PNAS
www.pnas.org

Supplementary Information for

Sensitivity analysis of individual treatment effects: a robust conformal inference approach

Ying Jin, Zhimei Ren, Emmanuel J. Candès

Emmanuel J. Candès.
E-mail: candes@stanford.edu

This PDF file includes:

1. Tables and figures

Counterfactual	Bound	ATE-type	ATT-type	ATC-type	General
$Y(1)$	$\ell(x)$	$p_1 \cdot \left(1 + \frac{1}{\Gamma \cdot r(x)}\right)$	1	$\frac{p_1}{p_0} \cdot \left(\frac{1}{\Gamma \cdot r(x)}\right)$	$p_1 \cdot \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x) \cdot \left(1 + \frac{1}{\Gamma \cdot r(x)}\right)$
	$u(x)$	$p_1 \cdot \left(1 + \frac{\Gamma}{r(x)}\right)$	1	$\frac{p_1}{p_0} \cdot \frac{\Gamma}{r(x)}$	$p_1 \cdot \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x) \cdot \left(1 + \frac{\Gamma}{r(x)}\right)$
$Y(0)$	$\ell(x)$	$p_0 \cdot \left(1 + \frac{r(x)}{\Gamma}\right)$	$\frac{p_0}{p_1} \cdot \frac{r(x)}{\Gamma}$	1	$p_0 \cdot \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x) \cdot \left(1 + \frac{r(x)}{\Gamma}\right)$
	$u(x)$	$p_0 \cdot \left(1 + \Gamma \cdot r(x)\right)$	$\frac{p_0}{p_1} \cdot \Gamma \cdot r(x)$	1	$p_0 \cdot \frac{\mathrm{d}\mathbb{Q}_X}{\mathrm{d}\mathbb{P}_X}(x) \cdot \left(1 + \Gamma \cdot r(x)\right)$

Table S1. Summary of the upper and lower bounds of the likelihood ratio for different inferential targets. For $t \in \{0,1\}$, $p_t = \mathbb{P}(T = t)$ and $r(x) = e(x)/(1-e(x))$ is the odds ratio of the propensity score. The training distribution for $Y(t)$ is always $\mathbb{P}_{X,Y(t)\,|\,T=t}$. For target distributions, ATE-type refers to $\mathbb{P}_{X,Y(t)}$; ATT-type refers to $\mathbb{P}_{X,Y(t)\,|\,T=1}$; ATC-type refers to $\mathbb{P}_{X,Y(t)\,|\,T=0}$; General refers to $\mathbb{Q}_X \times \mathbb{P}_{Y(t)\,|\,X}$.

2. Additional results

A. Constructing $\hat{G}_n(\cdot)$ in the PAC-type procedure. In this part, we provide the construction of $\hat{G}_n(t)$ in Section A with Waudby-Smith–Ramdas bound (1). The proof is a slight modification of (1) and included in SI Appendix D for completeness.

Proposition 2.1 (Waudby-SmithRamdas lower confidence bound for c.d.f.s)**.** *Suppose $\sup_x \hat{u}(x) \le M$ for some constant $M > 0$. For $t \in \mathbb{R}$ being any constant or any random variable in $\sigma(\mathcal{D}_{\text{train}})$, and any $\delta \in (0,1)$, we define $\hat{G}_n(t) = \max\left\{\hat{G}_n^L(t), \hat{G}_n^U(t)\right\}$, where*

$$\hat{G}_n^L(t) = M \cdot \inf\left\{g \ge 0 \colon \max_{1\le i\le n} \mathcal{K}_i^L(g) \le 2/\delta\right\},$$
$$\hat{G}_n^U(t) = 1 - M + M \cdot \inf\left\{g \ge 0 \colon \max_{1\le i\le n} \mathcal{K}_i^U(g) \le 2/\delta\right\}.$$

For any $g \ge 0$, the thresholding functions for $i = 1, \ldots, n$ are defined as

$$\mathcal{K}_i^L(g) = \prod_{j=1}^{i}\left(1 + \nu_j^L \cdot \left[\mathbb{1}_{\{V_j \le t\}}\, \hat{\ell}(X_j)/M - g\right]\right), \quad \mathcal{K}_i^U(g) = \prod_{j=1}^{i}\left(1 + \nu_j^U \cdot \left[1 - \mathbb{1}_{\{V_j > t\}}\, \hat{u}(X_j)/M - g\right]\right),$$

where $\nu_j^L = \min\left\{1, \sqrt{2\log(2/\delta)/[n(\hat{\sigma}_{j-1}^L)^2]}\right\}$, $\nu_j^U = \min\left\{1, \sqrt{2\log(2/\delta)/[n(\hat{\sigma}_{j-1}^U)^2]}\right\}$, and

$$(\hat{\sigma}_i^L)^2 = \frac{\frac{1}{4} + \sum_{j=1}^{i}\left(\mathbb{1}_{\{V_j \le t\}}\, \frac{\hat{\ell}(X_j)}{M} - \hat{\mu}_j^L\right)^2}{1+i}, \quad \hat{\mu}_i^L = \frac{\frac{1}{2} + \sum_{j=1}^{i} \mathbb{1}_{\{V_i \le t\}}\, \frac{\hat{\ell}(X_i)}{M}}{1+i},$$
$$(\hat{\sigma}_i^U)^2 = \frac{\frac{1}{4} + \sum_{j=1}^{i}\left(1 - \mathbb{1}_{\{V_j > t\}}\, \frac{\hat{u}(X_j)}{M} - \hat{\mu}_j^U\right)^2}{1+i}, \quad \hat{\mu}_i^U = \frac{\frac{1}{2} + \sum_{j=1}^{i}(1 - \mathbb{1}_{\{V_i > t\}}\, \frac{\hat{u}(X_i)}{M})}{1+i}.$$

Then it holds that $\mathbb{P}_{\mathcal{D}_{\text{calib}}}\left(\hat{G}_n(t) \le G(t)\right) \ge 1 - \delta$ for $G(t)$ defined in [13].

B. Validity of prediction intervals for ITE. In this part, we provide the coverage guarantee for prediction of ITEs omitted in Section A.

Proposition 2.2. *Consider a new test sample where we observe $(X_{n+1}, T_{n+1}, Y_{n+1}(T_{n+1}))$, with $T_{n+1} = w$ for $w \in \{0,1\}$. Let $\mathcal{D}$ be the calibration data generated from $\mathbb{P}^{\text{sup}}$ under confounding level Γ. If $\hat{C}_{1-w}(X_{n+1}, \Gamma, 1-\alpha)$ is constructed by Algorithm 1, then*

$$\mathbb{P}\left(Y_{n+1}(1) - Y_{n+1}(0) \in \hat{C}(X_{n+1}, \Gamma) \,\middle|\, T_{n+1} = w\right) \ge 1 - \alpha - \hat{\Delta},$$

for the prediction set $\hat{C}(X_{n+1}, \Gamma, 1-\alpha)$ in [A], where the probability is over $\mathcal{D}_{\text{calib}}$ as well as the test point, and $\hat{\Delta}$ is the gap of coverage for $Y_{n+1}(1-w)$ in Theorem 1.2. If $\hat{C}_{1-w}(X_{n+1}, \Gamma, 1-\alpha)$ is constructed by Algorithm 2 with confidence level $\delta \in (0,1)$, then with probability at least $1-\delta$ with respect to $\mathcal{D}_{\text{calib}}$, we have

$$\mathbb{P}\left(Y_{n+1}(1) - Y_{n+1} \in \hat{C}(X_{n+1}, \Gamma) \,\middle|\, T_{n+1} = w, \mathcal{D}_{\text{calib}}\right) \ge 1 - \alpha - \hat{\Delta},$$

where $\hat{\Delta}$ is the gap of coverage for $Y_{n+1}(1-w)$ in Theorem 3.1.

Proposition 2.3. *Consider a new test sample X_{n+1}. Let $\mathcal{D}$ be the observations for which $\mathbb{P}^{\text{sup}}$ is under confounding level Γ. If for $w \in \{0,1\}$, $\hat{C}_w(X_{n+1}, \Gamma, 1-\alpha/2, \delta/2)$ is constructed by Algorithm 1, then*

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \in \hat{C}(X_{n+1}, \Gamma)\big) \geq 1 - \alpha - \hat{\Delta}_1 - \hat{\Delta}_0,$$

where the probability is over $\mathcal{D}$ as well as the test point; $\hat{\Delta}_0$, $\hat{\Delta}_1$ is the coverage gap in Theorem 1.2 for counterfactual prediction of $Y_{n+1}(1), Y_{n+1}(0)$ when the bound functions are estimated. If $\hat{C}_{1-w}(X_{n+1}, \Gamma, 1-\alpha/2, \delta/2)$ is constructed by Algorithm 2, then with probability at least $1-\delta$ with respect to $\mathcal{D}_{\text{calib}}$, we have

$$\mathbb{P}\big(Y_{n+1}(1) - Y_{n+1} \in \hat{C}(X_{n+1}, \Gamma) \,\big|\, \mathcal{D}_{\text{calib}}\big) \geq 1 - \alpha - \hat{\Delta}_1 - \hat{\Delta}_0,$$

where $\hat{\Delta}_0$ and $\hat{\Delta}_1$ are the coverage gaps in Theorem 1.2 for counterfactual prediction of $Y_{n+1}(1), Y_{n+1}(0)$ when the bound functions are estimated.

C. Types of null hypotheses. With the general recipe of assessing robustness of causal conclusions on ITE, we provide concrete examples of the target set C and the corresponding forms of $\hat{C}(X_{n+1}, \Gamma)$.

Sharp null One might be interested in the sharp null, i.e., whether the individual treatment effect is zero. In this case, one could let $C = \{0\}$, and the prediction set can take any form. Rejecting $H_0(\Gamma)$ is saying that $Y_{n+1}(1) \neq Y_{n+1}(0)$ unless the true confounding level obeys $\Gamma^* > \Gamma$.

Directional null If we presume the stochastic nature of ITEs, the sharp null might be implausible. In this case, one may be interested in the directional null with $Y_{n+1}(1) - Y_{n+1}(0) \leq 0$, which is equivalent to choosing $C = (-\infty, 0]$. Rejecting $H_0(\Gamma)$ here is saying that the individual treatment effect is positive unless $\Gamma^* > \Gamma$. More generally, one might consider $C = (-\infty, a]$ for some $a \in \mathbb{R}$ to test whether the ITE is above a certain value. To make sense of the multiple testing procedure, one-sided prediction intervals are constructed for ITE, i.e., $\hat{C}(X_{n+1}, \Gamma) = [\hat{I}(X_{n+1}, \Gamma), \infty)$ for some $\hat{I}(X_{n+1}, \Gamma) \in \mathbb{R}$. It can be achieved by one-sided prediction intervals $\hat{C}_0(X_{n+1}, \Gamma, 1-\alpha, 1-\delta) = (-\infty, \hat{Y}_0(X_{n+1}, \Gamma)]$ if $Y_{n+1}(0)$ is missing and $\hat{C}_1(X_{n+1}, \Gamma, 1-\alpha, 1-\delta) = [\hat{Y}_1(X_{n+1}, \Gamma), \infty)$ if $Y_{n+1}(1)$ is missing, with a Bonferroni correction as introduced in Section A if both outcomes are missing. The Γ-value is hence the smallest Γ such that the two prediction intervals overlap.

D. Sharper results under parametric models. From the sharpness results in Section 3.C.2 (main text), we see that in the worst case scenario, the odds ratio [3] $\frac{\mathbb{P}(T=1 \mid X=x, U=u)\mathbb{P}(T=0 \mid X=x)}{\mathbb{P}(T=0 \mid X=x, U=u)\mathbb{P}(T=1 \mid X=x)}$ is equal to either Γ or $1/\Gamma$. Admittedly, we are unlikely to encounter this in practice. We now discuss how to get sharper prediction intervals by positing a parametric model.

Let $s\colon \mathcal{X} \to \mathbb{R}^{d_1}$ and $t\colon \mathcal{Y} \to \mathbb{R}^{d_2}$ be known functions. Imagine we are interested in predicting $Y(0)$ for a treated unit. Now assume that the true (confounded) treatment rule $\mathbb{P}(T = 1 \mid X, Y(0))$ follows the logistic model

$$\mathbb{P}(T = 1 \mid X, Y(0)) = \text{logit}^{-1}\big(\alpha^\top s(X) + \gamma^\top t(Y(0))\big), \tag{1}$$

where we define $\text{logit}^{-1}(x) = e^x/(1+e^x)$, and $\alpha \in \mathbb{R}^{d_1}$, $\gamma \in \mathbb{R}^{d_2}$ are unknown parameters. This parametric model may apply to healthcare data in which a doctor assigns a treatment to patients according to a logistic model depending on a full set of characteristics while only a few of those are included as covariates in the dataset due to privacy issues. The undocumented characteristics thus become the unmeasured confounder, and it may still be reasonable to assume that the treatment rule reduced to the outcome level, i.e., $\mathbb{P}(T = 1 \mid X, Y(0))$, approximately follows a logistic model [1].

Under the parametric model, the Γ-selection condition translates to

$$\frac{1}{\Gamma} \cdot \frac{e(X)}{1+e(X)} \leq \frac{\mathrm{d}\mathbb{P}_{X,Y(0) \mid T=1}}{\mathrm{d}\mathbb{P}_{X,Y(0) \mid T=0}} = e^{\alpha^\top s(X) + \gamma^\top t(Y(0))} \cdot \frac{\mathbb{P}(T=0)}{\mathbb{P}(T=1)} \leq \Gamma \cdot \frac{e(X)}{1+e(X)}.$$

Now let $\hat{e}(\cdot)$ be an estimator for $e(\cdot)$, and $\hat{p} = \frac{1}{n}\sum_{i=1}^n T_i$. We can modify the optimization problem in Remark 1.1 of the main text by replacing the constraints on W_i's with those on (α, γ). Specifically, for each k, we now solve

$$\begin{aligned}
\underset{\alpha,\gamma}{\text{minimize}} \quad & \frac{\sum_{i=1}^k \exp\{\alpha^\top s(X_{[i]}) + \gamma^\top t(Y_{[i]}(0))\}}{\sum_{i=1}^n \exp\{\alpha^\top s(X_i) + \gamma^\top t(Y_i(0))\} + W_{n+1}} \\
\text{subject to} \quad & \frac{\hat{e}(X_{n+1})}{\Gamma(1+\hat{e}(X_{n+1}))} \leq W_{n+1} \leq \frac{\Gamma \cdot \hat{e}(X_{n+1})}{1+\hat{e}(X_{n+1})} \\
& \frac{\hat{e}(X_i)}{\Gamma(1+\hat{e}(X_i))} \leq e^{\alpha^\top s(X_i) + \gamma^\top t(Y_i(0))} \cdot \frac{1-\hat{p}}{\hat{p}} \leq \frac{\Gamma \cdot \hat{e}(X_i)}{1+\hat{e}(X_i)}, \quad \forall i \in \mathcal{D}_{\text{calib}}.
\end{aligned}$$

Here, as before, $[1], [2], \ldots, [n]$ is the permutation of $\{1, \ldots, n\}$ such that $V_{[1]} \leq V_{[2]} \leq \cdots \leq V_{[n]}$. If we believe in the treatment assignment model, then this gives an extension of Algorithm 1 which provides sharper prediction intervals with valid coverage (up to estimation error in $\hat{e}$ and $\hat{p}$).

E. Nested method for inference on ITEs under unmeasured confounding. We provide details of the nested method for constructing less conservative ITE prediction intervals introduced in Section 4.

Assume the calibration data consist of both treated and control units. The nested method first uses the robust conformal inference methods in the main text to construct prediction intervals for the ITEs of all calibration data; it then learns and calibrates these intervals to produces a prediction interval for the ITE of a test point. Below, Algorithm 1 (resp. Algorithm 2) constructs a two-sided (resp. one-sided) prediction interval for the ITE of a test point. The output prediction interval achieves $(1 - \alpha - \gamma)$ coverage (up to estimation error of the propensity score), where γ corresponds to the confidence of the second calibration step. Depending on the input preference, such coverage is either marginalized over the calibration and test data, or calibration-conditionally valid with probability at least $1 - \delta$ over the calibration data.

Algorithms 1 and 2 use a subroutine to construct two-sided and one-sided robust prediction intervals for ITEs with either marginal or PAC-type coverage. The choice of the subroutine is flexible and depends on the target. Algorithm 3 (resp. Algorithm 4) provides a concrete procedure the constructs two-sided (resp. one-sided) prediction intervals for ITEs, where we use Conformalized Quantile Regression (CQR) (2) to compute the nonconformity score, and assume we have estimates $\hat{e}(x)$ for the propensity score $e(x) = \mathbb{P}(T = 1 \,|\, X = x)$, $\hat{p}$ for the treatment probability $\mathbb{P}(T = 1)$. The $\alpha/2$-th and $(1 - \alpha/2)$-th quantiles for $\mathbb{P}_{Y_i \,|\, X_i, T_i=1}$ and $\mathbb{P}_{Y_i \,|\, X_i, T_i=0}$ on $\mathcal{D}_{\text{train}}$ are denoted by $\hat{q}^{(1)}_{\alpha/2}(x)$, $\hat{q}^{(1)}_{1-\alpha/2}(x)$, $\hat{q}^{(0)}_{\alpha/2}(x)$, and $\hat{q}^{(0)}_{1-\alpha/2}(x)$, respectively.

Algorithm 1 Nested (two-sided) robust conformal prediction of ITEs

Input: Calibration data $\{(X_i, Y_i, T_i)\}_{i \in \mathcal{D}_{\text{calib}}}$, test covariate x, target level $\alpha, \gamma \in (0, 1)$, confidence level $\delta \in (0, 1)$.

Step I: sample splitting

Randomly split $\mathcal{D}_{\text{calib}}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Randomly split $\mathcal{D}_2$ into $\mathcal{D}_{2,1}$ and $\mathcal{D}_{2,2}$.

Step II: robust prediction of ITEs for $\mathcal{D}_2$

if target = marginal coverage **then**

  For all $i \in \mathcal{D}_2$, use $\mathcal{D}_1$ as calibration data to obtain two-sided prediction interval $[\hat{C}^L_i, \hat{C}^R_i] := [\hat{C}^L(X_i), \hat{C}^R(X_i)]$ for ITE with marginal coverage $1 - \alpha$ (e.g., from Algorithm 3).

else if target = $1 - \delta$ PAC-tpye guarantee **then**

  For all $i \in \mathcal{D}_2$, use $\mathcal{D}_1$ as calibration data to obtain two-sided prediction interval $[\hat{C}^L_i, \hat{C}^R_i] := [\hat{C}^L(X_i), \hat{C}^R(X_i)]$ for ITE with PAC-type coverage $1 - \alpha$ at confidence level $\delta/3$ (e.g., from Algorithm 3).

Step III: nested prediction interval for ITE of the test point

Fit $\hat{m}^L(x)$ and $\hat{m}^R(x)$, the conditional mean/median of C^L, C^R given $X_i = x$ using $\mathcal{D}_{2,1}$.

For $i \in \mathcal{D}_{2,2}$, compute $V_i = \max\{\hat{m}^L(X_i) - C^L_i, C^R_i - \hat{m}^R(X_i)\}$.

if target = marginal coverage **then**

  Compute η as the $(1 - \gamma)(1 + 1/|\mathcal{D}_{2,2}|)$-th quantile of the empirical distribution of $\{V_i \colon i \in \mathcal{D}_{2,2}\}$.

else if target = $1 - \delta$ PAC-tpye guarantee **then**

  Compute η as a $(1 - \delta/3)$ lower confidence bound for the $(1 - \gamma)$-th population quantile viewing $\{V_i \colon i \in \mathcal{D}_{2,2}\}$ as i.i.d. copies (using, e.g., Heoffding's inequality or W-S–R inequality).

Output: Prediction set $\hat{C}(x) = [\hat{m}^L(x) - \eta, \hat{m}^R(x) + \eta]$.

3. Technical proofs

A. Proof of Lemma 2.1.

Proof of Lemma 2.1. For any measurable subset $A \subset \mathcal{U}$, any $u \in A$ and any $x \in \mathcal{X}$, by the marginal Γ-selection condition [3],

$$\frac{1}{\Gamma} \cdot \mathbb{P}(T = 0 \,|\, X = x, U = u) \leq \mathbb{P}(T = 1 \,|\, X = x, U = u) \cdot \frac{\mathbb{P}(T = 0 \,|\, X = x)}{\mathbb{P}(T = 1 \,|\, X = x)} \leq \Gamma \cdot \mathbb{P}(T = 0 \,|\, X = x, U = u). \quad [2]$$

Algorithm 2 Nested (one-sided) robust conformal prediction of ITEs

Input: Calibration data $\{(X_i, Y_i, T_i)\}_{i\in\mathcal{D}_{\text{calib}}}$, test covariate x, target level $\alpha, \gamma \in (0,1)$, confidence level $\delta \in (0,1)$.

Step I: sample splitting

Randomly split $\mathcal{D}_{\text{calib}}$ into $\mathcal{D}_1$ and $\mathcal{D}_2$.

Randomly split $\mathcal{D}_2$ into $\mathcal{D}_{2,1}$ and $\mathcal{D}_{2,2}$.

Step II: robust prediction of ITEs for $\mathcal{D}_2$

if desire marginal coverage **then**

    For all $i \in \mathcal{D}_2$, use $\mathcal{D}_1$ as calibration data, obtain one-sided prediction interval $[\hat{C}_i^L, +\infty) := [\hat{C}^L(X_i), +\infty)$ for ITEs with marginal coverage $1-\alpha$ (e.g., from Algorithm 4).

else if desire $1-\delta$ PAC-tpye guarantee **then**

    For all $i \in \mathcal{D}_2$, use $\mathcal{D}_1$ as calibration data, obtain one-sided prediction interval $[\hat{C}_i^L, +\infty) := [\hat{C}^L(X_i), +\infty)$ for ITE with PAC-type coverage $1-\alpha$ at confidence level $\delta/3$ (e.g., from Algorithm 4).

Step III: nested prediction interval for ITE of test point

Fit $\hat{m}^L(x)$, the conditional mean/median of C^L given $X_i = x$ using $\mathcal{D}_{2,1}$.

For $i \in \mathcal{D}_{2,2}$, compute $V_i = C_i^L - \hat{m}^L(X_i)$.

if desire marginal coverage **then**

    Compute η as the $1-(1-\gamma)(1+1/|\mathcal{D}_{2,2}|)$-th quantile of the empirical distribution of $\{V_i : i \in \mathcal{D}_{2,2}\}$.

else if desire $1-\delta$ PAC-tpye guarantee **then**

    Compute η as a $(1-\delta/3)$ lower confidence bound for the γ-th population quantile viewing $\{V_i : i \in \mathcal{D}_{2,2}\}$ as i.i.d. copies (using, e.g., Heoffding's inequality or W-S–R inequality).

Output: Prediction set $\hat{C}(x) = [\hat{m}^L(x) + \eta, \infty)$.

Marginalizing over $u \in A$ yields

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(U \in A \,|\, X = x, T = 1)}{\mathbb{P}(U \in A \,|\, X = x, T = 0)} \le \Gamma \tag{3}$$

for $\mathbb{P}$-almost $x \in \mathcal{X}$. Since [3] holds for any measurable set in $\mathcal{U}$, we have

$$\frac{1}{\Gamma} \le \frac{\mathrm{d}\mathbb{P}_{U\,|\,X,T=1}}{\mathrm{d}\mathbb{P}_{U\,|\,X,T=0}}(u,x) \le \Gamma,$$

for $\mathbb{P}$-almost all $u \in \mathcal{U}$ and $x \in \mathcal{X}$. Meanwhile, for any measurable set $B \subset \mathcal{Y}$, by the tower property, we have for any $t \in \{0,1\}$ that

$$\begin{aligned}\mathbb{P}\big(Y(1) \in B, T = t \,\big|\, X\big) &= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\{Y(1)\in B\}}\mathbb{1}_{\{T=t\}} \,\big|\, X, U\big] \,\Big|\, X\Big] \\ &= \mathbb{E}\Big[\mathbb{E}\big[\mathbb{1}_{\{Y(1)\in B\}} \,\big|\, X, U\big] \cdot \mathbb{E}\big[\mathbb{1}_{\{T=t\}} \,\big|\, X, U\big] \,\Big|\, X\Big].\end{aligned} \tag{4}$$

Rewriting [2], we have $\mathbb{P}$-almost surely that

$$\frac{1}{\Gamma} \cdot \mathbb{E}\big[\mathbb{1}_{\{T=0\}} \,\big|\, X, U\big] \cdot \frac{\mathbb{E}\big[\mathbb{1}_{\{T=1\}} \,\big|\, X\big]}{\mathbb{E}\big[\mathbb{1}_{\{T=0\}} \,\big|\, X\big]} \le \mathbb{E}\big[\mathbb{1}_{\{T=1\}} \,\big|\, X, U\big] \le \Gamma \cdot \mathbb{E}\big[\mathbb{1}_{\{T=0\}} \,\big|\, X, U\big] \cdot \frac{\mathbb{E}\big[\mathbb{1}_{\{T=1\}} \,\big|\, X\big]}{\mathbb{E}\big[\mathbb{1}_{\{T=0\}} \,\big|\, X\big]}.$$

Multiplying all sides by $\mathbb{E}\big[\mathbb{1}_{\{Y(1)\in B\}} \,\big|\, X, U\big]$ and using [4], we know

$$\frac{1}{\Gamma} \cdot \mathbb{P}\big(Y(1) \in B, T = 0 \,\big|\, X\big) \cdot \frac{\mathbb{E}\big[\mathbb{1}_{\{T=1\}} \,\big|\, X\big]}{\mathbb{E}\big[\mathbb{1}_{\{T=0\}} \,\big|\, X\big]} \le \mathbb{P}\big(Y(1) \in B, T = 1 \,\big|\, X\big) \le \Gamma \cdot \mathbb{P}\big(Y(1) \in B, T = 0 \,\big|\, X\big) \cdot \frac{\mathbb{E}\big[\mathbb{1}_{\{T=1\}} \,\big|\, X\big]}{\mathbb{E}\big[\mathbb{1}_{\{T=0\}} \,\big|\, X\big]}$$

