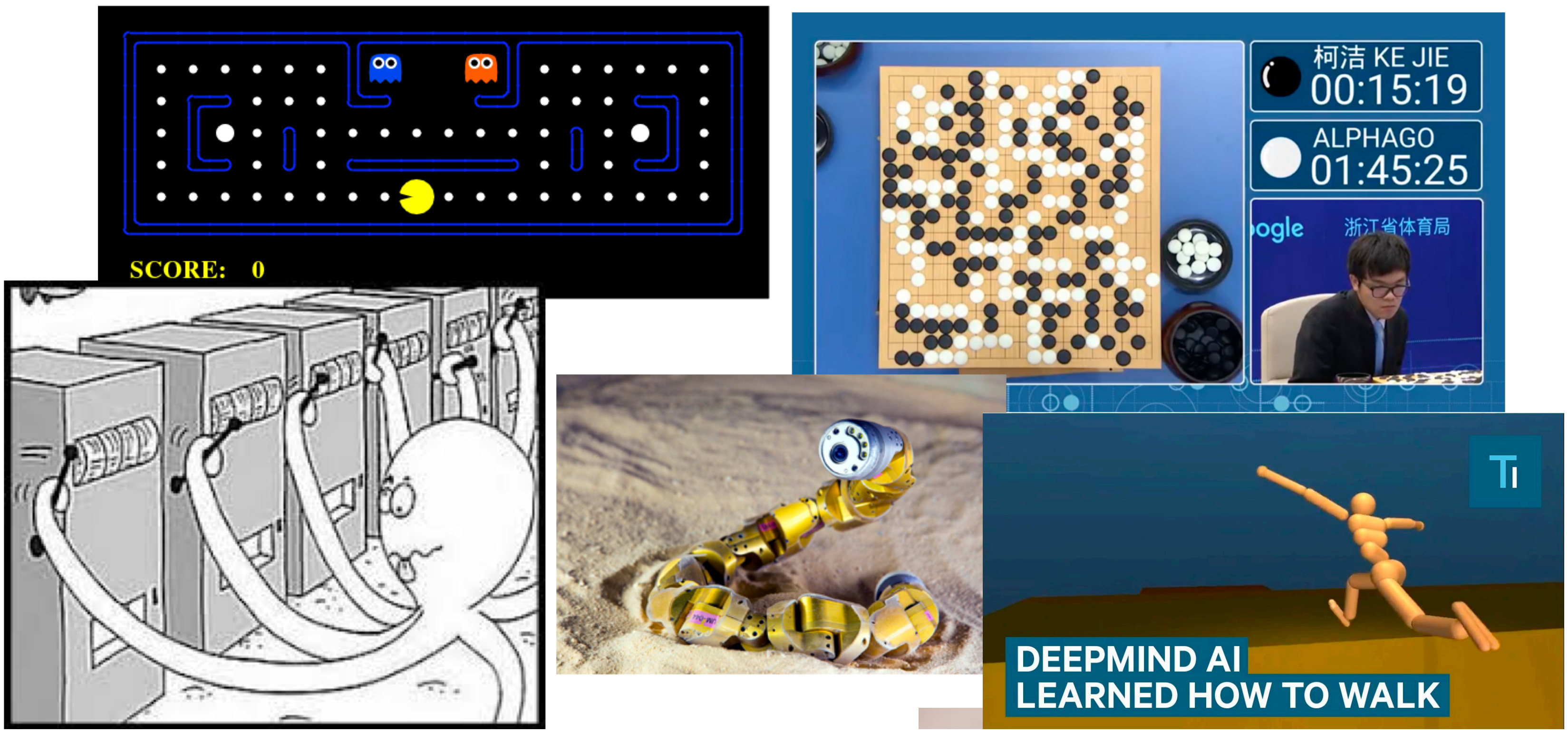
Uniform Offline Policy Evaluation (OPE) and Offline Learning in Tabular RL

Yu-Xiang Wang

Joint work with my student
Ming Yin and my collaborator Yu Bai

COMPUTER SCIENCE
UC SANTA BARBARA
Computing. ReInvented.

