

# Supplementary Materials of “Higher-Order Debiased Estimators for General Treatment Models”

Yulin Zhang<sup>1</sup>, Lin Liu<sup>2</sup>, Zheng Zhang<sup>1</sup>

<sup>1</sup>Institute of Statistics & Big Data, Renmin University of China; <sup>2</sup>Institute of Natural Sciences, MOE–LSC, School of Mathematical Sciences, CMA–Shanghai, SJTU–Yale Joint Center for Biostatistics and Data Science, Shanghai Jiao Tong University

## S1 Comparisons of General Treatment Models and Marginal Structural Models

In this section, we further discuss the similarities and differences between the *General Treatment Models (GTM)* (Ai et al., 2021) and the *Marginal Structural Models (MSM)* (See Example 6), a well-known class of semiparametric causal models developed by Robins and colleagues (Robins et al., 2000). MSM was developed, in part, to overcome the curse-of-dimensionality when drawing causal conclusions from interventional or observational studies when the treatment regimes under study are complex. In its simplest form, MSM directly models the counterfactual mean  $\mathbb{E}Y(t)$  as a parameterized function  $g(t; \beta^*)$  of  $t \in \mathcal{T}$  with  $\beta^* := (\beta_0^*, \beta_1^*)^\top$ .

When  $\mathcal{T} = \{0, 1\}$ ,  $g$  can only be a linear model, and this linear model is saturated/nonparametric: MSM (3) coincides with GTM in Example 1. For example, one can posit the following nonparametric (or saturated) linear MSM:

$$\mathbb{E}Y(t) = g(t; \beta^*) = \beta_0^* + \beta_1^* t. \quad (\text{S1.1})$$

This model is nonparametric because the model imposed on  $\mathbb{E}Y(t)$  does not induce any restrictions on the observed data distribution.

When  $t$  is a continuous variable, however, finite-dimensional parametric MSM is often specified. For example, the statistician could insist on modeling  $\mathbb{E}Y(t)$  by linear model (S1.1) and believe that the model is the truth. Under unconfoundedness, linear MSM (S1.1) imposes the following restriction on the observed data distribution:

$$\mathbb{E}Y(t) = \mathbb{E}[\mathbb{E}(Y|X, T=t)] = \mathbb{E}[\xi(X, T)Y|T=t] = \beta_0^* + \beta_1^* t \equiv g(t; \beta^*), \quad (\text{S1.2})$$

where  $\xi(x, t) = \mathbf{p}_T(x)/\mathbf{p}_{T|X}(t|x)$ . Under unconfoundedness, Model (S1.2) imposes the following moment/equality constraint on the data generating process:

$$\mathbb{E}[\xi(X, T)(Y - g(T; \beta^*))\omega(T)] \equiv 0, \forall \omega : \mathcal{T} \rightarrow \mathbb{R}, \quad (\text{S1.3})$$

so induces a conditional moment constraint in the sense of [Ai and Chen \(2003\)](#), leading to a non-trivial orthocomplement to the model tangent space. We also refer readers to Section 3 of [Chen and Xie \(2025\)](#) for a related discussion. It should be noted that one can also parameterize the MSM using nonparametric models when  $t$  is continuous, as in [Bonvini and Kennedy \(2022\)](#) or as we have done in Section 4.

## S2 Verification of Assumption 7 for ATE and QTE

In this section, we demonstrate that Assumptions 7(i) and 7(iii) hold for both ATE and QTE from Examples 1 and 2. We first recall the following definition from the empirical process theory ([van der Vaart and Wellner, 2023](#)).

**Definition 2** (VC-subgraph class). A collection  $\mathcal{F}$  of measurable functions is called a VC-subgraph class, or a VC-class, if the collection of all subgraphs of the functions (see Section 2.6.2 of [van der Vaart and Wellner \(2023\)](#)) in  $\mathcal{F}$  forms a VC-class of sets (see Section 2.6.1 of [van der Vaart and Wellner \(2023\)](#)).

It suffices to verify Assumption 7 for the first coordinate of  $\Gamma(\cdot, \cdot, \beta)$ , which corresponds to the estimating equation for  $\beta_0^*$  when  $T = 0$ . We first verify Assumption 7(i):

- ATE in Example 1. We need only to consider the function class  $\mathcal{F}_{\text{ATE}} := \{Y - \beta_0 : \beta_0 \in \mathcal{B}_0\}$ , where  $\mathcal{B}_0$  is the parameter space for  $\beta_0^*$ . By Lemma 2.6.17 in [van der Vaart and Wellner \(2023\)](#),  $\mathcal{F}_{\text{ATE}}$  is of VC-subgraph with index 1. Theorem 2.6.7 of [van der Vaart and Wellner \(2023\)](#) further implies that, for any probability measure  $Q$ ,

$$N(\mathcal{F}_{\text{ATE}}, \|\cdot\|_{Q,2}, \epsilon \|F_{\text{ATE}}\|_{Q,2}) \leq C(16e)\epsilon^{-2}, \quad 0 < \epsilon < 1,$$

where  $F_{\text{ATE}} = \sup_{\beta_0 \in \mathcal{B}_0} |y - \beta_0|$  is the envelope and  $C$  is a universal constant. Hence, by Definition 1,  $\mathcal{F}_{\text{ATE}}$  is of VC-type with characteristics  $(4\sqrt{Ce}, 2)$ .

- QTE in Example 2. In this case, the relevant function class  $\mathcal{F}_{\text{QTE}} := \{\tau - \mathbb{1}(Y \leq \beta_0) : \beta_0 \in \mathcal{B}_0\}$ . Denote  $\rho'(u) := \tau - \mathbb{1}(u \leq 0)$ , for  $u \in \mathbb{R}$ . Then  $\rho'(\cdot)$  is monotone and  $\mathcal{F}_{\text{QTE}} = \rho' \circ \mathcal{F}_{\text{ATE}} := \{\rho'(f) : f \in \mathcal{F}_{\text{ATE}}\}$ . Lemma 9.9 (viii) of [Kosorok \(2008\)](#) implies that applying a monotone function to a VC-subgraph class does not increase the VC index. Therefore,  $\mathcal{F}_{\text{QTE}}$  is also of VC-subgraph with index 1, and thus of VC-type with characteristics  $(4\sqrt{Ce}, 2)$ .

For verification of Assumption 7(iii), denote the first coordinate of  $\Gamma(\cdot, \cdot, \beta)$  as  $\Gamma^{(0)}(\cdot, \cdot, \beta_0)$ .

- ATE in Example 1. The difference is  $\Gamma^{(0)}(Y, T, \beta_0) - \Gamma^{(0)}(Y, T, \beta'_0) = (\beta_0 - \beta'_0)(1 - T)$ , which is linear in  $\beta_0 - \beta'_0$ . Thus,  $\Gamma^{(0)}(\cdot, \cdot, \beta_0)$  meets this condition with index  $\alpha_0 = 1$ .

- QTE in Example 2. We have  $\Gamma^{(0)}(Y, T, \beta_0) - \Gamma^{(0)}(Y, T, \beta'_0) = [\mathbb{1}(Y \leq \beta'_0) - \mathbb{1}(Y \leq \beta_0)](1 - T)$ . Then,

$$\begin{aligned}
& \mathbb{E} \left[ \sup_{|\beta_0 - \beta'_0| \lesssim \delta} |\Gamma^{(0)}(Y, T, \beta_0) - \Gamma^{(0)}(Y, T, \beta'_0)|^2 \right] \\
&= \mathbb{E} \left[ (1 - T)^2 \cdot \sup_{|\beta_0 - \beta'_0| \lesssim \delta} \mathbb{1}(\beta_0 \wedge \beta'_0 < Y \leq \beta_0 \vee \beta'_0) \right] \\
&= \mathbb{E} \left[ \sup_{|\beta_0 - \beta'_0| \lesssim \delta} \mathbb{1}(\beta_0 \wedge \beta'_0 < Y \leq \beta_0 \vee \beta'_0) \right] \cdot \mathbb{P}(T = 0) \\
&= \mathbb{E} \left[ \sup_{|\beta_0 - \beta'_0| \lesssim \delta} |\mathbb{P}_Y(\beta_0) - \mathbb{P}_Y(\beta'_0)| \right] \cdot \mathbb{P}(T = 0) \\
&\lesssim \delta,
\end{aligned}$$

where  $\mathbb{P}_Y(\beta) := \mathbb{P}(Y \leq \beta)$  is the marginal distribution of  $Y$  and the last inequality holds if  $\mathbb{P}_Y(\cdot)$  is Lipschitz continuous. Thus, in this case,  $\Gamma^{(0)}(\cdot, \cdot, \beta_0)$  satisfies this condition with index  $\alpha_0 = 1/2$ .

### S3 Higher-Order Estimators for Parameters of GTMs

As mentioned in Remark 5, for the sake of completeness, the following lemma characterizes the HOIFs of  $\psi(\beta)$  (or more precisely, of  $\bar{\psi}_k(\beta)$ ) in the sense of Robins et al. (2016), which paves the way towards further reducing the bias due to estimating  $\Sigma_k$  by  $\hat{\Sigma}_k$ . However, deriving the statistical properties of these higher-order estimators is beyond the scope of this paper, as the proof will eventually need to handle  $U$ -processes of order growing at a rate poly-log in sample size  $n$ .

**Lemma 1.** *For any  $\beta \in \mathcal{B}$ , under Assumption 1, the  $m^{\text{th}}$ -order influence functions of  $B_{\psi,k}$  for  $m \geq 2$  are of the following forms:*

$$\begin{aligned}
& \text{IF}_{B_{\psi,k}}^{(2)}(\beta) = -(\hat{\xi}(X_1, T_1) - \xi(X_1, T_1))\bar{\phi}_k(X_1, T_1)^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2)(\Gamma(Y_2, T_2, \beta) - \hat{b}_{\beta_{\text{init}}}(X_2, T_2)), \\
& \text{IF}_{B_{\psi,k}}^{(m)}(\beta) - \text{IF}_{B_{\psi,k}}^{(m-1)}(\beta) = \\
& + (-1)^{m+1}(\hat{\xi}(X_1, T_1) - \xi(X_1, T_1))\bar{\phi}_k(X_1, T_1)^\top \Sigma_k^{-1} \prod_{s=3}^m \left\{ (\bar{\phi}(X_s, T_s)\bar{\phi}(X_s, T_s)^\top - \Sigma_k)\Sigma_k^{-1} \right\} \\
& \times \bar{\phi}_k(X_2, T_2)(\Gamma(Y_2, T_2, \beta) - \hat{b}_{\beta_{\text{init}}}(X_2, T_2)).
\end{aligned}$$

Then the  $m^{\text{th}}$ -order influence functions of  $\bar{\psi}_k(\beta)$  for  $m \geq 2$  simply take the forms

$$\text{IF}_{\bar{\psi}_k}^{(m)}(\beta) = \text{IF}_{\bar{\psi}}^{(1)}(\beta) + \text{IF}_{B_{\psi,k}}^{(m)}(\beta).$$

Since (1) is only used for defining the treatment effect parameter  $\beta^*$  and there is no restriction

on the observed data tangent space, the derivations of HOIFs of  $\bar{\psi}_k(\boldsymbol{\beta})$  will follow the strategy presented in Liu et al. (2024) and Robins et al. (2016). When referring to Liu et al. (2024) in the proof, we always mean the [arXiv version](#) of Liu et al. (2024).

*Proof of Lemma 1.* We first derive the HOIFs of  $B_{\psi,k}(\boldsymbol{\beta})$  and  $\bar{\psi}_k(\boldsymbol{\beta})$ . To ease presentation, we introduce some additional notation useful locally in the derivation.

$$\begin{aligned}\alpha_k &:= \Sigma_k^{-1} \mathbb{E}[(\widehat{\xi}(X, T) - \xi(X, T))\bar{\phi}_k(X, T)], \eta_{\boldsymbol{\beta},k} := \Sigma_k^{-1} \mathbb{E}[(\Gamma(Y, T, \boldsymbol{\beta}) - \widehat{b}_{\boldsymbol{\beta}_{\text{init}}}(X, T))\bar{\phi}_k(X, T)], \\ \bar{\xi}_k(\cdot, \cdot) &:= \widehat{\xi}(\cdot, \cdot) - \alpha_k^\top \bar{\phi}_k(\cdot, \cdot), \bar{b}_{\boldsymbol{\beta},k}(\cdot, \cdot) := \widehat{b}_{\boldsymbol{\beta}_{\text{init}}}(\cdot, \cdot) + \eta_{\boldsymbol{\beta},k}^\top \bar{\phi}_k(\cdot, \cdot).\end{aligned}$$

So  $B_{\psi,k}(\boldsymbol{\beta}) \equiv \alpha_k^\top \Sigma_k \eta_{\boldsymbol{\beta},k}$  and  $\bar{\psi}_k(\boldsymbol{\beta}) = \widehat{\psi}^{(1)}(\boldsymbol{\beta}) - \alpha_k^\top \Sigma_k \eta_{\boldsymbol{\beta},k}$ . By Lemma A.2 of Liu et al. (2024), we have

$$\begin{aligned}\text{IF}_{\alpha_k}^{(1)} &= \Sigma_k^{-1} \left( \bar{\xi}_k(X, T) - \xi(X, T) \right) \bar{\phi}_k(X, T), \\ \text{IF}_{\eta_{\boldsymbol{\beta},k}}^{(1)} &= \Sigma_k^{-1} \left( \Gamma(Y, T, \boldsymbol{\beta}) - \bar{b}_{\boldsymbol{\beta},k}(X, T) \right) \bar{\phi}_k(X, T).\end{aligned}$$

It is a standard calculation to show that the influence function of  $B_{\psi,k}(\boldsymbol{\beta})$  is:

$$\begin{aligned}\text{IF}_{B_{\psi,k}}^{(1)}(\boldsymbol{\beta}) &= (\widehat{\xi}(X, T) - \xi(X, T))\bar{\phi}_k(X, T)^\top \Sigma_k^{-1} \eta_{\boldsymbol{\beta},k} + \alpha_k^\top \Sigma_k^{-1} \bar{\phi}_k(X, T)(\Gamma(Y, T, \boldsymbol{\beta}) - \widehat{b}_{\boldsymbol{\beta}_{\text{init}}}(X, T)) \\ &\quad - \alpha_k^\top \Sigma_k^{-1} \bar{\phi}_k(X, T)\bar{\phi}_k(X, T)^\top \Sigma_k^{-1} \eta_{\boldsymbol{\beta},k} - \alpha_k^\top \Sigma_k^{-1} \eta_{\boldsymbol{\beta},k}.\end{aligned}$$

By applying the same calculus rule from page 50 to page 51 of Liu et al. (2024), we obtain the second-order influence function of  $B_{\psi,k}$  takes the form

$$\text{IF}_{B_{\psi,k}}^{(2)}(\boldsymbol{\beta}) = -(\widehat{\xi}(X_1, T_1) - \xi(X_1, T_1))\bar{\phi}_k(X_1, T_1)^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2)(\Gamma(Y_2, T_2, \boldsymbol{\beta}) - \widehat{b}_{\boldsymbol{\beta}_{\text{init}}}(X_2, T_2)).$$

The HOIFs of  $B_{\psi,k}(\boldsymbol{\beta})$  then read as follows: for  $j \geq 3$

$$\begin{aligned}\text{IF}_{B_{\psi,k}}^{(j)}(\boldsymbol{\beta}) - \text{IF}_{B_{\psi,k}}^{(j-1)}(\boldsymbol{\beta}) &= (-1)^{j-1} (\widehat{\xi}(X_1, T_1) - \xi(X_1, T_1))\bar{\phi}_k(X_1, T_1)^\top \Sigma_k^{-1} \prod_{s=3}^j \left\{ \left( \bar{\phi}_k(X_s, T_s)\bar{\phi}_k(X_s, T_s)^\top \Sigma_k \right) \Sigma_k^{-1} \right\} \\ &\quad \times \bar{\phi}_k(X_2, T_2)(\Gamma(Y_2, T_2, \boldsymbol{\beta}) - \widehat{b}_{\boldsymbol{\beta}_{\text{init}}}(X_2, T_2)).\end{aligned}$$

The HOIFs of  $\bar{\psi}_k(\boldsymbol{\beta})$  follows from the HOIFs of  $B_{\psi,k}(\boldsymbol{\beta})$  and the linearity of influence functions.  $\square$

Armed with these results, we can construct higher-order estimators of  $\psi(\boldsymbol{\beta})$  and of  $\boldsymbol{\beta}$  itself as follows. Similar to the first-order estimator of  $\psi(\boldsymbol{\beta})$ , which is a second-order  $U$ -statistic, the  $m$ -th

order estimator of  $\psi(\beta)$  is given by the following  $(m+1)$ -th order  $U$ -statistic:

$$\hat{\psi}_k^{(m)}(\beta) := \hat{\psi}^{(1)}(\beta) + \mathbb{U}_{n,m+1}[\hat{\mathbb{F}}_{\mathbf{B}_{\psi,k}}^{(m)}]$$

where  $\hat{\mathbb{F}}_{\mathbf{B}_{\psi,k}}^{(m)}$  is defined in the same way as  $\mathbb{F}_{\mathbf{B}_{\psi,k}}^{(m)}$ , except that (i)  $\xi(X_1, T_1)\bar{\phi}_k(X_1, T_1)$  is replaced by  $\bar{\phi}_k(X_1, T_3)$ , (ii)  $\prod_{s=3}^m$  becomes  $\prod_{s=4}^{m+1}$  and (iii)  $\Sigma_k$  is replaced by  $\hat{\Sigma}_k$ . It is worth noting that  $-\hat{\mathbb{B}}_{\psi,k}(\beta) \equiv \mathbb{U}_{n,3}[\hat{\mathbb{F}}_{\mathbf{B}_{\psi,k}}^{(2)}(\beta)]$ .  $\hat{\beta}^{(m)}$  is then defined as the solution to the system of  $U$ -statistic equations

$$\hat{\psi}_k^{(m)}(\beta) \equiv 0.$$

Next, the following result bounds the bias and variance of  $\hat{\psi}_k^{(m)}(\beta)$  as an estimator of the truncated parameter  $\bar{\psi}_k(\beta)$ , indicating the potential benefit of using a higher-order estimator to debias the estimation error due to estimating  $\Sigma_k$ . The proof follows immediately from [Liu et al. \(2017\)](#), hence omitted.

**Lemma 2.** *For any  $\beta \in \mathcal{B}$  and fixed positive integer  $m \geq 2$ , we have*

$$\left\| \mathbb{E} \left[ \hat{\psi}_k^{(m)}(\beta) - \bar{\psi}_k(\beta) \right] \right\| \lesssim \|\hat{\xi} - \xi\|_{\mathbb{P},2} \cdot \|\hat{b}_{\beta_{\text{init}}} - b_{\beta}\|_{\mathbb{P},2} \cdot \|\hat{\Sigma}_k - \Sigma_k\|_{\text{op}}^{m-1}. \quad (\text{S3.1})$$

As a consequence of Assumption 6(ii), we in turn have

$$\left\| \mathbb{E} \left[ \hat{\psi}_k^{(m)}(\beta) - \bar{\psi}_k(\beta) \right] \right\| \lesssim \|\hat{\xi} - \xi\|_{\mathbb{P},2} \cdot \left( \|\beta_{\text{init}} - \beta\| + \|\hat{b}_{\beta_{\text{init}}} - b_{\beta_{\text{init}}}\|_{\mathbb{P},2} \right) \cdot \|\hat{\Sigma}_k - \Sigma_k\|_{\text{op}}^{m-1}. \quad (\text{S3.2})$$

If we further impose the same conditions as in Theorem 1 and additionally assume (i)  $m \lesssim \sqrt{\log n}$  and  $k \lesssim n/\log^3(n)$  and (ii)  $\|\hat{\Sigma}_k - \Sigma_k\|_{\text{op}} = o_{\mathbb{P}}(1)$ , we also have

$$\text{cov}[\hat{\psi}_k^{(m)}(\beta)] \preceq \frac{C}{n} \left( \frac{k}{n} + \left\{ \|\Pi[\hat{b}_{\beta_{\text{init}}} - b_{\beta}|\bar{\phi}_k]\|_{\mathbb{P},2}^2 + \|\Pi[\hat{\xi} - \xi|\bar{\phi}_k]\|_{\mathbb{P},2}^2 \right\} + \min_{(p,q): p,q \geq 1, \frac{1}{p} + \frac{1}{q} = 1} \|\Pi[\hat{b}_{\beta_{\text{init}}} - b_{\beta}|\bar{\phi}_k]\|_{\mathbb{P},p}^2 \cdot \|\Pi[\hat{b}_{\beta_{\text{init}}} - b_{\beta}|\bar{\phi}_k]\|_{\mathbb{P},q}^2 \right), \quad (\text{S3.3})$$

for some absolute constant  $C > 0$ .

Lemma 2 can be served as the starting point of deriving statistical properties of  $\hat{\beta}^{(m)}$ , were results related to higher-order  $U$ -processes established. We leave this important technical problem to future work.

### S3.1 Convergence Rates of the Empirical Gram Matrix Estimator

When  $\beta^*$  is  $n^{-1/2}$ -estimable, we can appeal to the methodology developed in [Liu et al. \(2017\)](#) and take  $k = o(n)$  or more precisely  $k \lesssim n/\log^3(n)$  without assuming regularity conditions on the

marginal density  $p_X(x)$  of  $X$ . Specifically, we use the same nuisance sample for estimating the nuisance parameters to construct the following second-order  $U$ -statistic estimator of  $\Sigma_k$ :

$$\begin{aligned}\hat{\Sigma}_k &:= \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \{\bar{z}_{k_x}(X_i) \bar{z}_{k_x}(X_i)^\top\} \otimes \{\bar{w}_{k_t}(T_j) \bar{w}_{k_t}(T_j)^\top\} \\ &= \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \bar{\phi}_k(X_i, T_j) \bar{\phi}_k(X_i, T_j)^\top.\end{aligned}$$

Obviously,  $\mathbb{E}[\hat{\Sigma}_k] = \Sigma_k$  so it is an unbiased estimator of  $\Sigma_k$ . We note that throughout the article  $\hat{\Sigma}_k$  denotes a generic estimator of  $\Sigma_k$  and only in this section  $\hat{\Sigma}_k$  denotes the empirical Gram matrix estimator.

Let  $\bar{\phi}_k(X_i, T_j) \bar{\phi}_k(X_i, T_j)^\top =: Q_i \otimes R_j$  to further simplify the presentation. Corollary 3.1 of [Minsker and Wei \(2019\)](#), which we cite below, provides a user-friendly moment bound on the second-order degenerate  $U$ -statistic component of the following Hoeffding decomposition:

$$\begin{aligned}\hat{\Sigma}_k - \Sigma_k &\equiv \sum_{i=1}^n \frac{1}{n} \left\{ \mathbb{E} \left[ \frac{Q_i \otimes R_j + Q_j \otimes R_i}{2} | O_i \right] - \mathbb{E} Q \otimes \mathbb{E} R \right\} \\ &\quad + \sum_{1 \leq i < j \leq n} \frac{1}{n(n-1)} \{ (Q_i \otimes R_j + Q_j \otimes R_i) - \mathbb{E}(Q_i \otimes R_j + Q_j \otimes R_i | O_i) \} \\ &=: \sum_{i=1}^n G(O_i) + \sum_{1 \leq i < j \leq n} H(O_i, O_j),\end{aligned}$$

where  $G$  and  $H$  are appropriately chosen  $k \times k$ -*symmetric matrix*-valued functions. For more recent development on tail bounds for matrix-valued  $U$ -statistics, we refer readers to [Bandeira et al. \(2025\)](#).

**Lemma 3** (Corollary 3.1 of [Minsker and Wei \(2019\)](#)). *Let  $\{O_i^{(l)}, i = 1, \dots, n\}$  for  $l = 1, 2$  be two independent copies of  $\{O_i, i = 1, \dots, n\}$  and let  $\mathbb{E}_l$  denote the expectation with respect to only the copies indexed by  $l$ . Then*

$$\begin{aligned}\mathbb{E} \left\| \sum_{1 \leq i < j \leq n} H(O_i, O_j) \right\|_{\text{op}} &\lesssim (\log k)^2 \left( \sum_{i=1}^n \mathbb{E} \left\| \sum_{j: j \neq i} \mathbb{E}_2 H^2(O_i^{(1)}, O_j^{(2)}) \right\|_{\text{op}} \right) \\ &\quad + (\log k)^3 \left( \mathbb{E} \max_i \left\| \sum_{j: j \neq i} H^2(O_i^{(1)}, O_j^{(2)}) \right\|_{\text{op}} \right).\end{aligned}\tag{S3.4}$$

Thus to obtain a useful bound of  $\mathbb{E} \left\| \sum_{1 \leq i < j \leq n} H(O_i, O_j) \right\|_{\text{op}}$ , we need to control each of the two terms on the RHS of (S3.4), respectively denoted as  $(\log k)^2 I_1$  and  $(\log k)^3 I_2$  here. We first simplify

the following term that will be useful to control both  $I_1$  and  $I_2$ . Then

$$\begin{aligned}
& \sum_{j:j \neq i} H^2(O_i^{(1)}, O_j^{(2)}) \\
&= \frac{1}{n^2(n-1)^2} \sum_{j:j \neq i} \left\{ \begin{aligned} & \{Q_i^{(1)} \otimes (R_j^{(2)} - \mathbb{E}R)\} \{Q_i^{(1)\top} \otimes (R_j^{(2)} - \mathbb{E}R)^\top\} \\ & + \{(Q_j^{(2)} - \mathbb{E}Q) \otimes R_i^{(1)}\} \{(Q_j^{(2)} - \mathbb{E}Q)^\top \otimes R_i^{(1)\top}\} \\ & + \{Q_i^{(1)} \otimes (R_j^{(2)} - \mathbb{E}R)\} \{(Q_j^{(2)} - \mathbb{E}Q)^\top \otimes R_i^{(1)\top}\} \\ & + \{(Q_j^{(2)} - \mathbb{E}Q) \otimes R_i^{(1)}\} \{Q_i^{(1)\top} \otimes (R_j^{(2)} - \mathbb{E}R)^\top\} \end{aligned} \right\} \\
&= \frac{1}{n^2(n-1)} \left\{ \begin{aligned} & \{Q_i^{(1)} Q_i^{(1)\top}\} \otimes \left\{ \frac{1}{n-1} \sum_{j:j \neq i} (R_j^{(2)} - \mathbb{E}R)(R_j^{(2)} - \mathbb{E}R)^\top \right\} \\ & + \left\{ \frac{1}{n-1} \sum_{j:j \neq i} (Q_j^{(2)} - \mathbb{E}Q)(Q_j^{(2)} - \mathbb{E}Q)^\top \right\} \otimes \{R_i^{(1)} R_i^{(1)\top}\} \\ & + \frac{1}{n-1} \sum_{j:j \neq i} \{Q_i^{(1)} (Q_j^{(2)} - \mathbb{E}Q)^\top\} \otimes \{(R_j^{(2)} - \mathbb{E}R) R_i^{(1)\top}\} \\ & + \frac{1}{n-1} \sum_{j:j \neq i} \{(Q_j^{(2)} - \mathbb{E}Q) Q_i^{(1)\top}\} \otimes \{R_i^{(1)} (R_j^{(2)} - \mathbb{E}R)^\top\} \end{aligned} \right\}.
\end{aligned}$$

As for  $I_1$ , we have

$$\begin{aligned}
& \sum_{j:j \neq i} \mathbb{E}_2 H^2(O_i^{(1)}, O_j^{(2)}) \\
&= \frac{1}{n^2(n-1)} \left\{ \begin{aligned} & \{Q_i^{(1)} Q_i^{(1)\top}\} \otimes \mathbb{E}_2 (R_j^{(2)} - \mathbb{E}R)(R_j^{(2)} - \mathbb{E}R)^\top \\ & + \mathbb{E}_2 (Q_j^{(2)} - \mathbb{E}Q)(Q_j^{(2)} - \mathbb{E}Q)^\top \otimes \{R_i^{(1)} R_i^{(1)\top}\} \\ & + \mathbb{E}_2 \{Q_i^{(1)} (Q_j^{(2)} - \mathbb{E}Q)^\top\} \otimes \{(R_j^{(2)} - \mathbb{E}R) R_i^{(1)\top}\} \\ & + \mathbb{E}_2 \{(Q_j^{(2)} - \mathbb{E}Q) Q_i^{(1)\top}\} \otimes \{R_i^{(1)} (R_j^{(2)} - \mathbb{E}R)^\top\} \end{aligned} \right\}.
\end{aligned}$$

The first two terms in the above display are relatively easy to control. For the last two terms, we use the third term as an example and by symmetry the fourth term follows from the same analysis.

$$\begin{aligned}
& \mathbb{E}_2 \{Q_i^{(1)} (Q_j^{(2)} - \mathbb{E}Q)^\top\} \otimes \{(R_j^{(2)} - \mathbb{E}R) R_i^{(1)\top}\} = \mathbb{E}_2 \{Q_i^{(1)} \otimes (R_j^{(2)} - \mathbb{E}R)\} \{(Q_j^{(2)} - \mathbb{E}Q) \otimes R_i^{(1)}\}^\top \\
&= Q_i^{(1)} R_i^{(1)\top} \otimes \mathbb{E}_2 (R_j^{(2)} - \mathbb{E}R)(Q_j^{(2)} - \mathbb{E}R)^\top.
\end{aligned}$$

Then

$$I_1 \lesssim \frac{\mathbb{E} \|QQ^\top\|_{\text{op}} k_t + \mathbb{E} \|RR^\top\|_{\text{op}} k_x + \mathbb{E} \|QR^\top\|_{\text{op}} + \mathbb{E} \|RQ^\top\|_{\text{op}}}{n(n-1)} \lesssim \frac{k}{n^2}.$$

As for  $I_2$ , we have, by repeatedly applying the matrix Bernstein inequality as in Lemma 4,

$$I_2 \equiv \mathbb{E} \max_i \left\| \sum_{j:j \neq i} H^2(O_i^{(1)}, O_j^{(2)}) \right\|_{\text{op}}$$

$$\lesssim \frac{k}{n^3}.$$

Thus combining the above together, we have that the second-order degenerate  $U$ -statistic part is bounded by

$$\left(\frac{k \log k}{n}\right)^2 + \left(\frac{\sqrt{k} \log k}{n}\right)^2 + \left(\frac{k^{1/3} \log k}{n}\right)^3.$$

As for the linear term in the above decomposition, one can directly apply the standard matrix Bernstein inequality ([Tropp, 2015](#)), cited below.

