rademacher complexity argument.

Uniform convergence theorem for near-empirically optimal policies

Value function on $\hat{M}_\epsilon$

Theorem 3.7: Let $\Pi_1 := \{\pi : s.t. \, || \hat{V}_t^{\pi} - \hat{V}_t^{\hat{\pi}^*} ||_\infty \leq \epsilon_{opt}, \forall \, t \in [H]\}$. Assume $\epsilon_{opt} \leq \sqrt{H}/S$, and also let $n \gtrsim H^2/d_m$. Then w.p. $\geq 1-\delta$,

$$\sup_{\pi \in \Pi_1} \left\| \widehat{Q}_1^{\pi} - Q_1^{\pi} \right\|_\infty \leq c_2 \sqrt{\frac{H^3 \log(HSA/\delta)}{n \cdot d_m}}.$$

$\langle \hat{Q}_1^\pi, \hat{d}_1^\pi \rangle = \hat{V}^\pi$

- Optimal in all parameters.
- Implies optimal learning bounds for ERM by taking ϵ_{opt} = 0

- Proof idea: A cute argument that takes the empirical optimal policy as an anchor point.

Local uniform convergence is sufficient for

Assume generative model

Offline Policy Learning

Simulation lemma (Kearns and Singh, 1998)	MSBO (Xie and Jiang, 2020)	Variance-Reduction (Sidford et al, 19), (Wainwright, 19)	Model-based (Agarwal, Kakade, Yang, 20)	Model-based via UniformOPE
$\sqrt{\frac{H^4 S^2}{n d_m}}$	$\sqrt{\frac{H^4}{n d_m}}$	$\sqrt{\frac{H^3 SA}{n}}$	$\sqrt{\frac{H^3 SA}{n}} + H \cdot \epsilon_{opt}$	$\sqrt{\frac{H^3}{n\, d_m}} + \epsilon_{opt}$

Converted from infinite horizon case…

Proof sketch:

$\cancel{\sum_t \langle \hat{d}_t^{\pi} - d_t^{\bar{\pi}}, r_t \rangle}$

- Apply Simulation Lemma in a different way

$$\widehat{Q}_t^{\pi} - Q_t^{\pi} = \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} (\widehat{P}_h^{\pi} - P_h^{\pi}) \widehat{Q}_h^{\pi}$$
$$= \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} (\widehat{P}_h - P_h) \widehat{V}_h^{\pi}$$

HW3 Q5

$\hat{Q}^{\pi} - Q^{\pi}$

or $Q^{\pi} - \hat{Q}^{\pi}$

- Error decomposition

$$\left| \widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}} \right| \leq \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left| (\widehat{P}_h - P_h) \widehat{V}_h^{\widehat{\pi}} \right|$$
$$\leq \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left| (\widehat{P}_h - P_h) \widehat{V}_h^{\widehat{\pi}^\star} \right|}_{(***)} + \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left| (\widehat{P}_h - P_h) (\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}}) \right|}_{(****)}$$

Bounding (****)

$$\left\| \sum_{h=t+1}^{H} \Gamma^{\widehat{\pi}}_{t+1:h-1} \cdot \left| (\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}}) \right| \right\|_\infty \le H \cdot \sup_h \left\| \Gamma^{\widehat{\pi}}_{t+1:h-1} \right\|_\infty \left\| |(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})| \right\|_\infty$$

$$\le H \cdot \sup_h \left\| |(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})| \right\|_\infty$$

$$\sup_h \left\| |(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})| \right\|_\infty \le \epsilon_{\text{opt}} \cdot \sup_h \left\| |\widehat{P}_h - P_h| \cdot \mathbf{1} \right\|_\infty$$

$\widehat{\pi}$ is in the local-policy class

fix s, a

$$\langle \widehat{P}_h(\cdot|s,a) - P_h(\cdot|s,a), \mathbb{1} \rangle \le \|\widehat{P}_h(\cdot|s,a) - P_h(\cdot|s,a)\|_1 \cdot 1 \le \sqrt{\frac{S \log(\cdot)}{n_{s,a}}}$$

- Apply the local-uniform assumption
- Apply L1-norm error bound (Recall from Lecture 3)

Slight improvement over the parameters in the original paper…

Bounding (***) $\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\widehat{\pi}} \left| (\widehat{P}_h - P_h) \widehat{V}_h^{\widehat{\pi}^\star} \right|$.