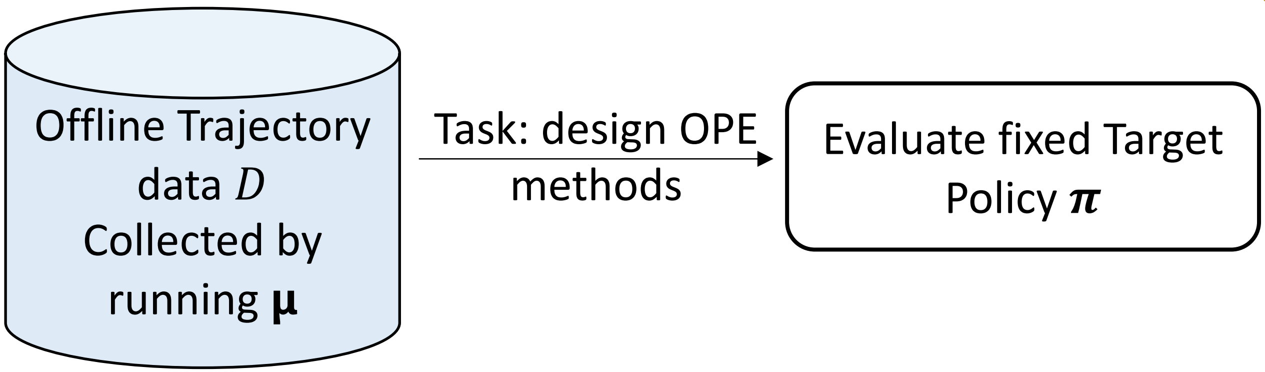
Reinforcement learning is among the hottest area of research in ML!



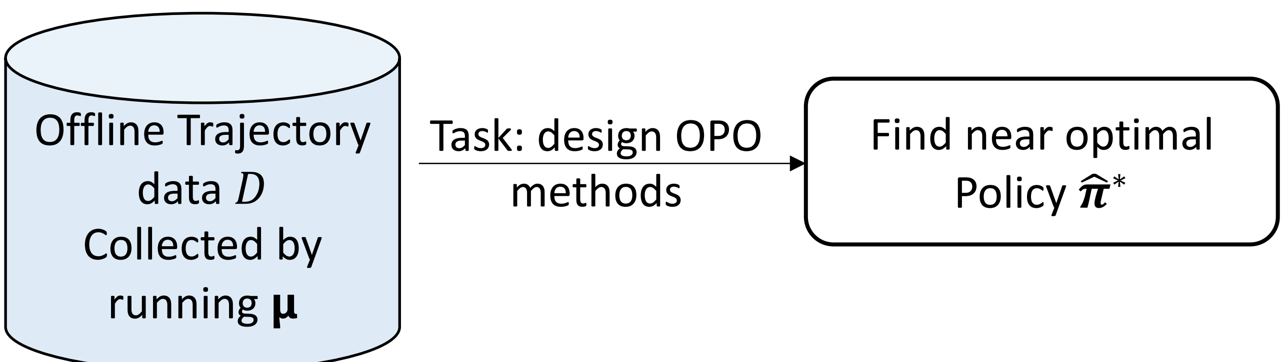
200+ papers on RL at NeurIPS'2019!

Topic today: Offline Reinforcement Learning, aka. Batch RL

- Task 1: Offline Policy Evaluation. (OPE)



- Task 2: Offline Policy Learning. (OPL)



Example applications of Offline RL

- Medical treatment / recommender systems
  - Cannot afford to run new experiments
  - Need safe policy improvements
- New material discovery / Learning self-driving car
  - Easy to parallelize the experiments
  - But hard to have many iterations
- Connections for online RL
  - Decomposing into offline epochs.
  - Each epoch is an offline learning problem

Outline of the talk

1. Notations and problem setup

2. Our contribution in OPE and OPL

3. Uniform convergence theorems

4. Key technical components + open problems

Formal problem setup: Episodic, Tabular, Non-Stationary MDPs

- Number of states, actions, horizon: S,A,H
- Number of offline trajectories: n

Translation: N = nH
- Number of "**steps**" in online RL
- Or number of "**generator calls**"

- Time-varying transition kernels:

$$P_t : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \mapsto [0, 1]$$

- Time-varying expected reward: $r_t : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$
- Policy $\pi := (\pi_1, \pi_2, \ldots, \pi_H)$ Logging policy: μ

