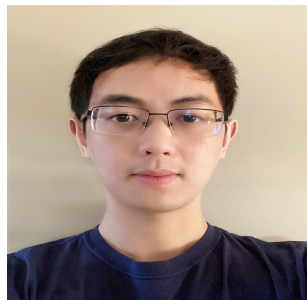# On Instance-Dependent Bounds for Offline Reinforcement Learning with Linear Function Approximation*

Thanh Nguyen-Tang[1], Ming Yin[2], Sunil Gupta[3], Svetha Venkatesh[3], Raman Arora[1]

[1]: Department of Computer Science, Johns Hopkins University

[2]: Department of Computer Science, Department of Statistics and Applied Probability, UC Santa Barbara

[3]: Applied AI Institute, Deakin University

*: https://arxiv.org/abs/2211.13208

Why Offline RL?

Reinforcement Learning with Online Interactions

Online Agent

Environment

- Can be extremely costly to run
- Can lead to unsafe/unethical behaviors

Offline Reinforcement Learning

Offline Agent

Environment

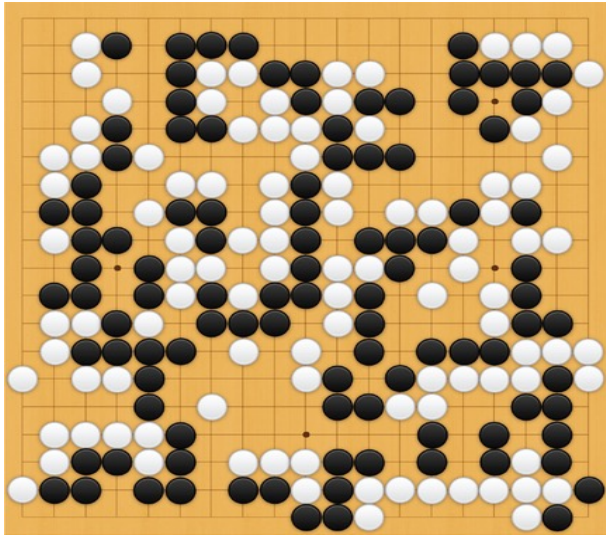- Can be more feasible in many domains
- Can enable better generalization by utilizing large datasets & diverse prior experiences

Image credit: Agarwal and Norouzi (2020)

Why (offline) RL with function approximation?

- We will never get enough data to learn each state individually

Game of Go: $> 10^{172}$

$\tilde{\mathcal{O}}(S)$ samples: impractical,
where S = state space size

We need a new mechanism:
- generalize from collected states to unvisited states

Episodic MDP

Episodic time-inhomogeneous Markov decision process
$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathrm{H}, \mathrm{P}, \mathrm{r}, \mathrm{d}_1)$$

state $s_h \in S$

Agent

action $a_h \in A$

reward $r_h = r_h(s_h, a_h) \in [0,1]$
next state $s_{h+1} \sim P_h(\cdot \mid s_h, a_h) \in S$

Environment

- State space $\mathcal{S}$

- Action space $\mathcal{A}$

- Episode length H:
  - Agent interacts with MDP for H steps and then restart the episode

- Transition kernels $\mathrm{P} = (\mathrm{P}_1, \ldots, \mathrm{P}_H)$, where $\mathrm{P}_h : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$

- Reward functions $\mathrm{r} = (\mathrm{r}_1, \ldots, \mathrm{r}_H)$, where $\mathrm{r}_h : \mathcal{S} \times \mathcal{A} \to [0,1]$

