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

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

for some positive definite kernel $\kappa(\cdot;\cdot)$. Similar to Theorem 2 of [6], 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 the class of neural networks. One could use different parameters to factorize different ω_i such that each $\widehat{\omega}_i$ is computed separately. Alternatively, one could allow different ω_i to share some common parameters. 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 RKHS is used 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

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

Note 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 and lemmas

B.1 Technical conditions

Let $Q^*_{i,\boldsymbol{\pi}}$ denote the function such that $Q^*_{i,\boldsymbol{\pi}}(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) = Q^*_i(\pi_i, \widetilde{A}_i(\boldsymbol{\pi}), S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$ almost surely for any t and i. Similarly, let $\widehat{Q}_{i,\boldsymbol{\pi}}$ denote the function such that $\widehat{Q}_{i,\boldsymbol{\pi}}(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) = \widehat{Q}_i(\pi_i, \widetilde{A}_i(\boldsymbol{\pi}), S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$ almost surely for any t and i.

(A5)(i) $\sum_{i=1}^N |V_i^*(\boldsymbol{\pi}) - \widehat{V}_i(\boldsymbol{\pi})|/N = o_p(1)$; (ii) $\widehat{Q}_{i,\boldsymbol{\pi}} \in \mathcal{Q}$, $\widehat{\omega}_i \in \mathcal{W}$ almost surely for any i. $\mathcal{Q}$ and $\mathcal{W}$ satisfy $\sup_Q N(\mathcal{Q}, e_Q, \varepsilon\|F\|_{Q,2}) \le (A/\varepsilon)^\nu$, $\sup_Q N(\mathcal{W}, e_Q, \varepsilon\|F\|_{Q,2}) \le (A/\varepsilon)^\nu$ for some $e \le A = O(1)$, $\nu = O(NT)$, and their envelope functions are bounded by some constant M. (iii) $\max_i |\widehat{Q}_{i,\boldsymbol{\pi}}(s_0, s_i, \widetilde{s}_i) - Q^*_{i,\boldsymbol{\pi}}(s_0, s_i, \widetilde{s}_i)|^2 p(b, s_0, s_i, \tilde{s}_i) ds_0 ds_i d\tilde{s}_i = o_p(1)$, $\max_i |\widehat{\omega}_i(s_0, s_i, \widetilde{s}_i) - \omega^*_i(s_0, s_i, \widetilde{s}_i)|^2 p(b, s_0, s_i, \tilde{s}_i) ds_0 ds_i d\tilde{s}_i = o_p(1)$.

(A6)(i) $\max_i |V_i^*(\boldsymbol{\pi}) - \widehat{V}_i(\boldsymbol{\pi})|^2 = o_p((NT)^{-1/2})$; (ii) $\max_i |\widehat{Q}_{i,\boldsymbol{\pi}}(s_0, s_i, \widetilde{s}_i) - Q^*_{i,\boldsymbol{\pi}}(s_0, s_i, \widetilde{s}_i)|^2 p(b, s_0, s_i, \tilde{s}_i) ds_0 ds_i d\tilde{s}_i = o_p((NT)^{-1/2})$; (iii) $\max_i |\widehat{\omega}_i(s_0, s_i, \widetilde{s}_i) - \omega^*_i(s_0, s_i, \widetilde{s}_i)|^2 p(b, s_0, s_i, \tilde{s}_i) ds_0 ds_i d\tilde{s}_i = o_p((NT)^{-1/2})$; (iv) $T \gg N\nu^2 \log^4(NT)$.

B.2 An auxiliary lemma

We briefly introduce our setup before presenting the lemma. Let $\{Z_t : t \ge 0\}$ be a stationary β-mixing process whose β-mixing coefficients are given by $\{\beta(q) : q \ge 0\}$. Let $\mathcal{F}$ be a pointwise measurable class of functions that take Z_t as input with a measurable envelope function F. For any $f \in \mathcal{F}$, suppose $\mathbb{E}f(Z_0) = 0$. Let $\sigma^2 > 0$ be a positive constant such that $\sup_{f\in\mathcal{F}} \mathbb{E}f^2(Z_0) \le \sigma^2 \le \mathbb{E}F^2(Z_0)$. In the following, we focus providing an exponential inequality for the empirical process $\sup_{f\in\mathcal{F}} |\sum_{t=0}^{T-1} f(Z_t)|$.