- Value functions: $V_t^\pi(s) = \mathbb{E}_\pi[\sum_{t'=t}^{H} r_{t'} | s_t = s]$

$$Q_t^\pi(s, a) = \mathbb{E}_\pi[\sum_{t'=t}^{H} r_{t'} | s_t = s, a_t = a] \qquad v^\pi = \mathbb{E}_\pi\left[\sum_{t=1}^{H} r_t\right].$$

A few more notations

- Trajectory data:

$$(s_1, a_1, r_1, s_2, ..., s_H, a_H, r_H, s_{H+1})$$

$$\text{where } s_1 \sim d_1,\ a_t \sim \pi_t(\cdot|s_t),\ s_{t+1} \sim P_t(\cdot|s_t, a_t)$$

$$\mathcal{D} = \left\{ (s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)}) \right\}_{i\in[n]}^{t\in[H]}$$

- Marginal state-action distribution:

$$d_t^\pi(s_t, a_t) = d_t^\pi(s_t) \cdot \pi(a_t|s_t).$$

- State-action transition matrix:

$$(P_t^\pi)_{(s,a),(s',a')} := P_t(s'|s,a)\pi_t(a'|s')$$

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.

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

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

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

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:

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

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

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

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

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

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

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

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?

TL;DR: Our main contributions are: Optimal OPE and near optimal OPL

1. Characterizing the OPE for any fixed policy:

$$\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2] \leq \frac{1}{n}\sum_{h=0}^{H}\sum_{s_h,a_h}\frac{d_h^{\pi}(s_h)^2}{d_h^{\mu}(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\cdot\mathrm{Var}\left[(V^{\pi}_{h+1}(s^{(1)}_{h+1}) + r^{(1)}_h)\middle|s^{(1)}_h = s_h, a^{(1)}_h = a_h\right] + O(n^{-1.5})$$

Or if in a simplified expression: $\boldsymbol{\epsilon \asymp \sqrt{\frac{H^2}{n\, d_m^{\mu}}} \asymp \sqrt{\frac{H^2 SA}{n}}}$

(Xie, Ma & W., NeurIPS'19)

(Yin & W., AISTATS-20)

2. Advances in Uniform OPE that allows for near optimal offline learning

The ERM solution: $\hat{\pi} = \arg\max_{\pi\in\Pi} \hat{v}^{\pi}_{\mathrm{TMIS}}$

Obeys that $\boldsymbol{v^{\pi^*} - v^{\hat{\pi}} \lesssim \sqrt{\frac{H^3}{n\, d_m^{\mu}}} \asymp \sqrt{\frac{H^3 SA}{n}}}$

(Yin, Bai & W., on arxiv)

Comparing with prior results

Per-instance optimal. (TMIS, Fitted Q-Iteration)

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)
$\sqrt{\frac{H^4 S^2}{n d_m}}$	$\sqrt{\frac{e^H poly(S,A)}{n}}$	$\sqrt{\frac{H^3}{n\, d_m}}$	$\sqrt{\frac{H^2}{n\, d_m}}$	$\sqrt{\frac{H^2}{n\, d_m}}$

Offline Policy Learning Assume generative model

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 Ours
$\sqrt{\frac{H^4 S^2}{n d_m}}$	$\sqrt{\frac{H^4}{n d_m}}$	$\sqrt{\frac{H^3 SA}{n}}$	$\sqrt{\frac{H^3 SA}{n}} + H \cdot \epsilon_{opt}$	$\sqrt{\frac{H^3}{n\, d_m}} + \epsilon_{opt}$

Converted from infinite horizon case…

Our result is the first that achieves optimal rates in the offline setting

- And also the first that achieves the optimal rates via a (local) uniform convergence argument
  - So it is not specific to one algorithm

