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 University, [2]Department of Computer Science, UCLA,

NEURAL INFORMATION PROCESSING SYSTEMS

Off-policy Evaluation

Off-policy evaluation (OPE) refers to the problem of evaluating the performance of a target policy π given offline data generated by a behavior policy $\bar{\pi}$.

- Most existing theoretical works on OPE are in the setting of tabular MDPs (Precup, 2000; Li et al., 2011; Dudík et al., 2011; Jiang & Li, 2016; Xie et al., 2019; Yin & Wang, 2020; Yin et al., 2021), where the state space $\mathcal{S}$ and the action space $\mathcal{A}$ are both finite.
- Real-world applications often have high-dimensional or even infinite-dimensional state and action spaces, where function approximation is required for computational tractability and generalization.

In this work, we theoretically study the OPE problem for time-inhomogeneous **linear MDPs** (Yang & Wang, 2019; Jin et al., 2020) where the transition probability and reward function are assumed to be linear functions of a known feature mapping and may vary from stage to stage.

Problem Setting

We consider the time-inhomogeneous episodic MDP $M(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$:

- a **known feature mapping** $\boldsymbol{\phi}: \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$,
- for any $h \in [H]$, there exists $\boldsymbol{\gamma}_h$ and $\boldsymbol{\mu}_h \in \mathbb{R}^d$, such that for any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, it holds that

$$\mathbb{P}_h(\cdot \mid s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\mu}_h(\cdot)\rangle, \qquad r_h(s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\gamma}_h\rangle.$$

- Without loss of generality, we assume that $\|\boldsymbol{\gamma}_h\|_2 \le 1$ and $\|\boldsymbol{\phi}(s, a)\|_2 \le 1$ for all $(s, a) \in \mathcal{S} \times \mathcal{A}$.
- We assume that at any stage h, for any state-action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, the reward received by the agent is given by $r = r_h(s, a) + \epsilon_h(s, a)$, where $r_h(s, a) \in [0, 1]$ is the expected reward and $\epsilon_h(s, a)$ is the random noise.

Important property: for a linear MDP, for any policy π, there exist weights $\{\mathrm{w}_h^\pi, h \in [H]\}$ such that for any $(s, a, h) \in \mathcal{S} \times \mathcal{A} \times [H]$, we have $Q_h^\pi(s, a) = \langle \boldsymbol{\phi}(s, a), \mathrm{w}_h^\pi\rangle$. Moreover, we have $\|\mathrm{w}_h^\pi\|_2 \le 2H\sqrt{d}$ for all $h \in [H]$ (Jin et al., 2020).

Our Contributions

- We develop `VA-OPE` (Variance-Aware Off-Policy Evaluation), an algorithm for OPE that effectively utilizes the variance information from the offline data.
- We show that our algorithm achieves $\tilde{O}(\sum_h (\mathrm{v}_h^\top \Lambda_h^{-1} \mathrm{v}_h)^{1/2}/\sqrt{K})$ policy evaluation error, where v_h is the expectation of the feature vectors under target policy and Λ_h is the uncentered covariance matrix under behavior policy weighted by the conditional variance of the value function.
- Our analysis is based on a novel two-step proof technique. We also establish a uniform convergence result over all possible choices of the initial state.
- **Compared with the previous work** `FQI-OPE` (Duan et al., 2020), our algorithm achieves a tighter error bound and milder dependence on H, and provides a tighter characterization of the distribution shift between the behavior policy and the target policy, which is also verified by extensive numerical experiments.

Main Results

Theorem

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

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

where $\boldsymbol{b}_h^\pi = \mathbb{E}_{\pi,h}[\boldsymbol{\phi}(s_h, a_h)]$ and $\boldsymbol{\Lambda}_h = \mathbb{E}_{\bar{\pi},h}\left[\sigma_h(s, a)^{-2}\boldsymbol{\phi}(s, a)\boldsymbol{\phi}(s, a)^\top\right]$.

