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
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)
  - 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
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: $n$
- Time-varying transition kernels: $P_t : S \times A \times S \mapsto [0, 1]$
- Time-varying expected reward: $r_t : S \times A \mapsto \mathbb{R}$
- Policy $\pi := (\pi_1, \pi_2, \ldots, \pi_H)$

Logging policy: $\mu$

- Value functions: $V_t^\pi(s) = \mathbb{E}_\pi \left[ \sum_{t'=t}^H r_{t'} \mid s_t = s \right]$
- $Q_t^\pi(s, a) = \mathbb{E}_\pi \left[ \sum_{t'=t}^H r_{t'} \mid s_t = s, a_t = a \right]$
- $v^\pi = \mathbb{E}_\pi \left[ \sum_{t=1}^H r_t \right]$
A few more notations

- **Trajectory data:**
  \[
  (s_1, a_1, r_1, s_2, \ldots, s_H, a_H, r_H, s_{H+1})
  \]
  where \( s_1 \sim d_1, a_t \sim \pi_t(\cdot|s_t), s_{t+1} \sim P_t(\cdot|s_t, a_t) \)
  
  \[
  D = \left\{ (s_t^{(i)}, a_t^{(i)}, r_t^{(i)}, s_{t+1}^{(i)}) \right\}_{i \in [n]}
  \]

- **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 $\mu$ is out of our control

- Need to make assumptions about it

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

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

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

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

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

\[
|\hat{v}^\pi - v^\pi| \leq \epsilon \quad \text{w.h.p.} \\
\mathbb{E}[|\hat{v}^\pi - v^\pi|^2] \leq \epsilon^2 \\
v^{\pi^*} - v^{\hat{\pi}} \leq 2\epsilon \quad \text{w.h.p} \\
v^{\pi^*} - \mathbb{E}[v^{\hat{\pi}}] \leq 2\epsilon
\]
Not quite this easy, the learned policy \( \hat{\pi} \) depends on the data

\[
\sup_{\pi \in \Pi} |\hat{\nu}^{\pi} - \nu^{\pi}| \leq \epsilon \quad \text{w.h.p.} \\
E \left[ \sup_{\pi \in \Pi} |\hat{\nu}^{\pi} - \nu^{\pi}|^2 \right] \leq \epsilon^2
\]

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

\[
\nu^{\pi^*} - E[\nu^{\hat{\pi}}] \leq 2\epsilon
\]

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

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

1. Characterizing the OPE for any fixed policy:

\[
E[(\hat{v}_\text{TMIS}^\pi - v^\pi)^2] \leq \frac{1}{n} \sum_{h=0}^{H} \sum_{s_h, a_h} \frac{d_h^\pi(s_h)^2}{d_h^\mu(s_h)} \frac{\pi(a_h|s_h)^2}{\mu(a_h|s_h)} \cdot \text{Var} \left[ \left( V_{h+1}(s_{h+1}) + r_{h}^{(1)} \right) | s_h^{(1)} = s_h, a_h^{(1)} = a_h \right] + O(n^{-1.5})
\]

Or if in a simplified expression: 
\[
\mathcal{E} \approx \sqrt{\frac{H^2}{n d_m^\mu}} \approx \sqrt{\frac{H^2SA}{n}}
\]

(Xie, Ma & W., NeurIPS’19)
(Yin & W., AISTATS-20)

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

The ERM solution: 
\[
\hat{\pi} = \arg \max_{\pi \in \Pi} \hat{v}_\text{TMIS}^\pi
\]

Obey that 
\[
\nu^{\pi^*} - \nu^{\hat{\pi}} \approx \sqrt{\frac{H^3}{n d_m^\mu}} = \sqrt{\frac{H^3SA}{n}}
\]

(Yin, Bai & W., on arxiv)
Comparing with prior results

Offline Policy Evaluation

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation lemma (1998)</td>
<td>$\sqrt{\frac{H^4 S^2}{n d_m}}$</td>
</tr>
<tr>
<td>IS / DR (2016)</td>
<td>$\sqrt{\frac{e^H \text{poly}(S, A)}{n}}$</td>
</tr>
<tr>
<td>MIS (2019)</td>
<td>$\sqrt{\frac{H^3}{n d_m}}$</td>
</tr>
<tr>
<td>TMIS (2020)</td>
<td>$\sqrt{\frac{H^2}{n d_m}}$</td>
</tr>
<tr>
<td>Fitted Q-Iteration (2020)</td>
<td>$\sqrt{\frac{H^2}{n d_m}}$</td>
</tr>
</tbody>
</table>

Offline Policy Learning

<table>
<thead>
<tr>
<th>Method</th>
<th>Formula</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simulation lemma (1998)</td>
<td>$\sqrt{\frac{H^4 S^2}{n d_m}}$</td>
</tr>
<tr>
<td>MSBO (2020)</td>
<td>$\sqrt{\frac{H^4}{n d_m}}$</td>
</tr>
<tr>
<td>Variance-Reduction (19)</td>
<td>$\sqrt{\frac{H^3 S A}{n}}$</td>
</tr>
<tr>
<td>Model-based (2020)</td>
<td>$\sqrt{\frac{H^3 S A}{n}} + H \cdot \epsilon_{opt}$</td>
</tr>
<tr>
<td>Model-based Ours</td>
<td>$\sqrt{\frac{H^3}{n d_m}} + \epsilon_{opt}$</td>
</tr>
</tbody>
</table>

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, \( \mu \)), 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 \( \Rightarrow \) 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...

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

- 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 \( \geq 1 - \delta \)

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

- 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 $\geq 1 - \delta$

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

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

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

**Theorem 3.7**: Let $$\Pi_1 := \{\pi : \text{s. t. } \|\hat{V}_t^\pi - \hat{V}_t^{\pi*}\|_\infty \leq \epsilon_{opt}, \forall t \in [H]\}$$. 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.
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 $\pi$
  • 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 \( \sqrt{\frac{H^3}{ndm}} \) ?

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:


Supplementary slides
We believe very strongly that the optimal bound should be unified in a way for every finite sample. A powerful feature of the proposed research is that the upper and lower bound perspective handles them in an equal support with the same size. Such a result resembles the Bernstein-type generalization error bound in the statistical learning theory [see, e.g., 10].

Besides obtaining the optimal rate, we also plan to search for uniform convergence results that adapt to the best policy at any time. A uniform convergence result ensures that we can do all of these on the same band with a total number of steps. Ultimately, dropping the assumption of constant factor for this problem. The important research question is how to characterize its entropy.

Figure 3: An illustration of the anticipated uniform confidence band for OPE. Note that we can search for the best certifiable policy by maximizing the lower bound, find the policy with the most potential by logging policy μ, and the MDP. In particular, we are hoping to prove an exact adaptive bound that says with high probability

\[ \sup_{\pi} \left( v^\pi - \hat{v}^\pi \right) = O\left( \frac{1}{n} \right) \]

is immediate through a high-probability OPE bound but the dependence on \( v^\pi \) and the MDP. In particular, we are hoping to prove an exact adaptive bound that says with high probability

\[ \sup_{\pi} \left( v^\pi - \hat{v}^\pi \right) = O\left( \frac{1}{n^{\alpha}} \right) \]

for some \( \alpha > 0 \). Note that the width of the confidence band will go to zero as \( n \to \infty \).

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 \( \pi, \mu, \text{MDP} M \) parameters.

*You may choose your target policy \( \pi \) arbitrarily using the same dataset!
Lower bound construction

The transition dynamics are defined as follows:

- For bandit states $b_{h,i}$, there is probability $1 / H$ to transition to $b_{h+1,i}$ regardless of the action chosen. For the rest of probability, optimal action $a^*$ will have probability $(1 / 2 + \epsilon) / H$ or $(1 / 2 - \epsilon) / H$ transition to $g_{h+1}$ or $b_{h+1}$ and all other actions $a$ will have equal probability $(1 / 2) / H$ for either $g_{h+1}$ or $b_{h+1}$, where $\epsilon$ is a parameter will be decided later. Or equivalently,

$$
P(s_{h+1,i} | s_{h,i}, a^*) = \begin{cases} 
1 / H & \text{if } s_{h} = s_{h+1,i} \text{ and } a^* \ni \frac{1}{2} + \epsilon, \\
1 / H & \text{if } s_{h} = g_{h+1}, \\
1 / H & \text{if } s_{h} = b_{h+1}, \\
(1 / 2) / H & \text{if } s_{h} = s_{h+1,i} \text{ and } a^* \ni \frac{1}{2} - \epsilon. 
\end{cases}
$$

- For $h = H, \ldots, 2H$, all states will always transition to the same type of states for the next step, i.e. $\forall a \in A$, $P(g_{h+1} | g_{h}, a) = 1$, $P(b_{h+1} | b_{h}, a) = 1$, $P(s_{h+1,i} | s_{h,i}, a) = 1$, $\forall i \in [S - 2]$.

$S - 2$ bandit states
Fictitious estimator technique

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

Idea: hypothetically plug in the ground truth occasionally

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

\[
\tilde{v}_\pi := \sum_{t=1}^H \langle \tilde{d}_t^\pi, \tilde{r}_t^\pi \rangle, \text{ with } \tilde{d}_t^\pi = \tilde{P}_t^\pi \tilde{d}_{t-1}^\pi
\]
The fictitious estimator is easier to analyze, because:

• Always unbiased.
• Has an *epistemical* Bellman-equation of variance
• Has nice martingale decompositions
• Moreover: Lemma C.1

