SUPPLEMENT TO "CONCORDANCE AND VALUE INFORMATION CRITERIA FOR OPTIMAL TREATMENT DECISION"

By Chengchun Shi, Rui Song and Wenbin Lu

North Carolina State University

In this supplement, we extend our information criteria to multiple stages and present some technical conditions for Theorem 4.2 and proofs of Theorem 3.4, Theorem 4.2, Theorem 7.1 and Lemma 7.1. We omit the proof of Theorem 3.1 since it is similar to the proof of Theorem 4.2.

10. Extensions to multiple stages.

10.1. *Two-stage study.* To illustrate the idea, we first consider a two-stage study. Let $O_0 = (X_0^{(1)}, A_0^{(1)}, X_0^{(2)}, A_0^{(2)}, Y_0)$ where Y_0 denotes the final response, $A_0^{(1)}$ and $A_0^{(2)}$ refer to the treatments patient receives at time point t_1 and t_2 respectively, $X_0^{(1)} \in \mathbb{R}^{p_1}$ stands for the baseline covariates and $X_0^{(2)}$ denotes some intermediate covariates collected on the patient between t_1 and t_2. Let $\overline{X}_0^{(2)} = \{(X_0^{(1)})^T, A_0^{(1)}, (X_0^{(2)})^T\}^T$. Assume $A_0^{(1)}$ and $A_0^{(2)}$ are binary treatments. For any $a^{(1)}, a^{(2)} \in \{0,1\}$, let $X_0^{(2)*}(a^{(1)})$ and $Y_0^*(a^{(1)}, a^{(2)})$ be the potential outcomes of the patient if he/she receives treatment a_1 at t_1 and treatment a_2 at t_2. For a given treatment regime $\bar{d} = (d_1, d_2)$, define the potential outcome

$$Y^*(\bar{d}) = \sum_{a^{(1)}, a^{(2)} \in \{0,1\}} Y_0^*(a^{(1)}, a^{(2)}) I\{d_1(X_0^{(1)}) = a^{(1)}, d_2(\overline{X}_0^{(2)*}) = a^{(2)}\},$$

where $\overline{X}_0^{(2)*}$ is a shorthand for $[(X_0^{(1)})^T, a^{(1)}, \{X_0^{(2)*}(a^{(1)})^T\}]^T$. Let $\bar{d}^{opt} = (d_1^{opt}, d_2^{opt}) = \arg\max_{\bar{d}} \mathrm{E} Y^*(\bar{d})$. Define the Q-function and the contrast function as

$$\begin{aligned}
&Q^{(2)}(\bar{x}^{(2)}, a^{(2)}) = \mathrm{E}(Y | \overline{X}_0^{(2)} = \bar{x}^{(2)}, A_0^{(2)} = a^{(2)}),\\
&\tau^{(2)}(\bar{x}^{(2)}) = Q^{(2)}(\bar{x}^{(2)}, 1) - Q^{(2)}(\bar{x}^{(2)}, 0),\\
&Q^{(1)}(x^{(1)}, a^{(1)}) = \mathrm{E}\left(Q^{(2)}\{\overline{X}_0^{(2)}, d_2^{opt}(\overline{X}_0^{(2)})\} \mid X_0^{(1)} = x^{(1)}, A_0^{(1)} = a^{(1)}\right),\\
&\tau^{(1)}(x^{(1)}) = Q^{(1)}(x^{(1)}, 1) - Q^{(1)}(x^{(1)}, 0).
\end{aligned}$$

Under the following three assumptions,

(C1.) $Y_0 = \sum_{a^{(1)},a^{(2)}} Y_0^*(a^{(1)}, a^{(2)}) I(A_0^{(1)} = a^{(1)}, A_0^{(2)} = a^{(2)})$,
$X_0^{(2)} = \sum_{a^{(1)}} X_0^{(2)*}(a^{(1)}) I(A_0^{(1)} = a^{(1)})$,
(C2.) For any $a^{(1)}, a^{(2)} \in \{0,1\}$, $A_0^{(2)} \perp\!\!\!\perp \{Y_0^*(a^{(1)}, a^{(2)}), X_0^{(2)*}(a_1)\} | \{\overline{X}_0^{(2)}, A_0^{(1)}\}$, and $A_0^{(1)} \perp\!\!\!\perp \{Y_0^*(a^{(1)}, a^{(2)}), X_0^{(2)*}(a^{(1)})\} | X_0^{(1)}$,
(C3.) For any $a^{(1)}, a^{(2)}, x^{(1)}, x^{(2)}$, $\Pr(A_0^{(2)} = a^{(2)} | \overline{X}_0^{(2)} = \bar{x}^{(2)}) > 0$ and $\Pr(A_0^{(1)} = a^{(1)} | X_0^{(1)} = x^{(1)}) > 0$,
we have

$$d_1^{opt}(x^{(1)}) = I\{\tau^{(1)}(x^{(1)}) > 0\} \text{ and } d_2^{opt}(\bar{x}^{(2)}) = I\{\tau^{(2)}(\bar{x}^{(2)}) > 0\}.$$

Assumption (C2) is the sequential randomization assumption, which automatically holds in the sequential multiple assignment randomized trial (SMART) studies.

For simplicity, we focus on the fixed-p scenario. The case where the dimension of the covariates grows much faster than the sample size can be similarly discussed. The observed data are summarized as

$$\left\{ O_i = \left(X_i^{(1)}, A_i^{(1)}, X_i^{(2)}, A_i^{(2)}, Y_i \right) \right\}_{i=1}^n.$$

We focus on the class of linear decision rules $d_1 = I(\beta_1^T x^{(1)} + c_1 > 0)$, $d_2 = I(\beta_2^T \bar{x}^{(2)} + c_2 > 0)$ and select the support of β_1 and β_2 via backward induction. Assume

$$\tau^{(2)}(\bar{x}^{(2)}) = G^{(2)}(\beta_{0,2}^T \bar{x}^{(2)} + c_{0,2}),$$

for some $\beta_{0,2} \in \mathbb{R}^{p_2}$, $c_{0,2} \in \mathbb{R}$ and some monotonically increasing function $G^{(2)}$ with $G^{(2)}(0) = 0$. Let Ω_2 be the space of all candidate models for β_2. For a given model $M_2 \in \Omega_2$, let $\hat{\theta}_{2,M_2} = (\hat{c}_{2,M_2}, \hat{\beta}_{2,M_2}^T)^T$ be some estimator for $\theta_{0,2} = (c_{0,2}, \beta_{0,2}^T)^T$ on the restricted model space. These estimators can be obtained via robust learning or CAL. For any $\theta_2 = (c_2, \beta_2^T)^T$, define

$$\begin{aligned}
\widehat{V}^{(2)}(\theta_2) &= \frac{1}{n} \sum_{i=1}^n \left\{ \frac{A_i^{(2)} I\{\beta_2^T \overline{X}_i^{(2)} > -c_2\}}{\pi_i^{(2)}} + \frac{(1 - A_i^{(2)}) I\{\beta_2^T \overline{X}_i^{(2)} \le -c_2\}}{1 - \pi_i^{(2)}} \right\} Y_i, \\
\widehat{C}^{(2)}(\beta_2) &= \frac{1}{n(n-1)} \sum_{i \neq j} \left\{ \omega_i^{(2)} - \omega_j^{(2)} \right\} I(\beta_2^T \overline{X}_i^{(2)} > \beta_2^T \overline{X}_j^{(2)}),
\end{aligned}$$

where

$$\omega_i^{(2)} = \left(\frac{A_i^{(2)}}{\pi_i^{(2)}} - \frac{1 - A_i^{(2)}}{1 - \pi_i^{(2)}} \right) Y_i, \quad \pi_i^{(2)} = \Pr\left(A_i^{(2)} = 1 \mid \overline{X}_i^{(2)} \right).$$

Let

$$\mathrm{VIC}^{(2)}(\theta_2) = n\widehat{V}^{(2)}(\theta_2) - \kappa_n^{(2)}\|\theta_2\|_0, \quad \mathrm{CIC}^{(2)}(\theta_2) = n\widehat{C}^{(2)}(\theta_2) - \kappa_n^{(2)}\|\theta_2\|_0.$$

We first use $\mathrm{VIC}^{(2)}$ and $\mathrm{CIC}^{(2)}$ to estimate $M_{0,2}$, the support of $\theta_{0,2}$. Define

$$\begin{aligned}\widehat{M}_2^V &= \arg\max_{M_2\in\Omega_2} \mathrm{VIC}^{(2)}(\hat{\theta}_{2,M_2}),\\ \widehat{M}_2^C &= \arg\max_{M_2\in\Omega_2} \mathrm{CIC}^{(2)}(\hat{\theta}_{2,M_2}).\end{aligned}$$

Similar to Theorem 3.1, we can show $\mathrm{VIC}^{(2)}$ and $\mathrm{CIC}^{(2)}$ are consistent under certain conditions.

Let Ω_1 be the space of all candidate models for β_1. For any $M_1 \in \Omega_1$, let $\hat{\theta}_{M_1}^{(1)} = (\hat{c}_{1,M_1}, \hat{\beta}_{1,M_1}^T)^T$ be some estimator on the restricted model space. To introduce our information criteria, we define the pseudo outcomes

$$\begin{aligned}Y_i^{(1),V} &= \left\{\frac{A_i^{(2)}\hat{d}_2^V(\overline{X}_i^{(2)})}{\pi_i^{(2)}} + \frac{(1-A_i^{(2)})I\hat{d}_2^V(\overline{X}_i^{(2)})}{1-\pi_i^{(2)}}\right\} Y_i,\\ Y_i^{(1),C} &= \left\{\frac{A_i^{(2)}\hat{d}_2^C(\overline{X}_i^{(2)})}{\pi_i^{(2)}} + \frac{(1-A_i^{(2)})I\hat{d}_2^C(\overline{X}_i^{(2)})}{1-\pi_i^{(2)}}\right\} Y_i,\end{aligned}$$

where

$$\hat{d}_2^V(\bar{x}^{(2)}) = I(\hat{\beta}_{\widehat{M}_V^{(2)}}^T \bar{x}^{(2)} > -\hat{c}_{\widehat{M}_V^{(2)}}), \quad \hat{d}_2^C(\bar{x}^{(2)}) = I(\hat{\beta}_{\widehat{M}_C^{(2)}}^T \bar{x}^{(2)} > -\hat{c}_{\widehat{M}_C^{(2)}}).$$

For any $\theta_1 = (c_1, \beta_1^T)^T$, let

$$\begin{aligned}\widehat{V}^{(1)}(\theta_1) &= \frac{1}{n}\sum_{i=1}^n \left\{\frac{A_i^{(1)} I\{\beta_1^T X_i^{(1)} > -c_1\}}{\pi_i^{(1)}} + \frac{(1-A_i^{(1)}) I\{\beta_1^T X_i^{(1)} \le -c_1\}}{1-\pi_i^{(1)}}\right\} Y_i^{(1),V},\\ \widehat{C}^{(1)}(\beta_1) &= \frac{1}{n(n-1)}\sum_{i\neq j}\left\{\omega_i^{(1)} - \omega_j^{(1)}\right\} I(\beta_1^T X_i^{(1)} > \beta_1^T X_j^{(1)}),\end{aligned}$$

where

$$\omega_i^{(1)} = \left(\frac{A_i^{(1)}}{\pi_i^{(1)}} - \frac{1-A_i^{(1)}}{1-\pi_i^{(1)}}\right) Y_i^{(1),C}, \quad \pi_i^{(1)} = \Pr\left(A_i^{(1)} = 1 \mid X_i^{(1)}\right).$$

Define

$$\mathrm{VIC}^{(1)}(\theta_1) = n\widehat{V}^{(1)}(\theta_1) - \kappa_n^{(1)}\|\theta_1\|_0, \quad \mathrm{CIC}^{(1)}(\theta_1) = n\widehat{C}^{(1)}(\theta_1) - \kappa_n^{(1)}\|\theta_1\|_0,$$

and

$$\begin{aligned}\widehat{M}_1^V &= \arg\max_{M_1\in\Omega_1} \mathrm{VIC}^{(1)}(\hat{\theta}_{1,M_1}),\\ \widehat{M}_1^C &= \arg\max_{M_1\in\Omega_1} \mathrm{CIC}^{(1)}(\hat{\theta}_{1,M_1}).\end{aligned}$$

When $\tau^{(1)}(x^{(1)}) = G^{(1)}(\beta_{0,1}^T x^{(1)} + c_{0,1})$ for some $\beta_{0,1}$, $c_{0,1}$ and monotonically increasing function $G^{(1)}$ with $G^{(1)}(0) = 0$, then $\widehat{M}_1^V$ and $\widehat{M}_1^C$ are consistent to $M_{0,1} = \mathrm{support}(\beta_{0,1})$ under proper choice of $\kappa_n^{(1)}$, provided that $(\hat{c}_{M_{0,1}}, \hat{\beta}_{M_{0,1}}^T)^T$ are consistent to $(c_{0,1}, \beta_{0,1}^T)^T$. Otherwise, $\mathrm{VIC}^{(1)}$ and $\mathrm{VIC}^{(1)}$ may select different models. More specifically, let

$$\begin{aligned}\theta_1^V &= \{c_1^V, (\beta_1^V)^T\}^T = \arg\max_{\theta_1=(c_1,\beta_1^T)^T} \mathrm{E}\left\{\tau^{(1)}(X_0^{(1)}) I(\beta_1^T X_0^{(1)} + c_1 > 0)\right\},\\ \beta_1^C &= \arg\max_{\beta_1} \mathrm{E}\left\{\tau^{(1)}(X_i^{(1)}) - \tau^{(1)}(X_j^{(1)})\right\} I(\beta_1^T X_i^{(1)} > \beta_1^T X_j^{(1)}), \quad i \neq j.\end{aligned}$$

Define M_1^V and M_1^C to be the support of θ_1^V and β_1^C, respectively. In the following, we show that $\widehat{M}_1^V$ and $\widehat{M}_1^C$ are consistent to M_1^V and M_1^C.

Let $\pi^{(2)}(\bar{x}^{(2)}) = \Pr(A_0^{(2)} = 1 | \overline{X}_0^{(2)} = \bar{x}^{(2)})$ and $\pi^{(1)}(x^{(1)}) = \Pr(A_0^{(1)} = 1 | X_0^{(1)} = x^{(1)})$. Let $V(\theta_1, \theta_2)$ be the average potential outcome of patients following the regime $d_1(x^{(1)}) = I(\theta_1^T x^{(1)} > 0)$, $d_2(\bar{x}^{(2)}) = I(\theta_2^T \bar{x}^{(2)} > 0)$, and $V^{(1)}(\theta_1) = V(\theta_1, \theta_{0,2})$. Define

$$C^{(1)}(\beta_1) = \mathrm{E}\left\{\tau^{(1)}(X_i^{(1)}) - \tau^{(1)}(X_j^{(1)})\right\} I(\beta_1^T X_i^{(1)} > \beta_1^T X_j^{(1)}), \quad i \neq j.$$

Let $\delta_V = \min_{j\in M_1^V} |\theta_j|$ and $\delta_C = \min_{j\in M_1^C} |\beta_j|$. For any $\delta > 0$, define

$$N_\delta^V = \{\theta_1^T \in \mathbb{R}^{p_1+1} : \|\theta_1 - \theta_1^V\|_2 \le \delta\}, \quad N_\delta^C = \{\beta_1^T \in \mathbb{R}^{p_1} : \|\beta_1 - \beta_1^C\|_2 \le \delta\},$$

and

$$S^V = \{\theta_1^T \in \mathbb{R}^{p_1+1} : \|\theta_1\|_2 = \|\theta_1^V\|_2\}, \quad S^C = \{\beta_1^T \in \mathbb{R}^{p_1} : \|\beta_1\|_2 = \|\beta_1^C\|_2\}.$$

For any $o = (x^{(1)}, a^{(1)}, x^{(2)}, a^{(2)}, y)$,

$$\begin{aligned}g(o, \beta_1) &= \frac{1}{2}\mathrm{E}\left\{\frac{(A_0^{(1)} - \pi_0^{(1)})Y_0^{(1)}}{\pi_0^{(1)}(1-\pi_0^{(1)})} - \frac{(a^{(1)} - \pi^{(1)}(x^{(1)}))y^{(1)}}{\pi^{(1)}(x^{(1)})(1-\pi^{(1)}(x^{(1)}))}\right\} I(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)})\\ &+ \frac{1}{2}\mathrm{E}\left\{\frac{(a^{(1)} - \pi^{(1)}(x^{(1)}))y^{(1)}}{\pi^{(1)}(x^{(1)})(1-\pi^{(1)}(x^{(1)}))} - \frac{(A_0^{(1)} - \pi_0^{(1)})Y_0^{(1)}}{\pi_0^{(1)}(1-\pi_0^{(1)})}\right\} I(\beta_1^T x^{(1)} > \beta_1^T X_0^{(1)}),\end{aligned}$$

where

$$y_1^{(1)} = \left\{ \frac{a^{(2)}}{\pi^{(2)}(\bar{x}^{(2)})} d_2^{opt}(\bar{x}^{(2)}) + \frac{1-a^{(2)}}{1-\pi^{(2)}(\bar{x}^{(2)})} \{1 - d_2^{opt}(\bar{x}^{(2)})\} \right\} y,$$

$$Y_0^{(1)} = \left\{ \frac{A_0^{(2)}}{\pi_0^{(2)}} d_2^{opt}(\overline{X}_0^{(2)}) + \frac{1-A_0^{(2)}}{1-\pi_0^{(2)}} \{1 - d_2^{opt}(\overline{X}_0^{(2)})\} \right\} Y_0,$$

and $\pi_0^{(1)} = \pi^{(1)}(X_0^{(1)})$, $\pi_0^{(2)} = \pi^{(2)}(\overline{X}_0^{(2)})$. Let

$$\phi_1(x^{(1)}, \beta_1) = \Pr(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)}), \quad \phi_2(x^{(1)}, \beta_1) = \Pr(\beta_1^T X_0^{(1)} < \beta_1^T x^{(1)}).$$

We first introduce some conditions.

(C4.) There exist some constants c_1, c_2 such that $0 < c_1 \le \pi^{(1)}(x^{(1)}), \pi^{(2)}(\bar{x}^{(2)}) \le c_2 < 1$ for any $x^{(1)}$ and $\bar{x}^{(2)}$. Besides, $\sup_{x^{(1)}} \mathrm{E}(Y_0^2 | X_0^{(1)} = x^{(1)}) < \infty$.
(C5.) With probability tending to 1, $\|\hat{\theta}_{M_{0,2}} - \theta_{0,2}\|_2 = O(R_{n,2})$ for some $n^{-1/2} \le R_{n,2} \to 0$. Besides, assume

$$\mathrm{E}\left(\sup_{\substack{\|\theta_2 - \theta_{0,2}\|_2 \le \varepsilon \\ \theta_2 = (c_2, \beta_2^T)^T}} |I(\beta_2^T X_0^{(2)} > -c_2) - I(\beta_{0,2}^T X_0^{(2)} > -c_{0,2})| \right) = O(\varepsilon),$$

as $\varepsilon \to 0$.
(C6.) With probability tending to 1, $\|\hat{\theta}_{M_1^V} - \theta_{0,1}\|_2 = O(R_{n,1})$ for some $n^{-1/2} \le R_{n,1} \to 0$.
(C7.) (i)Assume $V(\theta_1, \theta_2)$ is twice continuously differentiable in a small neighborhood around $(\theta_1^V, \theta_{0,2})$. (ii)Assume $V^{(1)}(\theta_1^V) > V^{(1)}(0)$ and $V^{(1)}(\theta_1^V) > \sup_{\theta_1 \in (N_{\varepsilon_0}^V)^c \cap S^V} V^{(1)}(\theta_1)$ for some $0 < \varepsilon_0 \le \delta^V$. (iii)Assume there exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that for any $\theta_1 \in N_{\varepsilon_0}^V \cap S^V$,

$$\bar{c}_1 \|\theta_1 - \theta_1^V\|_2^2 \le V^{(1)}(\theta_1^V) - V^{(1)}(\theta_1) \le \bar{c}_2 \|\theta_2 - \theta_2^V\|_2^2.$$

(iv)Assume

$$\mathrm{E}\left(\sup_{\substack{\|\theta_1 - \theta_1^V\|_2 \le \varepsilon \\ \theta_1 = (c_1, \beta_1^T)^T}} |I(\beta_1^T X_0^{(1)} > -c_1) - I((\beta_1^V)^T X_0^{(1)} > -c_1^V)| \right) = O(\varepsilon),$$

as $\varepsilon \to 0$.
(C8.) With probability tending to 1, $\|\hat{\beta}_{M_1^C} - \beta_{0,1}\|_2 = O(R_{n,1}^{(1)})$ for some

$n^{-1/2} \leq R_{n,1}^{(1)} \to 0$.