Lemma B.1 *Suppose the envelop function is uniformly bounded by some constant $M > 0$. In addition, suppose $\mathcal{F}$ belongs to the class of VC-type class such that $\sup_Q N(\mathcal{F}, e_Q, \varepsilon\|F\|_{Q,2}) \le (A/\varepsilon)^\nu$ [see Definition 2.1 in 3, for details] for some $A \ge e, \nu \ge 1$. Then there exist some constants $c, C > 0$ such that*

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{T-1} f(Z_t)\right| > c\sqrt{\nu q \sigma^2 T \log\left(\frac{AM}{\sigma}\right)} + c\nu M \log\left(\frac{AM}{\sigma}\right) + cq\tau + Mq\right) \le Cq\exp\left(-\frac{\tau^2 q}{CT\sigma^2}\right) + Cq\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},$$

for any $\tau > 0$, $1 \le q < T/2$.

C Proofs

We use c and C to denote some generic constants whose values are allowed to vary from place to place. For any two positive sequences $\{a_t\}_{t\ge1}$ and $\{b_t\}_{t\ge1}$, we write $a_t \preceq b_t$ if there exists some constant $C > 0$ such that $a_t \le Cb_t$ for any t. The notation $a_t \preceq 1$ means $a_t = O(1)$.

Lemma 1 can thus be proven in a similar manner as Theorem 1 of [6]. Lemma 2 can be similarly proven as Lemma 1 of [7]. Theorem 2 can be proven in a similar manner as Theorem 3. In the following, we focus on proving Theorems 1, 3 and Lemma B.1.

C.1 Proof of Theorem 1

To prove Theorem 1, we apply the central limit theorem for mixing triangle arrays developed in [5]. 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 [5] 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 $V_i^*(\boldsymbol{\pi}) = V_i(\boldsymbol{\pi})$. By 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) holds. Then we have $\omega_{i,t}^* = \omega_{i,t}$ for any i, t where $\omega_{i,t}$ is a shorthand for $\omega_i(\boldsymbol{\pi}, S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})$. As a result, 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) 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

$$\begin{aligned}
\mathbb{E}\left\{\omega_{i,t}^* \frac{\mathbb{I}(A_{i,t}=\pi_i, \widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} Q_{i,t}^*\right\} &= \mathbb{E}\left\{\omega_{i,t}^* \frac{\mathbb{I}(A_{i,t}=\pi_i, \widetilde{A}_{i,t}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} Q_{i,t+1}^*(\boldsymbol{\pi})\right\} \\
&= \int_{s_0, s_i, \tilde{s}_i} Q_i^*(\pi_i, \widetilde{A}_i(\boldsymbol{\pi}), s_0, s_i, \tilde{s}_i) p(\boldsymbol{\pi}, s_0, s_i, \tilde{s}_i) ds_0 ds_i d\tilde{s}_i.
\end{aligned}$$

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) or (A4) holds. The proof is hence completed.

C.2 Proof of Theorem 3

By Theorem 1, it suffices to show $\sqrt{NT}\widehat{V}^{\mathrm{DR}}(\boldsymbol{\pi})$ is asymptotically equivalent to $\sqrt{NT}\widehat{V}^{\mathrm{DR}*}(\boldsymbol{\pi})$. 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 show $|\eta_j|=o_p((NT)^{-1/2})$, 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

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

When (A3) holds, we have $\omega^*_{i,t}=\omega_{i,t}$ for any i,t. 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 Bernstein's inequality for exponential β-mixing processes [2] to bound $|\eta_1|$.

