# Off-Policy Fitted Q-Evaluation with Differentiable Function Approximators: Z-Estimation and Inference Theory

Joint work with Xuezhou Zhang(Princeton), Chengzhuo Ni(Princeton), Mengdi Wang(Princeton and DeepMind)

Ruiqi Zhang

June, 2022

## OPE(Off Policy Evaluation)

- Environment: $\mathcal{MDP}(\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \xi, H)$
- Using behavior policy $\overline{\pi}$ to generate batch data $\mathcal{D} = \{(s_n, a_n, s_{n+1}, r_n)\}_{n \in [N]}$.
- Estimate policy value $v_\pi$ under target policy $\pi$.

Common Methods:
- Importance Sampling: IS, WIS, MIS, etc.
- Hybrid Method: Doubly Robust, etc.
- Direct Method: **Fitted Q-Evaluation(FQE),** etc.

## FQE with Function Approximation.

- **Initialize:** $\widehat{Q}^\pi_{H+1}(s, a) = 0, \forall (s, a) \in \mathcal{S} \times \mathcal{A}$.
- **For** $h = H, H-1, ..., 1$, **Solve** ($\lambda > 0$ and $\rho$ is a regularizer)

$$\widehat{Q}^\pi_h = \underset{f \in \mathcal{F}}{\arg\min} \left\{ \frac{1}{N} \sum_{n=1}^{N} \left[ f(s_n, a_n) - y_n \right]^2 + \lambda \rho(f) \right\}, \qquad (1)$$

  where $y_n = r(s_n, a_n) + \int_{\mathcal{A}} \widehat{Q}^\pi_{h+1}(s_{n+1}, a)\pi(a \mid s_{n+1}) \mathrm{d}a$.
- **Return** $\widehat{v}_\pi = \int_{\mathcal{S} \times \mathcal{A}} \widehat{Q}_1(s, a)\pi(a|s)\xi(s) \mathrm{d}a \mathrm{d}s$.

**Problem:** What is $\mathcal{F}$ ?

Many OPE methods leverage **function approximation** to avoid exponentially large variance, that is, using a function class $\mathcal{F}$ to approximate $Q_h^{\pi}(s,a)$.

**The choice of** $\mathcal{F}$ :

- (Xie et al. 19), (Yin et al. 20): Tabular Class: $\mathcal{S}$ and $\mathcal{A}$ are finite.
- (Duan et al. 20), (Hao et al. 21): Linear Class $\mathcal{F} = \left\{w^{\top} \cdot \phi(s,a)\right\}$
- (Kallus et al. 20), (Chen et al. 22), (Ji et al. 22): Non-parametric class.
- **Our work: General Differentiable Parametric Class**

$$\mathcal{F} = \{f(\theta;\phi(s,a)) : \theta \in \Theta\}, \quad \text{where } f \text{ is third-time differentiable.}$$

FQE $\Longrightarrow$ **Optimization in the parameter space** $\Theta$.

Denote $\widehat{Q}_h = f(\widehat{\theta}_h, \phi(s,a))$ and let $\rho(f) = \rho(\theta)$, then (1) can be written as

$$\widehat{\theta}_h = \operatorname*{arg\,min}_{\theta\in\Theta}\left\{\frac{1}{2N}\sum_{n=1}^{N}\left[f(\theta,\phi(s_n,a_n)) - y_n(\widehat{\theta}_{h+1})\right]^2 + \lambda\rho(\theta)\right\}, \qquad (2)$$

where $y_n(\theta') = r(s_n,a_n) + \int_{\mathcal{A}} f(\theta',\phi(s_{n+1},a))\pi(a \mid s_{n+1})\mathrm{d}a$.

## Asymptotic Optimality

**Theorem (Asymptotic Normality)**

*We have*

$$\sqrt{K}\left(\widehat{v}_{\pi} - v_{\pi}\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2), \qquad \text{when } K \to \infty, \lambda = o(K^{-1/2}),$$

*Here,* $\sigma^2$ **has a closed** *form dependent on* $f, \phi, , \theta_h^*, \xi, \pi$ *and* $\overline{\pi}$. *Here* $\theta_h^*$ *is the ground true parameter where* $Q_h^{\pi}(s, a) = f(\theta_h^*, \phi(s, a))$.

- Generalize results in linear case (Hao et al. 21) and tabular case (Yin et al. 20).
- The convergence rate of $|\widehat{v}_{\pi} - v_{\pi}|$ is $O\left(\frac{1}{\sqrt{K}}\right)$.

**Theorem (Cramer-Rao Lower Bound)**

*The variance of any unbiased estimator of* $v_{\pi}$ *is lower bounded by* $\sigma^2$.

- Asymptotically Efficiency of FQE.

## Finite Sample Upper Bound

**Theorem (Finite Sample Upper Bound)**

*We denote $\mu$ and $\bar{\mu}$ as the state-action occupation measure generated by policy $\pi$ and $\bar{\pi}$ respectively. With high probability, we have*

*(i). Variance-aware error bound:* $|\widehat{v}_\pi - v_\pi| \leq \sqrt{\frac{2\log(6/\delta)\sigma^2}{K}} + O\left(\frac{1}{K}\right),$

*(ii). Reward-free error bound:*

$$|\widehat{v}_\pi - v_\pi| \leq \left[\sum_{h=1}^{H} \sqrt{1 + \chi^2_{\mathcal{G}_h}(\mu, \bar{\mu})}\right] \cdot \sqrt{\frac{H}{2K} \log\left(\frac{12}{\delta}\right)} + O\left(\frac{1}{K}\right),$$

*where* $\mathcal{G}_h := \left\{\left(\nabla_{\theta_h} f\left(\theta_h^*, \phi(s,a)\right)\right) \cdot \boldsymbol{x} : \boldsymbol{x} \in \mathbb{R}^d\right\}$.

$\mathcal{F}$**-Restricted chi-square:** Measuring the distribution shift in the function class.

$$\chi^2_{\mathcal{F}}(p_1, p_2) := \sup_{f \in \mathcal{F}} \frac{\mathbb{E}_{p_1}[f(x)]^2}{\mathbb{E}_{p_2}[f(x)^2]} - 1. \quad (3)$$

- $\chi^2_{\mathcal{G}_h}(\mu, \bar{\mu}) \ll \chi^2(\mu, \bar{\mu}) \ll \|\mu/\bar{\mu}\|_\infty$.
- In linear case, $\mathcal{G}_h = \mathcal{F}$ and the result is minimax optimal(Duan et al. 20).