CS292F StatRL Lecture 14
From OPE to Uniform OPE

Instructor: Yu-Xiang Wang
Spring 2021
UC Santa Barbara
Recap: Offline Reinforcement Learning, aka. Batch RL

• Task 1: Offline Policy Evaluation. (OPE)
  - Offline Trajectory data $D$
  - Collected by running $\mu$
  - Task: design OPE methods
  - Evaluate fixed Target Policy $\pi$

• Task 2: Offline Policy Learning. (OPL)
  - Offline Trajectory data $D$
  - Collected by running $\mu$
  - Task: design OPO methods
  - Find near optimal Policy $\hat{\pi}^*$

Via Uniform OPE
Recap: Lecture 13

• Offline Policy Evaluation for RL
  • Focused on Finite-Horizon Episodic MDPs
  • Offline data collected by running logging / behavioral policy
  • Evaluate a target policy

• OPE RL estimators:
  • Trajectory-IS, Per-Step IS, DR
  • Curse of Horizon
  • Overcoming the curse of horizon with MIS / TMIS
  • Connections to DM --- a model-based approach.
Recap: Notations

• Finite Horizon Episodic MDP

• Offline Data

• Policies / importance weights
Recap: Main challenge of OPE in RL: The curse of Horizon

\[ \hat{v}^{\pi}_{IS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} \left[ \prod_{t=1}^{h} \frac{\pi(a_t^{(i)}|s_t^{(i)})}{\mu(a_t^{(i)}|s_t^{(i)})} \right] r_{h}^{(i)}. \]

The curse of horizon. (Liu et al, 2018 NeurIPS)

• The variance is exponential in H!
Recap: Marginalized Importance Sampling

\[
\hat{v}_{\text{IS}}^\pi = \frac{1}{n} \sum_{i=1}^{n} \sum_{h=1}^{H} \left[ \prod_{t=1}^{H} \frac{\pi(a_t^{(i)} \mid s_t^{(i)})}{\mu(a_t^{(i)} \mid s_t^{(i)})} \right] r_h^{(i)}.
\]

\[
\hat{v}_{\text{MIS}}^\pi = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{H} \frac{\hat{d}_t^{\pi}(s_t^{(i)})}{\hat{d}_t^{\mu}(s_t^{(i)})} \hat{r}_t^{\pi}(s_t^{(i)}).
\]

Recap: Recursive estimation of State-visitation of target policy

\[ d_t^\pi(s_t) = \sum_{s_{t-1}} P_t^\pi(s_t|s_{t-1})d_{t-1}^\pi(s_{t-1}), \]

\[ \hat{d}_t^\pi = \hat{P}_t^\pi \hat{d}_{t-1}, \]

where \( \hat{P}_t^\pi(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}} \sum_{i=1}^n \frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\mu(a_{t-1}^{(i)}|s_{t-1})} \mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}) = (s_{t-1}, s_t)); \]

\[ \hat{r}_t^\pi(s_t) = \frac{1}{n_{s_t}} \sum_{i=1}^n \frac{\pi(a_t^{(i)}|s_t)}{\mu(a_t^{(i)}|s_t)} r_t^{(i)} \mathbf{1}(s_t^{(i)} = s_t), \]

Recap: Tabular MIS (finite action space)

\[ \hat{v}_{\text{MIS}}^{\pi} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{H} \frac{\hat{d}_t^{\pi}(s_t^{(i)})}{\hat{d}_t^{\mu}(s_t^{(i)})} \hat{r}_t^{\pi}(s^{(i)}). \]

- With a minor change to the following recursive estimation

\[ \hat{P}_t^{\pi}(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}} \sum_{i=1}^{n} \pi(a_{t-1}^{(i)}|s_{t-1}) \cdot 1((s_{t-1}^{(i)}, s_t^{(i)}, a_t^{(i)}) = (s_{t-1}, s_t, a_t)); \]

\[ \hat{r}_t^{\pi}(s_t) = \frac{1}{n_{s_t}} \sum_{i=1}^{n} \pi(a_t^{(i)}|s_t) \cdot \mu(a_t^{(i)}|s_t) \cdot 1(s_t^{(i)} = s_t). \]