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 by some constant $c>0$. As a result,

$$\max_{t_1,t_2}\mathbb{E}\left|\omega_{i,t_1}\frac{\mathbb{I}(A_{i,t_1}=\pi_i,\widetilde{A}_{i,t_1}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t},S_{i,t_1},\widetilde{S}_{i,t_1})}-1\right|\left|\omega_{i,t_2}\frac{\mathbb{I}(A_{i,t_2}=\pi_i,\widetilde{A}_{i,t_2}=\widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t_2},S_{i,t_2},\widetilde{S}_{i,t_2})}-1\right|=O(1).$$

It thus follows from Theorem 4.2 of [2] that there exists some constant $C > 0$ such that there exists some constant $C > 0$ such that for any $\tau \geq 0$ and integer $1 < q < T$,

$$\max_i \mathbb{P}\left(\left| \sum_{t=0}^{T-1} \left\{ \omega_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - 1 \right\} \right| \geq 6\tau \right) \leq \frac{T}{q}\beta(q)$$
$$+ \quad \max_i \mathbb{P}\left(\left| \sum_{t\in\mathcal{I}_r} \left\{ \omega_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - 1 \right\} \right| \geq \tau \right) + 4\exp\left\{ -\frac{\tau^2}{Cq(T+\tau)} \right\} \tag{5}$$

where $\mathcal{I}_r = \{q\lfloor T/q \rfloor, q\lfloor T/q \rfloor + 1, \cdots, T-1\}$. Suppose $\tau \geq qc$. Notice that $|\mathcal{I}_r| \leq q$. It follows that

$$\max_i \mathbb{P}\left(\left| \sum_{t\in\mathcal{I}_r} \left\{ \omega_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - 1 \right\} \right| \geq \tau \right) = 0. \tag{6}$$

Under (A2), $\beta(q) = O(\rho^q)$. Set $q = -3\log(NT)/\log\rho$, we obtain $T\beta(q)/q = O(N^{-3}T^{-2})$. Set $\tau = \max\{2\sqrt{CqT\log(NT)}, 4Cq\log(NT)\}$, we obtain as $T \to \infty$ that

$$\frac{\tau^2}{2} \geq 2CqT\log(NT) \quad \text{and} \quad \frac{\tau^2}{2} \geq 2Cq\tau\log(nT) \quad \text{and} \quad \tau \geq qc.$$

Since $\sqrt{CqT\log(NT)} \gg 2Cq\log(NT)$, it follows from (5) and (6) that

$$\max_i \mathbb{P}\left(\left| \sum_{t=0}^{T-1} \left\{ \omega_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - 1 \right\} \right| \geq 12\sqrt{CqT\log(NT)} \right) \preceq N^{-2}T^{-2}.$$

By Bonferroni's inequality, we obtain the following event occurs with probability at least $1 - O(N^{-1}T^{-1})$,

$$\max_i \left| \sum_{t=0}^{T-1} \left\{ \omega_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} - 1 \right\} \right| \leq 12\sqrt{CqT\log(NT)}.$$

It follows that

$$|\eta_1| \leq \frac{1}{N}\sum_{i=1}^N |\eta_{1,i}| \preceq \frac{\log(NT)}{\sqrt{T}} \left(\frac{1}{N}\sum_{i=1}^N |V_i^*(\boldsymbol{\pi}) - \widehat{V}_i(\boldsymbol{\pi})| \right), \tag{7}$$

with probability approaching 1. Under (A6) and the condition that $T \gg N\log^4(NT)$, we obtain $\eta_1 = o_p((NT)^{-1/2})$.

