A. Proofs of Main Theorems

A.1. Full Algorithm of General FQE

Algorithm 2 Fitted Q-Evaluation (Le et al., 2019)

input Dateset $\mathcal{D} = \{\mathcal{D}_1, \ldots, \mathcal{D}_K\}$, target policy π, function class $\mathcal{F}$, initial state distribution ξ_0.

1: Initialize $\widehat{Q}^\pi_{H+1} = 0$.

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

3: Compute regression targets for any $k \in [K], h' \in [H]$:

$$y^k_{h,h'} = r^k_{h'} + \int_a \widehat{Q}^\pi_{h+1}(s^k_{h'+1}, a)\pi(a|s^k_{h'+1})\mathrm{d}a.$$

4: Build training set $\{(s^k_{h'}, a^k_{h'}), y^k_{h,h'}\}_{k\in[K],h'\in[H]}$.

5: Solve a supervised learning problem:

$$\widehat{Q}^\pi_h = \operatorname*{argmin}_{f\in\mathcal{F}} \left\{ \frac{1}{K}\sum_{k=1}^K \frac{1}{H}\sum_{h'=1}^H \left(f(s^k_{h'}, a^k_{h'}) - y^k_{h,h'}\right)^2 + \lambda\rho(f) \right\},$$

where $\rho(f)$ is a proper regularizer.

6: **end for**

output $\widehat{v}^\pi = \int_s \int_a \widehat{Q}^\pi_1(s,a)\xi_1(s)\pi(a|s)\mathrm{d}s\mathrm{d}a$.

We restate the full algorithm of FQE in Algorithm 2. Here we simply assume the initial state distribution ξ_1 is known. In practice, we always have the access to sample from ξ_1 and thus we can approximate it by Monte Carlo sampling.

A.2. Equivalence between FQE and model-based plug-in estimator

We show that the FQE in Algorithm 2 with linear function class $\mathcal{F}$ is equivalent to a plug-in estimator. This equivalence is helpful to derive the asymptotic normality of FQE and bootstrapping FQE. Define

$$\widehat{M}_\pi = \widehat{\Sigma}^{-1}\sum_{n=1}^N \phi(s_n,a_n)\phi^\pi(s_{n+1})^\top, \widehat{R} = \widehat{\Sigma}^{-1}\sum_{n=1}^N r_n\phi(s_n,a_n), \widehat{\Sigma} = \sum_{n=1}^N \phi(s_n,a_n)\phi(s_n,a_n)^\top + \lambda I_d, \tag{A.1}$$

where $\phi^\pi(s) = \int_a \phi(s,a)\pi(a|s)\mathrm{d}a$, s_{N+1} is the terminal state and λ is the regularization parameter. Choosing $\rho(f) = \lambda I$, the FQE is equivalent to, for $h = H, \ldots, 1$, $\widehat{Q}_h(s,a) = \phi(s,a)^\top \widehat{w}^\pi_h$ with

$$\begin{aligned}
\widehat{w}^\pi_h &= \widehat{\Sigma}^{-1}\sum_{n=1}^N \phi(s_n,a_n)\Big(r_n + \int_a \widehat{Q}^\pi_{h+1}(s_{n+1},a)\pi(a|s_{n+1})\mathrm{d}a\Big) \\
&= \widehat{\Sigma}^{-1}\sum_{n=1}^N \phi(s_n,a_n)\Big(r_n + \int_a \phi(s_{n+1},a)^\top \widehat{w}^\pi_{h+1}\pi(a|s_{n+1})\mathrm{d}a\Big) \\
&= \widehat{\Sigma}^{-1}\sum_{n=1}^N \phi(s_n,a_n)r_n + \widehat{\Sigma}^{-1}\sum_{n=1}^N \phi(s_n,a_n)\phi^\pi(s_{n+1})^\top \widehat{w}_{h+1} \\
&= \widehat{R} + \widehat{M}_\pi \widehat{w}^\pi_{h+1}.
\end{aligned}$$

This gives us a recursive form of $\widehat{w}^\pi_h$. Denoting $\widehat{w}^\pi_{H+1} = 0$ and $\nu^\pi_1 = \mathbb{E}_{s\sim\xi_1, a\sim\pi(\cdot|s)}[\phi(s,a)]$, the FQE estimator can be written into

$$\widehat{v}_\pi = \int_s\int_a \widehat{Q}_1(s,a)\xi_1\pi(a|s)\mathrm{d}a\mathrm{d}s = (\nu^\pi_1)^\top \widehat{w}^\pi_1 = (\nu^\pi_1)^\top \sum_{h=0}^{H-1} (\widehat{M}_\pi)^h \widehat{R}. \tag{A.2}$$

On the other hand, from Condition 4.1, there exists some $w_r, w_h^\pi \in \mathbb{R}^d$ such that $Q_h^\pi(\cdot,\cdot) = \phi(\cdot,\cdot)^\top w_h^\pi$ for each $h \in [H]$ and $r(\cdot,\cdot) = \phi(\cdot,\cdot)^\top w_r$ and there exists $M_\pi \in \mathbb{R}^{d\times d}$ such that $\phi(s,a)^\top M_\pi = \mathbb{E}[\phi^\pi(s')^\top | s,a]$. From Bellman equation and Condition 4.1,

$$\begin{aligned} Q_h^\pi(s,a) &= r(s,a) + \mathbb{E}\Big[\int_a Q_{h+1}^\pi(s',a)\pi(a|s')\mathrm{d}a|s,a\Big] \\ &= \phi(s,a)^\top w_r + \phi(s,a)^\top \mathbb{E}[\phi^\pi(s')^\top|s,a] w_{h+1}^\pi = \phi(s,a)^\top\Big(w_r + M_\pi w_{h+1}^\pi\Big) \\ &= \phi(s,a)^\top \sum_{h=0}^{H-h} (M_\pi)^h w_r. \end{aligned} \tag{A.3}$$

Therefore, the true scalar value function can be written as

$$v_\pi = \mathbb{E}_{s\sim\xi_1, a\sim\pi(\cdot|s)}\Big[Q_1^\pi(s,a)\Big] = (\nu_1^\pi)^\top \sum_{h=0}^{H-1}(M_\pi)^h w_r,$$

which implies Eq. (A.2) is a plug-in estimator.

A.3. Proof of Theorem 4.2: Asymptotic normality of FQE

Recall $\nu_h^\pi = \mathbb{E}^\pi[\phi(x_h,a_h)|x_1\sim\xi_1]$ and denote $(\widehat{\nu}_h^\pi)^\top = (\nu_1^\pi)^\top\big(\widehat{M}_\pi\big)^{h-1}$. We follow Lemma B.3 in Duan & Wang (2020) to decompose the error term into following three parts:

$$\sqrt{N}(v_\pi - \widehat{v}_\pi) = E_1 + E_2 + E_3,$$

where

$$\begin{aligned} E_1 &= \frac{1}{\sqrt{N}}\sum_{n=1}^N\sum_{h=1}^H (\nu_h^\pi)^\top\Sigma^{-1}\phi(s_n,a_n)\Big(Q_h^\pi(s_n,a_n) - \big(r_n + V_{h+1}^\pi(s_{n+1})\big)\Big), \\ E_2 &= \sum_{h=1}^H\Big(N(\widehat{\nu}_h^\pi)^\top\widehat{\Sigma}^{-1} - (\nu_h^\pi)^\top\Sigma^{-1}\Big)\Big(\frac{1}{\sqrt{N}}\sum_{n=1}^N\phi(s_n,a_n)\Big(Q_h^\pi(s_n,a_n) - \big(r_n + V_{h+1}^\pi(s_{n+1})\big)\Big)\Big), \\ E_3 &= \lambda\frac{1}{\sqrt{N}}\sum_{h=0}^H(\widehat{\nu}_h^\pi)^\top\widehat{\Sigma}^{-1}w_h^\pi. \end{aligned}$$

To prove the asymptotic normality of $\sqrt{N}(v_\pi - \widehat{v}_\pi)$, we will first prove the asymptotic normality of E_1 and then show both E_1 and E_2 are asymptotically negligible.

For $n = 1, 2, \ldots, N$, we denote

$$e_n = \frac{1}{\sqrt{N}}\sum_{h=1}^H(\nu_h^\pi)^\top\Sigma^{-1}\phi(s_n,a_n)\Big(Q_h^\pi(s_n,a_n) - \big(r_{n+1} + V_{h+1}^\pi(s_{n+1})\big)\Big).$$

Then $E_1 = \sum_{n=1}^N e_n$. Define a filtration $\{\mathcal{F}_n\}_{n=1,\ldots,N}$ with $\mathcal{F}_n$ generated by $(s_1,a_1,s_2),\ldots,(s_{n-1},a_{n-1},s_n)$ and (s_n,a_n). From the definition of value function, it is easy to see $\mathbb{E}[e_n\,|\,\mathcal{F}_n] = 0$ that implies that $\{e_n\}_{n\in[N]}$ is a martingale difference sequence. To show the asymptotic normality, we use the following martingale central limit theorem for triangular arrays.

