Supplement to "Spatiotemporal Causal Effects Evaluation: A Multi-Agent Reinforcement Learning Framework"

Anonymous Author(s)
Affiliation
Address
email

A More on the learning procedure

A.1 Estimation of the weight

Consider the following optimization problem

$$\widehat{\omega}_i = \arg\min_{\omega_i \in \Omega} \sup_{f \in \mathcal{F}} \left| \sum_{t=0}^{T-1} \Delta_{i,t}(\omega_i) f(S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) \right|^2. \tag{1}$$

In our implementation, 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 $\mathcal{H}$ corresponds to an RFHS such that

$$\mathcal{H} = \left\{ f(\cdot) = \sum_{t=0}^{T-1} b_t \kappa(S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}; \cdot) : \{b_t\}_{t=0}^{T-1} \in \mathbb{R}^T \right\},$$

for some positive definite kernel $\kappa(\cdot;\cdot)$. Similar to Theorem 2 of [3], the optimization problem in (1) is then reduced to

$$\widehat{\omega}_i = \arg\min_{\omega_i \in \Omega} \sum_{t_1=0}^{T-1} \sum_{t_2=0}^{T-1} \Delta_{i,t_1}(\omega_i) \Delta_{i,t_2}(\omega_i) \kappa(S_{0,t_1+1}, S_{i,t_1+1}, \widetilde{S}_{i,t_1+1}; S_{0,t_2+1}, S_{i,t_2+1}, \widetilde{S}_{i,t_2+1}).$$

We set Ω to be the class of neural networks. One could use different parameters to factorize different ω_i's such that each $\widehat{\omega}_i$ is computed separately. Alternatively, one could share some parameters in common to estimate ω_i's jointly. We detail our procedure in Algorithm 1.

A.2 Estimation of the Q-function and the value

We now describe methods to estimate Q_i and $V_i(\boldsymbol{\pi})$. For two given function classes $\mathcal{G}$ and $\mathcal{Q}$, define the following penalized estimator

$$\begin{aligned} \widehat{g}_i(\cdot,\cdot,\cdot,\cdot,\cdot;\eta,Q_i) = \arg\min_{g \in \mathcal{G}} \frac{1}{T} \sum_{t=0}^{T-1} \{ & R_{i,t} + Q_i(\pi_i, \widetilde{A}_i(\boldsymbol{\pi}), S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) \\ & -\eta - Q_i(A_{i,t}, \widetilde{A}_{i,t}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) - g(A_{i,t}, \widetilde{A}_{i,t}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})\}^2 + \mu J_2^2(g), \end{aligned}$$

$$(\widehat{V}_i(\boldsymbol{\pi}), \widehat{Q}_i) = \arg\min_{(\eta, Q_i) \in \mathbb{R} \times \mathcal{Q}} \frac{1}{T} \sum_{t=0}^{T-1} \widehat{g}_i^2(A_{i,t}, \widetilde{A}_{i,t}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}; \eta, Q_i) + \lambda J_1^2(Q_i),$$

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Algorithm 1 Estimation of the weight.

Input: The data $\{(S_{0,j}, S_{i,j}, A_{i,j}, R_{i,j}, S_{0,j+1}, S_{i,j+1}) : 1 \le i \le N, 0 \le j < T\}$. A target policy $\boldsymbol{\pi}$.

Initial: Initial the density ratio $\omega_i = \omega_{i,\theta}$ for $1 \le i \le N$, to be neural networks parameterized by θ.

for iteration $= 1, 2, \cdots$ **do**

a Randomly sample a batch $\mathcal{M}$ from $\{0, 1, \cdots, T-1\}$.

b **Update** the parameter θ by $\theta \leftarrow \theta - \epsilon N^{-1} \sum_{i=1}^{N} \nabla_\theta D_i(\omega_{i,\theta}/z_{\omega_{i,\theta}})$ where $D_i(\omega_{i,\theta})$ is equal to

$$\frac{1}{|\mathcal{M}|} \sum_{t_1, t_2 \in \mathcal{M}} \Delta_{i,t_1}(\omega_{i,\theta}) \Delta_{i,t_2}(\omega_{i,\theta}) \kappa(S_{0,t_1+1}, S_{i,t_1+1}, \widetilde{S}_{i,t_1+1}; S_{0,t_2+1}, S_{i,t_2+1}, \widetilde{S}_{i,t_2+1}),$$

and $z_{\omega_{i,\theta}}$ is a normalization constant $z_{\omega_{i,\theta}} = |\mathcal{M}|^{-1} \sum_{t \in \mathcal{M}} \omega_{i,\theta}(S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1})$.

