A Fine-grained Analysis of Fitted Q-evaluation: Beyond Parametric Models

Jiayi Wang [1] Zhengling Qi [2] Raymond K. W. Wong [3]

[1]University of Texas at Dallas [2] The George Washington University [3] Texas A&M University

Background

- In reinforcement learning (RL), off-policy evaluation (OPE) is an important topic that focuses on estimating the expected total reward of a target policy based on data collected from a potentially different and unknown policy.
- Among various algorithms for OPE, fitted Q-evaluation (FQE) is arguably one of the most popular algorithms.
- FQE has demonstrated significant empirical success in many applications, the theoretical analysis of FQE is less explored in current literature.
- We delve deeply into the analysis of FQE estimators within the framework of a finite-horizon, time-inhomogeneous Markov Decision Process (MDP).

Set up

- Finite-horizon episodic Markov Decision Process (MDP): T is the length of horizon, $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\Pr_t(\cdot \mid s, a)$ representing the transition kernel (probability) at step t given the state $s \in \mathcal{S}$ and the action $a \in \mathcal{A}$, R_t is the immediate reward at step t.
- Given pre-collected training data consist of n independent and identically distributed trajectories $\mathcal{D}_n = \left\{\{(S_{i,t}, A_{i,t}, R_{i,t})\}_{1\le t<T}\right\}_{1\le i\le n}$, OPE aims to estimate the value of π defined as
$$\nu(\pi) = \mathbb{E}^\pi\left[\sum_{t=1}^T R_t\right].$$
- Define $\rho_t^\pi(s,a)$ and $\rho_t^b(s,a)$ as the marginal density of (S_t, A_t) at $(s,a) \in \mathcal{S}\times\mathcal{A}$ under the target policy π and behavior policy π^b respectively. Define the probability ratio function w_t^π and Q-function as
$$w_t^\pi(s,a) = \rho_t^\pi(s,a)/\rho_t^b(s,a), \qquad Q_t^\pi(s,a) = \mathbb{E}^\pi\left[\sum_{t'=t}^T R_t \mid S_t = s, A_t = a\right].$$
- FQE: Recursively apply a regression technique to learn $Q_T^\pi, Q_{T-1}^\pi, \ldots, Q_1^\pi$ in a sequential and backward order. Let $\hat{Q}_{T+1}^\pi = 0$, and for $t = T, T-1, \ldots, 1$, one can compute
$$\hat{Q}_t^\pi = \arg\min_{Q\in\mathcal{Q}^{(t)}} \frac{1}{n}\sum_{i=1}^n\left\{Q(S_{i,t}, A_{i,t}) - \left[R_{i,t} + \sum_{a'\in\mathcal{A}_{t+1}} \pi_t(a \mid S_{i,t+1})\hat{Q}_{t+1}^\pi(S_{i,t+1}, a')\right]\right\}^2$$
and $\hat{\nu}(\pi) = \int_{(s,a)\in\mathcal{S}\times\mathcal{A}} \hat{Q}_1^\pi(s,a)\rho_1^\pi(s,a)d(s,a)$.
- One can let $\mathcal{Q}^{(t)} = \{\phi_K(\cdot)^\intercal\beta : \beta \in \mathbb{R}^{K|\mathcal{A}|}\}$, where ϕ_K are pre-specified features. We consider two scenarios:
  1. K is fixed $\to$ parametric setting.
  2. K is growing with n (and T) $\to$ nonparametric setting.

Three key questions

1. For the fixed horizon T, how does the convergence rate depend on the number of episodes n given the completeness assumption for Q-functions? Is the optimal convergence rate ($n^{-1/2}$) still achievable under nonparametric models of Q-functions?
2. How does the convergence rate depend on the growing horizon T?
3. What is the role of the probability ratio functions w_t^π in improving the convergence rate for FQE estimators?

Comparison on the error bounds

Table 1: Comparison on the error bound for the first-order term in existing works. κ is defined as an overlap constant. $\tilde{\kappa}$ is the upper bound for the probability ratio functions; D is the dimension of space and action. d is the intrinsic dimension of the state-action space. Some logarithmic orders are omitted in the error bounds.