Changed to Empirical Estimate.
Recap: OPE error bound of MIS and Tabular MIS

- The MSE of MIS estimator obeys:

\[
\frac{1}{n} \sum_{t=1}^{H} \mathbb{E}_\mu \left[ \frac{d^\pi_t(s_t)^2}{d^\mu_t(s_t)^2} \text{Var}_\mu \left[ \frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)} (V^\pi_{t+1}(s_{t+1}) + r_t) \bigg| s_t \right] \right] + \tilde{O}(n^{-1.5})
\]


- TMIS obeys:

\[
\mathbb{E}[\left( \tilde{v}^\pi_{\text{TMIS}} - v^\pi \right)^2] \\
\leq \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h,a_h} \frac{d^\pi_h(s_h)^2}{d^\mu_h(s_h)^2} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \text{Var} \left[ (V^\pi_{h+1}(s^{(1)}_{h+1}) + r^{(1)}_h) \bigg| s^{(1)}_h = s_h, a^{(1)}_h = a_h \right] \\
+ O(n^{-1.5})
\]

Recap: Asymptotic rates for MIS and TMIS

• Assumptions:

**MIS:** \[ \frac{1}{n} \sum_{t=1}^{H} \mathbb{E}_{\mu} \left[ \frac{d_{t}^{\pi}(s_{t})^{2}}{d_{t}^{\mu}(s_{t})^{2}} \frac{\pi_{t}(a_{t}|s_{t})}{\mu_{t}(a_{t}|s_{t})} \left( V_{t+1}^{\pi}(s_{t+1}) + r_{t} \right) | s_{t} \right] \] + \tilde{O}(n^{-1.5})

**TMIS** \[ \frac{1}{n} \sum_{h=0}^{H} \sum_{s_{h},a_{h}} \frac{d_{h}^{\pi}(s_{h})^{2}}{d_{h}^{\mu}(s_{h})^{2}} \frac{\pi(a_{h}|s_{h})^{2}}{\mu(a_{h}|s_{h})} \text{Var} \left[ (V_{h+1}^{\pi}(s_{h+1}) + r_{h}^{(1)}) | s_{h}^{(1)} = s_{h}, a_{h}^{(1)} = a_{h} \right] \] + \tilde{O}(n^{-1.5})
Recap: TMIS vs on-policy evaluation

Lemma 3.4. For any policy $\pi$ and any MDP.

$$
\Var_\pi \left[ \sum_{t=1}^{H} r_t^{(1)} \right] = \sum_{t=1}^{H} \left( \E_\pi \left[ \Var \left[ r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)} | s_t^{(1)}, a_t^{(1)}) \right] \right] \\
+ \E_\pi \left[ \Var \left[ \E[r_t^{(1)} + V_{t+1}^{\pi}(s_{t+1}^{(1)} | s_t^{(1)}, a_t^{(1)}) | s_t^{(1)}] \right] \right] \right).
$$

Combined with the previous observation:

1. TMIS has an error that is linear in $H$.
2. TMIS is better than MC even when we are doing on-policy evaluation
How do MIS / TMIS overcome the curse of horizon?

• Leverage the model assumption: MDP

• We visit the same state many times.

• As a matter of fact, TMIS is equivalent to a model-based approach
Recap: TMIS is equivalent to DM -- a model-based approach

\[
\hat{V}_{MS} = \frac{1}{n} \sum_{i=1}^{n} \sum_{\tau} \frac{d}{d\tau} \left( \frac{1}{\sigma_{\tau}^2} \right) \left( \frac{c_{\tau}(s_{\tau}^{(i)})}{\sigma_{\tau}^2} \right) = \sum_{s_{\tau}} \frac{1}{n} \sum_{i=1}^{n} \frac{d}{d\tau} \left( \frac{1}{\sigma_{\tau}^2} \right) \frac{c_{\tau}(s_{\tau}^{(i)})}{\sigma_{\tau}^2} = \hat{V}_{DM}
\]
This lecture