**Lemma 4** (Rephrasing Theorem 6.1.1 of [Tropp \(2015\)](#)). *Let  $M := \sum_{i=1}^n G(O_i)$ . If for each  $i = 1, \dots, n$ ,  $\|G(O_i)\|_{\text{op}} \lesssim k/n$ , then*

$$\mathbb{E}\|M\|_{\text{op}} \lesssim \sqrt{\frac{k \log k}{n}} + \frac{k \log k}{n}.$$

Thus in summary, we have

$$\mathbb{E}\|\hat{\Sigma}_k - \Sigma_k\|_{\text{op}} \lesssim \sqrt{\frac{k \log k}{n}} + \frac{k \log k}{n} + \left(\frac{k \log k}{n}\right)^2 + k \left\{ \left(\frac{\log k}{n}\right)^2 + \left(\frac{\log k}{n}\right)^3 \right\}.$$

We can then combine this result with the empirical HOIF framework developed in [Liu et al. \(2017\)](#) to construct higher-order estimating equations for  $\beta^*$  as:

$$\hat{\psi}_k^{(m)}(\beta) = 0,$$

such that the final estimator  $\hat{\beta}^{(m)}$  with order  $m \asymp \log n$  is  $n^{-1/2}$ -consistent and asymptotic normal with no extra smoothness assumption imposed on the marginal density of  $X$ .

## S4 Discussion of Assumption 8

In this section, we briefly show how the localized conditions in Assumption 8 imply the standard uniform eigenvalue bounds for the Gram matrices:

**Lemma 5.** *For every  $k_x$  and  $k_t$ , the eigenvalues of  $\mathbb{E}[\bar{\mathbf{z}}_{k_x}(X)^{\otimes 2}]$  and  $\mathbb{E}[\bar{\mathbf{w}}_{k_t}(T)^{\otimes 2}]$  are bounded from above and away from zero uniformly in  $k_x$  and  $k_t$ .*

*Proof.* We focus on  $\mathbb{E}[\bar{\mathbf{w}}_{k_t}(T)^{\otimes 2}]$ ; the argument for  $\mathbb{E}[\bar{\mathbf{z}}_{k_x}(X)^{\otimes 2}]$  is analogous. For any unit vector

$\mathbf{b} \in \mathbb{R}^{k_t}$ , by definition

$$\mathbf{b}^\top \mathbb{E}[\bar{\mathbf{w}}_{k_t}(T)^{\otimes 2}] \mathbf{b} = \int_{\mathcal{T}} \left( \sum_{i=1}^{k_t} b_i \mathbf{w}_i(t) \right)^2 \mathbf{p}_T(t) dt.$$

By Assumption 8(ii), for any  $t \in \mathcal{T}$ , the number of nonzero terms in the basis function  $\bar{\mathbf{w}}_{k_t}(t)$  is at most  $c_8$ . Denoted them by  $(\mathbf{w}_{i_1}(t), \dots, \mathbf{w}_{i_{c_8}}(t))^\top$ . Then,

$$\left( \sum_{i=1}^{k_t} b_i \mathbf{w}_i(t) \right)^2 \lesssim \sum_{j=1}^{c_8} b_{i_j}^2 \mathbf{w}_{i_j}(t)^2.$$

It follows from Assumptions 8(i) and (iv) that

$$\begin{aligned} \int_{\mathcal{T}} \left( \sum_{i=1}^{k_t} b_i \mathbf{w}_i(t) \right)^2 \mathbf{p}_T(t) dt &\lesssim \int_{\mathcal{T}} \sum_{j=1}^{c_8} b_{i_j}^2 \mathbf{w}_{i_j}(t)^2 \mathbf{p}_T(t) dt \\ &\lesssim \sum_{j=1}^{c_8} b_{i_j}^2 \int_{\mathcal{S}_{T,i_j}} \zeta_t(k_t)^2 \mathbf{p}_T(t) dt \\ &\lesssim \sum_{j=1}^{c_8} b_{i_j}^2 \cdot k_t \cdot 1/k_t \lesssim 1, \end{aligned}$$

which yields the desired uniform upper bound on the largest eigenvalue.

For the lower bound, note that  $\sum_{i=1}^{k_t} \mathbb{1}\{t \in \mathcal{S}_{T,i}\} \leq c_8$  for all  $t \in \mathcal{T}$ . Hence,

$$\begin{aligned} \sum_{i=1}^{k_t} \int_{\mathcal{S}_{T,i}} \left( \sum_{i=1}^{k_t} b_i \mathbf{w}_i(t) \right)^2 \mathbf{p}_T(t) dt &= \int_{\mathcal{T}} \left( \sum_{i=1}^{k_t} \mathbb{1}\{t \in \mathcal{S}_{T,i}\} \right) \left( \sum_{i=1}^{k_t} b_i \mathbf{w}_i(t) \right)^2 \mathbf{p}_T(t) dt \\ &\leq c_8 \int_{\mathcal{T}} \left( \sum_{i=1}^{k_t} b_i \mathbf{w}_i(t) \right)^2 \mathbf{p}_T(t) dt \\ &\lesssim \mathbf{b}^\top \mathbb{E} \left[ \bar{\mathbf{w}}_{k_t}(T) \bar{\mathbf{w}}_{k_t}(T)^\top \right] \mathbf{b}. \end{aligned}$$

Then taking sum over all  $\mathcal{S}_{T,i}$  and using the assumption that  $\mathbf{b}^\top \int_{\mathcal{S}_{X,i}} \bar{\mathbf{w}}_{k_t}(t) \bar{\mathbf{w}}_{k_t}(t)^\top dt \mathbf{b} \gtrsim b_i^2$ , we obtain the lower bound

$$\mathbf{b}^\top \mathbb{E} \left[ \bar{\mathbf{w}}_{k_t}(T) \bar{\mathbf{w}}_{k_t}(T)^\top \right] \mathbf{b} \gtrsim 1.$$

□

## S5 On the Relationship Between Higher-Order Estimators and Other Related Proposals in Econometrics

As suggested by a referee, in this section, we illustrate the connection between the higher-order estimators studied in this paper and several other bias correction methods in the recent econometrics literature, including Cattaneo et al. (2019), Cattaneo and Jansson (2018), and Cattaneo et al. (2025).

### S5.1 Discussion of Cattaneo et al. (2019)

Cattaneo et al. (2019) study two-step semiparametric estimation and inference in settings where the first-step linear regression may include many covariates. To keep the exposition simple, we focus on the ATE case and take the target parameter to be  $\beta^* = \mathbb{E}[Y(1)]$ . To facilitate comparison with our construction, we characterize  $\beta^*$  via the moment condition

$$\psi(\beta^*) = \mathbb{E}[\xi(X, T)\{Y - b(X, T = 1)\} + b(X, T = 1) - \beta^*] = 0, \quad (\text{S5.5})$$

where  $\xi(x, t) = t/\mathbb{P}(T = 1|X = x)$  and  $b(x, t) = t\mathbb{E}[Y|X = x, T = 1]$  are nuisance functions.

In this example, we first would like to emphasize that Cattaneo et al. (2019) and our paper target different parameters, which will be discussed in detail shortly after. However, Cattaneo et al. (2019) and our paper share a common feature, which is that both consider to use linear models (of a possibly nonlinear  $k$ -dimensional feature map/basis  $\bar{\phi}_k$  of  $(X, T)$ ) to approximate certain quantities related to nuisance parameters, such that the target parameters of Cattaneo et al. (2019) and ours both reduce to bilinear functionals, with their precise forms to be revealed thereafter. In particular, Cattaneo et al. (2019) use linear models to directly approximate the nuisance parameters, whereas we use a linear approximation to represent the nuisance estimation *errors*; when the nuisance estimates are zeros, these two approaches collapse. For such bilinear functionals, a naïve plug-in estimator can have non-negligible leave-in bias in high-dimensional settings when the dimension  $k \gtrsim \sqrt{n}$  (Cattaneo et al., 2019). In terms of bias correction strategies, Cattaneo et al. (2019) start from the plug-in estimator and apply a jackknife-based correction, whereas we exploit the bilinear structure directly and construct an unbiased  $U$ -statistic estimator. We show below that, for bilinear functionals, these two strategies yield the same estimator.

To facilitate the analysis below, let  $\bar{\mathbf{z}}_k(X) \in \mathbb{R}^k$  be a dictionary for  $X$ . Since  $T$  is binary, define  $\bar{\phi}_k(X, T) = T\bar{\mathbf{z}}_k(X)$  and the  $k \times k$  population Gram matrix  $\Sigma_k := \mathbb{E}[\bar{\phi}_k(X, T)^{\otimes 2}] = \mathbb{E}[T\bar{\mathbf{z}}_k(X)^{\otimes 2}]$ .

Strictly following Cattaneo et al. (2019), they are interested in solving for  $\beta^*$  of the estimating equation  $\psi(\beta^*) = 0$  in (S5.5), and postulate linear approximations for the nuisance components  $\xi$  and  $b$ . Specifically, we assume that

$$\xi(X_i, T_i) = \bar{\phi}_k(X_i, T_i)^\top \gamma_\xi + e_{1i}, \quad \mathbb{E}[\bar{\phi}_k(X_i, T_i)e_{1i}] = 0,$$

$$b(X_i, T_i) = \bar{\phi}_k(X_i, T_i)^\top \gamma_b + e_{2i}, \quad \mathbb{E}[\bar{\phi}_k(X_i, T_i)e_{2i}] = 0,$$

where  $e_{1i}$  and  $e_{2i}$  are the best linear approximation error, and  $(\gamma_\xi, \gamma_b)$  are the population least-squares coefficients, i.e.,

$$\begin{aligned} \gamma_\xi &= \Sigma_k^{-1} \mathbb{E}[\xi(X, T) \bar{\mathbf{z}}_k(X)] = \Sigma_k^{-1} \mathbb{E}[\bar{\mathbf{z}}_k(X)] \\ \gamma_b &= \Sigma_k^{-1} \mathbb{E}[b(X, T) \bar{\mathbf{z}}_k(X)] = \Sigma_k^{-1} \mathbb{E}[TY \bar{\mathbf{z}}_k(X)]. \end{aligned}$$

Let  $\tilde{\xi}(X, T) = \bar{\phi}_k(X, T)^\top \gamma_\xi$  and  $\tilde{b}(X, T) = \bar{\phi}_k(X, T)^\top \gamma_b$ . Plugging  $(\tilde{\xi}, \tilde{b})$  into the moment equation (S5.5) yields the following population quantity

$$\begin{aligned} \tilde{\beta} &= \mathbb{E} \left[ \tilde{\xi}(X, T) \{Y - \tilde{b}(X, T=1)\} + \tilde{b}(X, T=1) \right] \\ &= \mathbb{E} [T \bar{\mathbf{z}}_k(X) Y]^\top \gamma_\xi - \gamma_\xi^\top \mathbb{E} [T \bar{\mathbf{z}}_k(X) \bar{\mathbf{z}}_k(X)^\top] \gamma_b + \mathbb{E} [\bar{\mathbf{z}}_k(X)]^\top \gamma_b \\ &= \mathbb{E} [T \bar{\mathbf{z}}_k(X) Y]^\top \Sigma_k^{-1} \mathbb{E} [\bar{\mathbf{z}}_k(X)] - \gamma_\xi^\top \Sigma_k \gamma_b + \gamma_\xi^\top \Sigma_k \gamma_b \\ &= \mathbb{E} [T \bar{\mathbf{z}}_k(X) Y]^\top \Sigma_k^{-1} \mathbb{E} [\bar{\mathbf{z}}_k(X)]. \end{aligned}$$

Hence, under this linear-approximation regime, the target parameter  $\beta^*$  actually reduces to the above *bilinear functional*  $\tilde{\beta}$ , which has exactly the same structure as  $B_{\psi, k}$  (see e.g. (10)), which will also be reintroduced later specifically for ATE  $\beta^*$ .

In contrast, our higher-order construction instead targets the approximation bias of first-order estimating equation  $\hat{\psi}^{(1)}(\beta) := \frac{1}{n} \sum_{i=1}^n [\hat{\xi}(X_i, T_i) \{Y_i - \hat{b}(X_i, 1)\} + \hat{b}(X_i, 1) - \beta]$ , and uses linear approximation only to approximate the nuisance residuals  $\hat{\xi} - \xi$  and  $\hat{b} - b$ . The bias of  $\hat{\psi}^{(1)}(\beta)$  has a product form involving two first-step estimation errors:

$$\mathbb{E}\{\hat{\psi}^{(1)}(\beta)\} - \psi(\beta) = \mathbb{E}[\{\hat{\xi}(X, T) - \xi(X, T)\} \{b(X, T) - \hat{b}(X, T)\}]. \quad (\text{S5.6})$$

Consistent with the main text, we employ sample splitting: the nuisance estimators  $\hat{\xi}$  and  $\hat{b}$  are obtained from an independent nuisance sample and all expectations  $\mathbb{E}[\cdot]$  heretofore in this section are taken with respect to the main sample conditional on the nuisance sample. Hence  $\hat{\xi}$  and  $\hat{b}$  can be treated as fixed functions.

Our framework imposes no linearity restriction on  $\xi$  and  $b$ . We allow  $\hat{\xi}$  and  $\hat{b}$  to be obtained using flexible methods (e.g., kernel smoothing, series estimation, or machine learning), although to obtain sharp theoretical results, we sometimes impose certain smoothness assumptions on  $\xi - \hat{\xi}$  and  $b - \hat{b}$  such that the truncation bias that cannot be captured by the basis  $\bar{\phi}_k$  is sufficiently small. We project the nuisance approximation residuals  $\hat{\xi} - \xi$  and  $\hat{b} - b$  onto the sieve space spanned by  $\bar{\phi}_k(x, t)$ . As described in Section 3.1, this projection leads to a decomposition of (S5.6) into (i) a truncation bias  $\text{TB}_{\psi, k}$  and (ii) a part of the bias (S5.6) captured by the  $k$ -dimensional dictionary

$\bar{\phi}_k$ . The latter can be written as a bilinear functional

$$B_{\psi,k} \equiv \alpha_k^\top \Sigma_k^{-1} \eta_k,$$

where

$$\alpha_k := \mathbb{E}[(\hat{\xi}(X, T) - 1)\bar{z}_k(X)], \quad \eta_k := \mathbb{E}[T(Y - \hat{b}(X, T))\bar{z}_k(X)]$$

are  $k$ -dimensional vectors; See Section 3.1 and 3.2 for a detailed derivation.

Next, we use the bilinear functional  $B_{\psi,k} = \alpha_k^\top \Sigma_k^{-1} \eta_k$  as a concrete example to show that the two constructions based on jackknife bias correction and  $U$ -statistics lead to the same estimator. Specifically, we take  $B_{\psi,k}$  as the parameter of interest and  $\alpha_k$  and  $\eta_k$  are high-dimensional nuisance parameters. We also assume that  $\Sigma_k := \mathbb{E}[\bar{\phi}_k(X, T)^{\otimes 2}]$  is known for convenience. Our strategy estimates  $B_{\psi,k}$  by leveraging its equivalent bilinear form:

$$B_{\psi,k} \equiv \mathbb{E}[\{\hat{\xi}(X, T) - 1\}\bar{z}_k(X)]^\top \Sigma_k^{-1} \mathbb{E}[T\{Y - \hat{b}(X, T)\}\bar{z}_k(X)],$$

which can be unbiasedly estimated by the following second-order  $U$ -statistic:

$$\hat{B}_{\psi,k} := \mathbb{U}_{n,2}[\{\hat{\xi}(X_i, T_i) - 1\}\bar{z}_k(X_i)^\top \Sigma_k^{-1} \bar{z}_k(X_j) T_j \{Y_j - \hat{b}(X_j, T_j)\}]. \quad (\text{S5.7})$$

The jackknife-based bias correction procedure in Cattaneo et al. (2019) starts from the plug-in estimator of  $B_{\psi,k}$

$$\hat{B}_{\psi,k,\text{plug-in}} = \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k,$$

where

$$\hat{\alpha}_k = \frac{1}{n} \sum_{i=1}^n \{\hat{\xi}(X_i, T_i) - 1\}\bar{z}_k(X_i), \quad \hat{\eta}_k := \frac{1}{n} \sum_{i=1}^n T_i \{Y_i - \hat{b}(X_i, T_i)\}\bar{z}_k(X_i).$$

The estimator  $\hat{B}_{\psi,k,\text{plug-in}}$ , which is a  $V$ -statistic, will have a bias of an order greater than or equal to  $n^{-1/2}$  when  $k \gtrsim \sqrt{n}$ . To implement the jackknife-based bias correction procedure in Cattaneo et al. (2019), for each  $\ell \in \{1, \dots, n\}$ , we compute the leave-one-out analogs of  $\hat{B}_{\psi,k,\text{plug-in}}$  as

$$\hat{B}_{\psi,k}^{(-\ell)} = \hat{\alpha}_k^{(-\ell)\top} \Sigma_k^{-1} \hat{\eta}_k^{(-\ell)},$$

where

$$\hat{\alpha}_k^{(-\ell)} := \frac{1}{n-1} \sum_{i=1, i \neq \ell}^n (\hat{\xi}(X_i, T_i) - 1)\bar{z}_k(X_i) = \frac{n}{n-1} \hat{\alpha}_k - \frac{1}{n-1} (T_\ell \hat{\xi}(X_\ell, T_\ell) - 1)\bar{z}_k(X_\ell), \quad (\text{S5.8})$$

$$\hat{\eta}_k^{(-\ell)} := \frac{n}{n-1} \hat{\eta}_k - \frac{1}{n-1} T_\ell(Y_\ell - \hat{b}(X_\ell, T_\ell)) \bar{z}_k(X_\ell). \quad (\text{S5.9})$$

Let  $\hat{B}_{\psi,k}^{(\cdot)} = n^{-1} \sum_{\ell=1}^n \hat{B}_{\psi,k}^{(-\ell)} = \frac{1}{n} \sum_{\ell=1}^n \hat{\alpha}_k^{(-\ell)\top} \Sigma_k^{-1} \hat{\eta}_k^{(-\ell)}$  be the average of the leave-one-out estimators. Then, the jackknife bias-corrected estimator is defined by

$$\hat{B}_{\psi,k,\text{jackknife}} = \hat{B}_{\psi,k,\text{plug-in}} - (n-1)(\hat{B}_{\psi,k}^{(\cdot)} - \hat{B}_{\psi,k,\text{plug-in}}).$$

Lemma 6 in the end of this subsection shows that the jackknife bias-corrected estimator coincides exactly with our second-order  $U$ -statistic estimator, i.e.,

$$\hat{B}_{\psi,k,\text{jackknife}} = \hat{B}_{\psi,k}.$$

The above derivation shows that, in the special case where the target functional admits the bilinear representation  $B_{\psi,k} = \alpha_k^\top \Sigma_k^{-1} \eta_k$ , the two constructions  $\hat{B}_{\psi,k}$  and  $\hat{B}_{\psi,k,\text{jackknife}}$  yield the same  $U$ -statistic. The key ingredient of our approach is to first construct a representation like  $B_{\psi,k}$  and estimate  $B_{\psi,k}$  by a natural  $U$ -statistic. By contrast, the jackknife-based bias correction method of Cattaneo et al. (2019) can be viewed as a general bias correction scheme for parameters such as  $B_{\psi,k}$ , for which standard plug-in estimators (such as the standard  $M$ -estimation) have non-negligible bias when the dimension is large relative to  $\sqrt{n}$ . It is worth mentioning that the approach in Cattaneo et al. (2019) can deal with parameters of a more generic nonlinear form than the bilinear form  $B_{\psi,k}$ , but the theoretical results are restricted to  $k = O(\sqrt{n})$  due to the nonlinearity of the more general problems considered. For  $B_{\psi,k}$ , we have seen that the  $U$ -statistic estimator is unbiased for any  $k$  when  $\Sigma_k$  is known. Finally, if  $\Sigma_k$  is unknown and is replaced by  $\hat{\Sigma}_k$ , our Supplementary Material Section S3 derives HOIFs of  $B_{\psi,k}$  to use higher-order  $U$ -statistics for further bias correction. To our knowledge, systematically extending jackknife-based corrections to control such higher-order remainder terms seems to be beyond the scope of Cattaneo et al. (2019), but we suspect that constructions similar to those in HOIFs can be considered. We also refer interested readers to Lin et al. (2024) for a recent result in this direction.

Finally, we record Lemma 6 and its proof below.

**Lemma 6.** *The following algebraic identity holds:*

$$\hat{B}_{\psi,k,\text{jackknife}} = \hat{B}_{\psi,k}.$$

*Proof.* Substituting the explicit leave-one-out formulas for  $\hat{\alpha}_k^{(-\ell)}$  and  $\hat{\eta}_k^{(-\ell)}$  in (S5.8)–(S5.9) into the definition of  $\hat{B}_{\psi,k}^{(\cdot)}$ , we have

$$\hat{B}_{\psi,k}^{(\cdot)} = \frac{1}{n} \sum_{\ell=1}^n \hat{\alpha}_k^{(-\ell)\top} \Sigma_k^{-1} \hat{\eta}_k^{(-\ell)}$$

$$\begin{aligned}
&= \frac{1}{n} \sum_{\ell=1}^n \left\{ \frac{n}{n-1} \hat{\alpha}_k - \frac{1}{n-1} (\hat{\xi}(X_\ell, T_\ell) - 1) \bar{z}_k(X_\ell) \right\}^\top \Sigma_k^{-1} \left\{ \frac{n}{n-1} \hat{\eta}_k - \frac{1}{n-1} T_\ell(Y_\ell - \hat{b}(X_\ell, T_\ell)) \bar{z}_k(X_\ell) \right\} \\
&= \frac{n^2}{(n-1)^2} \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k - \frac{2n}{(n-1)^2} \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k \\
&\quad + \frac{1}{n(n-1)^2} \sum_{\ell=1}^n (\hat{\xi}(X_\ell, T_\ell) - 1) \bar{z}_k(X_\ell)^\top \Sigma_k^{-1} \bar{z}_k(X_\ell) T_\ell(Y_\ell - \hat{b}(X_\ell, T_\ell)) \\
&= \left\{ 1 - \frac{1}{(n-1)^2} \right\} \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k + \frac{1}{n(n-1)^2} \sum_{\ell=1}^n (\hat{\xi}(X_\ell, T_\ell) - 1) \bar{z}_k(X_\ell)^\top \Sigma_k^{-1} \bar{z}_k(X_\ell) T_\ell(Y_\ell - \hat{b}(X_\ell, T_\ell)).
\end{aligned}$$

Then, we obtain

$$\begin{aligned}
&(n-1)(\hat{B}_{\psi,k}^{(\cdot)} - \hat{B}_{\psi,k,\text{plug-in}}) \\
&= -\frac{1}{n-1} \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k + \frac{1}{n(n-1)} \sum_{\ell=1}^n (\hat{\xi}(X_\ell, T_\ell) - 1) \bar{z}_k(X_\ell)^\top \Sigma_k^{-1} \bar{z}_k(X_\ell) T_\ell(Y_\ell - \hat{b}(X_\ell, T_\ell)).
\end{aligned}$$

Hence, using the above result, we derive

$$\begin{aligned}
\hat{B}_{\psi,k,\text{jackknife}} &= \hat{B}_{\psi,k,\text{plug-in}} - (n-1)(\hat{B}_{\psi,k}^{(\cdot)} - \hat{B}_{\psi,k,\text{plug-in}}) \\
&= \left( 1 + \frac{1}{n-1} \right) \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k - \frac{1}{n(n-1)} \sum_{\ell=1}^n (\hat{\xi}(X_\ell, T_\ell) - 1) \bar{z}_k(X_\ell)^\top \Sigma_k^{-1} \bar{z}_k(X_\ell) T_\ell(Y_\ell - \hat{b}(X_\ell, T_\ell)) \\
&= \frac{n}{(n-1)n^2} \sum_{i=1}^n \sum_{j=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{z}_k(X_i)^\top \Sigma_k^{-1} \bar{z}_k(X_j) T_j(Y_j - \hat{b}(X_j, T_j)) \\
&\quad - \frac{1}{n(n-1)} \sum_{\ell=1}^n (\hat{\xi}(X_\ell, T_\ell) - 1) \bar{z}_k(X_\ell)^\top \Sigma_k^{-1} \bar{z}_k(X_\ell) T_\ell(Y_\ell - \hat{b}(X_\ell, T_\ell)) \\
&= \frac{1}{(n-1)n} \sum_{1 \leq i \neq j \leq n} (\hat{\xi}(X_i, T_i) - 1) \bar{z}_k(X_i)^\top \Sigma_k^{-1} \bar{z}_k(X_j) T_j(Y_j - \hat{b}(X_j, T_j)) \\
&= \hat{B}_{\psi,k},
\end{aligned}$$

which completes the proof.  $\square$

## S5.2 Discussion of Cattaneo and Jansson (2018)

We next discuss how the bootstrap-based bias correction approach developed in Cattaneo and Jansson (2018) is also connected to our higher-order estimators.

Cattaneo and Jansson (2018) study two-step semiparametric estimation and inference with kernel-based first-step estimators, allowing the first step to converge slowly when the bandwidth is small. To facilitate comparison with our framework, we again use the ATE example in Section S5.2, where the target moment  $\psi(\beta)$  is linear in the nuisance functions  $\xi$  and  $b$ . In this linear case, the bias analyzed in Cattaneo and Jansson (2018) is mainly driven by the leave-in bias that results from

the use of the same sample in both steps, when the bandwidth is small. They show that, certain nonparametric bootstrap procedures can also recover the leading bias term and hence deliver valid inference.

Our construction allows the nuisance estimators  $\hat{\xi}$  and  $\hat{b}$  to be obtained by generic learners, such as random forests, or neural networks. By using sample splitting, we focus on the approximation bias. We first approximate this bias by a bilinear functional  $B_{\psi,k}$  induced by a dictionary-based projection, and then estimate this functional directly via a  $U$ -statistic estimator.

Despite this difference, the two approaches are closely connected. As discussed in Section S5.2, once the target is reduced to the bilinear functional  $B_{\psi,k}$ , we will show that the resulting bootstrap bias-corrected estimator is asymptotically equivalent to our  $U$ -statistics-based estimator  $\hat{B}_{\psi,k}$ .

Let  $\{O_i = (Y_i, T_i, X_i)\}_{i=1}^n$  be the original sample and let  $O_1^*, \dots, O_n^*$  be a bootstrap sample drawn with replacement from  $\{O_i\}_{i=1}^n$ . Define

$$\hat{\alpha}_k^* = \frac{1}{n} \sum_{j=1}^n (\hat{\xi}(X_j^*, T_j^*) - 1) \bar{z}_k(X_j^*), \quad \hat{\eta}_k^* = \frac{1}{n} \sum_{j=1}^n T_j^* (Y_j^* - \hat{b}(X_j^*, T_j^*)) \bar{z}_k(X_j^*).$$

The bootstrap plug-in estimator of  $B_{\psi,k} = \alpha_k^\top \Sigma_k^{-1} \eta_k$  is

$$\hat{B}_{\psi,k}^* = (\hat{\alpha}_k^*)^\top \Sigma_k^{-1} \hat{\eta}_k^*.$$

Let  $M_{ni}$  be the number of times that  $O_i$  is “redrawn” from  $\{O_1, \dots, O_n\}$  to form  $O_1^*, \dots, O_n^*$ . By construction, the vector of counts  $(M_{n1}, \dots, M_{nn})$  is independent of  $O_1, \dots, O_n$  and multinomially distributed with parameters  $n$  and (probabilities)  $1/n, \dots, 1/n$ . This representation allows us to rewrite any bootstrap empirical average as a weighted average over the original sample, with random weights  $\{M_{ni}\}_{i=1}^n$ . Then

$$\hat{\alpha}_k^* = \frac{1}{n} \sum_{j=1}^n (\hat{\xi}(X_j^*, T_j^*) - 1) \bar{z}_k(X_j^*) = \frac{1}{n} \sum_{i=1}^n M_{ni} (T_i \hat{\xi}(X_i, T_i) - 1) \bar{z}_k(X_i)$$

and similarly

$$\hat{\eta}_k^* = \frac{1}{n} \sum_{i=1}^n M_{ni} T_i (Y_i - \hat{b}(X_i, T_i)) \bar{z}_k(X_i).$$

The difference  $\hat{B}_{\psi,k}^* - \hat{B}_{\psi,k,\text{plug-in}}$  provides a natural estimator of the bias of the plug-in estimator  $\hat{B}_{\psi,k,\text{plug-in}}$  for  $B_{\psi,k}$ . For simplicity, we define the oracle bootstrap bias corrected estimator as

$$\hat{B}_{\psi,k,\text{bootstrap}} = \hat{B}_{\psi,k,\text{plug-in}} - \mathbb{E}^*[\hat{B}_{\psi,k}^* - \hat{B}_{\psi,k,\text{plug-in}}], \quad (\text{S5.10})$$

where  $\mathbb{E}^*[\cdot]$  denotes expectation conditional on the original sample. We refer to this estimator as oracle because it involves the conditional expectation  $\mathbb{E}^*[\cdot]$ . In practice, one can approximate this

conditional expectation by the Monte Carlo bootstrap average over  $B$  resamples.

Lemma 7 below shows that

$$\widehat{B}_{\psi,k,\text{bootstrap}} = \widehat{B}_{\psi,k} + O_{\mathbb{P}}\left(\frac{k}{n^2}\right).$$

Under  $k = o(n^2)$ , it follows that

$$\frac{k}{n^2} \cdot \frac{n}{\sqrt{k+n}} = \frac{k}{n\sqrt{k+n}} = o(1).$$

Hence, the remainder term is asymptotically negligible under the normalization used in our distributional results. Therefore,  $\widehat{B}_{\psi,k,\text{bootstrap}}$  and  $\widehat{B}_{\psi,k}$  are asymptotically equivalent up to an error term of order  $k/n^2$ , which is  $o(n^{-1/2})$  (negligible compared to the parametric rate) if  $k = o(n^{3/2})$ .