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

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

\[
\sup_{\pi \in \Pi} |\tilde{v}^\pi - v^\pi| = \sup_{\pi \in \Pi} \left| \sum_{t=1}^{H} \langle \tilde{d}_t^\pi, \tilde{r}_t \rangle - \sum_{t=1}^{H} \langle d_t^\pi, r_t \rangle \right|
\]

\[
= \sup_{\pi \in \Pi} \left| \sum_{t=1}^{H} \langle \tilde{d}_t^\pi, \tilde{r}_t \rangle - \sum_{t=1}^{H} \langle \tilde{d}_t^\pi, r_t \rangle + \sum_{t=1}^{H} \langle \tilde{d}_t^\pi, r_t \rangle - \sum_{t=1}^{H} \langle d_t^\pi, r_t \rangle \right|
\]

\[
\leq \sup_{\pi \in \Pi} \left| \sum_{t=1}^{H} \langle \tilde{d}_t^\pi - d_t^\pi, r_t \rangle \right| + \sup_{\pi \in \Pi} \left| \sum_{t=1}^{H} \langle d_t^\pi, \tilde{r}_t - r_t \rangle \right| \tag{\ast}
\]

\[
\tag{\ast\ast}
\]

Lemma C.2: \( \ast\ast \) \( \leq \sqrt{H^2/(nd_m)} \)

Therefore, it suffices to consider the case with deterministic rewards.
Martingale decomposition of the error \( \tilde{\nu}^\pi - \nu^\pi \)

Primal representation (Marginal distribution style):

\[
\sum_{t=1}^{H} \langle \tilde{d}^\pi_t - d^\pi_t, r_t \rangle
\]

Dual representation (Value function style):

\[
\langle \nu^\pi_1(s), (\tilde{d}^\pi_1 - d^\pi_1)(s) \rangle + \sum_{h=2}^{H} \langle \nu^\pi_h(s), ((\tilde{T}_h - T_h)d^\pi_{h-1})(s) \rangle
\]
Two implications of the Martingale Decomposition

1. Optimal *pointwise* convergence with high probability for fixed $\pi$
   - *(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

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

(Somewhat technical construction of a perturbation.)

• Step 2: Bound the expectation

(by the martingale decomposition)

\[
\leq \sum_{h=2}^{H} \mathbb{E} \left[ \sup_{\pi \in \Pi} \left| \langle v^\pi_h, (\hat{T}_h - T_h)\tilde{d}^\pi_{h-1} \rangle \cdot 1(E) \right| \right] + \mathbb{E} \left[ \sup_{\pi \in \Pi} \left| \langle v^\pi_1, \tilde{d}^\pi_1 - d^\pi_1 \rangle \cdot 1(E) \right| \right]
\]

\[
\leq O \left( \sqrt{H^4 S \log(HSA)/(nd_m)} \right)
\]

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

\[ Q^\pi_t = r_t + P^\pi_{t+1} Q^\pi_{t+1} = r_t + P_{t+1} v^\pi_{t+1}, \]

Also, the same Bellman equation for empirical MDP...
Ideas behind local uniform convergence result

• Taking differences of the empirical / true MDP’s Bellman equations

\[ \hat{Q}^\pi_t - Q^\pi_t = \hat{P}^\pi_{t+1} \hat{Q}^\pi_{t+1} - P^\pi_{t+1} Q^\pi_{t+1} \]

\[ = (\hat{P}^\pi_{t+1} - P^\pi_{t+1}) \hat{Q}^\pi_{t+1} + P^\pi_{t+1} (\hat{Q}^\pi_{t+1} - Q^\pi_{t+1}) \]

Back up recursively from the last step ...

\[ \hat{Q}^\pi_t - Q^\pi_t = \sum_{h=t+1}^{H} \Gamma^\pi_{t+1:h-1} (\hat{P}_h - P_h) \hat{v}^\pi_h \]

Multi-step transition matrix
Now take the empirically optimal policy as an anchor point...

\[ |\hat{Q}_t^\pi - Q_t^\pi| \leq \sum_{h=t+1}^{H} \left[ \sum_{s,a,h} \Gamma_{t+1:h-1} |(\hat{P}_h - P_h)\hat{v}_h^\pi| + \Gamma_{t+1:h-1} |(\hat{P}_h - P_h)(\hat{v}_h^\pi - \hat{v}_h)| \right] \]

Key observation:
\[ \hat{P}_h \perp \hat{v}_h^\pi \mid n_{s,a,h} \]
Save a factor of \( S \)

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

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

Apply the assumption of near-empirical optimality

\[ \leq \epsilon_{opt} \cdot \tilde{\epsilon} \left( \sqrt{\frac{H^2 S^2}{n \cdot d_m}} \right) \cdot 1 \]

Choose \( \epsilon_{opt} < \sqrt{H} / S \)
Comparing to Agarwal, Kakade, Yang (2020), we made some improvements

• Optimal local uniform convergence, when:

<table>
<thead>
<tr>
<th>Lemma 10 (AKY-20)</th>
<th>Our result:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\epsilon_{opt} &lt; \sqrt{\frac{H^5}{n \cdot d_m}}$</td>
<td>$\epsilon_{opt} &lt; \sqrt{H/S}$</td>
</tr>
</tbody>
</table>

• Comparison in terms of offline learning

<table>
<thead>
<tr>
<th>Theorem 1 (AKY-20)</th>
<th>Our result:</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sqrt{\frac{H^3}{n \cdot d_m}} + H \epsilon_{opt}$</td>
<td>$\sqrt{\frac{H^3}{n \cdot d_m}} + \epsilon_{opt}$</td>
</tr>
</tbody>
</table>