- Initial state distribution $\mathrm{d}_1 \in \Delta(\mathcal{S})$
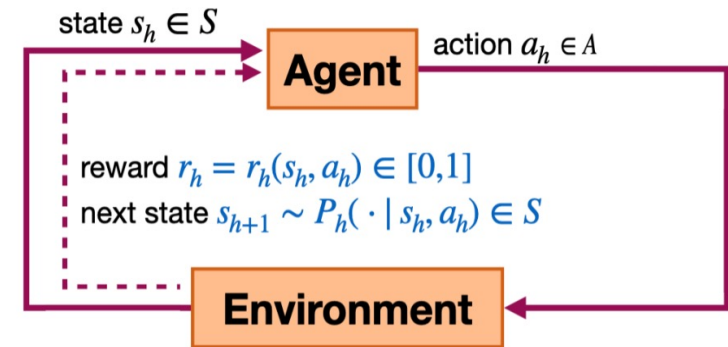
Episodic MDP
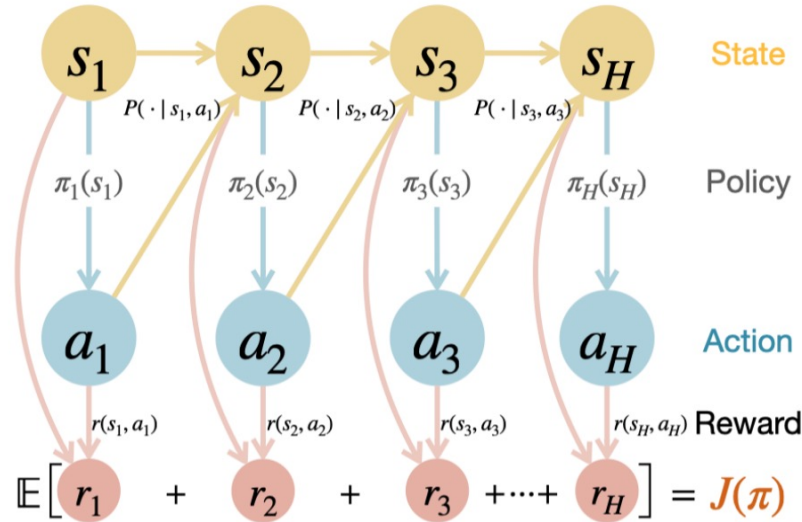
- A policy $\pi = \{\pi_h\}_{h \in [H]}$ where $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$
- Action-value functions:

$$Q_h^\pi(s, a) = \mathbb{E}_\pi\left[\sum_{i=1}^H r_i \mid (s_h, a_h) = (s, a)\right]$$

- Value functions

$$V_h^\pi(s) = \mathbb{E}_\pi\left[\sum_{i=1}^H r_i \mid s_h = s\right]$$

- The optimal policy π^* maximizes V_1^π

Dynamic Programming and Bellman Equation

- Optimal action-value functions $Q^* = \{Q_h^*\}_{h \in [H]}$

- Optimal value functions $V^* = \{V_h^*\}_{h \in [H]}$, $V_h^*(s) = \max_a Q_h^*(s, a)$

- Optimal policy π^* is greedy wrt Q^*

$$Q_h^*(s, a) = \underbrace{r_h(s, a) + \mathbb{E}_{s' \sim P_h(\cdot|s,a)}[V_{h+1}^*(s')]}_{\text{Bellman operator } \mathbb{B}_h: \quad \mathbb{B}_h V_{h+1}^*}$$

- Bellman equation

$$Q_h^* = \mathbb{B}_h Q_{h+1}^* , \, Q_{H+1}^* = 0$$

Offline RL with Value Function Approximation

- Offline dataset: collected a priori, $\mathcal{D} = \left\{\left(s_h^t, a_h^t, r_h^t\right)\right\}_{h \in [H]}^{t \in [K]}$
  - $a_h^t \sim \mu_h^t(\cdot | s_h^t), s_{h+1}^t \sim P_h(\cdot | s_h^t, a_h^t)$
  - $\mu = (\mu^1, \dots, \mu^K)$ is the behavior policy
  - Data were adaptively collected
  - K: # number of episodes
- No further interactions with MDP
- Learning objective: value suboptimality
$$\text{SubOpt}(\hat{\pi}; s_1) = V_1^*(s_1) - V_1^{\hat{\pi}}(s_1)$$
where $\hat{\pi} = \text{OfflineRLAlgo}(\mathcal{D}, \mathcal{F})$, $\mathcal{F}$ is some function class (e.g., neural networks)

This talk

How to design provably (instance-)efficient offline RL algorithms

in the function approximation setting

with the mildest data collection assumption possible?