**Lemma 7.** *The following hold under Assumptions 6–8:*

$$\widehat{B}_{\psi,k,\text{bootstrap}} = \widehat{B}_{\psi,k} + O_{\mathbb{P}}\left(\frac{k}{n^2}\right).$$

*Proof.* Using the definitions of  $\widehat{B}_{\psi,k}^*$  and  $\widehat{B}_{\psi,k,\text{plug-in}}$ , we have

$$\begin{aligned} \widehat{B}_{\psi,k}^* - \widehat{B}_{\psi,k,\text{plug-in}} &= (\widehat{\alpha}_k^*)^\top \Sigma_k^{-1} \widehat{\eta}_k^* - \widehat{\alpha}_k^\top \Sigma_k^{-1} \widehat{\eta}_k \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \{M_{ni}M_{nj} - 1\} (\widehat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \widehat{b}(X_j, T_j)) \\ &= \frac{1}{n^2} \sum_{i=1, i \neq j}^n \sum_{j=1}^n \{M_{ni}M_{nj} - 1\} (\widehat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \widehat{b}(X_j, T_j)) \\ &\quad + \frac{1}{n^2} \sum_{i=1}^n \{M_{ni}^2 - 1\} (\widehat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \widehat{b}(X_i, T_i)). \end{aligned}$$

Since

$$\begin{aligned} \mathbb{E}^*[M_{ni}M_{nj}] &= \frac{n(n-1)}{n^2} = 1 - \frac{1}{n}, \\ \mathbb{E}^*[M_{ni}^2] &= \frac{n(n-1)}{n^2} + 1 = \frac{n-1}{n} + 1 = 2 - \frac{1}{n}, \end{aligned}$$

taking  $\mathbb{E}^*[\cdot]$  conditional on the original sample yields

$$\begin{aligned} \mathbb{E}^*[\widehat{B}_{\psi,k}^* - \widehat{B}_{\psi,k,\text{plug-in}}] &= -\frac{1}{n^3} \sum_{i=1, i \neq j}^n \sum_{j=1}^n (\widehat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \widehat{b}(X_j, T_j)) \\ &\quad + \frac{n-1}{n^3} \sum_{i=1}^n (\widehat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \widehat{b}(X_i, T_i)) \end{aligned}$$

$$\begin{aligned}
&= -\frac{1}{n^3} \sum_{i=1}^n \sum_{j=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \hat{b}(X_j, T_j)) \\
&\quad + \frac{1}{n^2} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)) \\
&= -\frac{1}{n} \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k + \frac{1}{n^2} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)).
\end{aligned}$$

Hence,

$$\begin{aligned}
\hat{\mathbf{B}}_{\psi, k, \text{bootstrap}} &= \hat{\mathbf{B}}_{\psi, k, \text{plug-in}} - \mathbb{E}^*[\hat{\mathbf{B}}_{\psi, k}^* - \hat{\mathbf{B}}_{\psi, k, \text{plug-in}}] \\
&= \frac{n+1}{n} \hat{\alpha}_k^\top \Sigma_k^{-1} \hat{\eta}_k - \frac{1}{n^2} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)) \\
&= \frac{n+1}{n^3} \sum_{i=1, i \neq j}^n \sum_{j=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \hat{b}(X_j, T_j)) \\
&\quad + \left( \frac{n+1}{n^3} - \frac{1}{n^2} \right) \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)) \\
&= \hat{\mathbf{B}}_{\psi, k} - \frac{1}{n^3(n-1)} \sum_{i=1, i \neq j}^n \sum_{j=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \hat{b}(X_j, T_j)) \\
&\quad + \frac{1}{n^3} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)) \\
&= \hat{\mathbf{B}}_{\psi, k} + O_{\mathbb{P}}\left(\frac{k}{n^2}\right),
\end{aligned}$$

where the last equality holds by Lemma 8.  $\square$

**Lemma 8.** *The following hold under Assumptions 6–8:*

$$\begin{aligned}
&\frac{1}{(n-1)n^2} \sum_{i=1}^n \sum_{i \neq j, j=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \hat{b}(X_j, T_j)) = O_{\mathbb{P}}(1), \\
&\frac{1}{n^2} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)) = O_{\mathbb{P}}\left(\frac{k}{n}\right).
\end{aligned}$$

*Proof.* Denote

$$\begin{aligned}
I &:= \frac{1}{n^2} \sum_{i=1}^n \sum_{i \neq j, j=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_j) T_j (Y_j - \hat{b}(X_j, T_j)), \\
II &:= \frac{1}{n^2} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)).
\end{aligned}$$

For the term  $I$ , the same variance calculation as in Lemma 15 yields  $\text{var}(I) = O(1/n + k/n^2)$ . For

the mean,

$$\mathbb{E}[I] = \mathbb{E}[(\hat{\xi}(X, T) - 1)\bar{\mathbf{z}}_k(X)]^\top \Sigma_k^{-1} \mathbb{E}[\bar{\mathbf{z}}_k(X)T(Y - \hat{b}(X, T))] = \alpha_k^\top \Sigma_k \eta_k = O(1),$$

where the last equality holds by the definitions of  $\alpha_k$  and  $\eta_k$ , and the condition that the largest eigenvalue of  $\Sigma_k$  is bounded. Combining mean and variance bounds,

$$I = O_{\mathbb{P}}\left(\frac{1}{\sqrt{n}} + \frac{\sqrt{k}}{n} + 1\right) = O_{\mathbb{P}}(1).$$

For the term  $II$ , by  $\bar{\mathbf{z}}_k(x)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(x) \leq \|\Sigma_k^{-1}\|_{\text{op}} \|\bar{\mathbf{z}}_k(x)\|^2 \lesssim \zeta_k^2$ ,

$$\begin{aligned} & \mathbb{E}\left[\left|\frac{1}{n^2} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i))\right|\right] \\ & \lesssim \frac{\zeta_k^2}{n} \mathbb{E}\left[\left|(\hat{\xi}(X_i, T_i) - 1) T_i (Y_i - \hat{b}(X_i, T_i))\right|\right] = O\left(\frac{\zeta_k^2}{n}\right) = O\left(\frac{k}{n}\right). \end{aligned}$$

Thus, by Markov's inequality, we have

$$II = \frac{1}{n^2} \sum_{i=1}^n (\hat{\xi}(X_i, T_i) - 1) \bar{\mathbf{z}}_k(X_i)^\top \Sigma_k^{-1} \bar{\mathbf{z}}_k(X_i) T_i (Y_i - \hat{b}(X_i, T_i)) = O_{\mathbb{P}}\left(\frac{k}{n}\right).$$

Combining the results for terms  $I$  and  $II$ , we complete the proof.  $\square$

## S6 Technical Results Related to $U$ -Processes

In this section, we establish refined local maximal inequalities for  $U$ -processes that can be used to establish convergence rates and asymptotic normality of higher-order estimators considered in this paper. These are key technical tools for the main results in Sections 3.3 and 4.2, and may also be of independent interest for other related applications. We begin by introducing the necessary notation.

### S6.1 Notation and Review of Concepts from $U$ -Process Theory

We first recall the key notation introduced in Section 1.3, and then introduce additional terminologies for  $U$ -statistics and  $U$ -processes that will be used in the technical lemmas that follow.

Recall that  $\mathbb{U}_{n,m}$  is the  $m$ -th order  $U$ -statistic operator, i.e.,

$$\mathbb{U}_{n,m} h \equiv \mathbb{U}_{n,m} h(O_1, \dots, O_m) = \frac{(n-m)!}{n!} \sum_{1 \leq i_1 \neq \dots \neq i_m \leq n} h(O_{i_1}, \dots, O_{i_m})$$

for any measurable function (kernel)  $h : \mathcal{O}^m \rightarrow \mathbb{R}$ . So,  $\mathbb{U}_{n,1} \equiv \mathbb{P}_n$  is understood to be the empirical

mean operator. Let  $S_m$  denote the symmetrization of a function of  $m$  variables,

$$S_m h(O_1, \dots, O_m) = \frac{1}{m!} \sum_{1 \leq i_1 \neq \dots \neq i_m \leq m} h(O_{i_1}, \dots, O_{i_m}).$$

Because a  $U$ -statistic averages over unordered subsets,  $\mathbb{U}_{n,m} h = \mathbb{U}_{n,m}(S_m h)$ . Accordingly, all subsequent notation and results are stated for symmetric kernels; when a kernel is asymmetric, they apply by default to its symmetrized version.

For any symmetric measurable function  $h : \mathcal{O}^m \rightarrow \mathbb{R}$  and  $j = 1, \dots, m$ , let  $\mathbb{P}^{m-j} h$  denote the function on  $\mathcal{O}^j$  defined by

$$\begin{aligned} \mathbb{P}^{m-j} h(o_1, \dots, o_j) &= \mathbb{E}[h(o_1, \dots, o_j, O_{j+1}, \dots, O_m)] \\ &= \int \dots \int h(o_1, \dots, o_j, o_{j+1}, \dots, o_m) d\mathbb{P}(o_{j+1}) \dots d\mathbb{P}(o_m), \end{aligned} \tag{S6.1}$$

where we always take  $O_1, \dots, O_m$  to be i.i.d. drawn from a common law  $\mathbb{P}$  in this paper. Thus,  $\mathbb{P}^0 h \equiv h$  and  $\mathbb{P}^m h \equiv \mathbb{E}[h(O_1, \dots, O_m)]$ . We also introduce the following shorthand notation for Hoeffding projection of a symmetric measurable function  $h$ , for  $j \leq m$ :

$$\pi_j h(o_1, \dots, o_j) := (\delta_{o_1} - \mathbb{P}) \dots (\delta_{o_j} - \mathbb{P}) \mathbb{P}^{m-j} h(o_1, \dots, o_m). \tag{S6.2}$$

For a non-symmetric kernel  $h$  we work with its symmetrised form  $S_m h$ ; all the above definitions are understood with that replacement, i.e.,  $\mathbb{P}^{m-j} h \equiv \mathbb{P}^{m-j} S_m h$  and  $\pi_j h \equiv \pi_j S_m h$ . Then the Hoeffding decomposition of  $\mathbb{U}_{n,m} h$  of any measurable functions  $h$  is given by

$$\mathbb{U}_{n,m} h - \mathbb{P}^m h = \sum_{j=1}^m \binom{m}{j} \mathbb{U}_{n,j}(\pi_j h).$$

We next review some basic terminologies about  $U$ -processes used throughout. Let  $\mathcal{F}$  be a class of symmetric measurable functions  $f : \mathcal{O}^m \rightarrow \mathbb{R}$ . We assume that there is a symmetric measurable envelope  $F$  for  $\mathcal{F}$ , that is  $\sup_{f \in \mathcal{F}} |f| \leq F$ , such that  $\mathbb{P}^m F^2 < \infty$ . We will assume certain uniform covering number conditions for the function class  $\mathcal{F}$ . Let  $\|\cdot\|_{\mathcal{F}}$  denote the supremum norm over  $\mathcal{F}$ , i.e.  $\|f\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} |f|$ . For a probability measure  $Q$  and a constant  $p > 1$ , let  $\|\cdot\|_{Q,p}$  denote the  $L_p(Q)$ -seminorm, i.e.,  $\|f\|_{Q,p} := \{\int |f(x)|^p dQ(x)\}^{1/p}$  for finite  $q$  while  $\|f\|_{Q,\infty}$  denotes the essential supremum of  $|f|$  with respect to  $Q$ . For any function class  $\mathcal{F}$  with  $\|F\|_{Q,p} > 0$ , we use  $N(\mathcal{F}, \|\cdot\|_{Q,p}, \epsilon \|F\|_{Q,p})$  to denote the minimal number of  $\|\cdot\|_{Q,p}$ -balls of radius  $\epsilon \|F\|_{Q,p}$  needed to cover  $\mathcal{F}$ . See Section 2.1 of [van der Vaart and Wellner \(2023\)](#) for details. For  $j = 1, \dots, m$ , we define the uniform entropy integral

$$J_j(\delta, \mathcal{F}, F) := \int_0^\delta \sup_Q \left\{ 1 + \log N(\mathbb{P}^{m-j} \mathcal{F}, \|\cdot\|_{Q,2}, \tau \|\mathbb{P}^{m-j} F\|_{Q,2}) \right\}^{j/2} d\tau,$$

where  $\mathbb{P}^{m-j}\mathcal{F} = \{\mathbb{P}^{m-j}f : f \in \mathcal{F}\}$  and  $\sup_Q$  is taken over all finitely discrete distributions on  $\mathcal{O}^j$ . It is straightforward to verify that  $\mathbb{P}^{m-j}F$  is an envelope for  $\mathbb{P}^{m-j}\mathcal{F}$ . To avoid measurability difficulties, we will assume that  $F$  is pointwise measurable. If  $F$  is pointwise measurable and  $\mathbb{P}^m F < \infty$  (which we have assumed), then  $\pi_j \mathcal{F} := \{\pi_j f : f \in \mathcal{F}\}$  and  $\mathbb{P}^{m-j}F$  for  $j = 1, \dots, m$  are all pointwise measurable by the dominated convergence theorem.

For completeness, we recall Definitions 1 and 2.

**Definition 1** (VC-subgraph class). A collection  $\mathcal{F}$  of measurable functions is called a VC-subgraph class, or a VC-class, if the collection of all subgraphs of the functions (see Section 2.6.2 of [van der Vaart and Wellner \(2023\)](#)) in  $\mathcal{F}$  forms a VC-class of sets (see Section 2.6.1 of [van der Vaart and Wellner \(2023\)](#)).

**Definition 2** (VC type class). A function class  $\mathcal{F}$  with envelope  $F$  is said to be of VC type with characteristic  $(A, v)$  if  $\sup_Q \mathbb{N}(\mathcal{F}, \|\cdot\|_{Q,2}, \epsilon \|F\|_{Q,2}) \leq (A/\epsilon)^v$  for all  $\epsilon \in (0, 1]$ , where  $\sup_Q$  denotes the supremum over all finitely discrete probability measures.

## S6.2 Local Maximal Inequalities for $U$ -Processes

In this section we allow the function class to depend explicitly on the sample size  $n$ . Throughout, we denote a generic function class by  $\mathcal{F}_n$  to emphasize this dependence and derive non-asymptotic bounds that hold for each fixed  $n$ .

[Chen and Kato \(2020\)](#) establish local maximal inequalities for  $U$ -processes indexed by a VC-type class  $\mathcal{F}_n := \{f : \mathcal{O}^2 \rightarrow \mathbb{R}\}$ , whose envelope is denoted by  $F_n$ . Their bounds are expressed in terms of the VC characteristics  $(A_n, v_n)$  and quantities including  $\sup_{f \in \mathcal{F}_n} \|\mathbb{P}^{2-k}f\|_{\mathbb{P}^k,2}$  and  $\|\mathbb{P}^{2-k}F_n\|_{\mathbb{P}^k,2}$  for  $k = 1, 2$  (see their Corollary 5.5). Because their bound uses the projected envelope  $\mathbb{P}F_n(\cdot)$  (i.e. the conditional mean of  $F_n$  in one of the arguments), it can be unnecessarily loose when  $\mathbb{P}F_n$  is of greater magnitude in order than the quantity of actual interest,  $\sup_{f \in \mathcal{F}_n} |\mathbb{P}f|$ . To sharpen the analysis, we work directly with  $\sup_{f \in \mathcal{F}_n} |\mathbb{P}f|$  and derive a refined maximal inequality tailored specifically for our problem. The next lemma records this result; as indicated, it is particularly useful when  $\sup_{f \in \mathcal{F}_n} |\mathbb{P}f|$  is much less than  $\mathbb{P}F_n$ .

**Lemma 9.** *Suppose that (1)  $\mathcal{F}_n := \{f : \mathcal{O}^2 \rightarrow \mathbb{R}\}$  is a collection of symmetric pointwise measurable functions from  $\mathcal{O}^2$  to  $\mathbb{R}$  with envelope function  $F_n$ , and (2)  $\mathcal{F}_n$  is of VC type with characteristics  $(A_n, v_n)$  with  $A_n \geq e$  and  $v_n \geq 1$  (recall from Definition 1). Then*

$$(i) \quad \mathbb{E} \left[ \|\mathbb{U}_{n,2}(f) - \mathbb{P}^2 f\|_{\mathcal{F}_n} \right] \lesssim \frac{v_n \log A_n \cdot \|F_n\|_{\mathbb{P}^2,2}}{n} + \frac{(v_n \log A_n)^{1/2} \cdot \|\mathbb{P}F_n\|_{\mathbb{P}^2,2}}{\sqrt{n}}.$$

(ii) *If we further assume (3)  $\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f(\cdot)\|_\infty \leq m_n$ , and  $\|\mathbb{P}F_n(\cdot)\|_\infty \leq M_n$ , then*

$$\mathbb{E} \left[ \|\mathbb{U}_{n,2}(f) - \mathbb{P}^2 f\|_{\mathcal{F}_n} \right] \lesssim \frac{\left( \|F_n\|_{\mathbb{P}^2,2} + m_n \right) \cdot v_n \log(A'_n \vee n)}{n}$$

$$+ \frac{\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2} \cdot \sqrt{v_n \log(A'_n \vee n)}}{\sqrt{n}},$$

where  $A'_n := \sqrt{A_n} M_n / m_n$ .

**Remark 14.** For the class  $\mathcal{F}_n = \mathcal{H}_{2,n}$  defined in (S7.8) in Section S7.1,  $\mathcal{F}_n$  is of VC type with fixed characteristics  $(4A, v)$  and envelope  $F_n = H_{2,n}$ . In this case, the following hold:  $\|\mathbb{P}F_n\|_{\mathbb{P},2} \leq \|F_n\|_{\mathbb{P}^2,2} \lesssim \sqrt{k}$ ,  $\|F_n(\cdot)\|_\infty \lesssim k$ ,  $\sup_{f \in \mathcal{F}_n} \|f\|_{\mathbb{P}^2,2} \lesssim \sqrt{k}$  with  $k \lesssim n^2$ . Moreover, because each  $\mathbb{P}f(\cdot)$  is a conditional expectation (recall from the notation (S6.1)), (S7.1) implies  $\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f(\cdot)\|_\infty \lesssim 1$  and  $\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2} \lesssim 1$ . Plugging these bounds into Corollary 5.5 of Chen and Kato (2020) renders all higher-order projection terms negligible, leaving the first order term:

$$\begin{aligned} \mathbb{E} \left[ \|\mathbb{U}_{n,2}(f) - \mathbb{P}^2 f\|_{\mathcal{F}_n} \right] &\lesssim \frac{\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2} \sqrt{v \log(A \vee n)} + \|\mathbb{P}F_n\|_{\mathbb{P},2} \cdot v \log(A \vee n)}{\sqrt{n}} \\ &\lesssim \frac{\sqrt{k} \log(n)}{\sqrt{n}}. \end{aligned}$$

Instead, our maximal inequality replaces the projected envelope  $\mathbb{P}F_n$  by the tighter quantity  $\sup_{f \in \mathcal{F}_n} |\mathbb{P}f|$ , and  $\mathbb{P}F_n$  only influences the bound through the VC characteristics. Thus,  $\mathbb{P}F_n$  only appears in a logarithmic factor, which in turn is dominated by  $\log n$ . Specifically, Lemma 9 (ii) implies that:

$$\mathbb{E} \left[ \|\mathbb{U}_{n,2}(f) - \mathbb{P}^2 f\|_{\mathcal{F}_n} \right] \lesssim \frac{\log(n)}{\sqrt{n}} + \frac{\sqrt{k} \log(n)}{n}.$$

Even when  $n \lesssim k \lesssim n^2$ , the right-hand side of the above display still shrinks to zero. Hence, the refined inequality remains informative in the regime where the bound by directly applying results in Chen and Kato (2020) is not sufficiently sharp.  $\blacksquare$

*Proof of Lemma 9.* For any  $f$  defined on  $\mathcal{O}^2$ , we have the following decomposition

$$\mathbb{U}_{n,2}(f) - \mathbb{P}^2 f = \mathbb{U}_{n,2}(\pi_2 f) + 2\mathbb{U}_{n,1}(\pi_1 f).$$

Here,  $\pi_1$  is the Hajék projections defined in Section S6.1. Since the function class  $\mathcal{F}_n$  is of VC type with characteristics  $(A_n, v_n)$  satisfying  $A_n \geq e$  and  $v_n \geq 1$ , applying Corollary 5.6 of Chen and Kato (2020) to function class  $\mathcal{F}_n$  with their  $k = 1, 2$  and  $p = 1$ , we obtain

$$\mathbb{E} \left[ \|\mathbb{U}_{n,2}(\pi_2 f)\|_{\mathcal{F}_n} \right] \lesssim \frac{J_2(1, \mathcal{F}_n, F_n) \cdot \|F_n\|_{\mathbb{P}^2,2}}{n} \lesssim \frac{v_n \log A_n \cdot \|F_n\|_{\mathbb{P}^2,2}}{n}, \quad (\text{S6.3})$$

$$\mathbb{E} \left[ \|\mathbb{U}_{n,1}(\pi_1 f)\|_{\mathcal{F}_n} \right] \lesssim \frac{J_1(1, \mathcal{F}_n, F_n) \cdot \|\mathbb{P}F_n\|_{\mathbb{P},2}}{\sqrt{n}} \lesssim \frac{(v_n \log A_n)^{1/2} \cdot \|\mathbb{P}F_n\|_{\mathbb{P},2}}{\sqrt{n}}. \quad (\text{S6.4})$$

Combining (S6.3)–(S6.4), we prove part (i) of this lemma.

If we further assume condition (3) stated in this lemma, we can sharpen the bound for the first-order term  $\mathbb{U}_{n,1}(\pi_1 f)$ . Define the function class  $\pi_1 \mathcal{F}_n := \{f = \pi_1 f : f \in \mathcal{F}_n\}$ . For any  $f \in \mathcal{F}_n$ , by definition of  $\pi_1$ , we have  $\pi_1 f = \mathbb{P}f - \mathbb{P}^2 f$ . Up to a universal constant, centering does not alter the VC-type characteristics. Consequently, using Lemma 14, we obtain  $\pi_1 \mathcal{F}_n$  is still of VC type with characteristics of order  $(4\sqrt{A'_n}M_n/m_n, 2v_n)$  and envelope  $m_n$ . Since  $\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2}/m_n \leq 1$ , applying Lemma 13 to the function class  $\pi_1 \mathcal{F}_n$ , we obtain

$$\mathbb{E} \left[ \|\mathbb{U}_{n,1}(f)\|_{\pi_1 \mathcal{F}_n} \right] \lesssim \frac{\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2} \cdot \sqrt{v_n \log(A'_n \vee n)}}{\sqrt{n}} + \frac{m_n v_n \log(A'_n \vee n)}{n}, \quad (\text{S6.5})$$

where  $A'_n = \sqrt{A_n}M_n/m_n$ . Since  $A'_n \geq A_n$ , combining (S6.3) and (S6.5), we obtain

$$\begin{aligned} \mathbb{E} \left[ \|\mathbb{U}_{n,2}(f) - \mathbb{P}^2 f\|_{\mathcal{F}_n} \right] &\lesssim \mathbb{E} \left[ \|\mathbb{U}_{n,2}(\pi_2 f)\|_{\mathcal{F}_n} \right] + \mathbb{E} \left[ \|\mathbb{U}_{n,1}(\pi_1 f)\|_{\mathcal{F}_n} \right] \\ &\lesssim \frac{(\|\mathbb{F}_n\|_{\mathbb{P}^2,2} + m_n) \cdot v_n \log(A'_n \vee n)}{n} + \frac{\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2} \cdot \sqrt{v_n \log(A'_n \vee n)}}{\sqrt{n}}, \end{aligned}$$

which completes the proof.  $\square$

The function class encountered in this paper generally has a product structure  $\mathcal{H}_n = \{f \cdot g_n : f \in \mathcal{F}\}$ , where  $\mathcal{F}$  is of a fixed VC-type class with characteristics  $(A, v)$ , and  $g_n$  is a single function that depends on  $n$ . By Lemma 12 below,  $\mathcal{H}_n$  is again of a VC-type class whose characteristics remain independent of  $n$ , whereas its envelope,  $F_n = \sup_{f \in \mathcal{F}} |f| \cdot |g_n|$ , can vary with  $n$ . For such a function class, we develop a new and readily useful maximal inequality for third-order  $U$ -processes, which are heavily used in our higher-order estimators. The proof parallels that of Lemma 9, with additional steps to handle the third-order Hoeffding projection term  $\mathbb{U}_{n,3}(\pi_3 f)$ .

**Lemma 10.** *Suppose that (1)  $\mathcal{F}_n := \{f : \mathcal{O}^3 \rightarrow \mathbb{R}\}$  is a collection of symmetric functions from  $\mathcal{O}^3$  to  $\mathbb{R}$  with envelope function  $F_n$ , (2)  $\mathcal{F}_n$  is of VC type with bounded characteristics  $(A, v)$  (recall from Definition 1), and further (3)  $\sup_{f \in \mathcal{F}_n} \|\mathbb{P}^2 f(\cdot)\|_\infty \leq m_n$ , and  $\|\mathbb{P}^2 F_n(\cdot)\|_\infty \leq M_n$ . Then*

$$\mathbb{E} \left[ \|\mathbb{U}_{n,3}(f) - \mathbb{P}^3 f\|_{\mathcal{F}_n} \right] \lesssim \frac{(\|\mathbb{F}_n\|_{\mathbb{P}^3,2} + m_n) \cdot \log(A'_n \vee n)}{n} + \frac{\sup_{f \in \mathcal{F}_n} \|\mathbb{P}^2 f\|_{\mathbb{P},2} \cdot \sqrt{\log(A'_n \vee n)}}{\sqrt{n}},$$

where  $A'_n := M_n/m_n$ .

*Proof.* Similar to the proof of Lemma 9, we first apply the Hoeffding decomposition:

$$\mathbb{U}_{n,3}(f) - \mathbb{P}^3 f = \mathbb{U}_{n,3}(\pi_3 f) + 3\mathbb{U}_{n,2}(\pi_2 f) + 3\mathbb{U}_{n,1}(\pi_1 f),$$

where  $\pi_j$  for  $j = 1, 2, 3$  represent the Hoeffding projections as defined in Section S6.1. By Jensen's inequality, we have  $\|\mathbb{P}^3 - 2\mathbb{F}_n\|_{\mathbb{P}^2,2} = O(\|\mathbb{F}_n\|_{\mathbb{P}^3,2})$ . Since the function class  $\mathcal{F}_n$  is of VC type with

bounded characteristics  $(A, v)$  independent of the sample size  $n$ , we have  $J_3(1, \mathcal{F}_n, F_n) \lesssim 1$  and  $J_2(1, \mathcal{F}_n, F_n) \lesssim 1$ . Thus, applying Corollary 5.6 of [Chen and Kato \(2020\)](#) to function class  $\mathcal{F}$  with their  $k = 2, 3$  and  $p = 1$ , we obtain

$$\mathbb{E} [\|\mathbb{U}_{n,3}(\pi_3 f)\|_{\mathcal{F}_n}] \lesssim \frac{J_3(1, \mathcal{F}_n, F_n) \cdot \|F_n\|_{\mathbb{P}^3,2}}{n^{3/2}} \lesssim \frac{\|F_n\|_{\mathbb{P}^3,2}}{n^{3/2}} \quad (\text{S6.6})$$

and

$$\mathbb{E} [\|\mathbb{U}_{n,2}(\pi_2 f)\|_{\mathcal{F}_n}] \lesssim \frac{J_2(1, \mathcal{F}_n, F_n) \cdot \|F_n\|_{\mathbb{P}^2,2}}{n} \lesssim \frac{\|F_n\|_{\mathbb{P}^3,2}}{n}. \quad (\text{S6.7})$$

Following the same analysis for the empirical process  $\pi_1 \mathcal{F}_n := \{f = \pi_1 f : f \in \mathcal{F}_n\}$  as in Lemma 9, we obtain

$$\mathbb{E} [\|\mathbb{U}_{n,1}(f)\|_{\pi_1 \mathcal{F}_n}] \lesssim \frac{\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2} \cdot \sqrt{\log(A'_n \vee n)}}{\sqrt{n}} + \frac{m_n \log(A'_n \vee n)}{n}, \quad (\text{S6.8})$$

where  $A'_n = M_n/m_n$ .

Combining (S6.6)–(S6.8), we conclude

$$\begin{aligned} \mathbb{E} [\|\mathbb{U}_{n,3}(f)\|_{\mathcal{F}_n}] &\lesssim \mathbb{E} [\|\mathbb{U}_{n,3}(\pi_3 f)\|_{\mathcal{F}_n}] + \mathbb{E} [\|\mathbb{U}_{n,2}(\pi_2 f)\|_{\mathcal{F}_n}] + \mathbb{E} [\|\mathbb{U}_{n,1}(\pi_1 f)\|_{\mathcal{F}_n}] \\ &\lesssim \frac{\|F_n\|_{\mathbb{P}^3,2} + m_n \log(A'_n \vee n)}{n} + \frac{\sup_{f \in \mathcal{F}_n} \|\mathbb{P}f\|_{\mathbb{P},2} \cdot \sqrt{\log(A'_n \vee n)}}{\sqrt{n}}, \end{aligned}$$

which completes the proof.  $\square$

We next quote a result from Theorem 3 of [Andrews \(1994\)](#) that will be useful in the proof.