Upper bounds on $|\eta_2|$: When (A3) holds, we have $\omega^*_{i,t} = \omega_{i,t}$ for any i and t. As discussed in the proof of Theorem 1, we have $\mathbb{E}\eta_{2,i} = 0$ for any i where

$$\eta_{2,i} = \frac{1}{T}\sum_{t=0}^{T-1} \omega^*_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} \{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t} - Q^*_{i,t+1}(\boldsymbol{\pi}) + Q^*_{i,t}\}.$$

In addition, notice that $\eta_{2,i}$ can be written as

$$\eta_{2,i} = \frac{1}{T}\sum_{t=0}^{T-1} \omega^*_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} \{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - Q^*_{i,t+1}(\boldsymbol{\pi}) + Q^*_{i,t}(\boldsymbol{\pi})\}.$$

We apply Lemma B.1 to bound $\max_i |\eta_{2,i}|$. Define the class of functions $\mathcal{Q}_{i,\varepsilon}$ by

$$\left\{ f \in \mathcal{Q} : \max_i \int_{s_0,s_i,\widetilde{s}_i} |f(s_0, s_i, \widetilde{s}_i) - Q^*_{i,\boldsymbol{\pi}}(s_0, s_i, \widetilde{s}_i)|^2 p(b, s_0, s_i, \tilde{s}_i) ds_0 ds_i d\tilde{s}_i \leq \varepsilon \right\},$$

where $\varepsilon = \epsilon N^{-1/2}T^{-1/2}$ for some sufficiently small $\epsilon > 0$. It then follows from (A5)(ii) and (iii) that $\widehat{Q}_{i,\boldsymbol{\pi}} \in \mathcal{Q}_\varepsilon$ for any i with probability tending to 1. As such, we have

$$\eta_{2,i} \leq T^{-1} \sup_{Q_i \in \mathcal{Q}_{i,\varepsilon}} \left| \sum_{t=0}^{T-1} \omega^*_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})} \{f(S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) \right.$$
$$\left. - f(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) - Q^*_{i,t+1}(\boldsymbol{\pi}) + Q^*_{i,t}(\boldsymbol{\pi})\} \right|.$$

Consider the process $\{(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}, A_{i,t}, \widetilde{A}_{i,t}, S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) : t \geq 0\}$. Under (A2), such a process has β-mixing coefficients $\{\beta^*(q) : q \geq 0\}$ that satisfies $\beta^*(q) = O(\rho^q)$ as well. For any f, define the function $g = g(f)$ such that

$$g(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}, A_{i,t}, \widetilde{A}_{i,t}, S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) = \omega^*_{i,t} \frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})}$$
$$\times \quad \{f(S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) - f(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}) - Q^*_{i,t+1}(\boldsymbol{\pi}) + Q^*_{i,t}(\boldsymbol{\pi})\},$$

almost surely. Consider the class of functions $\mathcal{G}_{i,\varepsilon} = \{g(f) : f \in \mathcal{Q}_{i,\varepsilon}\}$. Since $\mathcal{Q}_{i,\varepsilon}$ belongs to the class of VC-type class, so does $\mathcal{G}_{i,\varepsilon}$. Moreover, the VC-index of $\mathcal{G}_{i,\varepsilon}$ is the same as $\mathcal{Q}_{i,\varepsilon}$. Under the boundedness assumption in Theorem 2, we have

$$\mathbb{E} g^2(S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}, A_{i,t}, \widetilde{A}_{i,t}, S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1}) \leq O(1)\varepsilon,$$

for some constant $O(1)$. In addition, the envelope function of $\mathcal{G}_{i,\varepsilon}$ is uniformly bounded.

Let $Z_{i,t} = (S_{0,t}, S_{i,t}, \widetilde{S}_{i,t}, A_{i,t}, \widetilde{A}_{i,t}, S_{0,t+1}, S_{i,t+1}, \widetilde{S}_{i,t+1})$. Applying Lemma B.1, we obtain

$$\max_i \mathbb{P}\left(\sup_{g \in \mathcal{G}_{i,\varepsilon}} \left| \sum_{t=0}^{T-1} g(Z_{i,t}) \right| > c\sqrt{\nu q \varepsilon T \log\left(\frac{1}{\varepsilon}\right)} + c\nu \log\left(\frac{1}{\varepsilon}\right) + cq\tau + cq \right)$$
$$\leq cq \exp\left(-\frac{\tau^2 q}{cT\varepsilon}\right) + cq\exp\left(-\frac{\tau}{c}\right) + \frac{T\beta(q)}{q},$$