Output $\omega_{i,\theta}$ for $1 \le i \le N$.

where J_1 and J_2 denote some penalty functions, μ and λ stand for some tuning parameters. Next we derive the close-form expressions of $(\widehat{V}_i(\boldsymbol{\pi}), \widehat{Q}_i)$ when using RKHS to model Q_i and g_i.

Define vectors $Z_{i,t} = (A_{i,t}, \widetilde{A}_{i,t}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})^\top$ and $Z^*_{i,t} = (\pi_i, \widetilde{A}_i(\boldsymbol{\pi}), S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1})^\top$. Let K_g and K_Q denote the reproducing kernels used to model g and Q, respectively. In practice, we can use gaussian RBF kernels to model these two functions. For a given Q_i and η, the optimizer of $\widehat{g}_i$ can be represented by $\sum_{t=0}^{T-1} \widehat{\beta}_{i,t} K_g(Z_{i,t}, \cdot)$. As such, we obtain

$$\widehat{\boldsymbol{\beta}}_i = \arg\min_{\boldsymbol{\beta}} \frac{1}{T} \sum_{t=0}^{T-1} \left\{ R_{i,t} + Q_i(Z^*_{i,t}) - \eta - Q_i(Z_{i,t}) - \sum_{j=0}^{T-1} \beta_j K_g(Z_{i,j}, Z_{i,t}) \right\}^2 + \mu \boldsymbol{\beta}^\top \boldsymbol{K}_g \boldsymbol{\beta}$$

$$= \frac{1}{T} \boldsymbol{\beta}^\top \{\boldsymbol{K}_g \boldsymbol{K}_g^\top + T\mu \boldsymbol{K}_g\} \boldsymbol{\beta} - \frac{2}{T} \boldsymbol{\beta}^\top \boldsymbol{K}_g (\boldsymbol{R} + \boldsymbol{Q}^*_i - \boldsymbol{Q}_i - \eta \mathbf{1}) + \text{some terms that are independent of } \boldsymbol{\beta},$$

where $\boldsymbol{K}_g = \{K_g(Z_{i,j_1}, Z_{i,j_2})\}_{j_1,j_2}$ and $\boldsymbol{R}$, $\boldsymbol{Q}^*_i$ and $\boldsymbol{Q}_i$ the column vectors formed by elements in R_t, $Q_i(Z^*_{i,t})$ and $Q_i(Z_{i,t})$, respectively. Notice that $\boldsymbol{K}_g$ is symmetric, by some calculations, we obtain

$$\widehat{\boldsymbol{\beta}}_i = (\boldsymbol{K}_g \boldsymbol{K}_g^\top + T\mu \boldsymbol{K}_g)^{-1} \boldsymbol{K}_g (\boldsymbol{R} + \boldsymbol{Q}^*_i - \boldsymbol{Q}_i - \eta \mathbf{1}) = (\boldsymbol{K}_g + T\mu \boldsymbol{I})^{-1} (\boldsymbol{R} + \boldsymbol{Q}^*_i - \boldsymbol{Q}_i - \eta \mathbf{1}).$$

As a result, for a given Q_i and η, we have

$$\widehat{g}_i(Z_{i,t}; \eta, Q_i) = \widehat{\boldsymbol{\beta}}_i^\top \boldsymbol{K}_g \boldsymbol{e}_t,$$

where $\boldsymbol{e}_t$ denotes the column vector with the t-th element equals to one and other elements equal to zero. As such,

$$\frac{1}{T} \sum_{t=0}^{T-1} \widehat{g}_i^2(A_{i,t}, \widetilde{A}_{i,t}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}; \eta, Q_i) = \frac{1}{T} \widehat{\boldsymbol{\beta}}_i^\top \boldsymbol{K}_g \boldsymbol{K}_g^T \widehat{\boldsymbol{\beta}}_i.$$