**Lemma 11** (Uniform entropy for sums and products). *Let  $\mathcal{F}$  and  $\mathcal{G}$  be two classes of functions with envelopes  $F$  and  $G$ , respectively. Then, for any  $\epsilon > 0$ , the uniform entropy numbers of  $\mathcal{F}\mathcal{G} = \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$  and  $\mathcal{F} + \mathcal{G} = \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$  satisfy*

$$\begin{aligned} \sup_Q \mathbf{N}(\mathcal{F} + \mathcal{G}, \|\cdot\|_{Q,2}, 2\epsilon\|F + G\|_{Q,2}) &\leq \sup_Q \mathbf{N}(\mathcal{F}, \|\cdot\|_{Q,2}, \epsilon\|F\|_{Q,2}) \cdot \sup_Q \mathbf{N}(\mathcal{G}, \|\cdot\|_{Q,2}, \epsilon\|G\|_{Q,2}), \\ \sup_Q \mathbf{N}(\mathcal{F} \cdot \mathcal{G}, \|\cdot\|_{Q,2}, 2\epsilon\|F \cdot G\|_{Q,2}) &\leq \sup_Q \mathbf{N}(\mathcal{F}, \|\cdot\|_{Q,2}, \epsilon\|F\|_{Q,2}) \cdot \sup_Q \mathbf{N}(\mathcal{G}, \|\cdot\|_{Q,2}, \epsilon\|G\|_{Q,2}). \end{aligned}$$

**Lemma 12** (VC-type characteristics of sums and products). *Suppose  $\mathcal{F}$  and  $\mathcal{G}$  are pointwise measurable measurable and VC type with characteristics  $(A_{\mathcal{F}}, v_{\mathcal{F}})$  and  $(A_{\mathcal{G}}, v_{\mathcal{G}})$ , and envelopes  $F$  and  $G$ , respectively. Then,  $\mathcal{F} + \mathcal{G} = \{f + g : f \in \mathcal{F}, g \in \mathcal{G}\}$  and  $\mathcal{F}\mathcal{G} = \{fg : f \in \mathcal{F}, g \in \mathcal{G}\}$  are VC type with characteristics  $A' := 2(A_{\mathcal{F}}^{v_{\mathcal{F}}} A_{\mathcal{G}}^{v_{\mathcal{G}}})^{1/v'}$  and  $v' := v_{\mathcal{F}} + v_{\mathcal{G}}$ . Specially, when  $\mathcal{G} = \{g\}$ ,  $\mathcal{F} + \mathcal{G}$  and  $\mathcal{F}\mathcal{G}$  will be VC type with characteristics  $(2A_{\mathcal{F}}, v_{\mathcal{F}})$ .*

*Proof.* Using Lemma 11, we have

$$\begin{aligned}
\sup_Q \mathbf{N}(\mathcal{F} + \mathcal{G}, \|\cdot\|_{Q,2}, \epsilon \|F + G\|_{Q,2}) &\leq \sup_Q \mathbf{N}(\mathcal{F}, \|\cdot\|_{Q,2}, \epsilon/2 \|F\|_{Q,2}) \sup_Q \mathbf{N}(\mathcal{G}, \|\cdot\|_{Q,2}, \epsilon/2 \|G\|_{Q,2}) \\
&\leq \left(\frac{2A_{\mathcal{F}}}{\epsilon}\right)^{v_{\mathcal{F}}} \left(\frac{2A_{\mathcal{G}}}{\epsilon}\right)^{v_{\mathcal{G}}} \\
&= \left(\frac{2(A_{\mathcal{F}}^{v_{\mathcal{F}}} A_{\mathcal{G}}^{v_{\mathcal{G}}})^{1/(v_{\mathcal{F}}+v_{\mathcal{G}})}}{\epsilon}\right)^{v_{\mathcal{F}}+v_{\mathcal{G}}},
\end{aligned}$$

which implies  $\mathcal{F} + \mathcal{G}$  is of VC type with characteristics  $A' = 2(A_{\mathcal{F}}^{v_{\mathcal{F}}} A_{\mathcal{G}}^{v_{\mathcal{G}}})^{1/v'}$  and  $v' = v_{\mathcal{F}} + v_{\mathcal{G}}$ .  $\square$

**Lemma 13.** Suppose that function class  $\mathcal{F}_n : \mathcal{O} \rightarrow \mathbb{R}$  is pointwise measurable with envelope function  $F_n$  and is of VC type with characteristics  $(A_n, v_n)$ , satisfying  $A_n \geq e$  and  $v_n \geq 1$ . Suppose that  $\log(\sup_{f \in \mathcal{F}} \|f_n\|_{\mathbb{P},2}/\|F_n\|_{\mathbb{P},2}) \lesssim \log n$  and  $\|F_n(\cdot)\|_{\infty} \lesssim M_n$ . Then

$$\mathbb{E} \left[ \|\mathbb{U}_{n,1}f - \mathbb{P}f\|_{\mathcal{F}_n} \right] \lesssim \frac{\sup_{f \in \mathcal{F}_n} \|f\|_{\mathbb{P},2} \cdot \sqrt{v_n \log(A_n \vee n)}}{\sqrt{n}} + \frac{M_n \cdot v_n \log(A_n \vee n)}{n}.$$

*Proof.* For any constant sequence satisfying  $\log a_n/b_n \lesssim \log n$ , we have the following inequality holds:

$$\begin{aligned}
J_1(a_n/b_n, \mathcal{F}, F) &= \int_0^{a_n/b_n} \sup_Q \{1 + \log \mathbf{N}(\mathcal{F}_n, \|\cdot\|_{Q,2}, \epsilon \|F_n\|_{Q,2})\}^{1/2} d\epsilon \\
&\leq \int_0^{a_n/b_n} \sqrt{1 + v_n \log(A_n/\epsilon)} d\epsilon \\
&\leq \left( \int_0^{a_n/b_n} 1^2 d\epsilon \right)^{1/2} \cdot \left( \int_0^{a_n/b_n} \{1 + v_n \log(A_n/\epsilon)\} d\epsilon \right)^{1/2} \\
&= \sqrt{\frac{a_n}{b_n}} \cdot \left\{ \frac{a_n}{b_n} (1 + v_n \log A_n) + v_n (-\epsilon \log \epsilon + \epsilon) \Big|_0^{a_n/b_n} \right\}^{1/2} \\
&= \sqrt{\frac{a_n}{b_n}} \cdot \left\{ \frac{a_n}{b_n} (1 + v_n \log A_n) + v_n \frac{a_n}{b_n} \{\log(a_n/b_n) + 1\} \right\}^{1/2} \\
&\lesssim \frac{a_n}{b_n} \cdot \sqrt{v_n \log(A_n a_n/b_n)} \lesssim \frac{a_n}{b_n} \cdot \sqrt{v_n \log(A_n \vee n)},
\end{aligned}$$

where the second inequality holds by Cauchy–Schwarz inequality, the penultimate inequality by  $A_n \geq e$ ,  $v_n \geq 1$ , and the last by  $\log a_n/b_n \lesssim \log n$ . Then, applying this result and Theorem 5.1 in Chen and Kato (2020), we obtain

$$\begin{aligned}
\mathbb{E} \left[ \|\mathbb{U}_{n,1}f - \mathbb{P}f\|_{\mathcal{F}_n} \right] &= \frac{J_1(\sup_{f \in \mathcal{F}_n} \|f\|_{\mathbb{P},2}/\|F_n\|_{\mathbb{P},2}, \mathcal{F}_n, F_n) \cdot \|F_n\|_{\mathbb{P},2}}{\sqrt{n}} \\
&\quad + \frac{J_1(\sup_{f \in \mathcal{F}_n} \|f\|_{\mathbb{P},2}/\|F_n\|_{\mathbb{P},2}, \mathcal{F}_n, F_n)^2 \cdot \|F_n(\cdot)\|_{\infty}}{n \cdot \sup_{f \in \mathcal{F}_n} \|f\|_{\mathbb{P},2}^2 / \|F_n\|_{\mathbb{P},2}^2}
\end{aligned}$$

$$\lesssim \frac{\sup_{f \in \mathcal{F}_n} \|f\|_{\mathbb{P},2} \cdot \sqrt{v_n \log(A_n \vee n)}}{\sqrt{n}} + \frac{M_n \cdot v_n \log(A_n \vee n)}{n}.$$

□

**Lemma 14.** *If  $\mathcal{F}_n := \{f : \mathcal{O}^r \rightarrow \mathbb{R}\}$  is of VC type with characteristics  $(A_n, v_n)$  and envelope  $F_n$ , then for every  $k = 1, \dots, r-1$ ,*

(i)  *$\mathbb{P}^{r-k}\mathcal{F}_n$  is also of VC type with characteristics  $(4\sqrt{A_n}, 2v_n)$  and envelope  $\mathbb{P}^{r-k}F_n$ .*

(ii) *Suppose further that  $\sup_{f \in \mathcal{F}_n} \|\mathbb{P}^{r-k}f(\cdot)\|_\infty \leq m_n$  and  $\|\mathbb{P}^{r-k}F_n(\cdot)\|_\infty \leq M_n$ . Then,  $\mathbb{P}^{r-k}\mathcal{F}_n$  is of VC type with characteristics  $(4\sqrt{A_n}M_n/m_n, 2v_n)$  and envelope  $m_n$ , i.e.,*

$$\sup_Q N(\mathbb{P}^{r-k}\mathcal{F}_n, \|\cdot\|_{Q,2}, \tau m_n) \leq \left( \frac{4M_n\sqrt{A_n}}{m_n\tau} \right)^{2v_n}, \text{ for all } 0 < \tau \leq 1.$$

*Proof.* Lemma 5.4 of [Chen and Kato \(2020\)](#) states that for any VC-type function class  $\mathcal{F}$  with characteristics  $(A_n, v_n)$ ,  $\mathbb{P}^{r-k}\mathcal{F}$  satisfies

$$\sup_Q N(\mathbb{P}^{r-k}\mathcal{F}_n, \|\cdot\|_{Q,2}, \epsilon \|\mathbb{P}^{r-k}F_n\|_{Q,2}) \leq (4\sqrt{A_n}/\epsilon)^{2v_n}, 0 < \forall \epsilon \leq 1,$$

which establishes part (i).

For part (ii), set  $\varepsilon = \tau m_n / \|\mathbb{P}^{r-k}F_n\|_{Q,2}$  in the display above. Then, since  $\|\mathbb{P}^{r-k}F_n\|_{Q,2} \leq \|\mathbb{P}^{r-k}F_n(\cdot)\|_\infty \leq M_n$ , we obtain

$$\sup_Q N(\mathbb{P}^{r-k}\mathcal{F}_n, \|\cdot\|_{Q,2}, \tau m_n) \leq \left( \frac{4\|\mathbb{P}^{r-k}F_n\|_{Q,2}\sqrt{A_n}}{m_n\tau} \right)^{2v_n} \leq \left( \frac{4M_n\sqrt{A_n}}{m_n\tau} \right)^{2v_n}.$$

□

## S7 Proof of Results in Section 3.3

Before starting the proof, we first summarize some frequently used results. Under Assumptions 6(iii) and 8(i)–(iii), Lemma 16 of [Liu et al. \(2017\)](#) implies the following uniform bound: for any function  $f \in \mathcal{C}(\mathcal{X} \times \mathcal{T})$ ,

$$\left\| \Pi[f \mid \bar{\phi}_k](\cdot, \cdot) \right\|_\infty \leq \|f(\cdot, \cdot)\|_\infty. \quad (\text{S7.1})$$

Recall that  $\Pi[f \mid \bar{\phi}_k]$  is the projection of  $f$  onto the linear span of basis  $\bar{\phi}_k$ , defined in Section 1.3.

Because the proof carefully tracks changes in  $\beta$ , we now make this dependence explicit by moving  $\beta$  out of subscripts into argument parentheses. For notational simplicity, hereafter we sometimes write  $\Gamma_\beta := \Gamma(\cdot, \cdot, \beta)$  and denote  $\hat{\Xi}_\beta := \hat{\Xi}(O_1, O_2; \beta)$  if this causes no ambiguity.

### S7.1 Proof of Theorem 1

*Proof of Theorem 1.* First, we will show, for any  $\beta \in \mathcal{B}$ ,

$$\|\psi(\beta)\| \gtrsim \|\beta - \beta^*\| \wedge c \quad (\text{S7.2})$$

for some constant  $c$ . Given (S7.2), we obtain  $\mathbb{E}[\|\tilde{\beta}^{(2)} - \beta^*\|] \lesssim \tilde{r}_{n,\beta}$ , if we show that  $\|\psi(\tilde{\beta}^{(2)})\| \lesssim \tilde{r}_{n,\beta}$  with probability approaching one.

By Assumption 5(ii), for any  $\epsilon \leq c_3/2p$  ( $c_3$  is as defined in Assumption 5(ii)), there exists  $\delta > 0$  such that for any  $\beta$  satisfying  $\|\beta - \beta^*\| \leq \delta$ , we have  $\|\nabla_{\beta}\psi(\beta) - \nabla_{\beta}\psi(\beta^*)\|_{\infty} \leq \epsilon$ . Then, for any  $\beta$  satisfying  $\|\beta - \beta^*\| \leq \delta$ , since  $\psi(\beta^*) = 0$ , by Taylor expansion, we have

$$\psi(\beta) = \nabla_{\beta}\psi(\beta^{\dagger})(\beta - \beta^*),$$

where  $\beta^{\dagger}$  lies between  $\beta^*$  and  $\beta$ . Then, by Assumption 5(iii), we have

$$\begin{aligned} \|\nabla_{\beta}\psi(\beta^{\dagger})(\tilde{\beta}^{(2)} - \beta^*)\| &\geq \|\nabla_{\beta}\psi(\beta^*)(\tilde{\beta}^{(2)} - \beta^*)\| - \|\{\nabla_{\beta}\psi(\beta^{\dagger}) - \nabla_{\beta}\psi(\beta^*)\}(\tilde{\beta}^{(2)} - \beta^*)\| \\ &\geq \|\nabla_{\beta}\psi(\beta^*)(\tilde{\beta}^{(2)} - \beta^*)\| - \epsilon p \cdot \|\tilde{\beta}^{(2)} - \beta^*\| \\ &\geq \|\nabla_{\beta}\psi(\beta^*)(\tilde{\beta}^{(2)} - \beta^*)\| - \frac{1}{2}\|\nabla_{\beta}\psi(\beta^*)(\tilde{\beta}^{(2)} - \beta^*)\| \\ &= \frac{1}{2}\|\nabla_{\beta}\psi(\beta^*) \cdot (\tilde{\beta}^{(2)} - \beta^*)\| \geq \frac{c_3}{2}\|\tilde{\beta}^{(2)} - \beta^*\|. \end{aligned}$$

For any  $\|\beta - \beta^*\| \geq \delta$ , by Assumption 4,  $\|\psi(\beta)\| \geq c$  for some constant  $c$ . Hence, combining these two cases, we obtain (S7.2).

Next, we show  $\|\psi(\tilde{\beta}^{(2)})\| \lesssim \tilde{r}_{n,\beta}$  with probability approaching one. Since  $\tilde{\psi}_k^{(2)}(\tilde{\beta}^{(2)}) \equiv 0$  and  $\psi(\beta^*) = 0$ , we decompose  $\|\psi(\tilde{\beta}^{(2)})\|$  as follows

$$\begin{aligned} \|\psi(\tilde{\beta}^{(2)})\| &\leq \|\psi(\tilde{\beta}^{(2)}) - \bar{\psi}_k(\tilde{\beta}^{(2)})\| + \|\bar{\psi}_k(\tilde{\beta}^{(2)}) - \tilde{\psi}_k^{(2)}(\tilde{\beta}^{(2)})\| + \|\tilde{\psi}_k^{(2)}(\tilde{\beta}^{(2)}) - \tilde{\psi}_k^{(2)}(\beta^*)\| \\ &\quad + \|\tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*)\| + \|\bar{\psi}_k(\beta^*) - \psi(\beta^*)\|. \end{aligned}$$

Since  $\tilde{\psi}_k^{(2)}(\tilde{\beta}^{(2)}) \equiv 0$  by the definition of  $\tilde{\beta}^{(2)}$ , we derive

$$\|\tilde{\psi}_k^{(2)}(\tilde{\beta}^{(2)}) - \tilde{\psi}_k^{(2)}(\beta^*)\| = \|\tilde{\psi}_k^{(2)}(\beta^*)\| \leq \|\tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*)\| + \|\bar{\psi}_k(\beta^*) - \psi(\beta^*)\|.$$

Thus

$$\|\psi(\tilde{\beta}^{(2)})\| \leq 3 \sup_{\beta \in \mathcal{B}} \|\psi(\beta) - \bar{\psi}_k(\beta)\| + 3 \sup_{\beta \in \mathcal{B}} \|\tilde{\psi}_k^{(2)}(\beta) - \bar{\psi}_k(\beta)\| =: 3I + 3II.$$

**Convergence Rate for  $I$ :**

$$\begin{aligned}
& \psi(\beta) - \bar{\psi}_k(\beta) \\
&= \mathbb{E} [\xi(X, T) \Gamma(Y, T, \beta)] - \mathbb{E} [\hat{\xi}(X, T) \{ \Gamma(Y, T, \beta) - \hat{b}_{\beta_{\text{init}}}(X, T) \} + \xi(X, T) \hat{b}_{\beta_{\text{init}}}(X, T)] \\
&\quad - \mathbb{E} [\{ \xi(X, T) - \hat{\xi}(X, T) \} \bar{\phi}_k(X, T)^\top] \Sigma_k^{-1} \mathbb{E} [\bar{\phi}_k(X, T) \{ b_\beta(X, T) - \hat{b}_{\beta_{\text{init}}}(X, T) \}] \\
&= \mathbb{E} [\{ \xi(X, T) - \hat{\xi}(X, T) \} \cdot \{ b_\beta(X, T) - \hat{b}_{\beta_{\text{init}}}(X, T) \}] \\
&\quad - \mathbb{E} [\{ \xi(X, T) - \hat{\xi}(X, T) \} \bar{\phi}_k(X, T)^\top] \Sigma_k^{-1} \mathbb{E} [\bar{\phi}_k(X, T) \{ b_\beta(X, T) - \hat{b}_{\beta_{\text{init}}}(X, T) \}] \\
&= \mathbb{E} [\Pi^\perp [\xi - \hat{\xi} \mid \bar{\phi}_k] (X, T) \cdot \Pi^\perp [b_\beta - \hat{b}_{\beta_{\text{init}}} \mid \bar{\phi}_k] (X, T)] = \text{TB}_{\psi, k}(\beta). \tag{S7.3}
\end{aligned}$$

Hence, we obtain

$$I \leq \sup_{\beta \in \mathcal{B}} \|\text{TB}_{\psi, k}(\beta)\| \lesssim \|\Pi^\perp(\hat{\xi} - \xi \mid \bar{\phi}_k)\|_{\mathbb{P}, 2} \cdot \sup_{\beta \in \mathcal{B}} \|\Pi^\perp[b_\beta - \hat{b}_{\beta_{\text{init}}} \mid \bar{\phi}_k]\|_{\mathbb{P}, 2}. \tag{S7.4}$$

**Convergence Rate for  $II$ :** Recalling the definition of  $\text{IF}_{\bar{\psi}_k}^{(2)}$  from Lemma 1 in Section S3, we consider the following decomposition:

$$\begin{aligned}
\tilde{\psi}_k^{(2)}(\beta) - \bar{\psi}_k(\beta) &= \mathbb{U}_{n, 3} \text{IF}_{\bar{\psi}_k}^{(2)} - \bar{\psi}_k(\beta) \\
&= \mathbb{U}_{n, 3} \left( \text{IF}_{\text{B}_{\psi, k}}^{(2)} - \mathbb{E}[\text{IF}_{\text{B}_{\psi, k}}^{(2)}] \right) + \mathbb{U}_{n, 2} \left( \hat{\Xi}_\beta - \mathbb{E}[\hat{\Xi}_\beta] \right). \tag{S7.5}
\end{aligned}$$

We will derive the convergence rate uniformly in  $\beta$  for each term in (S7.5). For the second term in (S7.5), consider the function class  $\mathcal{H}_1$  defined as follows

$$\mathcal{H}_1 := \left\{ h(O_1, O_2, \beta) = \hat{\Xi}(O_1, O_2; \beta) : \beta \in \mathcal{B} \right\}. \tag{S7.6}$$

By Assumption 7,  $\{\Gamma(y, t, \beta) : \beta \in \mathcal{B}\}$  is of VC type with characteristics  $(A, v)$ . By Lemma 12,  $\mathcal{H}_1$  is of VC type with bounded characteristics  $(4A, v)$  and envelope  $H_1(O_1, O_2) = \sup_{\beta \in \mathcal{B}} \|\hat{\Xi}(O_1, O_2; \beta)\|$ . By Assumptions 6 and 7, we have  $\|H_1\|_{\mathbb{P}^2, 2} \lesssim 1$ . By Jensen's inequality, we have  $\|\mathbb{P}H_1\|_{\mathbb{P}, 2} \lesssim \|H_1\|_{\mathbb{P}^2, 2} \lesssim 1$ . Then, using Lemma 9 for function class  $\mathcal{H}_1$ , we obtain

$$\mathbb{E} \left[ \sup_{\beta \in \mathcal{B}} \left\| \mathbb{U}_{n, 2}(\hat{\Xi}_\beta) - \mathbb{P}^2(\hat{\Xi}_\beta) \right\| \right] \lesssim \frac{1}{\sqrt{n}}. \tag{S7.7}$$

To control the first term in (S7.5), we need to consider the function class

$$\mathcal{H}_{2, n} := \left\{ h(O_1, O_2, O_3, \beta) = \text{IF}_{\text{B}_{\psi, k}}^{(2)}(O_1, O_2, O_3; \beta) : \beta \in \mathcal{B} \right\}. \tag{S7.8}$$

Then, by Assumption 7 and Lemma 12,  $\mathcal{H}_{2, n}$  is of VC type with bounded characteristics  $(4A, v)$

and envelope  $H_{2,n}$ , which is defined as follows

$$H_{2,n}(O_1, O_2, O_3) := \left| \left\{ \widehat{\xi}(X_1, T_1) \bar{\phi}_k(X_1, T_1) - \bar{\mathbf{z}}_{k_x}(X_1) \otimes \bar{\mathbf{w}}_{k_t}(T_3) \right\}^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \right| \\ \times \sup_{\beta \in \mathcal{B}} \left\| \Gamma(Y_2, T_2, \beta) - \widehat{b}_{\beta_{\text{init}}}(X_2, T_2) \right\|.$$

Using Assumptions 6, 7 and 8, we have  $\|\mathbb{P}^2 H_{2,n}(\cdot)\|_\infty \lesssim \|H_{2,n}(\cdot, \cdot, \cdot)\|_\infty \lesssim \zeta(k)^2 \lesssim k$ , and

$$\|H_{2,n}\|_{\mathbb{P}^3, 2} = \left\{ \mathbb{E} \left[ \|H_{2,n}(O_1, O_2, O_3)\|^2 \right] \right\}^{1/2} \\ \lesssim \left\{ \mathbb{E} \left[ \bar{\phi}_k(X_2, T_2)^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \sup_{\beta \in \mathcal{B}} \left\| \Gamma(Y_2, T_2, \beta) - \widehat{b}_{\beta_{\text{init}}}(X_2, T_2) \right\|^2 \right] \right\}^{1/2} \\ \lesssim \zeta(k) \lesssim \sqrt{k}.$$

For any  $h \in \mathcal{H}_{2,n}$ ,  $\mathbb{P}^2 h$  can be written explicitly as

$$\mathbb{P}^2 h(O) = \frac{1}{3} \widehat{\xi}(X, T) \Pi[b_\beta - \widehat{b}_{\beta_{\text{init}}} | \bar{\phi}_k](X, T) - \frac{1}{3} \int_{t \in \mathcal{T}} \Pi[b_\beta - \widehat{b}_{\beta_{\text{init}}} | \bar{\phi}_k](X, t) \mathbf{p}_T(t) dt \\ + \frac{1}{3} \Pi[\widehat{\xi} - \xi | \bar{\phi}_k](X, T) (\Gamma_\beta - \widehat{b}_{\beta_{\text{init}}}) - \frac{1}{3} \int_{x \in \mathcal{X}} \Pi[b_\beta - \widehat{b}_{\beta_{\text{init}}} | \bar{\phi}_k](x, T) \mathbf{p}_X(x) dx, \quad (\text{S7.9})$$

which is a combination of projection terms. By Assumptions 6 and 7, we have  $\|\xi(\cdot, \cdot)\|_\infty \lesssim 1$ ,  $\|\widehat{\xi}(\cdot, \cdot)\|_\infty \lesssim 1$ ,  $\sup_{\beta \in \mathcal{B}} \|\Gamma(\cdot, \cdot, \beta)\|_\infty \lesssim 1$ , and  $\sup_{\beta \in \mathcal{B}} \|\widehat{b}_\beta(\cdot, \cdot)\|_\infty \lesssim 1$ . Thus, using the result in (S7.1), we obtain

$$\sup_{h \in \mathcal{H}_{2,n}} \|\mathbb{P}^2 h(\cdot)\|_\infty \lesssim \sup_{\beta \in \mathcal{B}} \left\| \Pi[b_\beta - \widehat{b}_{\beta_{\text{init}}} | \bar{\phi}_k](\cdot, \cdot) \right\|_\infty + \|\Pi[\widehat{\xi} - \xi | \bar{\phi}_k](\cdot, \cdot)\|_\infty \\ \lesssim \sup_{\beta \in \mathcal{B}} \|(b_\beta - \widehat{b}_{\beta_{\text{init}}})(\cdot, \cdot)\|_\infty + \|(\widehat{\xi} - \xi)(\cdot, \cdot)\|_\infty \lesssim 1.$$

This also implies  $\sup_{h \in \mathcal{H}_{2,n}} \|\mathbb{P}^2 h\|_{\mathbb{P}, 2} \lesssim 1$ . Hence, applying Lemma 10 to  $\mathcal{H}_{2,n}$ , we obtain

$$\mathbb{E} \left[ \sup_{\beta \in \mathcal{B}} \left\| \mathbb{U}_{n,3} \left( \text{IF}_{\text{B}_{\psi,k}}^{(2)} \right) - \mathbb{P}^3 \left( \text{IF}_{\text{B}_{\psi,k}}^{(2)} \right) \right\| \right] \lesssim \frac{\sqrt{k} \log n}{n} + \frac{\log n}{\sqrt{n}}. \quad (\text{S7.10})$$

Hence, by (S7.5), (S7.7), (S7.10), we have

$$II = \sup_{\beta \in \mathcal{B}} \left\| \widetilde{\psi}_k^{(2)}(\beta) - \bar{\psi}_k(\beta) \right\| = O_{\mathbb{P}} \left( \frac{\sqrt{k} \log n}{n} \vee \frac{\log n}{\sqrt{n}} \right). \quad (\text{S7.11})$$

Combining (S7.4) and (S7.11), we obtain

$$\begin{aligned}\|\psi(\tilde{\beta}^{(2)})\| &= O_{\mathbb{P}}\left(\frac{\sqrt{k} \log n}{n} \vee \frac{\log n}{\sqrt{n}} + \|\Pi^{\perp}(\hat{\xi} - \xi|\bar{\phi}_k)\|_{\mathbb{P},2} \cdot \sup_{\beta \in \mathcal{B}} \|\Pi^{\perp}[b_{\beta} - \hat{b}_{\beta_{\text{init}}|\bar{\phi}_k}]\|_{\mathbb{P},2}\right) \\ &= O_{\mathbb{P}}(\tilde{r}_{n,\beta}),\end{aligned}$$

which completes the proof.  $\square$

## S7.2 Proof of Theorem 3

*Proof of Theorem 3.* By construction, we have  $\hat{\psi}_k^{(2)}(\hat{\beta}^{(2)}) = 0$  and  $\psi(\beta^*) = 0$ . Consider the following decomposition:

$$\begin{aligned}0 &= \hat{\psi}_k^{(2)}(\hat{\beta}^{(2)}) = \hat{\psi}_k^{(2)}(\hat{\beta}^{(2)}) - \mathbb{E}[\hat{\psi}_k^{(2)}(\hat{\beta}^{(2)})] + \mathbb{E}[\hat{\psi}_k^{(2)}(\hat{\beta}^{(2)})] - \mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] \\ &\quad + \mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] - \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)] + \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)] - \psi(\beta^*) \\ &\leq \sup_{\beta \in \mathcal{B}} \left\| \hat{\psi}_k^{(2)}(\beta) - \mathbb{E}[\hat{\psi}_k^{(2)}(\beta)] \right\| + \left\| \mathbb{E}[\hat{\psi}_k^{(2)}(\hat{\beta}^{(2)})] - \mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] \right\| \\ &\quad + \left\| \mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] - \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)] \right\| + \left\| \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)] - \psi(\beta^*) \right\|.\end{aligned}$$

Using the condition that for some constant  $c > 0$ ,  $\|\mathbb{E}[\hat{\psi}_k^{(2)}(\beta_1)] - \mathbb{E}[\hat{\psi}_k^{(2)}(\beta_2)]\| \geq c\|\beta_1 - \beta_2\|$ , it follows that

$$\begin{aligned}\|\hat{\beta}^{(2)} - \tilde{\beta}^{(2)}\| &\lesssim \sup_{\beta \in \mathcal{B}} \left\| \hat{\psi}_k^{(2)}(\beta) - \mathbb{E}[\hat{\psi}_k^{(2)}(\beta)] \right\| + \left\| \mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] - \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)] \right\| \\ &\quad + \left\| \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)] - \psi(\beta^*) \right\|.\end{aligned}$$

Conditioning on the nuisance sample, the analysis of  $\sup_{\beta \in \mathcal{B}} \|\hat{\psi}_k^{(2)}(\beta) - \mathbb{E}[\hat{\psi}_k^{(2)}(\beta)]\|$  closely parallels that of the term  $\sup_{\beta \in \mathcal{B}} \|\tilde{\psi}_k^{(2)}(\beta) - \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta)]\|$ , so both terms attain the same convergence rate. Thus, by the result in (S7.11), we have, with probability approaching one,

$$\sup_{\beta \in \mathcal{B}} \left\| \hat{\psi}_k^{(2)}(\beta) - \mathbb{E}[\hat{\psi}_k^{(2)}(\beta)] \right\| \lesssim \frac{\sqrt{k} \log n}{n} \vee \frac{\log n}{\sqrt{n}} \lesssim \tilde{r}_{n,\beta}.$$

By (S7.3), we have  $\mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] - \psi(\beta^*) = \text{TB}_{\psi,k}(\beta^*)$ . Hence  $\|\mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] - \psi(\beta^*)\| = \|\text{TB}_{\psi,k}(\beta^*)\| \lesssim \tilde{r}_{n,\beta}$ . It remains to control the term  $\mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] - \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)]$ .

$$\begin{aligned}&\left\| \mathbb{E}[\hat{\psi}_k^{(2)}(\beta^*)] - \mathbb{E}[\tilde{\psi}_k^{(2)}(\beta^*)] \right\| \\ &= \left\| \mathbb{E} \left[ \left\{ \xi(X, T) - \hat{\xi}(X, T) \right\} \bar{\phi}_k(X, T)^{\top} \right] \left\{ \hat{\Sigma}_k^{-1} - \Sigma_k^{-1} \right\} \mathbb{E} \left[ \bar{\phi}_k(X, T) \left\{ b_{\beta^*}(X, T) - \hat{b}_{\beta_{\text{init}}}(X, T) \right\} \right] \right\|\end{aligned}$$

$$\leq \|\xi - \hat{\xi}\|_{\mathbb{P},2} \cdot \|b_{\beta^*} - \hat{b}_{\beta_{\text{init}}}\|_{\mathbb{P},2} \cdot \|\hat{\Sigma}_k - \Sigma\|_{\text{op}}.$$

Combining all terms, we obtain the desired bound:

$$\mathbb{E}[\|\hat{\beta}^{(2)} - \tilde{\beta}^{(2)}\|] \lesssim \tilde{r}_{n,\beta} + \|\xi - \hat{\xi}\|_{\mathbb{P},2} \cdot \|b_{\beta^*} - \hat{b}_{\beta_{\text{init}}}\|_{\mathbb{P},2} \cdot \|\hat{\Sigma}_k - \Sigma_k\|_{\text{op}}.$$

By Assumption 6(ii), we have

$$\|b_{\beta^*} - \hat{b}_{\beta_{\text{init}}}\|_{\mathbb{P},2} \leq \|b_{\beta^*} - b_{\beta_{\text{init}}}\|_{\mathbb{P},2} + \|b_{\beta_{\text{init}}} - \hat{b}_{\beta_{\text{init}}}\|_{\mathbb{P},2} \lesssim \|\beta^* - \beta_{\text{init}}\| + \|b_{\beta_{\text{init}}} - \hat{b}_{\beta_{\text{init}}}\|_{\mathbb{P},2}.$$

Hence, when Assumption 9 holds,

$$\|\xi - \hat{\xi}\|_{\mathbb{P},2} \cdot \|b_{\beta^*} - \hat{b}_{\beta_{\text{init}}}\|_{\mathbb{P},2} \cdot \|\hat{\Sigma}_k - \Sigma_k\|_{\text{op}} = o(n^{-1/2}) \leq r_{n,\tilde{\beta}^{(2)}},$$

which implies  $\mathbb{E}[\|\hat{\beta}^{(2)} - \beta^*\|] \lesssim \tilde{r}_{n,\beta}$ . □

### S7.3 Proof of Theorem 2

The matrices  $V_1(\beta)$  and  $V_2(\beta)$  that appear in Theorem 2 are defined by

$$V_1(\beta) := \lim_{n \rightarrow \infty} \frac{n}{n+k} \mathbb{E} \left[ \left\{ 2\pi_1 \hat{\Xi}_\beta + 3\pi_1 \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) \right\}^{\otimes 2} \right] \quad (\text{S7.12})$$

$$\text{and } V_2(\beta) := \lim_{n \rightarrow \infty} \frac{1}{k+n} \mathbb{E} \left[ \left\{ \pi_2 \hat{\Xi}_\beta + 3\pi_2 \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) \right\}^{\otimes 2} \right]. \quad (\text{S7.13})$$