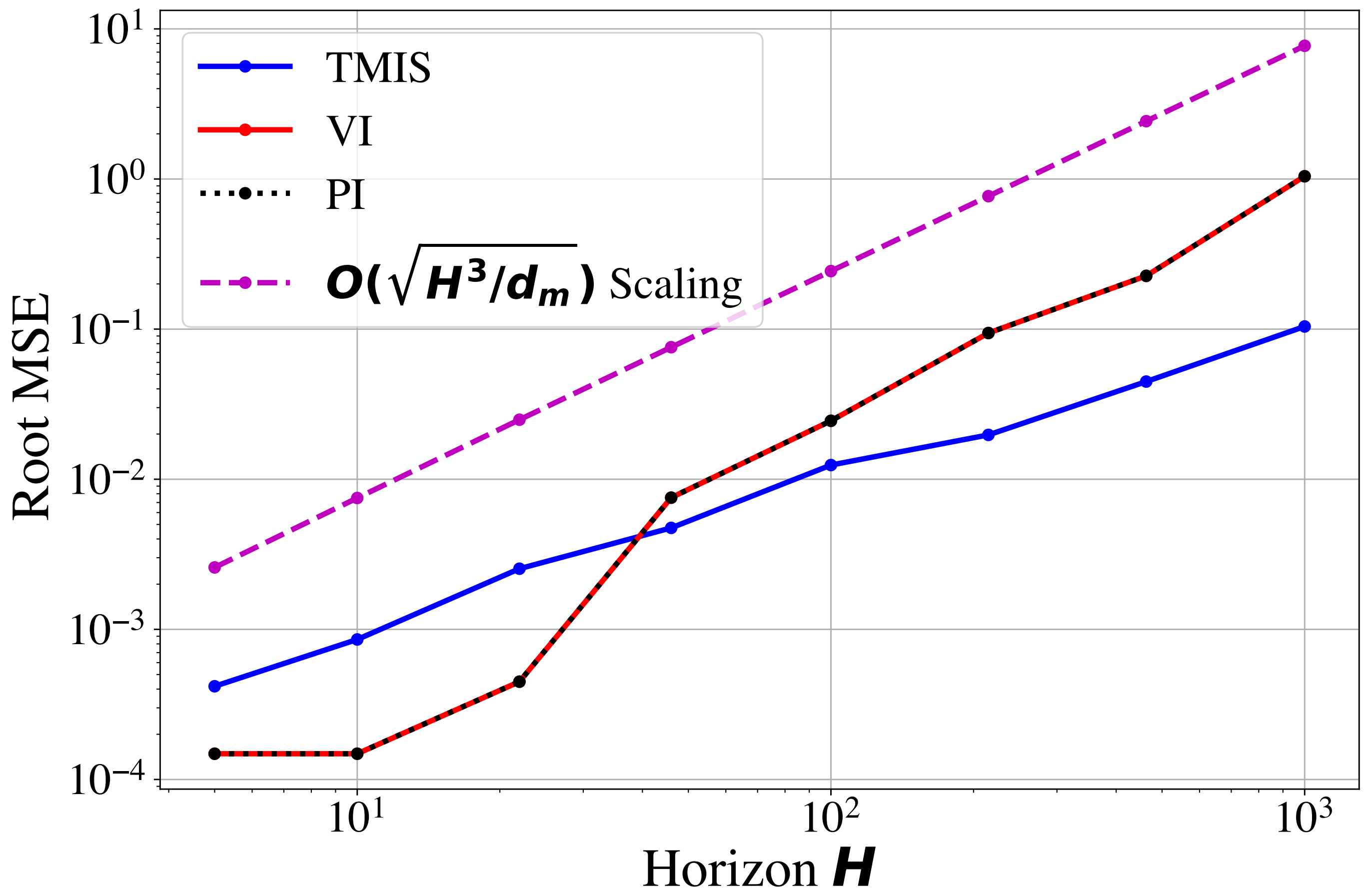
- On the side: we also include a lower bound

Theorem 3.8: Any estimator, exists (MDP, μ), s.t., with constant probability

$$\sup_{\pi \in \Pi} |\hat{v}^{\pi} - v^{\pi}| \gtrsim \sqrt{H^3/d_m n}$$

  - Idea: If faster rate => ERM breaks learning lower bounds.

Some simulation results: H^3 is the right scaling



Why is uniform convergence in RL a nontrivial problem?