- Key observation: Conditioning on D_{h-1}
  - $\widehat{P}_h$ depends only on the data at step h
  - $\widehat{V}_h^{\widehat{\pi}^\star}$ (s_h) depends only on the data after h
- Results in us saving a factor of S!

$$\left| (\widehat{P}_h - P_h) \widehat{V}_h^{\widehat{\pi}^\star} \right| (s_{t-1}, a_{t-1}) \leq 4\sqrt{\frac{\log(1/\delta)}{N}} \sqrt{\mathrm{Var}(\widehat{V}_h^{\widehat{\pi}^\star})(s_{t-1}, a_{t-1})} + \frac{4(H-t)}{3N} \log(\frac{1}{\delta})$$

$$= \frac{1}{n_{s,a}} \sum_{i \text{ s.t. } s_h^{(i)} = s,\, a_h^{(i)} = a} \hat{V}_h^{\hat{\pi}^\star}(s_{h+1}^{(i)})$$

$$N \leq n_{s,a} = \frac{d_h \cdot n}{2}$$

Putting things together

$$\left|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right| \le \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left|(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}}\right|$$

$$\le \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left|(\widehat{P}_h - P_h)\widehat{V}_h^{\widehat{\pi}^\star}\right|}_{(***)} + \underbrace{\sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\pi} \left|(\widehat{P}_h - P_h)(\widehat{V}_h^{\widehat{\pi}^\star} - \widehat{V}_h^{\widehat{\pi}})\right|}_{(****)}$$

$$\le \sum_{h=t+1}^{H} \Gamma_{t+1:h-1}^{\widehat{\pi}} \left(4\sqrt{\frac{\log(HSA/\delta)}{N}}\sqrt{\mathrm{Var}(\widehat{V}_h^{\widehat{\pi}^\star})} + \frac{4(H-t)}{3N}\log(\frac{HSA}{\delta}) \cdot \mathbf{1}\right)$$

$$+ c_1\epsilon_{\mathrm{opt}} \cdot \sqrt{\frac{H^2S^2\log(HSA/\delta)}{N}} \cdot \mathbf{1}$$

$< \widehat{Q}_t^{\widehat{\pi}} - Q^{\widehat{\pi}}$

$\mathrm{Var}(\widehat{V}_h^{\widehat{\pi}^\star}) < \mathrm{Var}(V_h^{\widehat{\pi}^\star})$

- Bounding the sum of variance
  - Triangular inequality then similar to that of Lemma 3.4

Lastly, backward recursion from H to 1, using the following theorem

Theorem F.4. *Conditional on $N > 0$, then with probability $1-\delta$, we have for all $t = 1, \ldots, H-1$*

$$\left\|\widehat{Q}_t^{\widehat{\pi}} - Q_t^{\widehat{\pi}}\right\|_\infty \le 4\sqrt{\frac{H^3 \log(HSA/\delta)}{N}} + 4\sqrt{\frac{\log(HSA/\delta)}{N}} \sum_{h=t+1}^{H} \left\|\widehat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}}\right\|_\infty + \frac{4H^2}{3N}\log(\frac{HSA}{\delta})$$
$$+ c_2 \epsilon_{opt} \cdot \sqrt{\frac{H^2 S^2 \log(HSA/\delta)}{N}}.$$

- Assume: $N \ge 64H^2 \cdot \log(HSA/\delta)$ *and* $\epsilon_{opt} \le \sqrt{H}/S$

- We obtain the final result:

$$\left\|\widehat{Q}_1^{\widehat{\pi}} - Q_1^{\widehat{\pi}}\right\|_\infty \le 2(9 + c_2)\sqrt{\frac{H^3 \log(HSA/\delta)}{N}}$$

Summary

$\sqrt{\frac{S H^3}{n d_m}}$ global

- We finished Offline RL for the tabular setting

- Via local uniform convergence, we showed that the model-based approach / ERM is minimax optimal

$\sqrt{\frac{H^3}{n d_m}}$ optimal

- We covered all essential tricks in the theoretical analysis

What I did not cover

- Optimal offline RL in
  - Infinite horizon case
  - Finite horizon stationary case
  - (Yin and W., 2021a)

- Optimal rates in global uniform OPE
  - (Yin and W., 2021b)

$$\sqrt{\frac{H^3 S}{n d_m}}$$

- Offline RL with function approximation