*Proof of Theorem 2.* By Assumption 5,  $\psi(\beta)$  is differentiable with respect to  $\beta$ . Using Mean Value Theorem, we obtain

$$0 = \psi(\beta^*) = \psi(\tilde{\beta}^{(2)}) + \nabla_\beta \psi(\beta^\dagger) \cdot (\tilde{\beta}^{(2)} - \beta^*),$$

where  $\beta^\dagger$  lies between  $\beta^*$  and  $\tilde{\beta}^{(2)}$ . Because  $\nabla_\beta \psi(\beta)$  is continuous in  $\beta$  at  $\beta^*$ , and  $\|\tilde{\beta}^{(2)} - \beta^*\| = o_{\mathbb{P}}(1)$ , we have

$$n/\sqrt{k+n}(\tilde{\beta}^{(2)} - \beta^*) = -[\nabla_\beta \psi(\beta^*)]^{-1} \cdot n/\sqrt{k+n} \cdot \psi(\tilde{\beta}^{(2)}) + o_{\mathbb{P}}(1). \quad (\text{S7.14})$$

Since  $0 = \psi(\beta^*)$  and  $\tilde{\psi}_k^{(2)}(\tilde{\beta}^{(2)}) = 0$ , we have the following decomposition

$$\frac{n}{\sqrt{k+n}} \psi(\tilde{\beta}^{(2)}) = \frac{n}{\sqrt{k+n}} \left\{ \psi(\tilde{\beta}^{(2)}) - \bar{\psi}_k(\tilde{\beta}^{(2)}) \right\} \quad (\text{S7.15})$$

$$+ \frac{n}{\sqrt{k+n}} \left\{ \bar{\psi}_k(\tilde{\beta}^{(2)}) - \tilde{\psi}_k^{(2)}(\tilde{\beta}^{(2)}) - \left( \bar{\psi}_k(\beta^*) - \tilde{\psi}_k^{(2)}(\beta^*) \right) \right\} \quad (\text{S7.16})$$

$$- \frac{n}{\sqrt{k+n}} \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\}. \quad (\text{S7.17})$$

By (S7.3), (11) and  $\|\tilde{\beta}^{(2)} - \beta^*\| = O_{\mathbb{P}}(\tilde{r}_{n,\beta})$ , the right-hand side of (S7.15) is  $o_{\mathbb{P}}(1)$ . By (11) and Lemma 16, (S7.17) converges to a normal distribution. Combining this and (S7.14), we obtain the desired distribution of  $\tilde{\beta}^{(2)}$  as stated in Theorem 2. To complete the proof, it remains to show (S7.16), which is a  $U$ -process indexed by  $\tilde{\beta}^{(2)}$ , is  $o_{\mathbb{P}}(1)$ . Since we have shown in Section S7.1 that  $\|\tilde{\beta}^{(2)} - \beta^*\| \lesssim \tilde{r}_{n,\beta}$  with probability  $1 - o(1)$ , it remains to show that, conditioned on this event, we have

$$\begin{aligned} & \mathbb{P} \left( \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \frac{n}{\sqrt{k+n}} \cdot \left\| \left\{ \bar{\psi}_k(\beta) - \tilde{\psi}_k^{(2)}(\beta) \right\} - \left\{ \bar{\psi}_k(\beta^*) - \tilde{\psi}_k^{(2)}(\beta^*) \right\} \right\| > \epsilon \right) \\ &= o(1). \end{aligned} \quad (\text{S7.18})$$

Consider the following decomposition

$$\begin{aligned} & \left\{ \tilde{\psi}_k^{(2)}(\beta) - \bar{\psi}_k(\beta) \right\} - \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} \\ &= \mathbb{U}_{n,3} \left( \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) - \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta^*) - \mathbb{E}[\text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) - \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta^*)] \right) + \mathbb{U}_{n,2} \left( \hat{\Xi}_{\beta} - \hat{\Xi}_{\beta^*} - \mathbb{E}[\hat{\Xi}_{\beta} - \hat{\Xi}_{\beta^*}] \right). \end{aligned} \quad (\text{S7.19})$$

First, we define the function class

$$\mathcal{H}_{1,\delta} := \left\{ h(O, \beta) = \hat{\Xi}_{\beta} - \hat{\Xi}_{\beta^*} = \hat{\xi}(X, T)[\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta^*)] : \beta \in \mathcal{B}, \|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta} \right\}. \quad (\text{S7.20})$$

By Assumption 7 and Lemma 12, the function class  $\mathcal{H}_{1,\delta}$  of VC type with bounded characteristics  $(8A, v)$  and envelope  $H_{1,\delta}(O) := \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \|\hat{\xi}(X, T)[\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta^*)]\|$ . Since  $\|\hat{\xi}(\cdot, \cdot)\|_{\infty} \lesssim 1$  under Assumption 6, we obtain

$$\|H_{1,\delta}\|_{\mathbb{P}^2,2} \lesssim \mathbb{E}^{1/2} \left[ \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \|\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta^*)\|^2 \right] \lesssim (\tilde{r}_{n,\beta})^{\alpha_0},$$

where the last inequality holds by Assumptions 7. Applying Theorem 2.14.1 of van der Vaart and Wellner (2023) to  $\mathcal{H}_{1,\delta}$ , we obtain

$$\mathbb{E} \left[ \left\| \mathbb{U}_{n,1}(h) - \mathbb{P}^2(h) \right\|_{\mathcal{H}_{1,\delta}} \right] \lesssim \frac{J_1(1, \mathcal{H}_{1,\delta}, H_{1,\delta}) \cdot \|H_{1,\delta}\|_{\mathbb{P},2}}{\sqrt{n}} \lesssim \frac{(\tilde{r}_{n,\beta})^{\alpha_0}}{\sqrt{n}}. \quad (\text{S7.21})$$

Similarly, we define the following function class

$$\mathcal{H}_{2,n,\delta} := \left\{ h(O_1, O_2, O_3, \beta) = \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) - \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta^*) : \beta \in \mathcal{B}, \|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta} \right\}. \quad (\text{S7.22})$$

Using arguments analogous to those used for  $\mathcal{H}_{1,\delta}$ , the function class is also VC type with bounded

characteristics, and its envelope function  $H_{2,n,\delta}$  is defined as

$$H_{2,n,\delta}(O_1, O_2, O_3) := \left| \left\{ \widehat{\xi}(X_1, T_1) \bar{\phi}_k(X_1, T_1) - \bar{\mathbf{z}}_{k_x}(X_1) \otimes \bar{\mathbf{w}}_{k_t}(T_3) \right\}^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \right| \\ \times \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \|\Gamma(Y_2, T_2, \beta) - \Gamma(Y_2, T_2, \beta^*)\|.$$

Note that,  $\|\xi(\cdot, \cdot)\|_\infty \lesssim 1$  and  $\|\widehat{\xi}(\cdot, \cdot)\|_\infty \lesssim 1$  under Assumption 6. Using Assumptions 7 and 8, we have  $\|\mathbb{P}^2 H_{2,n,\delta}(\cdot)\|_\infty \lesssim \|H_{2,n,\delta}(\cdot, \cdot, \cdot)\|_\infty \lesssim \zeta(k)^2 \lesssim k$  and

$$\begin{aligned} & \mathbb{E} \left[ \|H_{2,n,\delta}(O_1, O_2, O_3)\|^2 \right] \\ & \lesssim \mathbb{E} \left[ \bar{\phi}_k(X_2, T_2)^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \|\Gamma(Y_2, T_2, \beta) - \Gamma(Y_2, T_2, \beta^*)\|^2 \right] \\ & \lesssim \zeta(k)^2 \cdot \mathbb{E} \left[ \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \|\Gamma(Y_2, T_2, \beta) - \Gamma(Y_2, T_2, \beta^*)\|^2 \right] \\ & \lesssim k \cdot (\tilde{r}_{n,\beta})^{2\alpha_0}, \end{aligned}$$

which implies  $\|H_{2,n,\delta}\|_{\mathbb{P}^3,2} \lesssim \sqrt{k} \cdot (\tilde{r}_{n,\beta})^{\alpha_0}$ .

For any  $h \in \mathcal{H}_{2,n,\delta}$ , the explicit expression of  $\mathbb{P}^2 h$  is given by

$$\begin{aligned} \mathbb{P}^2 h(O) &= \frac{1}{3} \widehat{\xi}(X, T) \Pi[\Gamma_\beta - \Gamma_{\beta^*} | \bar{\phi}_k](X, T) - \frac{1}{3} \int_{t \in \mathcal{T}} \Pi[\Gamma_\beta - \Gamma_{\beta^*} | \bar{\phi}_k](X, t) \mathbf{p}_T(t) dt \\ &\quad + \frac{1}{3} \Pi[\widehat{\xi} - \xi | \bar{\phi}_k](X, T) (\Gamma_\beta - \Gamma_{\beta^*}) - \frac{1}{3} \int_{x \in \mathcal{X}} \Pi[\Gamma_\beta - \Gamma_{\beta^*} | \bar{\phi}_k](x, T) \mathbf{p}_X(x) dx. \end{aligned}$$

Since  $\|\xi(\cdot, \cdot)\|_\infty \lesssim 1$ ,  $\|\widehat{\xi}(\cdot, \cdot)\|_\infty \lesssim 1$  and projections do not increase length, using Jensen's inequality, we obtain

$$\sup_{h \in \mathcal{H}_{2,n,\delta}} \|\mathbb{P}^2 h\|_{\mathbb{P},2} \lesssim \mathbb{E}^{1/2} \left[ \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \|\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta^*)\|^2 \right] \lesssim (\tilde{r}_{n,\beta})^{\alpha_0},$$

where the last inequality holds by Assumption 7. By Assumptions 6, 7, and (S7.1), we have

$$\sup_{\beta} \left\| \Pi[\Gamma_\beta | \bar{\phi}_k](\cdot, \cdot) \right\|_\infty \leq \sup_{\beta} \|\Gamma_\beta(\cdot, \cdot)\|_\infty \lesssim 1, \text{ and } \left\| \Pi[\widehat{\xi} - \xi | \bar{\phi}_k](\cdot, \cdot) \right\|_\infty \lesssim \|(\widehat{\xi} - \xi)(\cdot, \cdot)\|_\infty \lesssim 1,$$

which implies  $\sup_{h \in \mathcal{H}_{2,n,\delta}} \|\mathbb{P}^2 h(\cdot)\|_\infty \lesssim 1$ . Hence, applying Lemma 10, we obtain

$$\mathbb{E} \left[ \left\| \mathbb{U}_{n,3}(h) - \mathbb{P}^3 h \right\|_{\mathcal{H}_{2,n,\delta}} \right] \lesssim \frac{\sqrt{k} \cdot (\tilde{r}_{n,\beta})^{\alpha_0} \log n + \log n}{n} + \frac{(\tilde{r}_{n,\beta})^{\alpha_0} \sqrt{\log n}}{\sqrt{n}}. \quad (\text{S7.23})$$

Hence, combining (S7.19), (S7.21), and (S7.23), we obtain

$$\begin{aligned}
& \mathbb{E} \left[ \sup_{\|\beta - \beta^*\| \lesssim \tilde{r}_{n,\beta}} \frac{n}{\sqrt{k+n}} \cdot \left\| \left( \bar{\psi}_k(\beta) - \tilde{\psi}_k^{(2)}(\beta) \right) - \left( \bar{\psi}_k(\beta^*) - \tilde{\psi}_k^{(2)}(\beta^*) \right) \right\| \right] \\
& \lesssim \frac{n}{\sqrt{k+n}} \cdot \left[ \frac{1}{\sqrt{n}} + \frac{\sqrt{k}}{n} \right] \cdot (\tilde{r}_{n,\beta})^{\alpha_0} \log n \\
& = \left( \sqrt{\frac{n}{k+n}} + \sqrt{\frac{k}{k+n}} \right) \cdot (\tilde{r}_{n,\beta})^{\alpha_0} \log n = o_{\mathbb{P}}(1),
\end{aligned}$$

where the last equality holds since we assume  $(\tilde{r}_{n,\beta})^{\alpha_0} \log n \rightarrow 0$ . Hence, we have shown that (S7.18) is  $o(1)$ , which completes the proof.  $\square$

#### S7.4 Technical Lemmas for Results in Section 3.3

**Lemma 15.** *Suppose that Assumptions 6–8 hold. Then, for a given  $\beta \in \mathcal{B}$ , conditioning on the nuisance sample,*

$$\left\| \text{cov} \left[ \mathbb{U}_{n,3} \left( \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) \right) \right] \right\|_{\text{op}}^2 = O \left( \frac{1}{n} + \frac{k}{n^2} \right),$$

where  $\text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta)$  is defined in Lemma 1.

*Proof.* Taking a unit-norm vector  $v \in \mathbb{R}^q$  ( $\|v\| = 1$ ), we need to show

$$v^\top \text{cov} \left[ \mathbb{U}_{n,3} \left( \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) \right) \right] v = \text{var} \left[ \mathbb{U}_{n,3} \left( v^\top \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) \right) \right] = O \left( \frac{1}{n} + \frac{k}{n^2} \right).$$

By using Lemma 15 in Liu et al. (2017),

$$\begin{aligned}
& \text{var} \left[ \mathbb{U}_{n,3} \left( v^\top \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) \right) \right] = \text{var} \left[ \mathbb{U}_{n,3} \left\{ S_3 \left( v^\top \text{IF}_{\text{B}_{\psi,k}}^{(2)}(\beta) \right) \right\} \right] \\
& \lesssim \sum_{j=1}^3 \frac{1}{n^j} \mathbb{E} \left[ \left\{ \mathbb{E} \left[ S_3 \left( v^\top \text{IF}_{\text{B}_{\psi,k}}^{(2)}(O_1, O_2, O_3; \beta) \right) \mid O_1, \dots, O_j \right] \right\}^2 \right] =: \sum_{j=1}^3 \zeta_j,
\end{aligned}$$

where  $S_3$  is the symmetrization operator defined in Section S6.1. For  $\zeta_1$ , first note that

$$\begin{aligned}
& \mathbb{E}[\text{IF}_{\text{B}_{\psi,k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_1] \\
& = \left\{ \hat{\xi}(X_1, T_1) \bar{\phi}_k(X_1, T_1) - \bar{\mathbf{z}}_{k_x}(X_1) \otimes \mathbb{E}[\bar{\mathbf{w}}_{k_t}(T)] \right\}^\top \Sigma_k^{-1} \mathbb{E} \left[ \bar{\phi}_k(X, T) \left\{ \hat{b}_\beta(X, T) - b_\beta(X, T) \right\}^\top v \right] \\
& = \hat{\xi}(X_1, T_1) \Pi[\{\hat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](X_1, T_1) - \int_{t \in \mathcal{T}} \Pi[\{\hat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](X_1, t) \mathbf{p}_T(t) dt.
\end{aligned}$$

Then

$$\begin{aligned} & \mathbb{E} \left[ \left| \widehat{\xi}(X_1, T_1) \Pi[\{\widehat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](X_1, T_1) \right|^2 \right] \\ & \lesssim \mathbb{E} \left[ \left| \Pi[\{\widehat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](X_1, T_1) \right|^2 \right] \lesssim \mathbb{E} \left[ \left| \{\widehat{b}_\beta - b_\beta\}^\top v \right|^2 \right] \lesssim 1, \end{aligned}$$

where the second inequality follows by the  $L_2$ -norm contraction property of  $L_2$ -projection and the last inequality follows by Assumption 6. Similarly, by Jensen's inequality, we have

$$\begin{aligned} & \mathbb{E} \left[ \left| \int_{\mathcal{T}} \Pi[\{\widehat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](X_1, t) \mathbf{p}_T(t) dt \right|^2 \right] \\ & \leq \iint_{\mathcal{X} \times \mathcal{T}} \left| \Pi[\{\widehat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](x, t) \right|^2 \mathbf{p}_T(t) \mathbf{p}_X(x) dt dx \\ & = \mathbb{E} \left[ \xi(X_1, T_1) \left| \Pi[\{\widehat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](X_1, T_1) \right|^2 \right] \\ & \lesssim \mathbb{E} \left[ \left| \Pi[\{\widehat{b}_\beta - b_\beta\}^\top v | \bar{\phi}_k](X_1, T_1) \right|^2 \right] \lesssim 1. \end{aligned}$$

Hence,

$$\mathbb{E} \left[ \left\{ \mathbb{E}[\mathbf{IF}_{B_{\psi,k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_1] \right\}^2 \right] = O(1).$$

By symmetry, we also have

$$\begin{aligned} & \mathbb{E} \left[ \left\{ \mathbb{E}[\mathbf{IF}_{B_{\psi,k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_2] \right\}^2 \right] = O(1), \\ & \mathbb{E} \left[ \left\{ \mathbb{E}[\mathbf{IF}_{B_{\psi,k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_3] \right\}^2 \right] = O(1). \end{aligned}$$

Taking these together implies  $\zeta_1 \lesssim \frac{1}{n}$ .

For  $\zeta_2$ , first note that

$$\begin{aligned} & \mathbb{E}[\mathbf{IF}_{B_{\psi,k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_1, O_2] \\ & = \left\{ \widehat{\xi}(X_1, T_1) \bar{\phi}_k(X_1, T_1) - \bar{\mathbf{z}}_{k_x}(X_1) \otimes \int_{\mathcal{T}} \bar{\mathbf{w}}_{k_t}(t) \mathbf{p}_T(t) dt \right\}^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \left\{ \widehat{b}_\beta(X_2, T_2) - \Gamma(Y_2, T_2, \beta) \right\}^\top v. \end{aligned}$$

By Assumptions 6–8, we have

$$\begin{aligned} & \mathbb{E} \left[ \left| \left\{ \bar{\mathbf{z}}_{k_x}(X_1) \otimes \int_{\mathcal{T}} \bar{\mathbf{w}}_{k_t}(t) \mathbf{p}_T(t) dt \right\}^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \right|^2 v^\top \left\{ \widehat{b}_\beta(X_2, T_2) - \Gamma(Y_2, T_2, \beta) \right\}^{\otimes 2} v \right] \\ & \lesssim \mathbb{E} \left[ \left| \left\{ \bar{\mathbf{z}}_{k_x}(X_1) \otimes \int_{\mathcal{T}} \bar{\mathbf{w}}_{k_t}(t) \mathbf{p}_T(t) dt \right\}^\top \Sigma_k^{-1} \left\{ \bar{\mathbf{z}}_{k_x}(X_1) \otimes \int_{\mathcal{T}} \bar{\mathbf{w}}_{k_t}(t) \mathbf{p}_T(t) dt \right\} \right|^2 \right] \end{aligned}$$

$$\begin{aligned}
&\lesssim \lambda_{\max}(\Sigma_k^{-1}) \mathbb{E} \left[ \left( \bar{\mathbf{z}}_{k_x}(X_{i_1}) \otimes \int_{\mathcal{T}} \bar{\mathbf{w}}_{k_t}(t) \mathbf{p}_T(t) dt \right)^\top \left( \bar{\mathbf{z}}_{k_x}(X_{i_1}) \otimes \int_{\mathcal{T}} \bar{\mathbf{w}}_{k_t}(t) \mathbf{p}_T(t) dt \right) \right] \\
&\lesssim \lambda_{\max}(\Sigma_k^{-1}) \mathbb{E} \left[ \bar{\mathbf{z}}_{k_x}(X)^\top \bar{\mathbf{z}}_{k_x}(X) \right] \mathbb{E} \left[ \bar{\mathbf{w}}_{k_t}(T)^\top \bar{\mathbf{w}}_{k_t}(T) \right] \lesssim k_x k_t = k.
\end{aligned}$$

where  $\lambda_{\max}(\Sigma_k^{-1})$  is the largest eigenvalue of matrix  $\Sigma_k^{-1}$ . Similarly, we derive

$$\begin{aligned}
&\mathbb{E} \left[ \left| \bar{\phi}_k(X_1, T_1)^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \right|^2 v^\top \left\{ \hat{b}_\beta(X_2, T_2) - \Gamma(Y_2, T_2, \beta) \right\}^{\otimes 2} v \right] \\
&\lesssim \mathbb{E} \left[ \bar{\phi}_k(X_1, T_1)^\top \Sigma_k^{-1} \bar{\phi}_k(X_1, T_1) \right] \\
&\lesssim \lambda_{\max}(\Sigma_k^{-1}) \mathbb{E} \left[ \bar{\phi}_k(X, T)^\top \bar{\phi}_k(X, T) \right] \\
&= \lambda_{\max}(\Sigma_k^{-1}) \iint_{\mathcal{X} \times \mathcal{T}} \xi(x, t)^{-1} \left( \bar{\mathbf{z}}_{k_x}(x)^\top \bar{\mathbf{z}}_{k_x}(x) \right) \otimes \left( \bar{\mathbf{w}}_{k_t}(t)^\top \bar{\mathbf{w}}_{k_t}(t) \right) \mathbf{p}_T(t) \mathbf{p}_X(x) dt dx \\
&\lesssim \lambda_{\max}(\Sigma_k^{-1}) \mathbb{E} \left[ \bar{\mathbf{z}}_{k_x}(X)^\top \bar{\mathbf{z}}_{k_x}(X) \right] \mathbb{E} \left[ \bar{\mathbf{w}}_{k_t}(T)^\top \bar{\mathbf{w}}_{k_t}(T) \right] \lesssim k.
\end{aligned}$$

Thus

$$\mathbb{E} \left[ \left| \mathbb{E}[\mathbf{IF}_{\mathbf{B}_{\psi, k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_1, O_2] \right|^2 \right] = O(k).$$

Similarly, we have

$$\begin{aligned}
&\mathbb{E} \left[ \left| \mathbb{E}[\mathbf{IF}_{\mathbf{B}_{\psi, k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_1, O_3] \right|^2 \right] = O(k), \\
&\mathbb{E} \left[ \left| \mathbb{E}[\mathbf{IF}_{\mathbf{B}_{\psi, k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \mid O_2, O_3] \right|^2 \right] = O(k).
\end{aligned}$$

Hence,

$$\zeta_2 \lesssim \frac{k}{n^2}.$$

For  $\zeta_3$ , we have

$$\begin{aligned}
\zeta_3 &\lesssim \frac{1}{n^3} \mathbb{E} \left[ \left( \mathbf{IF}_{\mathbf{B}_{\psi, k}}^{(2)}(O_1, O_2, O_3; \beta)^\top v \right)^2 \right] \\
&= \frac{1}{n^3} \mathbb{E} \left[ \left| \left\{ \hat{\xi}(X_1, T_1) \bar{\phi}_k(X_1, T_1) - \bar{\mathbf{z}}_{k_x}(X_1) \otimes \bar{\mathbf{w}}_{k_t}(T_3) \right\}^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) \right|^2 \right. \\
&\quad \left. \times v^\top \left\{ \hat{b}_\beta(X_2, T_2) - \Gamma(Y_2, T_2, \beta) \right\}^{\otimes 2} v \right] \\
&\lesssim \frac{1}{n^3} \mathbb{E} \left[ \left\{ \hat{\xi}(X, T) - \xi(X, T) \right\}^2 \bar{\phi}_k(X, T)^\top \Sigma_k^{-1} \bar{\phi}_k(X, T) \right] \\
&\lesssim \frac{1}{n^3} \lambda_{\max}(\Sigma_k^{-1}) \mathbb{E} \left[ \bar{\phi}_k(X, T)^\top \bar{\phi}_k(X, T) \right] \lesssim \frac{k}{n^3}.
\end{aligned}$$

Combining the above analyses, we complete the proof.  $\square$

Our next result characterizes the asymptotic distribution of  $\tilde{\psi}_k^{(2)}(\beta^*)$  (defined in Section S7). Before stating the lemma, recall the definitions in (S7.12) and (S7.13):

$$\begin{aligned} V_1(\beta^*) &:= \lim_{n \rightarrow \infty} \frac{n}{n+k} \mathbb{E} \left[ \left\{ 2\pi_1 \hat{\Xi}_{\beta^*} + 3\pi_1 \text{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right\}^{\otimes 2} \right] \\ V_2(\beta^*) &:= \lim_{n \rightarrow \infty} \frac{1}{k+n} \mathbb{E} \left[ \left\{ \pi_2 \hat{\Xi}_{\beta^*} + 3\pi_2 \text{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right\}^{\otimes 2} \right]. \end{aligned}$$

**Lemma 16.** *Conditioning on the nuisance sample, suppose that Assumptions 6–8 hold.*

*If  $k \ll n$ ,*

$$\sqrt{n} \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} = \sqrt{n} \cdot \mathbb{U}_{n,1} \left( 2\pi_1 \hat{\Xi}_{\beta^*} + 3\pi_1 \text{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right) + o_{\mathbb{P}}(1) \xrightarrow{d} N(0, V_1(\beta^*)).$$

*If  $k \gg n$ ,*

$$\sqrt{\frac{n^2}{k}} \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} = \sqrt{\frac{n^2}{k}} \cdot \mathbb{U}_{n,2} \left( \pi_2 \hat{\Xi}_{\beta^*} + 3\pi_2 \text{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right) + o_{\mathbb{P}}(1) \xrightarrow{d} N(0, V_2(\beta^*)/2).$$

*If  $k/n \rightarrow \tau$  for some  $\tau \in (0, \infty)$ ,*

$$\begin{aligned} \sqrt{n} \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} &= \sqrt{n} \cdot \mathbb{U}_{n,1} \left( 2\pi_1 \hat{\Xi}_{\beta^*} + 3\pi_1 \text{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right) \\ &\quad + \sqrt{n} \cdot \mathbb{U}_{n,2} \left( \pi_2 \hat{\Xi}_{\beta^*} + 3\pi_2 \text{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right) + o_{\mathbb{P}}(1) \\ &\xrightarrow{d} N \left( 0, (1+\tau) \left\{ V_1(\beta^*) + \frac{1}{2} V_2(\beta^*) \right\} \right). \end{aligned}$$

Here,  $\Pi^\perp[f \mid \bar{\phi}_k]$  denotes the orthocomplement of the projection of  $f$  onto the linear span of  $\bar{\phi}_k$ , as defined in Section 1.3.  $\tilde{\psi}_k^{(2)}(\beta^*)$  and  $\bar{\psi}_k(\beta^*)$  are defined in Section S7.

**Remark 15.** Since this lemma is established conditional on the nuisance sample, and both  $\hat{\Xi}_{\beta^*}$  and  $\text{IF}_{B_{\psi,k}}^{(2)}$  depend on the initial estimator  $\beta_{\text{init}}$ , the asymptotic distributions of  $\tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*)$  established in this lemma also implicitly depend on  $\beta_{\text{init}}$ .  $\blacksquare$

**Remark 16.** In Assumptions 3 and 7, we have assumed that the observations are bounded. For the CLT established in Lemma 16, the boundedness assumption can be relaxed if one uses Corollary 1.4 in Bhattacharya and Ghosh (1992) instead of their Corollary 1.5. The former only requires certain moment conditions to be satisfied, avoiding making the stronger boundedness assumption. We decide to stick to the boundedness assumption, following the related literature on HOIFs (Liu et al., 2017; Robins et al., 2023).  $\blacksquare$

*Proof.* Take a unit-norm vector  $v \in \mathbb{R}^q$  ( $\|v\| = 1$ ) and define

$$g_n(O_1) := \frac{a^\top \left\{ 2\pi_1 \widehat{\Xi}_{\beta^*} + 3\pi_1 \mathbf{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right\}}{\sqrt{k+n}},$$

$$h_n(O_1, O_2) := \frac{a^\top \left\{ \pi_2 \widehat{\Xi}_{\beta^*} + 3\pi_2 \mathbf{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \right\}}{n\sqrt{k+n}}.$$

Then,  $\mathbb{E}[g_n(O_1)] = 0$ , and  $h_n(O_1, O_2)$  is degenerate in view of the definition of  $\pi_2$  given in (S6.2) in Section S6.1, and

$$\lim_{n \rightarrow \infty} n^2 \mathbb{E}[h_n(O_1, O_2) h_n(O_1, O_2)^\top] = a^\top V_2(\beta^*) a,$$

$$\lim_{n \rightarrow \infty} n \mathbb{E}[g_n(O_1) g_n(O_1)^\top] = a^\top V_1(\beta^*) a.$$

By Corollary 1.5 in Bhattacharya and Ghosh (1992), if the following conditions hold

$$\sup_o |g_n(o)| \rightarrow 0, \tag{S7.24a}$$

$$\sup_{o_1, o_2} |h_n(o_1, o_2)| \rightarrow 0, \tag{S7.24b}$$

$$n \sup_{o_1} \mathbb{E}[h_n(o_1, O_2)] \rightarrow 0, \tag{S7.24c}$$

then we obtain the following central limit theorem jointly for  $\sum_j g_n(O_j)$  and  $\sum_{1 \leq i_1 \neq i_2 \leq n} h_n(O_{i_1}, O_{i_2})$ :

$$\begin{pmatrix} \sum_j g_n(O_j) \\ \sum_{1 \leq i_1 \neq i_2 \leq n} h_n(O_{i_1}, O_{i_2}) \end{pmatrix} \xrightarrow{d} N \left( \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} a^\top V_1(\beta^*) a & 0 \\ 0 & \frac{1}{2} a^\top V_2(\beta^*) a \end{bmatrix} \right).$$

By the Hoeffding decomposition (recall from Section S6.1), we have

$$\begin{aligned} \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) &= \mathbb{U}_{n,3} \pi_3 \mathbf{IF}_{B_{\psi,k}}^{(2)}(\beta^*) + 3\mathbb{U}_{n,2} \pi_2 \mathbf{IF}_{B_{\psi,k}}^{(2)}(\beta^*) \\ &\quad + 3\mathbb{U}_{n,1} \pi_1 \mathbf{IF}_{B_{\psi,k}}^{(2)}(\beta^*) + \mathbb{U}_{n,2} \pi_2 \widehat{\Xi}_{\beta^*} + 2\mathbb{U}_{n,1} \pi_1 \widehat{\Xi}_{\beta^*}. \end{aligned}$$

Then, in view of  $n/\sqrt{k+n} \cdot \mathbb{U}_{n,3} \pi_3 \mathbf{IF}_{B_{\psi,k}}^{(2)}(\beta^*) = o_{\mathbb{P}}(1)$ , we have

$$\frac{n}{\sqrt{k+n}} \cdot a^\top \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} = \sum_{1 \leq i_1 \neq i_2 \leq n} h_n(O_{i_1}, O_{i_2}) + \sum_j g_n(O_j) + o_{\mathbb{P}}(1).$$

If  $k \ll n$ , then  $V_2 = 0$  and  $n/\sqrt{k+n} = \sqrt{n} + o(1)$ , which implies that

$$\sqrt{n} \cdot a^\top \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} = \sum_j g_n(O_j) + o_{\mathbb{P}}(1) \xrightarrow{d} N \left( 0, a^\top V_1(\beta^*) a \right).$$

If  $k \gg n$ , then  $V_1 = 0$  and  $n/\sqrt{k+n} = n/\sqrt{k} + o(1)$ , which implies that

$$\frac{n}{\sqrt{k}} \cdot a^\top \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} = \sum_{1 \leq i_1 \neq i_2 \leq n} h_n(O_{i_1}, O_{i_2}) + o_{\mathbb{P}}(1) \xrightarrow{d} N\left(0, a^\top V_2(\beta^*) a/2\right).$$

If  $k = \tau n$  for some constant  $\tau$ , then  $n/\sqrt{k+n} = \sqrt{n}/\sqrt{1+\tau}$  and

$$\begin{aligned} \sqrt{n} \cdot a^\top \left\{ \tilde{\psi}_k^{(2)}(\beta^*) - \bar{\psi}_k(\beta^*) \right\} &= \sqrt{1+\tau} \left\{ \sum_j g_n(O_j) + \sum_{1 \leq i_1 \neq i_2 \leq n} h_n(O_{i_1}, O_{i_2}) \right\} + o_{\mathbb{P}}(1) \\ &\xrightarrow{d} N\left(0, (1+\tau)a^\top \left\{ V_1(\beta^*) + \frac{1}{2}V_2(\beta^*) \right\} a\right). \end{aligned}$$

It remains to verify whether Conditions (S7.24a)–(S7.24c) hold.

