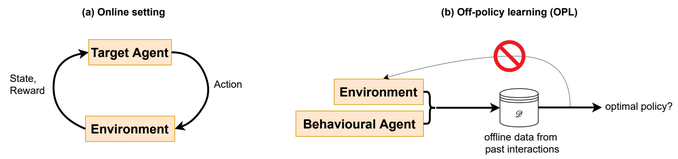
On Instance-Dependent Bounds for Offline Reinforcement Learning with Linear Function Approximation

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

[1]Johns Hopkins University [2]UC Santa Barbara [3]Deakin University (Australia)

Motivation

- **Offline Reinforcement Learning (RL)**: Learn an optimal policy from a fixed dataset collected a priori:



- **Function approximation**: We will never get enough data to learn each state individually → need to generalize from collected states to unseen states
- Existing results *[Rashidinejad et al., (2021), Yin et al., (2022), Jin et al., (2021)]* obtain finite value suboptimality $1/\sqrt{K}$, where K **is #of episodes in the offline data**, and nearly match lower minimax bounds
  - **Advantages**: hold for all instances, even in the worst case
  - **Limitations**: Assuming a worst-case setting is too conservative. In many natural problem instances, offline RL can be faster than $1/\sqrt{K}$.
- Can offline RL adaptively exploit **instance-dependent structure** of the underlying MDP to obtain faster rates than the minimax $1/\sqrt{K}$ rate?

Research Question:

How to design provably (instance-)efficient offline RL algorithms in the function approximation setting with the mildest data collection assumption possible?

Contributions

In this paper, we study instance-dependent bounds for offline RL with linear function approximation:

- We propose a new pessimistic algorithm that adapts to the "minimal gap" $\Delta_{\min}$, to achieve a fast rate of $\mathcal{O}(\frac{\log K}{K})$. Our result holds under the "single-policy concentrability" and *adaptively* collected data.
- Under an additional condition that the linear features for optimal actions in states reachable by the behavior policy span those in states reachable by an optimal policy, we show that our algorithm obtains an absolute zero value sub-optimality when K exceeds some problem-dependent constant.
- We provide information-theoretic lower bounds, which show that our gap-dependent bounds for offline RL are nearly optimal up to a polylog factor in terms of K and $\Delta_{\min}$.

Background

Time-inhomogenous Markov decision processes (MDPs):

- $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathbb{P} = \{\mathbb{P}_h\}_{h\in[H]}, r = \{r_h\}_{h\in[H]}, H, d_1)$: $\mathcal{S}, \mathcal{A}, H, \mathbb{P}, r, d_1$ are the state space, action space, episode horizon, transition kernels, reward functions, and initial state distribution, respectively.
- State-action visitation prob. $d_h^\pi(s,a) := \Pr((s_h, a_h) = (s,a) \mid \pi)$
- Value functions: $V_1^\pi(s_1) = \mathbb{E}_\pi\left[\sum_{h=1}^H r_h(s_h, a_h)\right]$ for policy π

Offline regime: Given historical data $\mathcal{D} = \{(s_h^t, a_h^t, r_h^t)\}_{h\in[H]}^{t\in[K]}$ generated by some **unknown behaviour policy** $\mu = \{\mu_h\}_{h\in[H]}$. The trajectory at any episode k can depend on the trajectories at all the previous episodes $t < k$.

Learning objective: Find $\hat{\pi}$ from $\mathcal{D}$ and a function class $\mathcal{F}$ to minimize $\mathrm{SubOpt}(\hat{\pi}) := \mathbb{E}_{s_1\sim d_1}\left[V_1^{\pi^*}(s_1) - V_1^{\hat{\pi}}(s_1)\right]$.

Linear MDPs: An MDP has a linear structure w.r.t. a known feature mapping $\phi_h : \mathcal{S}\times\mathcal{A} \to \mathbb{R}^d$ if for any (s, a, s', h),

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

for some unknown vectors $\theta_h \in \mathbb{R}^d$ and measures $\nu_h : \mathcal{S} \to \mathbb{R}^d$.

Algorithm: LSVI + LCB + Bootstrapping + Constrained Policy Extraction

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

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 $\{\Pi_h(\mu)\}_{h\in[H]}$.
for $k = 1, \ldots, K+1$ **do**
  $\hat{V}_{H+1}^k(\cdot) \leftarrow 0$.
  for step $h = H, H-1, ..., 1$ **do**
    $\Sigma_h^k \leftarrow \sum_{t=1}^{k-1} \phi_h(s_h^t, a_h^t)\cdot\phi_h(s_h^t, a_h^t)^T + \lambda\cdot I$.
    $\hat{w}_h^k \leftarrow (\Sigma_h^k)^{-1}\sum_{t=1}^{k-1}\phi_h(s_h^t, a_h^t)\cdot(r_h^t + \hat{V}_{h+1}^k(s_{h+1}^t))$.
    $b_h^k(\cdot,\cdot) \leftarrow \beta_k\cdot\|\phi_h(\cdot,\cdot)\|_{(\Sigma_h^k)^{-1}}$.
    $\bar{Q}_h^k(\cdot,\cdot) \leftarrow \langle\phi_h(\cdot,\cdot), \hat{w}_h^k\rangle - b_h^k(\cdot,\cdot)$.
    $\hat{Q}_h^k(\cdot,\cdot) \leftarrow \min\{\bar{Q}_h^k(\cdot,\cdot), H-h+1\}^+$.
    $\hat{\pi}_h^k \leftarrow \arg\max_{\pi_h\in\Pi_h(\mu)} \langle\hat{Q}_h^k, \pi_h\rangle$
    $\hat{V}_h^k(\cdot) \leftarrow \langle\hat{Q}_h^k(\cdot,\cdot), \pi_h^k(\cdot|\cdot)\rangle$.
  end for
end for
Output: Ensemble $\{\hat{\pi}^k : k \in [K+1]\}$.

- **Bootstrapping**: Split $\mathcal{D}$ into K subsets to mimic the data obtained by an online learner. Form an ensemble of value estimates from each subsets
- **Constrained policy extraction**: $\hat{\pi}_h^k \leftarrow \arg\max_{\pi:\pi \text{ supported by } \mu} \langle\hat{Q}_h^k, \pi\rangle$

Key Results

Gap-dependent bounds:

- Assumptions:
  - (Partial data coverage) $\forall(h,s,a), d_h^*(s,a) > 0 \implies d_h^\mu(s,a) > 0$
  - Let $\kappa_* = \max_{h\in[H]}\kappa_h$, where $\kappa_h^{-1} = \inf\{d_h^\mu(s,a) : d_h^\mu(s,a) > 0\}$
  - Let $\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a)$. Assume $\Delta_{\min} := \inf_{h,s,a}\{\Delta_h(s,a) : \Delta_h(s,a) > 0\}$ is strictly positive.
- Value suboptimality upper bound: $\tilde{\mathcal{O}}\left(d^3H^5\kappa^3\Delta_{\min}^1 K^{-1}\right)$
  - Independent of state space size $|\mathcal{S}|$
  - Fast rate K^{-1}, under partial data coverage and adaptively collected data
  - Lower bound: $\Omega(H^2\kappa_*\Delta_{\min}^1 K^{-1})$

Faster-than-K^{-1} rates:

- Assumptions
  - Let $\lambda_{\min}^+$ be the smallest positive eigenvalue of $\mathbb{E}_{(s,a)\sim d_h^*}\left[\phi_h(s,a)\phi_h(s,a)^T\right]$
  - Spanning features: $\mathrm{span}\{\phi_h^*(s_h) : \forall s_h \in \mathcal{S}_h^\mu\} \subseteq \mathrm{span}\{\phi_h^*(s_h) : \forall s_h \in \mathcal{S}_h^*\}$
- Results
  - Let $k_* = \Theta(d^6H^{10}\kappa^6\Delta_{\min}^1(\lambda_{\min}^+)^2 + \kappa^H(\lambda_{\min}^+)^1)$
  - We have: $\mathrm{SubOpt}(\pi^k) = 0, \forall k > k_*$.

We extend our results to linear mixture MDPs. Our results are summarized in the following table (Grey cells are our contributions):

Algorithm	Condition	Upper Bound	Lower Bound	Data
PEVI	Uniform	$\tilde{\mathcal{O}}\left(\frac{H^2d^{3/2}}{\sqrt{K}}\right)$	$\Omega\left(\frac{H}{\sqrt{K}}\right)$	Independent
BCP-VI	OPC	$\tilde{\mathcal{O}}\left(\frac{H^2d^{3/2}\kappa_*}{\sqrt{K}}\right)$	$\Omega\left(\frac{H\sqrt{\kappa_{\min}}}{\sqrt{K}}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(\frac{d^3H^5\kappa_*^3}{\Delta_{\min}\cdot K}\right)$	$\Omega\left(\frac{H^2\kappa_{\min}}{\Delta_{\min}\cdot K}\right)$	Adaptive
	OPC, Δ_{min}, UO-SF, $K \geq k^*$	0	0	Adaptive
BCP-VTR	OPC	$\tilde{\mathcal{O}}\left(\frac{H^2d\kappa_*}{\sqrt{K}}\right)$	$\Omega\left(\frac{H\sqrt{\kappa_{\min}}}{\sqrt{K}}\right)$	Adaptive
	OPC, Δ_{min}	$\tilde{\mathcal{O}}\left(\frac{d^2H^5\kappa_*^2}{\Delta_{\min}\cdot K}\right)$	$\Omega\left(\frac{H^2\kappa_{\min}}{\Delta_{\min}\cdot K}\right)$	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: $\tilde{\mathcal{O}}\left(d^3H^5\kappa^3\Delta_{\min}^1\epsilon^{-1}\right)$
  - "Good" linear features: $\tilde{\mathcal{O}}(d^6H^{10}\kappa^6\Delta_{\min}^1(\lambda_{\min}^+)^2 + \kappa^H(\lambda_{\min}^+)^1)$ (independent of ϵ)
- arXiv: `https://arxiv.org/abs/2211.13208`