Supplement to "Deeply-Debiased Off-Policy Interval Estimation"

Chengchun Shi [*1] **Runzhe Wan** [*2] **Victor Chernozhukov** [3] **Rui Song** [2]

A. Technical Definitions and Proofs

A.1. Third-Order Q-Estimator

We detail the form of $\widehat{Q}_k^{(3)}$. According to the definition, we have

$$\widehat{Q}_k^{(3)} = \frac{1}{|\mathbb{I}_k|T(|\mathbb{I}_k|T-1)} \sum_{\substack{i_1\in\mathbb{I}_k, 0\le t_1<T \\ i_2\in\mathbb{I}_k, 0\le t_2<T \\ (i_1,t_1)\neq(i_2,t_2)}} \mathcal{D}_k^{(i_1,t_1)}\mathcal{D}_k^{(i_2,t_2)}\widehat{Q}_k.$$

For any state-action pair (a,s), it follows that

$$\widehat{Q}_k^{(3)}(a,s) = \frac{(1-\gamma)^{-1}}{|\mathbb{I}_k|T(|\mathbb{I}_k|T-1)} \sum_{\substack{i_1\in\mathbb{I}_k, 0\le t_1<T \\ i_2\in\mathbb{I}_k, 0\le t_2<T \\ (i_1,t_1)\neq(i_2,t_2)}} \widehat{\tau}_k(A_{i_1,t_1}, S_{i_1,t_1}, a, s)\{R_{i_1,t_1} + \gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i_1,t_1+1})}\mathcal{D}_k^{(i_2,t_2)}\widehat{Q}_k(a', S_{i_1,t_1+1})$$
$$-\mathcal{D}_k^{(i_2,t_2)}\widehat{Q}_k(A_{i_1,t_1}, S_{i_1,t_1})\} + \frac{1}{|\mathbb{I}_k|T}\sum_{i_2\in\mathbb{I}_k, 0\le t_2<T}\mathcal{D}_k^{(i_2,t_2)}\widehat{Q}_k(a,s).$$

The right-hand-side is equal to

$$\widehat{Q}_k(a,s) + \frac{(1-\gamma)^{-1}}{|\mathbb{I}_k|T}\sum_{i\in\mathbb{I}_k, 0\le t<T}\widehat{\tau}_k(A_{i,t}, S_{i,t}, a, s)\{R_{i,t} + \gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}_k(a', S_{i,t+1}) - \widehat{Q}_k(A_{i,t}, S_{i,t})\}$$
$$+\frac{(1-\gamma)^{-2}}{|\mathbb{I}_k|T(|\mathbb{I}_k|T-1)} \sum_{\substack{i_1\in\mathbb{I}_k, 0\le t_1<T \\ i_2\in\mathbb{I}_k, 0\le t_2<T \\ (i_1,t_1)\neq(i_2,t_2)}} \widehat{\tau}_k(A_{i_1,t_1}, S_{i_1,t_1}, a, s)\{\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\widehat{\tau}_k(A_{i_2,t_2}, S_{i_2,t_2}, a', S_{i_1,t_1+1})$$
$$-\widehat{\tau}_k(A_{i_2,t_2}, S_{i_2,t_2}, A_{i_1,t_1}, S_{i_1,t_1}) + (1-\gamma)\widehat{\tau}_k(A_{i_2,t_2}, S_{i_2,t_2}, a, s)\}\{R_{i_2,t_2} + \gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i_2,t_2+1})}\widehat{Q}_k(a', S_{i_2,t_2+1}) - \widehat{Q}_k(A_{i_2,t_2}, S_{i_2,t_2})\}.$$

A.2. Definition of the L_2-norm Convergence

A sequence of variables $\{X_n\}_{n\ge0}$ is said to converge in L_2-norm to X if and only if $\mathbb{E}|X_n - X|^2 \to 0$ as $n\to\infty$.

A Q-estimator $\widehat{Q}$ is said to converge in L_2-norm to Q^π at a rate of $(nT)^{-\alpha}$ if

$$\sqrt{\mathbb{E}_{(a,s)\sim p_\infty}\mathbb{E}|\widehat{Q}(a,s) - Q^\pi(a,s)|^2} = O\{(nT)^{-\alpha}\}.$$

Similarly, a conditional density ratio estimator $\widehat{\tau}$ is said to converge in L_2-norm to τ^π at a rate of $(nT)^{-\alpha}$ if

$$\sqrt{\mathbb{E}_{(a,s)\sim p_\infty}\mathbb{E}_{(a^*,s^*)\sim p_\infty}\mathbb{E}|\widehat{\tau}(a,s,a^*,s^*) - \tau^\pi(a,s,a^*,s^*)|^2} = O\{(nT)^{-\alpha}\}.$$

Finally, a marginalized density ratio estimator $\widehat{\omega}$ is said to converge in L_2-norm to ω^π at a rate of $(nT)^{-\alpha}$ if

$$\sqrt{\mathbb{E}_{(a,s)\sim p_\infty}\mathbb{E}|\widehat{\omega}(a,s) - \omega^\pi(a,s)|^2} = O\{(nT)^{-\alpha}\}.$$

A.3. Proof of Lemma 3

To simplify the presentation, in the proof we assume the data consist of independent tuples in Lemma 1. With weakly dependent data, the aggregated bias will be upper bounded by the same order of magnitude (see the proof of Theorem 1 for details).

We first study the bias of the Q-estimator. We will prove a slightly stronger result, showing that

$$\mathbb{E}_{(a,s)\sim p_\infty}|\mathbb{E}\widehat{Q}_k^{(m)}(a,s)-Q^\pi(a,s)|^2=O\{(nT)^{-2\alpha_1-2(m-1)\alpha_2}\}. \tag{1}$$

We prove this assertion by induction. Consider the case where $m=2$. By the doubly-robustness property, we have $Q^\pi(a,s)=\mathbb{E}[\widehat{Q}_k(a,s)+\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)\{R_{i,t}+\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}_k(a',S_{i,t+1})-\widehat{Q}_k(A_{i,t},S_{i,t})\}]$. It follows that

$$\begin{aligned}\mathbb{E}\widehat{Q}_k^{(2)}(a,s)-Q^\pi(a,s)=\mathbb{E}\mathcal{D}_k^{(i,t)}\widehat{Q}_k(a,s)-Q^\pi(a,s)=\mathbb{E}\{\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-\tau^\pi(A_{i,t},S_{i,t},a,s)\}\\ \times\{Q^\pi(A_{i,t},S_{i,t})-\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}Q^\pi(a',S_{i,t+1})+\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\widehat{Q}_k(a',S_{i,t+1})-\widehat{Q}_k(A_{i,t},S_{i,t})\}.\end{aligned} \tag{2}$$

By Cauchy-Schwarz inequality, $\mathbb{E}_{(a,s)\sim p_\infty}|\mathbb{E}\widehat{Q}^{(2)}(a,s)-Q^\pi(a,s)|^2$ is upper bounded by

$$\begin{aligned}\mathbb{E}_{(a,s)\sim p_\infty}\mathbb{E}|\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-\tau^\pi(A_{i,t},S_{i,t},a,s)|^2\Big\{2\mathbb{E}|\widehat{Q}_k(A_{i,t},S_{i,t})-Q^\pi(A_{i,t},S_{i,t})|^2\\ +2\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\mathbb{E}|\widehat{Q}_k(a,S_{i,t+1})-Q^\pi(a,S_{i,t+1})|^2\Big\}.\end{aligned}$$

Under the convergence rate requirement, it is upper bounded by $\{(nT)^{-\alpha_1-\alpha_2}\}$. This proves the assertion with $m=2$.