**Condition (S7.24a):** Using (S7.9), Condition (S7.24a) holds by Assumption 6 and Assumption 7, and the result in (S7.1).

**Condition (S7.24b):** By Assumption 6 and Assumption 7, we have

$$\sup_{o_1, o_2} \left| a^\top \pi_2 \hat{\Xi}_{\beta^*}(o_1, o_2) \right| \leq \sup_{o_1, o_2} \left| a^\top \hat{\Xi}_{\beta^*}(o_1, o_2) \right| \lesssim 1$$

and

$$\begin{aligned} &\sup_{o_1, o_2} \left| a^\top \pi_2 \text{IF}_{\text{B}_{\psi, k}}^{(2)}(o_1, o_2; \beta^*) \right| \\ &\lesssim \sup_{o_1, o_2, o_3} \left| a^\top \text{IF}_{\text{B}_{\psi, k}}^{(2)}(o_1, o_2, o_3; \beta^*) \right| \\ &\lesssim \left\| (\hat{b}_{\beta_{\text{init}}} - \Gamma_{\beta^*})(\cdot, \cdot) \right\|_\infty \cdot \sup_{o_1, o_2, o_3} \left| \left\{ \hat{\xi}(x_1, t_1) \bar{\phi}_k(x_1, t_1) - \bar{\mathbf{z}}_{k_x}(x_1) \otimes \bar{\mathbf{w}}_{k_t}(t_3) \right\}^\top \Sigma_k^{-1} \bar{\phi}_k(x_2, t_2) \right| \\ &\lesssim \zeta(k)^2, \end{aligned}$$

where the last inequality holds by Assumption 8. Thus we have

$$\sup_{o_1, o_2} |h_n(o_1, o_2)| \lesssim \frac{\zeta(k)^2}{n\sqrt{k+n}} \rightarrow 0.$$

**Condition (S7.24c):** Denote  $\bar{\phi}_k := \{\phi_1, \dots, \phi_k\}$ . By Assumption 8, there are at most  $c_8$  ( $c_8$  is a positive integer that does not depend on  $n$ ) nonzero terms in the sum  $\sum_{i=1}^k |\phi_i(x, t) \phi_i(x_2, t_2)| \cdot \left| a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(x_2, t_2) - \Gamma(y_2, t_2, \beta^*) \right\} \right|$  for any  $(o, o_2) \in \mathcal{O}^2$  and all nonzero terms are bounded by

$$\zeta(k)^2 \left\| a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(\cdot, \cdot) - \Gamma(\cdot, \cdot, \beta^*) \right\} \right\|_\infty.$$

In addition, for any  $i$ , the probability of the set of  $o_2 \in \mathcal{O}$  such that  $|\phi_i(x, t) \phi_i(x_2, t_2)| \cdot |a^\top \{\hat{b}_{\beta_{\text{init}}}(x_2, t_2) -$

$|\Gamma(y_2, t_2, \beta^*)| \neq 0$  is bounded by  $1/k$ . Then, for any  $(x, t) \in \mathcal{X} \times \mathcal{T}$ ,

$$\begin{aligned}
& \mathbb{E} \left[ \left| \bar{\phi}_k(x, t)^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(X_2, T_2) - \Gamma(Y_2, T_2, \beta^*) \right\} \right| \right] \\
& \leq \lambda_{\max}(\Sigma_k^{-1}) \mathbb{E} \left[ \left| \bar{\phi}_k(x, t)^\top \bar{\phi}_k(X_2, T_2) \right| \cdot \left| a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(X_2, T_2) - \Gamma(Y_2, T_2, \beta^*) \right\} \right| \right] \\
& \leq \lambda_{\max}(\Sigma_k^{-1}) \mathbb{E} \left[ \sum_{i=1}^k |\phi_i(x, t) \phi_i(X_2, T_2)| \cdot \left| a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(X_2, T_2) - \Gamma(Y_2, T_2, \beta^*) \right\} \right| \right] \\
& \leq \frac{1}{k} \lambda_{\max}(\Sigma_k^{-1}) c_8 \cdot \sup_{(o, o_2) \in \mathcal{O}^2} |\phi_i(x, t) \phi_i(x_2, t_2)| \cdot \left| a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(x_2, t_2) - \Gamma(y_2, t_2, \beta^*) \right\} \right| \cdot \|\mathbf{p}_{X, T}(\cdot, \cdot)\|_\infty \\
& \leq \frac{1}{k} \lambda_{\max}(\Sigma_k^{-1}) \zeta(k)^2 \left\| a^\top (\hat{b}_{\beta_{\text{init}}} - \Gamma_{\beta^*})(\cdot, \cdot) \right\|_\infty = O(1),
\end{aligned}$$

where  $\lambda_{\max}(\Sigma_k^{-1})$  denotes the largest eigenvalue of  $\Sigma_k^{-1}$  and the last equality holds by Assumptions 6–8. As a result of Assumption 6 and Jensen's inequality,

$$\begin{aligned}
& \sup_o \mathbb{E} \left[ \left\| \left\{ \hat{\xi}(x, t) \bar{\phi}_k(x, t) - \bar{\mathbf{z}}_{k_x}(x) \otimes \int_{\mathcal{T}} \bar{\mathbf{w}}_{k_t}(u) \mathbf{p}_T(u) du \right\}^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(X_2, T_2) - \Gamma(Y_2, T_2, \beta^*) \right\} \right\| \right] \\
& \lesssim \sup_o \mathbb{E} \left[ \left\| (\bar{\mathbf{z}}_{k_x}(x) \otimes \bar{\mathbf{w}}_{k_t}(t))^\top \Sigma_k^{-1} \bar{\phi}_k(X_2, T_2) a^\top \left\{ \hat{b}_{\beta_{\text{init}}}(X_2, T_2) - \Gamma(Y_2, T_2, \beta^*) \right\} \right\| \right] = O(1).
\end{aligned}$$

Similarly, analogous results can be obtained for the remaining terms in  $\pi_2 \mathbf{IF}_{\mathbf{B}_{\psi, k}}^{(2)}(o_1, O_2; \beta^*)$ . Taking these together, we have:

$$n \sup_{o_1} \mathbb{E} \left[ \left\| \frac{a^\top \left\{ \pi_2 \mathbf{IF}_{\mathbf{B}_{\psi, k}}^{(2)}(o_1, O_2; \beta^*) \right\}}{n\sqrt{k} + n} \right\| \right] \leq \frac{n}{n\sqrt{k} + n} \cdot O(1) = o(1).$$

Finally, since  $\sup_{o_1, o_2} \left| a^\top \pi_2 \hat{\Xi}_{\beta^*}(o_1, o_2) \right| \lesssim 1$ , we obtain the following:

$$n \sup_{o_1} \mathbb{E} \|h_n(o_1, O_2)\| \leq \frac{n}{n\sqrt{k} + n} \cdot O(1) = o(1).$$

□

## S8 Proof of Results in Section 4

To aid the exposition, we define the following quantities:

$$1/\lambda(x) := \frac{\mathbf{p}_X(x)}{g(x)} = \int_{u \in \mathcal{T}} K_{t, h}(u) \mathbf{p}_{T|X}(u|x) du, \quad (\text{S8.1})$$

$$\Delta_1(x) := \int_{u \in \mathcal{T}} \{\hat{\xi}(x, u) - \xi(x, u) - \hat{\xi}(x, t) + \xi(x, t)\} K_{t, h}(u) \mathbf{p}_{T|X}(u|x) du, \quad (\text{S8.2})$$

$$\Delta_2(x) := \int_{u \in \mathcal{T}} \{b_\beta(x, u) - \hat{b}_{\beta_{\text{init}}}(x, u) - b_\beta(x, t) + \hat{b}_{\beta_{\text{init}}}(x, t)\} K_{t, h}(u) \mathbf{p}_{T|X}(u|x) du. \quad (\text{S8.3})$$

Using Assumption 1(ii) and Assumption 15, and a similar argument to that used in the proof of Theorem 1 in Kennedy et al. (2017), we have

$$0 < \underline{c}_\lambda \leq \lambda(x) \leq \bar{c}_\lambda < \infty \text{ for any } x, \text{ and } \sup_{x \in \mathcal{X}} |\Delta_j(x)| \lesssim h^{\alpha_1 \wedge \alpha_2}, \text{ for } j = 1, 2, \quad (\text{S8.4})$$

where  $\underline{c}_\lambda, \bar{c}_\lambda$  are two fixed constants. By the definitions of  $\lambda(x)$  and  $g(x)$  (defined in Section 4.1), we have, for any function  $h(x)$ ,

$$\mathbb{E}[h(X)] = \int_{x \in \mathcal{X}} h(x) \mathbf{p}_X(x) dx = \int_{x \in \mathcal{X}} h(x) \lambda(x) g(x) dx, \quad (\text{S8.5})$$

$$\mathbb{E}[h(X) K_{t,h}(T)] = \int_{x \in \mathcal{X}} h(x) g(x) dx. \quad (\text{S8.6})$$

By definition of  $\lambda(x)$  and (S8.4), we have

$$a^\top \tilde{\Omega}_{k_x} a = a^\top \int_{x \in \mathcal{X}} g(x) \bar{\mathbf{z}}_{k_x}(x) \bar{\mathbf{z}}_{k_x}(x)^\top dx a = a^\top \mathbb{E} [\lambda(X) \bar{\mathbf{z}}_{k_x}(X)^{\otimes 2}] a \asymp a^\top \mathbb{E} [\bar{\mathbf{z}}_{k_x}(X)^{\otimes 2}] a.$$

Hence, Assumption 8(iii) implies that the eigenvalues of  $\tilde{\Omega}_{k_x}$  are bounded from above and away from zero uniformly in  $k_x$ . Consequently, analogues to (S7.1), using Assumptions 8, we have: for any function  $f \in \mathcal{C}(\mathcal{X})$ ,

$$\|\Pi_g[f \mid \bar{\mathbf{z}}_{k_x}](\cdot)\|_\infty \leq \|f(\cdot)\|_\infty, \quad (\text{S8.7})$$

where  $\Pi_g[f \mid \bar{\mathbf{z}}_{k_x}](\cdot) = (\int_{\mathcal{X}} f(x) \bar{\mathbf{z}}_{k_x}(x) g(x) dx)^\top \tilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(\cdot)$ .

Under Assumption 6(iii), for any  $t \in \mathcal{T}$  and  $\delta > 0$ , we have

$$\begin{aligned} & \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \leq \delta} |\Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2)|^2 \mid T = t \right] \\ & \leq \frac{1}{c_5} \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \leq \delta} |\Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2)|^2 \mid T = t \right] \mathbf{p}_T(t) \\ & \leq \frac{1}{c_5} \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \leq \delta} |\Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2)|^2 \right]. \end{aligned}$$

Hence, applying Assumption 7(iii), we obtain

$$\mathbb{E}^{1/2} \left[ \sup_{|\beta_1 - \beta_2| \leq \delta} |\Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2)|^2 \mid T = t \right] \lesssim \delta^{\alpha_0}. \quad (\text{S8.8})$$

### S8.1 Proof of Theorem 6

Because the proof carefully tracks dependence on  $\beta$ , we now make that dependence explicit by moving  $\beta$  out of subscripts into argument parentheses. Denote population counterpart of  $\tilde{\psi}_{t,k}^{(2)}(\beta)$  as

given by  $\bar{\psi}_{t,k}(\beta) := \mathbb{E}[\tilde{\psi}_{t,k}^{(2)}(\beta)]$ .

*Proof of Theorem 6.* Assumption 14, together with an argument similar to that leading to (S7.2) in Section S7.1, implies

$$|\psi_t(\beta)| \gtrsim |\beta - \beta_t^*| \wedge c \quad (\text{S8.9})$$

for some constant  $c > 0$ . Consequently, to bound  $|\tilde{\beta}_t^{(2)} - \beta_t^*|$ , it suffices to control the quantity  $|\psi_t(\tilde{\beta}_t^{(2)})|$ .

First, we establish the consistency of  $\tilde{\beta}_t^{(2)}$  by showing that  $|\psi_t(\tilde{\beta}_t^{(2)})| \lesssim \tilde{r}_{n,\text{np}} \rightarrow 0$  with probability approaching one, where  $\tilde{r}_{n,\text{np}}$  is defined in Theorem 6. Note that  $\psi_t(\beta_t^*) \equiv 0$  and  $\tilde{\psi}_{t,k}(\tilde{\beta}_t^{(2)}) \equiv 0$ . Consider the following decomposition:

$$\begin{aligned} \left| \psi_t(\tilde{\beta}_t^{(2)}) \right| &\lesssim \left| \psi_t(\tilde{\beta}_t^{(2)}) - \bar{\psi}_{t,k}(\tilde{\beta}_t^{(2)}) \right| + \left| \bar{\psi}_{t,k}(\beta_t^*) - \psi_t(\beta_t^*) \right| \\ &\quad + \left| \left\{ \bar{\psi}_{t,k}(\tilde{\beta}_t^{(2)}) - \tilde{\psi}_{t,k}^{(2)}(\tilde{\beta}_t^{(2)}) \right\} - \left\{ \bar{\psi}_{t,k}(\beta_t^*) - \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right\} \right| \\ &\quad + \left| \tilde{\psi}_{t,k}^{(2)}(\tilde{\beta}_t^{(2)}) - \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right| \\ &\lesssim \sup_{\beta \in \mathcal{B}} \left| \psi_t(\beta) - \bar{\psi}_{t,k}(\beta) \right| \end{aligned} \quad (\text{S8.10})$$

$$+ \sup_{\beta \in \mathcal{B}} \left| \left\{ \bar{\psi}_{t,k}(\beta) - \tilde{\psi}_{t,k}^{(2)}(\beta) \right\} - \left\{ \bar{\psi}_{t,k}(\beta_t^*) - \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right\} \right| \quad (\text{S8.11})$$

$$+ \left| \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right|. \quad (\text{S8.12})$$

Here, the first term (S8.10) contributes to the bias, the third term (S8.12) captures the variance, and the second term (S8.11) is a  $U$ -process indexed by  $\beta$ .

First, we analyze the term (S8.10). Applying Lemmas 17 and 18, we obtain the following bias decomposition:

$$\begin{aligned} &\psi_t(\beta) - \bar{\psi}_{t,k}(\beta) \\ &= \mathbb{E}[\xi(X, T)b_\beta(X, T)|T = t]\mathbf{p}_T(t) - \mathbb{E} \left[ K_{t,h}(T_1)\hat{\xi}(X_1, T_1)\{\Gamma(Y_1, T_1, \beta) - \hat{b}_{\beta_{\text{init}}}(X_1, T_1)\} \right] \\ &\quad + \mathbb{E} \left[ \hat{b}_{\beta_{\text{init}}}(X_1, T_2)K_{t,h}(T_2) \right] \\ &\quad - \mathbb{E} \left[ \{K_{t,h}(T_1)\hat{\xi}(X_1, T_1) - K_{t,h}(T_3)\}\bar{\mathbf{z}}_{k_x}^\top(X_1)\tilde{\Omega}_{k_x}^{-1}\bar{\mathbf{z}}_{k_x}(X_2)\{\Gamma(Y_2, T_2, \beta) - \hat{b}_{\beta_{\text{init}}}(X_2, T_2)\}K_{t,h}(T_2) \right] \\ &= \mathbb{E} \left[ K_{t,h}(T) \left\{ \xi(X, t) - \hat{\xi}(X, t) \right\} \left\{ b_\beta(X, t) - \hat{b}_{\beta_{\text{init}}}(X, t) \right\} \right] \\ &\quad - \int_{x \in \mathcal{X}} \Pi_g[\hat{\xi}(\cdot, t) - \xi(\cdot, t)|\bar{\mathbf{z}}_{k_x}](x) \times \Pi_g[b_\beta(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t)|\bar{\mathbf{z}}_{k_x}](x)g(x)dx + O(h^{\alpha_1 \wedge \alpha_2}) \\ &= \int_{x \in \mathcal{X}} \left\{ \xi(x, t) - \hat{\xi}(x, t) \right\} \left\{ b_\beta(x, t) - \hat{b}_{\beta_{\text{init}}}(x, t) \right\} g(x)dx \\ &\quad - \int_{x \in \mathcal{X}} \Pi_g[\hat{\xi}(\cdot, t) - \xi(\cdot, t)|\bar{\mathbf{z}}_{k_x}](x) \times \Pi_g[b_\beta(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t)|\bar{\mathbf{z}}_{k_x}](x)g(x)dx + O(h^{\alpha_1 \wedge \alpha_2}) \end{aligned}$$

$$= \int_{x \in \mathcal{X}} \Pi_g^\perp [b_\beta(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}](x) \times \Pi_g^\perp [\xi(\cdot, t) - \hat{\xi}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}](x) g(x) dx + O(h^{\alpha_1 \wedge \alpha_2}), \quad (\text{S8.13})$$

where  $\Pi_g[f \mid \bar{\mathbf{z}}_{k_x}](\cdot) := \int_{\mathcal{X}} f(x) \bar{\mathbf{z}}_{k_x}(x)^\top g(x) dx \tilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(\cdot)$  denotes the projection (with respect to the weight  $g$ ) of  $f$  onto the linear span of  $\bar{\mathbf{z}}_{k_x}$  and  $\Pi_g^\perp[f \mid \bar{\mathbf{z}}_{k_x}](\cdot) := f(\cdot) - \Pi_g[f \mid \bar{\mathbf{z}}_{k_x}](\cdot)$  denotes the orthocomplement of this projection. This shows that the leading bias is determined by the product of the projection errors of  $b_\beta(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t)$  and  $\xi(\cdot, t) - \hat{\xi}(\cdot, t)$  onto the span of  $\bar{\mathbf{z}}_{k_x}$ , plus a kernel approximation error of order  $O(h^{\alpha_1 \wedge \alpha_2})$ . Thus, we conclude that

$$(\text{S8.10}) \lesssim \sup_{\beta \in \mathcal{B}} \left\| \Pi_g^\perp [b_\beta(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} \cdot \left\| \Pi_g^\perp [\xi(\cdot, t) - \hat{\xi}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} + O(h^{\alpha_1 \wedge \alpha_2}). \quad (\text{S8.14})$$

Here,  $\|\cdot\|_{g,2}$  denotes the norm  $\|f\|_{g,2} = \{\int_{\mathcal{X}} f(x)^2 g(x) dx\}^{1/2}$  for any function  $f(x)$ .

Next, we analyze the  $U$ -process term (S8.11). Since  $\mathcal{B}$  is compact, we may write  $\sup_{\beta \in \mathcal{B}}$  as  $\sup_{|\beta - \beta_t^*| \leq c}$  for some constant  $c$  sufficiently large to cover  $\mathcal{B}$ . Applying Lemma 23 with  $\epsilon = c$ , we obtain

$$(\text{S8.11}) = O_{\mathbb{P}} \left( \frac{\sqrt{k}}{nh} + \frac{1}{\sqrt{nh}} \right) \cdot \log n. \quad (\text{S8.15})$$

Last, for the term (S8.12), applying the decomposition in (S8.13) together with Lemma 22, we have

$$\begin{aligned} |\tilde{\psi}_{t,k}^{(2)}(\beta_t^*)| &= O \left( \left\| \Pi_g^\perp [b_{\beta_t^*}(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} \cdot \left\| \Pi_g^\perp [\xi(\cdot, t) - \hat{\xi}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} \right) \\ &\quad + O(h^{\alpha_1 \wedge \alpha_2}) + O_{\mathbb{P}} \left( \frac{\sqrt{k_x}}{nh} + \frac{1}{\sqrt{nh}} \right). \end{aligned}$$

Combining the bounds for terms (S8.10)–(S8.12), we have, with probability approaching one,

$$\left| \psi_t(\tilde{\beta}_t^{(2)}) \right| \lesssim (\text{S8.10}) + (\text{S8.11}) + (\text{S8.12}) \lesssim \tilde{r}_{n,\text{np}}.$$

This establishes the consistency of  $\tilde{\beta}_t^{(2)}$ .

Building on this consistency result, we now restrict our analysis to the region  $|\beta - \beta_t^*| \lesssim \tilde{r}_{n,\text{np}}$ . This refinement allows us to improve upon the initial rate  $\tilde{r}_{n,\text{np}}$  and obtain the final convergence rate stated in Theorem 6. Repeating the decomposition from the consistency argument, but now over the restricted region, we have

$$\begin{aligned} \left| \psi_t(\tilde{\beta}_t^{(2)}) \right| &\lesssim \sup_{|\beta - \beta_t^*| \lesssim \tilde{r}_{n,\text{np}}} \left| \psi_t(\beta) - \bar{\psi}_{t,k}(\beta) \right| + \left| \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right| \\ &\quad + \sup_{|\beta - \beta_t^*| \lesssim \tilde{r}_{n,\text{np}}} \left| \left\{ \bar{\psi}_{t,k}(\beta) - \tilde{\psi}_{t,k}^{(2)}(\beta) \right\} - \left\{ \bar{\psi}_{t,k}(\beta_t^*) - \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right\} \right|. \end{aligned}$$

The only change from the earlier argument is that the supremum is now taken over a shrinking neighborhood of  $\beta_t^*$ . Applying the bias decomposition (S8.13), Lemma 23 with  $\epsilon = \tilde{r}_{n,\text{np}}$ , and Lemma 22, we obtain, with probability approaching one,

$$\begin{aligned} \left| \psi_t(\tilde{\beta}_t^{(2)}) \right| &\lesssim \sup_{|\beta - \beta_t^*| \lesssim \tilde{r}_{n,\text{np}}} \left\| \Pi_g^\perp [b_\beta(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} \cdot \left\| \Pi_g^\perp [\xi(\cdot, t) - \hat{\xi}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} \\ &\quad + h^{\alpha_1 \wedge \alpha_2} + \frac{\sqrt{k_x}}{nh} + \frac{1}{\sqrt{nh}} + \left( \frac{\sqrt{k_x}}{nh} + \frac{1}{\sqrt{nh}} \right) \cdot \{(\tilde{r}_{n,\text{np}})^{\alpha_0} \vee \sqrt{h}\} \cdot \log n. \end{aligned}$$

Given the additional conditions  $h(\log n)^2 \rightarrow 0$  and  $(\tilde{r}_{n,\text{np}})^{\alpha_0} \log n \rightarrow 0$ , this further simplifies to

$$\begin{aligned} \left| \psi_t(\tilde{\beta}_t^{(2)}) \right| &\lesssim \frac{\sqrt{k_x}}{nh} + \frac{1}{\sqrt{nh}} + h^{\alpha_1 \wedge \alpha_2} \\ &\quad + \sup_{|\beta - \beta_t^*| \lesssim \tilde{r}_{n,\text{np}}} \left\| \Pi_g^\perp [b_\beta(\cdot, t) - \hat{b}_{\beta_{\text{init}}}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} \cdot \left\| \Pi_g^\perp [\xi(\cdot, t) - \hat{\xi}(\cdot, t) \mid \bar{\mathbf{z}}_{k_x}] \right\|_{g,2} \end{aligned}$$

with probability approaching one. Combining this with (S8.9) yields the final convergence rate for  $\tilde{\beta}_t^{(2)}$ , thus completing the proof.  $\square$

## S8.2 Technical Lemmas for Proving Theorem 6

**Lemma 17.** *Under Assumptions 13 and 15, we have*

$$\begin{aligned} \mathbb{E}[\xi(X, T)b_\beta(X, T)|T = t]\mathbf{p}_T(t) &= \mathbb{E}[\xi(X, t)b_\beta(X, t)K_{t,h}(T)] + O(h^{\alpha_1 \wedge \alpha_2}), \\ \mathbb{E} \left[ K_{t,h}(T_1)\hat{\xi}(X_1, T_1)\{\Gamma(Y_1, T_1, \beta) - \hat{b}_{\beta_{\text{init}}}(X_1, T_1)\} + \hat{b}_{\beta_{\text{init}}}(X_1, T_2)K_{t,h}(T_2) \right] \\ &= \mathbb{E} \left[ K_{t,h}(T)\hat{\xi}(X, t)\{b_\beta(X, t) - \hat{b}_{\beta_{\text{init}}}(X, t)\} + \xi(X, t)\hat{b}_{\beta_{\text{init}}}(X, t)K_{t,h}(T) \right] + O(h^{\alpha_1 \wedge \alpha_2}). \end{aligned}$$

*Proof.* Using a similar argument for proving (S8.4), we have

$$\sup_{x \in \mathcal{X}} \left| \int_{u \in \mathcal{T}} K_{t,h}(u) \mathbf{p}_{T|X}(u|x) du - \mathbf{p}_{T|X}(t|x) \right| = O(h^{\alpha_1 \wedge \alpha_2}).$$

Thus

$$\begin{aligned} \mathbb{E}[\xi(X, t)b_\beta(X, t)K_{t,h}(T)] &= \int_{x \in \mathcal{X}} \xi(x, t)b_\beta(x, t) \int_{u \in \mathcal{T}} K_{t,h}(u) \mathbf{p}_{T|X}(u|x) du \mathbf{p}_X(x) dx \\ &= \int_{x \in \mathcal{X}} \xi(x, t)b_\beta(x, t) \mathbf{p}_{T|X}(t|x) \mathbf{p}_X(x) dx + O(h^{\alpha_1 \wedge \alpha_2}) \\ &= \int_{x \in \mathcal{X}} b_\beta(x, t) \mathbf{p}_T(t) \mathbf{p}_X(x) dx + O(h^{\alpha_1 \wedge \alpha_2}) \\ &= \mathbb{E}[\xi(X, T)b_\beta(X, T)|T = t]\mathbf{p}_T(t) + O(h^{\alpha_1 \wedge \alpha_2}). \end{aligned}$$

Hence, we obtain the first result.