- **Instance-efficient:** the algorithm should be able to leverage instance-specific information to accelerate the learning
- Efficient:
  - **Sample-efficient**: required # of samples is independent of the state space size (and polynomial in other problem factors)
  - **Runtime-efficient**: the algorithm runs in polynomial time
- **Mild data collection assumption**: the offline data does not need to cover the entire state-action space and it can be collected adaptively by running some adaptive algorithm.

Offline RL with linear function approximation

- Existing algorithms obtain finite value suboptimality in K episodes and nearly match the lower bound

  - Lower bound: $\Omega(\mathrm{H}^{1.5}\mathrm{S}^{0.5}\kappa_*^{0.5}\mathrm{K}^{-0.5})$ where $\kappa_* = \sup\limits_{\mathrm{h},\mathrm{s_h},\mathrm{a_h}} \dfrac{\mathrm{d}^{\pi_\mathrm{h}^*}(\mathrm{s_h},\mathrm{a_h})}{\mathrm{d}_\mathrm{h}^\mu(\mathrm{s_h},\mathrm{a_h})}$ is single-policy concentrability coefficient [Rashidinejad et al., (2021)]

  - APVI [Yin et al., (2022)] achieves $\tilde{\mathcal{O}}(\mathrm{H}^{1.5}\mathrm{S}^{0.5}\kappa_*^{0.5}\mathrm{K}^{-0.5})$

- **Limitations**: inefficient when S is large (e.g., when $S = 10^{172}$ in the game of Go)

- **Solutions**: Linear MDP

A MDP $\mathcal{M}$ is a linear MDP if there exist some known feature mapping $\phi_h : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, unknown vectors $\{\theta_h\}_{h \in [H]}$ and unknown measures $\{\nu_h\}_{h \in [H]}$ such that

$$r_h(s,a) = \phi_h(s,a)^T \theta_h \text{ and } \mathbb{P}_h(s'|s,a) = \phi_h(s,a)^T \nu_h(s')$$

PEVI Algorithm (Jin et al., 2021): LSVI + LCB

Algorithm 2 Pessimistic Value Iteration (PEVI): Linear MDP

1: Input: Dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, r_h^\tau)\}_{\tau,h=1}^{K,H}$.
2: Initialization: Set $\widehat{V}_{H+1}(\cdot) \leftarrow 0$.
3: **for** step $h = H, H-1, \ldots, 1$ **do**
4:      Set $\Lambda_h \leftarrow \sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top + \lambda \cdot I$.
5:      Set $\widehat{w}_h \leftarrow \Lambda_h^{-1}(\sum_{\tau=1}^{K} \phi(x_h^\tau, a_h^\tau) \cdot (r_h^\tau + \widehat{V}_{h+1}(x_{h+1}^\tau)))$.   //Estimation
6:      Set $\Gamma_h(\cdot,\cdot) \leftarrow \beta \cdot (\phi(\cdot,\cdot)^\top \Lambda_h^{-1} \phi(\cdot,\cdot))^{1/2}$.   //Uncertainty
7:      Set $\overline{Q}_h(\cdot,\cdot) \leftarrow \phi(\cdot,\cdot)^\top \widehat{w}_h - \Gamma_h(\cdot,\cdot)$.   //Pessimism
8:      Set $\widehat{Q}_h(\cdot,\cdot) \leftarrow \min\{\overline{Q}_h(\cdot,\cdot), H - h + 1\}^+$.   //Truncation
9:      Set $\widehat{\pi}_h(\cdot|\cdot) \leftarrow \arg\max_{\pi_h} \langle \widehat{Q}_h(\cdot,\cdot), \pi_h(\cdot|\cdot) \rangle_{\mathcal{A}}$.   //Optimization
10:      Set $\widehat{V}_h(\cdot) \leftarrow \langle \widehat{Q}_h(\cdot,\cdot), \widehat{\pi}_h(\cdot|\cdot) \rangle_{\mathcal{A}}$.   //Evaluation
11: **end for**
12: Output: $\text{Pess}(\mathcal{D}) = \{\widehat{\pi}_h\}_{h=1}^{H}$.

- $\tilde{\mathcal{O}}(H^2 d^{1.5} K^{-0.5})$ (under uniform coverage) and $\tilde{\mathcal{O}}(H^2 d^{1.5} \kappa_*^{0.5} K^{-0.5})$ (under single-policy concentrability) (our work)

- Lower bound: $\Omega(H \kappa_*^{0.5} K^{-0.5})$ (our work)

- Independent of S

Minimax to instance-dependent bounds

- Minimax bounds:
  - <u>Advantages</u>: hold for all instances, even in the worst case
  - <u>Limitations</u>: assuming a worst-case setting is too pessimistic
    - In many natural settings, offline RL can be faster than $\frac{1}{\sqrt{K}}$
- We argue that to circumvent the minimax lower bounds and explain the rates we observe in practical settings, we should consider the intrinsic instance-dependent structure of the underlying MDP

Hu et al. (2021)

- Let $\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a)$.
- Let $\Delta_h(s) \in \inf\limits_{a}\{\Delta_h(s,a) : \Delta_h(s,a) > 0\}$ (if $\inf\limits_{a}\{\Delta_h(s,a) : \Delta_h(s,a) > 0\} = \emptyset, \Delta_h(s) = 0$)
- Probabilistic gap assumption: $\sup\limits_{\pi} P_{s \sim d^\pi}(0 < \Delta_h(s) \leq \delta) \leq \left(\dfrac{\delta}{\delta_0}\right)^\alpha$
- Fitted Q-Iteration (FQI) in linear case: $\mathcal{O}(K^{-1})$ (i.e., $\alpha = 1$)
- **Advantages**: hold for various function classes and use a "weak" version of gap assumption
- **Limitations**:
  - Strong assumption in data coverage: uniform feature coverage
  $\lambda_{\min}\left(\mathbb{E}_{(s,a) \sim d_h^\mu}[\phi_h(s,a)\,\phi_h(s,a)^T]\right) > 0$
  - Data were not collected adaptively

Wang et al. (2022)

Gap assumption (Simchowitz & Jamieson, 2019; Yang et al., 2021; He et al., 2021):
Let $\Delta_h(s, a) := V_h^*(s) - Q_h^*(s, a)$. Assume that $\Delta_{\min} := \inf\limits_{h,s,a}\{\Delta_h(s, a) : \Delta_h(s, a) > 0\}$ is strictly positive

- Gap assumption
  - Subsampled VI-LCB: $\tilde{\mathcal{O}}(H^4 S\, \kappa_* \, \Delta_{\min} K^{-1})$
  - Lower bound: $\Omega(H^2 S\, \kappa_* \Delta_{\min} K^{-1})$
- Zero value suboptimality when $K = \tilde{\mathcal{O}}\left(H^3 \Delta_{\min}^{-2} P^{-1}\right)$ where $P := \min\limits_{h,s,a:\, d_h^*(s,a)>0} d_h^\mu(s, a)$

  - Lower bound: $K = \Omega\left(H\, \Delta_{\min}^{-2}\, P^{-1}\right)$

- **Limitations**:
  - Scales with S and the techniques only works for tabular MDPs
  - Episodes were collected independently

Adaptively collected data

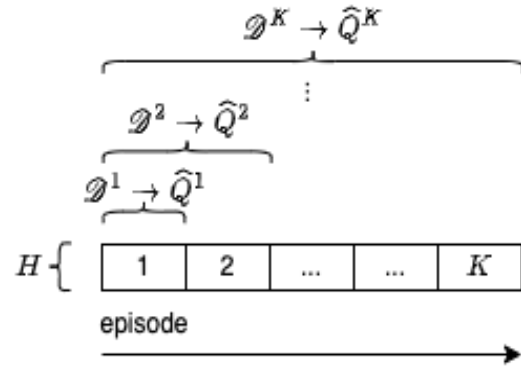- The dataset were collected by running an adaptive learning algorithm, e.g., in adaptive experiments, recommender's systems

- More formally, data at episode h, $(s_h^t, a_h^t, r_h^t)_{h \in [H]}$ is generated by μ_t which depends on all the data in the previous $t-1$ episodes

- Existing instance-dependent bounds either
  a) Scale with state space size S [Wang et al., 2022]
  b) Require strong uniform data coverage assumption [Wu et al., 2021]
  c) Require independently collected data [Wu et al., 2021; Wang et al., 2022]

We address these limitations with linear MDPs!

