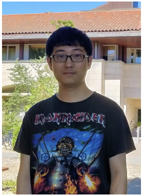
Variance-Aware Off-Policy Evaluation with Linear Function Approximation






Yifei Min[1] Tianhao Wang[1] Dongruo Zhou[2] Quanquan Gu[2]

[1]Department of Statistics and Data Science, Yale
[2]Department of Computer Science, UCLA

NeurIPS 2021

Outline

Reinforcement Learning

Reinforcement Learning (RL) research typically falls into either of the following two categories:

Reinforcement Learning

Reinforcement Learning (RL) research typically falls into either of the following two categories:

- Online RL, where the agent actively interacts with the environment to maximize some long-term cumulative rewards
    - E.g., episodic finite-horizon MDPs, discounted infinite-horizon MDPs, etc
- Offline RL (a.k.a, batch RL), where the goal is to extract useful information from the past data
    - E.g., offline policy optimization, **offline policy evaluation** (a.k.a., off-policy evaluation), etc

Reinforcement Learning

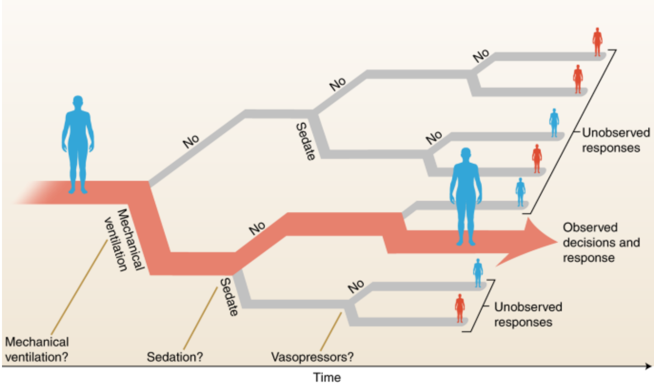
Reinforcement Learning (RL) research typically falls into either of the following two categories:

- Online RL, where the agent actively interacts with the environment to maximize some long-term cumulative rewards
  - E.g., episodic finite-horizon MDPs, discounted infinite-horizon MDPs, etc
- Offline RL (a.k.a, batch RL), where the goal is to extract useful information from the past data
  - E.g., offline policy optimization, **offline policy evaluation** (a.k.a., off-policy evaluation), etc

In this work, we study *off-policy evaluation* in the context of RL with *function approximation*, beyond the scope of traditional tabular MDPs.

(Offline) Reinforcement Learning

In offline Reinforcement Learning (RL), one of the most important tasks is to evaluate the value of an unobserved policy:



By Gottesman et al. – Guidelines for reinforcement learning in healthcare. https://www.nature.com/articles/s41591-018-0310-5

Off-policy evalution

In offline RL, *off-policy evaluation (OPE)* refers to a classic task which seeks to evaluate the performance of a **target policy** π given offline data generated by a **behavior policy** $\bar{\pi}$.

- Most existing works on OPE are for tabular MDPs (Precup, 2000; Jiang and Li, 2016; Yin et al., 2021)
- For linear MDPs (Yang and Wang, 2019; Jin et al., 2020), Duan et al. (2020) proposed a regression-based fitted Q-iteration method, `FQI-OPE`, which achieves a $\tilde{\mathcal{O}}(H^2\sqrt{(1+d(\pi,\bar{\pi}))/N})$ error bound where $d(\pi,\bar{\pi})$ is the distribution shift between π and $\bar{\pi}$
- Yet, the above error bound is *not* tight, because the **variance information** hidden in the data is not utilized

OPE for linear MDP

We consider the setting of linear MDP (Yang and Wang, 2019; Jin et al., 2020) where both the transition probabilities and reward functions can be linearly parametrized as

$$\mathbb{P}_h(s'|s,a) = \langle \phi(s,a), \boldsymbol{\mu}_h(s') \rangle, \;\; r_h(s,a) = \langle \phi(s,a), \boldsymbol{\theta}_h \rangle.$$

- The action-value function $Q_h^\pi(s,a)$ is also linear in the feature mapping ϕ (Jin et al., 2020), i.e., $\exists \mathrm{w}_h^\pi, Q_h^\pi(s,a) = \langle \phi(s,a), \mathrm{w}_h^\pi \rangle$

We assume that the offline data consists of K trajectories:

- Denote the dataset as $\mathcal{D}$ where $\mathcal{D} = \{\mathcal{D}_h\}_{h\in[H]}$. We assume $\mathcal{D}_{h_1}$ is independent of $\mathcal{D}_{h_2}$ for $h_1 \neq h_2$. For each stage h, we have $\mathcal{D}_h = \{(s_{k,h}, a_{k,h}, r_{k,h}, s'_{k,h})\}_{k\in[K]}$.