Theorem A.1 (Martingale CLT, Corollary 2.8 in (McLeish et al., 1974))**.** Let $\{X_{mn}; n = 1,\ldots,k_m\}$ be a martingale difference array (row-wise) on the probability triple $(\Omega,\mathcal{F},P)$. Suppose X_{mn} satisfy the following two conditions:

$$\max_{1\le n\le k_m}|X_{mn}| \xrightarrow{p} 0, \text{and } \sum_{n=1}^{k_m} X_{mn}^2 \xrightarrow{p} \sigma^2,$$

for $k_m\to\infty$. Then $\sum_{n=1}^{k_m} X_{mn} \xrightarrow{d} \mathcal{N}(0,\sigma^2)$.

Recall that the variance σ^2 is defined as

$$\sigma^2 = \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1} \Omega_{h,h} \Sigma^{-1} \nu_h^\pi + 2 \sum_{h_1<h_2} (\nu_{h_1}^\pi)^\top \Sigma^{-1} \Omega_{h_1,h_2} \Sigma^{-1} \nu_{h_2}^\pi, \tag{A.4}$$

and for any $h_1 \in [H], h_2 \in [H]$,

$$\Omega_{h_1,h_2} = \mathbb{E}\Big[\frac{1}{H} \sum_{h'=1}^{H} \phi(s_{h'}^1, a_{h'}^1) \phi(s_{h'}^1, a_{h'}^1)^\top \varepsilon_{h_1,h'}^1 \varepsilon_{h_2,h'}^1\Big],$$

where $\varepsilon_{h_1,h'}^1 = Q_{h_1}^\pi(s_{h'}^1, a_{h'}^1) - (r_{h'}^1 + V_{h_1+1}^\pi(s_{h'+1}^1))$. To apply Theorem A.1, we let $k_m = N$, $X_{mn} = e_n$ and we need to verify the following two conditions:

$$\max_{1\le n\le N} \Big| \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1} \left(\frac{1}{\sqrt{N}} \phi(s_n, a_n) \Big(Q_h^\pi(s_n, a_n) - \big(r_{n+1} + V_{h+1}^\pi(s_{n+1})\big)\Big)\right) \Big| \xrightarrow{p} 0, \text{ as } N \to \infty, \tag{A.5}$$

and

$$\sum_{n=1}^{N} \Big(\frac{1}{\sqrt{N}} \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1} \phi(s_n, a_n) \Big(Q_h^\pi(s_n, a_n) - \big(r_{n+1} + V_{h+1}^\pi(s_{n+1})\big)\Big)\Big)^2 \xrightarrow{p} \sigma^2, \text{ as } N \to \infty. \tag{A.6}$$

Verify Condition A.5: Since $r \in [0,1]$, we have $r_n + V_{h+1}^\pi(s_{n+1}) \in [0, H-h]$. For any $n \in [N]$, we have

$$\begin{aligned}
&\left| \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1} \left(\frac{1}{\sqrt{N}} \phi(s_n, a_n) \Big(Q_h^\pi(s_n, a_n) - \big(r_{n+1} + V_{h+1}^\pi(s_{n+1})\big)\Big)\right) \right| \\
&\le \frac{1}{\sqrt{N}} \sum_{h=1}^{H} \Big| (\nu_h^\pi)^\top \Sigma^{-1} \phi(s_n, a_n) \Big| \Big| Q_h^\pi(s_n, a_n) - \big(r_n + V_{h+1}^\pi(s_{n+1})\big) \Big| \\
&\le \frac{1}{\sqrt{N}} \sum_{h=1}^{H} (H-h+1) \Big| (\nu_h^\pi)^\top \Sigma^{-1} \phi(s_n, a_n) \Big|.
\end{aligned}$$

Note that $(\nu_h^\pi)^\top \Sigma^{-1} \phi(s_n, a_n)$ is independent of N. Then for fixed d, H, Condition A.5 is satisfied when $N \to \infty$.

Verify Condition A.6: Recall the definition of σ^2 in Eq. (A.4) and let $\sigma^2 = \sigma_1^2 + \sigma_2^2$ for

$$\begin{aligned}
\sigma_1^2 &= \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1} \Omega_{h,h} \Sigma^{-1} \nu_h^\pi, \\
\sigma_2^2 &= 2 \sum_{h_1<h_2} (\nu_{h_1}^\pi)^\top \Sigma^{-1} \Omega_{h_1,h_2} \Sigma^{-1} \nu_{h_2}^\pi.
\end{aligned}$$

Using the following decomposition,

$$\begin{aligned}
&\sum_{n=1}^{N} \Big(\frac{1}{\sqrt{N}} \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1} \phi(s_n, a_n) \Big(Q_h^\pi(s_n, a_n) - \big(r_{n+1} + V_{h+1}^\pi(s_{n+1})\big)\Big)\Big)^2 \\
&= \sum_{n=1}^{N} \frac{1}{N} \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1} \phi(s_n, a_n) \phi(s_n, a_n)^\top \Sigma^{-1} \nu_h^\pi \Big(Q_h^\pi(s_n, a_n) - (r_n + V_{h+1}^\pi(s_{n+1}))\Big)^2 \\
&\quad + \sum_{n=1}^{N} \frac{1}{N} 2 \sum_{h_1<h_2} (\nu_{h_1}^\pi)^\top \Sigma^{-1} \phi(s_n, a_n) (\nu_{h_2}^\pi)^\top \Sigma^{-1} \phi(s_n, a_n) \\
&\quad \cdot \Big(Q_{h_1}^\pi(s_n, a_n) - \big(r_{n+1} + V_{h_1+1}^\pi(s_{n+1})\big)\Big) \Big(Q_{h_2}^\pi(s_n, a_n) - \big(r_{n+1} + V_{h_2+1}^\pi(s_{n+1})\big)\Big).
\end{aligned}$$

We denote the first term as I_1, the second term as I_2 and separately bound $I_1 - \sigma_1^2$ and $I_2 - \sigma_2^2$ as follows:

- We rewrite I_1 in terms of episodes as

$$I_1 = \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1/2} \Big(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{H} \sum_{h=1}^{H} \Sigma^{-1/2} \phi(s_{h'}^k, a_{h'}^k) \phi(s_{h'}^k, a_{h'}^k)^\top (\varepsilon_{hh'}^k)^2 \Sigma^{-1/2} \Big) \Sigma^{-1/2} \nu_h^\pi.$$

  Moreover, denote

$$Z_h = \Sigma^{-1/2} \mathbb{E} \Big[\frac{1}{H} \sum_{h'=1}^{H} \phi(s_{h'}^1, a_{h'}^1) \phi(s_{h'}^1, a_{h'}^1)^\top (\varepsilon_{hh'}^1)^2 \Big] \Sigma^{-1/2} \in \mathbb{R}^{d \times d}.$$

  Then we have

$$\begin{aligned} |I_1 - \sigma_1^2| &= \sum_{h=1}^{H} (\nu_h^\pi)^\top \Sigma^{-1/2} \Big(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{H} \sum_{h'=1}^{H} \Sigma^{-1/2} \phi(s_{h'}^k, a_{h'}^k) \phi(s_{h'}^k, a_{h'}^k)^\top (\varepsilon_{hh'}^k)^2 \Sigma^{-1/2} - Z_h \Big) \Sigma^{-1/2} \nu_h^\pi \\ &\le \sum_{h=1}^{H} \Big\| (\nu_h^\pi)^\top \Sigma^{-1/2} \Big\|_2^2 \Big\| \frac{1}{K} \sum_{k=1}^{K} \Big(\frac{1}{H} \sum_{h'=1}^{H} \Sigma^{-1/2} \phi(s_{h'}^k, a_{h'}^k) \phi(s_{h'}^k, a_{h'}^k)^\top (\varepsilon_{hh'}^k)^2 \Sigma^{-1/2} - Z_h \Big) \Big\|_2, \end{aligned}$$

  where the last inequality is from Cauchy–Schwarz inequality. From Lemma B.7, we reach $I_1 \xrightarrow{p} \sigma_1^2$ as $K \to \infty$.