Suppose the assertion holds with $m=m_0\geq 2$. We aim to show it holds with $m=m_0+1$. Similar to (2), since the data tuples are i.i.d., we have

$$\begin{aligned}\mathbb{E}\widehat{Q}_k^{(m_0+1)}(a,s)-Q^\pi(a,s)=\mathbb{E}\mathcal{D}_k^{(i,t)}\mathbb{E}\widehat{Q}_k^{(m_0)}(a,s)-Q^\pi(a,s)=\mathbb{E}\{\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-\tau^\pi(A_{i,t},S_{i,t},a,s)\}\times\\ [Q^\pi(A_{i,t},S_{i,t})-\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}Q^\pi(a',S_{i,t+1})+\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}\mathbb{E}\{\widehat{Q}_k^{(m_0)}(a',S_{i,t+1})|S_{i,t+1}\}-\mathbb{E}\{\widehat{Q}_k^{(m_0)}(A_{i,t},S_{i,t})|A_{i,t},S_{i,t}\}].\end{aligned}$$

By Cauchy-Schwarz inequality, $\mathbb{E}_{(a,s)\sim p_\infty}|\mathbb{E}\widehat{Q}^{(m_0+1)}(a,s)-Q^\pi(a,s)|^2$ is upper bounded by

$$\begin{aligned}\mathbb{E}_{(a,s)\sim p_\infty}\mathbb{E}|\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-\tau^\pi(A_{i,t},S_{i,t},a,s)|^2\Big[2\mathbb{E}|\mathbb{E}\{\widehat{Q}_k^{(m_0)}(A_{i,t},S_{i,t})|A_{i,t},S_{i,t}\}-Q^\pi(A_{i,t},S_{i,t})|^2\\ +2\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\mathbb{E}|\mathbb{E}\{\widehat{Q}_k^{(m_0)}(a,S_{i,t+1})|S_{i,t+1}\}-Q^\pi(a,S_{i,t+1})|^2\Big].\end{aligned} \tag{3}$$

The above bound is of the order $O\{(nT)^{-2\alpha_1+2m_0\alpha_2}\}$. The assertion is thus proven.

We next consider the bias of the resulting value. Since $\eta_{\text{TR}}^{(m)}$ is a simple average of $\{\psi_{i,t}^{(m)}\}_{i,t}$, it suffices to provide an upper bound for $\psi_{i,t}^{(m)}$ for a given tuple $(i,t)\in\mathbb{I}_k$. We decompose $\widehat{Q}_k^{(m)}$ into the sum of the following two parts:

$$\begin{aligned}&\binom{|\mathbb{I}_k|T}{(m-1)}^{-1}\sum_{(i_l,t_l)=(i,t)\text{ for some }l}\mathcal{D}_k^{(i_1,t_1)}\cdots\mathcal{D}_k^{(i_{m-1},t_{m-1})}\widehat{Q}_k\\ +&\binom{|\mathbb{I}_k|T}{(m-1)}^{-1}\sum_{(i_l,t_l)\neq(i,t)\text{ for any }l}\mathcal{D}_k^{(i_1,t_1)}\cdots\mathcal{D}_k^{(i_{m-1},t_{m-1})}\widehat{Q}_k.\end{aligned}$$

Since the functions $\widehat{Q}_k,\widehat{\tau}_k$ and the immediate rewards are uniformly bounded, the first term is upper bounded by

$$c(m-1)\binom{|\mathbb{I}_k|T}{(m-1)}^{-1}\binom{|\mathbb{I}_k|T-1}{(m-2)}=\frac{c(m-1)^2}{|\mathbb{I}_k|T}=O(n^{-1}T^{-1}),$$

where c denotes some positive constant. Similarly, we can show the second term can be well-approximated by

$$\widehat{Q}_{k,i,t}^{(m)}=(m-1)\binom{|\mathbb{I}_k|T-1}{(m-2)}^{-1}\sum_{(i_l,t_l)\neq(i,t)\text{ for any }l}\mathcal{D}_k^{(i_1,t_1)}\cdots\mathcal{D}_k^{(i_{m-1},t_{m-1})}\widehat{Q}_k,$$

with the approximation error upper bounded by $O(n^{-1}T^{-1})$.

Since $\psi_{i,t}^{(m)}$ is a linear function $\widehat{Q}^{(m)}$, we have $\max_{i,t}|\psi_{i,t}^{(m)}-\phi_{i,t}^{(m)}|=O(n^{-1}T^{-1})$ where $\phi_{i,t}^{(m)}$ is a version of $\psi_{i,t}^{(m)}$ with $\widehat{Q}^{(m)}$ replaced with $\widehat{Q}_{i,t}^{(m)}$. It suffices to show the bias $\max_{i,t}|\mathbb{E}\phi_{i,t}^{(m)}-\eta^{\pi}|$ converges at a rate of $(nT)^{-\alpha_1-(m-1)\alpha_2-\alpha_3}$. Since the tuples of indices $(i,t),(i_1,t_1),\cdots,(i_m,t_m)$ are different, the corresponding data observations are independent. This assertion can be proven in a similar manner as (1).

A.4. Proof of Theorem 1

For any k, let r_1,r_2,r_3 denote the rate of convergence of $\widehat{Q}_k$, $\widehat{\tau}_k$ and $\widehat{\omega}_k$, respectively. These rates of convergence will approach zero when the corresponding nuisance estimators are consistent.

In Part 1, we prove a version Lemma 3 holds under the exponential β-mixing condition in (A1) as well. Specifically, the aggregated bias of the Q-estimator decays at a rate of $O(r_1r_2^{(m-1)})$, and the bias of the corresponding value estimator decays at a rate of $O(r_1r_2^{(m-1)}r_3)$. When one of the three estimated nuisance functions is consistent, the bias decays to zero.

In Part 2, we show the variance of the value estimator decays to zero. By Chebyshev's inequality, this implies that our value estimator is consistent. The proof is thus completed.

Part 1. To simplify the proof, we assume $\mathbb{I}_k$ contains a single element i. The bias is given by

$$\binom{T}{m-1}^{-1}\sum_{t_1<\cdots<t_{m-1}}(\mathbb{E}\mathcal{D}_k^{(i,t_1)}\cdots\mathcal{D}_k^{(i,t_{m-1})}\widehat{Q}_k-Q^{\pi}).$$

We next apply Berbee's coupling lemma (see e.g., Lemma 4.1 in Dedecker & Louhichi, 2002) to bound the bias. Consider a given ordered tuple $(t_1,t_2,\cdots,t_{m-1})$. Following the discussion below Lemma 4.1 in (Dedecker & Louhichi, 2002), we can construct i.i.d. data tuples $\{(S_{i,t_l}^0,A_{i,t_l}^0,R_{i,t_l}^0,S_{i,t_l+1}^0)\}_{1\le l\le m-1}$ such that the event

$$(S_{i,t_l}^0,A_{i,t_l}^0,R_{i,t_l}^0,S_{i,t_l+1}^0)=(S_{i,t_l},A_{i,t_l},R_{i,t_l},S_{i,t_l+1}),\quad\forall 1\le l\le m-1,$$

holds with probability at least $1-\sum_{l=1}^{m-2}\beta(t_{l+1}-t_l-1)$ where $\beta(\cdot)$ denotes the β-mixing coefficients of $\{(S_t,A_t,R_t)\}_{t\ge0}$. This allows us to decompose each of the individual bias $|\mathbb{E}\mathcal{D}_k^{(i,t_1)}\cdots\mathcal{D}_k^{(i,t_{m-1})}\widehat{Q}_m-Q^{\pi}|$ into the following two terms

$$\begin{aligned}&|\mathbb{E}\mathcal{D}_k^{(i,t_1)}\cdots\mathcal{D}_k^{(i,t_{m-1})}\widehat{Q}_k-Q^{\pi}|\mathcal{I}\{(S_{i,t_l}^0,A_{i,t_l}^0,R_{i,t_l}^0,S_{i,t_l+1}^0)=(S_{i,t_l},A_{i,t_l},R_{i,t_l},S_{i,t_l+1}),\quad\forall 1\le l\le m-1\}\\+&|\mathbb{E}\mathcal{D}_k^{(i,t_1)}\cdots\mathcal{D}_k^{(i,t_{m-1})}\widehat{Q}_k-Q^{\pi}|\mathcal{I}\{(S_{i,t_l}^0,A_{i,t_l}^0,R_{i,t_l}^0,S_{i,t_l+1}^0)\ne(S_{i,t_l},A_{i,t_l},R_{i,t_l},S_{i,t_l+1}),\quad\exists 1\le l\le m-1\}.\end{aligned}$$

Based on Lemma 3, the first term can be upper bounded by $O(T^{-\alpha_1-(m-1)\alpha_2})$. Under the boundedness property, the second term is upper bounded by $c\{\sum_{l=1}^{m-2}\beta(t_{l+1}-t_l-1)\}$ for some constant $c>0$. Averaging over all possible combinations of individual debiasing operators yields the following upper bound

$$O(T^{-\alpha_*})+c\binom{T}{m-1}^{-1}\sum_{t_1<\cdots<t_{m-1}}\sum_{l=1}^{m-2}\beta(t_{l+1}-t_l-1).$$

Under (A1), we have $\beta(t)=O(\rho^t)$ for some $0<\rho<1$ and any $t\ge0$. The second term is upper bounded by $O(T^{-1})$. This yields the upper bound $O(T^{-\alpha_*})$ when $\mathbb{I}_k$ consists of a single element. In general, we can show the bias is upper bounded by $O\{(nT)^{-\alpha_*}\}$. Using similar arguments, we can show the bias of the value is upper bounded by $O\{(nT)^{-\alpha}\}$. This completes the proof for Part 1.