Notation and Technical Assumptions

- We define the following uncentered covariance matrix under behavior policy for all $h \in [H]$:

$$\mathbf{\Sigma}_h = \mathbb{E}_{\bar{\pi},h}\left[\phi(s,a)\phi(s,a)^\top\right]. \tag{3.1}$$

Assumption (Coverage)

For all $h \in [H]$, $\kappa_h = \lambda_{\min}(\mathbf{\Sigma}_h) > 0$.

- We define the weighted version of the covariance matrices:

$$\mathbf{\Lambda}_h = \mathbb{E}_{\bar{\pi},h}\left[\sigma_h(s,a)^{-2}\phi(s,a)\phi(s,a)^\top\right], \tag{3.2}$$

where

$$\sigma_h(s,a) \approx \sqrt{\mathbb{V}_h V_{h+1}^{\pi}(s,a)},$$

$$[\mathbb{V}_h V_{h+1}^{\pi}](s,a) = [\mathbb{P}_h (V_{h+1}^{\pi})^2](s,a) - \left([\mathbb{P}_h V_{h+1}^{\pi}](s,a)\right)^2,$$

$$[\mathbb{P}_h f](s,a) = \int_{\mathcal{S}} f(s')\mathrm{d}\mathbb{P}_h(s'|s,a) = \phi(s,a)^\top \int_{\mathcal{S}} f(s')\,\mathrm{d}\boldsymbol{\mu}_h(s').$$

Recap of results in Duan et al. (2020)

The dominant term in the error bound in Duan et al. (2020) is $\tilde{\mathcal{O}}(\sum_{h=1}^{H}(H-h+1)\|\mathrm{v}_h^\pi\|_{\boldsymbol{\Sigma}_h^{-1}}/\sqrt{K})$ where $H-h+1$ is the trivial upper bound of $\sqrt{\mathbb{V}_h V_{h+1}}$

- We can do better by estimating $\mathbb{V}_h V_{h+1}$ more precisely

To demonstrate the intuition, suppose we have iid samples $\{(s_{k,h}, a_{k,h}, s'_{k,h})\}_{k\in[K]}$, and the regression error is:

$$\mathrm{e}_k = \phi(s_{k,h}, a_{k,h}) \frac{[\mathbb{P}_h V_{h+1}^\pi](s_{k,h}, a_{k,h}) - V_{h+1}^\pi(s'_{k,h})}{[\mathbb{V}_h V_{h+1}^\pi](s_{k,h}, a_{k,h})^2}$$

By CLT, $\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathrm{e}_k \xrightarrow{d} \mathcal{N}(0, \mathrm{Cov}(\mathrm{e}_k))$, so $\mathrm{Cov}(\mathrm{e}_k)$ is the 'correct measure' of error

- This implies that we should use weighted regression
- But, how to estimate the variance?

Estimate Variance via regression

Variance of the value function:

$$
\begin{aligned}
[\mathbb{V}_h V_{h+1}^{\pi}](s,a) &= [\mathbb{P}_h (V_{h+1}^{\pi})^2](s,a) - \big([\mathbb{P}_h V_{h+1}^{\pi}](s,a)\big)^2 \\
&= \underbrace{\boldsymbol{\phi}(s,a)^\top \int_{\mathcal{S}} V_{h+1}^{\pi}(s')^2 \,\mathrm{d}\boldsymbol{\mu}_h(s')}_{\text{linear in } \phi(s,a)} - (\underbrace{[\mathbb{P}_h V_{h+1}^{\pi}](s,a)}_{\text{linear in } \phi(s,a)})^2
\end{aligned}
$$

Again, **regression**!