- We rewrite I_2 as

$$I_2 = 2 \sum_{h_1 < h_2} (\nu_{h_1}^\pi)^\top \Sigma^{-1/2} \Big(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{H} \sum_{h'=1}^{H} \Sigma^{-1/2} \phi(s_{h'}^k, a_{h'}^k) \phi(s_{h'}^k, a_{h'}^k)^\top \varepsilon_{h_1 h'}^k \varepsilon_{h_2 h'}^k \Sigma^{-1/2} \Big) \Sigma^{-1/2} \nu_{h_2}^\pi,$$

  and denote

$$Z_{h_1 h_2} = \Sigma^{-1/2} \mathbb{E} \Big[\frac{1}{H} \sum_{h'=1}^{H} \phi(s_{h'}^1, a_{h'}^1) \phi(s_{h'}^1, a_{h'}^1)^\top \varepsilon_{h_1 h'}^1 \varepsilon_{h_2 h'}^1 \Big] \Sigma^{-1/2} \in \mathbb{R}^{d \times d}.$$

  Then we have

$$\begin{aligned} |I_2 - \sigma_2^2| &= 2 \sum_{h_1 < h_2} (\nu_{h_1}^\pi)^\top \Sigma^{-1/2} \Big(\frac{1}{K} \sum_{k=1}^{K} \frac{1}{H} \sum_{h'=1}^{H} \Sigma^{-1/2} \phi(s_{h'}^k, a_{h'}^k) \phi(s_{h'}^k, a_{h'}^k)^\top \varepsilon_{h_1 h'}^k \varepsilon_{h_2 h'}^k \Sigma^{-1/2} - Z_{h_1 h_2} \Big) \Sigma^{-1/2} \nu_{h_2}^\pi \\ &\le 2 \sum_{h_1 < h_2} \Big\| (\nu_{h_1}^\pi)^\top \Sigma^{-1/2} \Big\|_2 \Big\| (\nu_{h_2}^\pi)^\top \Sigma^{-1/2} \Big\|_2 \Big\| \frac{1}{K} \sum_{k=1}^{K} \Big(\frac{1}{H} \sum_{h'=1}^{H} \Sigma^{-1/2} \phi(s_{h'}^k, a_{h'}^k) \phi(s_{h'}^k, a_{h'}^k)^\top \varepsilon_{h_1 h'}^k \varepsilon_{h_2 h'}^k \Sigma^{-1/2} - Z_{h_1 h_2} \Big) \Big\|_2. \end{aligned}$$

  From Lemma B.7, we reach $I_2 \xrightarrow{p} \sigma_2^2$ as $K \to \infty$.

Putting the above two steps together, we have verified Condition A.6. Then applying Theorem A.1 we obtain that $E_1 \xrightarrow{d} \mathcal{N}(0, \sigma^2)$.

On the other hand, according to Lemmas B.6, B.10 in (Duan & Wang, 2020),

$$\begin{aligned} |E_2| &\le 15 \sqrt{(\nu_0^\pi)^\top (\Sigma^\pi)^{-1} \nu_0^\pi} \cdot \big\| (\Sigma^\pi)^{1/2} \Sigma^{-1/2} \big\|_2 \cdot \sqrt{C_1 \kappa_1} (2 + \kappa_2) \cdot \frac{\ln(8dH/\delta) dH^{3.5}}{\sqrt{N}} \\ |E_3| &\le \sqrt{(\nu_0^\pi)^\top (\Sigma^\pi)^{-1} \nu_0^\pi} \cdot \big\| (\Sigma^\pi)^{1/2} \Sigma^{-1/2} \big\|_2 \cdot \frac{5 \ln(8dH/\delta) C_1 dH^2}{\sqrt{N}}, \end{aligned}$$

with probability at least $1 - \delta$ and κ_1, κ_2 are some problem-dependent constants that do not depend on N. When $N \to \infty$, both $|E_2|, |E_3|$ converge in probability to 0. By Slutsky's theorem, we have proven the asymptotic normality of $\sqrt{N}(v_\pi - \widehat{v}_\pi)$.

■

A.4. Proof of Theorem 4.5: Efficiency bound

Influence function. Recall that our dataset $\mathcal{D}$ consists of K *i.i.d.* trajectories, each of which has length H. Denote

$$\boldsymbol{\tau} := \big(s_1, a_1, r_1, s_2, a_2, r_2, \ldots, s_H, a_H, r_H, s_{H+1}\big).$$

For simplicity, we assume that the reward r_h is deterministic given (s_h, a_h), *i.e.* $r_h = r(s_h, a_h)$ for some reward function r. The distribution of $\boldsymbol{\tau}$ is given by

$$\begin{aligned}\mathcal{P}(\mathrm{d}\boldsymbol{\tau}) =& \bar{\xi}_1(\mathrm{d}s_1, \mathrm{d}a_1)\mathcal{P}(\mathrm{d}s_2 \,\big|\, s_1, a_1)\bar{\pi}(\mathrm{d}a_2 \mid s_2)\mathcal{P}(\mathrm{d}s_3 \mid s_2, a_2)\\ &\ldots \mathcal{P}(\mathrm{d}s_H \mid s_{H-1}, a_{H-1})\bar{\pi}(\mathrm{d}a_H \mid s_H)\mathcal{P}(\mathrm{d}s_{H+1} \mid s_H, a_H).\end{aligned}$$

Define $\mathcal{P}_\eta := \mathcal{P} + \eta\Delta\mathcal{P}$ where $\Delta\mathcal{P}$ satisfies

$$(\Delta\mathcal{P})\mathcal{F} \subseteq \mathcal{F}$$

under condition 4.1. Denote score functions

$$g(\boldsymbol{\tau}) := \frac{\partial}{\partial\eta}\log\mathcal{P}_\eta(\mathrm{d}\boldsymbol{\tau}) \qquad \text{and} \qquad g(s' \mid s, a) := \frac{\partial}{\partial\eta}\log\mathcal{P}_\eta(\mathrm{d}s' \mid s, a).$$

Note that

$$g(\boldsymbol{\tau}) = \sum_{h=1}^{H} g(s_{h+1} \mid s_h, a_h).$$

We consider the pointwise estimation. The objective function ψ_{ξ_1} is defined as

$$\psi_{\xi_1}(\mathcal{P}_\eta) := \mathbb{E}\left[\sum_{h=1}^{H} r_\eta(s_h, a_h) \,\middle|\, (s_1, a_1) \sim \xi_1, \mathcal{P}_\eta, \pi\right].$$

We calculate the derivative $\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta)$ and have

$$\begin{aligned}\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta) =& \frac{\partial}{\partial\eta}\left[\sum_{h=1}^{H}\int_{(\mathcal{S}\times\mathcal{A})^h} r(s_h, a_h)\xi_1(\mathrm{d}s_1, \mathrm{d}a_1)\prod_{j=1}^{h-1}\mathcal{P}_\eta(\mathrm{d}s_{j+1} \mid s_j, a_j)\pi(\mathrm{d}a_{j+1} \mid s_{j+1})\right]\\ =& \sum_{h=1}^{H}\int_{(\mathcal{S}\times\mathcal{A})^h} r(s_h, a_h)\left(\sum_{j=1}^{h-1} g(s_{j+1} \,|\, s_j, a_j)\right)\xi_1(\mathrm{d}s_1, \mathrm{d}a_1)\prod_{j=1}^{h-1}\mathcal{P}_\eta(\mathrm{d}s_{j+1} \,|\, s_j, a_j)\pi(\mathrm{d}a_{j+1} \mid s_{j+1}).\end{aligned}$$

By using Q-functions $Q^\pi_{\eta,j}(s_j, a_j) := \mathbb{E}\big[\sum_{h=j}^{H} r_\eta(s_h, a_h) \,\big|\, (s_j, a_j), \mathcal{P}_\eta, \pi\big]$ for $j = 1, 2, \ldots, H$, $Q_{\eta,H+1} := 0$, we find that

$$\begin{aligned}\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta) =& \int_{(\mathcal{S}\times\mathcal{A})^H}\sum_{j=1}^{H-1} g(s_{j+1} \,|\, s_j, a_j)\sum_{h=j+1}^{H} r_\eta(s_h, a_h)\xi_1(\mathrm{d}s_1, \mathrm{d}a_1)\prod_{i=1}^{H-1}\mathcal{P}_\eta(\mathrm{d}s_{i+1} \,|\, s_i, a_i)\pi(\mathrm{d}a_{i+1} \mid s_{i+1})\\ =& \int_{(\mathcal{S}\times\mathcal{A})^H}\sum_{j=1}^{H-1} g(s_{j+1} \,|\, s_j, a_j)\xi_1(\mathrm{d}s_1, \mathrm{d}a_1)\prod_{i=1}^{j}\mathcal{P}_\eta(\mathrm{d}s_{i+1} \,|\, s_i, a_i)\pi(\mathrm{d}a_{i+1} \mid s_{i+1})\\ &\cdot\left(\sum_{h=j+1}^{H} r_\eta(s_h, a_h)\prod_{i=j+1}^{H-1}\mathcal{P}_\eta(\mathrm{d}s_{i+1} \,|\, s_i, a_i)\pi(\mathrm{d}a_{i+1} \mid s_{i+1})\right)\\ =& \sum_{j=1}^{H-1}\int_{(\mathcal{S}\times\mathcal{A})^{j+1}} g(s_{j+1} \,|\, s_j, a_j)Q^\pi_{\eta,j+1}(s_{j+1}, a_{j+1})\xi_1(\mathrm{d}s_1, \mathrm{d}a_1)\prod_{i=1}^{j}\mathcal{P}_\eta(\mathrm{d}s_{i+1} \,|\, s_i, a_i)\pi(\mathrm{d}a_{i+1} \mid s_{i+1})\\ =& \sum_{h=1}^{H}\mathbb{E}\big[g(s_{h+1} \,|\, s_h, a_h)V^\pi_{\eta,h+1}(s_{h+1}) \,\big|\, (s_1, a_1) \sim \xi_1, \mathcal{P}_\eta, \pi\big].\end{aligned}$$