holds $\mathbb{P}$-almost surely, and for $\mathbb{P}$-almost all $x \in \mathcal{X}$,

$$\frac{1}{\Gamma} \cdot \frac{1-e(x)}{e(x)} \le \frac{\mathbb{P}(Y(1) \in B, T = 0 \,|\, X = x)}{\mathbb{P}(Y(1) \in B, T = 1 \,|\, X = x)} \le \Gamma \cdot \frac{1-e(x)}{e(x)}.$$

Note that

$$\frac{\mathbb{P}(Y(1) \in B \,|\, X = x, T = 1)}{\mathbb{P}(Y(1) \in B \,|\, X = x, T = 0)} = \frac{\mathbb{P}(Y(1) \in B, T = 1 \,|\, X = x)}{\mathbb{P}(Y(1) \in B, T = 0 \,|\, X = x)} \cdot \frac{1-e(x)}{e(x)}.$$

Algorithm 3 Subroutine: Two-sided robust prediction for ITE

Input: Calibration data $\{(X_i, Y_i, T_i)\}_{i\in\mathcal{D}}$, estimated propensity score $\hat{e}(x)$, estimated $\hat{p}$ for $\mathbb{P}(T=1)$, estimated quantile functions $\hat{q}^{(1)}_{\alpha/2}(x)$, $\hat{q}^{(1)}_{1-\alpha/2}(x)$, test covariate x, test treatment $t \in \{0,1\}$, test outcome $y(t)$, target level $\alpha \in (0,1)$, confidence level $\delta \in (0,1)$ if necessary.
Define the nonconformity score $V_t(x,y) = \max\{\hat{q}^{(t)}_{\alpha/2}(x) - y, y - \hat{q}^{(t)}_{1-\alpha/2}(x)\}$ for $t = 0, 1$.
Set $\hat{\ell}(\cdot) = \frac{1}{\Gamma}\left[\frac{1-\hat{p}}{\hat{p}}\frac{\hat{e}(x)}{1-\hat{e}(x)}\right]^{2t-1}$, $\hat{u}(\cdot) = \Gamma\left[\frac{1-\hat{p}}{\hat{p}}\frac{\hat{e}(x)}{1-\hat{e}(x)}\right]^{2t-1}$.
if desire marginal coverage **then**
    Input $\{j \in \mathcal{D}\colon T_j = 1-t\}$, $\hat{\ell}(\cdot)$, $\hat{u}(\cdot)$, score V_{1-t}, and target α to Algorithm 1 (main text).
    Obtain $[\hat{C}^L_{1-t}(x), \hat{C}^R_{1-t}(x)] := \hat{C}(x)$, the output from Algorithm 1 (main text).
else if desire $1-\delta$ PAC-tpye guarantee **then**
    Input $\{j \in \mathcal{D}\colon T_j = 1-t\}$, $\hat{\ell}(\cdot)$, $\hat{u}(\cdot)$, score V_{1-t}, target α, δ to Algorithm 2 (main text).
    Obtain $[\hat{C}^L_{1-t}(x), \hat{C}^R_{1-t}(x)] := \hat{C}(x)$, the output from Algorithm 2 (main text).
Output: $[\hat{C}^L(x), \hat{C}^R(x)] = [y(1) - \hat{C}^R_0(x), y(1) - \hat{C}^L_0(x)]$ if $t = 1$,
        $[\hat{C}^L(x), \hat{C}^R(x)] = [\hat{C}^L_1(x) - y(0), \hat{C}^R_1(x) - y(0)]$ if $t = 0$.