Our algorithm: LSVI + LCB + Bootstrapping + Constrained policy

- Built upon PEVI algorithm [Jin et al., 2021] with

two additional modifications:

- Bootstrapping

Algorithm 1 Bootstrapped and Constrained Pessimistic Value Iteration (BCP-VI)

1: **Input:** Dataset $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h \in [H]}^{t \in [K]}$, uncertainty parameters $\{\beta_k\}_{k \in [K]}$, regularization hyperparameter λ, μ-supported policy class $\{\Pi_h(\mu)\}_{h \in [H]}$.

2: **for** $k = 1, \ldots, K+1$ **do**

3:      $\hat{V}_{H+1}^k(\cdot) \leftarrow 0$.

4:      **for** step $h = H, H-1, \ldots, 1$ **do**

5:          $\Sigma_h^k \leftarrow \sum_{t=1}^{k-1} \phi_h(s_h^t, a_h^t) \cdot \phi_h(s_h^t, a_h^t)^T + \lambda \cdot I$.

6:          $\hat{w}_h^k \leftarrow (\Sigma_h^k)^{-1} \sum_{t=1}^{k-1} \phi_h(s_h^t, a_h^t) \cdot (r_h^t + \hat{V}_{h+1}^k(s_{h+1}^t))$.

7:          $b_h^k(\cdot, \cdot) \leftarrow \beta_k \cdot \|\phi_h(\cdot, \cdot)\|_{(\Sigma_h^k)^{-1}}$.

8:          $\bar{Q}_h^k(\cdot, \cdot) \leftarrow \langle \phi_h(\cdot, \cdot), \hat{w}_h^k \rangle - b_h^k(\cdot, \cdot)$.

9:          $\hat{Q}_h^k(\cdot, \cdot) \leftarrow \min\{\bar{Q}_h^k(\cdot, \cdot), H-h+1\}^+$.

10:         $\hat{\pi}_h^k \leftarrow \arg\max_{\pi_h \in \Pi_h(\mu)} \langle \hat{Q}_h^k, \pi_h \rangle$

11:         $\hat{V}_h^k(\cdot) \leftarrow \langle \hat{Q}_h^k(\cdot, \cdot), \hat{\pi}_h^k(\cdot|\cdot) \rangle$.

12:      **end for**

13: **end for**

14: **Output:** Ensemble $\{\hat{\pi}^k : k \in [K+1]\}$.

- Constrained policy extraction: $\hat{\pi}_h^k \leftarrow \arg\max_{\pi: \pi \text{ supported by } \mu} \langle \hat{Q}_h^k, \pi \rangle_{\mathcal{A}}$

- Given the policy ensemble

$\{\hat{\pi}^k : k \in [K+1]\}$, we consider two execution policies:

$$\Pi_h(\mu) := \{\pi_h : \text{supp}(\pi_h(\cdot|s_h)) \subseteq \text{supp}(\mu_h(\cdot|s_h)), \forall s_h \in \mathcal{S}_h\}.$$

- Last-iteration policy: $\hat{\pi}^{\text{last}} = \hat{\pi}^{K+1}$

- Mixture policy: $\hat{\pi}^{\text{mix}} = \frac{1}{K} \sum_{k=1}^{K} \hat{\pi}^k$

Our results: Gap-dependent bounds

- Let $\kappa_* = \max_{h \in [H]} \kappa_h$ where $\kappa_h^{-1} = \inf\{ d_h^\mu(s,a) \mid d_h^\mu(s,a) > 0 \}$

- <u>Partial data coverage</u>: $\forall (h,s,a), d_h^*(s,a) > 0 \Rightarrow d_h^\mu(s,a) > 0$