(C9.) (i)Assume $\mathrm{E}|Q(\beta_2^T X_0^{(2)} + c_2 - c_{0,2}) - Q(\beta_{0,2}^T X_0^{(2)})|^2 = O(\|\beta_2 - \beta_{0,2}\|_2^2) + O(|c_2 - c_{0,2}|_2^2)$ for any $\theta_2 = (c_2, \beta_2^T)^T$ in a small neighborhood of $\theta_{2,0}$. (ii) $C^{(1)}(\beta_{0,1}) > C^{(1)}(0)$ and $C^{(1)}(\beta_{0,1}) > \sup_{\beta_1 \in (N_{\varepsilon_0}^C)^c \cap S^C} C^{(1)}(\beta_1)$ for some constants $0 < \varepsilon_0 \leq \delta^C$. (iii) There exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|\beta_{0,1} - \beta_1\|_2^2 \leq C(\beta_{0,1}) - C(\beta_1) \leq \bar{c}_2 \|\beta_{0,1} - \beta_1\|_2^2, \quad \forall \beta_1 \in N_{\varepsilon_0}^C \cap S^C.$$

(iv) There exist some ψ_1, ψ_2 such that $\mathrm{E}\psi_1^2(X_0^{(1)}), \mathrm{E}\psi_2^2(X_0^{(1)}) < \infty$ and $|\phi_j(X_0^{(1)}, \beta_1) - \phi_j(X_0^{(1)}, \beta_1^C)| \leq \psi_j(X_0^{(1)}) \|\beta_1 - \beta_1^C\|_2$ for all $\beta_1 \in N_{\varepsilon_0}^C$, $j = 1, 2$.
(v) Function $g(o, \beta_1)$ is twice continuously differentiable for all $\beta_1 \in N_{\varepsilon_0}^C$.
(vi) There is an integrable function $K(o)$ such that for all o and $\beta_1 \in N_{\varepsilon_0}^C$,

$$\|\Delta_2 g(o, \beta_1) - \Delta_2 g(o, \beta_{0,1})\|_2 \leq K(o) \|\beta_1 - \beta_{0,1}\|_2.$$

(vii) $\mathrm{E}|\partial_i g(O_0, \beta_{0,1})|^2 < \infty$, $\mathrm{E}|\partial_{ij} g(O_0, \beta_{0,1})| < \infty$.

THEOREM 10.1. *Assume (C1)-(C7) hold. If* $\kappa_n^{(1)} = o(n)$, *and*

$$\kappa_n^{(1)} \gg \max(nR_{n,1}^2, \sqrt{nR_{n,1}}, nR_{n,2}^2, \sqrt{nR_{n,2}}, n^{-1/3}).$$

Then conditional on the event $\widehat{M}_2^V = M_2^V$, *we have*

$$Pr\left(\widehat{M}_V^1 = M_V^1\right) \to 1.$$

Assume (C1)-(C5), (C8)-(C9) hold. If $\kappa_n^{(1)} = o(n)$, *and*

$$\kappa_n^{(1)} \gg \max(n(R_{n,1}^{(1)})^2, \sqrt{nR_{n,2}}, nR_{n,2}^2),$$

Then conditional on the event $\widehat{M}_2^C = M_2^C$, *we have*

$$Pr\left(\widehat{M}_C^1 = M_C^1\right) \to 1.$$

REMARK 10.1. *For a given model* M_1, *if we obtain* $\hat{\theta}_{M_1}$ *by maximizing* $\widehat{V}^{(1)}(\theta_1)$, *then Condition (C6) automatically holds. Similarly, Condition (C8) holds when we obtain* $\hat{\beta}_{M_1}$ *by maximizing* $\widehat{C}^{(1)}(\beta_1)$ *for any candidate model* M_1. *For simplicity, we assume the propensity scores are known for each patient. This assumption is satisfied for SMART studies. A doubly-robust version of VIC and CIC can be similarly constructed.*

REMARK 10.2. *Compared to Theorem 3.1, we can see that conditions on* $\kappa_n^{(1)}$ *are strengthened in the backward induction algorithm, due to the variability of* $\hat{\beta}_{\widehat{M}_2^V}$ *and* $\hat{\beta}_{\widehat{M}_2^C}$. *Take CIC as an example, when estimating* $\hat{\beta}_{M_2}$ *and* $\hat{c}_{M_2}$ *via CAL for any* M_2, *we have* $\hat{\beta}_{M_{0,2}} = \beta_{0,2} + O_p(n^{-1/2})$ *and* $\hat{c}_{M_{0,2}} = c_{0,2} + O_p(n^{-1/3})$. *Thus, CIC*$^{(2)}$ *is consistent when* $\kappa_n^{(2)} \to \infty$ *and* $\kappa_n^{(2)} = o(n)$ *while consistency of CIC*$^{(1)}$ *requires* $n^{1/3} \ll \kappa_n^{(1)} = o(n)$.

10.2. *Multi-stage study.* We generalize our information criteria to multi-stage studies. Assume treatment decisions are made at a finite number of time points $t_1, \ldots, t_K$. Data are summarized as

$$\left\{O_i = \left(X_i^{(1)}, A_i^{(1)}, \ldots, X_i^{(K)}, A_i^{(K)}, Y_i\right)\right\}_{i=1}^n,$$

where Y_i denotes the ith patient's response, $X_i^{(1)}$ stands for the baseline covariates, $X_i^{(j)}$ stands for the covariates collected between t_{j-1} and t_j for $2 \le j \le K$, and $A_i^{(j)}$ is the treatment received at time point t_j for $1 \le j \le K$.

Let $\overline{X}_i^{(k)} = (X_i^{(1)}, A_i^{(1)}, \ldots, X_i^{(k)})$ for $1 \le k \le K$. Similar to the two-stage study, we select models via backward induction. For any $\theta_K = (c_K, \beta_K^T)^T$, define

$$\mathrm{VIC}^{(K)}(\theta_K) = \widehat{V}^{(K)}(\theta_K) - \kappa_n^{(K)}\|\theta_K\|_0, \mathrm{CIC}^{(K)}(\theta_K) = \widehat{C}^{(K)}(\theta_K) - \kappa_n^{(K)}\|\theta_K\|_0,$$

where

$$\widehat{V}^{(K)}(\theta_K) = \frac{1}{n}\sum_{i=1}^n \left\{\frac{A_i^{(K)} I\{\beta_K^T \overline{X}_i^{(K)} > -c_K\}}{\pi_i^{(K)}} + \frac{(1-A_i^{(K)}) I\{\beta_K^T \overline{X}_i^{(K)} \le -c_K\}}{1-\pi_i^{(K)}}\right\} Y_i,$$

$$\widehat{C}^{(K)}(\beta_K) = \frac{1}{n(n-1)}\sum_{i\neq j}\left\{\omega_i^{(K)} - \omega_j^{(K)}\right\} I(\beta_K^T \overline{X}_i^{(K)} > \beta_K^T \overline{X}_j^{(K)}),$$

$$\omega_i^{(K)} = \left(\frac{A_i^{(K)}}{\pi_i^{(K)}} - \frac{1-A_i^{(K)}}{1-\pi_i^{(K)}}\right) Y_i, \quad \pi_i^{(K)} = \Pr\left(A_i^{(K)} = 1 \mid \overline{X}_i^{(K)}\right).$$

Based on $\mathrm{VIC}^{(K)}$ and $\mathrm{CIC}^{(K)}$, we can select models for the contrast function on the last stage, which we denoted by $\widehat{M}_V^{(K)}$ and $\widehat{M}_C^{(K)}$, accordingly.

Assume for now, we have $\widehat{M}_V^{(j)}$ and $\widehat{M}_C^{(j)}$ for $k+1 \le j \le K$. To obtain $\widehat{M}_V^{(k)}$ and $\widehat{M}_C^{(k)}$, we iteratively define the pseudo response $Y_i^{(k),V}$ and $Y_i^{(k),C}$

by

$$
\begin{aligned}
&Y_i^{(K),V} = Y_i^{(K),C} = Y_i,\\
&Y_i^{(j),V} = \left\{ \frac{A_i^{(j+1)} \hat{d}_{j+1}^V(\overline{X}_i^{(j+1)})}{\pi_i^{(j+1)}} + \frac{(1-A_i^{(j+1)}) I \hat{d}_{j+1}^V(\overline{X}_i^{(j+1)})}{1-\pi_i^{(j+1)}} \right\} Y_i^{(j+1),V},\\
&Y_i^{(j),C} = \left\{ \frac{A_i^{(j+1)} \hat{d}_{j+1}^C(\overline{X}_i^{(j+1)})}{\pi_i^{(j+1)}} + \frac{(1-A_i^{(j+1)}) I \hat{d}_{j+1}^C(\overline{X}_i^{(j+1)})}{1-\pi_i^{(j+1)}} \right\} Y_i^{(j+1),C},
\end{aligned}
$$

for $j = K-1, K-2, \ldots, k$, where

$$
\begin{aligned}
\hat{d}_{j+1}^V(\bar{x}^{(j+1)}) &= I(\hat{\beta}_{\widehat{M}_V^{(j+1)}}^T \bar{x}^{(j+1)} > -\hat{c}_{\widehat{M}_V^{(j+1)}}),\\
\hat{d}_{j+1}^C(\bar{x}^{(j+1)}) &= I(\hat{\beta}_{\widehat{M}_C^{(j+1)}}^T \bar{x}^{(j+1)} > -\hat{c}_{\widehat{M}_C^{(j+1)}}).
\end{aligned}
$$

Our information criteria on the kth stage is defined by

$$
\mathrm{VIC}^{(k)}(\theta_k) = \widehat{V}^{(k)}(\theta_k) - \kappa_n^{(k)} \|\theta_k\|_0, \mathrm{CIC}^{(k)}(\theta_k) = \widehat{C}^{(k)}(\theta_k) - \kappa_n^{(k)} \|\theta_k\|_0,
$$

where

$$
\begin{aligned}
&\widehat{V}^{(k)}(\theta_k) = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{A_i^{(k)} I\{\beta_k^T \overline{X}_i^{(k)} > -c_k\}}{\pi_i^{(k)}} + \frac{(1-A_i^{(k)}) I\{\beta_k^T \overline{X}_i^{(k)} \le -c_k\}}{1-\pi_i^{(k)}} \right\} Y_i^{(k),V},\\
&\widehat{C}^{(k)}(\beta_k) = \frac{1}{n(n-1)} \sum_{i \neq j} \left\{ \omega_i^{(k)} - \omega_j^{(k)} \right\} I(\beta_k^T \overline{X}_i^{(k)} > \beta_k^T \overline{X}_j^{(k)}),\\
&\omega_i^{(k)} = \left(\frac{A_i^{(k)}}{\pi_i^{(k)}} - \frac{1-A_i^{(k)}}{1-\pi_i^{(k)}} \right) Y_i^{(k),C}, \quad \pi_i^{(k)} = \Pr\left(A_i^{(k)} = 1 \mid \overline{X}_i^{(k)} \right).
\end{aligned}
$$

Our information criteria are consistent when the monotonic linear index assumption holds for the contrast function on each stage.

APPENDIX A: TECHNICAL CONDITIONS

A.1. Assumptions for Theorem 4.2. In this section, we present some technical conditions for Theorem 4.2. We use a shorthand and write $\pi_\alpha(x) = \pi(x, \alpha)$ and $h_\eta(x) = h(x, \eta)$. Define function $g(o, \beta, \alpha, \eta)$ as

$$
\begin{aligned}
&\frac{1}{2}\mathrm{E}\left\{ \frac{\{A_0 - \pi_\alpha(X_0)\}\{Y_0 - h_\eta(X_0)\}a}{\pi_\alpha(X_0)\{1-\pi_\alpha(X_0)\}\pi_\alpha(x)} - \frac{\{a - \pi_\alpha(x)\}\{y - h_\eta(x)\}A_0}{\pi_\alpha(x)\{1-\pi_\alpha(x)\}\pi_\alpha(X_0)} \right\} I(X_0^T\beta > x^T\beta)\\
+&\frac{1}{2}\mathrm{E}\left\{ \frac{\{a - \pi_\alpha(x)\}\{y - h_\eta(x)\}A_0}{\pi_\alpha(x)\{1-\pi_\alpha(x)\}\pi_\alpha(X_0)} - \frac{\{A_0 - \pi_\alpha(X_0)\}\{Y_0 - h_\eta(X_0)\}a}{\pi_\alpha(X_0)\{1-\pi_\alpha(X_0)\}\pi_\alpha(x)} \right\} I(x^T\beta > X_0^T\beta).
\end{aligned}
$$

For $\theta = (c, \beta^T)^T$, let

$$\begin{aligned} V(\theta, \alpha, \eta) &= \mathrm{E}\left\{ \frac{A_0 I(X_0^T\beta > -c)}{\pi(X_0, \alpha)} + \frac{(1-A_0) I(X_0^T\beta \le -c)}{1-\pi(X_0, \alpha)} \right\} Y_0 \\ &- \mathrm{E}\left\{ \frac{A_0 I(X_0^T\beta > -c)}{\pi(X_0, \alpha)} + \frac{(1-A_0) I(X_0^T\beta \le -c)}{1-\pi(X_0, \alpha)} - 1 \right\} h(X_0, \eta). \end{aligned}$$

We write $g(o, \beta, \alpha, \eta)$ as $g(o, \zeta)$ where ζ represents the $(p+q_1+q_2)$-dimensional parameter $(\beta^T, \alpha^T, \eta^T)^T$. Similarly, we write $V(\theta, \alpha, \eta)$ as $V(\bar{\zeta})$ where $\bar{\zeta}$ represents the $(p+1+q_1+q_2)$-dimensional parameter $(\theta^T, \alpha^T, \eta^T)^T$. Let $\partial_j g(o, \zeta)$ and $\partial_{jk} g(o, \zeta)$ stand for the derivatives

$$\frac{\partial g(o, \zeta)}{\partial \zeta^j} \text{ and } \frac{\partial^2 g(o, \zeta)}{\partial \zeta^j \partial \zeta^k}.$$

Let $\Delta_2 g(o, \zeta)$ be the Hessian matrix of $g(o, \zeta)$. Define $\zeta_0 = (\beta_0^T, (\alpha^*)^T, (\eta^*)^T)^T$ and $\bar{\zeta}_0 = (c_0, \zeta_0^T)^T$.

(A3') Assume there exists some constants $0 < c_1, c_2 < 1$ such that $c_1 \le \pi(x, \alpha) \le c_2$ for all x and α in a small neighborhood of α^*.
(A4') Assume $\sup_x \mathrm{E}(Y_0^2 | X_0 = x) = O(1)$, $\sup_x \sup_{\eta: \|\eta - \eta^*\|_2 \le \epsilon} |h_\eta(x)| = O(1)$ for some sufficiently small $\epsilon > 0$.
(A10)(i) Assume $V^{DR}(\theta_0) > V^{DR}(0), V^{DR}(\theta_0) > \sup_{\theta \in \tilde{N}_{\varepsilon_0}^c \cap \tilde{S}(\theta_0)} V^{DR}(\theta) > 0$ for some constant $0 < \varepsilon_0 \le \delta$. (ii) Assume

$$\mathrm{E}\left(\sup_{\substack{\|\theta - \theta_0\|_2 \le \varepsilon \\ \theta = (c, \beta^T)^T}} \left| I(X_0^T\beta > -c) - I(X_0^T\beta_0 > -c_0) \right| \right) = O(\varepsilon),$$

as $\varepsilon \to 0$. (iii) There exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|\theta_0 - \theta\|_2^2 \le V^{DR}(\theta_0) - V^{DR}(\theta) \le \bar{c}_2 \|\theta_0 - \theta\|_2^2, \quad \forall \theta \in \tilde{N}_{\varepsilon_0} \cap \tilde{S}(\theta_0).$$

(iv) Assume V is uniformly continuous. Besides, for any $\bar{\zeta}$ in a small neighborhood of $\bar{\zeta}_0$,

$$V(\bar{\zeta}) = V(\bar{\zeta}_0) + \frac{\partial V(\bar{\zeta}_0)}{\partial \bar{\zeta}} (\bar{\zeta} - \bar{\zeta}_0) + \frac{1}{2} (\bar{\zeta} - \bar{\zeta}_0)^T \Delta_2 V(\bar{\zeta}_0) (\bar{\zeta} - \bar{\zeta}_0) + o(1) \|\bar{\zeta} - \bar{\zeta}_0\|_2^2.$$

(A11)(i) Assume $C^{DR}(\beta_0) > C^{DR}(0)$, $C^{DR}(\beta_0) > \sup_{\beta \in N_{\varepsilon_0}^c \cap S(\beta_0)} C^{DR}(\beta)$ for some $0 < \varepsilon_0 \le \delta$. (ii) There exist some constants $\bar{c}_1, \bar{c}_2 > 0$ such that

$$\bar{c}_1 \|\beta_0 - \beta\|_2^2 \le C^{DR}(\beta_0) - C^{DR}(\beta) \le \bar{c}_2 \|\beta_0 - \beta\|_2^2, \quad \forall \beta \in N_{\varepsilon_0} \cap S(\beta_0).$$

(iii) Assume function $g(o,\zeta)$ is twice continuously differentiable for all ζ in a small neighborhood of ζ_0. (iv) Assume there's an integrable function $K(o)$ such that for all o and ζ in a small neighborhood of ζ_0,

$$\|\Delta_2 g(o,\zeta) - \Delta_2 g(o,\zeta_0)\|_2 \le K(o)\|\zeta - \zeta_0\|_2.$$

(v) $\max_j \mathrm{E}|\partial_j g(O_0,\zeta_0)| < \infty$, $\max_{ij} \mathrm{E}|\partial_{ij} g(O_0,\zeta_0)| < \infty$.

A.2. Discussion of (A6)(ii) and (A6')(ii). Condition (A6)(ii) and (A6')(ii) are related to the margin assumption (Qian and Murphy, 2011; Luedtke and van der Laan, 2016) which requires $\Pr(0 < |Q(X_0^T\beta_0)| < \varepsilon) = O(\varepsilon^\alpha)$ for some $\alpha > 0$, as $\varepsilon \to 0$. Notice that under (A6)(ii) or (A6')(ii), we have

$$\mathrm{E}\left(\sup_{|c-c_0|\le\varepsilon} |I(X_0^T\beta_0 > -c) - I(X_0^T\beta_0 > -c_0)|\right) = O(\varepsilon), \quad \text{as } \varepsilon \to 0.$$

It follows that

$$\mathrm{E}\sup_{t\le\varepsilon} I(0 < |X_0^T\beta_0 + c_0| \le t) = O(\varepsilon), \quad \text{as } \varepsilon \to 0,$$

and hence

$$\Pr(0 < |X_0^T\beta_0 + c_0| \le \varepsilon) = O(\varepsilon), \quad \text{as } \varepsilon \to 0.$$

Notice that the function $Q(\cdot)$ satisfies $Q(-c_0) = 0$. Assume

$$\left.\frac{dQ(x-c_0)}{dx}\right|_{x=0} \neq 0.$$

Then for sufficiently small $\varepsilon > 0$, the event $0 < |Q(X_0^T\beta_0)| \le \varepsilon$ is contained in the event $0 < |X_0^T\beta_0 + c_0| \le c_*\varepsilon$ for some constant $c_* > 0$. As a result, we have

$$\Pr(0 < |Q(X_0^T\beta_0)| \le \varepsilon) \le \Pr(0 < |X_0^T\beta_0 + c_0| \le c_*\varepsilon) = O(\varepsilon), \quad \text{as } \varepsilon \to 0.$$

Therefore, the margin assumption holds with $\alpha = 1$.

A.3. Discussion of (A7)(iii) and (A7')(iii). In (A7)(iii) and (A7')(iii), we require $g(o,\beta)$ to be twice continuously differentiable for any $\beta \in N_{\varepsilon_0}$. This condition is likely to be violated under treatment effect homogeneity, i.e, $\beta_0 = 0$, $c_0 = 0$, or under the nonregular scenarios where $\Pr(X_0^T\beta_0) > 0$. In this paper, we assume $\beta_0 \neq 0$. Without loss of generality, assume $\beta_0^1 > 0$.

Since $\varepsilon_0 < \beta_0^1$, we have $\beta^1 > 0$, $\forall \beta \in N_{\varepsilon_0}$. When $\tau(x) = Q(x^T\beta_0)$, we have by (A1) and (A2) that

$$\begin{aligned} g(o,\beta) &= \frac{1}{2}\mathrm{E}Q(X_0^T\beta_0)\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\} \\ &- \frac{\{a-\pi(x)\}y}{2\pi(x)\{1-\pi(x)\}}\{\Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta)\}. \end{aligned} \tag{A.1}$$

For any $\beta \in \mathbb{R}^p$, let $\beta^{(-1)}$ denote the last $p-1$ components of β. Let $F_{X_0^{(-1)}}(\cdot)$ denote the cumulative distribution function of the last $p-1$ components of X_0. Let $f_{X_0^1}(\cdot|z^{(-1)})$ denote the conditional density function of X_0^1 given $X_0^{(-1)} = z^{(-1)}$. With some calculations, we can show

$$\begin{aligned} g(o,\beta) = \int_{\mathbb{R}^{p-1}} \int_{\frac{\beta^T x - \beta^{(-1)T} z^{(-1)}}{\beta^1}}^{+\infty} \left(Q(z^1\beta_0^1 + z^{(-1)}\beta_0^{(-1)}) - \frac{\{a-\pi(x)\}y}{2\pi(x)\{1-\pi(x)\}} \right) \\ \times f_{X_0^1}(z^1|z^{(-1)}) dF_{X_0^{(-1)}}(z^{(-1)}) - \frac{1}{2}\left(\mathrm{E}Q(X_0^T\beta_0) - \frac{\{a-\pi(x)\}y}{\pi(x)\{1-\pi(x)\}} \right). \end{aligned}$$

Therefore, (A7)(iii) and (A7')(iii) hold under certain regularity conditions on $f_{X_0^1}$, $F_{X_0^{(-1)}}$ and their derivatives.

In Section A.3.1, we provide two examples assuming the covariates are jointly normal and show (A7)(iii) and (A7')(iii) hold under both examples. In Section A.3.2, we show both conditions are violated under the treatment effect homogeneity. In Section A.3.3, we provide two example and show both are violated under both examples.

A.3.1. *Gaussian covariates.* We assume $X_0 \sim N(0,\Sigma)$ for some positive definite covariance matrix Σ.

Example 1 (Linear contrast function) Assume $Q(z) = a_0 z + b_0$ for some $a_0, b_0 \in \mathbb{R}$. It follows from (A.1) that

$$\begin{aligned} g(o,\beta) &= \frac{1}{2}\mathrm{E}(a_0 X_0^T\beta_0 + b_0)\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\} \\ &- \frac{\{a-\pi(x)\}y}{2\pi(x)\{1-\pi(x)\}}\{\Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta)\}. \end{aligned}$$

To show (A7)(iii) and (A7')(iii) hold, it suffices to show functions

$$\begin{aligned} g_1(x,\beta) &= \mathrm{E}X_0^T\beta_0\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\}, \\ g_2(x,\beta) &= \Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta), \end{aligned}$$

are twice continuously differentiable for all $\beta \in N_{\varepsilon_0}$. Since $\beta_0 \neq 0$ and $\varepsilon_0 < \min_{j\in\mathcal{M}_\beta} |\beta_0^j|$, we have $\beta \neq 0$, for any $\beta \in N_{\varepsilon_0}$. The random variable $X_0^T\beta$ is normally distributed with variance $\beta^T\Sigma\beta > 0$. As a result,

$$g_2(x,\beta) = 1 - 2\Phi\left(\frac{x^T\beta}{\sqrt{\beta^T\Sigma\beta}}\right).$$

It is immediate to see that g_2 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$.

As for g_1, we have

$$\begin{aligned} g_1(x,\beta) &= \mathrm{E}\left(X_0^T\beta_0 - \frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta + \frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta\right)\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\} \\ &= \frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}\mathrm{E}X_0^T\beta\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\} = 2\frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}\mathrm{E}X_0^T\beta I(X_0^T\beta > x^T\beta) \\ &= \frac{\sqrt{2}\beta_0^T\Sigma\beta}{\sqrt{\pi\beta^T\Sigma\beta}}\int_{\frac{x^T\beta}{\sqrt{\beta^T\Sigma\beta}}}^{+\infty} z\exp(-z^2/2)dz = \frac{\sqrt{2}\beta_0^T\Sigma\beta}{\sqrt{\pi\beta^T\Sigma\beta}}\exp\left(-\frac{(x^T\beta)^2}{2\beta^T\Sigma\beta}\right), \end{aligned}$$

where the second equality is due to that $X_0^T\beta_0 - (\beta_0^T\Sigma\beta)(X_0^T\beta)/(\beta^T\Sigma\beta)$ is independent of $X_0^T\beta$. Hence, g_1 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$.

Example 2 (Nonlinear contrast function) Assume $Q(z) = \exp(a_0 z + b_0) - 1$ for some $a_0, b_0 \in \mathbb{R}$. It follows from (A.1) that

$$\begin{aligned} g(o,\beta) &= \frac{1}{2}\mathrm{E}\{\exp(a_0X_0^T\beta_0 + b_0) - 1\}\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\} \\ &- \frac{\{a - \pi(x)\}y}{2\pi(x)\{1-\pi(x)\}}\{\Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta)\}. \end{aligned}$$

To show (A7)(iii) and (A7')(iii) hold, it suffices to show functions

$$\begin{aligned} g_2(x,\beta) &= \Pr(X_0^T\beta > x^T\beta) - \Pr(X_0^T\beta < x^T\beta), \\ g_3(x,\beta) &= \mathrm{E}\exp(a_0X_0^T\beta_0)\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\}, \end{aligned}$$

are twice continuously differentiable for all $\beta \in N_{\varepsilon_0}$. We've shown in Example 1 that g_2 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$. As for g_3, since $X_0^T\beta_0 - (\beta_0^T\Sigma\beta)(X_0^T\beta)/(\beta^T\Sigma\beta)$ is independent of $X_0^T\beta$, we have

$$\begin{aligned} g_3(x,\beta) = \mathrm{E}\exp\left(a_0\frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta\right)\{I(X_0^T\beta > x^T\beta) - I(X_0^T\beta < x^T\beta)\} \\ \times\mathrm{E}\exp\left(a_0X_0^T\beta_0 - a_0\frac{\beta_0^T\Sigma\beta}{\beta^T\Sigma\beta}X_0^T\beta\right) \end{aligned}$$

With some calculations, we can show

$$g_3(x,\beta) = \exp\left(\frac{a_0^2\beta_0^T\Sigma\beta_0}{2}\right)\left\{1 - 2\Phi\left(\frac{x^T\beta - a_0\beta_0^T\Sigma\beta}{\sqrt{\beta^T\Sigma\beta}}\right)\right\}.$$

It is immediate to see that g_3 is twice continuously differentiable with respect to $\beta \in N_{\varepsilon_0}$.

A.3.2. *Treatment effect homogeneity.* Under the treatment effect homogeneity, we have $\beta_0 = 0$, $c_0 = 0$, and hence $Q(x^T\beta_0) = 0, \forall x$. It follows from (A.1) that

$$g(o,\beta) = \frac{\{a - \pi(x)\}y}{2\pi(x)(1-\pi(x))}\{\Pr(X_0^T\beta < x^T\beta) - \Pr(X_0^T\beta > x^T\beta)\}.$$

Let $\beta^{(i)}(\gamma) = (\underbrace{0,\ldots,0}_{i-1},\gamma,\underbrace{0,\ldots,0}_{p-i})^T$, we have

$$\lim_{\gamma\to 0^+} g(o,\beta^{(i)}(\gamma)) = \frac{\{a - \pi(x)\}y}{2\pi(x)(1-\pi(x))}\{\Pr(X_0^{(i)} < x^i) - \Pr(X_0^{(i)} > x^i)\},$$

and

$$\lim_{\gamma\to 0^-} g(o,\beta^{(i)}(\gamma)) = \frac{\{a - \pi(x)\}y}{2\pi(x)(1-\pi(x))}\{\Pr(X_0^{(i)} > x^i) - \Pr(X_0^{(i)} < x^i)\}.$$

As a result, $g(o,\beta)$ is even not continuous when $\Pr(X_0^{(i)} < x^i) \neq \Pr(X_0^{(i)} > x^i)$ for some $i \in \{1,\ldots,p\}$. Hence, $g(o,\beta)$ is not twice differentiable. Condition (A7)(iii) and (A7')(iii) are thus violated.

A.3.3. *Nonregular cases.* Below, we provide two nonregular examples where $\Pr(X_0^T\beta_0 = 0) > 0$ and show (A7)(iii), (A7')(iii) are violated under both examples.

Example 3 We assume the p-dimensional covariates X_0 consist of p independent Rademacher random variables. In addition, the contrast function takes the form $\tau(X_0) = X_0^1 - X_0^2$. It is immediate to see that

$$\Pr\{\tau(X_0) = 0\} = \Pr(X_0^1 = X_0^2 = 1) + \Pr(X_0^1 = X_0^2 = -1) = 0.5 > 0.$$

We show $g(o,\beta)$ is not continuous at $\beta_0 = (1,-1,0,0,\ldots,0)^T$ and $x = (1,1,0,0,\ldots,0)$. Condition (A7)(iii) and (A7')(iii) are thus violated. Let

$\beta_0^{(1)}(\gamma) = \beta_0 + (\gamma, 0, 0, 0, \ldots, 0)^T$, we have $x^T\beta_0^{(1)}(\gamma) = \gamma$. It follows that

$$g(o, \beta_0^{(1)}(\gamma)) = \frac{1}{2}\mathrm{E}(X_0^1 - X_0^2)[I\{(1+\gamma)X_0^1 - X_0^2 > \gamma\} - I\{(1+\gamma)X_0^1 - X_0^2 < \gamma\}]$$
$$-\frac{(a-\pi(x))y}{2\pi(x)(1-\pi(x))}[\Pr\{(1+\gamma)X_0^1 - X_0^2 > \gamma\} - \Pr\{(1+\gamma)X_0^1 - X_0^2 < \gamma\}].$$

In the following, we show

$$\lim_{\gamma\to 0}\mathrm{E}(X_0^1 - X_0^2)[I\{(1+\gamma)X_0^1 - X_0^2 > \gamma\} - I\{(1+\gamma)X_0^1 - X_0^2 < \gamma\}]$$
$$= \mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 > X_0^2) - I(X_0^1 < X_0^2)\}, \tag{A.2}$$

and

$$\lim_{\gamma\to 0^+}\left(\Pr\{(1+\gamma)X_0^1 - X_0^2 > \gamma\} - \Pr\{(1+\gamma)X_0^1 - X_0^2 < \gamma\}\right) \tag{A.3}$$
$$\neq \lim_{\gamma\to 0^-}\left(\Pr\{(1+\gamma)X_0^1 - X_0^2 > \gamma\} - \Pr\{(1+\gamma)X_0^1 - X_0^2 < \gamma\}\right).$$

Function $g(o, \beta)$ is thus not continuous at $\beta = \beta_0$ when $\{a - \pi(x)\}y \neq 0$.

With some calculations, we have for sufficiently small $\gamma > 0$,

$$\mathrm{E}(X_0^1 - X_0^2)[I\{(1+\gamma)X_0^1 - X_0^2 > \gamma\} - I\{(1+\gamma)X_0^1 - X_0^2 < \gamma\}]$$
$$= 2\Pr(X_0^1 = 1, X_0^2 = -1) + 2\Pr(X_0^1 = -1, X_0^2 = 1) = 1,$$
$$\mathrm{E}(X_0^1 - X_0^2)[I\{(1-\gamma)X_0^1 - X_0^2 > -\gamma\} - I\{(1-\gamma)X_0^1 - X_0^2 < -\gamma\}]$$
$$= 2\Pr(X_0^1 = 1, X_0^2 = -1) + 2\Pr(X_0^1 = -1, X_0^2 = 1) = 1,$$

and

$$\mathrm{E}(X_0^1 - X_0^2)[I\{X_0^1 - X_0^2 > 0\} - I\{X_0^1 - X_0^2 < 0\}]$$
$$= 2\Pr(X_0^1 = 1, X_0^2 = -1) + 2\Pr(X_0^1 = -1, X_0^2 = 1) = 1.$$

Equation (A.2) is thus satisfied.

As for (A.3), we have for sufficiently small $\gamma > 0$,

$$\begin{aligned}
&\Pr\{(1+\gamma)X_0^1 - X_0^2 > \gamma\} - \Pr\{(1+\gamma)X_0^1 - X_0^2 < \gamma\} \\
= \;& \Pr(X_0^1 = 1, X_0^2 = -1) - \Pr(X_0^1 = -1) = -\frac{1}{4}, \\
&\Pr\{(1-\gamma)X_0^1 - X_0^2 > -\gamma\} - \Pr\{(1-\gamma)X_0^1 - X_0^2 < -\gamma\} \\
= \;& \Pr(X_0^2 = -1) - \Pr(X_0^1 = -1, X_0^2 = 1) = \frac{1}{4}.
\end{aligned}$$

This shows (A.3).

Example 4 We assume the p-dimensional covariates X_0 consist of p independent truncated Gaussian random variables. Specifically, for any $j \in \{1, \ldots, p\}$,

$$\Pr(X_0^{(j)} = 0) = \frac{1}{2} \text{ and } \Pr(X_0^{(j)} \leq z) = \frac{1}{2} + \frac{1}{\sqrt{2\pi}} \int_0^z \exp\left(-\frac{z^2}{2}\right) dz, \quad \forall z > 0.$$

In addition, the contrast function takes the form $\tau(X_0) = X_0^1 - X_0^2$. It is immediate to see that

$$\Pr\{\tau(X_0) = 0\} = \Pr(X_0^1 = X_0^2) = \frac{1}{4}.$$

Let $\beta_0^{(3)}(\gamma) = \beta_0 + (0, 0, \gamma, 0, 0, \ldots, 0)^T$ and $x^1 = x^2 = x^3 = 0$. It follows that $x^T \beta_0^{(3)}(\gamma) = 0$. Notice that

$$\Pr(X_0^T \beta_0^{(3)}(\gamma) = 0) \geq \Pr\left(\bigcap_{j=1}^{3} \{X_0^j = 0\}\right) = \frac{1}{8}.$$

Similar to Example 1, we show

$$\begin{aligned} &\lim_{\gamma \to 0} \mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - I(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\} \\ &\qquad = \mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 - X_0^2 > 0) - I(X_0^1 - X_0^2 < 0)\}, \end{aligned} \tag{A.4}$$

and

$$\begin{aligned} &\lim_{\gamma \to 0^+} \left(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\right) \\ \neq &\lim_{\gamma \to 0^-} \left(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\right). \end{aligned} \tag{A.5}$$

Condition (A7)(iii) and (A7')(iii) are thus violated.

With some calculations, we have for sufficiently small $\gamma > 0$,

$$\begin{aligned} &\mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - I(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\} \\ = &\frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) dx - \frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) \left\{1 - 2\Phi\left(\frac{x}{\gamma}\right)\right\} dx \\ - &\frac{1}{2\pi} \int_0^{+\infty} \int_0^{+\infty} (x - y) \exp\left(-\frac{x^2 + y^2}{2}\right) \left\{1 - 2\Phi\left(\frac{y - x}{\gamma}\right)\right\} dxdy, \\ &\mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 - X_0^2 - \gamma X_0^3 > 0) - I(X_0^1 - X_0^2 - \gamma X_0^3 < 0)\} \\ = &\frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) \left\{2\Phi\left(\frac{x}{\gamma}\right) - 1\right\} dx - \frac{1}{2\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) dx \\ - &\frac{1}{2\pi} \int_0^{+\infty} \int_0^{+\infty} (x - y) \exp\left(-\frac{x^2 + y^2}{2}\right) \left\{1 - 2\Phi\left(\frac{x - y}{\gamma}\right)\right\} dxdy, \end{aligned}$$

and

$$\mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 - X_0^2 > 0) - I(X_0^1 - X_0^2 < 0)\} = \frac{1}{\sqrt{2\pi}} \int_0^{+\infty} x \exp\left(-\frac{x^2}{2}\right) dx$$
$$-\frac{1}{2\pi} \int_0^{+\infty} (x-y) \exp\left(-\frac{x^2+y^2}{2}\right) \{I(x > y) - I(x < y)\} dxdy.$$

It follows from the dominated convergence theorem that

$$\lim_{\gamma \to 0} \mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - I(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\}$$
$$= \mathrm{E}(X_0^1 - X_0^2)\{I(X_0^1 - X_0^2 > 0) - I(X_0^1 - X_0^2 < 0)\}.$$

This proves (A.4).

Similarly, we can show

$$\lim_{\gamma \to 0^+} \big(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\big)$$
$$= \Pr(X_0^1 = X_0^2 = 0, X_0^3 > 0) + \Pr(X_0^1 > X_0^2) - \Pr(X_0^1 < X_0^2) = \frac{1}{8},$$

and

$$\lim_{\gamma \to 0^-} \big(\Pr(X_0^1 - X_0^2 + \gamma X_0^3 > 0) - \Pr(X_0^1 - X_0^2 + \gamma X_0^3 < 0)\big)$$
$$= \Pr(X_0^1 > X_0^2) - \Pr(X_0^1 < X_0^2) - \Pr(X_0^1 = X_0^2 = 0, X_0^3 > 0) = -\frac{1}{8}.$$

This proves (A.5).

APPENDIX B: PROOF OF THEOREM 3.4

Similar to the proof of Theorem 3.3, in the following, we provide tail inequalities for

$$\Pr\left(\mathrm{CIC}(\hat{\beta}_{M(\lambda_0)}) \le \sup_{\lambda \in \Omega_-} \mathrm{CIC}(\hat{\beta}_{M(\lambda)})\right), \tag{B.1}$$

and

$$\Pr\left(\mathrm{CIC}(\hat{\beta}_{M(\lambda_0)}) \le \sup_{\lambda \in \Omega_+} \{n\widehat{C}(\tilde{\beta}_{M(\lambda)}) - \kappa_n \|\hat{\beta}_{M(\lambda)}\|_0\}\right). \tag{B.2}$$

B.1. Underfitted model space. Under Assumption (A7')(i) and (iii), using similar arguments in Section 9.0.1, we can show

$$C(\hat{\beta}_{M(\lambda_0)}) - \sup_{\lambda \in \Omega_-} C(\hat{\beta}_{M(\lambda)}) > 5\xi, \tag{B.3}$$

for some constant $\xi > 0$. This together with

$$\kappa_n(\|\hat{\beta}_{M(\lambda_0)}\|_0 - \inf_{\lambda \in \Omega_-} \|\hat{\beta}_{M(\lambda)}\|_0) \leq O(1)\kappa_n,$$

and the condition $\kappa_n = o(n)$ suggests that for sufficiently large n, we have

$$C(\hat{\beta}_{M(\lambda_0)}) - \kappa_n\|\hat{\beta}_{M(\lambda_0)}\|_0 - \sup_{\lambda \in \Omega_-} \left\{C(\hat{\beta}_{M(\lambda)}) - \kappa_n\|\hat{\beta}_{M(\lambda)}\|_0\right\} \geq 4\xi.$$

Therefore, the event defined in (B.1) happens if

$$\sup_{\lambda \in \Omega_-} \left|\left\{\widehat{C}(\hat{\beta}_{M(\lambda)}) - C(\hat{\beta}_{M(\lambda)}) - \widehat{C}(\hat{\beta}_{M(\lambda_0)}) + C(\hat{\beta}_{M(\lambda_0)})\right\}\right| \geq 4\xi,$$

or

$$\sup_{\|\beta\|_0 \leq s_n} \left|\widehat{C}(\beta) - C(\beta)\right| \geq 2\xi.$$

This means (B.1) is smaller than

$$\Pr\left(\sup_{\|\beta\|_0 \leq s_n} \left|\widehat{C}(\beta) - C(\beta)\right| \geq 2\xi\right). \tag{B.4}$$

Define

$$\begin{aligned} h(O_i, O_j, \beta) &= \frac{1}{2}\left(\frac{(A_i - \pi_i)Y_i}{\pi_i(1-\pi_i)} - \frac{(A_j - \pi_j)Y_j}{\pi_j(1-\pi_j)}\right) I(X_i^T\beta > X_j^T\beta) \\ &+ \frac{1}{2}\left(\frac{(A_j - \pi_j)Y_j}{\pi_j(1-\pi_j)} - \frac{(A_i - \pi_i)Y_i}{\pi_i(1-\pi_i)}\right) I(X_j^T\beta > X_i^T\beta), \end{aligned} \tag{B.5}$$

and

$$D(O_i, O_j, \beta) = h(O_i, O_j, \beta) - g(O_i, \beta) - g(O_j, \beta) + C(\beta). \tag{B.6}$$

Function D is symmetric. Besides, it satisfies

$$\mathrm{E}D(O_i, o, \beta) = \mathrm{E}D(o, O_j, \beta) = 0.$$

The process $\sum_{i\neq j} D(O_i, O_j, \beta)$ is a degenerate U-process. Define the event

$$\bar{\mathcal{A}} = \left\{ \left| \frac{1}{n(n-1)} \sum_{i\neq j} D(O_i, O_j, \beta) \right| \leq \xi \right\}.$$

Since

$$m_C(\beta) = \widehat{C}(\beta) - C(\beta) = \frac{1}{n(n-1)} \sum_{i\neq j} h(O_i, O_j, \beta) - C(\beta)$$

$$= \left\{ \frac{1}{n(n-1)} \sum_{i\neq j} D(O_i, O_j, \beta) \right\} + \left\{ \frac{2}{n} \sum_i g(O_i, \beta) - 2C(\beta) \right\},$$

under the event defined in $\bar{\mathcal{A}}$, (B.4) can be bounded by

$$\Pr\left(\sup_{\|\beta\|_0 \leq s_n} \left| \frac{1}{n} \sum_i \{g(O_i, \beta) - C(\beta)\} \right| \geq \frac{\xi}{2} \right),$$

or

$$\sum_{M\in\Omega^*} \Pr\left(\sup_{\beta\in B_M} \left| \frac{1}{n} \sum_i \{g(O_i, \beta) - C(\beta)\} \right| \geq \frac{\xi}{2} \right). \tag{B.7}$$

Let $O_{n+1}, \ldots, O_{2n}$ as i.i.d copies of O_0, independent of $\{O_1, \ldots, O_n\}$. Observe that $g(O_i, \beta) = \mathrm{E}\{h(O_i, O_{n+i}, \beta)|O_i\}$. Here the expectation is taken with respect to O_{n+i}. It follows by Jensen's inequality that

$$\begin{aligned} &\sup_{M\in\Omega^*} \mathrm{E}\left(\sup_{\beta\in B_M} \left| \frac{1}{n} \sum_i \{g(O_i, \beta) - C(\beta)\} \right| \right) \\ \leq\ &\sup_{M\in\Omega^*} \mathrm{E}\left(\sup_{\beta\in B_M} \left| \frac{1}{n} \sum_i \{h(O_i, O_{n+i}, \beta) - C(\beta)\} \right| \right). \end{aligned} \tag{B.8}$$

Similar to (9.15), RHS of (B.8) is of the order $O(\sqrt{s_n/n})$. This together with (B.8) and the assumption $s_n = o(n)$ suggests

$$\sup_{M\in\Omega^*} \mathrm{E}\left(\sup_{\beta\in B_M} \left| \frac{1}{n} \sum_i \{g(O_i, \beta) - C(\beta)\} \right| \right) \leq \frac{\xi}{6}, \tag{B.9}$$

for sufficiently large n. Under the condition $\|Y_0\|_{\psi_1} = O(1)$, we can show the ψ_1 Orlicz norm of the envelope function of the class $\{g(o, \beta)\}$ is $O(1)$.

Hence, it follows from Lemma G.4 that

$$\begin{aligned}&\sup_{M\in\Omega^*}\Pr\left(\sup_{\beta\in B_M}\left|\frac{1}{n}\sum_i\{g(O_i,\beta)-C(\beta)\}\right|\right.\\ \geq\ &\left.\frac{3}{2}\mathrm{E}\sup_{\beta\in B_M}\left|\frac{1}{n}\sum_i\{g(O_i,\beta)-C(\beta)\}\right|+\frac{\xi}{4}\right)\leq\exp\left(-\frac{\bar{c}n}{\log n}\right),\end{aligned}$$

for some constant $\bar{c}$. This together with (B.9) suggests

$$\sup_{M\in\Omega^*}\Pr\left(\sup_{\beta\in B_M}\left|\frac{1}{n}\sum_i\{g(O_i,\beta)-C(\beta)\}\right|\geq\frac{\xi}{2}\right)\leq\exp\left(-\frac{\bar{c}n}{\log n}\right). \tag{B.10}$$

Since $n/\log n\gg s_n\log p$, for sufficiently large n, (B.7) is bounded by

$$O(p^{s_n})\exp\left(-\frac{\bar{c}n}{\log n}\right)\leq\exp\left(-\frac{\bar{c}n}{2\log n}\right).$$

Therefore, we've shown (B.4) can be bounded by

$$\Pr(\bar{\mathcal{A}}^c)+\exp\left(-\frac{\bar{c}n}{2\log n}\right).$$

Besides, similar to (B.19)-(B.21) (appear a few pages later), we have

$$\Pr(\bar{\mathcal{A}}^c)\leq\exp\left(-\frac{\bar{k}n}{2\log n}\right), \tag{B.11}$$

for some constant $\bar{k}$. This gives the probability bound that CIC chooses an underfitted model.

B.2. Overfitted model space. Let $R_M^C=t_0n^{-1/2}|M|^{1/2}\log^{1/2}p$, and $B_M^C=\{\beta:\beta\in B_M,\|\beta-\beta_0\|_2\leq R_M^C\}$. Since $|M|\leq s_n$ and $n\gg s_n\log p\log n$, we obtain that $\sup_{M\in\Omega_+^*}R_M^C\to0$. Similar to (9.22), conditional on the event

$$\bigcap_{\lambda\in\Omega_+}\left\{\|\tilde{\beta}_{M(\lambda)}-\beta_0\|_2\leq R_M^C\right\}, \tag{B.12}$$

we can show that (B.2) is bounded by

$$\begin{aligned}&\Pr\left(\sup_{M\in\Omega_+^*}\sup_{\beta\in B_M^C}\frac{n}{|M|}|m_C(\beta)-m_C(\beta_0)|\geq4\bar{c}\kappa_n\right)\\ +\ &\Pr\left(n|m_C(\hat{\beta}_{M(\lambda_0)})-m_C(\beta_0)|\geq4\bar{c}\kappa_n\right),\end{aligned} \tag{B.13}$$

for some constant $\bar{c} > 0$, where $m_C(\beta)$ stands for the centered U-process $m_C = \widehat{C} - C$. Since

$$\begin{aligned} m_C(\beta) &= \widehat{C}(\beta) - C(\beta) = \frac{1}{n(n-1)} \sum_{i \neq j} h(O_i, O_j, \beta) - C(\beta) \\ &= \left\{ \frac{1}{n(n-1)} \sum_{i \neq j} D(O_i, O_j, \beta) \right\} + \left\{ \frac{2}{n} \sum_i g(O_i, \beta) - 2C(\beta) \right\}, \end{aligned}$$

the first term of (B.13) can be bounded by

$$\begin{aligned} & \Pr\left(\sup_{M \in \Omega_+^*} \sup_{\beta \in B_M} \frac{n}{|M|} \left| \frac{1}{n(n-1)} \sum_{i \neq j} D(O_i, O_j, \beta) \right| \geq \bar{c}\kappa_n \right) \\ + & \Pr\left(\sup_{M \in \Omega_+^*} \sup_{\beta \in B_M^C} \frac{n}{|M|} \left| \frac{1}{n} \sum_i \{ g(O_i, \beta) - g(O_i, \beta_0) - C(\beta) + C(\beta_0) \} \right| \geq \bar{c}\kappa_n \right) \\ \leq & \sum_{M \in \Omega_+^*} \Pr\left(\sup_{\beta \in B_M} \frac{n}{|M|} \left| \frac{1}{n(n-1)} \sum_{i \neq j} D(O_i, O_j, \beta) \right| \geq \bar{c}\kappa_n \right) \\ + & \Pr\left(\sup_{M \in \Omega_+^*} \sup_{\beta \in B_M^C} \frac{n}{|M|} \left| \frac{1}{n} \sum_i \{ g(O_i, \beta) - g(O_i, \beta_0) - C(\beta) + C(\beta_0) \} \right| \geq \bar{c}\kappa_n \right). \end{aligned}$$

We begin by providing an upper bound for

$$\text{(B.14)} \quad \sum_{M \in \Omega_+^*} \Pr\left(\sup_{\beta \in B_M} \frac{n}{|M|} \left| \frac{1}{n(n-1)} \sum_{i \neq j} D(O_i, O_j, \beta) \right| \geq \bar{c}\kappa_n \right).$$

Under Assumption (A3), we have

$$\begin{aligned} \sup_\beta |h(O_i, O_j, \beta)| &= \frac{1}{2} \sup_\beta \left| \left(\frac{(A_i - \pi_i) Y_i}{\pi_i(1-\pi_i)} - \frac{(A_j - \pi_j) Y_j}{\pi_j(1-\pi_j)} \right) I(X_i^T \beta > X_j^T \beta) \right| \\ &+ \frac{1}{2} \sup_\beta \left| \left(\frac{(A_j - \pi_j) Y_j}{\pi_j(1-\pi_j)} - \frac{(A_i - \pi_i) Y_i}{\pi_i(1-\pi_i)} \right) I(X_j^T \beta > X_i^T \beta) \right| \\ &\leq \frac{|Y_i| + |Y_j|}{c_1(1-c_2)} \triangleq H(O_i, O_j). \end{aligned}$$

Therefore, H is the envelope function of h. Under the assumption $\|Y_i\|_{\psi_1} = O(1)$, this also implies

$$\| \sup_\beta |h(O_i, O_j, \beta)| \|_{\psi_1} \leq \|H(O_i, O_j)\|_{\psi_1} = O(1).$$

Hence, it follows from Lemma 2.2.2 in van der Vaart and Wellner (1996) that

$$\|\max_{i,j}\sup_{\beta}|h(O_i,O_j,\beta)|\|_{\psi_1}=O(\log n). \tag{B.15}$$

By Jensen's inequality, we have

$$\begin{aligned}&\|\sup_{\beta}|D(O_i,O_j,\beta)|\|_{\psi_1}\le\|\sup_{\beta}|h(O_i,O_j,\beta)|\|_{\psi_1}+\|\sup_{\beta}|g(O_i,\beta)|\|_{\psi_1}\\ +\;&\|\sup_{\beta}|g(O_j,\beta)|\|_{\psi_1}+\|\sup_{\beta}|C(\beta)|\|_{\psi_1}\le 4\|\sup_{\beta}|h(O_i,O_j,\beta)|\|_{\psi_1}.\end{aligned}$$

This together with (B.15) implies that

$$\omega_n\triangleq\left\|\sum_{i\neq j}\sup_{\beta}|D(O_i,O_j,\beta)|\right\|_{\psi_1}=O(\log n). \tag{B.16}$$

We can also show $4H$ is the envelope function of D. Define $\epsilon_1,\ldots,\epsilon_n$ to be i.i.d Radamacher random variables independent of $\{O_1,\ldots,O_n\}$. It follows from Jensen's inequality and the degeneracy of D that

$$\begin{aligned}&\mathrm{E}\sup_{\beta}|\sum_{i\neq j}D(O_i,O_j,\beta)I(H(O_i,O_j)\le 8\omega_n)|\\ \le\;&\mathrm{E}\sup_{\beta}|\sum_{i\neq j}\epsilon_i\epsilon_jD(O_i,O_j,\beta)I(H(O_i,O_j)\le 8\omega_n)|\\ \le\;&4\mathrm{E}\sup_{\beta}|\sum_{i\neq j}\epsilon_i\epsilon_jh(O_i,O_j,\beta)I(H(O_i,O_j)\le 8\omega_n)|.\end{aligned}$$

The class of functions $\{D(O_i,O_j,\beta):\beta\in B_M\}$ has VC index $|M|+2$, so is the class of functions $\{D(O_i,O_j,\beta)I(H(O_i,O_j)\le 8\omega_n):\beta\in B_M\}$. Using similar arguments in the proof of Lemma 5 and Theorem 6 in Nolan and Pollard (1987), we can show

$$\mathrm{E}Z_\varepsilon\triangleq\mathrm{E}\sup_{\beta}|\sum_{i\neq j}D(O_i,O_j,\beta)I(H(O_i,O_j)\le 8\omega_n)|=O(|M|n). \tag{B.17}$$

Let $t=\bar{c}(n-1)|M|\kappa_n/2$. Define

$$\begin{aligned}U_\varepsilon&=\sup_{\beta\in B_M}\sup_{\alpha:\|\alpha\|_2\le 1}\sum_{i,j}\epsilon_i\alpha_jD(O_i,O_j,\beta)I(H(O_i,O_j)\le 8\omega_n),\\ M_\varepsilon&=\sup_{\beta\in B_M}\sup_{k=1,\ldots,n}\left|\sum_i\epsilon_iD(O_i,O_k,\beta)I(H(O_i,O_k)\le 8\omega_n)\right|.\end{aligned}$$

Since $\kappa_n \to \infty$, by (B.16) and (B.17), for sufficiently large n, Theorem 7.1 applies and we have

$$\begin{aligned}
(\text{B.18}) \quad & \Pr\left(\sup_{\beta\in B_M}\frac{n}{|M|}\left|\frac{1}{n(n-1)}\sum_{i\neq j}D(O_i,O_j,\beta)\right|\geq \bar{c}\kappa_n\right)\\
= \ & \Pr\left(\sup_{\beta\in B_M}\left|\sum_{i\neq j}D(O_i,O_j,\beta)\right|\geq \bar{c}(n-1)|M|\kappa_n\right)\\
\leq \ & \Pr\left(\sup_{\beta\in B_M}\left|\sum_{i\neq j}D(O_i,O_j,\beta)\right|\geq \bar{k}\mathrm{E}Z_\varepsilon + t\right)\\
\leq \ & 3\exp\left(-\frac{1}{\bar{k}}\min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2},\frac{t}{\mathrm{E}M_\varepsilon},\frac{t}{n\omega_n},\left(\frac{t}{\omega_n\sqrt{n}}\right)^{2/3},\sqrt{\frac{t}{\omega_n}}\right)\right),
\end{aligned}$$

for some constant $\bar{k} > 0$, since $t = \bar{c}(n-1)|M|\kappa_n - \bar{k}\mathrm{E}Z_\varepsilon \gg \bar{c}(n-1)|M|\kappa_n/2$.

Using similar arguments in the proof of Corollary 6 in Clémençon, Lugosi and Vayatis (2008), we can show $\mathrm{E}U_\varepsilon = O(\sqrt{|M|n})$. Besides, by definition, it is immediate to see $\mathrm{E}M_\varepsilon = O(n\omega_n)$. Hence, it follows from (B.18), (B.16) and $n \gg \kappa_n \gg \log(n)$ that for sufficiently large n,

$$\begin{aligned}
(\text{B.19}) \quad & \Pr\left(\sup_{\beta\in B_M}\frac{n}{|M|}\left|\frac{1}{n(n-1)}\sum_{i\neq j}D(O_i,O_j,\beta)\right|\geq \bar{c}\kappa_n\right)\\
\leq \ & 3\exp\left(-K\min\left(|M|\kappa_n^2,\frac{|M|\kappa_n}{\log n},\left(\frac{|M|\sqrt{n}\kappa_n}{\log n}\right)^{2/3},\sqrt{\frac{|M|n\kappa_n}{\log n}}\right)\right)\\
\leq \ & 3\exp\left(-K\min\left(\frac{|M|\kappa_n}{\log n},\left(\frac{|M|\sqrt{n}\kappa_n}{\log n}\right)^{2/3},\sqrt{\frac{|M|n\kappa_n}{\log n}}\right)\right),
\end{aligned}$$

for some constant $K > 0$. Recall $\Omega_s^* = \{M \in \Omega_+^*, |M| = s\}$. The number of elements $|\Omega_s^*|$ is bounded by $O(p^s)$. This with (B.19) suggests

$$\begin{aligned}
(\text{B.20}) \quad & \sum_{M\in\Omega_s^*}\Pr\left(\sup_{\beta\in B_M}\frac{n}{s}\left|\frac{1}{n(n-1)}\sum_{i\neq j}D(O_i,O_j,\beta)\right|\geq \bar{c}\kappa_n\right)\\
\leq \ & O(p^s)\exp\left(-K\min\left(\frac{s\kappa_n}{\log n},\left(\frac{s\sqrt{n}\kappa_n}{\log n}\right)^{2/3},\sqrt{\frac{sn\kappa_n}{\log n}}\right)\right)\\
\leq \ & \exp\left(-K\min\left(\frac{s\kappa_n}{\log n},\left(\frac{s\sqrt{n}\kappa_n}{\log n}\right)^{2/3},\sqrt{\frac{sn\kappa_n}{\log n}}\right)+O(s\log p)\right).
\end{aligned}$$

Under the given condition $\kappa_n \gg \log p \log n$, $n \gg s \log p$ for all $s \leq s_n$, we obtain

$$\frac{s\kappa_n}{\log n} \gg s \log p, \ \left(\frac{s\sqrt{n}\kappa_n}{\log n}\right)^{2/3} \gg s \log p, \ \sqrt{\frac{sn\kappa_n}{\log n}} \gg s \log p,$$

and hence (B.20) is bounded by

$$\begin{aligned} & \exp\left(-\frac{K}{2}\min\left(\frac{s\kappa_n}{\log n}, \left(\frac{s\sqrt{n}\kappa_n}{\log n}\right)^{2/3}, \sqrt{\frac{sn\kappa_n}{\log n}}\right)\right) \\ \leq \ & \exp\left(-\frac{K}{2}\min\left(\frac{\kappa_n}{\log n}, \left(\frac{\sqrt{n}\kappa_n}{\log n}\right)^{2/3}, \sqrt{\frac{n\kappa_n}{\log n}}\right)\right) \\ \leq \ & \exp\left(-\frac{K\kappa_n}{2\log n}\right), \end{aligned}$$

by $\kappa_n = o(n)$, for sufficiently large n. Since $\Omega_+^* = \cup_{s=1}^{s_n} \Omega_s^*$, this together with (B.20) suggests (B.14) is bounded by

$$\begin{aligned} |s_n| \exp\left(-\frac{K\kappa_n}{2\log n}\right) \ & \leq \ \exp\left(-\frac{K\kappa_n}{2\log n} + \log n\right) \\ & \leq \ \exp\left(-\frac{K\kappa_n}{3\log n}\right) \leq \exp(-K\log p), \end{aligned} \tag{B.21}$$

since $s_n = o(n)$, $\kappa_n \gg \log p \log n$. Thus, we've establish the upper bound for (B.14). Now, we provide an upper bound for

$$\Pr\left(\sup_{\substack{M \in \Omega_+^* \\ \beta \in B_M^C}} \frac{1}{|M|}\left|\sum_{i=1}^n \{g(O_i, \beta) - g(O_i, \beta_0) - C(\beta) + C(\beta_0)\}\right| \geq \bar{c}\kappa_n\right). \tag{B.22}$$

It follows from Assumption (A7')(iii) that

$$\sup_{\beta \in B_M^C} \frac{1}{|M|}|C(\beta) - C(\beta_0)| \leq \bar{c}_2 \frac{(R_M^C)^2}{|M|} \leq O(1)\frac{\log p \log n}{n}, \quad \forall M \in \Omega_+^*.$$

Since $\kappa_n \gg \log p \log n$, (B.22) is bounded by

$$\Pr\left(\sup_{\substack{M \in \Omega_+^* \\ \beta \in B_M^C}} \frac{1}{|M|}\left|\sum_{i=1}^n \{g(O_i, \beta) - g(O_i, \beta_0)\}\right| \geq \frac{\bar{c}}{2}\kappa_n\right), \tag{B.23}$$

for sufficiently large n.

By Assumption (A7')(ii), g is twice continuously differentiable around β_0. For any $\beta \in B_M^C$, a second-order Taylor expansion gives

$$\begin{aligned} g(O_i,\beta) &= g(O_i,\beta_0) + \frac{\partial g(O_i,\beta_0)}{\partial \beta}(\beta_0-\beta) \\ &+ \frac{1}{2}(\beta_0-\beta)^T \Delta_2 g(O_i,\beta^*)(\beta_0-\beta), \end{aligned} \tag{B.24}$$

for some β^* lying on the line segment joining β and β_0. It follows from $\sup_{M\in\Omega_+^*} R_M^C \to 0$ that $\|\beta^*-\beta_0\|_2 \to 0$. Therefore, by Assumption (A7')(iii),

$$\begin{aligned} &\|\Delta_2 g(O_i,\beta^*) - \Delta_2 g(O_i,\beta_0)\|_2 \le K(O_i)\|\beta^*-\beta_0\|_2 \\ \le\ & K(O_i)\|\beta-\beta_0\|_2 \le R_M^C K(O_i) = o(1)K(O_i), \end{aligned} \tag{B.25}$$

where the $o(1)$ term is uniform in $i = 1,\dots,n$. Combining (B.24) with (B.25) gives

$$\begin{aligned} g(O_i,\beta) &= g(O_i,\beta_0) + \frac{\partial g(O_i,\beta_0)}{\partial \beta}(\beta_0-\beta) \\ &+ \frac{1}{2}(\beta_0-\beta)^T \Delta_2 g(O_i,\beta_0)(\beta_0-\beta) + o(1)\|\beta-\beta_0\|_2^2 K(O_i). \end{aligned} \tag{B.26}$$

Since $\|\beta_0\|_0$ is fixed, it follows from Assumption (A7')(v) that

$$\begin{aligned} &\|\mathrm{E}\Delta_2 g(O_i,\beta_0)\|_2 \le \|\mathrm{E}\Delta_2 g(O_i,\beta_0)\|_\infty \\ \le\ & \max_k \sum_j |\mathrm{E}\partial_{kj} g(O_i,\beta_0)| = O(1). \end{aligned} \tag{B.27}$$

Besides, we have

$$\mathrm{E}K(O_0) \le \|K(O_i)\|_{\psi_1} = O(1). \tag{B.28}$$

Define the events

$$\begin{aligned} \mathcal{E}_0 &= \left\{ \sum_i K(O_i) \le 2n\mathrm{E}K(O_0) \right\}, \\ \mathcal{E}_1 &= \left\{ \sup_{kj} \left| \sum_i (\partial_{kj} g(O_i,\beta_0) - \mathrm{E}\partial_{kj} g(O_i,\beta_0)) \right| \le n \right\}. \end{aligned}$$

Assume $\mathcal{E}_0 \cup \mathcal{E}_1$ holds. Then similar to (B.27), we can show

$$\left\| \frac{1}{n}\sum_{i=1}^n \Delta_2 g(O_i,\beta) \right\|_2 = O(1),$$

which together with (B.26) and (B.28) implies

$$\sum_i g(O_i,\beta) - g(O_i,\beta_0) = \sum_i \frac{\partial g(O_i,\beta_0)}{\partial \beta}(\beta-\beta_0) + O(\|\beta-\beta_0\|_2^2).$$

Since $\kappa_n \gg (R_M^C)^2/|M| = O(\log p \log n)$, this implies under $\mathcal{E}_0, \mathcal{E}_1$, (B.23) is bounded by

$$\Pr\left(\sup_{M\in\Omega_+^*}\sup_{\beta\in B_M^C}\left|\sum_i \frac{1}{|M|}\frac{\partial g(O_i,\beta_0)}{\partial\beta}(\beta-\beta_0)^T\right| \geq \frac{\bar{c}\kappa_n}{4}\right),$$

or equivalently,

$$\Pr\left(\sup_{M\in\Omega_+^*}\sup_{\beta\in B_M^C}\sup_{j\in M}\left|\sum_i \partial_j g(O_i,\beta_0)\right|_2 \geq \frac{\bar{c}\sqrt{|M|}\kappa_n}{4R_M^C}\right). \tag{B.29}$$

Define

$$\mathcal{E}_2 = \left\{\sup_j\left|\sum_i \partial_j g(O_i,\beta_0)\right| \geq \frac{\bar{c}\sqrt{|M|}\kappa_n}{4R_M^C}\right\}. \tag{B.30}$$

The probability defined in (B.29) is bounded by $\Pr(\mathcal{E}_2)$. From the above discussion, we've shown (B.23) is bounded by $\Pr(\mathcal{E}_0\cup\mathcal{E}_1\cup\mathcal{E}_2)$.

Under Assumption (A7')(iii) and (v), we have

$$\|K(O_i)\|_{\psi_1} = O(1),\ \sup_j\|\partial_j g(O_i,\beta_0)\|_{\psi_1} = O(1),\ \sup_{jk}\|\partial_{jk} g(O_i,\beta_0)\|_{\psi_1} = O(1).$$

Since $C(\beta_0)$ maximizes C, we have $\partial C(\beta_0)/\partial\beta = 0$ and thus

$$\mathrm{E}\frac{\partial g(O_i,\beta_0)}{\partial\beta} = 0.$$

It follows from Lemma G.2 that there exists some constants $\bar{c}_0 > 0$ that

$$\sup_j \Pr\left(\left|\sum_i \partial_j g(O_i,\beta_0)\right| \geq \frac{\bar{c}\sqrt{|M|}\kappa_n}{4R_M^C}\right) \tag{B.31}$$

$$\leq 2\exp\left(-\frac{\bar{c}_0\kappa_n^2}{\log p\log n}\right) + 2\exp\left(-\frac{\bar{c}_0\sqrt{n}\kappa_n}{\sqrt{\log p\log n}}\right) \leq 4\exp(-\bar{c}_0\log p\log n),$$

where the last inequality is due to $\kappa_n \gg \log p\log n$ and $n \gg \log p\log n$. Similarly we can show

$$\sup_{jk}\Pr\left(\left|\sum_i \partial_{jk} g(O_i,\beta_0)\right| \geq n\right) \leq 2\exp(-\bar{c}_0\log p\log n), \tag{B.32}$$

$$\Pr\left(\sum_i K(O_i) \geq 2n\mathrm{E}K(O_0)\right) \leq 2\exp(-\bar{c}_0\log p\log n).$$

Using Bonferroni's inequality, it follows from (B.31) and (B.32) that $\Pr(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2)$ is bounded by

$$\begin{aligned} & \Pr(\mathcal{E}_0 \cup \mathcal{E}_1 \cup \mathcal{E}_2) \le 2(p^2+1)\exp\left(-\bar{c}_0 \log p \log n\right) + 4p\exp(-\bar{c}_0 \log p \log n) \\ \le\ & \exp(-\bar{c}_0 \log p \log n + \log p + \log 4) + \exp(-\bar{c}_0 \log p \log n + \log(2p^2+2)) \\ \le\ & 2\exp(-\bar{c}_0 \log p \log n/2) \le \exp(-\bar{c}_0 \log p/4). \end{aligned}$$

Combining this together with (B.21), we've shown the first term of (B.13) is bounded by

$$\exp(-K \log p) + \exp(-\bar{c}_0 \log p/4) \le 2\exp(-\bar{K} \log p) \le \exp(-\bar{K} \log p/2),$$

for some constant $\bar{K} > 0$ and sufficiently large n.

Similarly, we can show that conditional on (B.12), the second term of (B.13) is bounded by

$$\exp\left(-\frac{\bar{K}_0 \kappa_n^2}{n(R_n^{(1)})^2}\right) + \exp(-\bar{K}_0 \log p),$$

for some constant $\bar{K}_0 > 0$. By Lemma 7.1, the event in (B.12) happens with probability at least $1 - \exp(-K_0 \log p)$ for some $K_0 > 0$. Thus, (B.2) can be bounded by

$$\begin{aligned} & \exp\left(-\frac{\bar{K}_0 \kappa_n^2}{n(R_n^{(1)})^2}\right) + \exp(-\bar{K}_0 \log p) + \exp(-\bar{K} \log p/2) \\ \le\ & \exp(-\bar{c} \log p) + \exp\left(-\frac{\bar{c}\kappa_n^2}{n(R_n^{(1)})^2}\right), \end{aligned}$$

for some constant $\bar{c} > 0$. The proof is hence completed.

APPENDIX C: PROOF OF THEOREM 4.2

C.1. Consistency of VIC^{DR}. For any M, let

$$\tilde{\theta}_M = \underset{\substack{\theta=(c,\beta^T)^T:\beta^{M^c}=0 \\ \|\theta\|_2=\|\theta_0\|_2}}{\arg\max} \widehat{V}^{DR}(\theta).$$

It suffices to show with probability tending to 1,

$$\text{VIC}^{DR}(\hat{\theta}_{M_\beta}) > \sup_{M \in \Omega_-} \text{VIC}^{DR}(\hat{\theta}_M), \tag{C.1}$$

$$\text{VIC}^{DR}(\hat{\theta}_{M_\beta}) > \sup_{M \in \Omega_+} \{n\widehat{V}^{DR}(\tilde{\theta}_M) - \kappa_n \|\hat{\beta}_M\|_0\}. \tag{C.2}$$

C.1.1. *Underfitted model space.* Similar to (9.8), it follows from Assumption (A5) and (A10)(i) that there exists some constant $\xi > 0$,

$$V^{DR}(\hat{\theta}_{M_\beta}) \geq \sup_{M \in \Omega_-} V^{DR}(\hat{\theta}_M) + 3\xi, \tag{C.3}$$

for sufficiently large n. Since $\kappa_n = o(n)$, for sufficiently large n, (C.3) further implies

$$\begin{aligned} & nV^{DR}(\hat{\theta}_{M_\beta}) - \kappa_n \|\hat{\beta}_{M_\beta}\|_0 \\ - & \sup_{M \in \Omega_-} \left\{ nV^{DR}(\hat{\theta}_M) - \kappa_n \|\hat{\beta}_M\|_0 \right\} \geq 2n\xi. \end{aligned} \tag{C.4}$$

Observe that $V(\theta, \alpha^*, \eta^*) = V^{DR}(\theta)$. Under the events defined in (A8)(i), it follows from (A10)(iii) that V is uniformly continuous and hence

$$\sup_\theta \left| V(\theta, \hat{\alpha}, \hat{\eta}) - V^{DR}(\theta) \right| = o(1).$$

Combining this together with (C.4) implies

$$\begin{aligned} & nV(\hat{\theta}_{M_\beta}, \hat{\alpha}, \hat{\eta}) - \kappa_n \|\hat{\beta}_{M_\beta}\|_0 \\ - & \sup_{M \in \Omega_-} \left\{ nV(\hat{\theta}_M, \hat{\alpha}, \hat{\eta}) - \kappa_n \|\hat{\beta}_M\|_0 \right\} \geq n\xi, \end{aligned} \tag{C.5}$$

for sufficiently large n. In view of (C.5), the event defined in (C.1) holds when

$$\sup_\theta \sup_{\substack{\|\hat{\alpha} - \alpha^*\|_2 \leq \epsilon \\ \|\hat{\eta} - \eta^*\|_2 \leq \epsilon}} \left| \widehat{V}(\theta, \alpha, \eta) - V(\theta, \alpha, \eta) \right| \geq \frac{\xi}{2},$$

for some small $\epsilon > 0$. Therefore, it suffices to show

$$\Pr \left(\sup_\theta \sup_{\substack{\|\hat{\alpha} - \alpha^*\|_2 \leq \epsilon \\ \|\hat{\eta} - \eta^*\|_2 \leq \epsilon}} \left| \widehat{V}(\theta, \alpha, \eta) - V(\theta, \alpha, \eta) \right| \geq \frac{\xi}{2} \right) \to 0. \tag{C.6}$$

We now show for any $M \in \Omega^*$, the class of functions

$$\begin{aligned} f(o) \quad = \quad & \left\{ \frac{aI(x^T\beta > -c)}{\pi(x, \alpha)} + \frac{(1-a)I(x^T\beta \leq -c)}{1 - \pi(x, \alpha)} \right\} y \\ - \quad & \left\{ \frac{aI(x^T\beta > -c)}{\pi(x, \alpha)} + \frac{(1-a)I(x^T\beta \leq -c)}{1 - \pi(x, \alpha)} - 1 \right\} h(x, \eta), \end{aligned}$$

indexed by c, β, α, η belongs to the VC type class. That means, there exist some measurable envelope function F and positive constants K, $1 \le v < \infty$ such that

$$\sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \le (K/\varepsilon)^v, \forall 0 < \varepsilon \le 1, \tag{C.7}$$

where $\mathcal{F} = \{f(o) : c \in \mathbb{R}, \beta \in \mathbb{R}^p, \alpha \in \mathbb{R}^{q_1}, \eta \in \mathbb{R}^{q_2}\}$, $N(\cdot,\cdot,\cdot)$ stands for the entropy function (cf. Definition 2.2.3, van der Vaart and Wellner, 1996). The supremum in (C.7) is taken over all discrete measure Q such that $0 < QF^2 < \infty$, and $L_2(Q)$ is the norm on $\mathcal{F}$ defined as $\|f\|_{Q,2} = (\int |f|^2 dQ)^{1/2}$.

To prove this, we first show the class of functions

$$f_1(o) = \frac{aI(x^T\beta > -c)}{\pi(x,\alpha)} y$$

indexed by c, β, α belongs to the VC type class with the envelope function

$$F_1(o) = \frac{\sqrt{2|y|^2 + 2}}{c_1^2}.$$

Define

$$\mathcal{G}_1 = \{ayI(x^T\beta > -c) : c \in \mathbb{R}, \beta \in \mathbb{R}^p\}, \quad \mathcal{G}_2 = \{\pi(x,\alpha) : \alpha \in \mathbb{R}^{q_1}\}.$$

Class of functions $\mathcal{G}_1$ has finite VC index with the envelope function $G_1(o) = |y|$. By Assumption (A9), $\mathcal{G}_2$ also has VC index. Besides, $\mathcal{G}_2$ is uniformly bounded by 1. Consider function

$$\phi(x_1, x_2) = \frac{x_1}{x_2}.$$

Note that the class of function $\mathcal{F}_1 = \{f_1(o) : c \in \mathbb{R}, \beta \in B_{M_0}, \alpha \in A_{M_1}\}$ can be represented as $\{\phi(g_1, g_2) : g_1 \in \mathcal{G}_1, g_2 \in \mathcal{G}_2\}$. Besides, for any $g_1, g_3 \in \mathcal{G}_1$, $g_2, g_4 \in \mathcal{G}_2$, we have

$$\begin{aligned} |\phi(g_1,g_2) - \phi(g_3,g_4)|^2 &\le \frac{|g_1g_4 - g_2g_3|^2}{|g_2g_4|^2} \le \frac{1}{c_1^4}|g_1g_4 - g_2g_3|^2 \\ &\le \frac{2}{c_1^4}|g_1 - g_3|^2 + \frac{2}{c_1^4}|g_2 - g_4|^2, \end{aligned} \tag{C.8}$$

by Assumption (A3). It follows from (C.8) and Lemma A.6 in Chernozhukov, Chetverikov and Kato (2014) that

$$\sup_Q N(\varepsilon\|F_1\|_2, \mathcal{F}_1, L_2(Q)) \le \sup_Q N(\varepsilon\|G_1\|_2, \mathcal{G}_1, L_2(Q)) \sup_Q N(\varepsilon, \mathcal{G}_2, L_2(Q)).$$

The above entropy is bounded by $(K_0/\varepsilon)^{v_0}$ for some constants K_0 and $1 \le v_0 < \infty$. This shows $\mathcal{F}_1$ belongs to the VC type class with VC index bounded by v_0. Similarly one can show the following classes of functions

$$f_2(o) = \frac{(1-a)I(x^T\beta \le -c)}{1-\pi(x,\alpha)}y,$$
$$f_3(o) = \frac{aI(x^T\beta > -c)}{\pi(x,\alpha)}h(x,\eta),$$
$$f_4(o) = \frac{(1-a)I(x^T\beta \le -c)}{1-\pi(x,\alpha)}h(x,\eta),$$

belong to the VC type class. Then repeated applications of Lemma G.3 imply that $\mathcal{F}$ belongs to the VC type class.

Therefore, similar to (9.15), we can show

$$\mathrm{E}\left(\sup_{\theta}\sup_{\substack{\|\hat\alpha-\alpha^*\|_2\le\epsilon\\ \|\hat\eta-\eta^*\|_2\le\epsilon}}\left|\widehat{V}(\theta,\alpha,\eta)-V(\theta,\alpha,\eta)\right|\right)\to 0.$$

Assertion (C.6) thus follows. This proves (C.1).

C.1.2. *Overfitted model space.* Similar to Lemma 7.1, we can show that for any overfitted model space M, $\tilde\theta_M$ satisfies

$$\tilde\theta_M = \theta_0 + O_p(n^{-1/3}). \tag{C.9}$$

The class of functions

$$\left\{\{I(x^T\beta > -c) - I(x^T\beta_0 > -c_0)\}\left(\frac{a}{\pi_\alpha(x)}+\frac{(1-a)}{1-\pi_\alpha(x)}\right)\right\}y$$
$$-\left\{\{I(x^T\beta > -c) - I(x^T\beta_0 > -c_0)\}\left(\frac{a}{\pi_\alpha(x)}+\frac{(1-a)}{1-\pi_\alpha(x)}\right)-1\right\}h_\eta(x),$$

indexed by $\{\|\theta-\theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}, \|\alpha-\alpha^*\|_2 = O(n^{-1/2}), \|\eta-\eta^*\|_2 = O(n^{-1/2})\}$, belongs to the VC type class with finite VC index. Similar to (9.23), we can show its envelope function Ψ_V satisfies

$$\mathrm{E}|\Psi_V|^2 = O\{R_n, n^{-1/3}\}.$$

Therefore, similar to (9.24), we can show that conditional on $\hat\alpha$ and $\hat\eta$,

$$\mathrm{E}\left(\sup_{\|\theta-\theta_0\|_2=O\{\max(R_n,n^{-1/3})\}} n\,|m_V(\theta,\hat\alpha,\hat\eta)-m_V(\theta_0,\hat\alpha,\hat\eta)|\right)$$
$$= O_p\left(\sqrt{n\max(R_n,n^{-1/3})}\right),$$

where $m_V(\theta,\hat{\alpha},\hat{\eta}) = \widehat{V}(\theta,\hat{\alpha},\hat{\eta}) - V(\theta,\hat{\alpha},\hat{\eta})$.

By Markov's inequality, this means with probability tending to 1, we have

$$\begin{aligned}(\text{C.10}) \quad & \sup_{\|\theta-\theta_0\|_2=O\{\max(R_n,n^{-1/3})\}} n\,|m_V(\theta,\hat{\alpha},\hat{\eta}) - m_V(\theta_0,\hat{\alpha},\hat{\eta})| \\ = \; & O\left(\sqrt{n\max(R_n,n^{-1/3})}\right).\end{aligned}$$

We now show that for any θ that satisfies $\|\theta-\theta_0\|_2 = O\{\max(R_n,n^{-1/3})\}$,

$$\begin{aligned}(\text{C.11}) \quad & |V(\theta,\hat{\alpha},\hat{\eta}) - V(\theta_0,\hat{\alpha},\hat{\eta}) - V(\theta,\alpha^*,\theta^*) + V(\theta_0,\alpha^*,\theta^*)| \\ = \; & O\{\max(R_n^2,n^{-2/3})\}.\end{aligned}$$

Let $\bar{\zeta} = (\theta^T,\alpha^T,\eta^T)^T$ and $\bar{\zeta}_0 = (\theta_0^T,(\alpha^*)^T,(\eta^*)^T)^T$. It follows from Assumptions (A10)(iii) that

$$\begin{aligned}V(\theta,\alpha,\eta) \; = \; & V(\theta_0,\alpha^*,\eta^*) + \frac{\partial V}{\partial\theta_0}(\theta-\theta_0) + \frac{\partial V}{\partial\alpha^*}(\alpha-\alpha^*) + \frac{\partial V}{\partial\eta^*}(\eta-\eta^*) \\ (\text{C.12}) \quad + \; & \frac{1}{2}(\bar{\zeta}-\bar{\zeta}_0)^T\Delta_2 V(\bar{\zeta}_0)(\bar{\zeta}-\bar{\zeta}_0) + o(\|\bar{\zeta}-\bar{\zeta}_0\|_2^2).\end{aligned}$$

Specifically, take $\theta=\theta_0$ in (C.12), we obtain

$$\begin{aligned}V(\theta_0,\alpha,\eta) \; = \; & V(\theta_0,\alpha^*,\eta^*) + \frac{\partial V}{\partial\alpha^*}(\alpha-\alpha^*) + \frac{\partial V}{\partial\eta^*}(\eta-\eta^*) \\ (\text{C.13}) \quad + \; & O(\|\alpha-\alpha^*\|_2^2 + \|\eta-\eta^*\|_2^2),\end{aligned}$$

Besides, set $\alpha=\alpha^*$ and $\eta=\eta^*$ in (C.12), we have

$$(\text{C.14}) \quad V(\theta,\alpha^*,\eta^*) = V(\theta_0,\alpha^*,\eta^*) + \frac{\partial V}{\partial\theta_0}(\theta-\theta_0) + O(\|\theta-\theta_0\|_2^2).$$

The linear term $\partial V/\partial\theta_0$ in (B.26) is equal to 0 since θ_0 maximizes $V^{DR}(\theta) = V(\theta,\alpha^*,\eta^*)$. Under Assumption (A8)(i), combining (C.14) together with (C.12) and (C.13), we obtain (C.11).

Under Assumption (A10)(iii), we have for any θ that satisfies $\|\theta-\theta_0\|_2 = O\{\max(R_n,n^{-1/3})\}$,

$$|V(\theta_0,\alpha^*,\eta^*) - V(\theta,\alpha^*,\eta^*)| = O\{\max(R_n^2,n^{-2/3})\}.$$

This together with (C.11) implies that

$$(\text{C.15}) \quad |V(\theta,\hat{\alpha},\hat{\eta}) - V(\theta_0,\hat{\alpha},\hat{\eta})| = O\{\max(R_n^2,n^{-2/3})\},$$

with probability tending to 1, for θ such that $\|\theta-\theta_0\|_2 = O\{\max(R_n, n^{-1/3})\}$.

Under the given condition, we have $\kappa_n \gg \max(nR_n^2, \sqrt{nR_n}, n^{1/3})$. Observe that $\widehat{V}^{DR}(\theta) = \widehat{V}(\theta, \hat{\alpha}, \hat{\eta})$ and $V^{DR}(\theta) = V(\theta, \alpha^*, \theta^*)$. Combining (C.10) with (C.15), we have

$$\sup_{\|\theta-\theta_0\|_2=O\{\max(R_n,n^{-1/3})\}} n\left|\widehat{V}^{DR}(\theta) - V^{DR}(\theta) - \widehat{V}^{DR}(\theta_0) + V^{DR}(\theta_0)\right| = O\left(nR_n^2 + \sqrt{nR_n} + n^{1/3}\right) \ll \kappa_n.$$

Under Assumption (A5) and (C.9), this further implies

$$\text{(C.16)} \quad \sup_{M\in\Omega_+} n\left|\widehat{V}^{DR}(\hat{\theta}_M) - V^{DR}(\hat{\theta}_M) - \widehat{V}^{DR}(\hat{\theta}_{M_\beta}) + V^{DR}(\hat{\theta}_{M_\beta})\right| \ll \kappa_n,$$

with probability tending to 1. Recall that $\text{VIC}^{DR}(\theta) = n\widehat{V}^{DR}(\theta) - \kappa_n\|\beta\|_0$ for $\theta = (c, \beta^T)^T$. For any $M \in \Omega_+$, we have $\kappa_n\|\hat{\beta}_M\|_0 - \kappa_n\|\hat{\beta}_{M_\beta}\|_0 \geq \kappa_n$. Under the event defined in (C.16), this implies (C.2). The proof is hence completed.

C.2. Consistency of CIC^{DR}. Similar to the proof of Theorem 3.4, it suffices to show with probability tending to 1,

$$\text{(C.17)} \quad \text{CIC}^{DR}(\hat{\beta}_{M_\beta}) > \sup_{M\in\Omega_-} \text{CIC}^{DR}(\hat{\beta}_M),$$

$$\text{(C.18)} \quad \text{CIC}^{DR}(\hat{\beta}_{M_\beta}) > \sup_{M\in\Omega_+} \{n\widehat{C}^{DR}(\tilde{\beta}_M) - \kappa_n\|\hat{\beta}_M\|_0\},$$

where $\tilde{\beta}_M$ denotes the empirical maximizer of $\widehat{C}^{DR}$ on the restricted model space with $\|\tilde{\beta}_M\|_2 = \|\beta_0\|_2$. (C.17) can be proven using similar arguments in Section B.2. In the following, we focus on (C.18).

Using similar arguments in the proof of Theorem 4 in Fan et al. (2016), we can show that for any overfitted model M,

$$\text{(C.19)} \quad \tilde{\beta}_M = \beta_0 + O_p(n^{-1/2}).$$

The class of functions

$$\left\{\frac{\{A_i - \pi(X_i,\alpha)\}\{Y_i - h(X_i,\eta)\}A_j}{\pi(X_i,\alpha)\{1-\pi(X_i,\alpha)\}\pi(X_j,\alpha)} - \frac{\{A_j - \pi(X_j,\alpha)\}\{Y_j - h(X_j,\eta)\}A_i}{\pi(X_j,\alpha)\{1-\pi(X_j,\alpha)\}\pi(X_i,\alpha)}\right\} I(X_i^T\beta > X_j^T\beta),$$

belongs to the VC type class. Let $\tilde{\beta}_{M_\beta} = \hat{\beta}_{M_\beta}$. The maximal inequality for degenerate U-process implies for any model $M \in \Omega_+ \cup \{M_\beta\}$,

$$\text{(C.20)}\quad \widehat{C}^{DR}(\tilde{\beta}_M) = \frac{2}{n}\sum_{i=1}^{n} g(O_i, \tilde{\beta}_M, \hat{\alpha}, \hat{\eta}) - C(\tilde{\beta}_M, \hat{\alpha}, \hat{\eta}) + O_p\left(\frac{1}{n}\right).$$

Besides, it follows from Assumption (A11)(ii) and (C.19) that for any $M \in \Omega_+ \cup \{M_\beta\}$,

$$\begin{aligned}\text{(C.21)}\qquad C^{DR}(\tilde{\beta}_M) &= C^{DR}(\beta_0) + O_p\left((R_n^{(1)})^2\right) + O_p\left(\frac{1}{n}\right)\\ &= C^{DR}(\beta_0) + O_p\left((R_n^{(1)})^2\right),\end{aligned}$$

where the last equation is due to that $R_n^{(1)} \geq n^{-1/2}$. Since $\kappa_n \gg n(R_n^{(1)})^2$, it follows from (C.21) that with probability tending to 1,

$$\text{(C.22)}\qquad n|C^{DR}(\tilde{\beta}_M) - C^{DR}(\tilde{\beta}_{M_\beta})| \ll \kappa_n.$$

Assume for now we can show

$$\text{(C.23)}\qquad \left|\sum_i g(O_i, \tilde{\beta}_M, \hat{\alpha}, \hat{\eta}) - \sum_i g(O_i, \tilde{\beta}_{M_\beta}, \hat{\alpha}, \hat{\eta})\right| \ll \kappa_n,$$

with probability tending to 1. Combining (C.22), (C.23) together with (C.20) suggests that for any $M \in \Omega_+$,

$$\text{(C.24)}\qquad n|\widehat{C}^{DR}(\tilde{\beta}_M) - \widehat{C}^{DR}(\hat{\beta}_{M_\beta})| \ll \kappa_n.$$

By the definition of CIC^{DR}, (C.24) implies (C.18). Therefore, it remains to show (C.23).

Let $\hat{\zeta}_M = (\tilde{\beta}_M^T, \hat{\alpha}^T, \hat{\eta}^T)^T$ and $\zeta_0 = (\beta_0^T, (\alpha^*)^T, (\eta^*)^T)^T$ Under Assumption (A11)(iii), a second order Taylor expansion around ζ_0 implies

$$\begin{aligned}\sum_i g(O_i, \hat{\zeta}_M) &= \sum_i g(O_i, \zeta_0) + \sum_i \frac{\partial g(O_i, \zeta_0)}{\partial \zeta_0}(\hat{\zeta}_M - \zeta_0)\\ &+ \frac{1}{2}(\hat{\zeta}_M - \zeta_0)^T \sum_i \Delta_2 g(O_i, \zeta_M^*)(\hat{\zeta}_M - \zeta_0),\end{aligned}\qquad \text{(C.25)}$$

for some ζ_M^* lying on the line segment between ζ_0 and $\hat{\zeta}_M$. Besides, it follows from Assumption (A11)(iv) that

$$\text{(C.26)}\qquad \left\|\sum_i \Delta_2 g(O_i, \zeta_M^*) - \sum_i \Delta_2 g(O_i, \zeta_0)\right\|_2 \leq \sum_i K(O_i)\|\zeta_M^* - \zeta_0\|_2.$$

Note that function $K(O_0)$ is integrable, and each element in the Hessian matrix $\Delta_2 g(O_i, \zeta_0)$ is integrable. Combining (C.25) with (C.26) suggests

$$\begin{aligned} \sum_i g(O_i, \hat{\zeta}_M) &= \sum_i g(O_i, \zeta_0) + \sum_i \frac{\partial g(O_i, \zeta_0)}{\partial \zeta_0}(\hat{\zeta}_M - \zeta_0) \\ &+ O_p(n\|\hat{\zeta}_M - \zeta_0\|_2^2). \end{aligned} \tag{C.27}$$

Observe $\sum_i \partial_j g(O_i, \zeta_0)$ is the summation of mean zero i.i.d random variable and hence we have $\sum_i \partial_j g(O_i, \zeta_0) = O_p(\sqrt{n})$. Under Assumptions (A5) and (A8), this together with (C.27) implies

$$\begin{aligned} \sum_i g(O_i, \hat{\zeta}_M) &= \sum_i g(O_i, \zeta_0) + O_p(n^{1/2}(R_n^{(1)} + n^{-1/2})) + O_p(n(R_n^{(1)} + n^{-1/2})^2) \\ &= \sum_i g(O_i, \zeta_0) + O_p(n(R_n^{(1)})^2). \end{aligned}$$

Under the Assumption that $\kappa_n \gg n(R_n^{(1)})^2$, this proves (C.23). The proof is hence completed.

APPENDIX D: PROOF OF THEOREM 7.1

Let F be the envelope of the class of functions $\mathcal{F}$, i.e,

$$F(x, y) = \sup_{f \in \mathcal{F}} |f(x, y)|.$$

For a given $\rho > 0$, define $\mathcal{F}_1$ and $\mathcal{F}_2$ to be the following class of functions

$$\begin{aligned} \mathcal{F}_1 &= \{f(x, y) I(F(x, y) \le \rho), f \in \mathcal{F}\}, \\ \mathcal{F}_2 &= \{f(x, y) I(F(x, y) > \rho), f \in \mathcal{F}\}. \end{aligned}$$

It is immediate to see that

$$\begin{aligned} Z &\le \sup_{f_1 \in \mathcal{F}_1} \left| \sum_{i \ne j} \{f_1(X_i, X_j) - \mathrm{E} f_1(X_i, X_j)\} \right| \\ &+ \sup_{f_2 \in \mathcal{F}_2} \left| \sum_{i \ne j} \{f_2(X_i, X_j) - \mathrm{E} f_2(X_i, X_j)\} \right| \triangleq Z_1 + Z_2. \end{aligned} \tag{D.1}$$

Besides, it follows by Jensen's inequality that

$$\mathrm{E} Z_2 \le 2\mathrm{E} \sup_{f_2 \in \mathcal{F}_2} \left| \sum_{i \ne j}^{n} f_2(X_i, X_j) \right| \triangleq 2\mathrm{E} I. \tag{D.2}$$

Combining (D.1) with (D.2), we obtain

$$\begin{aligned}(\text{D.3}) \qquad & \Pr\left(Z \geq C\mathrm{E}Z_\varepsilon + 2t\right) \\ \leq\ & \Pr\left(Z_1 \geq C\mathrm{E}Z_\varepsilon + t\right) + \Pr\left(Z_2 \geq t\right).\end{aligned}$$

Hence, it suffices to provide upper bounds on the probabilities in the second line of (D.3). We first bound $\Pr(Z_2 \geq t)$.

Observe that for any function f, we can express the U-statistic as average of sums of i.i.d blocks,

$$(\text{D.4}) \quad \frac{1}{n(n-1)}\sum_{i\neq j} f(X_i, X_j) = \frac{1}{n!}\sum_{\pi}\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} f(X_{\pi(i)}, X_{\pi(i+\lfloor n/2\rfloor)}).$$

where π stands for all permutations over $[1,\ldots,n]$ and $\lfloor x\rfloor$ stands for the largest integer y such that $y \leq x$.

Let $\rho = 8\omega_n$. It follows by the definition of ω_n that

$$(\text{D.5}) \quad 8\mathrm{E}\max_{1\leq i\leq\lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor}) \leq 8\|\max_i F(X_i, X_{i+\lfloor n/2\rfloor})\|_{\psi_1} \leq \rho,$$

By Chebyshev inequality, we have

$$\begin{aligned}& \Pr\left(\max_{k\leq\lfloor n/2\rfloor}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{k} f_2(X_i, X_{i+\lfloor n/2\rfloor})\right| > 0\right) \\ \leq\ & \Pr\left(\max_{1\leq i\leq\lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor}) > \rho\right) \leq \frac{1}{8}.\end{aligned}$$

Combining this with the Hoffmann-Jørgensen inequality (cf. Ledoux and Talagrand, 2011, Proposition 6.8), we have

$$(\text{D.6}) \quad \mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor} f_2(X_i, X_{i+\lfloor n/2\rfloor})\right| \leq 8\mathrm{E}\max_{1\leq i\leq\lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor}).$$

Apply the decomposition (D.4) to the class of functions in $\mathcal{F}_2$, it follows by Lemma A.1 in Clémençon, Lugosi and Vayatis (2008) that

$$\begin{aligned}\mathrm{E}I \ &\leq\ \frac{1}{n!}\sum_{\pi}\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\frac{1}{\lfloor n/2\rfloor}\sum_{i=1}^{\lfloor n/2\rfloor} f_2(X_{\pi(i)}, X_{\pi(i+\lfloor n/2\rfloor)})\right| \\ &\leq\ \frac{1}{\lfloor n/2\rfloor}\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor} f_2(X_i, X_{i+\lfloor n/2\rfloor})\right|.\end{aligned}$$

Consequently, together with (D.6) and (D.5), we obtain

$$\begin{aligned} \mathrm{E}I &\leq 8\mathrm{E}\max_{1\leq i\leq \lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor}) \\ &\leq 8\left\|\max_{1\leq i\leq \lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor})\right\|_{\psi_1} \leq \rho. \end{aligned} \tag{D.7}$$

Due to the decomposition (D.4), we have

$$\begin{aligned} \|Z_2\|_{\psi_1} &\leq \frac{1}{n!}\sum_{\pi}\frac{n(n-1)}{\lfloor n/2\rfloor}\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_{\pi(i)}, X_{\pi(i+\lfloor n/2\rfloor)}) - \mathrm{E}f_2\}\right|\right\|_{\psi_1} \\ &\leq 2n\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right|\right\|_{\psi_1}. \end{aligned} \tag{D.8}$$

It follows by Theorem 6.21 in Ledoux and Talagrand (2011) that

$$\begin{aligned} &\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right|\right\|_{\psi_1} \\ \leq\ &\mathrm{E}\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right| + 2\left\|\max_{i=1}^{\lfloor n/2\rfloor} F(X_i, X_{i+\lfloor n/2\rfloor})\right\|_{\psi_1}. \end{aligned}$$

This together with (D.2), (D.5), (D.6) and (D.7) suggests

$$\left\|\sup_{f_2\in\mathcal{F}_2}\left|\sum_{i=1}^{\lfloor n/2\rfloor}\{f_2(X_i, X_{i+\lfloor n/2\rfloor}) - \mathrm{E}f_2(X_i, X_{i+\lfloor n/2\rfloor})\}\right|\right\|_{\psi_1} \leq K'\omega_n/2,$$

for some constant K'. Combining this with (D.8), we obtain

$$\|Z_2\|_{\psi_1} \leq nK'\omega_n.$$

Therefore, by Markov's inequality, we obtain

$$\Pr(Z_2 \geq t) \leq 2\exp\left(-\frac{t}{nK'\omega_n}\right). \tag{D.9}$$

It remains to bound

$$\Pr(Z_1 \geq C\mathrm{E}Z_\varepsilon + t).$$

Since each function in $\mathcal{F}_1$ is bounded, it follows by Lemma (G.5) that

$$\begin{aligned}(\text{D.10}) \quad & \Pr(Z_1 \geq C\mathrm{E}Z_1 + t) \\ \leq \quad & \exp\left(-\frac{1}{C}\min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2}, \frac{t}{\mathrm{E}M_\varepsilon}, \frac{t}{8n\omega_n}, \left(\frac{t}{8\sqrt{n}\omega_n}\right)^{2/3}, \sqrt{\frac{t}{8\omega_n}}\right)\right).\end{aligned}$$

The proof is hence completed by (D.9) and (D.10).

APPENDIX E: PROOF OF LEMMA 7.2

E.1. Uniform convergence of empirical maximizer of $\widehat{V}$. Recall that $\Omega_+^* = \{M \subseteq [1, \ldots, p] : M_\beta \subsetneq M, |M| \leq s_n\}$. Hence, we have

$$\{M(\lambda) : \lambda \in \Omega_+\} \subseteq \Omega_+^*. \tag{E.1}$$

Since $\tilde{\theta}_{M(\lambda_1)} = \tilde{\theta}_{M(\lambda_2)}$ whenever $M(\lambda_1) = M(\lambda_2)$, we have

$$\begin{aligned}(\text{E.2}) \qquad & \bigcap_{\lambda\in\Omega_+} \left\{\|\tilde{\theta}_{M(\lambda)} - \theta_0\|_2 \geq tn^{-1/3}|M(\lambda)|^{1/3}\log^{1/3} p\right\} \\ \subseteq \quad & \bigcap_{M\in\Omega_+^*} \left\{\|\tilde{\theta}_M - \theta_0\|_2 \geq tn^{-1/3}|M(\lambda)|^{1/3}\log^{1/3} p\right\}.\end{aligned}$$

Define

$$\mathcal{A}_0 = \bigcap_{M\in\Omega_+^*} \left\{\tilde{\theta}_M \in \tilde{N}_{\varepsilon_0}\right\},$$

where ε_0 is defined in (A6')(i). It follows from (E.2) that

$$\begin{aligned}(\text{E.3}) \qquad & \Pr\left(\bigcap_{\lambda\in\Omega_+} \left\{\|\tilde{\theta}_{M(\lambda)} - \theta_0\|_2 \geq tn^{-1/3}|M(\lambda)|^{1/3}\log^{1/3} p\right\}\right) \\ \leq \quad & \Pr\left(\bigcap_{M\in\Omega_+^*} \left\{\|\tilde{\theta}_M - \theta_0\|_2 \geq tn^{-1/3}|M|^{1/3}\log^{1/3} p\right\} \bigcap \mathcal{A}_0^c\right) \\ + \quad & \Pr\left(\bigcap_{M\in\Omega_+^*} \left\{\|\tilde{\theta}_M - \theta_0\|_2 \geq tn^{-1/3}|M|^{1/3}\log^{1/3} p\right\} \bigcap \mathcal{A}_0\right) \\ \leq \quad & \Pr(\mathcal{A}_0^c) + \Pr\left(\bigcap_{M\in\Omega_+} \left\{\|\tilde{\theta}_M - \theta_0\|_2 \geq \frac{t|M|^{1/3}\log^{1/3} p}{n^{1/3}}\right\} \bigcap \mathcal{A}_0\right).\end{aligned}$$

In view of (E.1) and (E.3), it suffices to bound

$$\Pr(\mathcal{A}_0^c) + \Pr\left(\bigcap_{M\in\Omega_+^*}\left\{\|\tilde{\theta}_M - \theta_0\|_2 \geq tn^{-1/3}|M|^{1/3}\log^{1/3} p\right\}\bigcap \mathcal{A}_0\right).$$

In the following, we break the proof into two steps. In the first step, we show

$$\Pr(\mathcal{A}_0^c) \leq \exp\left(-\frac{\bar{c}_0 n}{\log n}\right), \tag{E.4}$$

for some constant $\bar{c}_0 > 0$. In the second step, we show

$$\begin{aligned}\Pr&\left(\bigcap_{M\in\Omega_+^*}\left\{\|\tilde{\theta}_M - \theta_0\|_2 \geq tn^{-1/3}|M|^{1/3}\log^{1/3} p\right\}\bigcap \mathcal{A}_0\right) \\ &\leq \exp\left(-c_0^* t^3 \log p\right) + \exp\left(-\frac{c_0^* t^2 n^{1/3}\log^{2/3} p}{\log n}\right),\end{aligned} \tag{E.5}$$

for some constant $c_0^* > 0$ and all $t \geq t_0$, (7.5) thus holds by setting $\bar{c} = \min(\bar{c}_0, c_0^*)$.

E.1.1. *Proof of* (E.4). For any $M \in \Omega_+^*$, if $\tilde{\theta}_M \notin \tilde{N}_{\varepsilon_0}$, by definition, we have

$$\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0}^c\\ \beta^{M^c}=0}} \widehat{V}(\theta) \geq \sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0}\\ \beta^{M^c}=0}} \widehat{V}(\theta),$$

and hence

$$\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0}^c\\ \beta^{M^c}=0}} \widehat{V}(\theta) \geq \widehat{V}(\theta_0).$$

This further implies

$$\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0}^c\\ \beta^{M^c}=0}} \{m_V(\theta) - m_V(\theta_0)\} \geq V(\theta_0) - \sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0}^c\\ \beta^{M^c}=0}} V(\theta). \tag{E.6}$$

By Assumption (A6'), there exists some constants $\delta_0 > 0$ such that for sufficiently large n,

$$V(\theta_0) - \sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}_{\varepsilon_0}^c\\ \beta^{M^c}=0}} \geq 2\delta_0.$$

Combining this together with (E.6) gives

$$\sup_{\substack{\theta=(c,\beta^T)^T\in\tilde{S}(\theta_0)\cap\tilde{N}^c_{\varepsilon_0}\\ \beta^{M^c}=0}} \{m_V(\theta)-m_V(\theta_0)\}\geq 2\delta_0,$$

which further implies

$$\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{M^c}=0}} |m_V(\theta)|\geq\delta_0 \quad \text{or} \quad |m_V(\theta_0)|\geq\delta_0.$$

To summarize, we've shown

$$\begin{aligned}\Pr(\tilde{\theta}_M\notin\tilde{N}_{\varepsilon_0}) &\leq \Pr\left(\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{M^c}=0}}|m_V(\theta)|\geq\delta_0\right)+\Pr(|m_V(\theta_0)|\geq\delta_0)\\ &\leq 2\Pr\left(\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{M^c}=0}}|m_V(\theta)|\geq\delta_0\right),\end{aligned}$$

where the last inequality is due to that $\beta_0^{M^c}=0$. It follows from Bonferroni's inequality that

$$2\Pr(\mathcal{A}_0^c)\leq 2\sum_{M\in\Omega_+^*}\left(\sup_{\substack{\theta=(c,\beta^T)^T\\ \beta^{M^c}=0}}|m_V(\theta)|\geq\delta_0\right). \tag{E.7}$$

Similar to (9.10), we can show that RHS of (E.7) is upper bounded by $\exp(-\bar{c}n/\log n)$. This proves (E.4).

E.1.2. *Proof of* (E.5). On the set $\mathcal{A}_0$, it follows from Assumption (A6')(iii) that for all $M\in\Omega_+^*$, we have

$$V(\theta_0)-V(\tilde{\theta}_M)\geq\bar{c}_1\|\theta_0-\tilde{\theta}_M\|_2^2.$$

When $\|\theta_0-\tilde{\theta}_M\|_2\geq tn^{-1/3}|M|^{1/3}\log^{1/3}p$, we have

$$V(\theta_0)-V(\tilde{\theta}_M)\geq\bar{c}_1t^2n^{-2/3}|M|^{2/3}\log^{2/3}p. \tag{E.8}$$

It follows from the definition of $\tilde{\theta}_M$ that $\widehat{V}(\tilde{\theta}_M)\geq\widehat{V}(\theta_0)$, which together with (E.8) gives

$$\widehat{V}(\tilde{\theta}_M)-V(\tilde{\theta}_M)-\widehat{V}(\theta_0)+V(\theta_0)\geq\bar{c}_1t^2n^{-2/3}|M|^{2/3}\log^{2/3}p,$$

or equivalently

$$m_V(\tilde{\theta}_M) - m_V(\theta_0) \geq \bar{c}_1 t^2 n^{-2/3} |M|^{2/3} \log^{2/3} p.$$

Hence, LHS of (E.5) is bounded by

$$\Pr\left(\bigcap_{M \in \Omega_+^*} \left| m_V(\tilde{\theta}_M) - m_V(\theta_0) \right| \geq \bar{c}_1 t^2 n^{-2/3} |M|^{2/3} \log^{2/3} p \right).$$

For sufficiently large t_0, (E.5) follows using similar arguments in showing (9.21). This completes the proof.

E.2. Uniform convergence of empirical maximizer of $\widehat{C}$. Define

$$\mathcal{A}_0 = \bigcap_{M \in \Omega_+^*} \left\{ \tilde{\beta}_M \in N_{\varepsilon_0} \right\}.$$

Similar to Section E.1, it suffices to show

$$\Pr(\mathcal{A}_0^c) \leq \exp\left(-\frac{\bar{c} n}{\log n} \right), \tag{E.9}$$

and

$$\begin{aligned} \Pr & \left(\bigcap_{M \in \Omega_+^*} \left\{ \|\tilde{\beta}_M - \beta_0\|_2 \geq \frac{t |M|^{1/2} \log^{1/2} p \log^{1/2} n}{\sqrt{n}} \right\} \bigcap \mathcal{A}_0 \right) \\ \leq & \exp\left(-\bar{c} t^2 \log p \right) + \exp\left(-\bar{c} t \sqrt{n \log p} \right), \end{aligned} \tag{E.10}$$

for some constant $\bar{c} > 0$.

E.2.1. *Proof of* (E.9). When $\mathcal{A}_0^c$ holds, it follows from Assumption (A7')(i) that

$$\sup_{M \in \Omega_+^*} \left(C(\beta_0) - C(\tilde{\beta}_M) \right) \geq \xi_0,$$

for some constant $\xi_0 > 0$. By the definition of $\tilde{\beta}_M$, this further implies

$$\sup_{M \in \Omega_+^*} \left(\widehat{C}(\tilde{\beta}_M) - \widehat{C}(\beta_0) - C(\tilde{\beta}_M) + C(\beta_0) \right) \geq \xi_0.$$

Therefore, we have

$$\Pr(\mathcal{A}_0^c) \leq \Pr\left(\sup_{M \in \Omega_+^*} \left(\widehat{C}(\tilde{\beta}_M) - \widehat{C}(\beta_0) - C(\tilde{\beta}_M) + C(\beta_0) \right) \geq \xi_0 \right),$$

the RHS of which can be bounded in a similar manner as (B.4). (E.9) can therefore be proven.

E.2.2. *Proof of* (E.10). When the event defined in the LHS of (E.10) holds, by Assumption (A7')(iii), we have

$$\bigcup_{M\in\Omega_+^*}\left\{\left(C(\beta_0)-C(\tilde{\beta}_M)\right)\geq\frac{\bar{c}_1|M|t^2\log p\log n}{n}\right\}.$$

By the definition of $\tilde{\beta}_M$, this further implies

$$\bigcup_{M\in\Omega_+^*}\left\{\left(\widehat{C}(\tilde{\beta}_M)-\widehat{C}(\beta_0)-C(\tilde{\beta}_M)+C(\beta_0)\right)\geq\frac{\bar{c}_1|M|t^2\log p\log n}{n}\right\}.$$

Hence, LHS of (E.10) is bounded by

(E.11)
$$\Pr\left(\bigcup_{M\in\Omega_+^*}\left\{\left(\widehat{C}(\tilde{\beta}_M)-\widehat{C}(\beta_0)-C(\tilde{\beta}_M)+C(\beta_0)\right)\geq\frac{\bar{c}_1|M|t^2\log p\log n}{n}\right\}\right).$$

Using similar arguments in bounding (B.13), we can show (E.11) is bounded by

$$\begin{aligned}&\frac{1}{2}\exp\left(-\bar{c}t^2\log p\right)+\frac{1}{2}\exp\left(-\bar{c}t\sqrt{n\log p}\right)\\+\;&\frac{1}{2}\exp\left(-\bar{c}n^{1/3}t^{4/3}\log^{2/3}p\right)+\exp\left(\frac{\bar{c}n}{\log n}\right),\end{aligned}$$

for some constant $\bar{c}>0$.

Since $n^{1/3}t^{4/3}\log^{2/3}p=(t^2\log p)^{1/3}(t\sqrt{n\log p})^{2/3}$, we have

$$n^{1/3}t^{4/3}\log^{2/3}p\geq\min(t^2\log p,t\sqrt{n\log p}),$$

and therefore

$$\exp\left(-\bar{c}n^{1/3}t^{4/3}\log^{2/3}p\right)\leq\exp\left(-\bar{c}t^2\log p\right)+\exp\left(-\bar{c}t\sqrt{n\log p}\right).$$

This further implies (E.11) is bounded by

$$\exp\left(-\bar{c}t^2\log p\right)+\exp\left(-\bar{c}t\sqrt{n\log p}\right)+\exp\left(\frac{\bar{c}n}{\log n}\right).$$

The proof is hence completed.

APPENDIX F: PROOF OF THEOREM 10.1

F.1. Consistency of VIC$^{(1)}$. It suffices to show that with probability tending to 1,

$$\text{VIC}^{(1)}(\hat{\theta}_{M_1^V}) > \sup_{M_1 \in \Omega_-^V} \text{VIC}^{(1)}(\hat{\theta}_{M_1}), \tag{F.1}$$

$$\text{VIC}^{(1)}(\hat{\theta}_{M_1^V}) > \sup_{M_1 \in \Omega_+^V} \{n\widehat{V}^{(1)}(\tilde{\theta}_{M_1}) - \kappa_n^{(1)} \|\hat{\beta}_{M_1}\|\}, \tag{F.2}$$

where $\tilde{\theta}_{M_1}$ denotes the empirical maximizer of $\widehat{V}^{(1)}$ on the restricted model space with $\|\tilde{\theta}_{M_1}\|_2 = \|\theta_1^V\|_2$, and

$$\Omega_-^V = \{M_1 \in \Omega_1 : M_1^V \not\subset M_1\} \text{ and } \Omega_+^V = \{M_1 \in \Omega_1 : M_1^V \subsetneq M_1\}.$$

We focus on (F.2). (F.1) can be similarly proven.

Under Condition (C7)(i), we have that for any $(\theta_1^T, \theta_2^T)^T \to \{(\theta_1^V)^T, (\theta_2^V)^T\}^T$,

(F.3)

$$\begin{aligned} V(\theta_1, \theta_2) &= V(\theta_1^V, \theta_2^V) + \frac{1}{2}(\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_2^V)}{\partial (\theta_1^V)^2} (\theta_1 - \theta_1^V)^T \\ &+ \frac{1}{2}(\theta_2 - \theta_2^V)^T \frac{\partial^2 V(\theta_1^V, \theta_2^V)}{\partial (\theta_2^V)^2} (\theta_2 - \theta_2^V)^T + (\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_2^V)}{\partial \theta_1 \partial \theta_2^T} (\theta_2 - \theta_2^V)^T \\ &+ o(\|\theta_1 - \theta_1^V\|_2^2) + o(\|\theta_2 - \theta_2^V\|_2^2). \end{aligned}$$

Note that the first-order terms vanish in (F.3) since $\{(\theta_1^V)^T, (\theta_2^V)^T\}^T$ maximizes $V(\theta_1, \theta_2)$.

Set $\theta_2 = \theta_2^V$ in (F.3), we obtain

$$V(\theta_1, \theta_2^V) - V(\theta_1^V, \theta_2^V) = \frac{1}{2}(\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_2^V)}{\partial (\theta_1^V)^2} (\theta_1 - \theta_1^V)^T + o(\|\theta_1 - \theta_1^V\|_2^2).$$

Combining this together with (F.3), we obtain that for any $(\theta_1^T, \theta_2^T)^T \to \{(\theta_1^V)^T, (\theta_2^V)^T\}^T$,

$$\begin{aligned} V(\theta_1, \theta_2) &= V(\theta_1, \theta_2^V) + \frac{1}{2}(\theta_2 - \theta_2^V)^T \frac{\partial^2 V(\theta_1^V, \theta_2^V)}{\partial (\theta_2^V)^2} (\theta_2 - \theta_2^V)^T \\ &+ (\theta_1 - \theta_1^V)^T \frac{\partial^2 V(\theta_1^V, \theta_2^V)}{\partial \theta_1 \partial \theta_2^T} (\theta_2 - \theta_2^V)^T + o(\|\theta_1 - \theta_1^V\|_2^2) + o(\|\theta_2 - \theta_2^V\|_2^2). \end{aligned}$$

Conditional on the event that $\widehat{M}_2^V = M_{0,2}$ and those defined in (C5), we have for any $\theta_1 \to \theta_1^V$,

(F.4)

$$V(\theta_1, \theta_{\widehat{M}_2^V}) = V(\theta_1, \theta_2^V) + O(R_{n,2}^2) + O(R_{n,2})\|\theta_1 - \theta_1^V\|_2 + o(\|\theta_1 - \theta_1^V\|_2^2).$$

Besides, under Condition (C5), similar to (9.24), we can show that

$$\sup_{\theta_1}\sup_{\|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})} n\left|\widehat{V}(\theta_1,\theta_2)-\widehat{V}(\theta_1,\theta_{0,2})-V(\theta_1,\theta_2)+V(\theta_1,\theta_{0,2})\right|=O_p(\sqrt{nR_{n,2}}),$$

where

$$\widehat{V}(\theta_1,\theta_2)=\frac{1}{n}\sum_{i=1}^{n}\frac{I\{A_i^{(1)}=d_1^{\theta_1}(X_i^{(1)}),A_i^{(2)}=d_2^{\theta_2}(\overline{X}_i^{(2)})\}}{\Pr(A_i^{(1)}=d_1^{\theta_1}(X_i^{(1)})|X_i^{(1)})\Pr(A_i^{(2)}=d_2^{\theta_2}(\overline{X}_i^{(2)})|\overline{X}_i^{(2)})}Y_i,$$

where

$$d_1^{\theta_1}(x^{(1)})=I(\beta_1^Tx^{(1)}>-c_1)\quad\text{and}\quad d_2^{\theta_2}(\bar{x}^{(2)})=I(\beta_2^T\bar{x}^{(2)}>-c_2).$$

Conditional on the event that $\widehat{M}_2^V=M_{0,2}$ and those defined in (C5), this further implies

(F.5)
$$\sup_{\theta_1} n\left|\widehat{V}^{(1)}(\theta_1)-\widehat{V}(\theta_1,\theta_{0,2})-V(\theta_1,\theta_{\widehat{M}_2^V})+V(\theta_1,\theta_{0,2})\right|=O_p(\sqrt{nR_{n,2}}).$$

Based on (F.4) and (F.5), using similar arguments in the proof of Lemma 7.1, we can show that $\tilde{\theta}_{M_1}$ satisfies $\tilde{\theta}_{M_1}=\theta_1^V+O_p(R_{n,2})+O_p(n^{-1/4}R_{n,2}^{1/4})+O_p(n^{-1/3})$. (F.2) can thus be proven using similar arguments in the proof of Theorem 3.3.

F.2. Consistency of CIC$^{(1)}$. It suffices to show with probability tending to 1,

$$\text{CIC}^{(1)}(\hat{\beta}_{M_1^C})>\sup_{M_1\in\Omega_-^C}\text{CIC}^{(1)}(\hat{\beta}_{M_1}),\tag{F.6}$$

$$\text{CIC}^{(1)}(\hat{\beta}_{M_1^C})>\sup_{M_1\in\Omega_+^C}\{n\widehat{C}^{(1)}(\tilde{\beta}_{M_1})-\kappa_n^{(1)}\|\hat{\beta}_{M_1}\|\},\tag{F.7}$$

where $\tilde{\beta}_{M_1}$ denotes the empirical maximizer of $\widehat{C}^{(1)}$ on the restricted model space with $\|\tilde{\beta}_{M_1}\|_2=\|\beta_1^C\|_2$, and

$$\Omega_-^C=\{M_1\in\Omega_1:M_1^C\not\subset M_1\},\qquad\Omega_+^C=\{M_1\in\Omega_1:M_1^C\subsetneq M_1\}.$$

In the following, we focus on proving (F.7). (F.6) can be similarly proven.

For $\theta_2=(c_2,\beta_2^T)^T$, define

$$Y_i^{(1)}(\theta_2)=\left\{\frac{A_i^{(2)}}{\pi_i^{(2)}}I(\beta_2^T\overline{X}_i^{(2)}>-c_2)+\frac{1-A_i^{(2)}}{1-\pi_i^{(2)}}I(\beta_2^T\overline{X}_i^{(2)}\le-c_2)\right\}Y_i,$$

$$y^{(1)}(\theta_2)=\left\{\frac{a^{(2)}}{\pi^{(2)}(\bar{x}^{(2)})}I(\beta_2^T\bar{x}^{(2)}>-c_2)+\frac{1-a^{(2)}}{1-\pi^{(2)}(\bar{x}^{(2)})}I(\beta_2^T\bar{x}^{(2)}\le-c_2)\right\}y.$$

For any $o = (x^{(1)}, a^{(1)}, x^{(2)}, a^{(2)}, y)$, let

$$g(o, \beta_1, \theta_2) = \frac{1}{2}\mathrm{E}\left\{ \frac{(A_0^{(1)} - \pi_0^{(1)})Y_0^{(1)}(\theta_2)}{\pi_0^{(1)}(1-\pi_0^{(1)})} - \frac{(a^{(1)} - \pi^{(1)})y^{(1)}(\theta_2)}{\pi^{(1)}(1-\pi^{(1)})} \right\} I(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)})$$
$$+\frac{1}{2}\mathrm{E}\left\{ \frac{(a^{(1)} - \pi^{(1)})y^{(1)}(\theta_2)}{\pi^{(1)}(1-\pi^{(1)})} - \frac{(A_0^{(1)} - \pi_0^{(1)})Y_0^{(1)}(\theta_2)}{\pi_0^{(1)}(1-\pi_0^{(1)})} \right\} I(\beta_1^T x^{(1)} > \beta_1^T X_0^{(1)}),$$

where we use a shorthand and write $\pi^{(1)} = \pi^{(1)}(x^{(1)})$.