Similarly, we can represent Q_i as $\sum_{t=0}^{2T-1} \widehat{\alpha}_{i,t} K_Q(\widetilde{Z}_{i,t}, \cdot)$ where $\widetilde{Z}_{i,t}$ denotes the t-th element in the vector $(Z_{i,0}^\top, Z_{i,1}^\top, \cdots, Z_{i,T-1}^\top, Z_{i,0}^{*\top}, \cdots, Z_{i,T-1}^{*\top})^\top$. Let $\boldsymbol{K}_Q$ denotes the corresponding $2T \times 2T$ matrix, we have

$$Q_i(Z_{i,t}) = \boldsymbol{\alpha}_i^\top \boldsymbol{K}_Q \boldsymbol{e}_t \quad \text{and} \quad Q_i(Z^*_{i,t}) = \widehat{\boldsymbol{\alpha}}_i^\top \boldsymbol{K}_Q \boldsymbol{e}_{t+T+1}.$$

It follow that

$$\boldsymbol{Q}^*_i - \boldsymbol{Q}_i = \underbrace{[-\boldsymbol{I}_T, \boldsymbol{I}_T]}_{\boldsymbol{C}} \boldsymbol{K}_Q \widehat{\boldsymbol{\alpha}}_i,$$

noting that $\boldsymbol{K}_Q$ is symmetric. Let $\boldsymbol{E} = \boldsymbol{K}_g^\top (\boldsymbol{K}_g + T\mu \boldsymbol{I})^{-1}$, $\widehat{\boldsymbol{\alpha}}_i$ corresponds to the solution of the following optimization problem,

$$\widehat{\boldsymbol{\alpha}}_i = \arg\min_{\boldsymbol{\alpha}} (\boldsymbol{R} + \boldsymbol{C}\boldsymbol{K}_Q \boldsymbol{\alpha} - \eta \mathbf{1})^\top \boldsymbol{E}^\top \boldsymbol{E} (\boldsymbol{R} + \boldsymbol{C}\boldsymbol{K}_Q \boldsymbol{\alpha} - \eta \mathbf{1}) + T\lambda \boldsymbol{\alpha}^\top \boldsymbol{K}_Q \boldsymbol{\alpha}.$$

Taking derivatives with respect to $\boldsymbol{\alpha}$ and η, we obtain

$$(\widehat{\boldsymbol{\alpha}}_i, \widehat{V}_i(\boldsymbol{\pi}))^\top = -([\boldsymbol{C}\boldsymbol{K}_Q, -\mathbf{1}]^\top \boldsymbol{E}^\top \boldsymbol{E} [\boldsymbol{C}\boldsymbol{K}_Q, -\mathbf{1}] + [T\lambda \boldsymbol{K}_Q, \mathbf{0}; \mathbf{0}^\top, 0])^{-1} [\boldsymbol{C}\boldsymbol{K}_Q, -\mathbf{1}] \boldsymbol{E}^\top \boldsymbol{E}\boldsymbol{R}.$$

A.3 Estimation of the treatment assignment probability

Note that $b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t}) = \mathbb{E}\{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))|S_{0,t},S_{i,t},\widetilde{S}_{i,t}\}$. It can thus be learned by applying machine learning algorithms to datasets with responses $\{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi})):0\le t<T\}$ and predictors $\{(S_{0,t},S_{i,t},\widetilde{S}_{i,t}):0\le t<T\}$.

B Additional technical conditions

C Proofs

We use c and C to denote some generic constants whose values are allowed to vary from place to place. Lemma 1 can thus be proven in a similar manner as Theorem 1 of [3]. Lemma 2 can be similarly proven as Lemma 1 of [4]. In the following, we focus on proving Theorems 1, 2 and 3.

C.1 Proof of Theorem 1

To prove Theorem 1, we apply the central limit theorem for mixing triangle arrays developed in [1]. Define

$$\widehat{V}_t^{\text{DR}*}(\boldsymbol{\pi})=\frac{1}{N}\sum_{i=1}^{N}\left[V_i^*(\boldsymbol{\pi})+\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{R_{i,t}+Q_{i,t+1}^*(\boldsymbol{\pi})-Q_{i,t}^*-V_i^*(\boldsymbol{\pi})\}\right],$$

we have $\widehat{V}^{\text{DR}*}(\boldsymbol{\pi})=T^{-1}\sum_{t=0}^{T-1}\widehat{V}_t^{\text{DR}*}(\boldsymbol{\pi})$.

