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 $\boldsymbol{\mu}$ → 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 $\widehat{\boldsymbol{\pi}}^*$

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

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

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

- The variance is exponential in H!

Recap: Marginalized Importance Sampling

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

$$\downarrow$$

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

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

Recap: Recursive estimation of State-visitation of target policy

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

$$\widehat{d}_t^{\pi} = \widehat{P}_t^{\pi} \widehat{d}_{t-1}^{\pi},$$

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

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

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

Recap: Tabular MIS (finite action space)

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

- With a minor change to the following recursive estimation

$$\widehat{P}^{\pi}_{t}(s_t|s_{t-1}) = \frac{1}{n_{s_{t-1}}}\sum_{i=1}^{n}\frac{\pi(a_{t-1}^{(i)}|s_{t-1})}{\boxed{\mu(a_{t-1}^{(i)}|s_{t-1})}}\cdot \mathbf{1}((s_{t-1}^{(i)}, s_t^{(i)}, a_t^{(i)}) = (s_{t-1}, s_t, a_t));$$

$$\widehat{r}^{\pi}_{t}(s_t) = \frac{1}{n_{s_t}}\sum_{i=1}^{n}\frac{\pi(a_t^{(i)}|s_t)}{\boxed{\mu(a_t^{(i)}|s_t)}} r_t^{(i)}\cdot \mathbf{1}(s_t^{(i)} = s_t).$$

Changed to Empirical Estimate.

Recap: OPE error bound of MIS and Tabular MIS

- The MSE of MIS estimator obeys:

$$\frac{1}{n}\sum_{t=1}^{H}\mathbb{E}_\mu\left[\frac{d_t^\pi(s_t)^2}{d_t^\mu(s_t)^2}\mathrm{Var}_\mu\left[\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)}\left(V_{t+1}^\pi(s_{t+1})+r_t\right)\Big|s_t\right]\right]+\tilde{O}(n^{-1.5})$$

Xie, W., and Ma. (2019): Towards Optimal OPE for RL using Marginalized Importance Sampling. NeurIPS 2019.

- TMIS obeys:

Yin & W. (2020). Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS-2020*

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

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

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

Recap: Asymptotic rates for MIS and TMIS

- Assumptions:

- MIS: $$\frac{1}{n}\sum_{t=1}^{H}\mathbb{E}_\mu\left[\frac{d_t^\pi(s_t)^2}{d_t^\mu(s_t)^2}\mathrm{Var}_\mu\left[\frac{\pi_t(a_t|s_t)}{\mu_t(a_t|s_t)}\big(V_{t+1}^\pi(s_{t+1})+r_t\big)\middle|s_t\right]\right]+\tilde{O}(n^{-1.5})$$

- TMIS $$\frac{1}{n}\sum_{h=0}^{H}\sum_{s_h,a_h}\frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\mathrm{Var}\left[(V_{h+1}^\pi(s_{h+1}^{(1)})+r_h^{(1)})\middle|s_h^{(1)}=s_h,a_h^{(1)}=a_h\right]$$
$$+O(n^{-1.5})$$

Recap: TMIS vs on-policy evaluation

Lemma 3.4. *For any policy π and any MDP.*

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

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

$$\hat{V}_{MIS} = \frac{1}{n}\sum_{i=1}^{n}\sum_{t=1}^{H}\frac{\hat{d}^{\pi}(s_t^{(i)})}{\hat{d}^{\mu}(s_t^{(i)})}\hat{r}_t^{\pi}(s_t^{(i)}) = \underline{\sum_{s\in S}\sum_{t=1}^{H}\hat{d}_t^{\pi}(s)\hat{r}_t(s)} = \hat{V}_{DM}$$

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

$$\mathcal{T} f(s,a) := r(s,a) + \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} f(s',a').$$

- Given offline transition data and a function class

$$\text{FQI:} \quad f_t \in \operatorname{argmin}_{f \in \mathcal{F}} \sum_{i=1}^{n} \left(f(s_i', a_i) - r_i - \gamma \max_{a' \in \mathcal{A}} f_{t-1}(s_i, a_i) \right)^2.$$

Iteratively from some initialization.

- For the finite horizon episodic case:

Fitted Q iterations for OPE (Duan and Wang, 2020)

- Recall Bellman equation for a fixed policy

$$Q_{h-1}^{\pi}(s,a) = r(s,a) + \mathbb{E}\left[V_h^{\pi}(s') \,\middle|\, s,a\right]$$

- Given offline transition data and a function class

$\widehat{Q}_{H+1}^{\pi} := 0$ 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

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

- Let's work out the optimal solution!

In conclusion, in the tabular MDP case, they are all equivalent.