Part 2. For $1\le k\le\mathbb{K}$, let $\widehat{\eta}_{\text{TR},k}^{(m)}=(nT/\mathbb{K})^{-1}\sum_{i\in\mathbb{I}_k}\sum_{t=0}^{T-1}\psi_{i,t}^{(m)}$. By Cauchy-Schwarz inequality, it suffices to show the $\text{Var}(\widehat{\eta}_{\text{TR},k}^{(m)})\to0$ for each k. Using similar arguments in the proof of Lemma 3, we can show the difference $(nT/\mathbb{K})^{-1}\sum_{i\in\mathbb{I}_k}\sum_{t=0}^{T-1}(\psi_{i,t}^{(m)}-\phi_{i,t}^{(m)})$ is upper bounded by $O(n^{-1}T^{-1})$. Consequently, it suffices to upper bound the variance of $\widehat{\eta}_{\text{TR},k,U}^{(m)}=(nT/\mathbb{K})^{-1}\sum_{i\in\mathbb{I}_k}\sum_{t=0}^{T-1}\phi_{i,t}^{(m)}$.

A key observation is that, conditional on the estimators $\widehat{Q}_k$, $\widehat{\tau}_k$ and $\widehat{\omega}_k$, $\widehat{\eta}_{\text{TR},k,U}^{(m)}$ corresponds to an m-th order U-statistic. Under the given conditions, the kernel function associated with the U-statistic is uniformly bounded. We first consider the

variance of $\widehat{\eta}^{(m)}_{\text{TR},k,U}$ conditional on the nuisance estimators. To simplify the proof, we similarly assume that $\mathbb{I}_k$ consists of a single trajectory, as in Part 1. By definition, the conditional variance is given by

$$\left(\frac{m!}{T!}\right)^2 \sum_{\substack{\text{disjoint } t_1,\cdots,t_m \\ \text{disjoint } t_1',\cdots,t_m'}} \text{cov}\Bigg(\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\mathcal{D}_k^{(i,t_1)}\cdots\mathcal{D}_k^{(i,t_{m-1})}\widehat{Q}_k(a,s)+\frac{1}{1-\gamma}\widehat{\omega}_k(A_{i,t_m},S_{i,t_m})\{R_{i,m}$$
$$-\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,m+1})}\mathcal{D}_k^{(i,1)}\cdots\mathcal{D}_k^{(i,m-1)}\widehat{Q}_k(a,S_{i,m+1})+\mathcal{D}_k^{(i,1)}\cdots\mathcal{D}_k^{(i,m-1)}\widehat{Q}_k(A_{i,t_m},S_{i,t_m})\},$$
$$\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\mathcal{D}_k^{(i,t_1')}\cdots\mathcal{D}_k^{(i,t_{m-1}')}\widehat{Q}_k(a,s)+\frac{1}{1-\gamma}\widehat{\omega}_k(A_{i,t_m'},S_{i,t_m'})\{R_{i,t_m'}-\gamma$$
$$\times\mathbb{E}_{a\sim\pi(\cdot|S_{i,t_m'+1})}\mathcal{D}_k^{(i,1)}\cdots\mathcal{D}_k^{(i,m-1)}\widehat{Q}_k(a,S_{i,t_m'+1})+\mathcal{D}_k^{(i,t_1')}\cdots\mathcal{D}_k^{(i,t_m'-1)}\widehat{Q}_k(A_{i,t_m'},S_{i,t_m'})\}\Big|\,\widehat{Q}_k,\widehat{\tau}_k,\widehat{\omega}_k\Bigg),$$

where $\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}$ denotes the expectation by assuming $s\sim\mathbb{G}$ and $a\sim\pi(\cdot|s)$. Using similar arguments in Part 1, we can show that the above conditional variance decays to zero. In addition, $\mathbb{E}(\widehat{\eta}^{(m)}_{\text{TR},k,U}|\widehat{Q}_k,\widehat{\tau}_k,\widehat{\omega}_k)$ is to converge to η^π, when one of the nuisance estimator is consistent. Under the given conditions, $\widehat{\eta}^{(m)}_{\text{TR},k,U}$ is bounded. This further yields that $\text{Var}\{\mathbb{E}(\widehat{\eta}^{(m)}_{\text{TR},k,U}|\widehat{Q}_k,\widehat{\tau}_k,\widehat{\omega}_k)\}\to 0$. Together with the fact that the conditional variance of $\widehat{\eta}^{(m)}_{\text{TR},k,U}$ decays to zero, the variance of $\widehat{\eta}^{(m)}_{\text{TR},k,U}$ decays to zero. The proof is thus completed.

A.5. Proof of Theorem 2

In the proof of Theorem 1, we have shown that $\widehat{\eta}^{(m)}_{\text{TR},k}-\widehat{\eta}^{(m)}_{\text{TR},k,U}=O(n^{-1}T^{-1})$. This in turn implies that $\widehat{\eta}^{(m)}_{\text{TR}}-\widehat{\eta}^{(m)}_{\text{TR},U}=O(n^{-1}T^{-1})$ where $\widehat{\eta}^{(m)}_{\text{TR},U}$ is a simple average of $\{\widehat{\eta}^{(m)}_{\text{TR},k,U}\}_k$. It suffices to focus on $\widehat{\eta}^{(m)}_{\text{TR},U}$.

The rest of the proof is divided into three parts. We first define $\widehat{\eta}^{(m),*}_{\text{TR},U}$ as a version of $\widehat{\eta}^{(m)}_{\text{TR},U}$ with the Q-, marginalized density ratio and conditional density ratio estimators replaced by their oracle values, and prove that $\sqrt{nT}(\widehat{\eta}^{(m),*}_{\text{TR},U}-\eta^\pi)\xrightarrow{d}N(0,\sigma^2)$. We next show that the difference $\widehat{\eta}^{(m),*}_{\text{TR},U}-\widehat{\eta}^{(m)}_{\text{TR},U}+\mathbb{E}\widehat{\eta}^{(m)}_{\text{TR},U}-\eta^\pi$ is $o_p\{(nT)^{-1/2}\}$. The assertion thus follows from an application of Slutsky's theorem. Finally, in Part 3, we present the variance decomposition formula for $\text{Var}(\widehat{\eta}^{(m),*}_{\text{TR},U})$.

Part 1: A key observation is that, the oracle version of the estimator $\widehat{\eta}^{(m),*}_{\text{TR},U}-\eta^\pi$ corresponds to an m-th order U-statistic. The corresponding symmetric kernel function is given by

$$h(\{(S_{i_j,t_j},A_{i_j,t_j},R_{i_j,t_j},S_{i_j,t_j+1})\}_{j=1}^m)=\frac{1}{m(1-\gamma)}\sum_{j=1}^m\Bigg[\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\prod_{l\neq j}\mathcal{D}^{(i_l,t_l)}Q^\pi(a,s)+\frac{1}{1-\gamma}\omega^\pi(A_{i_j,t_j},S_{i_j,t_j})$$
$$\times\Bigg\{R_{i_j,t_j}+\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i_j,t_j+1})}\prod_{l\neq j}\mathcal{D}^{(i_l,t_l)}Q^\pi(a,S_{i_j,t_j+1})-\prod_{l\neq j}\mathcal{D}^{(i_l,t_l)}Q^\pi(A_{i_j,t_j},S_{i_j,t_j})\Bigg\}\Bigg]-\eta^\pi.$$

Here, $\mathcal{D}^{(i_1,t_1)}$ denotes a version of $\mathcal{D}_k^{(i_1,t_1)}$ by replacing the estimator $\widehat{\tau}_k$ with the oracle value τ^π. Under (A1) and the boundedness assumption in (A3), the conditions in Theorem 1 (c) of Denker & Keller (1983) are satisfied. The asymptotic normality of $\widehat{\eta}^{(m),*}_{\text{TR},U}$ is thus proven. In addition, the asymptotic variance of $\sqrt{nT}(\widehat{\eta}^{(m),*}_{\text{TR},U}-\eta^\pi)$ is given by $(nT)^{-1}m^2\mathbb{E}|\sum_{i,t}h_1(S_{i,t},A_{i,t},R_{i,t},S_{i,t+1})|^2$ where

$$h_1(s_1,a_1,r_1,s_1')=\mathbb{E}_{(s_2,a_2,r_2,s_2'),\cdots,(s_m,a_m,r_m,s_m')\overset{iid}{\sim}p_\infty}h(\{(s_j,a_j,r_j,s_j')\}_{j=1}^m).$$

Here, we use p_∞ to denote the limiting distribution of the stochastic process $\{(S_t,A_t,R_t,S_{t+1})\}_{t\geq 0}$.