It follows that

$$\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta)\Big|_{\eta=0} = \mathbb{E}\left[\sum_{h=1}^{H} g(s_{h+1} \mid s_h, a_h) V_{h+1}^{\pi}(s_{h+1}) \;\middle|\; (s_1, a_1) \sim \xi_1, \mathcal{P}, \pi\right].$$

Define $w_h(s,a) := \phi(s,a)^\top \Sigma^{-1} \nu_h^\pi = \phi(s,a)^\top \Sigma^{-1} \mathbb{E}\big[\phi(s_h,a_h) \;\big|\; (s_1,a_1) \sim \xi_1, \mathcal{P}, \pi\big]$ for $h = 1, 2, \ldots, H$. Let $\mathbb{H}$ For any $f \in \mathbb{H}$ with $f(s,a) = \phi(s,a)^\top w_f$, we have

$$\begin{aligned}
\mathbb{E}\big[f(s_h,a_h) \;\big|\; (s_1,a_1) \sim \xi_1, \mathcal{P}, \pi\big] =& \mathbb{E}\big[\phi(s_h,a_h)^\top w_f \;\big|\; (s_1,a_1) \sim \xi_1, \mathcal{P}, \pi\big] \\
=& \mathbb{E}\big[\phi(s_h,a_h)^\top \Sigma^{-1} \mathbb{E}_{(s,a)\sim\bar{\mu}}[\phi(s,a)\phi(s,a)^\top] w_f \;\big|\; (s_1,a_1) \sim \xi_1, \mathcal{P}, \pi\big] \\
=& \mathbb{E}_{(s,a)\sim\bar{\mu}}\Big[\mathbb{E}\big[\phi(s_h,a_h) \;\big|\; (s_1,a_1) \sim \xi_1, \mathcal{P}, \pi\big]^\top \Sigma^{-1} \phi(s,a)\phi(s,a)^\top w_f\Big] \\
=& \mathbb{E}_{(s,a)\sim\bar{\mu}}\big[w_h(s,a) f(s,a)\big],
\end{aligned}$$

where $\bar{\mu}$ is the distribution of dataset $\mathcal{D}$. Since the mapping $(s,a) \mapsto \mathbb{E}\big[g(s' \mid s,a) V_h^\pi(s') \;\big|\; s,a\big]$ belongs to $\mathbb{H}$, therefore,

$$\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta)\Big|_{\eta=0} = \mathbb{E}_{(s,a)\sim\bar{\mu}}\left[\sum_{h=1}^{H} w_h(s,a) g(s' \mid s,a) V_{h+1}^{\pi}(s')\right].$$

Note that $\mathbb{E}\big[g(s' \mid s,a) \;\big|\; s,a\big] = 0$, therefore,

$$\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta)\Big|_{\eta=0} = \mathbb{E}_{(s,a)\sim\bar{\mu}}\left[\sum_{h=1}^{H} w_h(s,a) g(s' \mid s,a)\Big(V_{h+1}^{\pi}(s') - \mathbb{E}\big[V_{h+1}^{\pi}(s') \;\big|\; s,a\big]\Big)\right].$$

By definition of μ, we have

$$\begin{aligned}
&\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta)\Big|_{\eta=0} \\
=& \frac{1}{H}\sum_{j=1}^{H} \mathbb{E}\left[\sum_{h=1}^{H} w_h(s_j,a_j) g(s_{j+1} \mid s_j,a_j)\Big(V_{h+1}^{\pi}(s_{j+1}) - \mathbb{E}\big[V_{h+1}^{\pi}(s_{j+1}) \;\big|\; s_j,a_j\big]\Big) \;\middle|\; (s_1,a_1) \sim \bar{\xi}_1, \mathcal{P}, \bar{\pi}\right].
\end{aligned}$$

We use the property $\mathbb{E}\big[g(s' \mid s,a) \;\big|\; s,a\big] = 0$ again and derive that

$$\begin{aligned}
&\frac{\partial}{\partial\eta}\psi_{\xi_1}(\mathcal{P}_\eta)\Big|_{\eta=0} \\
=& \frac{1}{H}\sum_{j=1}^{H} \mathbb{E}\left[\sum_{h=1}^{H} w_h(s_j,a_j)\Big(\sum_{l=1}^{H} g(s_{l+1} \mid s_l,a_l)\Big)\Big(V_{h+1}^{\pi}(s_{j+1}) - \mathbb{E}\big[V_{h+1}^{\pi}(s_{j+1}) \;\big|\; s_j,a_j\big]\Big) \;\middle|\; (s_1,a_1) \sim \bar{\xi}_1, \mathcal{P}, \bar{\pi}\right] \\
=& \frac{1}{H}\mathbb{E}\left[g(\boldsymbol{\tau})\sum_{h=1}^{H}\sum_{j=1}^{H} w_h(s_j,a_j)\Big(V_{h+1}^{\pi}(s_{j+1}) - \mathbb{E}\big[V_{h+1}^{\pi}(s_{j+1}) \;\big|\; s_j,a_j\big]\Big) \;\middle|\; (s_1,a_1) \sim \bar{\xi}_1, \mathcal{P}, \bar{\pi}\right].
\end{aligned}$$

We can conclude that

$$\dot{\psi}_{\mathcal{P}}(\boldsymbol{\tau}) := \frac{1}{H}\sum_{h=1}^{H}\sum_{h'=1}^{H} w_{h'}(s_h,a_h)\Big(V_{h'+1}^{\pi}(s_{h+1}) - \mathbb{E}\big[V_{h'+1}^{\pi}(s_{h+1}) \;\big|\; s_h,a_h\big]\Big),$$

is an influence function.

Efficiency bound. For notational convenience, we take shorthands

$$q(s,a,s') := \sum_{h=1}^{H} w_h(s,a)\Big(V_{h+1}^{\pi}(s') - \mathbb{E}\big[V_{h+1}^{\pi}(s') \;\big|\; s,a\big]\Big),$$

and rewrite

$$\dot{\psi}_{\mathcal{P}}(\boldsymbol{\tau}) = \frac{1}{H}\sum_{h=1}^{H} q(s_h, a_h, s_{h+1}).$$

Since $\mathbb{E}\big[q(s,a,s') \;\big|\; s,a\big] = 0$, we find that

$$\mathbb{E}\big[\dot{\psi}^2_{\mathcal{P}}(\boldsymbol{\tau})\big] = \frac{1}{H^2}\mathbb{E}\left[\left(\sum_{h=1}^{H} q(s_h,a_h,s_{h+1})\right)^2 \;\middle|\; \bar{\xi}_1, \mathcal{P}, \bar{\pi}\right] = \frac{1}{H^2}\sum_{h=1}^{H}\mathbb{E}\Big[q^2(s_h,a_h,s_{h+1}) \;\Big|\; \bar{\xi}_1, \mathcal{P}, \bar{\pi}\Big].$$

It follows that

$$\begin{aligned}\mathbb{E}\big[\dot{\psi}^2_{\mathcal{P}}(\boldsymbol{\tau})\big] =& \frac{1}{H^2}\mathbb{E}_{(s,a)\sim\bar{\mu}}\Big[\mathbb{E}\big[q^2(s,a,s') \;\big|\; s,a\big]\Big] = \frac{1}{H^2}\mathbb{E}_{(s,a)\sim\bar{\mu}}\Big[\mathbb{E}\big[q^2(s,a,s') \;\big|\; s,a\big]\Big] \\ =& \frac{1}{H^2}\mathbb{E}_{(s,a)\sim\bar{\mu}}\left[\left(\phi(s,a)^\top\Sigma^{-1}\sum_{h=1}^{H}\Big(V^\pi_{h+1}(s') - \mathbb{E}\big[V^\pi_{h+1}(s') \;\big|\; s,a\big]\Big)\nu^\pi_h\right)^2\right],\end{aligned}$$

which coincides with the asymptotic variance of OPE estimator defined in (4.2).