- TMIS

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

- Model-based Plugin

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

- Fitted Q Iteration

$$\hat{v}^{\pi}_{\text{FQI}} = \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}}\hat{d}_{1}(s)\pi(a|s)\hat{Q}_{1}(s,a)$$

Recap: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)

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

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 $\widehat{\boldsymbol{\pi}}^*$

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

$$\left|\hat{v}^{\pi} - v^{\pi}\right| \leq \epsilon$$ with high probability

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

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

(For some estimator $\hat{v}^{\pi}$)

  - Combine with OPE:

$$|\hat{v}^{\pi} - v^{\pi}| \leq \epsilon \quad \textit{w.h.p.} \qquad \overset{?}{\Rightarrow} \qquad v^{\pi^*} - v^{\hat{\pi}} \leq 2\epsilon \quad \textit{w.h.p}$$

$$\mathbb{E}\left[|\hat{v}^{\pi} - v^{\pi}|^2\right] \leq \epsilon^2 \qquad \Rightarrow \qquad v^{\pi^*} - \mathbb{E}[v^{\hat{\pi}}] \leq 2\epsilon$$

Not quite this easy, the learned policy $\hat{\pi}$ depends on the data

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

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

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

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

In standard statistical learning: $\epsilon \asymp \sqrt{d/n}$

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 μ is out of our control

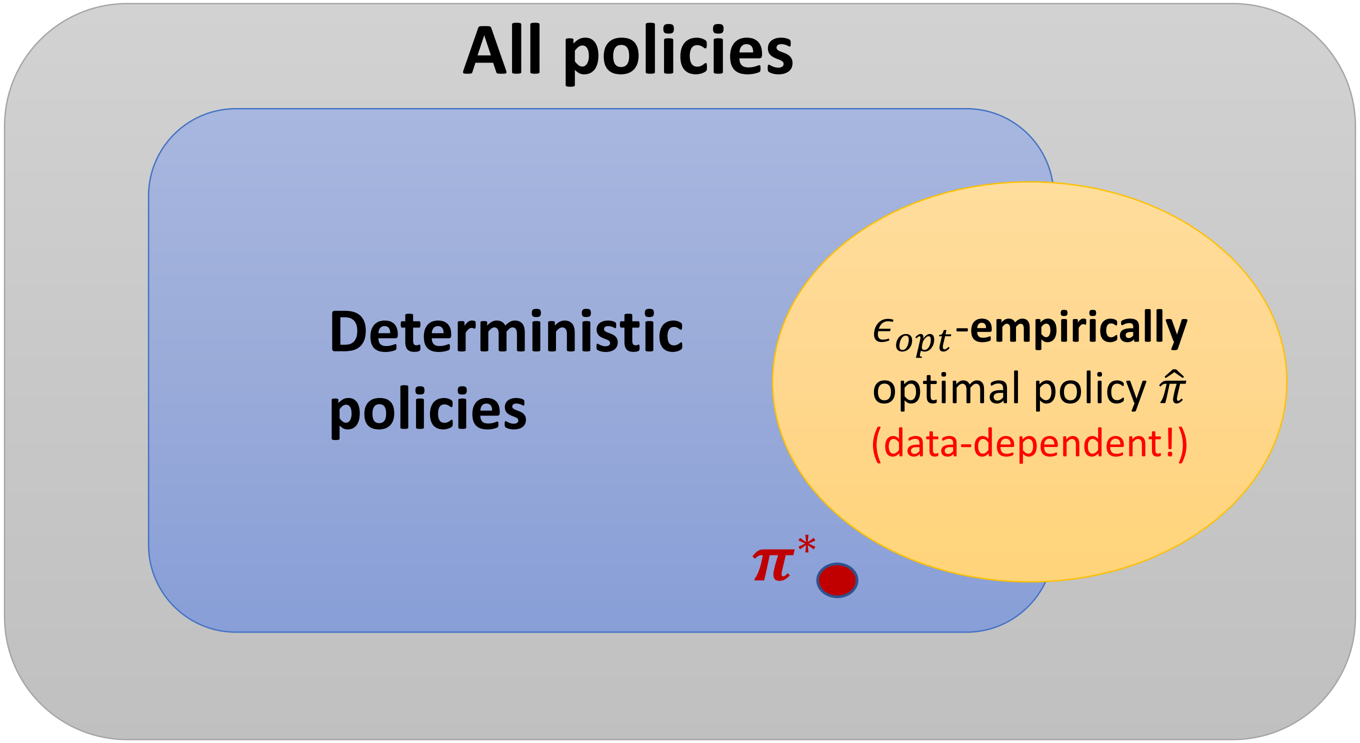
- Need to make assumptions about it

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

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

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

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

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)