• Fitted Q Iterations

• Uniform Convergence in OPE in RL
Fitted Q Iterations (Munos, 2003) (Munos & Szepesvari, 2008)

• Recall Bellman Optimality equation and the Bellman operator
  \[ \mathcal{T} f(s, a) := r(s, a) + \mathbb{E}_{s' \sim P(\cdot | s, a)} \max_{a' \in A} f(s', a'). \]

• Given offline transition data and a function class
  \[ \text{FQI: } f_t \in \arg \min_{f \in \mathcal{F}} \sum_{i=1}^{n} \left( f(s'_i, a_i) - r_i - \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s_i, a_i) \right)^2. \]
  
  Iteratively from some initialization.

• For the finite horizon episodic case:
Fitted Q iterations for OPE \( \text{(Duan and Wang, 2020)} \)

• Recall Bellman equation for a fixed policy

\[
Q_{h-1}^\pi(s, a) = r(s, a) + \mathbb{E}[V^\pi_h(s') | s, a]
\]

• Given offline transition data and a function class

\[
\hat{Q}_{H+1}^\pi := 0 \text{ and for } h = H, H - 1, \ldots, 0,
\]

\[
\hat{Q}_h^\pi = \arg \min_{f_h \in \mathcal{F}} \sum_{i=1}^n \left( f_h(s_h^{(i)}, a_h^{(i)}) - r_h^{(i)} - \sum_{a' \in \mathcal{A}} \pi(a'|s_{h+1}^{(i)}) f_{h+1}(s_{h+1}^{(i)}, a') \right)^2
\]
FQI in the tabular case

\[ \hat{Q}^\pi_h = \arg \min_{f_h \in \mathcal{F}} \sum_{i=1}^{n} \left( f_h(s^{(i)}_h, a^{(i)}_h) - r^{(i)}_h - \sum_{a' \in A} \pi(a' | s^{(i)}_{h+1}) f_{h+1}(s^{(i)}_{h+1}, a') \right)^2 \]

- Let’s work out the optimal solution!
In conclusion, in the tabular MDP case, they are all equivalent.

- **TMIS**
  \[
  \hat{v}^\pi_{\text{MIS}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{t=1}^{H} \frac{\hat{d}^\pi_t(s_t^{(i)})}{\hat{d}^\mu_t(s_t^{(i)})} \hat{r}^\pi_t(s^{(i)}). 
  \]

- **Model-based Plugin**
  \[
  \hat{v}^\pi_{\text{DM}} = \sum_{h=1}^{H} \sum_{s \in S} \hat{d}^\pi_h(s) \hat{r}^\pi_h(s) 
  \]

- **Fitted Q Iteration**
  \[
  \hat{v}^\pi_{\text{FQI}} = \sum_{s \in S} \sum_{a \in A} \hat{d}_1(s) \pi(a|s) \hat{Q}_1(s, a) 
  \]
Recap: Offline Reinforcement Learning, aka. Batch RL

• Task 1: Offline Policy Evaluation. (OPE)
  
  Offline Trajectory data $D$
  Collected by running $\mu$

  Task: design OPE methods

  Evaluate fixed Target Policy $\pi$

  Via Uniform OPE

• Task 2: Offline Policy Learning. (OPL)

  Offline Trajectory data $D$
  Collected by running $\mu$

  Task: design OPO methods

  Find near optimal Policy $\hat{\pi}^*$
Observation 1: OPE is in its essence a statistical estimation problem.

• But is slightly non-trivial because we are estimating a single number, when the number of parameters describing the distribution are numerous.

• Find functions of the data --- estimators, such that

\[ |\hat{u}^\pi - u^\pi| \leq \epsilon \] with high probability

\[ \mathbb{E}[|\hat{u}^\pi - u^\pi|^2] \leq \epsilon^2 \]
Observation 2: Offline Learning is a statistical learning problem

• But with a structured hypothesis class (the policy class), and structured observations (trajectories).