Work	Parametric?	Regularity on Q	Error Bound
Yin & Wang (2020)	✓	Tabular	$\mathcal{O}(T\tilde{\kappa}\sqrt{1/n})$
Duan et al. (2020)	✓	Linear	$\mathcal{O}(T^2\sqrt{\kappa/n})$
Zhang et al. (2022)	✓	Differentiable	$\mathcal{O}(T^2\sqrt{\kappa/n})$
Nguyen-Tang et al. (2021)	×	Besov	$\mathcal{O}(T^{2-\alpha/(2\alpha+2D)}\tilde{\kappa}n^{-\alpha/(2\alpha+2D)})$
Ji et al. (2022)	×	Besov	$\mathcal{O}(T^2\kappa n^{-\alpha/(2\alpha+d)})$
Our Work for parametric setting	✓	Linear	$\mathcal{O}(T^{1.5}\sqrt{\kappa/n})$ $\mathcal{O}(T\tilde{\kappa}\sqrt{1/n})$ when w_t^π are linear
Our Work for nonparametric setting	×	Hölder	$\mathcal{O}(T^{1.5}\sqrt{\kappa/n})$ when Q_t^π are smooth enough $\mathcal{O}(T\tilde{\kappa}\sqrt{1/n})$ when w_t^π are Hölder

Connection with MIS estimators

- MIS estimator under the tabular setting is shown to have an error bound that has a linear dependence on the horizon (Yin & Wang (2020)).
- There is an equivalence between the FQE estimator and MIS estimator when adopting linear modeling of Q-functions (Duan et al. (2020)).
- Is linear dependence on the horizon for more general linear modeling (with potentially continuous state space) achievable for FQE estimators?

Parametric setting

Define $(\mathcal{P}_t^\pi f)(s,a) = \mathbb{E}\left\{\sum_{a'}\pi_t(a' \mid S_{t+1})f(S_{t+1}, a') \mid S_t = s, A_t = a\right\}$, $\kappa := \frac{1}{T}\sum_{t=1}^T \sup_{f\in\mathcal{Q}^{(t)}}[\mathcal{E}_t^\pi f]^2/\|f\|_{\mathcal{L}_2}^2$.

- **Assumption 1**: $\mathbb{E}\{R_t \mid S_t = \cdot, A_t = \cdot\} \in \mathcal{Q}^{(t)}$, for $t = 1, \ldots, T$. For every $q \in \mathcal{Q}^{(t+1)}$, we have $\mathcal{P}_t^\pi q \in \mathcal{Q}^{(t)}$.
- **Theorem 1**: $\mathcal{Q}^{(t)} = \left\{\phi_K(\cdot,\cdot)^\intercal\beta : \beta \in \mathbb{R}^{K|\mathcal{A}|}\right\}$ for some pre-specified feature ϕ_K and K is a fixed constant, under Assumption 1 and some technical conditions, if $T = o([n/(\log n \log T)]^{1/2})$, we have
$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}_{\mathrm{p}}\left(\sqrt{\frac{T^3\kappa}{n}} + T^3\frac{\log n \log T}{n}\right).$$
- Compared with the bound in Duan et al. (2020), our first order term has an order of $T^{1.5}/\sqrt{n}$. We achieve a sharper horizon dependence by exploiting the fact that the variance of the first order term can be decomposed as a sum of T individual expectations of the conditional variance.
- **Assumption 2**: $w_t^\pi \in \left\{\phi_K(\cdot,\cdot)^\intercal\beta : \beta \in \mathbb{R}^{K|\mathcal{A}|}\right\}$, $t = 1, \ldots, T$.
- **Theorem 2**: Under conditions listed in Theorem 1, if we futher assume Assumption 2, we have
$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}_{\mathrm{p}}\left(T\sqrt{\frac{\kappa}{n}} + T^3\frac{\log n \log T}{n}\right).$$
- Theorem 2 shows that with an additional realizability assumption (Assumption 2) on the probability ratio functions, the convergence rate of the error will depend *linearly* with respect to the horizon T in the first-order term. This is a significant improvement in horizon dependence over the existing literature on the setting of using linear function approximation.

Nonparametric setting: a slower rate

Define the projection Π_t such that $\Pi_t g(s,a) = \phi_K(s,a)^\intercal(\Sigma_t)^{-1}\mathbb{E}[\phi_K(S_t, A_t)g(S_t, A_t)]$.