for some constant $c > 0$. Set $q = -2\log(NT)/\log\rho$, we have $T\beta(q)/q = O(N^{-2}T^{-1})$. Set $\tau = \max(2c\log(NT), \sqrt{2c\varepsilon T \log(NT)/q})$, the RHS is bounded by $O(N^{-2}T^{-1}\log(NT))$. By Bonferroni's inequality, we obtain with probability tending to 1 that

$$T|\eta_{2,i}| \leq c\sqrt{\nu q \varepsilon T \log\left(\frac{1}{\varepsilon}\right)} + c\nu\log\left(\frac{1}{\varepsilon}\right) + cq\tau + cq, \quad \forall i \in \{1, \cdots, N\},$$

or equivalently,

$$\max_i |\eta_{2,i}| \preceq \sqrt{\frac{\epsilon}{NT}} + o\left(\frac{1}{\sqrt{NT}}\right),$$

under the condition that $T \gg N\nu^2 \log^4(NT)$. Since ϵ can be chosen arbitrarily small, we obtain $\max_i |\eta_{2,i}| = o_p((NT)^{-1/2})$. This in turn implies $\eta_2 = o_p((NT)^{-1/2})$.

Upper bounds on $|\eta_3|$: Using similar arguments in proving $\eta_2 = o_p((NT)^{-1/2})$, we can show $\eta_3 = o_p((NT)^{-1/2})$. We omit the technical details to save space.

Upper bounds on $|\eta_4|$ and $|\eta_5|$: We show $\eta_4 = o_p((NT)^{-1/2})$ only. Using similar arguments, one can show $\eta_5 = o_p((NT)^{-1/2})$.

Note that

$$\eta_4 = \frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t} - \omega^*_{i,t})\frac{\mathbb{I}(A_{i,t} = \pi_i, \widetilde{A}_{i,t} = \widetilde{A}_i(\boldsymbol{\pi}))}{b_i(\boldsymbol{\pi}|S_{0,t}, S_{i,t}, \widetilde{S}_{i,t})}\{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - Q^*_{i,t+1}(\boldsymbol{\pi}) + Q^*_{i,t}(\boldsymbol{\pi})\}$$
$$\leq O(1)\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}|\widehat{\omega}_{i,t} - \omega^*_{i,t}||\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - Q^*_{i,t+1}(\boldsymbol{\pi}) + Q^*_{i,t}(\boldsymbol{\pi})|$$
$$\leq O(1)\left\{\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}[(\widehat{\omega}_{i,t} - \omega^*_{i,t})^2 + \{\widehat{Q}_{i,t+1}(\boldsymbol{\pi}) - \widehat{Q}_{i,t}(\boldsymbol{\pi}) - Q^*_{i,t+1}(\boldsymbol{\pi}) + Q^*_{i,t}(\boldsymbol{\pi})\}^2]\right\}$$
$$\leq O(1)\left\{\frac{1}{NT}\sum_{t=0}^{T-1}\sum_{i=1}^{N}(\widehat{\omega}_{i,t} - \omega^*_{i,t})^2\right\} + O(1)\left\{\frac{1}{NT}\sum_{t=0}^{T}\sum_{i=1}^{N}\{\widehat{Q}_{i,t}(\boldsymbol{\pi}) - Q^*_{i,t}(\boldsymbol{\pi})\}^2\right\},$$

where $O(1)$ denotes some universal constant, and the last two inequalities are due to Cauchy-Schwarz inequality.

To prove $\eta_4 = o_p((NT)^{-1/2})$, it suffices to show

$$\max_i \left[\frac{1}{T} \sum_{t=0}^{T} \{\widehat{Q}_{i,t}(\boldsymbol{\pi}) - Q^*_{i,t}(\boldsymbol{\pi})\}^2 \right] = o_p((NT)^{-1/2}), \tag{8}$$

and

$$\max_i \left\{ \frac{1}{T} \sum_{t=0}^{T-1} (\widehat{\omega}_{i,t} - \omega^*_{i,t})^2 \right\} = o_p((NT)^{-1/2}). \tag{9}$$