Suppose we have shown each $\widehat{V}_t^{\text{DR}*}(\boldsymbol{\pi})$ is an unbiased estimator for $V(\boldsymbol{\pi})$. For $t\in\{0,1,\cdots,T-1\}$, let $x_t=(NT)^{-1/2}\{\widehat{V}_t^{\text{DR}*}(\boldsymbol{\pi})-V(\boldsymbol{\pi})\}$. It suffices to show the conditions in (1)-(5) of [1] hold for $\{x_t:0\le t<T\}$. We next verify these conditions.

Condition (1). Note that $\{R_{i,t},Q_i^*,\omega_i,V_i(\boldsymbol{\pi}):1\le i\le N,t\ge 0\}$ are uniformly bounded from infinity, the set of functions $\{b_i:1\le i\le N\}$ are uniformly bounded from zero. As such, $\{x_t:0\le t<T\}$ are uniformly bounded. Condition (1) thus holds for any $\nu^*>0$.

Condition (2). This condition is automatically implied by the assumption that $NT\text{Var}\{\widehat{V}^{\text{DR}*}(\boldsymbol{\pi})\}\to\sigma^2>0$.

Condition (3). This condition holds by setting $\kappa=0$ and $T_n=0$ for any n.

Condition (4). Note that the strong mixing coefficients are upper bounded by the β-mixing coefficients. Under Condition (A2), we can take the sequence $\alpha(h)$ in Condition (4) by $\kappa_0\rho^h$.

Condition (5). Since $\kappa_0\rho^h$ decays to zero at an exponential rate as h grows to infinity, Condition (5) is automatically satisfied.

It remains to show $\mathbb{E}\widehat{V}_t^{\text{DR}*}(\boldsymbol{\pi})=V(\boldsymbol{\pi})$ for any t. Suppose (A4) holds. Under the given conditions, we have $V_i^*(\boldsymbol{\pi})=V_i(\boldsymbol{\pi})$. Under Lemma 2, we have

$$\mathbb{E}\{R_{i,t}+Q_{i,t+1}^*(\boldsymbol{\pi})-Q_{i,t}^*-V_i^*(\boldsymbol{\pi})|\boldsymbol{A}_t,\boldsymbol{S}_t\}=0,$$

and hence,

$$\mathbb{E}\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{R_{i,t}+Q_{i,t+1}^*(\boldsymbol{\pi})-Q_{i,t}^*-V_i^*(\boldsymbol{\pi})\}=0.$$

Consequently, $\mathbb{E}\widehat{V}_t^{\text{DR}*}(\boldsymbol{\pi})=N^{-1}\sum_{i=1}^N V_i(\boldsymbol{\pi})=V(\boldsymbol{\pi})$.

Suppose (A3)(i) holds. For any i,t, the expectation of the density ratio $\omega_{i,t}\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))/b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})$ equals one. As such, we have

$$\begin{aligned}&\mathbb{E}\left\{V_i^*(\boldsymbol{\pi})-\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}V_i^*(\boldsymbol{\pi})\right\}\\&=V_i^*(\boldsymbol{\pi})\mathbb{E}\left\{1-\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\right\}=0.\end{aligned}\tag{2}$$

In addition, using similar arguments in (2), we have by (A3)(i) that

$$\mathbb{E}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}R_{i,t}\right\}=V_i(\boldsymbol{\pi}). \tag{3}$$

Moreover, by some calculations, we have

$$\begin{aligned}\mathbb{E}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}Q^*_{i,t}\right\} &= \mathbb{E}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}Q^*_{i,t+1}(\boldsymbol{\pi})\right\}\\ &= \int_{s_0,s_i,\tilde{s}_i} Q^*_i(\pi_i,\widetilde{A}_i(\boldsymbol{\pi}),s_0,s_i,\tilde{s}_i)p(\boldsymbol{\pi},s_0,s_i,\tilde{s}_i)ds_0ds_id\tilde{s}_i.\end{aligned}$$

Consequently,

$$\mathbb{E}\left[\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{Q^*_{i,t+1}(\boldsymbol{\pi})-Q^*_{i,t}\}\right]=0.$$

This together with (2) and (3) yields

$$\mathbb{E}\left[V_i^*(\boldsymbol{\pi})+\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{R_{i,t}+Q^*_{i,t+1}(\boldsymbol{\pi})-Q^*_{i,t}-V_i^*(\boldsymbol{\pi})\}\right]=V_i(\boldsymbol{\pi}).$$

It follows that $\mathbb{E}\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})=V(\boldsymbol{\pi})$.