Using the definition of  $\xi(x, t)$ , we have

$$\begin{aligned}
& \mathbb{E} \left[ K_{t,h}(T_1) \widehat{\xi}(X_1, T_1) \{ \Gamma(Y_1, T_1, \beta) - \widehat{b}_{\beta_{\text{init}}}(X_1, T_1) \} + \widehat{b}_{\beta_{\text{init}}}(X_1, T_2) K_{t,h}(T_2) \right] \\
&= \int_{x \in \mathcal{X}} \int_{u \in \mathcal{T}} K_{t,h}(u) \left\{ \widehat{\xi}(x, u) (b_\beta(x, u) - \widehat{b}_{\beta_{\text{init}}}(x, u)) + \xi(x, u) \widehat{b}_{\beta_{\text{init}}}(x, u) \right\} \mathbf{p}_{T|X}(u|x) du \mathbf{p}_X(x) dx \\
&= \int_{x \in \mathcal{X}} \int_{u \in \mathcal{T}} K_{t,h}(u) \left\{ \widehat{\xi}(x, t) (b_\beta(x, t) - \widehat{b}_{\beta_{\text{init}}}(x, t)) + \xi(x, t) \widehat{b}_{\beta_{\text{init}}}(x, t) \right\} \mathbf{p}_{T|X}(u|x) du \mathbf{p}_X(x) dx + O(h^{\alpha_1 \wedge \alpha_2}) \\
&= \mathbb{E} \left[ K_{t,h}(T) \left\{ \widehat{\xi}(X, t) (b_\beta(X, t) - \widehat{b}_{\beta_{\text{init}}}(X, t)) + \xi(X, t) \widehat{b}_{\beta_{\text{init}}}(X, t) \right\} \right] + O(h^{\alpha_1 \wedge \alpha_2}),
\end{aligned}$$

where the second equality uses a similar argument for proving (S8.4).  $\square$

**Lemma 18.** *Under Assumptions 6–8, 13 and 15, we have*

$$\begin{aligned}
& \left| \mathbb{E} \left[ \{ \widehat{\xi}(X, T) - \xi(X, T) \} \bar{\mathbf{z}}_{k_x}(X)^\top K_{t,h}(T) \right] \tilde{\Omega}_{k_x}^{-1} \mathbb{E} \left[ \bar{\mathbf{z}}_{k_x}(X) \{ \Gamma(Y, T, \beta) - \widehat{b}_{\beta_{\text{init}}}(X, T) \} K_{t,h}(T) \right] \right. \\
& \quad \left. - \int_{x \in \mathcal{X}} \Pi_g[\widehat{\xi}(\cdot, t) - \xi(\cdot, t) | \bar{\mathbf{z}}_{k_x}](x) \times \Pi_g[b_\beta(\cdot, t) - \widehat{b}_{\beta_{\text{init}}}(\cdot, t) | \bar{\mathbf{z}}_{k_x}](x) g(x) dx \right| = O(h^{\alpha_1 \wedge \alpha_2}).
\end{aligned}$$

*Proof.* By the definition of  $\xi(X, T)$ ,

$$\begin{aligned}
& \mathbb{E} \left[ \{ K_{t,h}(T_1) \widehat{\xi}(X_1, T_1) - K_{t,h}(T_3) \} \bar{\mathbf{z}}_{k_x}^\top(X_1) \tilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(X_2) \{ \Gamma(Y_2, T_2, \beta) - \widehat{b}_{\beta_{\text{init}}}(X_2, T_2) \} K_{t,h}(T_2) \right] \\
&= \mathbb{E} \left[ \{ \widehat{\xi}(X, T) - \xi(X, T) \} \bar{\mathbf{z}}_{k_x}(X)^\top K_{t,h}(T) \right] \tilde{\Omega}_{k_x}^{-1} \mathbb{E} \left[ \bar{\mathbf{z}}_{k_x}(X) \{ b_\beta(X, T) - \widehat{b}_{\beta_{\text{init}}}(X, T) \} K_{t,h}(T) \right].
\end{aligned}$$

Recalling the definitions of  $\Delta_1(X)$  and  $\Delta_2(X)$  in (S8.2) and (S8.3), we have

$$\mathbb{E} \left[ \{ b_\beta(X, T) - \widehat{b}_{\beta_{\text{init}}}(X, T) - b_\beta(X, t) + \widehat{b}_{\beta_{\text{init}}}(X, t) \} K_{t,h}(T) \bar{\mathbf{z}}_{k_x}(X) \right] = \mathbb{E}[\Delta_2(X) \bar{\mathbf{z}}_{k_x}(X)^\top]$$

and

$$\mathbb{E}[\{ \widehat{\xi}(X, T) - \xi(X, T) - \widehat{\xi}(X, t) + \xi(X, t) \} \bar{\mathbf{z}}_{k_x}(X)^\top K_{t,h}(T)] = \mathbb{E}[\Delta_1(X) \bar{\mathbf{z}}_{k_x}(X)^\top].$$

Then, we obtain

$$\begin{aligned}
& \mathbb{E}[\{ \widehat{\xi}(X, T) - \xi(X, T) \} \bar{\mathbf{z}}_{k_x}(X)^\top K_{t,h}(T)] \tilde{\Omega}_{k_x}^{-1} \mathbb{E} \left[ \bar{\mathbf{z}}_{k_x}(X) \{ \Gamma(Y, T, \beta) - \widehat{b}_{\beta_{\text{init}}}(X, T) \} K_{t,h}(T) \right] \\
&= \left( \mathbb{E}[\{ \widehat{\xi}(X, t) - \xi(X, t) \} \bar{\mathbf{z}}_{k_x}(X)^\top K_{t,h}(T)] + \mathbb{E}[\Delta_1(X) \bar{\mathbf{z}}_{k_x}(X)^\top] \right) \times \tilde{\Omega}_{k_x}^{-1} \\
& \quad \times \left( \mathbb{E}[\{ b_\beta(X, t) - \widehat{b}_{\beta_{\text{init}}}(X, t) \} \bar{\mathbf{z}}_{k_x}(X) K_{t,h}(T)] + \mathbb{E}[\Delta_2(X) \bar{\mathbf{z}}_{k_x}(X)] \right) \\
&= \left[ \int_{x \in \mathcal{X}} \{ \widehat{\xi}(x, t) - \xi(x, t) \} \bar{\mathbf{z}}_{k_x}(x)^\top g(x) dx + \int_{x \in \mathcal{X}} \Delta_1(x) \lambda(x) \bar{\mathbf{z}}_{k_x}(x)^\top g(x) dx \right] \\
& \quad \times \tilde{\Omega}_{k_x}^{-1} \times \left[ \int_{x \in \mathcal{X}} \{ b_\beta(x, t) - \widehat{b}_{\beta_{\text{init}}}(x, t) \} \bar{\mathbf{z}}_{k_x}(x) g(x) dx + \int_{x \in \mathcal{X}} \Delta_2(x) \lambda(x) \bar{\mathbf{z}}_{k_x}(x)^\top g(x) dx \right]
\end{aligned}$$

$$\begin{aligned}
&= \int_{x \in \mathcal{X}} \left\{ \Pi_g[\widehat{\xi}(\cdot, t) - \xi(\cdot, t)|\bar{z}_{k_x}](x) + \Pi_g[\Delta_1 \lambda|\bar{z}_{k_x}](x) \right\} \\
&\quad \times \left\{ \Pi_g[b_\beta(\cdot, t) - \widehat{b}_{\beta_{\text{init}}}(\cdot, t)|\bar{z}_{k_x}](x) + \Pi_g[\Delta_2 \lambda|\bar{z}_{k_x}](x) \right\} g(x) dx.
\end{aligned}$$

where the second equality holds by using (S8.5) and (S8.6). Hence

$$\begin{aligned}
&\left| \mathbb{E}[\{\widehat{\xi}(X, T) - \xi(X, T)\} \bar{z}_{k_x}(X)^\top K_{t,h}(T)] \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) \{\Gamma(Y, T, \beta) - \widehat{b}_{\beta_{\text{init}}}(X, T)\} K_{t,h}(T)] \right. \\
&\quad \left. - \int_{x \in \mathcal{X}} \Pi_g[\widehat{\xi}(\cdot, t) - \xi(\cdot, t)|\bar{z}_{k_x}](x) \times \Pi_g[b_\beta(\cdot, t) - \widehat{b}_{\beta_{\text{init}}}(\cdot, t)|\bar{z}_{k_x}](x) g(x) dx \right| \\
&\leq \left| \int_{x \in \mathcal{X}} \Pi_g[\widehat{\xi}(\cdot, t) - \xi(\cdot, t)|\bar{z}_{k_x}](x) \times \Pi_g[\Delta_2 \lambda|\bar{z}_{k_x}](x) g(x) dx \right| \\
&\quad + \left| \int_{x \in \mathcal{X}} \Pi_g[\Delta_1 \lambda|\bar{z}_{k_x}](x) \times \Pi_g[b_\beta(\cdot, t) - \widehat{b}_{\beta_{\text{init}}}(\cdot, t)|\bar{z}_{k_x}](x) g(x) dx \right| \\
&\quad + \left| \int_{x \in \mathcal{X}} \Pi_g[\Delta_1 \lambda|\bar{z}_{k_x}](x) \times \Pi_g[\Delta_2 \lambda|\bar{z}_{k_x}](x) g(x) dx \right| \\
&\lesssim \|\Pi_g[\widehat{\xi}(\cdot, t) - \xi(\cdot, t)|\bar{z}_{k_x}]\|_{g,2} \cdot \|\Pi_g[\Delta_2 \lambda|\bar{z}_{k_x}]\|_{g,2} \\
&\quad + \|\Pi_g[\Delta_1 \lambda|\bar{z}_{k_x}]\|_{g,2} \cdot \|\Pi_g[b_\beta(\cdot, t) - \widehat{b}_{\beta_{\text{init}}}(\cdot, t)|\bar{z}_{k_x}]\|_{g,2} \\
&\quad + \|\Pi_g[\Delta_1 \lambda|\bar{z}_{k_x}]\|_{g,2} \cdot \|\Pi_g[\Delta_2 \lambda|\bar{z}_{k_x}]\|_{g,2} \\
&\lesssim \|\widehat{\xi}(X, t) - \xi(X, t)\|_{g,2} \cdot \|\Delta_2(X) \lambda(X)\|_{g,2} \\
&\quad + \|\Delta_1(X) \lambda(X)\|_{g,2} \cdot \|b_\beta(X, t) - \widehat{b}_{\beta_{\text{init}}}(X, t)\|_{g,2} \\
&\quad + \|\Delta_1(X) \lambda(X)\|_{g,2} \cdot \|\Delta_2(X) \lambda(X)\|_{g,2} \\
&= O(h^{\alpha_1 \wedge \alpha_2}), \tag{S8.16}
\end{aligned}$$

where the last equality holds by using Assumptions 6, 15, and (S8.4).  $\square$

**Lemma 19.** *Under Assumptions 6, 7, 13, and 15, we have*

$$\begin{aligned}
&\mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \lesssim \epsilon} |\{\Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2)\} K_{t,h}(T)|^2 \right] = O \left( \frac{\epsilon^{2\alpha_0} \vee h}{h} \right), \\
&a^\top \mathbb{E} \left[ K_{t,h}(T)^2 \{\widehat{\xi}(X, T) - 1\}^2 \xi(X, T) \bar{z}_{k_x}(X) \bar{z}_{k_x}(X)^\top \right] a \lesssim \frac{1}{h} a^\top \tilde{\Omega}_{k_x} a,
\end{aligned}$$

where  $a \in \mathbb{R}^{k_x}$  is a unit vector.

*Proof.* By Taylor expansion, Assumptions 6 and (S8.8), we have

$$\mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \lesssim \epsilon} |\{\Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2)\} K_{t,h}(T)|^2 \right] \tag{S8.17}$$

$$\begin{aligned}
&= \frac{1}{h^2} \int_{u \in \mathcal{T}} \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \lesssim \epsilon} \left| \widehat{\xi}(X, T) \{ \Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2) \} \right|^2 \mid T = u \right] \left| K \left( \frac{u - t}{h} \right) \right|^2 \mathbf{p}_T(u) du \\
&= \frac{1}{h} \int_{u \in \mathcal{T}} \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \lesssim \epsilon} \left| \widehat{\xi}(X, T) \{ \Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2) \} \right|^2 \mid T = t + hu \right] |K(u)|^2 \mathbf{p}_T(t + hu) du \\
&= \frac{1}{h} \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \lesssim \epsilon} \left| \widehat{\xi}(X, T) \{ \Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2) \} \right|^2 \mid T = t \right] \mathbf{p}_T(T = t) \cdot \kappa_{0,2} \\
&\quad + \frac{1}{h} \int_{u \in \mathcal{T}} \nabla_t \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \lesssim \epsilon} \left| \widehat{\xi}(X, T) \{ \Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2) \} \right|^2 \mid T = t + \tau hu \right] hu K(u)^2 du \\
&\lesssim \frac{1}{h} \left\{ \mathbb{E} \left[ \sup_{|\beta_1 - \beta_2| \lesssim \epsilon} |\Gamma(Y, T, \beta_1) - \Gamma(Y, T, \beta_2)|^2 \mid T = t \right] \mathbf{p}_T(T = t) \cdot \kappa_{0,2} + O(h) \right\} = O \left( \frac{\epsilon^{2\alpha_0} \vee h}{h} \right),
\end{aligned}$$

where  $\tau \in (0, 1)$  and  $\kappa_{0,2} = \int K(u)^2 du$ . Thus, we obtain the first result. Similarly, we have  $\mathbb{E}[K_{t,h}(T)^2] = O(1/h)$ . Then, we obtain

$$\begin{aligned}
&a^\top \mathbb{E} \left[ K_{t,h}(T)^2 \{ \widehat{\xi}(X, T) - 1 \}^2 \xi(X, T) \bar{\mathbf{z}}_{k_x}(X) \bar{\mathbf{z}}_{k_x}(X)^\top \right] a \\
&\lesssim a^\top \mathbb{E} \left[ K_{t,h}(T)^2 \xi(X, T) \bar{\mathbf{z}}_{k_x}(X) \bar{\mathbf{z}}_{k_x}(X)^\top \right] a \\
&= \mathbb{E} \left[ K_{t,h}(T)^2 \right] \cdot a^\top \mathbb{E} \left[ \bar{\mathbf{z}}_{k_x}(X) \bar{\mathbf{z}}_{k_x}(X)^\top \right] a \\
&= \mathbb{E} \left[ K_{t,h}(T)^2 \right] \cdot a^\top \int_{x \in \mathcal{X}} \bar{\mathbf{z}}_{k_x}(x) \bar{\mathbf{z}}_{k_x}(x)^\top g(x) \cdot \frac{\mathbf{p}_X(x)}{g(x)} dx a \\
&\leq \sup_{x \in \mathcal{X}} \left| \frac{\mathbf{p}_X(x)}{g(x)} \right| \cdot \mathbb{E} \{ K_{t,h}(T)^2 \} \cdot a^\top \widetilde{\Omega}_{k_x} a \lesssim \frac{1}{h} a^\top \widetilde{\Omega}_{k_x} a,
\end{aligned}$$

where the first inequality holds since we have  $\|\widehat{\xi}\| \lesssim 1$ ; the first equality holds by the definition of  $\xi(x, t)$ .  $\square$

**Lemma 20.** *Under Assumptions 6–8, 13, and 15, we have the following results:*

$$\begin{aligned}
\text{(i)} \quad &\sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| \mathbb{E} \left[ \{ \widehat{\xi}(X, T) - \xi(X, T) \} K_{t,h}(T) \bar{\mathbf{z}}_{k_x}(X)^\top \right] \widetilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(X) K_{t,h}(T) (\Gamma_\beta - \Gamma_{\beta_t^*})(Y, T) \right|^2 \right] \\
&\lesssim \frac{\epsilon^{2\alpha_0} \vee h}{h}; \\
\text{(ii)} \quad &\sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| K_{t,h}(T) \mathbb{E} [\bar{\mathbf{z}}_{k_x}(X)^\top] \widetilde{\Omega}_{k_x}^{-1} \mathbb{E} [\bar{\mathbf{z}}_{k_x}(X) K_{t,h}(T) \{ b_\beta(X, T) - b_{\beta_t^*}(X, T) \}] \right|^2 \right] \\
&\lesssim \frac{\epsilon^{2\alpha_0} \vee h^{2(\alpha_1 \wedge \alpha_2)}}{h}; \\
\text{(iii)} \quad &\sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| K_{t,h}(T) \widehat{\xi}(X, T) \bar{\mathbf{z}}_{k_x}(X)^\top \widetilde{\Omega}_{k_x}^{-1} \mathbb{E} [\bar{\mathbf{z}}_{k_x}(X) K_{t,h}(T) \{ b_\beta(X, T) - b_{\beta_t^*}(X, T) \}] \right|^2 \right] \\
&\lesssim \frac{\epsilon^{2\alpha_0} \vee h^{2(\alpha_1 \wedge \alpha_2)}}{h};
\end{aligned}$$

$$\begin{aligned}
\text{(iv)} \quad & \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| \mathbb{E}[K_{t,h}(T)] \bar{z}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) K_{t,h}(T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}] \right|^2 \right] \\
& \lesssim \epsilon^{2\alpha_0} \vee h^{2(\alpha_1 \wedge \alpha_2)}.
\end{aligned}$$

*Proof.* To aid the exposition, we define the following quantities:

$$\Delta_3(x) := \int_{u \in \mathcal{T}} \{\hat{\xi}(x, u) - \xi(x, u)\} K_{t,h}(u) \mathbf{p}_{T|X}(u|x) du, \quad (\text{S8.18})$$

$$\Delta_4(x) := \int_{u \in \mathcal{T}} \{b_\beta(x, u) - b_{\beta_t^*}(x, u)\} K_{t,h}(u) \mathbf{p}_{T|X}(u|x) du - \{b_\beta(x, t) - b_{\beta_t^*}(x, t)\} \mathbf{p}_{T|X}(t|x) \quad (\text{S8.19})$$

$$\begin{aligned}
\Delta_5(x) &:= \Delta_4(x) + \{b_\beta(x, t) - b_{\beta_t^*}(x, t)\} \mathbf{p}_{T|X}(t|x) \\
&= \int_{u \in \mathcal{T}} \{b_\beta(x, u) - b_{\beta_t^*}(x, u)\} K_{t,h}(u) \mathbf{p}_{T|X}(u|x) du
\end{aligned} \quad (\text{S8.20})$$

Note that both  $\Delta_4(x)$  and  $\Delta_5(x)$  depend on  $\beta$ , but, for notational convenience, we suppress this argument. Assumption 15 states that the functions  $b_\beta$  share the same degree of smoothness for every  $\beta$ , we can repeat the argument used to establish (S8.4) and obtain the following results:

$$\sup_{x \in \mathcal{X}} |\Delta_3(x)| \lesssim 1, \quad \sup_{\beta} \sup_{x \in \mathcal{X}} |\Delta_4(x)| \lesssim h^{\alpha_1 \wedge \alpha_2}, \quad \text{and} \quad \sup_{\beta} \sup_{x \in \mathcal{X}} |\Delta_5(x)| \lesssim 1. \quad (\text{S8.21})$$

First, we have

$$\begin{aligned}
& \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| \Pi_g[\Delta_5(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](X) \right|^2 \right] \\
&= \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \int_{x \in \mathcal{X}} \left| \Pi_g[\Delta_5(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](x) \right|^2 \lambda(x) g(x) dx \\
&\lesssim \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \int_{x \in \mathcal{X}} |\Delta_5(x) \lambda(x)|^2 g(x) dx \\
&\lesssim \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \sup_{x \in \mathcal{X}} |\Delta_4(x)|^2 + \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \xi(X, T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}^2 | T = t \right] \\
&\lesssim h^{2(\alpha_1 \wedge \alpha_2)} + \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)]^2 | T = t \right] \\
&= O \left( \epsilon^{2\alpha_0} \vee h^{2(\alpha_1 \wedge \alpha_2)} \right),
\end{aligned} \quad (\text{S8.22})$$

where the first equality holds by (S8.5); the first inequality by (S8.4) and the properties of projections; the second inequality holds by definition of  $\xi(x, t)$  and  $\sup_{x \in \mathcal{X}} |\lambda(x)| \lesssim 1$ ; the third inequality holds by (S8.21) and Jensen's inequality; the last equality by (S8.8).

Recalling the definition of  $\Delta_3(x)$  in (S8.18) and using (S8.5), we have

$$\mathbb{E}[\{\hat{\xi}(X, T) - \xi(X, T)\} \bar{z}_{k_x}(X) K_{t,h}(T)] = \mathbb{E}[\Delta_3(X) \bar{z}_{k_x}(X)] = \int_{x \in \mathcal{X}} \Delta_3(x) \lambda(x) \bar{z}_{k_x}(x) g(x) dx, \quad (\text{S8.23})$$

$$\mathbb{E}[\{b_\beta(X, T) - b_{\beta_t^*}(X, T)\} \bar{z}_{k_x}(X) K_{t,h}(T)] = \mathbb{E}[\Delta_5(X) \bar{z}_{k_x}(x)] = \int_{x \in \mathcal{X}} \Delta_5(x) \lambda(x) \bar{z}_{k_x}(x) g(x) dx. \quad (\text{S8.24})$$

Then, we have

$$\begin{aligned} & \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \mathbb{E}[\{\hat{\xi}(X, T) - \xi(X, T)\} K_{t,h}(T) \bar{z}_{k_x}(X)^\top] \tilde{\Omega}_{k_x}^{-1} \bar{z}_{k_x}(X) K_{t,h}(T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)] \right|^2 \right] \\ &= \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \int_{x \in \mathcal{X}} \Delta_3(x) \lambda(x) \bar{z}_{k_x}(x)^\top g(x) dx \tilde{\Omega}_{k_x}^{-1} \bar{z}_{k_x}(X) K_{t,h}(T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)] \right|^2 \right] \\ &= \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \Pi_g [\Delta_3(X) \lambda(X) \bar{z}_{k_x}(X)] K_{t,h}(T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)] \right|^2 \right] \\ &\lesssim \sup_{x \in \mathcal{X}} |\Pi_g [\Delta_3(\cdot) \lambda(\cdot) \bar{z}_{k_x}](x)|^2 \cdot \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |K_{t,h}(T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)]|^2 \right] \\ &\lesssim \sup_{x \in \mathcal{X}} |\Delta_3(x) \lambda(x)|^2 \cdot \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |K_{t,h}(T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)]|^2 \right] = O \left( \frac{\epsilon^{2\alpha_0} \vee h}{h} \right), \end{aligned} \quad (\text{S8.25})$$

where the last inequality holds by (S8.7) and the last equality holds by Lemma 19 and (S8.4). Hence, we obtain the result (i).

For the result (ii),

$$\begin{aligned} & \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| K_{t,h}(T) \mathbb{E}[\bar{z}_{k_x}(X)^\top] \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) K_{t,h}(T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}] \right|^2 \right] \\ &= \mathbb{E} \left[ |K_{t,h}(T)|^2 \right] \cdot \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \mathbb{E}[\bar{z}_{k_x}(X)^\top] \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) K_{t,h}(T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}] \right|^2 \\ &= \mathbb{E} \left[ |K_{t,h}(T)|^2 \right] \cdot \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \mathbb{E}[\bar{z}_{k_x}(X)^\top] \tilde{\Omega}_{k_x}^{-1} \int_{x \in \mathcal{X}} \bar{z}_{k_x}(x) \Delta_5(x) \lambda(x) g(x) dx \right|^2 \\ &= \mathbb{E} \left[ |K_{t,h}(T)|^2 \right] \cdot \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |\mathbb{E}[\Pi_g[\Delta_5(\cdot) \lambda(\cdot) \bar{z}_{k_x}](X)]|^2 \\ &\lesssim \mathbb{E} \left[ |K_{t,h}(T)|^2 \right] \cdot \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ |\Pi_g[\Delta_5(\cdot) \lambda(\cdot) \bar{z}_{k_x}](X)|^2 \right] \\ &\lesssim \frac{\epsilon^{2\alpha_0} \vee h^{2(\alpha_1 \wedge \alpha_2)}}{h}, \end{aligned}$$

where the second equality uses (S8.24) and the last inequality holds by (S8.22).

For the result (iii),

$$\sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| K_{t,h}(T) \hat{\xi}(X, T) \bar{z}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) K_{t,h}(T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}] \right|^2 \right]$$

$$\begin{aligned}
&= \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| K_{t,h}(T) \widehat{\xi}(X, T) \bar{z}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \int_{x \in \mathcal{X}} \bar{z}_{k_x}(x) \Delta_5(x) \lambda(x) g(x) dx \right|^2 \right] \\
&\lesssim \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ |K_{t,h}(T) \Pi_g[\Delta_5(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](X)|^2 \xi(X, T) \right] \\
&= \mathbb{E} \left[ |K_{t,h}(T)|^2 \right] \cdot \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ |\Pi_g[\Delta_5(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](X)|^2 \right] \\
&\lesssim \frac{\epsilon^{2\alpha_0} \vee h^{2(\alpha_1 \wedge \alpha_2)}}{h},
\end{aligned}$$

where the first inequality holds since  $\xi(x, t)$  and  $\widehat{\xi}(x, t)$  are uniformly bounded, and the last inequality holds by (S8.22).

Similarly, for the result (iv), we have

$$\begin{aligned}
&\sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| \mathbb{E}[K_{t,h}(T)] \bar{z}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) K_{t,h}(T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}] \right|^2 \right] \\
&= |\mathbb{E}[K_{t,h}(T)]|^2 \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ \left| \bar{z}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \int_{x \in \mathcal{X}} \bar{z}_{k_x}(x) \Delta_5(x) \lambda(x) g(x) dx \right|^2 \right] \\
&= |\mathbb{E}[K_{t,h}(T)]|^2 \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \mathbb{E} \left[ |\Pi_g[\Delta_5(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](X)|^2 \right] \lesssim \epsilon^{2\alpha_0} \vee h^{2(\alpha_1 \wedge \alpha_2)},
\end{aligned}$$

where the last inequality uses  $\mathbb{E}[K_{t,h}(T)] \lesssim 1$  and (S8.22).  $\square$

**Lemma 21.** Under Assumptions 6–8, 13, and 15, for any  $h \in \mathcal{H}_{4,n}$ , we have

$$\sup_{\beta \in \mathcal{B}} \sup_{o \in \mathcal{O}} |\mathbb{P}^2 h(o; \beta)| \lesssim \frac{1}{h}.$$

*Proof.* For any  $h \in \mathcal{H}_{4,n}$ , the explicit expression of  $\mathbb{P}^2 h$  is given by

$$\begin{aligned}
&\mathbb{P}^2 h(O; \beta) \\
&= \frac{1}{3} \{K_{t,h}(T) \widehat{\xi}(X, T) - \mathbb{E}[K_{t,h}(T)]\} \bar{z}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) K_{t,h}(T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}] \\
&\quad + \frac{1}{3} \mathbb{E}[\{\widehat{\xi}(X, T) - \xi(X, T)\} K_{t,h}(T) \bar{z}_{k_x}(X)^\top] \tilde{\Omega}_{k_x}^{-1} \bar{z}_{k_x}(X) K_{t,h}(T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)] \\
&\quad - \frac{1}{3} K_{t,h}(T) \mathbb{E}[\bar{z}_{k_x}(X)^\top] \tilde{\Omega}_{k_x}^{-1} \mathbb{E}[\bar{z}_{k_x}(X) K_{t,h}(T) \{b_\beta(X, T) - b_{\beta_t^*}(X, T)\}].
\end{aligned} \tag{S8.26}$$

Recalling the definitions of  $\Delta_3(x)$  and  $\Delta_4(x)$  in (S8.18) and (S8.20), for any  $h \in \mathcal{H}_{4,n}$ , we have:

$$\begin{aligned}
\mathbb{P}^2 h(O; \beta) &= \frac{1}{3} \{K_{t,h}(T) \widehat{\xi}(X, T) - \mathbb{E}[K_{t,h}(T)]\} \Pi_g[\Delta_5(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](X) \\
&\quad + \frac{1}{3} \Pi_g[\Delta_3(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](X) K_{t,h}(T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)] \\
&\quad - \frac{1}{3} K_{t,h}(T) \mathbb{E}[\Pi_g[\Delta_5(\cdot) \lambda(\cdot) | \bar{z}_{k_x}](X)].
\end{aligned}$$

By (S8.7), we have

$$\begin{aligned} \sup_{x \in \mathcal{X}} |\Pi_g[\Delta_3(\cdot)\lambda(\cdot)|\bar{z}_{k_x}](x)| &\lesssim \sup_{x \in \mathcal{X}} |\Delta_3(x)\lambda(x)| \lesssim 1, \\ \text{and } \sup_{x \in \mathcal{X}} |\Pi_g[\Delta_5(\cdot)\lambda(\cdot)|\bar{z}_{k_x}](x)| &\lesssim \sup_{x \in \mathcal{X}} |\Delta_5(x)\lambda(x)| \lesssim 1. \end{aligned}$$

Since  $\|\hat{\xi}(\cdot, \cdot)\|_\infty \lesssim 1$ ,  $\|\Gamma(\cdot, \cdot, \beta)\|_\infty \lesssim 1$ , and  $\|K_{t,h}(\cdot)\|_\infty \lesssim 1/h$ , it follows that

$$\sup_{\beta \in \mathcal{B}} \sup_{o \in \mathcal{O}} |\pi_1 h(o; \beta)| \lesssim \frac{1}{h},$$

which completes the proof.  $\square$

**Lemma 22.** *Suppose that Assumptions 6–8 and 13 hold and  $k_x, n \rightarrow \infty$  with  $k_x = o(n^2)$ . Then, for a given  $\beta$ , conditioning on the nuisance sample,*

$$\text{var} \left( \tilde{\psi}_{t,k}^{(2)}(\beta) \right) = O \left( \frac{1}{nh} + \frac{k_x}{n^2 h^2} \right).$$

*Proof.* First, we have:

$$\text{var} \left( \tilde{\psi}_{t,k}^{(2)}(\beta) \right) \lesssim \text{var} \left( \mathbb{U}_{n,2} \hat{\Upsilon}_\beta \right) + \text{var} \left( \mathbb{U}_{n,3} \left( \text{IF}_{\text{B}_{\psi,t,k}}^{(2)}(\beta) \right) \right),$$

where

$$\begin{aligned} \hat{\Upsilon}_\beta(O_1, O_2) &:= K_{t,h}(T_1) \hat{\xi}(X_1, T_1) \{ \Gamma_\beta - \hat{b}_{\beta_{\text{init}}}(X_1, T_1) \} + \hat{b}_{\beta_{\text{init}}}(X_1, T_2) K_{t,h}(T_2) \text{ and} \\ \text{IF}_{\text{B}_{\psi,t,k}}^{(2)}(\beta) &:= \{ K_{t,h}(T_1) \hat{\xi}(X_1, T_1) - K_{t,h}(T_3) \} \bar{z}_{k_x}^\top(X_1) \tilde{\Omega}_{k_x}^{-1} \bar{z}_{k_x}(X_2) \{ \Gamma(Y_2, T_2, \beta) - \hat{b}_{\beta_{\text{init}}}(X_2, T_2) \} K_{t,h}(T_2). \end{aligned}$$

Using Lemma 15 in Liu et al. (2017) and a Taylor expansion argument similar to that used in the proof of Lemma 19, we obtain

$$\begin{aligned} \text{var} \left( \mathbb{U}_{n,2} \hat{\Upsilon}_\beta \right) &= \text{var} \left[ \mathbb{U}_{n,2} \left\{ S_2 \left( \hat{\Upsilon}_\beta \right) \right\} \right] \\ &\lesssim \frac{1}{n} \mathbb{E} \left[ \left\{ \mathbb{E} \left[ S_2 \left( \hat{\Upsilon}_\beta \right) \mid O_1 \right] \right\}^2 \right] + \frac{1}{n^2} \mathbb{E} \left[ \left\{ \mathbb{E} \left[ S_2 \left( \hat{\Upsilon}_\beta \right) \mid O_1, O_2 \right] \right\}^2 \right] \lesssim \frac{1}{nh}, \end{aligned}$$

where  $S_2$  is the symmetrization operator defined in Section S6.1. Similarly, for  $\text{var} \left\{ \mathbb{U}_{n,3} \left( \text{IF}_{\text{B}_{\psi,t,k}}^{(2)}(\beta) \right) \right\}$ , we have

$$\begin{aligned} \text{var} \left( \mathbb{U}_{n,3} \left( \text{IF}_{\text{B}_{\psi,t,k}}^{(2)}(\beta) \right) \right) &= \text{var} \left[ \mathbb{U}_{n,3} \left\{ S_3 \left( \text{IF}_{\text{B}_{\psi,t,k}}^{(2)}(\beta) \right) \right\} \right] \\ &\lesssim \sum_{j=1}^3 \frac{1}{n^j} \mathbb{E} \left[ \left\{ \mathbb{E} \left[ S_3 \left( \text{IF}_{\text{B}_{\psi,t,k}}^{(2)}(\beta) \right) \mid O_1, \dots, O_j \right] \right\}^2 \right] =: \sum_{j=1}^3 \zeta_j. \end{aligned}$$

Following (S8.26) and Lemma 20, we have

$$\mathbb{E} \left[ \left\{ \mathbb{E}[S(\mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta)) \mid \mathcal{O}_1] \right\}^2 \right] \lesssim \frac{1}{h}, \text{ and hence } \zeta_1 \lesssim \frac{1}{nh}.$$

For  $\zeta_3$ , following the proof of (S8.31), we conclude

$$\zeta_3 \lesssim \frac{1}{n^3} \mathbb{E} \left[ \left( \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta) \right)^2 \right] \lesssim \frac{k_x}{n^3 h^2}.$$

For  $\zeta_2$ , by Jensen's inequality, we obtain:

$$\zeta_2 \lesssim \frac{1}{n^2} \mathbb{E} \left[ \left( \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta) \right)^2 \right] \lesssim \frac{k_x}{n^2 h^2}.$$

Combining all the analysis above, we obtain

$$\text{var} \left( \mathbb{U}_{n,3} \left( \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta) \right) \right) \lesssim \sum_{j=1}^3 \zeta_j \lesssim \frac{1}{nh} + \frac{k_x}{n^2 h^2},$$

which completes the proof.  $\square$

**Lemma 23.** *Under Assumptions 6–8, 13, and 15, for any  $\epsilon > 0$ ,*

$$\mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \left\{ \bar{\psi}_{t,k}(\beta) - \tilde{\psi}_{t,k}^{(2)}(\beta) \right\} - \left\{ \bar{\psi}_{t,k}(\beta_t^*) - \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right\} \right| \right] \lesssim \left( \frac{\sqrt{k}}{nh} + \frac{1}{\sqrt{nh}} \right) \cdot (\epsilon^{\alpha_0} \vee \sqrt{h}) \cdot \log n.$$