- Even pointwise convergence is nontrivial

- Union bound is not tight
  - Discrete policy class: $\log|\Pi| = HS \log A$
  - But we expect $\tilde{O}(H)$

- Most standard approaches lead to suboptimal dependence in S and H

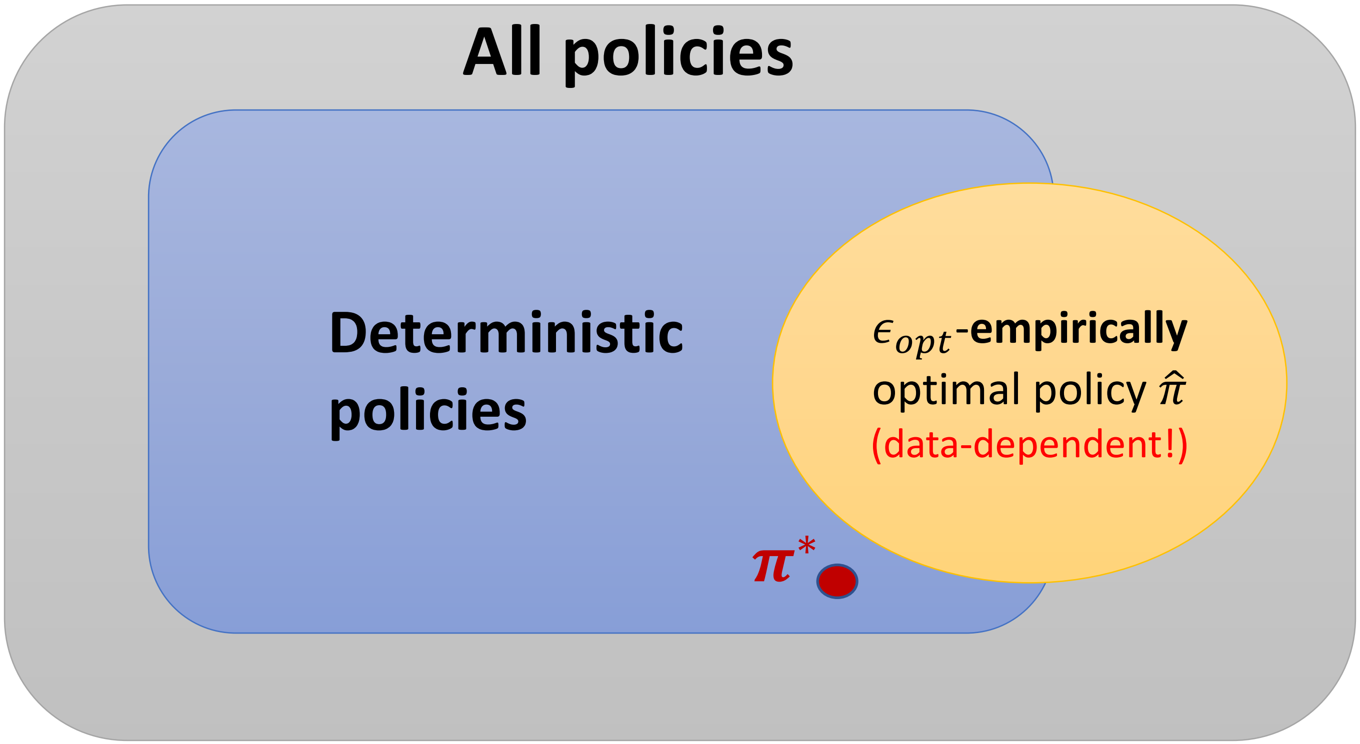
Obtaining optimal dependence in H is usually quite tricky…

$$\mathbb{E}[(\widehat{v}^{\pi}_{\mathrm{TMIS}} - v^{\pi})^2] \leq \frac{1}{n}\sum_{h=0}^{H}\sum_{s_h,a_h}\frac{d^{\pi}_h(s_h)^2}{d^{\mu}_h(s_h)}\frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)}\cdot\mathrm{Var}\left[(V^{\pi}_{h+1}(s^{(1)}_{h+1}) + r^{(1)}_h)\middle|s^{(1)}_h = s_h, a^{(1)}_h = a_h\right] + O(n^{-1.5})$$

- You are adding H terms that are potentially $O(H^2)$
- How do you see that the total is $O(H^2)$?

- See Lemma 3.4 in (Yin and W., 2020) for a cute proof.

The policy classes we consider



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

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 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 near-empirically optimal policies

Theorem 3.7: Let $\Pi_1 := \{\pi : s.t. \; || \hat{V}_t^{\pi} - \hat{V}_t^{\hat{\boldsymbol{\pi}}^*} ||_\infty \leq \epsilon_{opt}, \forall \, t \in [H]\}$. 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.