- **Assumption 3:** For every $a \in \mathcal{A}$ and $t = 1, \ldots, T$, $\{q(\cdot, a) : q \in \mathcal{Q}^{(t)}, \|q\|_\infty \le 1\}$ is a subset of Hölder space $\Lambda_\infty(p, L)$ with constants $p > d/2$ and $L > 0$.
- **Assumption 4:** There exists a constant $\beta_Q > 1/2$ (independent of T) such that $\sup_{q\in\mathcal{Q}^{(t)}(1)}\|q - \Pi_t q\|_\infty \lesssim K^{-\beta_Q}$ for $t = 1, \ldots, T$.
- **Theorem 3:** Under Assumption 1, 3, 4 and some technical conditions, if we further assume that $K = o(\min\{\sqrt{n/(\log n \log T)}, n/(T^2 \log n \log T)\})$, $T = o(K^{\beta_Q})$, then we have
$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}_{\mathrm{p}}\left(T^2K^{-\beta_Q} + \sqrt{\frac{T^3\kappa}{n}} + \frac{T^3K\log n \log T}{n}\right).$$
- **Corollary 1:** Under conditions listed in Theorem 3, we further assume that T is bounded.

  (i) If $1/2 < \beta_Q \le 1$, then by taking $K \asymp \sqrt{n}/(\log n \log T)$, we have
$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}_{\mathrm{p}}\left(n^{-\beta_Q/2}\log n\right).$$
  (ii) If $\beta_Q > 1$, then by taking $K \asymp (n/(\log n))^{1/(1+\beta_Q)}$, we have
$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}_{\mathrm{p}}\left(n^{-1/2}\right).$$
- When β_Q is large enough, i.e., Q functions are smooth enough, we can achieve the optimal convergence rate $n^{-1/2}$.
- When $1/2 < \beta_Q \le 1$, by choosing K appropriately, our bound is faster than the optimal convergence rate $n^{-\beta_Q/(1+2\beta_Q)}$ for nonparametrically estimating the Q-functions

Nonparametric setting: a faster rate with realizability on ratio function

- **Assumption 5:** There exists a constant $\beta_w > 1/2$ such that $\sup_t\|w_t^\pi - \Pi_t w_t^\pi\|_\infty \lesssim K^{-\beta_w}$ for $t = 1, \ldots, T$.
- **Theorem 4:** Under conditions listed in Theorem 3, we futher assume Assumption 5, we have
$$|\hat{\nu}(\pi) - \nu(\pi)| = \mathcal{O}\left(\frac{T}{\sqrt{n}} + T^2K^{-\beta_Q-\beta_w} + T^3K^{-2\beta_Q} + T^3K^{-\beta_Q}\sqrt{\frac{K\log n \log T}{n}} + \frac{T^3K\log n\log T}{n}\right).$$
- **Corollary 2:** Under conditions listed in Theorem 4, we further assume $\beta_Q = \beta_w = \beta > 1/2$, $T\log T = o\left((n/\log n)^{\beta/(1+2\beta)}\right)$, by taking the optimal order of K such that
$$K \asymp \{n/(\log n \log T)\}^{\frac{1}{1+2\beta}},$$
we have $|\hat{\nu}(\pi) - \nu(\pi)| =$
$$\begin{cases} \mathcal{O}_{\mathrm{p}}\left(\frac{T}{\sqrt{n}}\right), \text{if } T = o\left(n^{\frac{2\beta-1}{4(1+2\beta)}}(\log n)^{\frac{-\beta}{1+2\beta}}\right), \\ \mathcal{O}_{\mathrm{p}}\left(T^3\left(\frac{n}{\log n\log T}\right)^{\frac{-2\beta}{1+2\beta}}\right), \text{otherwise.}\end{cases}$$
- If the number of horizon T is bounded, we can achieve the optimal convergence rate ($n^{-1/2}$) for $|\hat{\nu}(\pi) - \nu(\pi)|$ even though we estimate Q functions nonparametricly. Compared to Corollary 1, we do not require $\beta_Q > 1$ to achieve such optimal convergence rate.
- In the scenario where T grows relatively slowly compared to n (case 1), the convergence exhibits a $n^{-1/2}$ dependence with respect to n, with a linear dependence on the horizon. To the best of our knowledge, this convergence rate aligns with the best-known rate for FQE in tabular settings Yin & Wang (2020) (necessarily parametric), despite our analysis is conducted under a much more challenging nonparametric setting.