*Proof.* We begin by decomposing the supremum as follows:

$$\begin{aligned} & \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \bar{\psi}_{t,k}(\beta) - \tilde{\psi}_{t,k}^{(2)}(\beta) - \bar{\psi}_{t,k}(\beta_t^*) + \tilde{\psi}_{t,k}^{(2)}(\beta_t^*) \right| \\ &= \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \bar{\psi}_{t,k}(\beta) - \hat{\psi}_t^{(1)}(\beta) - \mathbb{U}_{n,3}[\mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta)] - \bar{\psi}_{t,k}(\beta_t^*) + \hat{\psi}_t^{(1)}(\beta_t^*) + \mathbb{U}_{n,3}[\mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta_t^*)] \right| \\ &\leq \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \mathbb{U}_{n,3} \left\{ \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta) - \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta_t^*) - \mathbb{E} \left( \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta) - \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta_t^*) \right) \right\} \right| \end{aligned} \quad (\text{S8.27})$$

$$+ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \hat{\psi}_t^{(1)}(\beta) - \hat{\psi}_t^{(1)}(\beta_t^*) - \mathbb{E} \left[ \hat{\psi}_t^{(1)}(\beta) - \hat{\psi}_t^{(1)}(\beta_t^*) \right] \right|. \quad (\text{S8.28})$$

We first analyze the term (S8.28). Consider the function class

$$\mathcal{H}_3 := \left\{ h(O, \beta) = K_{t,h}(T) \hat{\xi}(X, T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)] : \beta \in \mathcal{B}, |\beta - \beta_t^*| \lesssim \epsilon \right\}.$$

with envelope function  $H_3(O) := \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |K_{t,h}(T) \hat{\xi}(X, T) [\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)]|$ . By an argument similar to that for  $\mathcal{H}_{1,\delta}$  (see (S7.20)), we find that  $\mathcal{H}_3$  is also of VC type with bounded

characteristics. Since  $\|\widehat{\xi}(\cdot, \cdot)\|_\infty \lesssim 1$ , Lemma 19 gives

$$\begin{aligned} \mathbb{E}[|H_3(O)|^2] &= \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} \left| \widehat{\xi}(X, T) \{ \Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*) \} K_{t,h}(T) \right|^2 \right] \\ &\lesssim \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} | \Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*) \} K_{t,h}(T) |^2 \right] = O \left( \frac{\epsilon^{2\alpha_0} \vee h}{h} \right). \end{aligned} \quad (\text{S8.29})$$

Thus,  $\|H_3\|_{\mathbb{P},2} \lesssim (\epsilon^{\alpha_0} \vee \sqrt{h})/\sqrt{h}$ . Applying Theorem 2.14.1 of van der Vaart and Wellner (2023) to  $\mathcal{H}_3$  yields

$$\mathbb{E} [ |(\text{S8.28})| ] = \mathbb{E} [ \|\mathcal{U}_n(h)\|_{\mathcal{H}_3} ] \lesssim \frac{J_1(1, \mathcal{H}_3, H_3) \cdot \|H_3\|_{\mathbb{P},2}}{\sqrt{n}} \lesssim \frac{\epsilon^{\alpha_0} \vee \sqrt{h}}{\sqrt{nh}}. \quad (\text{S8.30})$$

We now turn to the term (S8.27). Consider the function class

$$\mathcal{H}_{4,n} := \left\{ h(O_1, O_2, O_3, \beta) = \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta) - \mathbf{IF}_{\mathbf{B}_{\psi,t,k}}^{(2)}(\beta_t^*) : \beta \in \mathcal{B}, |\beta - \beta_t^*| \lesssim \epsilon \right\}.$$

Similarly, we can show that  $\mathcal{H}_{4,n}$  is also of VC type with bounded characteristics and envelope

$$\begin{aligned} H_{4,n}(O_1, O_2, O_3) &:= |\{ K_{t,h}(T_1) \widehat{\xi}(X_1, T_1) - K_{t,h}(T_3) \} \bar{\mathbf{z}}_{k_x}(X_1)^\top \tilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(X_2) K_{t,h}(T_2)| \\ &\quad \times \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |\Gamma(Y_2, T_2, \beta) - \Gamma(Y_2, T_2, \beta_t^*)|. \end{aligned}$$

Under Assumptions 6, 8 and 13, we have  $\|H_{4,n}(\cdot, \cdot, \cdot)\|_\infty \lesssim h^2 \zeta(k_x)^2 \lesssim k_x h^2$ . In addition,

$$\begin{aligned} &\mathbb{E}[|H_{4,n}(O_1, O_2, O_3)|^2] \quad (\text{S8.31}) \\ &= \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)|^2 K_{t,h}(T)^2 \bar{\mathbf{z}}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \right. \\ &\quad \times \mathbb{E} \left[ K_{t,h}(T)^2 \{ \widehat{\xi}(X, T) - 1 \}^2 \xi(X, T) \bar{\mathbf{z}}_{k_x}(X) \bar{\mathbf{z}}_{k_x}(X)^\top \right] \tilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(X) \left. \right] \\ &\lesssim \frac{1}{h} \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)|^2 K_{t,h}(T)^2 \bar{\mathbf{z}}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \cdot \tilde{\Omega}_{k_x} \cdot \tilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(X) \right] \\ &= \frac{1}{h} \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)|^2 |K_{t,h}(T)|^2 \bar{\mathbf{z}}_{k_x}(X)^\top \tilde{\Omega}_{k_x}^{-1} \bar{\mathbf{z}}_{k_x}(X) \right] \\ &\lesssim \frac{\zeta(k_x)^2}{h} \mathbb{E} \left[ \sup_{|\beta - \beta_t^*| \lesssim \epsilon} |\Gamma(Y, T, \beta) - \Gamma(Y, T, \beta_t^*)|^2 \cdot |K_{t,h}(T)|^2 \right] \\ &\lesssim \frac{k_x \cdot (\epsilon^{2\alpha_0} \vee h)}{h^2}, \end{aligned}$$

where the inequalities follow from Lemma 19 and Assumption 8. Thus,  $\|H_{4,n}\|_{\mathbb{P}^3,2} \lesssim \sqrt{k_x}(\epsilon^{\alpha_0} \vee \sqrt{h})/h$ . Further, by Lemma 20, we have  $\sup_{h \in \mathcal{H}_{4,n}} \|\mathbb{P}^2 h\|_{\mathbb{P},2} \lesssim (\epsilon^{\alpha_0} \vee \sqrt{h})/\sqrt{h}$  and by Lemma 21,

$\sup_{h \in \mathcal{H}_{4,n}} \|\mathbb{P}^2 h(\cdot)\|_\infty \lesssim 1/h$ . Thus, applying Lemma 10 yields

$$\mathbb{E}[(\text{S8.28})] = \mathbb{E}[\|\mathbb{U}_{n,3}(h)\|_{\mathcal{H}_{4,n}}] \lesssim \frac{\sqrt{k_x} \cdot (\epsilon^{\alpha_0} \vee \sqrt{h}) \cdot \log n}{nh} + \frac{(\epsilon^{\alpha_0} \vee \sqrt{h}) \cdot \sqrt{\log n}}{\sqrt{nh}}. \quad (\text{S8.32})$$

Therefore, combining (S8.30) and (S8.32) completes the proof.  $\square$

## S9 Supplementary Information on the Numerical Experiments

### S9.1 Supplementary simulation results

In this section, we collect supplementary information on the simulation study conducted in Section 5.1. In particular, we display the normal-QQ plots of our proposed higher-order estimators over various settings in the simulation (i.e. varying the sample size  $n$  and the smoothness parameter  $s$ ): see Figure 5 for  $\beta_0^*$  and Figure 6 for  $\beta_1^*$  below for Case 1 and Figure 7 for  $\beta_0^*$  and Figure 8 for  $\beta_1^*$  for Case 2.

It is quite clear that the higher-order estimator proposed here for QTE with binary treatment is asymptotically normally distributed across different sample sizes

$$n \in \{250, 500, 750, 1000, 1500, 2000, 2500, 3000\}$$

and different smoothness levels of the nuisance parameters with  $s = 0.25, 0.4, 0.6$ .

The GOE estimator (Ai et al., 2021) is obtained by solving (6) with  $\hat{\xi}$  estimated through the following quadratic programming problem:

$$\min_{\boldsymbol{\xi}} \sum_{i=1}^n \frac{(\xi_i - 1)^2}{2} \text{ s.t. } \frac{1}{n} \sum_{i=1}^n \xi_i \bar{\phi}_k(X_i, T_i) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \bar{\phi}_k(X_i, T_j).$$

Here,  $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$  with  $\xi_i \equiv \xi(X_i, T_i)$ , and  $\bar{\phi}_k$  is a power series basis of length  $k$ . The parameter  $k$  is selected by minimizing a cross-validation criterion.

Sensitivity analyses of our proposed higher-order estimators are also conducted, specifically on the choice of  $k$  and on the initial estimator  $\beta_{\text{init}}$ .

### S9.2 Supplementary results on real data analysis

This subsection provides additional numerical results for Section 5.2. Specifically, Tables 3–4 report sensitivity analyses of our higher-order estimator with respect to the choice of  $k$  and to the quality of the initial estimator  $\beta_{\text{init}}$ . A summary of the results is provided at the end of Section 5.2.

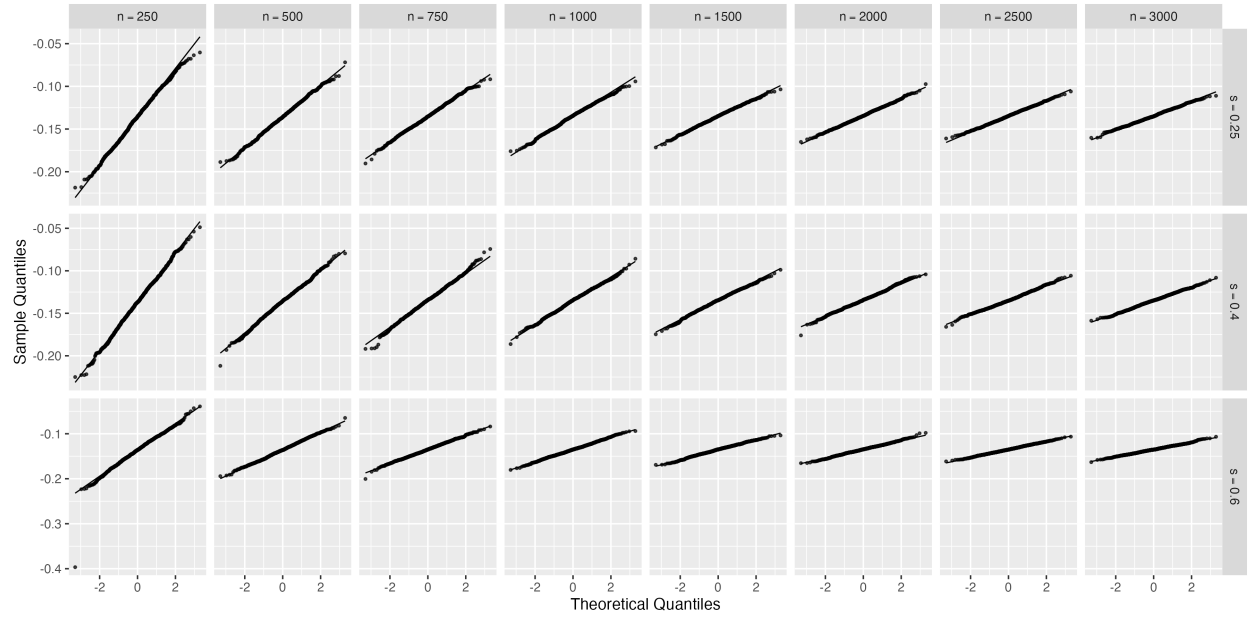


Figure 5: Normal-QQ plots for the first coordinate of  $\hat{\beta}^{(2)}$  across all simulation settings in Case 1, based on 1000 Monte Carlo replications.

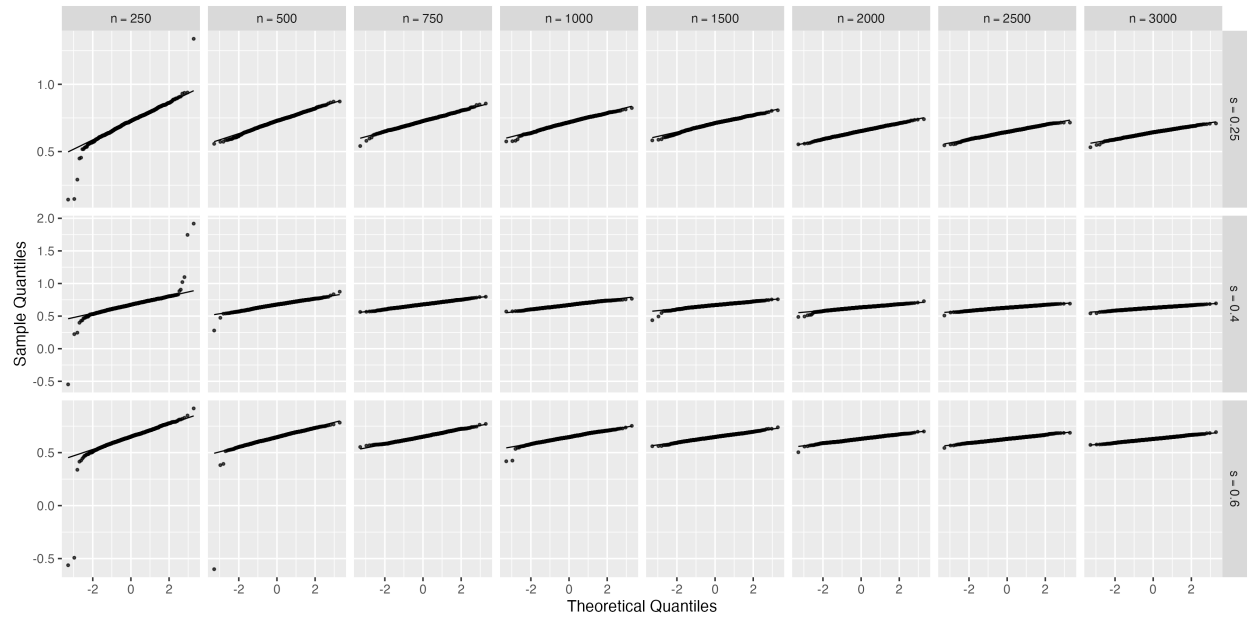


Figure 6: Normal-QQ plots for the second coordinate of  $\hat{\beta}^{(2)}$  across all simulation settings in Case 1, based on 1000 Monte Carlo replications.

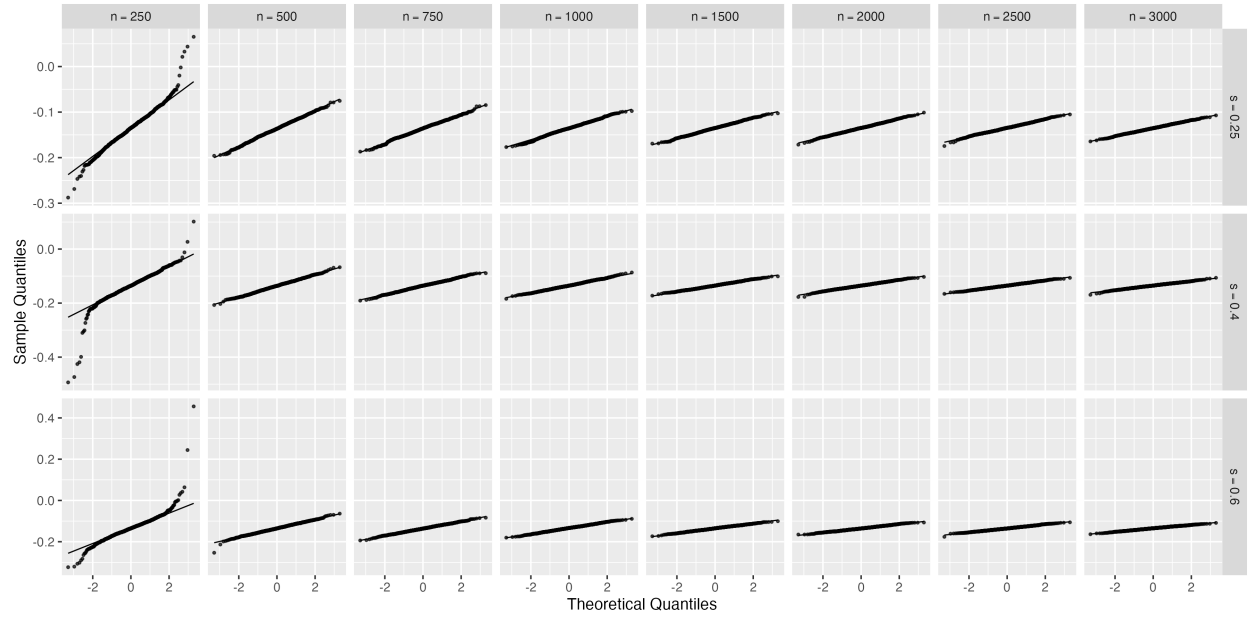


Figure 7: Normal-QQ plots for the first coordinate of  $\hat{\beta}^{(2)}$  across all simulation settings in Case 2, based on 1000 Monte Carlo replications.

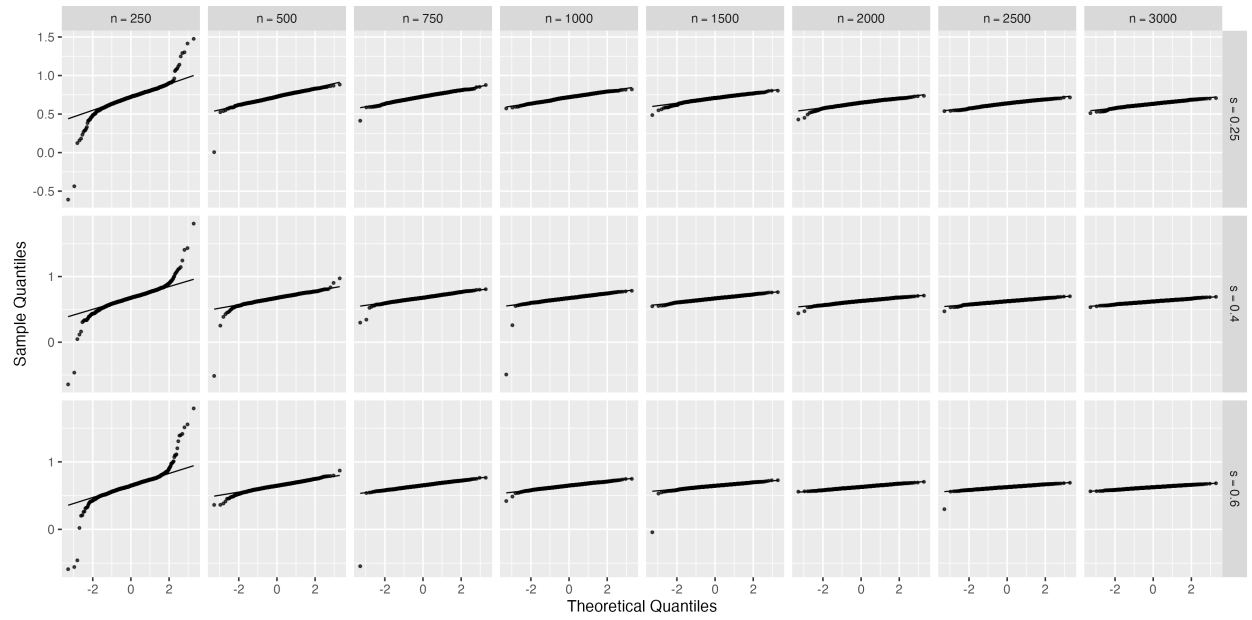


Figure 8: Normal-QQ plots for the second coordinate of  $\hat{\beta}^{(2)}$  across all simulation settings in Case 2, based on 1000 Monte Carlo replications.

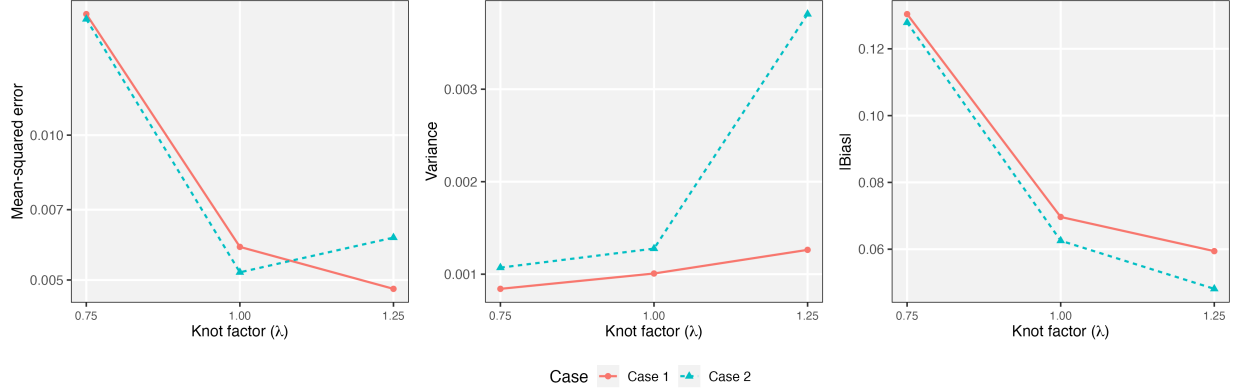


Figure 9: Sensitivity of the higher-order estimator for the QTE,  $\beta_1^* - \beta_0^*$ , to the choice of spline basis dimension  $k$ . The horizontal axis reports the knot factor  $\lambda$ , where the number of interior knots for each continuous covariate is set to  $\lceil \lambda n / 100 \rceil$ . The three panels report the empirical mean squared error, variance, and absolute bias across 1000 Monte Carlo replications for Cases 1 and 2 with  $n = 2000$  and  $s = 0.25$ .

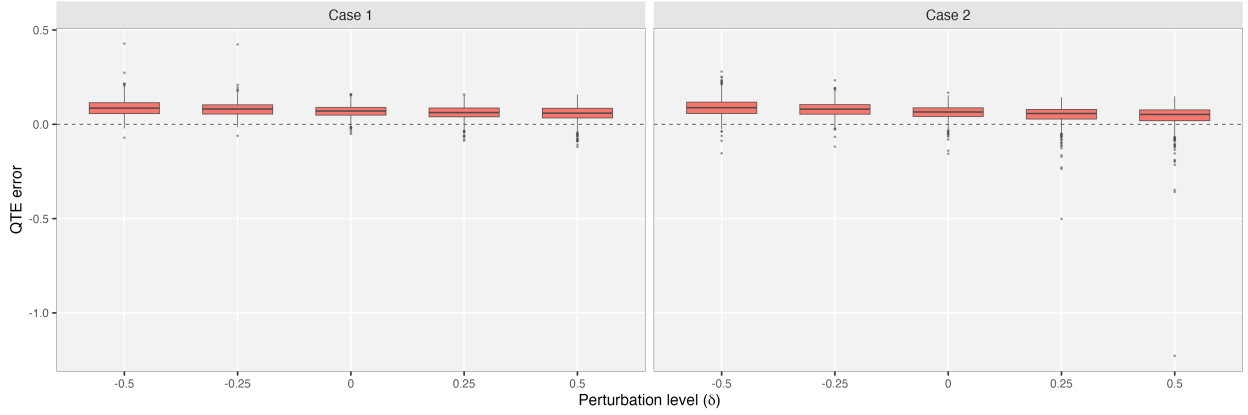


Figure 10: Sensitivity of the higher-order estimator for the QTE,  $\beta_1^* - \beta_0^*$ , to perturbations in the initial estimator. The horizontal axis reports the perturbation level  $\delta$ , where the perturbed initial estimator is defined by  $\beta_{\text{init}}^{(\delta)} = \beta_{\text{init}} + \delta(-1, 1)^\top$ . The boxplots display the resulting QTE estimation errors across 1000 Monte Carlo replications for Cases 1 and 2 with  $n = 2000$  and  $s = 0.25$ .

| $\tau$ | Knot Scale ( $\lambda$ ) | Forest | Neural Net | Boosting | LASSO |
|--------|--------------------------|--------|------------|----------|-------|
| 0.25   | 0.75                     | 1.00   | 1.09       | 1.09     | 1.11  |
|        | 1.00                     | 1.01   | 1.09       | 1.09     | 1.11  |
|        | 1.25                     | 1.00   | 1.09       | 1.06     | 1.11  |
| 0.50   | 0.75                     | 4.67   | 4.47       | 4.50     | 4.45  |
|        | 1.00                     | 4.74   | 4.37       | 4.40     | 4.35  |
|        | 1.25                     | 4.52   | 4.24       | 4.35     | 4.22  |
| 0.75   | 0.75                     | 13.08  | 11.70      | 12.04    | 11.79 |
|        | 1.00                     | 12.68  | 11.77      | 12.15    | 11.90 |
|        | 1.25                     | 12.81  | 11.75      | 12.15    | 11.90 |

Table 3: Sensitivity of the HOE estimates of the QTE of 401(k) eligibility (in thousand dollars) to the choice of spline basis size. The knot scale  $\lambda$ , defined in Section 5.2, controls the size of the spline basis, with a larger value corresponding to a larger basis. Here  $\tau \in \{25\%, 50\%, 75\%\}$  denotes the quantile level of the QTE.

| $\tau$ | Perturbation Level ( $\delta$ ) | Forest | Neural Net | Boosting | LASSO |
|--------|---------------------------------|--------|------------|----------|-------|
| 0.25   | -0.50                           | 1.10   | 1.12       | 1.11     | 1.12  |
|        | -0.25                           | 1.05   | 1.12       | 1.10     | 1.11  |
|        | 0.00                            | 1.01   | 1.09       | 1.09     | 1.11  |
|        | 0.25                            | 1.00   | 1.11       | 1.09     | 1.11  |
|        | 0.50                            | 1.00   | 1.09       | 1.06     | 1.11  |
| 0.50   | -0.50                           | 4.75   | 4.40       | 4.56     | 4.36  |
|        | -0.25                           | 4.85   | 4.30       | 4.53     | 4.32  |
|        | 0.00                            | 4.74   | 4.37       | 4.40     | 4.35  |
|        | 0.25                            | 4.67   | 4.26       | 4.42     | 4.35  |
|        | 0.50                            | 4.30   | 4.25       | 4.45     | 4.35  |
| 0.75   | -0.50                           | 12.68  | 11.90      | 12.10    | 11.90 |
|        | -0.25                           | 12.68  | 11.80      | 12.15    | 11.90 |
|        | 0.00                            | 12.68  | 11.77      | 12.15    | 11.90 |
|        | 0.25                            | 12.43  | 11.77      | 12.15    | 11.90 |
|        | 0.50                            | 12.63  | 11.77      | 12.15    | 11.90 |

Table 4: Sensitivity of the HOE estimates of the QTE of 401(k) eligibility (in thousand dollars) to deterministic perturbations of the initial estimator. The perturbation follows the same scheme as in the simulation study, namely  $\beta_{\text{init}}^{(\delta)} = \beta_{\text{init}} + \delta(-1, 1)^\top$ , where  $\delta$  denotes the perturbation level. Here  $\tau \in \{25\%, 50\%, 75\%\}$  denotes the quantile level of the QTE.

## References

- Chunrong Ai and Xiaohong Chen. Efficient estimation of models with conditional moment restrictions containing unknown functions. *Econometrica*, 71(6):1795–1843, 2003.
- Chunrong Ai, Oliver Linton, Kaiji Motegi, and Zheng Zhang. A unified framework for efficient

- estimation of general treatment models. *Quantitative Economics*, 12(3):779–816, 2021.
- Donald WK Andrews. Empirical process methods in econometrics. *Handbook of Econometrics*, 4: 2247–2294, 1994.
- Afonso S Bandeira, Kevin Lucca, Petar Nizic-Nikolac, and Ramon van Handel. Matrix chaos inequalities and chaos of combinatorial type. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 795–805, 2025.
- Rabi N Bhattacharya and Jayanta K Ghosh. A class of  $U$ -statistics and asymptotic normality of the number of  $k$ -clusters. *Journal of Multivariate Analysis*, 43(2):300–330, 1992.
- Matteo Bonvini and Edward H Kennedy. Fast convergence rates for dose-response estimation. *arXiv preprint arXiv:2207.11825*, 2022.
- Matias D Cattaneo and Michael Jansson. Kernel-based semiparametric estimators: Small bandwidth asymptotics and bootstrap consistency. *Econometrica*, 86(3):955–995, 2018.
- Matias D Cattaneo, Michael Jansson, and Xinwei Ma. Two-step estimation and inference with possibly many included covariates. *The Review of Economic Studies*, 86(3):1095–1122, 2019.
- Matias D Cattaneo, Max H Farrell, Michael Jansson, and Ricardo P Masini. Higher-order refinements of small bandwidth asymptotics for density-weighted average derivative estimators. *Journal of Econometrics*, 252(Part B):105855, 2025.
- Xiaohong Chen and Haitian Xie. Local overidentification and efficiency gains in modern causal inference and data combination. *arXiv preprint arXiv:2510.16683*, 2025.
- Xiaohui Chen and Kengo Kato. Jackknife multiplier bootstrap: Finite sample approximations to the  $U$ -process supremum with applications. *Probability Theory and Related Fields*, 176(3):1097–1163, 2020.
- Edward H Kennedy, Zongming Ma, Matthew D McHugh, and Dylan S Small. Non-parametric methods for doubly robust estimation of continuous treatment effects. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 79(4):1229–1245, 2017.
- Michael R Kosorok. *Introduction to Empirical Processes and Semiparametric Inference*. Springer, 2008.
- Licong Lin, Fangzhou Su, Wenlong Mou, Peng Ding, and Martin Wainwright. When is it worthwhile to jackknife? Breaking the quadratic barrier for  $Z$ -estimators. *arXiv preprint arXiv:2411.02909*, 2024.
- Lin Liu, Rajarshi Mukherjee, Whitney K Newey, and James M Robins. Semiparametric efficient empirical higher order influence function estimators. *arXiv preprint arXiv:1705.07577*, 2017.
- Lin Liu, Rajarshi Mukherjee, and James M Robins. Assumption-lean falsification tests of rate double-robustness of double-machine-learning estimators. *Journal of Econometrics*, 240(2):105500, 2024.

- Stanislav Minsker and Xiaohan Wei. Moment inequalities for matrix-valued  $U$ -statistics of order 2. *Electronic Journal of Probability*, 24(133):1–32, 2019.
- James Robins, Lingling Li, Eric Tchetgen Tchetgen, and Aad van der Vaart. Technical report: Higher order influence functions and minimax estimation of nonlinear functionals. *arXiv preprint arXiv:1601.05820*, 2016.
- James M Robins, Miguel A Hernán, and Babette Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, 2000.
- James M Robins, Lingling Li, Lin Liu, Rajarshi Mukherjee, Eric Tchetgen Tchetgen, and Aad van der Vaart. Minimax estimation of a functional on a structured high-dimensional model (Corrected version). *arXiv preprint arXiv:1512.02174*, 2023.
- Joel A Tropp. An introduction to matrix concentration inequalities. *Foundations and Trends® in Machine Learning*, 8(1-2):1–230, 2015.
- Aad W van der Vaart and Jon Wellner. *Weak Convergence and Empirical Processes: with Applications to Statistics*. Springer Science & Business Media, 2023.