Key techniques used in the proof

- Fictitious estimator technique
- Martingale Decomposition of the error
- Anchor around the empirically optimal policy
  - Statistical independence of the past and the future when conditioning on the number of observations

To reiterate the main points

- For fixed π
  - Model-based OPE is exact optimal up to low order terms

- For uniform convergence:
  - Model-based OPE achieves optimal uniform convergence in a large ball around ERM.
  - **Corollary:** ERM with on Model-based OPE is rate-optimal
  - Near optimal global uniform convergence in some restricted regimes.

- Getting tight dependence in H, S is nontrivial
  - Key proof techniques presented in our work

Future work / open problems

1. Is the rate for **global** uniform convergence $\sqrt{\frac{H^3}{nd_m}}$?

2. The **natural complexity measure** for RL policy classes that gives rise to the "dimension" being $O(H)$ rather than $O(HS)$?

3. Function approximation settings?

Thank you for your attention!

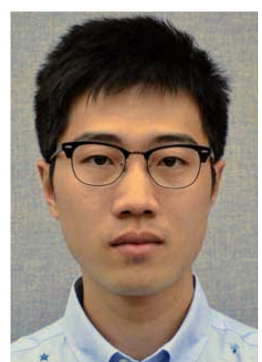
(Work supported by NSF # 2007117)

Reference and co-authors:

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

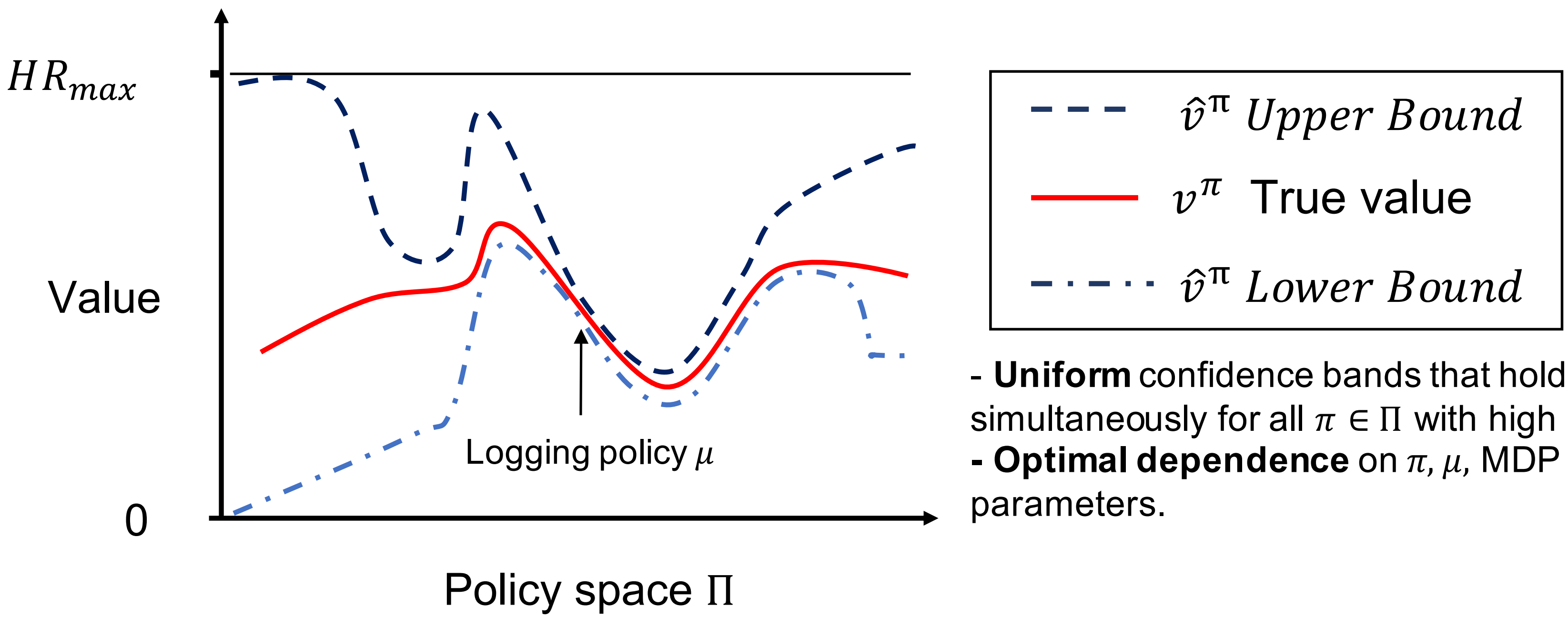
Yin and W. (2020) **Asymptotically Efficient Off-Policy Evaluation for Tabular Reinforcement Learning**. In AISTATS 2020.