■

A.5. Proof of Theorem 5.1: Distributional consistency of bootstrapping FQE

In order to simplify the derivation, we assume $\lambda = 0$ and the empirical covariance matrix $\sum_{n=1}^{N}\phi(s_n,a_n)\phi(s_n,a_n)^\top$ is invertible in this section since the effect of λ is asymptotically negligible. For a matrix $A \in \mathbb{R}^{m\times n}$, suppose the vec operator stacks the column of a matrix such that $\mathrm{vec}(A) \in \mathbb{R}^{mn\times 1}$. We use the equivalence form of FQE in Eq. (A.2) such that

$$\widehat{v}_\pi = (\nu^\pi_1)^\top \sum_{h=0}^{H-1}(\widehat{M}_\pi)^h \widehat{R}.$$

$\widehat{M}_\pi$ can be viewed as the solution of the following multivariate linear regression:

$$\phi^\pi(s_{n+1})^\top = \phi(s_n,a_n)^\top M_\pi + \eta_n,$$

where $\eta_n = \phi^\pi(s_{n+1})^\top - \phi(s_n,a_n)^\top M_\pi$. We first derive the asymptotic distribution of $\sqrt{N}\mathrm{vec}(\widehat{M}_\pi - M_\pi)$ that follows:

$$\begin{aligned}\sqrt{N}\mathrm{vec}(\widehat{M}_\pi - M_\pi) &= \mathrm{vec}\Big(\sqrt{N}\widehat{\Sigma}^{-1}\sum_{n=1}^{N}\phi(s_n,a_n)\Big(\phi^\pi(s'_n)^\top - \phi(s_n,a_n)^\top M_\pi\Big)\Big) \\ &= (N\widehat{\Sigma}^{-1}\otimes I_d)\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathrm{vec}\Big(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}\phi(s^k_h,a^k_h)\Big(\phi^\pi(s^{k'}_h)^\top - \phi(s^k_h,a^k_h)^\top M_\pi\Big)\Big),\end{aligned} \tag{A.7}$$

where $\otimes$ is kronecker product. Define $\xi^k_h = \phi^\pi(s^k_{h+1})^\top - \phi(s^k_h,a^k_h)^\top M_\pi$. From the definition of M_π, it is easy to see

$$\mathbb{E}[\phi^\pi(s^k_{h+1})^\top | s^k_h, a^k_h] = \int_s \mathbb{P}(s|s^k_h,a^k_h)\int_a \pi(a|s)\phi(s,a)^\top \mathrm{d}a\mathrm{d}s = \phi(s^k_h,a^k_h)^\top M_\pi.$$

Again with martingale central limit theorem and independence between each episode, we have as $K\to\infty$,

$$\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\mathrm{vec}\Big(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}\phi(s^k_h,a^k_h)\xi^k_h\Big) \xrightarrow{d} N(0,\Delta), \tag{A.8}$$

where $\Delta \in \mathbb{R}^{d^2\times d^2}$ is the covariance matrix defined as: for $j,k \in [d^2]$

$$\Delta_{jk} = \mathbb{E}\Big[\mathrm{vec}\Big(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}\phi(s^k_h,a^k_h)\xi^k_h\Big)_j \mathrm{vec}\Big(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}\phi(s^k_h,a^k_h)\xi^k_h\Big)_k\Big]. \tag{A.9}$$

Next we start to derive the conditional bootstrap asymptotic distribution. For notation simplicity, denote $\phi_{hk} = \phi(s_h^k, a_h^k)$ and $y_{hk} = \phi^\pi(s_{h+1}^k)^\top$. We rewrite the dataset combined with feature map $\phi(\cdot,\cdot)$ such that $\mathcal{D}_k = \{\phi_{hk}, y_{hk}, r_{hk}\}_{h=1}^H$. Recall that we bootstrap $\mathcal{D}$ by episodes such that each episode is sampled with replacement to form the starred data $\mathcal{D}_k^* = \{\phi_{hk}^*, y_{hk}^*, r_{hk}^*\}_{h=1}^H$ for $k \in [K]$. More specifically,

$$\phi_{hk}^* = \sum_{k=1}^K W_k^* \phi_{hk}, y_{hk}^* = \sum_{k=1}^K W_k^* y_{hk}, r_{hk}^* = \sum_{k=1}^K W_k^* r_{hk},$$

where $W^* = (W_1^*, \ldots, W_K^*)$ is the bootstrap weight. For example, W^* could be a multinomial random vector with parameters $(K; K^{-1}, \ldots, K^{-1})$ that forms the standard nonparametric bootstrap. Note that for different $h \in [H]$, they have the same bootstrap weight and given the original samples $\mathcal{D}_1, \ldots, \mathcal{D}_K$, the resampled vectors are independent. Define the corresponding starred quantity $\widehat{M}_\pi^*, \widehat{R}^*$ as

$$\widehat{M}_\pi^* = \widehat{\Sigma}^{*-1} \sum_{k=1}^K \sum_{h=1}^K \phi_{hk}^* y_{hk}^*, \ \widehat{R}^* = \widehat{\Sigma}^{*-1} \sum_{k=1}^K \sum_{h=1}^H r_{hk}^* \phi_{hk}^*,$$

where

$$\widehat{\Sigma}^* = \sum_{k=1}^K \sum_{h=1}^H \phi_{hk}^* \phi_{hk}^{*\top}.$$

We will derive the asymptotic distribution of $\sqrt{N}(\mathrm{vec}(\widehat{M}_\pi^* - \widehat{M}_\pi))$ by using the following decomposition:

$$\begin{aligned}\sqrt{N}\mathrm{vec}(\widehat{M}_\pi^* - \widehat{M}_\pi) &= \sqrt{N}\mathrm{vec}\Big(\widehat{\Sigma}^{*-1} \sum_{k=1}^K \sum_{h=1}^H \phi_{hk}^*(y_{hk}^* - \phi_{hk}^* \widehat{M}_\pi)\Big) \\ &= (N\widehat{\Sigma}^{*-1} \otimes I_d)\mathrm{vec}\Big(\frac{1}{\sqrt{K}} \sum_{k=1}^K \frac{1}{\sqrt{H}} \sum_{h=1}^H \phi_{hk}^*(y_{hk}^* - \phi_{hk}^* \widehat{M}_\pi)\Big).\end{aligned}$$

We denote

$$Z = \frac{1}{\sqrt{K}} \sum_{k=1}^K \frac{1}{\sqrt{H}} \sum_{h=1}^H \phi_{hk}(y_{hk} - \phi_{hk} M_\pi), Z^* = \frac{1}{\sqrt{K}} \sum_{k=1}^K \frac{1}{\sqrt{H}} \sum_{h=1}^H \phi_{hk}^*(y_{hk}^* - \phi_{hk}^* \widehat{M}_\pi).$$

Both Z and Z^* are the sum of independent $d \times d$ random matrices. We prove the bootstrap consistency using the Mallows metric as a central tool. The Mallows metric, relative to the Euclidean norm $\|\cdot\|$, for two probability measures μ, ν in $\mathbb{R}^d$ is defined as

$$\Lambda_l(\mu, \nu) = \inf_{U \sim \mu, V \sim \nu} \mathbb{E}^{1/l}(\|U - V\|^l),$$

where U and V are two random vectors that U has law μ and V has law ν. For random variables U, V, we sometimes write $\Lambda_l(U, V)$ as the Λ_l-distance between the laws of U and V. We refer Bickel & Freedman (1981); Freedman et al. (1981) for more details about the properties of Mallows metric. Suppose the common distribution of original K episodes $\{\mathcal{D}_1, \ldots, \mathcal{D}_K$ is μ and their empirical distribution is μ_K. Both μ and μ_K are probability in $\mathbb{R}^{2Hd+H}$. From Lemma B.1, we know that $\Lambda_4(\mu_K, \mu) \to 0$ a.e. as $K \to \infty$.

- **Step 1.** We prove $\widehat{\Sigma}^*/N$ converges in conditional probability to Σ. From the bootstrap design, $\frac{1}{H}\sum_{h=1}^H \phi_{kh}^* \phi_{kh}^{*\top}$ is independent of $\frac{1}{H}\sum_{h=1}^H \phi_{k'h}^* \phi_{k'h}^{*\top}$ for any $k \neq k'$. According to Lemma B.3, we have

$$\begin{aligned}\Lambda_1\Big(\sum_{k=1}^K \frac{1}{H}\sum_{h=1}^H \phi_{kh}^* \phi_{kh}^{*\top}, \sum_{k=1}^K \frac{1}{H}\sum_{h=1}^H \phi_{kh}\phi_{kh}^\top\Big) &\le \sum_{k=1}^K \Lambda_1\Big(\frac{1}{H}\sum_{h=1}^H \phi_{kh}^* \phi_{kh}^{*\top}, \frac{1}{H}\sum_{h=1}^H \phi_{kh}\phi_{kh}^\top\Big) \\ &= K\Lambda_1\Big(\frac{1}{H}\sum_{h=1}^H \phi_{kh}^* \phi_{kh}^{*\top}, \frac{1}{H}\sum_{h=1}^H \phi_{kh}\phi_{kh}^\top\Big).\end{aligned}$$

Both sides of the above inequality are random variables such that the distance is computed between the conditional distribution of the starred quantity and the unconditional distribution of the unstarred quantity. Define a mapping $f : \mathbb{R}^{Hd} \to \mathbb{R}^{d\times d}$ such that for any $x_1, \ldots, x_H \in \mathbb{R}^d$,

$$f(x_1, \ldots, x_H) = \frac{1}{H}\sum_{h=1}^{H} x_h x_h^\top.$$

From Lemma B.2 with f, we have as K goes to infinity

$$\Lambda_1\Big(\frac{1}{H}\sum_{h=1}^{H}\phi_{kh}^*\phi_{kh}^{*\top}, \frac{1}{H}\sum_{h=1}^{H}\phi_{kh}\phi_{kh}^\top\Big) \to 0\,.$$

This implies the conditional law of $\frac{1}{H}\sum_{h=1}^{H}\phi_{kh}^*\phi_{kh}^{*\top}$ is close to the unconditional law of $\frac{1}{H}\sum_{h=1}^{H}\phi_{kh}\phi_{kh}^\top$. By the law of large numbers:

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H}\sum_{h=1}^{H}\phi_{kh}\phi_{kh}^\top \xrightarrow{p} \Sigma. \tag{A.10}$$

This further implies the conditional on $\mathcal{D}$, we have $\widehat{\Sigma}^*/N \xrightarrow{p} \Sigma$.