Algorithm: VA-OPE

Algorithm 1 Variance-Aware Off-Policy Evaluation (VA-OPE)

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

$\quad \hat{\mathbf{\Sigma}}_h \leftarrow \sum_{k=1}^{K} \check{\phi}_{k,h} \check{\phi}_{k,h}^\top + \lambda \mathbf{I}_d$

$\quad \hat{\beta}_h \leftarrow \hat{\mathbf{\Sigma}}_h^{-1} \sum_{k=1}^{K} \check{\phi}_{k,h} \hat{V}_{h+1}^{\pi}(\check{s}'_{k,h})^2$ (estimate second moment)

$\quad \hat{\theta}_h \leftarrow \hat{\mathbf{\Sigma}}_h^{-1} \sum_{k=1}^{K} \check{\phi}_{k,h} \hat{V}_{h+1}^{\pi}(\check{s}'_{k,h})$ (estimate first moment)

$\quad \hat{\sigma}_h(\cdot,\cdot) \leftarrow \sqrt{\max\{1, \hat{\mathbb{V}}_h \hat{V}_{h+1}^{\pi}(\cdot,\cdot)\} + 1}$ (estimate variance)

$\quad \hat{\mathbf{\Lambda}}_h \leftarrow \sum_{k=1}^{K} \phi_{k,h} \phi_{k,h}^\top / \hat{\sigma}_{k,h}^2 + \lambda \mathbf{I}_d$ (backward

$\quad Y_{k,h} \leftarrow r_{k,h} + \langle \phi_h^{\pi}(s'_{k,h}), \hat{w}_{h+1}^{\pi} \rangle$ weighted

$\quad \hat{w}_h^{\pi} \leftarrow \hat{\mathbf{\Lambda}}_h^{-1} \sum_{k=1}^{K} \phi_{k,h} Y_{k,h} / \hat{\sigma}_{k,h}^2$ regression)

$\quad \hat{Q}_h^{\pi}(\cdot,\cdot) \leftarrow \langle \phi(\cdot,\cdot), \hat{w}_h^{\pi} \rangle, \quad \hat{V}_h^{\pi}(\cdot) \leftarrow \langle \phi_h^{\pi}(\cdot), \hat{w}_h^{\pi} \rangle$

end for

Output: $\hat{v}_1^{\pi} \leftarrow \int_{\mathcal{S}} \hat{V}_1^{\pi}(s)\, \mathrm{d}\xi_1(s)$

Error bound for VA-OPE

Theorem (M., Wang, Zhou, Gu)

There exists some C such that with probability at least $1-\delta$, the output of `VA-OPE` *satisfies*

$$|v_1^\pi - \hat{v}_1^\pi| \leq C \cdot \left[\sum_{h=1}^{H} \|\mathrm{v}_h^\pi\|_{\mathbf{\Lambda}_h^{-1}}\right] \cdot \sqrt{\frac{\log(16H/\delta)}{K}}$$

where $\mathrm{v}_h^\pi = \mathbb{E}_{\pi,h}[\phi(s_h, a_h)]$.

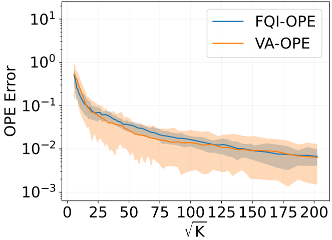
- $\sum_{h=1}^{H} \|\mathrm{v}_h^\pi\|_{\mathbf{\Lambda}_h^{-1}}$ characterizes the distribution shift between the target policy and behavior policy and is **instance-dependent** and **variance-aware**
- This recovers the result in Duan et al. (2020) in the worst case, and improves it by an order of $\Omega(H)$ in some cases

Numerical experiments

We test the performance of our algorithms on a hard-to-learn linear MDP instance (Zhou et al., 2021).

Numerical experiments

We test the performance of our algorithms on a hard-to-learn linear MDP instance (Zhou et al., 2021).



Comparison of `VA-OPE` and `FQI-OPE` under different settings of horizon length H. `VA-OPE`'s advantage becomes more significant as H increases, matching the theoretical prediction. The results are averaged over 50 trials and the error bars denote an empirical [10%,90%] confidence interval.

Conclusion

- For off-policy evaluation in RL with linear function approximation, we propose a weighted regression-based algorithm, `VA-OPE`
- Theoretical analysis demonstrates the superiority of our proposed method
- We also evaluate the performance of `VA-OPE` empirically via synthetic experiments, which corroborate our theory

Thank you!

Duan, Y., Jia, Z. and Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning*. PMLR.

Jiang, N. and Li, L. (2016). Doubly robust off-policy value evaluation for reinforcement learning. In *International Conference on Machine Learning*. PMLR.

Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR.

Precup, D. (2000). Eligibility traces for off-policy policy evaluation. *Computer Science Department Faculty Publication Series* 80.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*.

Yin, M., Bai, Y. and Wang, Y.-X. (2021). Near-optimal provable uniform convergence in offline policy evaluation for reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

ZHOU, D., HE, J. and GU, Q. (2021). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR.