Supplementary Material for "Integrative R-learner of heterogeneous treatment effects combining experimental and observational studies"

Outline of the appendices

- Appendix A: proof of Theorem 1.
- Appendix B: proof of Proposition 2.
- Appendix C: pseudocode of iterative learning algorithm used in the real data experiment.

Appendix A. Proof of Theorem 1

To prove Theorem 1, we show a useful lemma first.

Lemma 3 *Under Assumptions 3-4,* $\widehat{L}_N(\beta) = L_N(\beta) + O_P(a_N^2)$.

Proof For the simplicity of the notations, we denote

$$A_{\mu,i} = \mu(X_i, S_i) - \hat{\mu}^{-k(i)}(X_i, S_i),\ B_{\mu,i} = Y_i - \mu(X_i, S_i),$$
$$A_{e,i} = e(X_i, S_i) - \hat{e}^{-k(i)}(X_i, S_i),\ B_{e,i} = A_i - e(X_i, S_i).$$

By algebra, we have

$$\begin{aligned}
\widehat{L}_N(\beta) =& \mathbb{P}_N\{B_{\mu,i} + A_{\mu,i} - g(X_i, S_i; \beta)B_{e,i} - g(X_i, S_i; \beta)A_{e,i}\}^2 \\
=& \mathbb{P}_N\{B_{\mu,i} - g(X_i, S_i; \beta)B_{e,i}\}^2 + \mathbb{P}_N\{A_{\mu,i} - g(X_i, S_i; \beta)A_{e,i}\}^2 + \\
& 2\mathbb{P}_N\{B_{\mu,i} - g(X_i, S_i; \beta)B_{e,i}\}\mathbb{P}_N\{A_{\mu,i} - g(X_i, S_i; \beta)A_{e,i}\} \\
=& L_N(\beta) + \mathbb{P}_N A_{\mu,i}^2 + \mathbb{P}_N A_{e,i}^2 g^2(X_i, S_i) - 2\mathbb{P}_N A_{\mu,i}A_{e,i}g(X_i, S_i; \beta) + \\
& 2\mathbb{P}_N B_{\mu,i}A_{\mu,i} - 2\mathbb{P}_N B_{\mu,i}A_{e,i}g(X_i, S_i; \beta) - 2\mathbb{P}_N B_{e,i}A_{\mu,i}g(X_i, S_i; \beta) + 2\mathbb{P}_N B_{e,i}A_{e,i}g^2(X_i, S_i).
\end{aligned}$$

By Markov's inequality and Assumption 4, we have the second term above $\mathbb{P}_N A_{\mu,i}^2$ is $O_P(a_N^2)$. Plus Assumption 3, we have the third term $\mathbb{P}_N A_{e,i}^2$ is $O_P(a_N^2)$. As for the fourth term,

$$\mathbb{P}_N A_{\mu,i}A_{e,i}g(X_i, S_i; \beta) \le C_1 \mathbb{P}_N A_{\mu,i}A_{e,i} \le C_1\sqrt{\mathbb{P}_N A_{\mu,i}^2 \mathbb{P}_N A_{e,i}^2} = O_P(a_N^2),$$

for some positive constant C_1, and the last inequality is by Cauchy-Schwarz inequality.

Next, to deal with the last four terms, we tackle $\mathbb{P}_N B_{\mu,i}A_{\mu,i}$ first. Let

$$B_{\mu\mu}^k = \frac{\sum_{i:k(i)=k} B_{\mu,i}A_{\mu,i}}{|\{i : k(i) = k\}|},$$

and note that $|\mathbb{P}_N B_{\mu,i}A_{e,i}| \le \sum_{k=1}^K |B_{\mu\mu}^k|$, where K is finite, thus it is suffice to show $B_{\mu\mu}^k = O_P(a_N^2)$. Let $\mathcal{I}^{-k} = \{X_i, A_i, Y_i, S_i : k(i) \neq k\}$. Then we have

$$\begin{aligned}
\mathbb{E}(B_{\mu\mu}^k) &= \mathbb{E}(B_{\mu,i}A_{e,i}) = \mathbb{E}\{\mathbb{E}(B_{\mu,i}A_{e,i} \mid \mathcal{I}^{-k}, X_i, S_i)\} \\
&= \mathbb{E}\{A_{e,i}\mathbb{E}(B_{\mu,i} \mid \mathcal{I}^{-k}, X_i, S_i)\} = 0.
\end{aligned}$$

Then we have its variance

$$\text{var}(B^k_{\mu\mu}) = \mathbb{E}\{(B^k_{\mu\mu})^2\} = \frac{\mathbb{E}\{\sum_{i:k(i)=k} B^2_{\mu,i}A^2_{\mu,i} + \sum_{i\neq j:k(i)=k,k(j)=k} B_{\mu,i}B_{\mu,j}A_{\mu,i}A_{\mu,j}\}}{|\{i:k(i)=k\}|^2}$$
$$= \frac{\mathbb{E}(B^2_{\mu,i}A^2_{\mu,i})}{|\{i:k(i)=k\}|} + \frac{\sum_{i\neq j:k(i)=k,k(j)=k}\mathbb{E}(B_{\mu,i}B_{\mu,j}A_{\mu,i}A_{\mu,j})}{|\{i:k(i)=k\}|^2}.$$

By Assumption 3-4 we have $\mathbb{E}(B^2_{\mu,i}A^2_{e,i}) = \mathbb{E}\{\mathbb{E}(B^2_{\mu,i}A^2_{e,i} \mid \mathcal{I}^{-k}, X_i, S_i)\} = \mathbb{E}\{A^2_{e,i}\mathbb{E}(B^2_{\mu,i} \mid \mathcal{I}^{-k}, X_i, S_i)\} \leq C_2\mathbb{E}A^2_{e,i} = O(a^2_N)$, for some positive constant C_2. As for the interaction terms, we have $\mathbb{E}(B_{\mu,i}B_{\mu,j}A_{\mu,i}A_{\mu,j}) = \mathbb{E}\left\{A_{\mu,i}A_{\mu,j}\mathbb{E}(B_{\mu,i}B_{\mu,j} \mid \mathcal{I}^{-k}, X_i, S_i)\right\} = \mathbb{E}\left\{A_{\mu,i}A_{\mu,j}\mathbb{E}(B_{\mu,j})\mathbb{E}(B_{\mu,i} \mid \mathcal{I}^{-k}, X_i, S_i)\right\} = 0$. The second last equality is implied by $B_{\mu,i}$ and $B_{\mu,j}$ are independent for $i \neq j$, and the last equality comes from the definition of $B_{\mu,i}$. Therefore, we have $\text{var}(B^k_{\mu\mu}) = (K/N)O(a^2_N) = O(a^2_N/N)$, which is negligible with a faster diminishing rate than $O(a^2_N)$. Then, by Chebyshev' inequality, we have $B^k_{\mu\mu} = O_P(a^2_N/N)$, i.e., $\mathbb{P}_N B_{\mu,i}A_{\mu,i} = O_P(a^2_N/N)$. Similarly, under the Assumption 3 that $g(X_i, S_i; \beta)$ is uniformly bounded, we can get the same results for the left three terms equal to $O_P(a^2_N/N)$. Therefore, $\widehat{L}_N(\beta) - L_N(\beta)$ is dominated by the $O_P(a^2_N)$-term $\mathbb{P}_N A^2_{\mu,i} + \mathbb{P}_N A^2_{e,i}g^2(X_i, S_i) - 2\mathbb{P}_N A_{\mu,i}A_{e,i}g(X_i, S_i; \beta)$, which finally leads to the conclusion in the lemma $\widehat{L}_N(\beta) - L_N(\beta) = O_P(a^2_N)$. ■

Next, we are showing the proof of Theorem 1 with the help of Lemma 3.

Proof of Theorem 1:

First, we let $\mathbb{G}_N = \sqrt{N}(\mathbb{P}_N - \mathbb{P})$, $p^\tau = \{A - e(X, A)\}p_\tau(X)$, and $p^c = \{A - e(X, A)\}(1 - S)p_c(X)$. Recall $p_i^\intercal = \{(p^\tau)^\intercal, (p^c)^\intercal\}$, and then

$$\Gamma = \mathbb{P}(p_i p_i^\intercal) = \begin{Bmatrix} p^\tau(p^\tau)^\intercal & p^\tau(p^c)^\intercal \\ p^c(p^\tau)^\intercal & p^c(p^c)^\intercal \end{Bmatrix} \overset{\text{denoted as}}{=} \begin{pmatrix} \Gamma_{\tau\tau} & \Gamma_{\tau c} \\ \Gamma_{c\tau} & \Gamma_{cc} \end{pmatrix}.$$

By the definition of $\hat{\beta}$ and Lemma 3, we have

$$\hat{\beta} = \underset{b}{\text{argmin}}\, \mathbb{P}_N \left\{Y_i - \mu(X_i, S_i) - p_i^\intercal b\right\}^2 + O_P(a^2_N).$$

By the Pointwise Linearizaiton of the series method (see Lemma 4.1 in Belloni et al., 2015), under Assumptions 5-7 and the fact $\mathbb{E}(\epsilon \mid X, A, S) = 0$ which is shown under Assumptions 1-2 in Section 3, we have for any $\alpha := (\alpha_\tau^\intercal, \alpha_c^\intercal)^\intercal \in \mathbb{R}^{2D}$,

$$\sqrt{N}\alpha^\intercal(\hat{\beta} - \beta) = \alpha^\intercal \begin{pmatrix} \Gamma_{\tau\tau} & \Gamma_{\tau c} \\ \Gamma_{c\tau} & \Gamma_{cc} \end{pmatrix}^{-1} \mathbb{G}_N \begin{pmatrix} p^\tau\epsilon \\ p^c\epsilon \end{pmatrix} + o_P(1) + O_P\left(\sqrt{N}a^2_N\right). \tag{12}$$

Let

$$\Sigma = \begin{pmatrix} \Sigma_{\tau\tau} & \Sigma_{\tau c} \\ \Sigma_{c\tau} & \Sigma_{cc} \end{pmatrix} := \begin{pmatrix} \Gamma_{\tau\tau} & \Gamma_{\tau c} \\ \Gamma_{c\tau} & \Gamma_{cc} \end{pmatrix}^{-1}.$$

Since under Assumption 4, $a_N = O(N^{-r}), r > 1/4$, thus $O_P\left(\sqrt{N}a^2_N\right)$ is negligible compared to $o_p(1)$. Therefore, we have

$$\sqrt{N}\alpha_\tau^\intercal(\hat{\beta}_\tau - \beta_\tau) = \alpha_\tau^\intercal\mathbb{G}_N\left(\Sigma_{\tau\tau}p^\tau\epsilon + \Sigma_{\tau c}p^c\epsilon\right) + o_P(1). \tag{13}$$

Under Assumptions 5-7 and the fact $\mathbb{E}(\epsilon \mid X, A, S) = 0$, by the Pointwise Normality of the series method (see Theorem 4.2 in Belloni et al., 2015), we have

$$\sqrt{N}\frac{\alpha_\tau^\intercal(\hat{\beta}_\tau - \beta_\tau)}{\|\alpha_\tau^\intercal \Omega^{1/2}\|} \xrightarrow{d} \mathcal{N}(0,1) + o_P(1),$$

where $\Omega = \mathbb{E}\left\{(\Sigma_{\tau\tau}p^\tau\epsilon + \Sigma_{\tau c}p^c\epsilon)\,(\Sigma_{\tau\tau}p^\tau\epsilon + \Sigma_{\tau c}p^c\epsilon)^\intercal\right\}$. Then, we take $\alpha_\tau = p^\tau$, for any $x \in \mathcal{X}$,

$$\sqrt{N}\frac{(p^\tau)^\intercal(\hat{\beta}_\tau - \beta_\tau)}{\|(p^\tau)^\intercal \Omega^{1/2}\|} \xrightarrow{d} \mathcal{N}(0,1) + o_P(1).$$

Under Assumption 6*(d)*, the approximation error is negligible relative to the estimation error, then

$$\sqrt{N}\frac{\hat{\tau}(x) - \tau(x)}{\|(p^\tau)^\intercal \Omega^{1/2}\|} \xrightarrow{d} \mathcal{N}(0,1) + o_P(1),$$

which immediately arrives at the first part conclusion in Theorem 1, $\hat{\tau}(x) - \tau(x) = O(N^{-1/2})$, for any $x \in \mathcal{X}$. Besides, it also gives the asymptotic variance of $\hat{\tau}(x)$,

$$\mathbb{V}\{\hat{\tau}(x)\} = N^{-1}(p^\tau)^\intercal \Omega p^\tau. \tag{14}$$

Expanding Ω, under Assumption 7, we have

$$\begin{aligned}
\Omega =& \Sigma_{\tau\tau}\mathbb{E}\left\{p^\tau(p^\tau)^\intercal\epsilon^2\right\}\Sigma_{\tau\tau} + \Sigma_{\tau c}\mathbb{E}\left\{p^\tau(p^c)^\intercal\epsilon^2\right\}\Sigma_{\tau c} + \\
& \Sigma_{\tau c}\mathbb{E}\left\{p^c(p^c)^\intercal\epsilon^2\right\}\Sigma_{\tau c} + \Sigma_{\tau\tau}\mathbb{E}\left\{p^\tau(p^c)^\intercal\epsilon^2\right\}\Sigma_{\tau c} \\
=& \left(\Sigma_{\tau\tau}\Gamma_{\tau\tau} + \Sigma_{\tau c}\Gamma_{\tau c}^\intercal\right)\Sigma_{\tau\tau}\sigma^2 + \left(\Sigma_{\tau c}\Gamma_{cc} + \Sigma_{\tau\tau}\Gamma_{\tau c}\right)\Sigma_{\tau c}^\intercal\sigma^2 \\
=& \Sigma_{\tau\tau}\sigma^2 + \mathbf{0}_{D\times D} \\
=& \left(\Gamma_{\tau\tau} - \Gamma_{\tau c}\Gamma_{cc}^{-1}\Gamma_{\tau c}^\intercal\right)^{-1}\sigma^2 \\
=& \left[\mathbb{E}\{Sp^\tau(p^\tau)^\intercal\} + \mathbb{E}\{(1-S)p^\tau(p^\tau)^\intercal\} - \Gamma_{\tau c}\Gamma_{cc}^{-1}\Gamma_{\tau c}^\intercal\right]^{-1}\sigma^2.
\end{aligned} \tag{15}$$

Next, we aim to obtain the asymptotic variance of the HTE estimator $\hat{\tau}_{\text{rct}} = (p^\tau)^\intercal\hat{\beta}_{\text{rct}}$ with only the RCT, where $\hat{\beta}_{\text{rct}} = \text{argmin}_{b\in\mathbb{R}^D}\,\mathbb{P}_N S_i\left[Y_i - \hat{\mu}^{-k(i)}(X_i,S_i) - \{A_i - \hat{e}^{-k(i)}(X_i,S_i)\}p_\tau^\intercal(X_i)b\right]^2$. Similarly, we can replace the estimated nuisance functions with the true ones based on Lemma 3, and following the same strategy as the integrative R-learner above to obtain the asymptotic variance, $\mathbb{V}(\hat{\tau}_{\text{rct}}) = N^{-1}(p^\tau)^\intercal\Omega_{\text{rct}}p^\tau$, where

$$\Omega_{\text{rct}} = \left[\mathbb{E}\left\{Sp^\tau(p^\tau)^\intercal\right\}\right]^{-1}\sigma^2. \tag{16}$$

By Hölder's inequality, $\mathbb{E}\{(1-S)p^\tau(p^\tau)^\intercal\} - \Gamma_{\tau c}\Gamma_{cc}^{-1}\Gamma_{\tau c}^\intercal$ is non-negative definitive; i.e., for any $v \in \mathbb{R}^D$,

$$v^\intercal(\Omega^{-1} - \Omega_{\text{rct}}^{-1})v \geq 0, \tag{17}$$

where the inequality becomes an equality if and only if $p_\tau(X) = Mp_c(X)$ for some constant matrix M. From (14), we have

$$p^\tau\mathbb{V}\{\hat{\tau}(x)\}(p^\tau)^\intercal = N^{-1}p^\tau(p^\tau)^\intercal\Omega p^\tau(p^\tau)^\intercal$$

$$
\begin{aligned}
&\Longrightarrow \mathbf{I}_D \mathbb{V}\{\hat{\tau}(x)\} = N^{-1} p^\tau (p^\tau)^\intercal \Omega \\
&\Longrightarrow \mathbf{I}_D \mathbb{V}^{-1}\{\hat{\tau}(x)\} = N\Omega^{-1} \left\{p^\tau (p^\tau)^\intercal\right\}^{-1} \\
&\Longrightarrow (p^\tau)^\intercal \left\{p^\tau (p^\tau)^\intercal\right\}^{-1} \mathbb{V}^{-1}\{\hat{\tau}(x)\} p^\tau = N (p^\tau)^\intercal \left\{p^\tau (p^\tau)^\intercal\right\}^{-1} \Omega^{-1} \left\{p^\tau (p^\tau)^\intercal\right\}^{-1} p^\tau \\
&\overset{\text{by (17)}}{\Longrightarrow} (p^\tau)^\intercal \left\{p^\tau (p^\tau)^\intercal\right\}^{-1} \left[\mathbb{V}^{-1}\{\hat{\tau}(x)\} - \mathbb{V}^{-1}\{\hat{\tau}_{\text{rct}}(x)\}\right] p^\tau \geq 0 \\
&\qquad \text{(Multiply a positive number } (p^\tau)^\intercal p^\tau \text{ on the both sides and get the below formula)} \\
&\Longrightarrow (p^\tau)^\intercal p^\tau (p^\tau)^\intercal \left\{p^\tau (p^\tau)^\intercal\right\}^{-1} p^\tau \left[\mathbb{V}^{-1}\{\hat{\tau}(x)\} - \mathbb{V}^{-1}\{\hat{\tau}_{\text{rct}}(x)\}\right] \geq 0 \\
&\Longrightarrow (p^\tau)^\intercal p^\tau \left[\mathbb{V}^{-1}\{\hat{\tau}(x)\} - \mathbb{V}^{-1}\{\hat{\tau}_{\text{rct}}(x)\}\right] \geq 0 \\
&\Longrightarrow \mathbb{V}^{-1}\{\hat{\tau}(x)\} - \mathbb{V}^{-1}\{\hat{\tau}_{\text{rct}}(x)\} \geq 0 \\
&\Longrightarrow \mathbb{V}\{\hat{\tau}(x)\} \leq \mathbb{V}\{\hat{\tau}_{\text{rct}}(x)\},
\end{aligned}
$$

