Offline Reinforcement Learning with Linear Function Approximation

Wei Xiong

Mathematics, The Hong Kong University of Science and Technology

wxiongae@connect.ust.hk

Dec 9, 2022

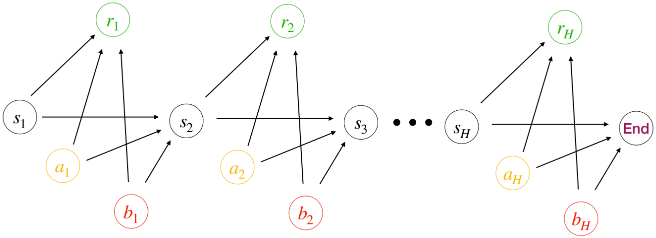
Outline

Table of Contents

Two-Player Zero-Sum Markov Game

Two-Player Zero-Sum Markov Game (MG): $\mathcal{M}(\mathcal{S}, \mathcal{A}_1, \mathcal{A}_2, H, \mathbb{P}, r)$

- $\mathcal{S}$: set of states; $\mathcal{A}_1, \mathcal{A}_2$: set of actions for the max-player[1] / min-player;
- H: time horizon, length of the game;
- $r_h(x_h, a_h, b_h) \in [0, 1]$: reward function of the max-player at step h;
- $\mathbb{P}_h(x_{h+1}|x_h, a_h, b_h)$: transition probability at step h.



[1] The player aims to maximize the cumulative rewards hence the name.

Policy, Value, and Nash Equilibrium

- Policy: mappings from state to a distribution of action: $\pi = \{\pi_h : \mathcal{S} \to \Delta_{\mathcal{A}_1}\}$ and $\nu = \{\nu_h : \mathcal{S} \to \Delta_{\mathcal{A}_2}\}$
- Value: Expected cumulative reward starting from step h:
  - V-value: $V_h^{\pi,\nu}(x_h) = \mathbb{E}_{\pi,\nu}[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}, b_{h'}) \mid x_h]$;
  - Q-value: $Q_h^{\pi,\nu}(x_h, a_h, b_h) = \mathbb{E}_{\pi,\nu}[\sum_{h'=h}^{H} r_{h'}(x_{h'}, a_{h'}, b_{h'}) \mid x_h, a_h, b_h]$.
- Best response (the strongest opponent):
  - $V_h^{\pi,*} = V_h^{\pi,\mathrm{br}(\pi)} = \inf_\nu V_h^{\pi,\nu}$;
  - $V_h^{*,\nu} = V_h^{\mathrm{br}(\nu),\nu} = \sup_\pi V_h^{\pi,\nu}$
- Nash Equilibrium (NE): (π^*, ν^*) is an NE if
  - they are best response to each other;
  - V_h^* is the Nash Value of (π^*, ν^*);
- Learning objective: finding a pair $(\widehat{\pi}, \widehat{\nu})$ such that for any $x \in \mathcal{S}$,

$$\mathrm{SubOpt}((\widehat{\pi}, \widehat{\nu}), x) := V_1^{*,\widehat{\nu}}(x) - V_1^{\widehat{\pi},*}(x) < \epsilon.$$

Offline Learning

Offline learning means that we learn the policy from a pre-determined dataset without further interaction with the environment.

- The dataset $\mathcal{D} = \{(x_h^\tau, a_h^\tau, b_h^\tau)\}_{\tau,h=1}^{K,H}$ is collected by some behavior policy independently;
- The MG possesses a linear structure with a known feature $\phi(x,a,b) \in \mathbb{R}^d$:

$$r_h(x,a,b) = \phi(x,a,b)^\top \theta_h, \qquad \mathbb{P}_h(\cdot|x,a,b) = \phi(x,a,b)^\top \mu_h(\cdot);$$

- Tabular MG with finite state and action spaces is a special case of Linear MG;
- Goal: learn an ϵ-approximate NE with a sample complexity polynomial in $\left(\frac{1}{\epsilon}, H, d\right)$;

Problem: what is the minimal dataset assumption that permits efficient learning?