• Lessons from statistical learning theory:
  • ERM suffices and almost necessary.
  • In RL context this is: \( \hat{\pi} = \arg \max_{\pi \in \Pi} \hat{v}^\pi \)

• Combine with OPE:

\[
\left| \hat{\nu}^\pi - \nu^\pi \right| \leq \epsilon \text{ w.h.p.}
\]

\[
\mathbb{E} \left[ \left| \hat{\nu}^\pi - \nu^\pi \right|^2 \right] \leq \epsilon^2
\]

\[
\nu^{\pi^*} - \nu^{\hat{\pi}} \leq 2\epsilon \text{ w.h.p.}
\]

\[
\nu^{\pi^*} - \mathbb{E}[\nu^{\hat{\pi}}] \leq 2\epsilon
\]
Not quite this easy, the learned policy \( \hat{\pi} \) depends on the data

\[
\sup_{\pi \in \Pi} |\hat{\pi} - \pi| \leq \varepsilon \quad \text{w.h.p.}
\]

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

\[
\nu^* - \hat{\nu} \leq 2\varepsilon \quad \text{w.h.p.}
\]

\[
\nu^* - \mathbb{E}[\hat{\nu}] \leq 2\varepsilon
\]

In standard statistical learning: \( \varepsilon \approx \sqrt{d/n} \)

Where \( d \) is VC-dimension / metric entropy \( \log |\Pi| \), or implied by Rademacher complexity, etc.

(Much older Empirical process theory, Glivenko-Cantelli style)

Vapnik (1995)

What is a natural complexity measure for the policy class in RL?
We will not deal with exploration in offline RL, because we can’t

• The logging policy $\mu$ is out of our control

• Need to make assumptions about it

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

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

• Assumed to simplify the discussion on optimality
• Sometimes appear only in low-order terms.
The policy classes we consider

All policies

Deterministic policies

$\pi^*$

$\epsilon_{opt}$-empirically optimal policy $\hat{\pi}$ (data-dependent!)

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:
Start with the family of all deterministic policies

• The optimal policy is deterministic

• There are a finite number of them

• We have strong pointwise convergence bound from TMIS (last week)
Exercise: counting the number of deterministic policies

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

• How many deterministic policies are there?
We need a high probability version of the bound we get

**Theorem (MSE bound)** (Yin & W., 2020)

\[
\mathbb{E}[(\hat{v}_{\text{TMIS}}^{\pi} - v^{\pi})^2] \\
\leq \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h,a_h} \frac{d_{h}^{\pi}(s_h)^2}{d_{h}^{\mu}(s_h)} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \text{Var} \left[ (V_{h+1}^{\pi}(s_{h+1}) + r_{h}^{(1)}) \bigg| s_{h}^{(1)} = s_h, a_{h}^{(1)} = a_h \right]
\]

What we need is something stronger:
Fictitious estimator technique

- Fictitious estimator
  - Nice event: \( E_t := \{ n_{s_t,a_t} \geq n d_{t}^\mu(s_t, a_t)/2 \} \)
  - Define
    \[
    \tilde{r}_t(s_t, a_t) = \hat{r}_t(s_t, a_t)1(E_t) + r_t(s_t, a_t)1(E_t^c)
    \]
    \[
    \hat{P}_{t+1}(\cdot|s_t, a_t) = \hat{P}_{t+1}(\cdot|s_t, a_t)1(E_t) + P_{t+1}(\cdot|s_t, a_t)1(E_t^c).
    \]

Idea: hypothetically plug in the ground truth occasionally

\[
\tilde{P}_t^\pi(s_t|s_{t-1}) = \sum_{a_{t-1}} \tilde{P}_t(s_t|s_{t-1}, a_{t-1})\pi(a_{t-1}|s_{t-1}).
\]

\[
\tilde{\nu}^\pi := \sum_{t=1}^H \langle \tilde{d}_t^\pi, \tilde{r}_t^\pi \rangle, \text{ with } \tilde{d}_t^\pi = \tilde{P}_t^\pi \tilde{d}_{t-1}^\pi
\]
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

\[ \sup_{\pi \in \Pi} |\tilde{\nu}^\pi - \hat{\nu}^\pi| = 0 \quad \text{w.h.p.} \]