Since the expectation of the temporal-difference error $r+\gamma\mathbb{E}_{a'\sim\pi(\cdot|s')}Q^\pi(a',s')-Q(a,s)$ is zero under the distribution p_∞, the function $h_1(s_1,a_1,r_1,s_1')$ equals

$$\frac{1}{m(1-\gamma)}\omega^\pi(a_1,s_1)\{r_1+\gamma\mathbb{E}_{a_1'\sim\pi(\cdot|s_1')}Q^\pi(a_1',s_1')-Q(a_1,s_1)\}.$$

Consequently, the asymptotic variance σ^2 equals

$$\frac{1}{nT(1-\gamma)^2}\mathbb{E}\left|\sum_{i,t}\omega^\pi(A_{i,t},S_{i,t})\{R_{i,t}+\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}Q^\pi(a',S_{i,t+1})-Q^\pi(A_{i,t},S_{i,t})\}\right|^2.$$

Under MA and CMIA, for any index i, the sequence of temporal-difference errors $\{\varepsilon_{i,t}\}_{t\geq 0} = \{R_{i,t} + \gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}Q^\pi(a',S_{i,t+1}) - Q^\pi(A_{i,t},S_{i,t})\}_{t\geq 0}$ forms a martingale difference sequence. As such, the elements in $\{\omega^\pi(A_{i,t},S_{i,t})\varepsilon_{i,t}\}_{t\geq 0}$ are pairwise uncorrelated. Consequently,

$$\sigma^2=\frac{1}{nT(1-\gamma)^2}\sum_{i,t}\mathbb{E}\left|\omega^\pi(A_{i,t},S_{i,t})\{R_{i,t}+\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}Q^\pi(a',S_{i,t+1})-Q^\pi(A_{i,t},S_{i,t})\}\right|^2,$$

and is equal to (3). This completes the proof for Part 1.

Part 2: For any $1\leq k\leq\mathbb{K}$, we similarly define $\widehat{\eta}^{(m),*}_{\text{TR},k,U}$ as the oracle version of $\widehat{\eta}^{(m)}_{\text{TR},k,U}$. In this Part, we focus on proving $\sqrt{nT}\{\widehat{\eta}^{(m)}_{\text{TR},k,U}-\widehat{\eta}^{(m),*}_{\text{TR},k,U}-\mathbb{E}(\widehat{\eta}^{(m)}_{\text{TR},k,U}|\widehat{Q}_k,\widehat{\tau}_k,\widehat{\omega}_k)+\eta^\pi\}=o_p(1)$. This in turn implies that $\sqrt{nT}(\widehat{\eta}^{(m)}_{\text{TR},k,U}-\widehat{\eta}^{(m),*}_{\text{TR},k,U}-\mathbb{E}\widehat{\eta}^{(m)}_{\text{TR},k,U}+\eta^\pi)=o_p(1)$ and hence $\sqrt{nT}(\widehat{\eta}^{(m)}_{\text{TR},U}-\widehat{\eta}^{(m),*}_{\text{TR},U}-\mathbb{E}\widehat{\eta}^{(m)}_{\text{TR},U}+\eta^\pi)=o_p\{(nT)^{-1/2}\}$.

We next show $\sqrt{nT}\{\widehat{\eta}^{(m)}_{\text{TR},k,U}-\widehat{\eta}^{(m),*}_{\text{TR},k,U}-\mathbb{E}(\widehat{\eta}^{(m)}_{\text{TR},k,U}|\widehat{Q}_k,\widehat{\tau}_k,\widehat{\omega}_k)+\eta^\pi\}=o_p(1)$. To simplify the proof, we assume $\mathbb{I}_k$ consists of a single element i. Note that $\widehat{\eta}^{(m)}_{\text{TR},k,U}-\widehat{\eta}^{(m),*}_{\text{TR},k,U}$ can be decomposed into the sum $\sum_{j=0}^m\widehat{\eta}_{j,k}$ where $\widehat{\eta}_{0,k}$ is the main effect term, $\widehat{\eta}_{1,k}$ is the first-order linear term and $\widehat{\eta}_{j,k}$ is the high-order U-statistic for any $j\geq 2$. Specifically,

$$\widehat{\eta}_{0,k}=\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\{\widehat{Q}_k(a,s)-Q^\pi(a,s)\},$$

corresponding to the difference between two plug-in estimators. Its conditional variance equals zero given $\widehat{Q}_k$ and we have $\widehat{\eta}_{0,k}=\mathbb{E}(\widehat{\eta}_{0,k}|\widehat{Q}_k)$. $(1-\gamma)\widehat{\eta}_{1,k}$ equals

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\widehat{\omega}_k(A_{i,t},S_{i,t})[Q^\pi(A_{i,t},S_{i,t})-\widehat{Q}_k(A_{i,t},S_{i,t})-\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\{Q^\pi(a,S_{i,t+1})-\widehat{Q}_k(a,S_{i,t+1})\}]\\
+&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)[Q^\pi(A_{i,t},S_{i,t})-\widehat{Q}_k(A_{i,t},S_{i,t})-\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\{Q^\pi(a,S_{i,t+1})-\widehat{Q}_k(a,S_{i,t+1})\}]\\
+&\frac{1}{T}\sum_{t=0}^{T-1}\{\widehat{\omega}_k(A_{i,t},S_{i,t})+\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-2\omega^\pi(A_{i,t},S_{i,t})\}\varepsilon_{i,t}.
\end{aligned}$$

Using similar arguments in the proof of Part 1, the conditional variance of the third line given $\widehat{\omega}_k$ and $\widehat{\tau}_k$ is equal to $T^{-1}\mathbb{E}\{\widehat{\omega}_k(A_{i,t},S_{i,t})+\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-2\omega^\pi(A_{i,t},S_{i,t})\}^2\varepsilon_{i,t}^2$. It is of the order $o_p(T^{-1})$ given that $\widehat{\omega}_k$ and $\widehat{\tau}_k$ coverages to ω^π and τ^π, respectively. As such, we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\{\widehat{\omega}_k(A_{i,t},S_{i,t})+\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-2\omega^\pi(A_{i,t},S_{i,t})\}\varepsilon_{i,t}\\
=&\mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1}\{\widehat{\omega}_k(A_{i,t},S_{i,t})+\mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t},S_{i,t},a,s)-2\omega^\pi(A_{i,t},S_{i,t})\}\varepsilon_{i,t}\middle|\widehat{\omega}_k,\widehat{\tau}_k\right]+o_p(T^{-1/2}).
\end{aligned}\tag{4}$$

As for the first line, similar to the proof of Theorem 1, we will apply Berbee's coupling lemma to bound its conditional variance. Specifically, following the discussion below Lemma 4.1 of (Dedecker & Louhichi, 2002), we can construct a sequence of data tuples $\{O^0_{i,t}=(S^0_{i,t_l},A^0_{i,t_l},R^0_{i,t_l},S^0_{i,t_l+1})\}_{1\leq l\leq m-1}$ such that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1}\widehat{\omega}_k(A_{i,t},S_{i,t})[Q^\pi(A_{i,t},S_{i,t})-\widehat{Q}_k(A_{i,t},S_{i,t})-\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\{Q^\pi(a,S_{i,t+1})-\widehat{Q}_k(a,S_{i,t+1})\}]\\
=&\frac{1}{T}\sum_{t=0}^{T-1}\widehat{\omega}_k(A^0_{i,t},S^0_{i,t})[Q^\pi(A^0_{i,t},S^0_{i,t})-\widehat{Q}_k(A^0_{i,t},S^0_{i,t})-\gamma\mathbb{E}_{a\sim\pi(\cdot|S^0_{i,t+1})}\{Q^\pi(a,S^0_{i,t+1})-\widehat{Q}_k(a,S^0_{i,t+1})\}],
\end{aligned}\tag{5}$$

with probability at least $1 - T\beta(q)/q$ such that the sequences $\{U^0_{i,2t} : i \geq 0\}$ and $\{U^0_{i,2t+1} : i \geq 0\}$ are i.i.d. where $U^0_i = (O^0_{i,tq}, O^0_{i,tq+1}, \cdots, O^0_{i,tq+q-1})$. Due to the independence, the conditional variance of (5) is upper bounded by $O_p(q^2 T^{-1-2\alpha_1})$, under Condition (A2). Take q to be proportional to $\log T$, the probability $1 - T\beta(q)/q$ will approach 1, under Condition (A1). As such, the conditional variance of (5) is $o_p(T^{-1})$ and we have