Table of Contents

Existing Results of Offline MDP

Single-policy (optimal policy) coverage is the necessary and sufficient condition for sample-efficient learning.

- Tabular MDP [4, 3, 2]: with b denoting the behavior policy:

$$\sup_{x,a,h} \frac{d_h^{\pi^*}(x,a)}{d_h^b(x,a)} \leq C^*$$

- Linear MDP [1]

$$\mathbb{E}_{\pi^*}[\sum_{h=1}^{H} \phi_h^\top \Lambda_h^{-1} \phi_h], \qquad \text{where} \Lambda_h = \sum_{k=1}^{K} \phi_h^k (\phi_h^k)^\top + \lambda I_d.$$

Q: Single-policy (NE) coverage is necessary and sufficient for Markov Games?

Single-policy (NE) coverage is Insufficient

Consider the matrix (bandit) game $\mathcal{M}_1$ and $\mathcal{M}_2$ with payoff matrices:

$$G_1 = \begin{pmatrix} 0.5 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix} \quad G_2 = \begin{pmatrix} 0 & 0 & -1 \\ 1 & 0 & -1 \\ 1 & 1 & 0 \end{pmatrix}$$

- Given a dataset that is consistent with both $\mathcal{M}_1$ and $\mathcal{M}_2$ and let $\widehat{\pi} = (p_1, p_2, p_3)$ and $\widehat{\nu} = (q_1, q_2, q_3)$ be the learned policy:

$$\mathrm{SubOpt}_{\mathcal{M}_1}((\widehat{\pi}, \widehat{\nu}), x) + \mathrm{SubOpt}_{\mathcal{M}_2}((\widehat{\pi}, \widehat{\nu}), x) \geq 2$$

- Either $\mathrm{SubOpt}_{\mathcal{M}_1}((\widehat{\pi}, \widehat{\nu}), x)$ or $\mathrm{SubOpt}_{\mathcal{M}_2}((\widehat{\pi}, \widehat{\nu}), x)$ is larger than 1;

Conclusion: Single-policy (NE) coverage is not sufficient for Markov Games.

What is the Sufficient Coverage Condition?

Suppose that $\mathcal{M}_1$ with $G_1 = \begin{pmatrix} 0.5 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$ is the ground truth.

$\{(\pi,\nu) : (\pi,\nu) \text{ is arbitrary}\}$

$\downarrow$

$\begin{pmatrix} 0.5 & -1 & 0 \\ 1 & 0 & 1 \\ 0 & -1 & 0 \end{pmatrix}$

$\downarrow$

$\{(\pi^*,\nu) : \nu \text{ is arbitrary}\} \cup \{(\pi,\nu^*) : \pi \text{ is arbitrary}\}$

$\downarrow$

$\begin{pmatrix} * & -1 & * \\ 1 & 0 & 1 \\ * & -1 & * \end{pmatrix}$

$\downarrow$

Intuition: the second unilateral concentration condition ensures us to verify that π^* and ν^* are the best response to each other (definition of NE).

Table of Contents

Pessimistic Minimax Value Iteration (PMVI)

Suppose that we have constructed $\overline{V}_{h+1}$ and $\underline{V}_{h+1}$. We employ the fact that the Bellman equation is linear in the feature.

- Estimate the linear coefficient by least-squares regression;

$$\underline{w}_h \leftarrow \operatorname*{argmin}_w \sum_{\tau=1}^{K} [r_h^\tau + \underline{V}_{h+1}(x_{h+1}^\tau) - (\phi_h^\tau)^\top w]^2 + \|w\|_2^2,$$

$$\overline{w}_h \leftarrow \operatorname*{argmin}_w \sum_{\tau=1}^{K} [r_h^\tau + \overline{V}_{h+1}(x_{h+1}^\tau) - (\phi_h^\tau)^\top w]^2 + \|w\|_2^2,$$

- Pessimistic Q value with penalty term $\Gamma_h(x,a,b) = \beta\sqrt{\phi(x,a,b)^\top \Lambda_h^{-1}\phi(x,a,b)}$:

$$\underline{Q}_h(\cdot,\cdot,\cdot) \leftarrow \Pi_{H-h+1}\{\phi(\cdot,\cdot,\cdot)^\top \underline{w}_h - \Gamma_h(\cdot,\cdot,\cdot)\},$$
$$\overline{Q}_h(\cdot,\cdot,\cdot) \leftarrow \Pi_{H-h+1}\{\phi(\cdot,\cdot,\cdot)^\top \overline{w}_h + \Gamma_h(\cdot,\cdot,\cdot)\}.$$

- Compute the output policy pair (NE subroutines):

$$(\widehat{\pi}_h(\cdot \mid \cdot), \nu_h'(\cdot \mid \cdot)) \leftarrow \mathrm{NE}(\underline{Q}_h(\cdot,\cdot,\cdot)), \qquad (\pi_h'(\cdot \mid \cdot), \widehat{\nu}_h(\cdot \mid \cdot)) \leftarrow \mathrm{NE}(\overline{Q}_h(\cdot,\cdot,\cdot)).$$

Main Result

Theorem ([7])

Let $\beta = \mathcal{O}\big(dH\sqrt{\log(2dKH/\delta)}\big)$, *it holds with probability* $1-\delta$ *that*

$$\mathrm{SubOpt}\big((\hat{\pi},\hat{\nu}),x\big) \le 4\beta \cdot \mathrm{RU}(\mathcal{D},x).$$

which features a new notion, Relative Uncertainty:

$$\mathrm{RU}(\mathcal{D},x) = \max\left\{\sup_{\nu}\sum_{h=1}^{H}\mathbb{E}_{\pi^*,\nu}\left[\sqrt{\phi_h^\top\Lambda_h^{-1}\phi_h}\,\middle|\,x_1=x\right], \sup_{\pi}\sum_{h=1}^{H}\mathbb{E}_{\pi,\nu^*}\left[\sqrt{\phi_h^\top\Lambda_h^{-1}\phi_h}\,\middle|\,x_1=x\right]\right\}.$$

- Data-dependent bound: Λ_h^{-1} is fully determined by the offline dataset;
- Unilateral Concentration: $\{(\pi^*,\nu),(\pi,\nu^*) : \pi,\nu \text{ are arbitrary}\}$.

Conclusion: Low relative uncertainty is sufficient for sample-efficient learning.

Low Relative Uncertainty is Necessary

Minimax Lower Bound:

$$\mathbb{E}_{\mathcal{D}} \frac{\mathrm{SubOpt}(\mathrm{Alg}(\mathcal{D}); x)}{\mathrm{RU}(\mathcal{D}, x)} \geq C',$$

where C' is an absolute constant and x is the initial state. The expectation is taken with respect to the dataset generalization.

Conclusion: Low relative uncertainty is necessary for sample-efficient learning.

Table of Contents

Conclusion and Future Directions

- We propose the first line of work studying the dataset condition that permits efficient multi-agent offline RL;
- We figure out that low relative uncertainty is the necessary and sufficient condition for achieving sample efficiency in offline linear MGs setup;
- The suboptimality bound is $\mathcal{O}(\sqrt{d}H)$ away from the minimax lower bound;
- Once can leverage (1) reference-advantage decomposition and (2) weighted regression to achieve an optimal sample complexity at a cost of stronger assumptions [5].

Rewrite the Problem

Given $\widehat{V}_{h+1}$, the essential problem is to construct an estimator of

$$\tilde{Q}_h(x,a) := \mathcal{T}_h \widehat{V}_{h+1}(x,a) := r_h(x,a) + \mathbb{E}_{x_{h+1}|x,a} \widehat{V}_{h+1}(x_{h+1}) = w_h^\top \phi(x,a),$$

with $\mathcal{D} := \{x_h^\tau, a_h^\tau\}_{h,\tau=1}^K$ such that the following inequality holds with high probability:

$$|\widehat{w}^\top \phi(x,a) - w_h^\top \phi(x,a)| \le \Gamma_h(x,a).$$

- A sharper estimator of the linear coefficient leads to a better regret bound:

$$\mathrm{SubOpt}(\widehat{\pi}, x) \le \mathbb{E}_{\pi^*|x_1=x} \sum_{h=1}^H \Gamma_h(x,a), \qquad \widehat{\pi} \text{ greedy in } \widehat{Q}_h;$$

- Let $\widehat{Q}_h$ be the least-squares solution. Hoeffding+ uniform concentration gives

$$\left|\tilde{Q}_h(x,a) - \widehat{Q}_h(x,a)\right| \lesssim \underbrace{||\sum_{\tau \in \mathcal{D}} \phi\left(x_h^\tau, a_h^\tau\right) \cdot \xi_h^\tau(\widehat{V}_{h+1})||_{\Lambda_h^{-1}}}_{\texttt{(A)} \le \beta = \tilde{O}(dH)} ||\phi(x,a)||_{\Lambda_h^{-1}},$$

with $\Lambda_h = \lambda I + \sum_{\tau=1}^K \phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top$, $\xi_h^\tau(f) = f(x_{h+1}^\tau) + r_h^\tau - (\mathcal{T}_h f)(x_h^\tau, a_h^\tau)$.

What Causes Suboptimality?

$$\left|\tilde{Q}_h(x,a) - \widehat{Q}_h(x,a)\right| \lesssim \underbrace{\|\sum_{\tau\in\mathcal{D}} \phi(x_h^\tau, a_h^\tau)\cdot \xi_h^\tau(\widehat{V}_{h+1})\|_{\Lambda_h^{-1}}}_{\text{(A)} \leq \beta = \tilde{O}\left(H(\sqrt{d} + \sqrt{\log \mathcal{N}(\widehat{V}_{h+1})})\right)} \|\phi(x,a)\|_{\Lambda_h^{-1}},$$

- $\widehat{V}_{h+1} \in \mathcal{F}_h \oplus \{\Gamma_{h+1}\}$ is computed by later least-square value iteration thus depending on the data at step h;
- The issue is solved by a uniform concentration over ϵ-net, paying for a covering number: improve the d-dependency:

$$\sqrt{\log \mathcal{N}(\mathcal{F}_h)} = \sqrt{d} \qquad \text{v.s.} \qquad \sqrt{\log \mathcal{N}(\mathcal{F}_h \oplus \{\Gamma_{h+1}\})} = d;$$

- Leverage the variance information to improve the Horizon-dependency:
  - Hoeffding: range H;
  - Bernstein: conditional variance of $\xi_h^\tau(\widehat{V}_{h+1})$: $\sigma = H$;
  - Directly using Bernstein-type inequality offers no advantage.

Improve the d-dependency

The key observation is that both the Bellman operator and the estimator are linear in the target:

- $\mathcal{T}_h(f+g) = \mathcal{T}_h f + \mathcal{T}_h g$;
- $\widehat{w}_h(f+g) = \widehat{w}_h(f) + \widehat{w}_h(g)$.

Reference-Advantage Decomposition by V_{h+1}^*:

$$|\langle \widehat{w}_h(\widehat{V}_{h+1}), \phi(x,a)\rangle - \mathcal{T}_h \widehat{V}_{h+1}(x,a)| \leq$$
$$\underbrace{|\langle \widehat{w}_h(V_{h+1}^*), \phi(x,a)\rangle - \mathcal{T}_h V_{h+1}^*(x,a)|}_{\text{Reference}} + \underbrace{|\langle \widehat{w}_h(\widehat{V}_{h+1} - V_{h+1}^*), \phi(x,a)\rangle - \mathcal{T}_h \widehat{V}_{h+1}(x,a)|}_{\text{Advantage}}$$

- Reference with deterministic V_{h+1}^*: no need for uniform concentration thus improving $\sqrt{d}$;
- Advantage: $\|\widehat{V}_{h+1} - V_{h+1}^*\|_\infty = \tilde{O}\big(\frac{\sqrt{d}H^2}{\sqrt{K\kappa}}\big)$[2]: leading to a high-order concentration error of advantage part.

[2] This requires a stronger coverage condition.