Yin, Bai and W. (2020) **Near Optimal Provable Uniform Convergence in Offline Policy Evaluation for Reinforcement Learning**. In arXiv:2007.03760



Supplementary slides

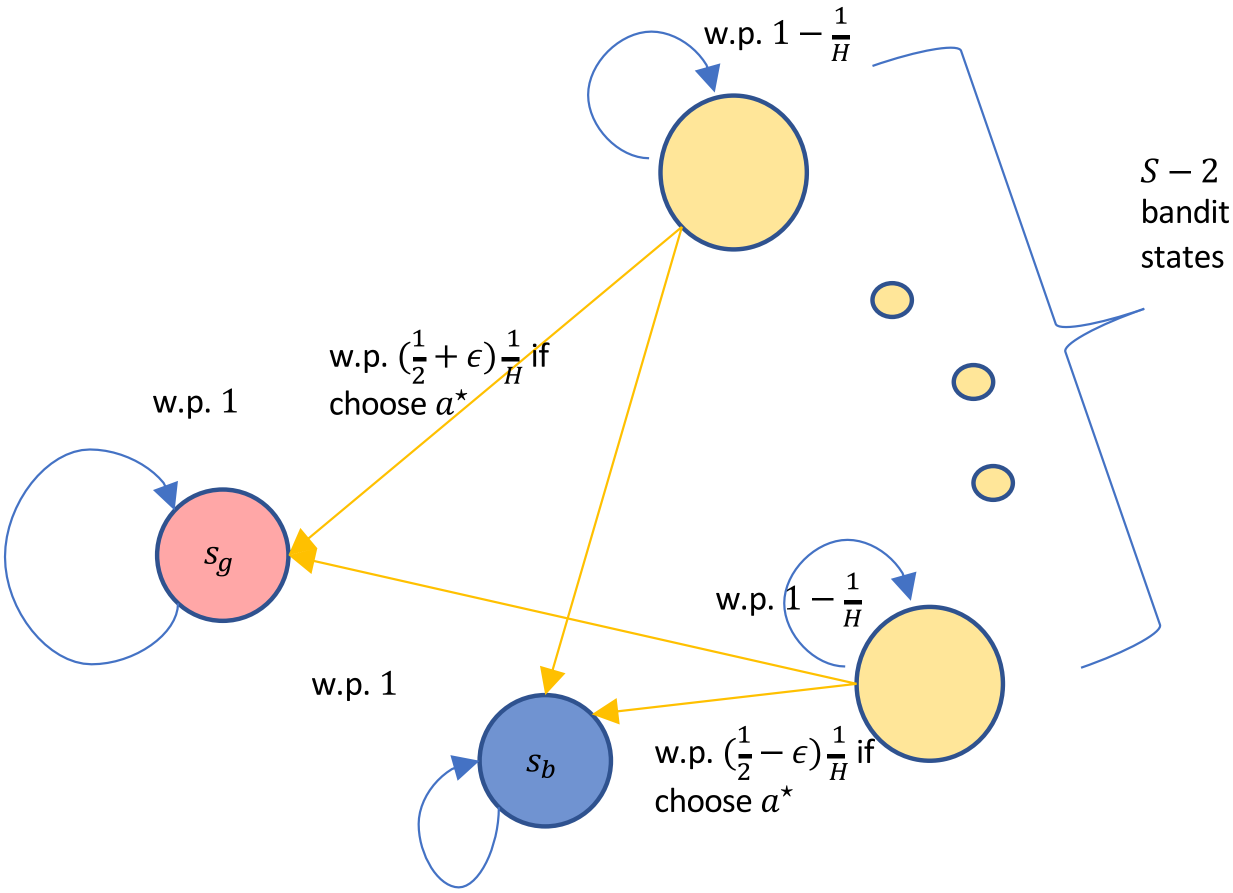
An illustration of what practical uniform-convergence looks like



- **Uniform** confidence bands that hold simultaneously for all $\pi \in \Pi$ with high probability.
- **Optimal dependence** on π, μ, MDP M parameters.