Thus, $\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$ is unbiased when either (A3)(i) or (A4) holds. The proof is hence completed.

C.2 Proof of Theorem 2

By Theorem 1, it suffices to show $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$ is asymptotically equivalent to $\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$. Note that $\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})-\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$ can be decomposed by $\eta_1+\eta_2+\eta_3+\eta_4+\eta_5$ where

$$\begin{aligned}\eta_1 &= \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}-1\right\}\{V_i^*(\boldsymbol{\pi})-\widehat{V}_i(\boldsymbol{\pi})\},\\ \eta_2 &= \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi})-\widehat{Q}_{i,t}-Q^*_{i,t+1}(\boldsymbol{\pi})+Q^*_{i,t}\},\\ \eta_3 &= \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t}-\omega_{i,t})\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{R_{i,t}+Q^*_{i,t+1}(\boldsymbol{\pi})-Q^*_{i,t}-V_i^*(\boldsymbol{\pi})\},\\ \eta_4 &= \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t}-\omega_{i,t})\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi})-\widehat{Q}_{i,t}-Q^*_{i,t+1}(\boldsymbol{\pi})+Q^*_{i,t}\},\\ \eta_5 &= \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t}-\omega_{i,t})\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}\{V_i^*(\boldsymbol{\pi})-\widehat{V}_i(\boldsymbol{\pi})\}.\end{aligned}$$

In the following, we provide upper bounds on each $|\eta_j|$, for $j=1,2,\cdots,5$.

Upper bounds on $|\eta_1|$: Note that $\eta_1=N^{-1}\sum_{i=1}^N\eta_{1,i}$ where

$$\eta_{1,i}=\{V_i^*(\boldsymbol{\pi})-\widehat{V}_i(\boldsymbol{\pi})\}\left[\frac{1}{T}\sum_{t=0}^{T-1}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}-1\right\}\right].$$

The expectation of the density ratio equals one. As a result, we have

$$\mathbb{E}\left\{\omega_{i,t}\frac{\mathbb{I}(A_{i,t}=\pi_i,\widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t},\widetilde{S}_{i,t})}-1\right\}=0,$$

for any i, t. In the following, we apply the

Under Condition (A2), the β-mixing coefficients of the sequence

$$\left\{ \omega_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi} | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - 1 : t \geq 0 \right\}, \tag{4}$$

decays to zero at an exponential rate. In addition, all the terms in (4) are uniformly bounded. It follows from Corollary 3.3 in [2] that

$$\mathbb{P}\left(\left| \sum_{t=0}^{T-1} \left\{ \omega_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi} | S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - 1 \right\} \right| \geq \epsilon \right) \leq c \exp\left(- \frac{CT\epsilon}{\log T \log\log T} \right),$$

for some constants $c, C > 0$.

Upper bounds on $|\eta_2|$:

Upper bounds on $|\eta_3|$:

Upper bounds on $|\eta_4|$ and $|\eta_5|$:

References

[1] Christian Francq and Jean-Michel Zakoïan. A central limit theorem for mixing triangular arrays of variables whose dependence is allowed to grow with the sample size. *Econometric Theory*, 21(6):1165–1171, 2005.

[2] Johannes T. N. Krebs. A large deviation inequality for β-mixing time series and its applications to the functional kernel regression model. *Statist. Probab. Lett.*, 133:50–58, 2018.

[3] Qiang Liu, Lihong Li, Ziyang Tang, and Dengyong Zhou. Breaking the curse of horizon: Infinite-horizon off-policy estimation. In *Advances in Neural Information Processing Systems*, pages 5356–5366, 2018.

[4] Chengchun Shi, Xiaoyu Wang, Shikai Luo, Rui Song, Hongtu Zhu, and Jieping Ye. A reinforcement learning framework for time-dependent causal effects evaluation in a/b testing. *arXiv preprint arXiv:2002.01711*, 2020.