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \widehat{\omega}_k(A^0_{i,t}, S^0_{i,t})[Q^\pi(A^0_{i,t}, S^0_{i,t}) - \widehat{Q}_k(A^0_{i,t}, S^0_{i,t}) - \gamma \mathbb{E}_{a\sim\pi(\cdot|S^0_{i,t+1})}\{Q^\pi(a, S^0_{i,t+1}) - \widehat{Q}_k(a, S^0_{i,t+1})\}] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \widehat{\omega}_k(A^0_{i,t}, S^0_{i,t})[Q^\pi(A^0_{i,t}, S^0_{i,t}) - \widehat{Q}_k(A^0_{i,t}, S^0_{i,t}) - \gamma \mathbb{E}_{a\sim\pi(\cdot|S^0_{i,t+1})}\{Q^\pi(a, S^0_{i,t+1}) - \widehat{Q}_k(a, S^0_{i,t+1})\}] \middle| \widehat{Q}_k, \widehat{\omega}_k\right] \\
&\qquad + o_p(T^{-1/2}).
\end{aligned}$$

This in turn implies that

$$\begin{aligned}
&\frac{1}{T}\sum_{t=0}^{T-1} \widehat{\omega}_k(A_{i,t}, S_{i,t})[Q^\pi(A_{i,t}, S_{i,t}) - \widehat{Q}_k(A_{i,t}, S_{i,t}) - \gamma \mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\{Q^\pi(a, S_{i,t+1}) - \widehat{Q}_k(a, S_{i,t+1})\}] \\
&= \mathbb{E}\left[\frac{1}{T}\sum_{t=0}^{T-1} \widehat{\omega}_k(A_{i,t}, S_{i,t})[Q^\pi(A_{i,t}, S_{i,t}) - \widehat{Q}_k(A_{i,t}, S_{i,t}) - \gamma \mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\{Q^\pi(a, S_{i,t+1}) - \widehat{Q}_k(a, S_{i,t+1})\}] \middle| \widehat{Q}_k, \widehat{\omega}_k\right] \\
&\qquad + o_p(T^{-1/2}).
\end{aligned} \tag{6}$$

Using similar arguments, we can show the second line satisfies a similar relation as well. This together with (4) and (6) yields that $\widehat{\eta}_{1,k} = \mathbb{E}(\widehat{\eta}_{1,k}|\widehat{Q}_k, \widehat{\omega}_k, \widehat{\tau}_k) + o_p(T^{-1/2})$.

$\widehat{\eta}_{2,k}$ equals $\{T(T-1)\}^{-1}\sum_{t_1\neq t_2}\widehat{\eta}_{2,t_1,t_2,k}$ where $(1-\gamma)^2\widehat{\eta}_{2,t_1,t_2,k}$ equals

$$\begin{aligned}
&\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t_1+1})}[\{\widehat{\omega}_k(A_{i,t_1}, S_{i,t_1}) + \mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t_1}, S_{i,t_1}, a, s)\}\widehat{\tau}_k(A_{i,t_2}, S_{i,t_2}, a, S_{i,t_1+1}) \\
&\qquad -2\omega^\pi(A_{i,t_1}, S_{i,t_1})\tau^\pi(A_{i,t_2}, S_{i,t_2}, a, S_{i,t_1+1})]\varepsilon_{i_2,t_2} \\
&-[\{\widehat{\omega}_k(A_{i,t_1}, S_{i,t_1}) + \mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t_1}, S_{i,t_1}, a, s)\}\widehat{\tau}_k(A_{i,t_2}, S_{i,t_2}, A_{i,t_1}, S_{i,t_1}) \\
&\qquad -2\omega^\pi(A_{i,t_1}, S_{i,t_1})\tau^\pi(A_{i,t_2}, S_{i,t_2}, A_{i,t_1}, S_{i,t_1})]\varepsilon_{i_2,t_2} \\
&+\{\widehat{\omega}_k(A_{i,t_1}, S_{i,t_1}) + \mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t_1}, S_{i,t_1}, a, s)\}\widehat{\tau}_k(A_{i,t_2}, S_{i,t_2}, A_{i,t_1}, S_{i,t_1}) \\
&\times\{Q^\pi(A_{i,t_2}, S_{i,t_2}) - \widehat{Q}_k(A_{i,t_2}, S_{i,t_2}) - \mathbb{E}_{a\sim\pi(\cdot|S_{i,t_2+1})}\{Q^\pi(a, S_{i,t_2+1}) - \widehat{Q}_k(a, S_{i,t_2+1})\}\} \\
&-\gamma\{\widehat{\omega}_k(A_{i,t_1}, S_{i,t_1}) + \mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}\widehat{\tau}_k(A_{i,t_1}, S_{i,t_1}, a, s)\}\mathbb{E}_{a\sim\pi(\cdot|S_{i,t_1+1})}\widehat{\tau}_k(A_{i,t_2}, S_{i,t_2}, a, S_{i,t_1+1}) \\
&\times\{Q^\pi(A_{i,t_2}, S_{i,t_2}) - \widehat{Q}_k(A_{i,t_2}, S_{i,t_2}) - \mathbb{E}_{a\sim\pi(\cdot|S_{i,t_2+1})}\{Q^\pi(a, S_{i,t_2+1}) - \widehat{Q}_k(a, S_{i,t_2+1})\}\}.
\end{aligned}$$

Other high-order terms can be similarly derived. Using similar arguments in proving $\widehat{\eta}_{1,k} = \mathbb{E}(\widehat{\eta}_{1,k}|\widehat{Q}_k) + o_p(T^{-1/2})$, we can show $\widehat{\eta}_{j,k} = \mathbb{E}(\widehat{\eta}_{j,k}|\widehat{Q}_k, \widehat{\omega}_k, \widehat{\tau}_k) + o_p(T^{-1/2})$ for any $j \geq 2$. This further implies that $\widehat{\eta}^{(m)}_{\text{TR},k,U} - \widehat{\eta}^{(m),*}_{\text{TR},k,U} - \mathbb{E}(\widehat{\eta}^{(m)}_{\text{TR},k,U}|\widehat{Q}_k, \widehat{\tau}_k, \widehat{\omega}_k) + \eta^\pi = o_p(T^{-1/2})$, since $\mathbb{E}\widehat{\eta}^{(m),*}_{\text{TR},k,U} = \eta^\pi$. More generally, when $\mathbb{I}_k$ consists of multiple trajectories, we can similarly show that $\widehat{\eta}^{(m)}_{\text{TR},k,U} - \widehat{\eta}^{(m),*}_{\text{TR},k,U} - \mathbb{E}(\widehat{\eta}^{(m)}_{\text{TR},k,U}|\widehat{Q}_k, \widehat{\tau}_k, \widehat{\omega}_k) + \eta^\pi = o_p(n^{-1/2}T^{-1/2})$. This completes the proof of Part 2.

