Sample Complexity of Offline RL with Deep ReLU Networks[1]

Thanh Nguyen-Tang
(Email: nguyent2792@gmail.com,
Website: thanhnguyentang.github.io)

Applied Artificial Intelligence Institute (A^2I^2), Deakin University, Australia

EEML,Virtual Budapest Hungary
7-15 July 2021,

[1]Full paper at https://arxiv.org/abs/2103.06671.

Motivation for offline RL with function approximation

- **Offline RL goal**: Learn an optimal policy from an offline data without any further exploration.
- Most theoretical results in offline RL focus on tabular environments with small finite state spaces [Yin and Wang, 2020, Yin et al., 2021, Yin and Wang, 2021], or linear MDP [Duan and Wang, 2020]
- In practice, most MDPs are complex with infinitely large state space $\rightarrow$ function approximation such as deep neural networks to generalize from observed states to unseen ones is necessary

Related work for offline RL with function approximation

- [Munos and Szepesvári, 2008]: A classical analysis of fitted Q-iteration (FQI)
  - They use a general function class (thus the bound depends on a so-called Bellman inherent error)
  - The bound is not tight
- [Le et al., 2019]: A modern analysis of FQI for offline RL
  - They use a general function class
  - The bound is tighter than that in [Munos and Szepesvári, 2008]
  - An improper analysis: it incorrectly ignores the data-dependent structure in FQI
- [Yang et al., 2019]: An analysis of FQI for Q-learning
  - They use deep ReLU network function approximation
  - Their algorithm does not reuse the offline data for different iterations; thus, the sample complexity scales with the number of iterations in the algorithm
  - Their result relies on a rather limited smoothness condition: Hölder smoothness. What about MDPs that are beyond Hölder smoothness?

Paper summary

In this work, we study sample complexity of offline RL with deep ReLU network function approximation:

1. We introduce a new general dynamic condition, namely **Besov dynamic closure** that allows *fractional* and *inhomogeneous* smoothness of the MDP, and encompasses/generalizes the prior conditions (Hölder and Sobolev smoothness)
2. Our sample complexity is established under a **data-dependent structure** that is ignored in prior algorithms [Yang et al., 2019] or improperly handled by prior analyses [Le et al., 2019]
3. We obtain a sample complexity of $\tilde{\mathcal{O}}\left(\kappa^{1+d/\alpha} \cdot \epsilon^{-2-2d/\alpha}\right)$ where κ is a distribution shift measure, d is the dimension of the state-action space, α is the (possibly fractional) smoothness parameter of the underlying MDP, and ϵ is a user-specified error

Deep ReLU networks as function approximation

Denote $\Phi(L, m, S, B)$ the space of L-height, m-width (fully connected) ReLU networks with "sparsity constraint" S and "norm constraint" B

- "sparsity constraint" S: The total number of parameters $\leq S$
- "norm constraint" B: The maximum value of the network parameters $\leq B$.

The unit ball of ReLU network function space $\mathcal{F}_{NN}$:

$$\mathcal{F}_{NN} := \left\{ f \in \Phi(L, m, S, B) : \|f\|_\infty \leq 1 \right\}.$$

Besov dynamic closure

- Let $B^{\alpha}_{p,q} := \{f \in L^p(\mathcal{X}) : \|f\|_{B^{\alpha}_{p,q}} < \infty\}$ be the Besov space with smoothness α and regularities p and q where $\|\cdot\|_{B^{\alpha}_{p,q}}$ is the Besov norm (more technical details in our paper).
- Hölder spaces and Sobolev spaces are special cases of Besov spaces

Assumption (*Besov dynamic closure*)

$\forall f \in \mathcal{F}_{NN}(\mathcal{X}), \forall \pi, T^{\pi} f \in \bar{B}^{\alpha}_{p,q}(\mathcal{X})$ for some $p, q \in [1, \infty]$ and $\alpha > \frac{d}{p \wedge 2}$ where $\bar{B}^{\alpha}_{p,q}(\mathcal{X})$ the ∞-norm unit ball of $B^{\alpha}_{p,q}(\mathcal{X})$ and T^{π} is Bellman operator.

Intuition:

- only requires the boundedness of a very general notion of local oscillations of the underlying MDP
- the underlying MDP could be discontinuous, non-differentiable or have inhomogeneous smoothness.

$\rightarrow$ the most general dynamic assumption in (offline) RL with function approximation.

Algorithm and data-dependent structure

- **Algorithm**: A simple variant of FQI with deep ReLU network function approximation

Algorithm 1 Least-squares value iteration (LSVI)

1: Initialize $Q_0 \in \mathcal{F}_{NN}$.
2: **for** $k = 1$ **to** K **do**
3: If **OPE** (for a fixed policy π): $y_i \leftarrow r_i + \gamma \int_{\mathcal{A}} Q_{k-1}(s_i', a)\pi(da|s_i'), \forall i$
4: If **OPL**: $y_i \leftarrow r_i + \gamma \max_{a' \in \mathcal{A}} Q_{k-1}(s_i', a'), \forall i$
5: $Q_k \leftarrow \arg\min_{f \in \mathcal{F}_{NN}} \frac{1}{n} \sum_{i=1}^{n} (f(s_i, a_i) - y_i)^2$
6: **end for**
7: If **OPE**, return

$$V_K = \|Q_K\|_{\rho^\pi} = \sqrt{\mathbb{E}_{\rho(s)\pi(a|s)} [Q_K(s,a)^2]}$$

8: If **OPL**, return the greedy policy π_K w.r.t. Q_K.