*You may choose your target policy π arbitrarily using the same dataset !

Lower bound construction



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.1

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

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

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

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.3)

Dual representation (Value function style):

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

Two implications of the Martingale Decomposition

1. Optimal *pointwise* convergence with high probability for fixed π
   - *(Chung & Lu, 2006)* Special Freedman's inequality + Fine grained variance calculations from *(Yin & W, AISTATS'20)*

2. Allow us to handle uniform convergence using Rademacher complexity-style arguments

Rademacher Complexity based approaches to uniform convergence

- Step 1: Concentration via McDiarmid

$$\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle \widetilde{d_t^\pi} - d_t^\pi, r_t\rangle\right| \leq O(\sqrt{\frac{H^4\log(HSA/\delta)}{nd_m}}) + \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\sum_{t=1}^{H}\langle \widetilde{d_t^\pi} - d_t^\pi, r_t\rangle\right|\right]$$

(Somewhat technical construction of a perturbation.)

- Step 2: Bound the expectation

(by the martingale decomposition)

$$\leq \sum_{h=2}^{H}\mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle v_h^\pi, (\widehat{T}_h - T_h)\widehat{d}_{h-1}^\pi\rangle\right|\cdot \mathbb{1}(E)\right] + \mathbb{E}\left[\sup_{\pi\in\Pi}\left|\langle v_1^\pi, \widehat{d}_1^\pi - d_1^\pi\rangle\right|\cdot \mathbb{1}(E)\right]$$

$$\leq O\left(\sqrt{H^4 S\log(HSA)/(nd_m)}\right)$$ By Rademacher complexity for each time step.

Main challenge: regrouping the things into < f(Policy), g(Data) >

Ideas behind local uniform convergence result

- Borrow ideas from the generative model literature
  - Specifically Agarwal, Kakade, Yang (2020)

- Recall: Bellman equations

$$Q_t^\pi = r_t + P_{t+1}^\pi Q_{t+1}^\pi = r_t + P_{t+1} v_{t+1}^\pi,$$

Also, the same Bellman equation for empirical MDP…

Ideas behind local uniform convergence result

- Taking differences of the empirical / true MDP's Bellman equations

$$
\begin{aligned}
\widehat{Q}_t^{\pi} - Q_t^{\pi} &= \widehat{P}_{t+1}^{\pi}\widehat{Q}_{t+1}^{\pi} - P_{t+1}^{\pi}Q_{t+1}^{\pi} \\
&= (\widehat{P}_{t+1}^{\pi} - P_{t+1}^{\pi})\widehat{Q}_{t+1}^{\pi} + P_{t+1}^{\pi}(\widehat{Q}_{t+1}^{\pi} - Q_{t+1}^{\pi})
\end{aligned}
$$

Back up recursively from the last step …

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

Multi-step transition matrix

Now take the empirically optimal policy as an anchor point…

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

Key observation:
$\widehat{P}_h \perp \hat{v}_h^{\widehat{\pi}^*} \mid n_{s,a,h}$
Save a factor of S

$$\leq O\left(\sqrt{\frac{H^3}{n\, d_m}} + \sqrt{\frac{1}{n\, d_m}} \sum_{h=t+1}^{H} |\hat{Q}_h^{\widehat{\pi}} - Q_h^{\widehat{\pi}}|\right) \cdot \mathbf{1}$$

Back-up recursively from t = H to 1
Tight variance calculation saves a factor of H

Apply the assumption of near-empirical optimality

$$\leq \epsilon_{opt} \cdot \tilde{O}\left(\sqrt{\frac{H^2 S^2}{n\, d_m}}\right) \cdot \mathbf{1}$$

Choose $\epsilon_{opt} < \sqrt{H}/S$

Comparing to Agarwal, Kakade, Yang (2020), we made some improvements

- Optimal local uniform convergence, when:

Lemma 10 (AKY-20)	Our result:
$\epsilon_{opt} < \sqrt{\frac{H^5}{n\, d_m}}$	$\epsilon_{opt} < \sqrt{H}/S$

- Comparison in terms of offline learning

Theorem 1 (AKY-20)	Our result:
$\sqrt{\frac{H^3}{n\, d_m}} + H\, \epsilon_{opt}$	$\sqrt{\frac{H^3}{n\, d_m}} + \epsilon_{opt}$