Under mild condition: \( n \geq \frac{1}{d_m} \log \frac{HSA}{\delta} \)
The noise in the reward is straightforward to handle.

\[
\sup_{\pi \in \Pi} |\tilde{v}^\pi - v^\pi| = \sup_{\pi \in \Pi} |\sum_{t=1}^{H} \langle \tilde{d}_t^\pi, \tilde{r}_t \rangle - \sum_{t=1}^{H} \langle d_t^\pi, r_t \rangle| \\
= \sup_{\pi \in \Pi} |\sum_{t=1}^{H} \langle \tilde{d}_t^\pi, \tilde{r}_t \rangle - \sum_{t=1}^{H} \langle \tilde{d}_t^\pi, r_t \rangle + \sum_{t=1}^{H} \langle \tilde{d}_t^\pi, r_t \rangle - \sum_{t=1}^{H} \langle d_t^\pi, r_t \rangle| \\
\leq \sup_{\pi \in \Pi} |\sum_{t=1}^{H} \langle \tilde{d}_t^\pi - d_t^\pi, r_t \rangle| + \sup_{\pi \in \Pi} |\sum_{t=1}^{H} \langle d_t^\pi, \tilde{r}_t - r_t \rangle| \\ 
\text{(**) \quad Lemma C.4: (**) } \approx \sqrt{H^2/(nd_m)}
\]

Therefore, it suffices to consider the case with \textbf{deterministic rewards}.
Dealing with the reward noise

**Lemma C.4.** We have with probability $1 - \delta$:

\[
\sup_{\pi \in \Pi} \left| \sum_{t=1}^{H} \langle \tilde{d}_t^\pi, \tilde{r}_t - r_t \rangle \right| \leq O\left( \sqrt{\frac{H^2 \log(\text{HSA}/\delta)}{n \cdot d_m}} \right)
\]
Martingale decomposition of the error $\tilde{\nu}^\pi - \nu^\pi$

**Primal representation (Marginal distribution style):**

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

**Dual representation (Value function style):**

$$\langle \nu_1^\pi(s), (\tilde{d}_1^\pi - d_1^\pi)(s) \rangle + \sum_{h=2}^{H} \langle \nu_h^\pi(s), ((\tilde{T}_h - T_h)d_{h-1}^\pi)(s) \rangle$$

(Lemma C.5)
Let’s derive the Martingale Decomposition

\[
\begin{align*}
\tilde{d}_t^\pi &= \pi_t \tilde{T}_t \tilde{d}_{t-1} \\
\tilde{d}_t^\pi &= \pi_t T_t d_{t-1} \\
T_t &\in \mathbb{R}^{S \times (S \cdot A)} \\
(T_t)_{s_t, (s_{t-1}, a_{t-1})} &= P_t(s_t|s_{t-1}, a_{t-1})
\end{align*}
\]

Take the difference of the two

\[
\tilde{d}_t^\pi - d_t^\pi = \pi_t (\tilde{T}_t - T_t) \tilde{d}_{t-1} + \pi_t T_t (d_{t-1}^\pi - d_{t-1}^\pi)
\]

Recursively apply the above

\[
\tilde{d}_t^\pi - d_t^\pi = \sum_{h=2}^{t} \Gamma_{h+1:t} \pi_h (\tilde{T}_h - T_h) \tilde{d}_{h-1}^\pi + \Gamma_{1:t} (d_1^\pi - d_1^\pi)
\]
Dual representation (Value function style):

\[
\langle v_1^\pi(s), (\tilde{d}_1^\pi - d_1^\pi)(s) \rangle + \sum_{h=2}^{H} \langle v_h^\pi(s), ((\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi)(s) \rangle
\]

\[
X = \sum_{t=1}^{H} \left( \sum_{h=2}^{H} \langle r_t, \Gamma_{h+1:t}^\pi_h(\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle + \langle r_t, \Gamma_{1:t}(\tilde{d}_1^\pi - d_1^\pi) \rangle \right)
\]