- Value suboptimality upper bound: $\tilde{\mathcal{O}}\left(d^3\, H^5\, \kappa_*^3\, \Delta_{\min}^{-1} K^{-1} \right)$
  - Independent of state space size S
  - The first result that scales with K^{-1} under linear MDP, gap assumption, partial data coverage, and adaptively collected data

- Lower bound: $\Omega(H^2 \kappa_{\min} \Delta_{\min}^{-1} K^{-1})$
  - Our upper bound is tight in K and $\Delta_{\min}$

- Techniques: count the number of times the empirical gaps exceed a certain value + peeling technique

Our results: Leverage "good" linear features for faster-than-K^{-1} rates

- Let λ_{min}^+ be the smallest positive eigenvalue of $\mathbb{E}_{(s,a)\sim d_h^*}[\phi_h(s,a)\phi_h(s,a)^T]$

- Let $k_* = \Omega(d^6 H^{10} \kappa_*^6 \Delta_{min}^{-1}(\lambda_{min}^+)^{-2} + \kappa_*^H(\lambda_{min}^+)^{-1})$

Assumption 4.4 (Unique Optimality and Spanning features). *We assume that*

1. (Unique Optimality - UO): The optimal actions are unique, i.e.

$$|supp(\hat{\pi}_h^*(\cdot|s_h))| = 1, \forall (h, s_h) \in [H] \times \mathcal{S}_h^*.$$

2. (Spanning Features - SF): Let $\phi_h^*(s) := \phi_h(s, \pi_h^*(s))$. For any $h \in [H]$,

$$span\{\phi_h^*(s_h) : \forall s_h \in \mathcal{S}_h^\mu\} \subseteq span\{\phi_h^*(s_h) : \forall s_h \in \mathcal{S}_h^*\}.$$

- We have: $\mathrm{SubOpt}(\hat{\pi}^k) = 0 \ \forall k \geq k_*$

Our other results

Algorithm	Condition	Upper Bound	Lower Bound	Data
PEVI	Uniform	$\tilde{\mathcal{O}}\left(\frac{H^2 d^{3/2}}{\sqrt{K}}\right)$	$\Omega\left(\frac{H}{\sqrt{K}}\right)$	Independent
BCP-VI	OPC	$\tilde{\mathcal{O}}\left(\frac{H^2 d^{3/2}\kappa_*}{\sqrt{K}}\right)$	$\Omega\left(\frac{H\sqrt{\kappa_{\min}}}{\sqrt{K}}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(\frac{d^3 H^5 \kappa_*^3}{\Delta_{\min}\cdot K}\right)$	$\Omega\left(\frac{H^2 \kappa_{\min}}{\Delta_{\min}\cdot K}\right)$	Adaptive
	OPC, Δ_{min}, UO-SF, $K \geq k^*$	0	0	Adaptive
BCP-VTR	OPC	$\tilde{\mathcal{O}}\left(\frac{H^2 d\kappa_*}{\sqrt{K}}\right)$	$\Omega\left(\frac{H\sqrt{\kappa_{\min}}}{\sqrt{K}}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(\frac{d^2 H^5 \kappa_*^3}{\Delta_{\min}\cdot K}\right)$	$\Omega\left(\frac{H^2 \kappa_{\min}}{\Delta_{\min}\cdot K}\right)$	Adaptive

Summary

We now have a provably (instance-)efficient algorithm for linear function approximation with polynomial sample and runtime

Algorithm: LSVI + LCB + Bootstrapping + Constrained policy extraction, under linear assumptions

Sample complexity:

- Gap-dependent: $\tilde{\mathcal{O}}\big(d^3\,H^5\,\kappa_*^3\,\Delta_{\min}^{-1}\epsilon^{-1}\big)$

- "Good" linear features: $\tilde{\mathcal{O}}\big(d^6 H^{10}\kappa_*^6\Delta_{\min}^{-1}(\lambda_{\min}^+)^{-2} + \kappa_*^H(\lambda_{\min}^+)^{-1}\big)$

Thank you

See our poster and arXiv version (https://arxiv.org/abs/2211.13208) for more details