with the equality holding when $p_\tau(X) = M p_c(X)$ for some constant matrix M. ■

Appendix B. Proof of Proposition 2

Proof The proof of (10) is mainly based on the inverse probability weights (IPW) component, thus we tackle it first.

a) IPW-adjusted outcomes: We have

$$
\begin{aligned}
&\mathbb{E}\left\{\frac{AY}{e(X,S)} \mid X, S=0\right\} \\
=&\frac{1}{e(X,0)}[\mathbb{E}\{AY \mid X, A=1, S=0\} e(X,0) + \mathbb{E}\{AY \mid X, A=0, S=0\}\{1-e(X,0)\}] \\
=&\mathbb{E}\{Y \mid X, A=1, S=0\}.
\end{aligned}
$$

Similarly, we have

$$
\mathbb{E}\left\{\frac{(1-A)Y}{1-e(X,S)} \mid X, S=0\right\} = \mathbb{E}\{Y \mid X, A=0, S=0\}.
$$

b) Augmented IPW-adjusted outcomes: Then we have

$$
\begin{aligned}
&\mathbb{E}\left[\frac{A\{Y - Q_S(X,1)\}}{e(X,S)} + Q_S(X,1) \mid X, S=0\right] \\
=&\mathbb{E}\left\{\frac{AY}{e(X,S)} \mid X, S=0\right\} + Q_0(X,1)\mathbb{E}\left[\left\{1 - \frac{A}{e(X,S)}\right\} \mid X, S=0\right] \\
=&\mathbb{E}(Y \mid X, A=1, S=0).
\end{aligned}
$$

Similarly, we have

$$
\mathbb{E}\left[\frac{(1-A)\{Y - Q_S(X,0)\}}{1-e(X,S)} + Q_S(X,0) \mid X, S=0\right] = \mathbb{E}(Y \mid X, A=0, S=0).
$$

Finally, by taking the difference of the above two formulas and based on the definition of $c(X)$ in (1), we arrive at the conclusion $\mathbb{E}(\widetilde{Y} \mid X, S=0) = \tau(X) + c(X)$ ■

Appendix C. Iterative learning for the integrative R-learner in the real data experiment

Algorithm 2 Iterative learning for the integrative R-learner

Data: The RCT and the OS data $\{(X_i, A_i, Y_i, S_i)\}_{i\in\mathcal{I}_n\cup\mathcal{I}_m}$; the number of iterations B

Result: $\tau(\cdot)$

Initialize confounding function $c(X_i)$.

Estimate the nuisance functions with cross-fitting based on Xgboost, denoted as $\hat{\mu}(X_i, S_i)$ and $\hat{e}(X_i.S_i)$, resulting in $\hat{\epsilon}_{Y_i} = Y_i - \hat{\mu}(X_i, S_i)$ and $\hat{\epsilon}_{A_i} = A_i - \hat{e}(X_i, S_i)$.

for $b = 1, \ldots, B$ **do**

Calculate the pseudo outcomes $\widetilde{Y}_i \leftarrow \hat{\epsilon}_{Y_i}/\hat{\epsilon}_{A_i} - (1 - S_i)c(X_i), i \in \mathcal{I}_n \cup \mathcal{I}_m$;

Obtain $\tau(\cdot)$ by fitting $\widetilde{Y}_i$ on X_i using weighted random forest with weights equal to $\hat{\epsilon}^2_{A_i}, i \in \mathcal{I}_n \cup \mathcal{I}_m$;

Update the pseudo outcomes $\widetilde{Y}_i \leftarrow \hat{\epsilon}_{Y_i}/\hat{\epsilon}_{A_i} - \tau(X_i), i \in \mathcal{I}_m$;

Update $c(X_i)$ by fitting $\widetilde{Y}_i$ on X_i using weighted random forest with weights equal to $\hat{\epsilon}^2_{A_i}, i \in \mathcal{I}_m$.

end