The left-hand-side (LHS) of (8) can be upper bounded by

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left[\frac{1}{T} \sum_{t=0}^{T} \{f(Z_{i,t}) - Q^*_{i,t}(\boldsymbol{\pi})\}^2 \right],$$

with probability tending to 1. Using similar arguments in proving $\eta_2 = o_p((NT)^{-1/2})$, we can show

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left| \frac{1}{T} \sum_{t=0}^{T} \{f(Z_{i,t}) - Q^*_{i,t}(\boldsymbol{\pi})\}^2 - \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}\{f(Z_{i,t}) - Q^*_{i,t}(\boldsymbol{\pi})\}^2 \right| \preceq \frac{\epsilon}{\sqrt{NT}} + o\left(\frac{1}{\sqrt{NT}}\right),$$

with probability tending to 1. Under (A6), we have

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left| \frac{1}{T} \sum_{t=0}^{T} \mathbb{E}\{f(Z_{i,t}) - Q^*_{i,t}(\boldsymbol{\pi})\}^2 \right| \preceq \frac{\epsilon}{\sqrt{NT}}.$$

It follows that

$$\max_i \sup_{f \in \mathcal{Q}_{i,\varepsilon}} \left[\frac{1}{T} \sum_{t=0}^{T} \{f(Z_{i,t}) - Q^*_{i,t}(\boldsymbol{\pi})\}^2 \right] \preceq \frac{\epsilon}{\sqrt{NT}} + o\left(\frac{1}{\sqrt{NT}}\right),$$

with probability tending to 1. Let $\epsilon \to 0$, we obtain (8). Similarly, we can show (9) holds. The proof is hence completed.

C.3 Proof of Lemma B.1

We break the proof into three steps. In the first step, we use Berbee's coupling lemma [see Lemma 4.1 in 4] to approximate $\sup_{f \in \mathcal{F}} |\sum_{t=0}^{T-1} f(Z_t)|$ by sum of i.i.d. variables. In the second step, we apply the tail inequality in Lemma 1 of [1] to bound the derivation between the empirical process and its mean. Finally, we apply the maximal inequality in Corollary 5.1 of [3] to bound the expectation of the empirical process.

Step 1. Following the discussion below Lemma 4.1 of [4], we can construct a sequence of random variables $\{Z_t^0 : t \geq 0\}$ such that

$$\sup_{f \in \mathcal{F}} \left| \sum_{t=0}^{T-1} f(Z_t) \right| = \sup_{f \in \mathcal{F}} \left| \sum_{t=0}^{T-1} f(Z_t^0) \right|, \tag{10}$$

with probability at least $1 - T\beta(q)/q$, and that the sequences $\{U_{2i}^0 : i \geq 0\}$ and $\{U_{2i+1}^0 : i \geq 0\}$ are i.i.d. where $U_i^0 = (Z_{iq}^0, Z_{iq+1}^0, \cdots, Z_{iq+q-1}^0)$.

Recall that $\mathcal{I}_r = \{q\lfloor T/q \rfloor, q\lfloor T/q \rfloor + 1, \cdots, T-1\}$, we have

$$\sup_{f \in \mathcal{F}} \left| \sum_{t=0}^{T-1} f(Z_t^0) \right| \leq \sum_{j=0}^{q-1} \sup_{f \in \mathcal{F}} \left| \sum_{t=0}^{\lfloor T/q \rfloor} f(Z_{tq+j}^0) \right| + \sup_{f \in \mathcal{F}} \left| \sum_{t \in \mathcal{I}_r} f(Z_t^0) \right|.$$

Under the boundedness assumption on F, the second term on the right-hand-side (RHS) is bounded from above by Mq. Without loss of generality, suppose $\lfloor T/q \rfloor$ is an even number. The first term on the RHS can be bounded from above by $\sum_{j=0}^{2q-1} \sup_{f\in\mathcal{F}} |\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq+j})|$. To summarize, we have shown

$$\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{T-1} f(Z_t^0)\right| \le \sum_{j=0}^{2q-1} \sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq+j})\right| + Mq.$$