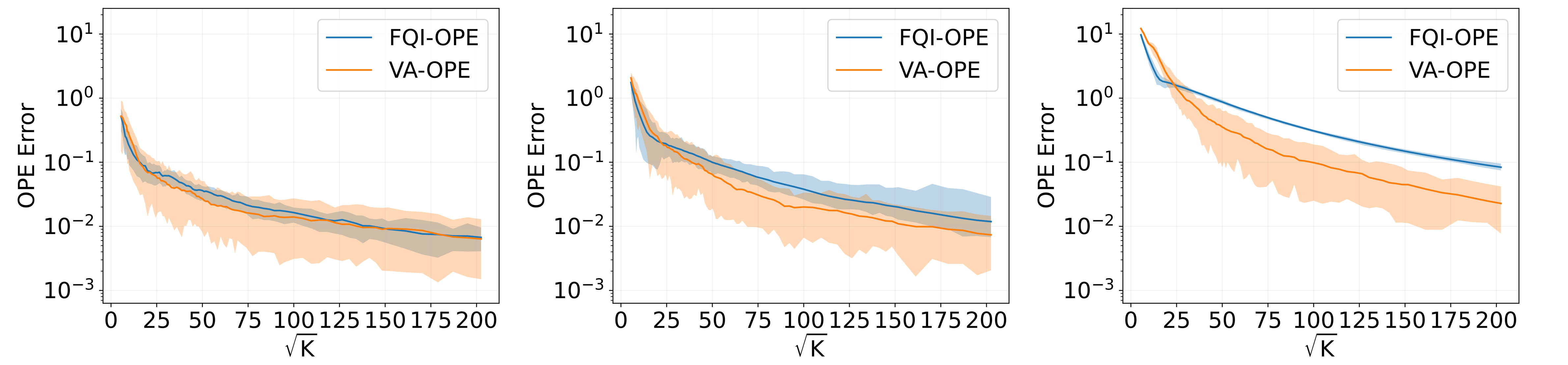
Remark: Compared with (Duan et al., 2020), their error upper bound is always $\tilde{O}(H^2)$, while ours is in between $\tilde{O}(H) \sim \tilde{O}(H^2)$, and is **instance-dependent**.

Algorithm

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

1: **for** $h = H, H-1, \ldots, 1$ **do**
2: $\quad \check{\boldsymbol{\Sigma}}_h \leftarrow \sum_{k=1}^K \check{\boldsymbol{\phi}}_{k,h} \check{\boldsymbol{\phi}}_{k,h}^\top + \lambda \mathbf{I}_d$
3: $\quad \hat{\boldsymbol{\beta}}_h \leftarrow \check{\boldsymbol{\Sigma}}_h^{-1} \sum_{k=1}^K \check{\boldsymbol{\phi}}_{k,h} \hat{V}_{h+1}^\pi(\check{s}'_{k,h})^2$ (estimate second moment)
4: $\quad \hat{\boldsymbol{\theta}}_h \leftarrow \check{\boldsymbol{\Sigma}}_h^{-1} \sum_{k=1}^K \check{\boldsymbol{\phi}}_{k,h} \hat{V}_{h+1}^\pi(\check{s}'_{k,h})$ (estimate first moment)
5: $\quad \hat{\sigma}_h(\cdot, \cdot) \leftarrow \sqrt{\max\{1, \hat{\mathbb{V}}_h \hat{V}_{h+1}^\pi(\cdot, \cdot)\} + 1}$ (estimate variance)
6: $\quad \hat{\boldsymbol{\Lambda}}_h \leftarrow \sum_{k=1}^K \boldsymbol{\phi}_{k,h} \boldsymbol{\phi}_{k,h}^\top / \hat{\sigma}_{k,h}^2 + \lambda \mathbf{I}_d$ (backward
7: $\quad Y_{k,h} \leftarrow r_{k,h} + \langle \boldsymbol{\phi}_h^\pi(s'_{k,h}), \hat{\mathrm{w}}_{h+1}^\pi\rangle$ weighted
8: $\quad \hat{\mathrm{w}}_h^\pi \leftarrow \hat{\boldsymbol{\Lambda}}_h^{-1} \sum_{k=1}^K \boldsymbol{\phi}_{k,h} Y_{k,h} / \hat{\sigma}_{k,h}^2$ regression)
9: $\quad \hat{Q}_h^\pi(\cdot, \cdot) \leftarrow \langle \boldsymbol{\phi}(\cdot, \cdot), \hat{\mathrm{w}}_h^\pi\rangle, \quad \hat{V}_h^\pi(\cdot) \leftarrow \langle \boldsymbol{\phi}_h^\pi(\cdot), \hat{\mathrm{w}}_h^\pi\rangle$
10: **end for**
11: **Output:** $\hat{v}_1^\pi \leftarrow \int_{\mathcal{S}} \hat{V}_1^\pi(s)\, \mathrm{d}\xi_1(s)$

Experiments



(a) **Small** h (b) **Medium** h (c) **Large** h

Duan, Y., Jia, Z., & Wang, M. (2020). Minimax-optimal off-policy evaluation with linear function approximation. In *International Conference on Machine Learning* (pp. 2701–2709). PMLR.

yifei.min@yale.edu