Improve the H-dependency

Weighted Regression [6]: assigning sample-dependent weights in the regression subroutine.

$$\operatorname*{argmin}_{w\in\mathbb{R}^d} \sum_{\tau\in\mathcal{D}} \frac{\left[\phi(x_h^\tau, a_h^\tau)^\top w - r_h^\tau - f_{h+1}\left(x_{h+1}^\tau\right)\right]^2}{\widehat{\sigma}_h^2(x_h^\tau, a_h^\tau)} + \lambda\|w\|_2^2$$

Suppose that $\widehat{\sigma}_h^2(\cdot,\cdot) \approx \mathrm{Var}[r_h^\tau + f_{h+1}(x_{h+1}^\tau) - (\mathcal{T}_h f_{h+1})(x_h^\tau, a_h^\tau)|x_h^\tau, a_h^\tau]$ [3]

- The conditional variance of $\xi_h^\tau(f_{h+1}) = \frac{r_h^\tau + f_{h+1}(x_{h+1}^\tau) - (\mathcal{T}_h f_{h+1})(x_h^\tau, a_h^\tau)}{\widehat{\sigma}_h(x_h^\tau, a_h^\tau)}$ is $O(1)$;
- The Bernstein's inequality implies a $\tilde{O}(\sqrt{d}\cdot 1)||\phi(x,a)||_{\Sigma_h^{-1}}$ with

$$\Sigma_h^{-1} = \Big(\sum_{\tau\in\mathcal{D}} \frac{\phi(x_h^\tau, a_h^\tau)\phi(x_h^\tau, a_h^\tau)^\top}{\widehat{\sigma}_h^2(x_h^\tau, a_h^\tau)} + \lambda I\Big)^{-1} \preccurlyeq H^2\Lambda_h^{-1};$$

- The new bonus with weighted regression is never worse than the regular $\sqrt{d}H||\phi(x,a)||_{\Lambda_h^{-1}}$;
- The new bonus is in an instance-dependent manner and can provide faster rates for many instances.

[3] This holds for a stronger coverage condition.

Thank you for Listening!

Paper:

- Han Zhong*, Wei Xiong*, Jiyuan Tan*, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang, "Pessimistic Minimax Value Iteration: Provably Efficient Equilibrium Learning from Offline Datasets", ICML 2022;
- Wei Xiong*, Han Zhong*, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang, "Nearly Minimax Optimal Offline Reinforcement Learning with Linear Function Approximation: Single-Agent MDP and Markov Game", Preprint.

[1] Ying Jin, Zhuoran Yang, and Zhaoran Wang. Is pessimism provably efficient for offline rl? In *International Conference on Machine Learning*, pages 5084–5096. PMLR, 2021.

[2] Gen Li, Laixi Shi, Yuxin Chen, Yuejie Chi, and Yuting Wei. Settling the sample complexity of model-based offline reinforcement learning. *arXiv preprint arXiv:2204.05275*, 2022.

[3] Laixi Shi, Gen Li, Yuting Wei, Yuxin Chen, and Yuejie Chi. Pessimistic q-learning for offline reinforcement learning: Towards optimal sample complexity. *arXiv preprint arXiv:2202.13890*, 2022.

[4] Tengyang Xie, Nan Jiang, Huan Wang, Caiming Xiong, and Yu Bai. Policy finetuning: Bridging sample-efficient offline and online reinforcement learning. *Advances in neural information processing systems*, 34:27395–27407, 2021.

[5] Wei Xiong, Han Zhong, Chengshuai Shi, Cong Shen, Liwei Wang, and Tong Zhang. Nearly minimax optimal offline reinforcement learning with linear function approximation: Single-agent mdp and markov game. *arXiv preprint arXiv:2205.15512*, 2022.

[6] Ming Yin, Yaqi Duan, Mengdi Wang, and Yu-Xiang Wang. Near-optimal offline reinforcement learning with linear representation: Leveraging variance information with pessimism. *arXiv preprint arXiv:2203.05804*, 2022.

[7] Han Zhong, Wei Xiong, Jiyuan Tan, Liwei Wang, Tong Zhang, Zhaoran Wang, and Zhuoran Yang. Pessimistic minimax value iteration: Provably efficient equilibrium learning from offline datasets. *arXiv preprint arXiv:2202.07511*, 2022.