Part 3: Finally, we discuss the variance decomposition formula. Similar to Step 2, we can decompose $\widehat{\eta}^{(m),*}_{\text{TR},U}$ into the sum $\sum_{j=0}^m \widehat{\eta}^*_j$ where $\widehat{\eta}^*_0$ is the main effect $\eta^\pi = \mathbb{E}_{(a,s)\sim(\pi,\mathbb{G})}Q^\pi(a,s)$, $\widehat{\eta}^*_1$ is the first-order term

$$\frac{1}{nT(1-\gamma)}\sum_{i=1}^n\sum_{t=0}^{T-1}\omega^\pi(A_{i,t}, S_{i,t})\{R_{i,t} + \gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}Q^\pi(a, S_{i,t+1}) - Q^\pi(A_{i,t}, S_{i,t})\}.$$

For any $j \geq 2$, $\widehat{\eta}^*_j$ corresponds to a degenerate U-statistic whose explicit form is given by

$$\binom{m}{j}\frac{j!}{(nT)!}\sum_{\text{disjoint } (i_1,t_1),\cdots,(i_j,t_j)} h_r(\{(S_{i_l,t_l}, A_{i_l,t_l}, R_{i_l,t_l}, S_{i_l,t_{l+1}})\}_{l=1}^j),$$

where

$$h_r(\{(s_l,a_l,r_l,s_l')\}_{l=1}^j)=\sum_{r=1}^{j}\binom{j}{r}(-1)^{j-r}\mathbb{E}_{(s_{l+1},a_{l+1},r_{l+1},s_{l+1}'),\cdots,(s_m,a_m,r_m,s_m')\overset{iid}{\sim}p_\infty}h(\{(s_j,a_j,r_j,s_j')\}_{j=1}^m),$$

where the kernel h is defined in Part 1. For instance,

$$\widehat{\eta}_2^*=\frac{1}{(1-\gamma)^2nT(nT-1)}\sum_{(i_1,t_1)\neq(i_2,t_2)}\Big[\omega^\pi(A_{i_1,t_1},S_{i_1,t_1})\{\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i_1,t_1+1})}\tau^\pi(A_{i_2,t_2},S_{i_2,t_2},a',S_{i_1,t_1+1})$$
$$-\tau^\pi(A_{i_2,t_2},S_{i_2,t_2},A_{i_1,t_1},S_{i_1,t_1})\}+(1-\gamma)\omega^\pi(A_{i_2,t_2},S_{i_2,t_2})\Big]\epsilon_{i_2,t_2}.$$

Other high-order terms can be similarly derived.

A.6. Proof of Theorem 3

By Theorem 2, we have $\sqrt{nT}(\widehat{\eta}^{(m)}-\mathbb{E}\widehat{\eta}^{(m)})\xrightarrow{d}N(0,\sigma^2)$ for any m. Under the given conditions, using similar arguments in Part 1 of the proof of Theorem 1, $\mathbb{E}\widehat{\eta}^{(m)}$ converges to η^π at a rate of $o\{(nT)^{-1/2}\}$. This further implies that $\sqrt{nT}(\widehat{\eta}^{(m)}-\eta^\pi)\xrightarrow{d}N(0,\sigma^2)$.

To prove the validity of our CI, it suffices to show the sampling variance estimator $(\widehat{\sigma}^{(m)})^2$ is consistent. The consistency can be proven using similar arguments in Part 2 of the proof of Theorem 2. We omit the details to save space.

B. More on the estimation of the nuisance functions

B.1. Fitted-Q evaluation

We review the fitted-Q evaluation (FQE) algorithm proposed in Le et al. (2019), which is the subroutine we use to learn the Q-function. FQE is an iterative algorithm based on the Bellman's equation:

$$Q(a,s)=\mathbb{E}_{a'\sim\pi(\cdot|s)}\left(R_t+\gamma Q(a'|S_{t+1})|A_t=a,S_t=s\right).$$

Based on this equation, we iteratively update the estimate by

$$Q_m(a,s)=\arg\min_Q\sum_{i'\in\mathbb{I}_k}\sum_{t<T}(\gamma\mathbb{E}_{a'\sim\pi(\cdot|S_{i,t+1})}Q_{m-1}(a'|S_{i,t+1})$$
$$+R_{i,t}-Q(A_{i,t},S_{i,t}))^2,$$

for $m=1,2,\cdots$. The optimization problem can be solved with various supervised learning algorithms. We summarize FQE in Algorithm 1.

Algorithm 1 Fitted-Q evaluation

Input: Data $\{S_{j,t},A_{j,t},R_{j,t},S_{j,t+1}\}_{j,t}$, policy π, function class $\mathcal{F}$, decay rate γ, number of iterations M
Randomly pick $Q_0\in\mathcal{F}$
for $m=1,\ldots,M$ **do**
  Update target values $Z_{j,t}=R_{j,t}+\gamma Q_{m-1}(S_{j,t+1},\pi(S_{j,t+1}))$ for all (j,t);
  Solve a regression problem to update the Q-function:
  $Q_m=\arg\min_{Q\in\mathcal{F}}\frac{1}{n}\sum_{i=1}^n\{Q(S_{j,t},A_{j,t})-Z_{j,t}\}^2$
end for
Output: The estimated Q-function $Q_M(\cdot,\cdot)$

B.2. Learning the density ratio ω

The estimation of the density ratio ω is based on the following key observation.

Lemma 1 *For any function f, we have $L(\omega, f) = 0$, where $L(\omega, f)$ is*

$$\begin{aligned}\mathbb{E}_{a\sim\pi(\cdot|S_{i,t+1})}\{\omega(A_{i,t}, S_{i,t})(\gamma f(a, S_{i,t+1}) - f(A_{i,t}, S_{i,t}))\} \\ + (1-\gamma)\mathbb{E}_{S_0\sim\mathbb{G}, a\sim\pi(\cdot|S_0)} f(a, S_0).\end{aligned} \tag{7}$$

Conversely, ω is the only function satisfying this condition.

Therefore, as suggested in Uehara et al. (2019), ω can be learned by solving the following mini-max problem

$$\arg\min_{\omega\in\Omega} \sup_{f\in\mathcal{F}} L(\omega, f)^2, \tag{8}$$

for some functional class Ω and $\mathcal{F}$. The expectation in (7) is approximated by the sample mean. To simplify the calculation, we can choose $\mathcal{F}$ to be a reproducing kernel Hilbert space (RKHS) , with which the inner maximization has a closed form solution, and then ω can be learned by solving the outer minimization via stochastic gradient descent. Let $\kappa(\cdot,\cdot;\cdot,\cdot)$ be the kernel function of the RKHS. Consider sampling a random minibatch $\{S_{i_g,t_g}, A_{i_g,t_g}, S_{i_g,t_g+1} : g \in \mathcal{M}\}$ from a data subset $\mathbb{I}_k$. We form the objective function $D(\omega)$ as $\binom{|\mathcal{M}|}{2}^{-1}\sum_{g_1,g_2\in\mathcal{M},g_1\neq g_2} D(\omega, g_1, g_2)$ where $D(\omega, g_1, g_2)$ is equal to

$$\begin{aligned}2(1-\gamma)\omega(X_{i_{g_1},t_{g_1}})\Big\{\gamma\mathbb{E}_{\substack{a\sim\pi(\bullet|S_{i_{g_1},t_{g_1}+1})\\ s'\sim\mathbb{G},a'\sim\pi(\bullet|s')}}\kappa(S_{i_{g_1},t_{g_1}+1}, a; s', a') - \mathbb{E}_{s'\sim\mathbb{G},a'\sim\pi(\bullet|s')}\kappa(X_{i_{g_1},t_{g_1}}; s', a')\Big\} \\ +\omega(X_{i_{g_1},t_{g_1}})\omega(X_{i_{g_2},t_{g_2}})\Big\{\gamma^2\mathbb{E}_{\substack{a_1\sim\pi(\bullet|S_{i_{g_1},t_{g_1}+1})\\ a_2\sim\pi(\bullet|S_{i_{g_2},t_{g_2}+1})}}\kappa(S_{i_{g_2},t_{g_2}+1}, a_2; S_{i_{g_1},t_{g_1}+1}, a_1) \\ -2\gamma\mathbb{E}_{a\sim\pi(\bullet|S_{i_{g_1},t_{g_1}+1})}\kappa(S_{i_{g_1},t_{g_1}+1}, a; X_{i_{g_2},t_{g_2}}) + \kappa(X_{i_2,t_2}; X_{i_1,t_1})\Big\} \\ +(1-\gamma)^2\mathbb{E}_{\substack{s',s''\sim\mathbb{G}\\ a'\sim\pi(\bullet|s'),a''\sim\pi(\bullet|s'')}}\kappa(a', s'; a'', s''),\end{aligned}$$

where $X_{i,t}$ denotes the state-action pair $(A_{i,t}, S_{i,t})$. Thus, in each step, we take a random minibatch from the observed data. Then we update the model parameter

$$\theta \leftarrow \theta - \epsilon\Delta_\theta D(\omega_\theta/z_{\omega_\theta}),$$

where z_{ω_θ} is a normalizing constant such that

$$z_{\omega_\theta} = \frac{1}{|\mathcal{M}|}\sum_{g\in\mathcal{M}}\omega_\theta(A_{i_g,t_g}, S_{i_g,t_g}).$$

Note that ω satisfies $\mathbb{E}\omega(\pi, A_t, S_t) = 1$. For a given $\widehat{\omega}_k$, we can further normalize the density ratio by $\widehat{\omega}_k(\bullet) = \widehat{\omega}_k(\bullet)/\{\sum_{j,t}\widehat{\omega}_k(A_{j,t}, S_{j,t})/(nT)\}$. This yields the final estimates.