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

$$\leq \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h, a_h} \frac{d^{\pi}_h(s_h)^2}{d^{\mu}_h(s_h)} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \mathrm{Var}\left[(V^{\pi}_{h+1}(s^{(1)}_{h+1}) + r^{(1)}_h) \middle| s^{(1)}_h = s_h, a^{(1)}_h = a_h \right]$$

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

$$\widetilde{r}_t(s_t,a_t) = \widehat{r}_t(s_t,a_t)\mathbf{1}(E_t) + r_t(s_t,a_t)\mathbf{1}(E_t^c)$$
$$\widetilde{P}_{t+1}(\cdot|s_t,a_t) = \widehat{P}_{t+1}(\cdot|s_t,a_t)\mathbf{1}(E_t) + P_{t+1}(\cdot|s_t,a_t)\mathbf{1}(E_t^c).$$

Idea: hypothetically plug in the ground truth occasionally

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

$$\widetilde{v}^\pi := \sum_{t=1}^H \langle \widetilde{d}_t^\pi, \widetilde{r}_t^\pi \rangle, \text{ with } \widetilde{d}_t^\pi = \widetilde{P}_t^\pi \widetilde{d}_{t-1}^\pi$$

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

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

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

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**.

Dealing with the reward noise

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

$$\sup_{\pi \in \Pi} \left| \sum_{t=1}^{H} \langle \widetilde{d}_t^{\pi}, \widetilde{r}_t - r_t \rangle \right| \leq O\left(\sqrt{\frac{H^2 \log(HSA/\delta)}{n \cdot d_m}}\right)$$

Martingale decomposition of the error $\tilde{v}^\pi - v^\pi$

Primal representation (Marginal distribution style):

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

|| (Lemma C.5)

Dual representation (Value function style):

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

Let's derive the Martingale Decomposition

$$\widetilde{d}_t^\pi = \pi_t \widetilde{T}_t \widetilde{d}_{t-1}^\pi \qquad\qquad T_t \in \mathbb{R}^{S\times(S\cdot A)}$$

$$d_t^\pi = \pi_t T_t d_{t-1}^\pi. \qquad\qquad (T_t)_{s_t,(s_{t-1},a_{t-1})} = P_t(s_t|s_{t-1},a_{t-1})$$

Take the difference of the two

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

Recursively apply the above

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

Dual representation (Value function style):

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

$$\begin{aligned}
X &= \sum_{t=1}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle + \langle r_t, \Gamma_{1:t}(\widetilde{d}_1^\pi - d_1^\pi)\rangle\right)\\
&= \sum_{t=1}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle\right) + \sum_{h=1}^{H}\langle r_t, \Gamma_{1:t}(\widetilde{d}_1^\pi - d_1^\pi)\rangle\\
&= \sum_{t=2}^{H}\left(\sum_{h=2}^{t}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle\right) + \sum_{h=1}^{H}\langle r_t, \Gamma_{1:t}(\widetilde{d}_1^\pi - d_1^\pi)\rangle\\
&= \sum_{h=2}^{H}\left(\sum_{t=h}^{H}\langle r_t, \Gamma_{h+1:t}\pi_h(\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle\right) + \sum_{h=1}^{H}\langle(\pi_1^T\Gamma_{1:t}^T r_t)(s), (\widetilde{d}_1^\pi - d_1^\pi)(s)\rangle\\
&= \sum_{h=2}^{H}\left(\langle\underbrace{\sum_{t=h}^{H}\pi_h^T\Gamma_{h+1:t}^T r_t}_{V_h^\pi(s)}, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi\rangle\right) + \langle\underbrace{(\sum_{h=1}^{H}\pi_1^T\Gamma_{1:t}^T r_t)(s)}_{V_1^\pi(s)}, (\widetilde{d}_1^\pi - d_1^\pi)(s)\rangle
\end{aligned}$$

Let's check that this is a Martingale

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

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

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

$$\begin{aligned}
&\mathbb{E}\left[\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi \rangle \middle| \mathcal{D}_t \right] \\
=&\mathbb{E}\left[\mathbb{E}\left[\langle V_h^\pi, (\widetilde{T}_h - T_h)\widetilde{d}_{h-1}^\pi \rangle \middle| \mathcal{D}_{h-1} \right] \middle| \mathcal{D}_t \right] \\
=&\mathbb{E}\left[\langle V_h^\pi, \mathbb{E}\left[(\widetilde{T}_h - T_h) \middle| \mathcal{D}_{h-1} \right] \widetilde{d}_{h-1}^\pi \rangle \middle| \mathcal{D}_t \right] = 0
\end{aligned}$$

Final results after applying a complex martingale concentration

- And a union bound.

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

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

where $O(\cdot)$ absorbs only the absolute constants.

Uniform convergence theorem for all deterministic policies

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

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

- Optimal in H, suboptimal in S.

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

Uniform convergence theorem for all policies

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

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

- 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

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

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

- 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.