- **Step 2.** We prove Z^* conditionally converges to a multivariate Gaussian distribution. From Lemma B.4,

$$\Lambda_2(\mathrm{vec}(Z^*), \mathrm{vec}(Z))^2 \le \Lambda_2\Big(\mathrm{vec}(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}\phi_{hk}^*(y_{hk}^* - \phi_{hk}^*\widehat{M}_\pi)), \mathrm{vec}(\frac{1}{\sqrt{H}}\sum_{h=1}^{H}\phi_{hk}(y_{hk} - \phi_{hk}M_\pi))\Big)^2.$$

Using Lemma B.5, we have the right side converges to 0, a.e. as $K \to \infty$. This means the conditional law of $\mathrm{vec}(Z^*)$ is close to the unconditional law of $\mathrm{vec}(Z)$, and the latter essentially converges to a multivariate Gaussian distribution with zero mean and covariance matrix Δ from Eq. (A.8).

By Slutsky's theorem, we have conditional on $\mathcal{D}$,

$$\sqrt{N}\mathrm{vec}(\widehat{M}_\pi^* - \widehat{M}_\pi) \xrightarrow{d} N\Big(0, (\Sigma^{-1}\otimes I_d)\Delta(\Sigma^{-1}\otimes I_d)\Big), \tag{A.11}$$

where Δ is defined in Eq. (A.9).

According to the equivalence between FQE and plug-in estimator in Section A.2,

$$\widehat{v}_\pi^* = (\nu_1^\pi)^\top\sum_{h=0}^{H-1}(\widehat{M}_\pi^*)^h\widehat{R}^*, \widehat{v}_\pi = (\nu_1^\pi)^\top\sum_{h=0}^{H-1}(\widehat{M}_\pi)^h\widehat{R}.$$

Define a function $g : \mathbb{R}^{d\times d} \to \mathbb{R}$ as

$$g(M) := (\nu_1^\pi)^\top\sum_{h=0}^{H-1}(M)^h w_r.$$

By the high-order matrix derivative (Petersen & Pedersen, 2008), we have

$$\frac{\partial}{\partial M}(\nu_1^\pi)^\top(M)^h w_r = \sum_{r=1}^{h-1}(M^r)^\top\nu_1^\pi w_r^\top(M^{h-1-r})^\top \in \mathbb{R}^{d\times d}.$$

This implies the gradient of g at $\mathrm{vec}(M_\pi)$

$$\begin{aligned}\nabla g(\mathrm{vec}(M_\pi)) &= \mathrm{vec}\Big(\sum_{h=0}^{H-1}\sum_{r=1}^{h-1}(M_\pi^r)^\top\nu_1^\pi w_r^\top(M_\pi^{h-1-r})^\top\Big)\\ &= \mathrm{vec}\Big(\sum_{h=1}^{H}\nu_h^\pi w_r^\top\sum_{h'=1}^{H-h}(M_\pi^{h'-1})^\top\Big) \in \mathbb{R}^{d^2\times 1}.\end{aligned}$$

Applying multivariate delta theorem (Theorem B.6) for Eq. (A.11), we have conditional on $\mathcal{D}$

$$\sqrt{N}\Big(g(\widehat{M}_\pi^*) - g(\widehat{M}_\pi)\Big) \xrightarrow{d} \mathcal{N}\Big(0, \nabla^\top g(\mathrm{vec}(\widehat{M}_\pi))(\Sigma^{-1}\otimes I_d)\Delta(\Sigma^{-1}\otimes I_d)\nabla g(\mathrm{vec}(\widehat{M}_\pi))\Big),$$

where Δ is defined in Eq. (A.9). From Eq. (A.10), we have $\widehat{\Sigma}/N \xrightarrow{p} \Sigma$. Using Slutsky's theorem and Eqs. (A.7)-(A.8), we have

$$\sqrt{N}\Big(\mathrm{vec}(\widehat{M}_\pi - M_\pi)\Big) \xrightarrow{d} \mathcal{N}(0, (\Sigma^{-1}\otimes I_d)\Delta(\Sigma^{-1}\otimes I_d)).$$

This further implies $\widehat{M}_\pi \xrightarrow{p} M_\pi$. By continuous mapping theorem,

$$\sqrt{N}\Big(g(\widehat{M}_\pi^*) - g(\widehat{M}_\pi)\Big) \xrightarrow{d} \mathcal{N}\Big(0, \nabla^\top g(\mathrm{vec}(M_\pi))(\Sigma^{-1}\otimes I_d)\Delta(\Sigma^{-1}\otimes I_d)\nabla g(\mathrm{vec}(M_\pi))\Big).$$

Now we simplify the variance term as follows:

$$\begin{aligned}
&\nabla^\top g(\mathrm{vec}(M_\pi))(\Sigma^{-1}\otimes I_d)\Delta(\Sigma^{-1}\otimes I_d)\nabla g(\mathrm{vec}(M_\pi))\\
=&\sum_{h=1}^{H}(\nu_h^\pi)^\top\Sigma^{-1}w_r^\top\sum_{h'=1}^{H-h}(M_\pi)^{h'-1}\mathbb{E}\Big[\frac{1}{H}\sum_{h=1}^{H}\xi_h^\top\phi(s_h^1,a_h^1)\phi(s_h^1,a_h^1)^\top\xi_h\Big]\sum_{h=1}^{H}\sum_{h'=1}^{H-h}(M_\pi)^{h'-1}w_r\Sigma^{-1}(\nu_h^\pi)^\top\\
=&\sum_{h=1}^{H}(\nu_h^\pi)^\top\Sigma^{-1}w_r^\top\sum_{h'=1}^{H-h}(M_\pi)^{h'-1}\mathbb{E}\Big[\frac{1}{H}\sum_{h=1}^{H}\xi_h^\top\phi(s_h^1,a_h^1)\phi(s_h^1,a_h^1)^\top\xi_h\Big]\sum_{h'=1}^{H-h}(M_\pi)^{h'-1}w_r\Sigma^{-1}(\nu_h^\pi)^\top\\
&+2\sum_{h_1<h_2}(\nu_{h_1}^\top)^\top\Sigma^{-1}w_r^\top\sum_{h'=1}^{H-h_1}(M_\pi)^{h'-1}\mathbb{E}\Big[\frac{1}{H}\sum_{h=1}^{H}\xi_h^\top\phi(s_h^1,a_h^1)\phi(s_h^1,a_h^1)^\top\xi_h\Big]\sum_{h'=1}^{H-h_2}(M_\pi)^{h'-1}w_r\Sigma^{-1}(\nu_{h_2}^\top)^\top\\
=&\sum_{h=1}^{H}(\nu_h^\pi)^\top\Sigma^{-1}\mathbb{E}\Big[\frac{1}{H}\sum_{h=1}^{H}w_r^\top\sum_{h'=1}^{H-h}(M_\pi)^{h'-1}\xi_h^\top\phi(s_h^1,a_h^1)\phi(s_h^1,a_h^1)^\top\xi_h\sum_{h'=1}^{H-h}(M_\pi)^{h'-1}w_r\Big]\Sigma^{-1}(\nu_h^\pi)^\top\\
&+2\sum_{h_1<h_2}(\nu_{h_1}^\top)^\top\Sigma^{-1}\mathbb{E}\Big[\frac{1}{H}\sum_{h=1}^{H}w_r^\top\sum_{h'=1}^{H-h_1}(M_\pi)^{h'-1}\xi_h^\top\phi(s_h^1,a_h^1)\phi(s_h^1,a_h^1)^\top\xi_h\sum_{h'=1}^{H-h_2}(M_\pi)^{h'-1}w_r\Big]\Sigma^{-1}(\nu_{h_2}^\top)^\top,
\end{aligned}$$