B.3. Learning the conditional sampling ratio τ

Following the same analogy, our algorithm for estimating τ is motivated by the following key observation.

Lemma 2 *For any two pairs (i,t) and (i',t') such that $O_{i,t}$ and $O_{i',t'}$ are independent, we have for any function f that $\mathbb{E}\Delta(\tau, f, \pi; i, t, i', t') = 0$, where $\Delta(\tau, f, \pi; i, t, i', t')$ is*

$$\begin{aligned}\tau(S_{i',t'}, A_{i',t'}; A_{i,t}, S_{i,t})\Big\{\gamma\mathbb{E}_{a\sim\pi(\cdot|S_{i',t'+1})} f(S_{i',t'+1}, a; A_{i,t}, S_{i,t}) \\ - f(S_{i',t'}, A_{i',t'}; A_{i,t}, S_{i,t})\Big\} + (1-\gamma)f(A_{i,t}, S_{i,t}, A_{i,t}, S_{i,t}).\end{aligned}$$

Conversely, τ is the only function satisfying this condition.

Therefore, τ can be learned by solving the following mini-max problem

$$\arg\min_{\omega\in\Omega}\sup_{f\in\mathcal{F}}\left|\sum_{(i,t)\neq(i',t')}\Delta(\omega, f, \pi; i, t, i', t')\right|^2, \tag{9}$$

Algorithm 2 Estimation of the density ratio.

Input: The data subset in $\mathcal{I}_\ell$.
Initial: Initial the density ratio $\omega = \omega_\beta$ to be a neural network parameterized by β.
for iteration $= 1, 2, \cdots$ **do**

a. Randomly sample batches $\mathcal{M}$, $\mathcal{M}^*$ from the data transitions.
b. **Update** the parameter β by

$$\beta \leftarrow \beta - \epsilon \binom{|\mathcal{M}|}{2}^{-2} \sum_{\substack{(i_1,t_1),(i_1',t_1')\in\mathcal{M} \\ (i_1,t_1)\neq(i_1',t_1')}} \sum_{\substack{(i_2,t_2),(i_2',t_2')\in\mathcal{M} \\ (i_2,t_2)\neq(i_2',t_2')}} \nabla_\beta D(\frac{\omega_\beta}{z_{\omega_\beta}}, \pi; i_1, t_1, i_1', t_1', i_2, t_2, i_2', t_2'),$$

where z_{ω_β} is a normalization constant

$$z_{\omega_\beta}(\cdot; A_{i,t}, S_{i,t}) = \frac{1}{|\mathcal{M}^*|} \sum_{(i',t')\in\mathcal{M}^*} \omega_\beta(X_{i',t'}; X_{i,t}).$$

end for
Output: the density ratio ω_β.

for some functional class Ω and $\mathcal{F}$. The optimization for τ can be implemented in a similar way as that for ω. Specifically, We set $\mathcal{F}$ to a unit ball of a reproducing kernel Hilbert space (RFHS), i.e., $\mathcal{F} = \{f \in \mathcal{H} : \|f\|_{\mathcal{H}} = 1\}$, where

$$\mathcal{H} = \left\{ f(\cdot) = \sum_{(i,t)\neq(i',t')} b_{i,t,i',t'} \kappa(X_{i',t'}, X_{i,t}; \cdot) : b_{i,t,i',t'} \in \mathbb{R} \right\},$$

for some positive definite kernel $\kappa(\cdot;\cdot)$, where $X_{i,t}$ is a shorthand for the state-action pair $(A_{i,t}, S_{i,t})$. The optimization problem in (9) is then reduced to

$$\arg\min_{\omega\in\Omega} \sum_{\substack{(i_1,t_1)\neq(i_1',t_1') \\ (i_2,t_2)\neq(i_2',t_2')}} D(\omega, \pi; i_1, t_1, i_1', t_1', i_2, t_2, i_2', t_2'),$$

where $D(\omega, \pi; i_1, t_1, i_1', t_1', i_2, t_2, i_2', t_2')$ is given by

$$\begin{aligned}
&\frac{\omega(X_{i_1',t_1'}; X_{i_1,t_1})}{(1-\gamma)^{-1}} \Big\{ \gamma \mathbb{E}_{a\sim\pi(\bullet|S_{i_1',t_1'+1})} \kappa(S_{i_1',t_1'+1}, a, X_{i_1,t_1}; X_{i_2,t_2}, X_{i_2,t_2}) - \kappa(X_{i_1',t_1'}, X_{i_1,t_1}; X_{i_2,t_2}, X_{i_2,t_2}) \Big\} \\
+&\frac{\omega(X_{i_2',t_2'}; X_{i_2,t_2})}{(1-\gamma)^{-1}} \Big\{ \gamma \mathbb{E}_{a\sim\pi(\bullet|S_{i_2',t_2'+1})} \kappa(S_{i_2',t_2'+1}, a, X_{i_2,t_2}; X_{i_1,t_1}, X_{i_1,t_1}) - \kappa(X_{i_2',t_2'}, X_{i_2,t_2}; X_{i_1,t_1}, X_{i_1,t_1}) \Big\} \\
&+\omega(X_{i_1',t_1'}; X_{i_1,t_1}) \omega(X_{i_2',t_2'}; X_{i_2,t_2}) \Big\{ \gamma^2 \mathbb{E}_{\substack{a_1\sim\pi(\bullet|S_{i_1',t_1'+1}) \\ a_2\sim\pi(\bullet|S_{i_2',t_2'+1})}} \kappa(S_{i_2',t_2'+1}, a, X_{i_2,t_2}; S_{i_1',t_1'+1}, a_1, X_{i_1,t_1}) \\
-&\gamma \mathbb{E}_{a_1\sim\pi(\bullet|S_{i_1',t_1'+1})} \kappa(S_{i_1',t_1'+1}, a, X_{i_1,t_1}; X_{i_2',t_2'}, X_{i_2,t_2}) - \gamma \mathbb{E}_{a_2\sim\pi(\bullet|S_{i_2',t_2'+1})} \kappa(S_{i_2',t_2'+1}, a, X_{i_2,t_2}; X_{i_1',t_1'}, X_{i_1,t_1}) \\
&+\kappa(X_{i_2',t_2'}, X_{i_2,t_2}; X_{i_1',t_1'}, X_{i_1,t_1}) \Big\} + (1-\gamma)^2 \kappa(X_{i_1,t_1}, X_{i_1,t_1}; X_{i_2,t_2}, X_{i_2,t_2}).
\end{aligned}$$

In our implementation, we set Ω to the class of neural networks. The detailed estimating procedure is given in Algorithm 2.

C. Additional numerical details

In this section, we report more details of the simulation environments and the algorithm implementations.

C.1. More about the toy example

The behaviour policy is chosen as a Bernoulli distribution with equal probabilities, and the target policy is chosen as follows: if the agent is at state A, then it takes action to transit to B or C with equal probabilities, while if it is at state B or C, it takes action to transit to A with probability 1.0. The movement is uncertain: with probability 0.9 the transition will follow the action, and with 0.1 the agent will just stay where it is. The initial states are equally distributed over the three states. In Figure 1, when the convergence rate of nuisance estimators is set as $(nT)^{-\alpha}$, to inject noises in the nuisance functions, we add a noise following $\mathcal{N}(0, (0.2n^{-\alpha})^2)$ to $Q(s,a)$ when Q is contaminated, and add a noise following $\mathcal{N}(0, (0.04n^{-\alpha})^2)$ to the corrsponding density ratio when ω or τ is contaminated. In Figure 2, to inject noises in the nuisance functions, we add a fixed noise following $\mathcal{N}(0, 0.2^2)$ to $Q(s,a)$ when Q is contaminated, and add a fixed noise following $\mathcal{N}(0, 0.04^2)$ to the corresponding density ratio when ω or τ is contaminated. The length of trajectories is fixed as 50 for all settings.