It follows from the maximal inequality for U-process that

$$\sup_{\theta_2}\sup_{\beta_1}\left| \widehat{C}(\beta_1, \theta_2) - \frac{2}{n}\sum_{i=1}^n g(O_i, \beta_1, \theta_2) + C(\beta_1, \theta_2) \right| = O_p\left(\frac{1}{n}\right),$$

where

$$\widehat{C}(\beta_1, \theta_2) = \frac{1}{n(n-1)}\sum_{i\neq j}\left\{\omega_i^{(2)}(\theta_2) - \omega_j^{(2)}(\theta_2)\right\} I(\beta_2^T \overline{X}_i^{(2)} > \beta_2^T \overline{X}_j^{(2)}),$$

$$C(\beta_1, \theta_2) = \mathrm{E}\widehat{C}(\beta_1, \theta_2), \quad \omega_i^{(2)}(\theta_2) = \left(\frac{A_i^{(2)}}{\pi_i^{(2)}} - \frac{1 - A_i^{(2)}}{1 - \pi_i^{(2)}}\right) Y_i^{(1)}(\theta_2).$$

This further implies that

$$\sup_{\beta_1}\left| \widehat{C}^{(1)}(\beta_1) - \frac{2}{n}\sum_{i=1}^n g(O_i, \beta_1, \hat{\theta}_{\widehat{M}_2^C}) + C(\beta_1, \hat{\theta}_{\widehat{M}_2^C}) \right| = O_p\left(\frac{1}{n}\right). \tag{F.8}$$

In the following, we focus on proving

$$\mathrm{E}\sup_{\substack{\|\beta_1 - \beta_1^C\|_2 \le \varepsilon_n \\ \|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})}} \left| \sum_{i=1}^n \{g(O_i, \beta_1, \theta_2) - g(O_i, \beta_1) - g(O_i, \beta_1^C, \theta_2) + g(O_i, \beta_1^C)\} \right|$$
$$= O\left(\sqrt{nR_{n,2}} + nR_{n,2}\varepsilon_n\right), \tag{F.9}$$

for any sequence $\varepsilon_n \to 0$. When this holds, it follows from Jensen's inequality that

$$\sup_{\substack{\|\beta_1 - \beta_1^C\|_2 \le \varepsilon_n \\ \|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})}} \left| C(\beta_1, \theta_2) - C(\beta_1, \theta_2^C) - C(\beta_1^C, \theta_2) + C(\beta_1^C, \theta_2^C) \right|$$
$$= O\left(\sqrt{nR_{n,2}} + nR_{n,2}\varepsilon_n\right). \tag{F.10}$$

Under the event defined in (C5) and that $\widehat{M}_2^C = M_2^C$, it follows from (F.9) and (F.10) that

$$\sup_{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n}\left|\sum_{i=1}^n\{g(O_i,\beta_1,\hat{\theta}_{\widehat{M}_2^C})-g(O_i,\beta_1)-g(O_i,\beta_1^C,\hat{\theta}_{\widehat{M}_2^C})+g(O_i,\beta_1^C)\}\right| = O_p\left(\sqrt{nR_{n,2}}+nR_{n,2}\varepsilon_n\right), \tag{F.11}$$

and

$$\sup_{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n} n\left|C^{(1)}(\beta_1)-C(\beta_1,\theta_2^C)-C^{(1)}(\beta_1^C)+C(\beta_1^C,\theta_2^C)\right| = O_p\left(\sqrt{nR_{n,2}}+nR_{n,2}\varepsilon_n\right). \tag{F.12}$$

Based on (F.8),(F.11) and (F.12), using similar arguments in the proof of Lemma 7.1, we can show that $\tilde{\beta}_{M_1}$ converges at a rate of $O_p(R_{n,2}) + O_p(n^{-1/4}R_{n,2}^{1/4})$ for any $M_1 \in \Omega_+^C$. Similar to the proof of Theorem 3.4, we can show (F.7) holds. It remains to show (F.9).

For any $o = (x^{(1)}, a^{(1)}, x^{(2)}, a^{(2)}, y)$, observe that $\{g(o,\beta_1,\theta_2) - g(o,\beta_1^C,\theta_2) - g(o,\beta_1) + g(o,\beta_1^C)\}$ can be decomposed as

$$g(o,\beta_1,\theta_2)-g(o,\beta_1^C,\theta_2)-g(o,\beta_1)+g(o,\beta_1^C)=\sum_{j=1}^4\{g_j(o,\beta_1,\theta_2)-g_j(o,\beta_1^C,\theta_2)\},$$

where

$$\begin{aligned}
g_1(o,\beta_1,\theta_2) &= \frac{1}{2}\mathrm{E}\frac{(A_0^{(1)}-\pi_0^{(1)})\{Y_0^{(1)}(\theta_2)-Y_0^{(1)}\}}{\pi_0^{(1)}(1-\pi_0^{(1)})}I(\beta_1^T X_0^{(1)} > \beta_1^T x^{(1)}),\\
g_2(o,\beta_1,\theta_2) &= -\frac{1}{2}\frac{(a^{(1)}-\pi^{(1)})\{y^{(1)}(\theta_2)-y^{(1)}\}}{\pi^{(1)}(1-\pi^{(1)})}\phi_1(x^{(1)},\beta_1),\\
g_3(o,\beta_1,\theta_2) &= \frac{1}{2}\frac{(a^{(1)}-\pi^{(1)})\{y^{(1)}(\theta_2)-y^{(1)}\}}{\pi^{(1)}(1-\pi^{(1)})}\phi_2(x^{(1)},\beta_1),\\
g_4(o,\beta_1,\theta_2) &= -\frac{1}{2}\mathrm{E}\frac{(A_0^{(1)}-\pi_0^{(1)})\{Y_0^{(1)}(\theta_2)-Y_0^{(1)}\}}{\pi_0^{(1)}(1-\pi_0^{(1)})}I(\beta_1^T x^{(1)} > \beta_1^T X_0^{(1)}).
\end{aligned}$$

Therefore, it suffices to show for $j = 1, 2, 3, 4$,

$$\mathrm{E}\sup_{\substack{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n\\ \|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})}}\left|\sum_{i=1}^n\{g_j(O_i,\beta_1,\theta_2)-g_j(O_i,\beta_1^C,\theta_2)\}\right| = O\left(\sqrt{nR_{n,2}}+nR_{n,2}\varepsilon_n\right),$$

We first show

$$\mathrm{E} \sup_{\substack{\|\beta_1-\beta_1^C\|_2\le\varepsilon_n \\ \|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})}} \left|\sum_{i=1}^{n}\{g_1(O_i,\beta_1,\theta_2)-g_1(O_i,\beta_1^C,\theta_2)\}\right|$$

$$= O\left(\sqrt{nR_{n,2}}+nR_{n,2}\varepsilon_n\right). \tag{F.13}$$

Note that

$$\begin{aligned}
&\left|\mathrm{E}\{g_1(O_i,\beta_1,\theta_2)-g_1(O_i,\beta_1^C,\theta_2)\}\right| \\
= &\frac{1}{2}\left|\mathrm{E}\left\{\frac{(A_0^{(1)}-\pi_0^{(1)})\{Y_0^{(1)}(\theta_2)-Y_0^{(1)}\}}{\pi_0^{(1)}(1-\pi_0^{(1)})}\{\phi_2(X_0^{(1)},\beta_1)-\phi_2(X_0^{(1)},\beta_1^C)\}\right\}\right| \\
\le &\frac{\varepsilon_n}{2}\mathrm{E}\left|\frac{(A_0^{(1)}-\pi_0^{(1)})Q(\beta_{0,2}^TX_0^{(2)})}{\pi_0^{(1)}(1-\pi_0^{(1)})}\right||I(\beta_2^TX_0^{(2)}>-c_2)-I(\beta_{0,2}^TX_0^{(2)}>-c_{0,2})|\psi_2(X_0^{(2)}) \\
\le &\frac{\varepsilon_n}{2c_1(1-c_2)}\mathrm{E}|Q(\beta_{0,2}^TX_0^{(2)})||I(\beta_2^TX_0^{(2)}>-c_2)-I(\beta_{0,2}^TX_0^{(2)}>-c_{0,2})|\psi_2(X_0^{(2)}) \\
\le &\frac{\varepsilon_n}{2c_1(1-c_2)}\sqrt{\mathrm{E}Q^2(\beta_{0,2}^TX_0^{(2)})|I(\beta_2^TX_0^{(2)}>-c_2)-I(\beta_{0,2}^TX_0^{(2)}>-c_{0,2})|}\sqrt{\mathrm{E}\psi_2^2(X_0^{(2)})} \\
\le &O(\varepsilon_n)\sqrt{\mathrm{E}Q^2(\beta_{0,2}^TX_0^{(2)})|I(\beta_2^TX_0^{(2)}>-c_2)-I(\beta_{0,2}^TX_0^{(2)}>-c_{0,2})|}.
\end{aligned}$$

where the first inequality is due to Condition (C9)(iv), the second inequality is due to Condition (C4), the third inequality is due to Cauchy-Schwarz inequality and the last inequality is due to Condition (C9)(iv).

Now we claim for any $\theta_2=(c_2,\beta_2^T)^T$ such that $\|\theta_2-\theta_{0,2}\|_2=O(R_{n,2})$,

$$\mathrm{E}Q^2(\beta_{0,2}^TX_0^{(2)})|I(\beta_2^TX_0^{(2)}>-c_2)-I(\beta_{0,2}^TX_0^{(2)}>-c_{0,2})|=O(R_{n,2}^2). \tag{F.14}$$

This implies

$$\left|\mathrm{E}\{g_1(O_i,\beta_1,\theta_2)-g_1(O_i,\beta_1^C,\theta_2)\}\right|=O(\varepsilon_nR_{n,2}). \tag{F.15}$$

Note that

$$I(\beta_2^TX_0^{(2)}>-c_2)=I(\beta_2^TX_0^{(2)}+c_2-c_{0,2}>-c_{0,2})=I(Q(\beta_2^TX_0^{(2)}+c_2-c_{0,2})>0),$$

and

$$I(\beta_{0,2}^TX_0^{(2)}>-c_{0,2})=I(Q(\beta_{0,2}^TX_0^{(2)})>0).$$

To prove (F.14), it suffices to show

$$\mathrm{E}Q^2(\beta_{0,2}^T X_0^{(2)})|I(Q(\beta_2^T X_0^{(2)} + c_2 - c_{0,2}) > 0) - I(Q(\beta_{0,2}^T X_0^{(2)}) > 0)| = O(R_{n,2}^2).$$

Note that when $I(Q(\beta_2^T X_0^{(2)} + c_2 - c_{0,2}) > 0) \neq I(Q(\beta_{0,2}^T X_0^{(2)}) > 0)$, we have

$$|Q(\beta_2^T X_0^{(2)} + c_2 - c_{0,2}) - Q(\beta_{0,2}^T X_0^{(2)})| \geq |Q(\beta_{0,2}^T X_0^{(2)})|.$$

By Markov's inequality, This implies

$$\begin{aligned} & \mathrm{E}Q^2(\beta_{0,2}^T X_0^{(2)})|I(Q(\beta_2^T X_0^{(2)} + c_2 - c_{0,2}) > 0) - I(Q(\beta_{0,2}^T X_0^{(2)}) > 0)| \\ \leq\ & \mathrm{E}|Q(\beta_2^T X_0^{(2)} + c_2 - c_{0,2}) - Q(\beta_{0,2}^T X_0^{(2)})|^2 = O(R_{n,2}^2), \end{aligned}$$

for any $\theta_2 = (c_2, \beta_2^T)^T$ such that $\|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})$, where the last equality is due to Condition (C9)(i). This proves (F.14). To show (F.13), in view of (F.15), it suffices to show

$$\begin{aligned} \mathrm{E} \sup_{\substack{\|\beta_1 - \beta_1^C\|_2 \leq \varepsilon_n \\ \|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})}} & \left| \sum_{i=1}^n \{g_1(O_i, \beta_1, \theta_2) - g_1(O_i, \beta_1^C, \theta_2) \right. \\ & \left. - \mathrm{E}g_1(O_i, \beta_1, \theta_2) + \mathrm{E}g_1(O_i, \beta_1^C, \theta_2)\} \right| = O\left(\sqrt{nR_{n,2}}\right), \end{aligned}$$

or

$$\mathrm{E} \sup_{\beta_1} \sup_{\|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})} \left| \sum_{i=1}^n \{g_1(O_i, \beta_1, \theta_2) - \mathrm{E}g_1(O_i, \beta_1, \theta_2)\} \right| = O\left(\sqrt{nR_{n,2}}\right).$$

This can be proven in a similar manner as (B.9).

Similarly, we can show that for $j = 2, 3, 4$,

$$\begin{aligned} \mathrm{E} \sup_{\substack{\|\beta_1 - \beta_1^C\|_2 \leq \varepsilon_n \\ \|\theta_2 - \theta_{0,2}\|_2 = O(R_{n,2})}} & \left| \sum_{i=1}^n \{g_j(O_i, \beta_1, \theta_2) - g_j(O_i, \beta_1^C, \theta_2)\} \right| \\ & = O\left(\sqrt{nR_{n,2}} + nR_{n,2}\varepsilon_n\right). \end{aligned}$$

This proves (F.9). The proof is thus completed.

APPENDIX G: TECHNICAL LEMMAS AND DEFINITIONS

DEFINITION G.1. *For any random variable Y, define the Orlicz norm $\|Y\|_{\psi_p}$ as*

$$\|Y\|_{\psi_p} \triangleq \inf_{C>0} \left\{ E\exp\left(\frac{|Y|^p}{C^p}\right) \leq 2 \right\}.$$

LEMMA G.1. *For any random variable X, if $\omega = \|X\|_{\psi_1} < \infty$, then for any integer $p \geq 1$, $E|X|^p \leq p!\omega^p$.*

Proof: It follows by the definition of Orlicz norm that

$$1 + \frac{\mathrm{E}|X|^p}{\omega^p} \leq \mathrm{E} \exp\left(\frac{|X|}{\omega}\right) \leq 2.$$

The assertion thus follows.

LEMMA G.2 (Bernstein inequality). *Let $X_1, \ldots, X_n$ be independent mean zero random variables, if $\omega = \max_i \|X_i\|_{\psi_1} < \infty$, there exists some constant c such that for all $t > 0$,*

$$Pr\left(|\sum_{i=1}^{n} X_i| > t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{n\omega^2}, \frac{t}{\omega}\right)\right).$$

Proof: See Theorem 3.1 in Klartag and Mendelson (2005).

LEMMA G.3. *Let $\mathcal{F}$ and $\mathcal{G}$ be class of measurable functions $S \to \mathbb{R}$, to which measurable envelopes F and G are attached, respectively. Assume there exists some constant K and $v \geq 1$ such that*

$$\sup_Q N(\varepsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq (K/\varepsilon)^v, \quad \sup_Q N(\varepsilon\|G\|_{Q,2}, \mathcal{G}, L_2(Q)) \leq (K/\varepsilon)^v,$$

for all $0 < \varepsilon \leq 1$. Denoted by $\mathcal{F} + \mathcal{G}$ the pointwise sum of $\mathcal{F}$ and $\mathcal{G}$. Then

$$\sup_Q N(\varepsilon\|\sqrt{2F^2 + 2G^2}\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq (K/\varepsilon)^{2v}.$$

Proof: For any $f_1, f_2 \in \mathcal{F}$ and $g_1, g_2 \in \mathcal{G}$, by Cauchy-Schwarz inequality, we have

$$|f_1 + g_1 - f_2 - g_2|^2 \leq 2|f_1 - f_2|_2^2 + 2|g_1 - g_2|_2^2.$$

The assertion then follows from application of Lemma A.6 in Chernozhukov, Chetverikov and Kato (2014).

LEMMA G.4. *Let $X_1, \ldots, X_n$ be i.i.d random variables with values in a measurable space $(\mathcal{S}, \mathcal{B})$, and let $\mathcal{F}$ be a countable class of measurable functions $f : \mathcal{S} \to \mathbb{R}$. Assume $Ef(X_i) = 0$ for all f, and $\omega = \|\sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_1} <$*

∞. *Then for all* $0 < \eta < 1$ *and* $\delta > 0$, *there exists some constant* $C = C(\eta, \delta)$ *such that for all* $t \geq 0$,

$$Pr\left(\sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} f(X_i)| \geq (1+\eta)E\|\sum_{i=1}^{n} f(X_i)\|_{\mathcal{F}} + t\right)$$
$$\leq \exp\left(-\frac{t^2}{2(1+\delta)n\sigma^2}\right) + 3\exp\left(-\frac{t}{C\omega\log(n)}\right),$$

where $\sigma^2 = \sup_{f\in\mathcal{F}} Ef^2(X_i)$.

Proof: It follows by Lemma 2.2.2 in van der Vaart and Wellner (1996) that

$$\|\max_i \sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_1} \leq K\log(n) \max_i \|\sup_{f\in\mathcal{F}} |f(X_i)|\|_{\psi_1},$$

for some constant K. The assertion follows by Theorem 4 in Adamczak (2008).

LEMMA G.5. *Assume* $X_1, \ldots, X_n$ *are i.i.d random variables. Assume function* f *is symmetric and satisfies* $Ef(X_i, x) = Ef(x, X_i) = 0$, $f(x, x) = 0, \forall x$, $\sup_f |f(x, y)| \leq F, \forall x, y$. *Define the following degenerate U-process,*

$$Z = \sup_{f\in\mathcal{F}} |\sum_{i\neq j} f(X_i, X_j)|$$

Let $\varepsilon_1, \ldots, \varepsilon_n$ *be i.i.d Rademacher random variables independent of* $\{X_1, \ldots, X_n\}$, *and introduce the random variables:*

$$Z_\varepsilon = \sup_{f\in\mathcal{F}} \left|\sum_{i,j} \varepsilon_i\varepsilon_j f(X_i, X_j)\right|,$$
$$U_\varepsilon = \sup_{f\in\mathcal{F}} \sup_{\alpha:\|\alpha\|_2\leq 1} \sum_{i,j} \varepsilon_i\alpha_j f(X_i, X_j),$$
$$M_\varepsilon = \sup_{f\in\mathcal{F}} \sup_{k=1,\ldots,n} \left|\sum_i \varepsilon_i f(X_i, X_k)\right|.$$

Then there exists some constants $C > 0$ *such that for all* n *and* $t > 0$,

$$Pr(Z > CEZ_\varepsilon + t)$$
$$\leq \exp\left(-\frac{1}{C}\min\left(\frac{t^2}{(EU_\varepsilon)^2}, \frac{t}{EM_\varepsilon}, \frac{t}{nF}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right). \tag{G.1}$$

Proof: It follows by Theorem 11 in Clémençon, Lugosi and Vayatis (2008) that (G.1) is bounded by

$$\text{(G.2)}\ \exp\left(-\frac{1}{C'}\min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2}, \frac{t}{nF+\mathrm{E}M_\varepsilon}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right),$$

for some constant C'. Note that (G.2) is smaller than

$$\exp\left(-\frac{1}{C'}\min\left(\frac{t^2}{(\mathrm{E}U_\varepsilon)^2}, \frac{t}{2nF}, \frac{t}{2\mathrm{E}M_\varepsilon}, \left(\frac{t}{F\sqrt{n}}\right)^{2/3}, \sqrt{\frac{t}{F}}\right)\right).$$

Lemma G.5 thus holds by setting $C = 2C'$.

REFERENCES

Adamczak, R. (2008). A tail inequality for suprema of unbounded empirical processes with applications to Markov chains. *Electron. J. Probab.* **13** no. 34, 1000–1034.

Chernozhukov, V., Chetverikov, D. and Kato, K. (2014). Gaussian approximation of suprema of empirical processes. *Ann. Statist.* **42** 1564–1597.

Clémençon, S., Lugosi, G. and Vayatis, N. (2008). Ranking and empirical minimization of U-statistics. *Ann. Statist.* **36** 844–874.

Fan, C., Lu, W., Song, R. and Zhou, Y. (2016). Concondance-Assisted learning for estimating optimal individualized treatment regimes. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* n/a–n/a.

Klartag, B. and Mendelson, S. (2005). Empirical processes and random projections. *J. Funct. Anal.* **225** 229–245.

Ledoux, M. and Talagrand, M. (2011). *Probability in Banach spaces. Classics in Mathematics*. Springer-Verlag, Berlin Isoperimetry and processes, Reprint of the 1991 edition.

Luedtke, A. R. and van der Laan, M. J. (2016). Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Ann. Statist.* **44** 713–742.

Nolan, D. and Pollard, D. (1987). U-processes: rates of convergence. *Ann. Statist.* **15** 780–799.

Qian, M. and Murphy, S. A. (2011). Performance guarantees for individualized treatment rules. *Ann. Statist.* **39** 1180–1210.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak convergence and empirical processes. Springer Series in Statistics*. Springer-Verlag, New York With applications to statistics.

Chengchun Shi
Department of Statistics,
North Carolina State University,
Raleigh NC, U.S.A.
E-mail: cshi4@ncsu.edu

Rui Song
Department of Statistics,
North Carolina State University,
Raleigh NC, U.S.A.
E-mail: rsong@ncsu.edu

Wenbin Lu
Department of Statistics,
North Carolina State University,
Raleigh NC, U.S.A.
E-mail: wlu4@ncsu.edu