where $\xi_h^1 = \phi(s_h^1,a_h^1)^\top M_\pi - \phi^\pi(s_{h+1}^1)^\top$. Recall that we define

$$\begin{aligned}
\varepsilon_{h,h'}^1 &= Q_h^\pi(s_{h'}^1,a_{h'}^1) - (r_{h'}^1 + V_{h+1}^\pi(s_{h+1}^1))\\
&= \sum_{h=1}^{H-h_1}\Big(\phi(s_h^1,a_h^1)^\top M_\pi - \phi^\pi(s_{h+1}^1)^\top\Big)(M_\pi)^{h'-1}w_r = \sum_{h=1}^{H-h}\xi_h^1(M_\pi)^{h-1}w_r,
\end{aligned}$$

where the second equation is from Eq. (A.3). This implies

$$\begin{aligned}
&\nabla^\top g(\mathrm{vec}(M_\pi))(\Sigma^{-1}\otimes I_d)\Delta(\Sigma^{-1}\otimes I_d)\nabla g(\mathrm{vec}(M_\pi))\\
=&\sum_{h=1}^{H}(\nu_h^\pi)^\top\Sigma^{-1}\mathbb{E}\Big[\frac{1}{H}\sum_{h'=1}^{H}\phi(s_{h'}^1,a_{h'}^1)\phi(s_{h'}^1,a_{h'}^1)^\top(\varepsilon_{h,h'}^1)^2\Big]\Sigma^{-1}(\nu_h^\pi)^\top\\
&+2\sum_{h_1<h_2}(\nu_{h_1}^\top)^\top\Sigma^{-1}\mathbb{E}\Big[\frac{1}{H}\sum_{h'=1}^{H}\phi(s_{h'}^1,a_{h'}^1)\phi(s_{h'}^1,a_{h'}^1)^\top\varepsilon_{h_1,h'}^1\varepsilon_{h_2,h'}^1\Big]\Sigma^{-1}(\nu_{h_2}^\top)^\top = \sigma^2.
\end{aligned}$$

Therefore, we have proven that

$$\sqrt{N}\Big(g(\widehat{M}_\pi^*) - g(\widehat{M}_\pi)\Big) \xrightarrow{d} \mathcal{N}\big(0,\sigma^2\big).$$

On the other hand,

$$\widehat{R}^* = (\widehat{\Sigma}^*)^{-1}\sum_{k=1}^{K}\sum_{h=1}^{H}r_{hk}^*\phi_{hk}^* = KH(\widehat{\Sigma}^*)^{-1}\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H}\sum_{h=1}^{H}r_{hk}^*\phi_{hk}^*.$$

Using Lemma B.3, we have

$$\Lambda_1\Big(\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H}\sum_{h=1}^{H} r_{hk}^*\phi_{hk}^*, \frac{1}{K}\sum_{k=1}^{K}\frac{1}{H}\sum_{h=1}^{H} r_{hk}\phi_{hk}\Big) \le \Lambda_1\Big(\frac{1}{H}\sum_{h=1}^{H} r_{hk}^*\phi_{hk}^*, \frac{1}{H}\sum_{h=1}^{H} r_{hk}\phi_{hk}\Big).$$

The right hand side of the display goes to 0 as $K \to \infty$. From the law of large number,

$$\frac{1}{K}\sum_{k=1}^{K}\frac{1}{H}\sum_{h=1}^{H} r_{hk}\phi_{hk} \xrightarrow{p} \mathbb{E}\Big[\frac{1}{H}\sum_{h=1}^{H}\phi(s_h^1,a_h^1)\phi(s_h^1,a_h^1)^\top\Big] w_r$$

Combining with the fact that the conditional laws of $\widehat{\Sigma}^*$ concentrates around Σ, this ends the proof. ■

A.6. Proofs of Corollary 5.2 and Corollary 5.4

We prove the consistency of bootstrap confidence interval by using Lemma 23.3 in Van der Vaart (2000). Suppose $\Psi(t) = \mathbb{P}(\mathcal{N}(0,\sigma^2) \le t)$. Combining Theorem 4.2 and Theorem 5.1, we have

$$\mathbb{P}_{\mathcal{D}}\Big(\sqrt{N}(\widehat{v}_\pi - v_\pi) \le t\Big) \to \Psi(t),\ \mathbb{P}_{W^*|\mathcal{D}}\Big(\sqrt{N}(\widehat{v}_\pi^* - \widehat{v}_\pi) \le t\Big) \to \Psi(t).$$

Using the quantile convergence theorem (Lemma 21.1 in Van der Vaart (2000)), it implies $q_\delta^\pi \to \Psi^{-1}(t)$ almost surely. Therefore,

$$\begin{aligned}\mathbb{P}_{\mathcal{D}W^*}\Big(v_\pi \le \widehat{v}_\pi - q_{\delta/2}^\pi\Big) &= \mathbb{P}_{\mathcal{D}W^*}\Big(\sqrt{N}(\widehat{v}_\pi - v_\pi) \ge q_{\delta/2}^\pi\Big)\\ &\to \mathbb{P}_{\mathcal{D}W^*}\Big(\mathcal{N}(0,\sigma^2) \ge \Psi^{-1}(\delta/2)\Big) = 1-\delta/2.\end{aligned}$$

This finishes the proof of Corollary 5.2.

It is well known that the convergence in distribution implies the convergence in moment under the uniform integrability condition. The proof of the consistency of bootstrap moment estimation is straightforward since the condition $\limsup_{N\to\infty}\mathbb{E}_{W^*|\mathcal{D}}[(\sqrt{N}(\widehat{v}_\pi^* - \widehat{v}_\pi))^q] < \infty$ for some $q > 2$ ensures a similar uniform integrability condition. Together with the distributional consistency in Theorem 5.1, we apply Lemma 2.1 in Kato (2011) then we reach the conclusion. ■

B. Supporting Results

We present a series of useful lemmas about Mallows metric.

Lemma B.1 (Lemma 8.4 in Bickel & Freedman (1981))**.** Let $\{X_i\}_{i=1}^n$ be independent random variables with common distribution μ. Let μ_n be the empirical distribution of $X_1,\dots,X_n$. Then $\Lambda_l(\mu_n,\mu) \to 0$ a.e..

Lemma B.2 (Lemma 8.5 in Bickel & Freedman (1981))**.** Suppose X_n, X are random variables and $\Lambda_l(X_n,X) \to 0$. Let f be a continuous function. Then $\Lambda_l(f(X_n),f(X)) \to 0$.

Lemma B.3 (Lemma 8.6 of Bickel & Freedman (1981))**.** Let $\{U_i\}_{i=1}^n, \{V_i\}_{i=1}^n$ be independent random vectors. Then we have

$$\Lambda_1\Big(\sum_{i=1}^n U_i, \sum_{i=1}^n V_i\Big) \le \sum_{i=1}^n \Lambda_1\Big(U_i, V_i\Big).$$

Lemma B.4 (Lemma 8.7 of Bickel & Freedman (1981))**.** Let $\{U_i\}_{i=1}^n, \{V_i\}_{i=1}^n$ be independent random vectors and $\mathbb{E}[U_j] = \mathbb{E}[V_j]$. Then we have

$$\Lambda_2\Big(\sum_{i=1}^n U_i, \sum_{i=1}^n V_i\Big)^2 \le \sum_{i=1}^n \Lambda_2\Big(U_i, V_i\Big)^2.$$

Let μ_K and μ be probabilities on $\mathbb{R}^{2Hd}$. A data point in $\mathbb{R}^{2Hd}$ can be written as $(x_1, \ldots, x_H, y_1, \ldots, y_H)$ where $x_h \in \mathbb{R}^d$ and $y_h \in \mathbb{R}^d$. Denote

$$\Sigma(\mu) = \int \frac{1}{H} \sum_{h=1}^{H} x_h x_h^\top \mu(dx_1, \ldots, dx_H, dy_1, \ldots, dy_H),$$

$$M(\mu) = \Sigma(\mu)^{-1} \int \sum_{h=1}^{H} x_h y_h^\top \mu(dx_1, \ldots, dx_H, dy_1, \ldots, dy_H),$$

$$\varepsilon(\mu, x_1, \ldots, x_H, y_1, \ldots, y_H) = \sum_{h=1}^{H} (y_h - x_h^\top M(\mu)).$$

Lemma B.5 (Lemma 7 in Eck (2018))**.** If $\Lambda_4(\mu_K, \mu) \to 0$ as $K \to \infty$, then we have the μ_K-law of $\text{vec}(\sum_{h=1}^{H} \varepsilon(\mu_K, x_1, \ldots, x_H, y_1, \ldots, y_H) x_h^\top)$ converges to the μ-law of $\text{vec}(\sum_{h=1}^{H} \varepsilon(\mu, x_1, \ldots, x_H, y_1, \ldots, y_H) x_h^\top)$ in Λ_2.