C.2. More about the simulation settings

C.2.1. The modified Cartpole environment

Following Uehara et al. (2019), we slightly modified the original Cartpole environment in Brockman et al. (2016) to better fit the off-policy evaluation task. Specifically, we add small Gaussian noise with mean zero and standard deviation 0.02 on the original deterministic transition dynamics, and define a new state-action-dependent reward as $(1-(x^2)/11.52-(\theta^2)/288)$, where x is the cart position and θ is the pole angle, to replace the original constant rewards.

C.2.2. The Diabetes environment

We use the simulation environment about an mobile health application on diabetes control calibrated in Shi et al. (2020). The state vector is 15-dimensional and it contains the measurements of four hourly covariates and the hourly amounts of insulin injected in the past four hours, and the action space is discrete with 5 levels on different amounts of insulin injection. The reward is a deterministic function of the glucose level, the state transition for the glucose is a linear function estimated from real data, and the noise for the glucose is set to have standard deviation 10 in our experiment. The objective is to learn an optimal policy that maps patients time-varying coefficients into the amount of insulin injected to optimal patients health status. More details can be found in Shi et al. (2020).

C.2.3. Construction of the Behaviour and target policies

For both environment, we first run deep-Q network to get a near-optimal Q-function $Q(s,a)$, and then apply softmax on its Q-value divided by an adjustable temperature τ to define the action probability of a behaviour policy as

$$\pi_b(a|s) \propto exp(\frac{Q(s,a)}{\tau})$$

For Cartpole, we model the Q-function as a dense neural network with 2 hidden layers of dimension 256, and set the optimizer as Adam with batch size 64 and learning rate 0.01. For Diabetes, we model the Q-function as a dense neural network with 2 hidden layers of dimension 64, and set the optimizer as Adam with batch size 128 and learning rate 0.0001.

C.3. Implementation details

For the Cartpole experiment, to implement our method, we set $\mathbb{K} = 2$ and sample 5% of the total pairs in calculation of the incomplete U-statistics. To estimate the Q-function, we use random forests to model the Q-function, with the number of trees set as 1000 and their max depth as 20. To estimate ω, we model it as a dense neural network with 5 hidden layers of dimension 512, connected via ReLu, and model the kernel $k(\cdot,\cdot)$ as a Laplacian kernel with bandwidth chosen by the median heuristic. We optimize the problem via Adam with batch size 256 and learning rate 0.001. To estimate τ, we model it as a dense neural network with 3 hidden layers of dimension 512, and optimize the problem via Adam with batch size 32 and learning rate 0.0001, with the other hyper-parameters the same with those of ω.

For the Diabetes experiment, to implement our method, we keep the other hyper-parameters the same with those for Cartpole, except that we sample 20% of the total pairs in calculation of the incomplete U-statistics, adjust the number of trees as 1000 and their max depth as 50, and adjust the learning rate for ω as 0.0001 and the learning rate for τ as 0.00005.

To implement the IS-based CI construction methods, for simplicity, we directly use the true behaviour policies. The open-

source code [1] is used to implement CoinDice. We use the default hyper-parameters, except for the following adjustments to get a better results for CoinDice. For CartPole, we set the learning rate as 0.005, batch size as 32, distribution regularizer as 0.05, neural network regularizers as 1, and set the neural networks as having one hidden layer of dimension 64. For Diabetes, we adjust the distribution regularizer as 2.5 and set the neural networks as having two hidden layers of dimension 256. In our experiments, we find Coindice is sensitive to these hyper-parameters, and tuned intensively to report results with the best combination.

C.4. Computational complexity

In this section, we analyze the computational complexity for the proposed value estimator $\hat{\eta}_{TR}^{(m)}$. The construction of the CI is straightforward and has the same complexity. Let $N = nT$ and let the dimension of the action plus that of the state be p. There are four main dominating parts of the computation: the calculation of $\hat{Q}$, $\hat{\omega}$, and $\hat{\omega}^*$, and the construction of the final estimator. For simplicity, we assume the standard dense networks with feedforward pass and back-propagation are used for the first three parts, and let the maxmium latent layer width and the depth for all the neural networks be w and d. For calculation of $\hat{Q}$, assume FQE converges in M_1 iterations, then according to the theory of neural networks, the complexity for the part is $O(NM_1w^dp)$. For calculation of $\hat{\omega}$ and $\hat{\omega}^*$, assume the training iterations of neural networks be M_2, then we have the complexity for these two part is $O(NM_2w^dp)$. For the last part, to calculate $\hat{\eta}_{TR}^{(m)}$, suppose we sample M_3 states from the reference distribution and use M_4 samples in the calculation of the incomplete U-statistics, the complexity is $O((M_3 + N)M_4)$. Putting the above results together, the total complexity for calculating $\hat{\eta}_{TR}^{(m)}$ and its CI is

$$O(nT(M_1 + M_2)w^dp + (M_3 + nT)M_4)$$

Note that the computation for the last part can be easily implemented in parallel, and for computing estimates of different order, the first three parts can be shared.

D. More on the CoinDice method

We discuss why CoinDice will fail to achieve valid CI estimation in this section. As we have commented in the introduction, CoinDice uses the empirical likelihood approach for interval estimation, assuming the data transactions are i.i.d. It is known that directly applying the empirical likelihood method without further adjustment will fail to handle weakly dependent data.

To elaborate this, let us consider a simple example. Given a sequence of stationary random variables $\{Z_t\}_{1\le t\le n}$, we aim to construct a CI for its mean. The CI based on the empirical likelihood method is given as follows

$$\{\mathbb{E}_{\mathbb{P}}Z : D_f(\mathbb{P}||\mathbb{P}_n) \le \rho/n\},$$

for some $\rho > 0$, where $\mathbb{P}_n$ denotes the empirical distribution of $\{Z_t\}_t$.

Here, the choice of ρ is essential to the validity of the resulting CI. When the observations $\{Z_t\}_t$ are i.i.d., one may set ρ to $\mathbb{P}(\chi_1^2 \le \rho) = 1 - \alpha$ for a given significance α. However, such a choice of ρ would fail with weakly dependent observations. More specifically, ρ shall be chosen such that

$$\mathbb{P}\left(\chi_1^2 \le \frac{\rho \mathrm{Var}(Z_1)}{\mathrm{Var}(Z_1) + 2\sum_{j=2}^{+\infty}\mathrm{cov}(Z_1, Z_j)}\right) = 1 - \alpha,$$

to ensure the validity of the resulting CI. See Theorem 5 and Theorem 11 of Duchi et al. (2016) for details.

When the observations are weakly dependent, the factor $\mathrm{Var}(Z_1)/\{\mathrm{Var}(Z_1) + 2\sum_{j=2}^{+\infty}\mathrm{cov}(Z_1, Z_j)\}$ is not equal to one in general. Consequently, directly applying the empirical likelihood method by assuming the data are i.i.d. will result in an invalid CI. CoinDice estimates the value via the marginalized important-sampling estimator instead of the doubly-robust estimator. As such, the summands in their estimator are positively corrected. The corresponding factor is much smaller than 1. Hence, applying CoinDice leads to a very narrow but invalid CI.

References

Brockman, G., Cheung, V., Pettersson, L., Schneider, J., Schulman, J., Tang, J., and Zaremba, W. Openai gym. *arXiv preprint arXiv:1606.01540*, 2016.

[1] https://github.com/google-research/dice_rl

Dedecker, J. and Louhichi, S. Maximal inequalities and empirical central limit theorems. In *Empirical process techniques for dependent data*, pp. 137–159. Springer, 2002.

Denker, M. and Keller, G. On u-statistics and v. mise'statistics for weakly dependent processes. *Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete*, 64(4):505–522, 1983.

Duchi, J., Glynn, P., and Namkoong, H. Statistics of robust optimization: A generalized empirical likelihood approach. *arXiv preprint arXiv:1610.03425*, 2016.

Le, H. M., Voloshin, C., and Yue, Y. Batch policy learning under constraints. *arXiv preprint arXiv:1903.08738*, 2019.

Shi, C., Wan, R., Song, R., Lu, W., and Leng, L. Does the markov decision process fit the data: testing for the markov property in sequential decision making. In *International Conference on Machine Learning*, pp. 8807–8817. PMLR, 2020.

Uehara, M., Huang, J., and Jiang, N. Minimax weight and q-function learning for off-policy evaluation. *arXiv preprint arXiv:1910.12809*, 2019.