\[
= \sum_{t=1}^{H} \left( \sum_{h=2}^{H} \langle r_t, \Gamma_{h+1:t}^\pi_h(\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle \right) + \sum_{h=1}^{H} \langle r_t, \Gamma_{1:t}(\tilde{d}_1^\pi - d_1^\pi) \rangle
\]

\[
= \sum_{t=2}^{H} \left( \sum_{h=2}^{H} \langle r_t, \Gamma_{h+1:t}^\pi_h(\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle \right) + \sum_{h=1}^{H} \langle r_t, \Gamma_{1:t}(\tilde{d}_1^\pi - d_1^\pi) \rangle
\]

\[
= \sum_{h=2}^{H} \left( \sum_{t=h}^{H} \langle r_t, \Gamma_{h+1:t}^\pi_h(\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle \right) + \sum_{h=1}^{H} \langle (\pi_1^T \Gamma_{1:t}^T r_t)(s), (\tilde{d}_1^\pi - d_1^\pi)(s) \rangle
\]

\[
= \sum_{h=2}^{H} \left( \sum_{t=h}^{H} \langle \pi_h^T \Gamma_{h+1:t}^T r_t, (\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle \right) + \sum_{h=1}^{H} \langle (\sum_{t=1}^{H} \pi_1^T \Gamma_{1:t}^T r_t)(s), (\tilde{d}_1^\pi - d_1^\pi)(s) \rangle
\]
Let’s check that this is a Martingale

\[ X_t := \mathbb{E}[X | \mathcal{D}_t] = \sum_{h=2}^{H} \langle V_h^\pi, (\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle + \langle V_1^\pi, \tilde{d}_1^\pi - d_1^\pi \rangle. \]

\[ \mathbb{E}[X|\mathcal{D}_t] = \sum_{h=t+1}^{H} \mathbb{E} \left[ \langle V_h^\pi, (\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle | \mathcal{D}_t \right] + \sum_{h=2}^{t} \langle V_h^\pi, (\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle + \langle V_1^\pi, (\tilde{d}_1^\pi - d_1^\pi) \rangle. \]

Note for \( h \geq t + 1 \), \( \mathcal{D}_t \subset \mathcal{D}_{h-1} \), so by total law of expectation (tower property) we have

\[
\mathbb{E} \left[ \langle V_h^\pi, (\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle | \mathcal{D}_t \right] \\
= \mathbb{E} \left[ \mathbb{E} \left[ \langle V_h^\pi, (\tilde{T}_h - T_h)\tilde{d}_{h-1}^\pi \rangle | \mathcal{D}_{h-1} \right] | \mathcal{D}_t \right] \\
= \mathbb{E} \left[ \langle V_h^\pi, \mathbb{E} \left[ (\tilde{T}_h - T_h) | \mathcal{D}_{h-1} \right] \tilde{d}_{h-1}^\pi \rangle | \mathcal{D}_t \right] = 0
\]
Final results after applying a complex martingale concentration

• And a union bound.

**Theorem E.6.** With probability $1 - \delta$, we have

$$
\left| \sum_{t=1}^{H} \langle \tilde{d}_t^\pi - d_t^\pi, r_t \rangle \right| \leq O\left( \sqrt{\frac{H^2 \log(HSA/\delta)}{nd_m}} + \sqrt{\frac{H^4 SA \cdot \log(H^2 S^2 A^2/\delta) \log(HSA/\delta)}{n^2 d_m^2}} \right)
$$

where $O(\cdot)$ absorbs only the absolute constants.
Uniform convergence theorem for all deterministic policies

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

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

- Optimal in $H$, suboptimal in $S$.

- Proof: Union bound with a high-probability pointwise OPE bound.
Uniform convergence theorem for all policies

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

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

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

Uniform convergence theorem for near-empirically optimal policies

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

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

- Optimal in all parameters.
- Implies optimal learning bounds for ERM by taking \( \epsilon_{opt} = 0 \)
- Proof idea: A cute argument that takes the empirical optimal policy as an anchor point.