Theorem B.6 (Multivariate delta theorem)**.** Suppose $\{T_n\}$ is a sequence of k-dimensional random vectors such that $\sqrt{n}(T_n - \theta) \xrightarrow{d} N(0, \Sigma(\theta))$. Let $g : \mathbb{R}^k \to \mathbb{R}$ be once differentiable at θ with the gradient matrix $\nabla g(\theta)$. Then

$$\sqrt{n}(g(T_n) - g(\theta)) \xrightarrow{d} N(0, \nabla^\top g(\theta) \Sigma(\theta) \nabla g(\theta)).$$

We restate Lemma B.5 in Duan & Wang (2020) in the following that is proven using matrix Bernstein inequality.

Lemma B.7. Under the assumption $\phi(s,a)^\top \Sigma^{-1} \phi(s,a) \le C_1 d$ for all $(s,a) \in \mathcal{X}$, with probability at least $1 - \delta$,

$$\left\| \Sigma^{-1/2} \left(\frac{1}{N} \sum_{n=1}^{N} \phi(s_n, a_n) \phi(s_n, a_n)^\top \right) \Sigma^{-1/2} - I \right\|_2 \le \sqrt{\frac{2\ln(2d/\delta) C_1 dH}{N}} + \frac{2\ln(2d/\delta) C_1 dH}{3N}. \tag{B.1}$$

C. Supplement for Experiments

C.1. Experiment details

The original CliffWalking environment from OpenAI gym has deterministic state transitions. That is, for any state-action pair (s, a), there exists a corresponding $s' \in \mathcal{S}$ such that $\mathbb{P}(\cdot \mid s, a) = \delta_{s'}(\cdot)$. We modify the environment in order to make it stochastic. Specifically, we introduce randomness in state transitions such that given a state-action pair (s, a), the transition takes place in the same way as in the deterministic environment with probability $1 - \epsilon$ and takes place as if the action were a random action a', instead of the intended a, with probability ϵ. This is an episodic tabular MDP and the agent stops when falling from the cliff or reaching the terminal point. We also reduce the penalty of falling off the cliff from -100 to -50.

The original MountainCar environment from OpenAI gym has deterministic state transitions. We modify the environment in order to make it stochastic. Specifically, we introduce randomness in state transitions by adding a Gaussian random force, namely, $\mathcal{N}(0, \frac{1}{10})$ multiplied by the constant-magnitude force from the original environment. We also increase the gravity parameter from 0.0025 to 0.008, the force parameter from 0.001 to 0.008 and the maximum allowed speed from 0.07 to 0.2.

Empirical coverage probability. The preceding discussion leads to the simulation method for estimating the coverage probability of a confidence interval. The simulation method has three steps:

1. Simulate many fresh dataset of episode size K following the behavior policy.
2. Compute the confidence interval for each dataset.
3. Compute the proportion of dataset for which the true value of target policy is contained in the confidence interval. That proportion is an estimate for the empirical coverage probability for the confidence interval.

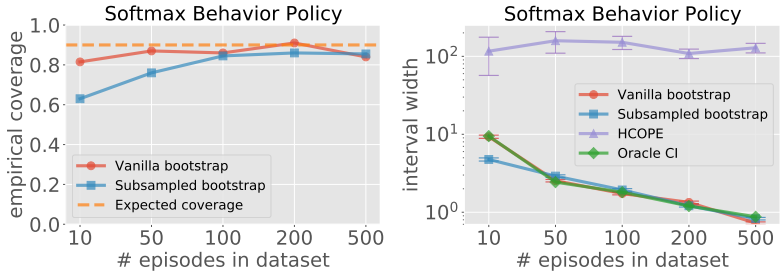

Figure 7. Left: Empirical coverage probability of CI; Right: CI width under different behavior policies.

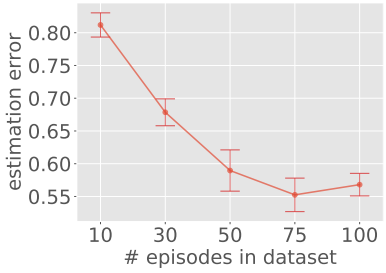

Figure 8. Error of correlation estimates, as data size increases.

The true value of target policy is computed through Monte Carlo rollouts with sufficient number of samples (10000 in our experiments).

With linear function approximation, we use the confidence interval proposed in Section 6 in Duan & Wang (2020) as a baseline since it is only available confidence interval based on FQE. In particular, it shows that with probability at least $1-\delta$,

$$|\widehat{v}^{\pi} - v^{\pi}| \leq \sum_{h=1}^{H} (H-h+1)\sqrt{(\widehat{\nu}_h^{\pi})^{\top}\widehat{\Sigma}^{-1}\widehat{\nu}_h^{\pi}}\left(\sqrt{2\lambda} + 2\sqrt{2d\log\left(1+\frac{N}{\lambda d}\right)\log\left(\frac{3N^2H}{\delta}\right)} + \frac{4}{3}\log\left(\frac{3N^2H}{\delta}\right)\right),$$

where $(\widehat{\nu}_h^{\pi})^{\top} = (\nu_1^{\pi})^{\top}(\widehat{M}^{\pi})^h$ and $\widehat{M}_{\pi}$ is defined in Eq. (A.1).

C.2. Additional experiments

In Figure 7, we include the result for soft-max behavior policy in the Cliff Walking environment. In Figure 8, we include the result for correlation estimation in Cliff Walking environment. The behavior policy is 0.1 ϵ-greedy policy while two target policies are optimal policy and 0.1 ϵ-greedy policy.

In order to better understand the tradeoff between computational efficiency and accuracy with finite samples, we conduct some empirical demonstrations based on Cliffwalking. We set $s = K^{\gamma}$ and the true coverage probability is 0.9. It is relatively safe to set $\gamma > 0.5$. Note that $\gamma = 1$ corresponds to the vanilla bootstrap that has the highest accuracy but heaviest computation.

We argue that bootstrapping sample transitions (which are dependent) would lead to inconsistent estimations of the error distribution and thus output wrong confidence interval and variance estimation. We further run one additional test using the taxi environment and further compute the CI and variance estimations based on different bootstrap distributions in Cliffwalking (CW) and taxi environments. This is already in an asymptotic regime since both the number of episodes and the number of bootstrap samples are 10e+6. It is clear that bootstrapping by sample transition gives an incorrect distribution, thus it is inconsistent.

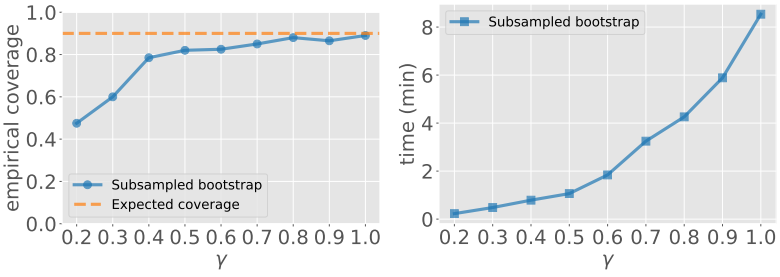

Figure 9. Subsampled bootstrap with $s = K^\gamma$

	True distribution	By episodes	By sample transition
CI (CW)	(-20.44, -19.74)	(-20.40, -19.74)	(-20.52, -19.58)
Variance (CW)	0.45	0.44	0.082
CI (Taxi)	(2.49, 3.82)	(2.45, 3.73)	(2.01, 4.09)
Variance (Taxi)	0.16	0.16	0.39

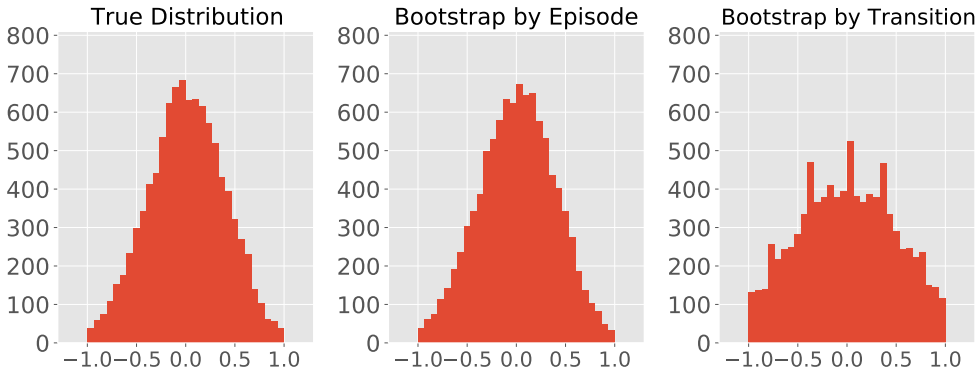

Figure 10. Taxi environment (Dietterich 2000)