This together with (10) yields that

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{T-1} f(Z_t)\right| > 2\tau q + Mq\right) \le \mathbb{P}\left(\sum_{j=0}^{2q-1} \sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq+j})\right| > 2\tau q\right) + \frac{T\beta(q)}{q}, \tag{11}$$

for any $\tau > 0$. By Bonferroni's inequality, we obtain

$$\mathbb{P}\left(\sum_{j=0}^{2q-1} \sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq+j})\right| > 2\tau q\right) \le \sum_{j=0}^{2q-1} \mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq+j})\right| > \tau\right),$$

for any $\tau > 0$. Since the process is stationary, we obtain

$$\mathbb{P}\left(\sum_{j=0}^{2q-1} \sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq+j})\right| > 2\tau q\right) \le 2q\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| > \tau\right).$$

Combining this together with (11) yields

$$\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{T-1} f(Z_t)\right| > 2\tau q + Mq\right) \le 2q\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| > \tau\right) + \frac{T\beta(q)}{q}. \tag{12}$$

By construction, $\{Z^0_{2tq} : t \ge 0\}$ are i.i.d. This completes the proof of the first step.

Step 2. In the second step, we focus on relating the empirical process $\sup_{f\in\mathcal{F}} |\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})|$ to its expectation. Without loss of generality, assume $T = kq$ for some integer $k > 0$. Set the constants η and δ in Lemma 1 of [1] to 1, we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| > 2\mathbb{E}\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| + \tau\right) \\ \le 4\exp\left(-\frac{\tau^2}{2T\sigma^2/q}\right) + \exp\left(-\frac{\tau}{CM}\right),\end{aligned}$$

for some constant $C > 0$. Combining this together with (12), we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{T-1} f(Z_t)\right| > 4q\mathbb{E}\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| + 2\tau q + Mq\right) \\ \le 8q\exp\left(-\frac{\tau^2}{2T\sigma^2/q}\right) + 2q\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},\end{aligned} \tag{13}$$

for any $\tau > 0$. This completes the proof of the second step.

Step 3. It remains to bound $\mathbb{E}\sup_{f\in\mathcal{F}} |\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})|$. By Corollary 5.1 of [3], we obtain

$$\mathbb{E}\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{\lfloor T/(2q)\rfloor} f(Z^0_{2tq})\right| \preceq \sqrt{\frac{\nu\sigma^2 T}{q}\log\left(\frac{AM}{\sigma}\right)} + \nu M\log\left(\frac{AM}{\sigma}\right).$$

Combining this together with (13), we obtain

$$\begin{aligned}\mathbb{P}\left(\sup_{f\in\mathcal{F}} \left|\sum_{t=0}^{T-1} f(Z_t)\right| > c\sqrt{\nu q\sigma^2 T\log\left(\frac{AM}{\sigma}\right)} + c\nu M\log\left(\frac{AM}{\sigma}\right) + cq\tau + Mq\right) \\ \le Cq\exp\left(-\frac{\tau^2 q}{CT\sigma^2}\right) + Cq\exp\left(-\frac{\tau}{CM}\right) + \frac{T\beta(q)}{q},\end{aligned}$$

for some constants $c, C > 0$ and any $\tau > 0, 1 \le q < T/2$. The proof is hence completed.

References

[1] Radosł aw Adamczak. A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.*, 13:no. 34, 1000–1034, 2008.

[2] Xiaohong Chen and Timothy M. Christensen. Optimal uniform convergence rates and asymptotic normality for series estimators under weak dependence and weak conditions. *J. Econometrics*, 188(2):447–465, 2015.

[3] Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Gaussian approximation of suprema of empirical processes. *Ann. Statist.*, 42(4):1564–1597, 2014.

[4] Jérôme Dedecker and Sana Louhichi. Maximal inequalities and empirical central limit theorems. In *Empirical process techniques for dependent data*, pages 137–159. Birkhäuser Boston, Boston, MA, 2002.

[5] 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.

[6] 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.

[7] 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.