- **Data-dependent structure**: the regression targets y_i depend on Q_{k-1} which in turn depend on the offline data $\{(s_i, a_i)\}_{i=1}^{n}$

Main theorem

Under *the Besov dynamic closure* and the *finite concentration coefficient*, for any $\epsilon > 0, \delta \in (0,1], K > 0$, and for $n \gtrsim \left(\frac{1}{\epsilon^2}\right)^{1+\frac{d}{\alpha}} \log^6 n + \frac{1}{\epsilon^2}(\log(1/\delta) + \log\log n)$, with probability at least $1-\delta$, the sup-optimality of Algorithm LSVI is

$$\begin{cases} \mathrm{SubOpt}(V_K;\pi) \leq \frac{\sqrt{\kappa_\mu}}{1-\gamma}\epsilon + \frac{\gamma^{K/2}}{(1-\gamma)^{1/2}} & \text{for OPE,} \\ \mathrm{SubOpt}(\pi_K) \leq \frac{4\gamma\sqrt{\kappa_\mu}}{(1-\gamma)^2}\epsilon + \frac{4\gamma^{1+K/2}}{(1-\gamma)^{3/2}} & \text{for OPL.} \end{cases}$$

In addition, the optimal deep ReLU network $\Phi(L,m,S,B)$ that obtains such sample complexity (for both OPE and OPL) satisfies

$$L \asymp \log N, m \asymp N\log N, S \asymp N, \text{ and } B \asymp N^{1/d+(2\iota)/(\alpha-\iota)},$$

where $\iota := d(p^{-1} - (1+\lfloor\alpha\rfloor)^{-1})_+, N \asymp n^{\frac{(\beta+1/2)d}{2\alpha+d}}$, and $\beta = (2+\frac{d^2}{\alpha(\alpha+d)})^{-1}$.

Key result summary

Work	Functions	Regularity	Tasks	Sample complexity	Remark
Yin and Wang [2020]	Tabular	Tabular	OPE	$\tilde{\mathcal{O}}\left(\kappa \cdot \lvert\mathcal{S}\rvert^2 \cdot \lvert\mathcal{A}\rvert^2 \cdot \epsilon^{-2}\right)$	minimax-optimal
Duan and Wang [2020]	Linear	Linear	OPE	$\tilde{\mathcal{O}}\left(\kappa \cdot d \cdot \epsilon^{-2}\right)$	minimax-optimal
Le et al. [2019]	General	General	OPE/OPL	N/A	improper analysis
Yang et al. [2019]	ReLU nets	Hölder	OPL	$\tilde{\mathcal{O}}\left(K \cdot \kappa^{2+d/\alpha} \cdot \epsilon^{-2-d/\alpha}\right)$	no data reuse
Ours	**ReLU nets**	**Besov**	OPE/OPL	$\tilde{\mathcal{O}}\left(\kappa^{1+d/\alpha} \cdot \epsilon^{-2-2d/\alpha}\right)$	**data reuse**

Key insights

- We get rid of the algorithmic iteration K, an improvement over [Yang et al., 2019]
- Importantly, our sample complexity is established under the most general conditions so far: Besov dynamic closure and the data-dependent structure
- Technical proof: a uniform-convergence argument + local Rademacher complexity + a localization argument + upper bound minimization

Conclusion: An substantial improvement in generality, statistical efficiency and technique over the prior results.

References I

Yaqi Duan and Mengdi Wang. Minimax-optimal off-policy evaluation with linear function approximation. *CoRR*, abs/2002.09516, 2020.

Hoang Minh Le, Cameron Voloshin, and Yisong Yue. Batch policy learning under constraints. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 3703–3712. PMLR, 2019.

Rémi Munos and Csaba Szepesvári. Finite-time bounds for fitted value iteration. *J. Mach. Learn. Res.*, 9:815–857, 2008.

Zhuoran Yang, Yuchen Xie, and Zhaoran Wang. A theoretical analysis of deep q-learning. *CoRR*, abs/1901.00137, 2019.

Ming Yin and Yu-Xiang Wang. Asymptotically efficient off-policy evaluation for tabular reinforcement learning. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 3948–3958. PMLR, 2020.

References II

Ming Yin and Yu-Xiang Wang. Characterizing uniform convergence in offline policy evaluation via model-based approach: Offline learning, task-agnostic and reward-free, 2021.

Ming Yin, Yu Bai, and Yu-Xiang Wang. Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In Arindam Banerjee and Kenji Fukumizu, editors, *The 24th International Conference on Artificial Intelligence and Statistics, AISTATS 2021, April 13-15, 2021, Virtual Event*, volume 130 of *Proceedings of Machine Learning Research*, pages 1567–1575. PMLR, 2021. URL http://proceedings.mlr.press/v130/yin21a.html.