Algorithm 4 Subroutine: One-sided robust prediction for ITE

Input: Calibration data $\{(X_i, Y_i, T_i)\}_{i\in\mathcal{D}}$, estimated propensity score $\hat{e}(x)$, estimated $\hat{p}$ for $\mathbb{P}(T=1)$, estimated quantile functions $\hat{q}^{(0)}_{\alpha}(x)$, $\hat{q}^{(1)}_{1-\alpha}(x)$, test covariate x, test treatment $t \in \{0,1\}$, test outcome $y(t)$, target level $\alpha \in (0,1)$, confidence level $\delta \in (0,1)$ if necessary.
Define the nonconformity scores $V_0(x,y) = y - \hat{q}^{(0)}_{\alpha}(x)$ and $V_1(x,y) = \hat{q}^{(1)}_{1-\alpha}(x) - y$ for $t = 0, 1$.
Set $\hat{\ell}(\cdot) = \frac{1}{\Gamma}\left[\frac{1-\hat{p}}{\hat{p}}\frac{\hat{e}(x)}{1-\hat{e}(x)}\right]^{2t-1}$, $\hat{u}(\cdot) = \Gamma\left[\frac{1-\hat{p}}{\hat{p}}\frac{\hat{e}(x)}{1-\hat{e}(x)}\right]^{2t-1}$.
if desire marginal coverage **then**
    Input $\{j \in \mathcal{D}\colon T_j = 1-t\}$, $\hat{\ell}(\cdot)$, $\hat{u}(\cdot)$, score V_{1-t}, test covariate x, and target α to Algorithm 1 (main text).
    Obtain $(-\infty, \hat{C}_0(x)] := \hat{C}(x)$ if $t = 1$ or $[\hat{C}_1(x), +\infty) := \hat{C}(x)$ if $t = 0$, from Algorithm 1 (main text).
else if desire $1-\delta$ PAC-tpye guarantee **then**
    Input $\{j \in \mathcal{D}\colon T_j = 1-t\}$, $\hat{\ell}(\cdot)$, $\hat{u}(\cdot)$, score V_{1-t}, test covariate x, target α, δ to Algorithm 2 (main text).
    Obtain $(-\infty, \hat{C}_0(x)] := \hat{C}(x)$ if $t = 1$ or $[\hat{C}_1(x), +\infty) := \hat{C}(x)$ if $t = 0$, from Algorithm 2 (main text).
Output: $[\hat{C}^L(x), +\infty) = [y(1) - \hat{C}_0(x), +\infty)$ if $t = 1$, and $[\hat{C}^L(x), +\infty) = [\hat{C}^L_1(x) - y(0), +\infty)$ if $t = 0$.

Consequently,

$$\frac{1}{\Gamma} \le \frac{\mathbb{P}(Y(1) \in B \,|\, X = x, T = 1)}{\mathbb{P}(Y(1) \in B \,|\, X = x, T = 0)} \le \Gamma$$

holds for $\mathbb{P}$-almost all $x \in \mathcal{X}$. By the arbitrariness of B, we have

$$\frac{1}{\Gamma} \le \frac{\mathrm{d}\mathbb{P}_{Y(1)\,|\,X,T=1}}{\mathrm{d}\mathbb{P}_{Y(1)|X,T=0}}(x,y) \le \Gamma.$$

Repeating the above steps for $Y(0)$ we conclude the proof of Lemma 2.1. □

B. Proof of Theorem 1.2.

Proof of Theorem 1.2. Fixing any $\widetilde{\mathbb{P}} \in \mathcal{P}(\mathbb{P}, \ell, u)$, we denote the likelihood ratio $w(x,y) = \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(x,y)$. Recall that the calibration data is $\{(X_i, Y_i)\}_{i\in\mathcal{D}_{\mathrm{calib}}}$ with $\mathcal{D}_{\mathrm{calib}} = \{1, \ldots, n\}$, and the test data point is $(X_{n+1}, Y_{n+1}) \sim \widetilde{\mathbb{P}}$. We denote the random variables $Z_i = (X_i, Y_i)$ and realized values $z_i = (x_i, y_i)$ for $i = 1, \ldots, n$.

As a starting point, we elaborate on the weighted conformal inference introduced in (3), which paves the way for the analysis of marginal coverage later on. Following (3), the random variables $\{Z_i\}_{i=1}^{n+1}$ are *weighted exchangeable*, meaning that the density of their joint distribution can be factorized as

$$f(z_1, \ldots, z_{n+1}) = \prod_{i=1}^{n+1} w_i(z_i) \cdot g(z_1, \ldots, z_{n+1}), \qquad [5]$$

where g is some permutation-invariant function, i.e., $g(z_{\sigma(1)},\dots,z_{\sigma(n+1)}) = g(z_1,\dots,z_{n+1})$ for any permuation σ of $1,\dots,n+1$. Specifically, here $w_i(z)=1$ for $1\le i\le n$ and $w_{n+1}(z)=w(x,y)$. For a set of values $z_1,\dots,z_{n+1}$ where there may be repeated elements, we denote the unordered set $z=[z_1,\dots,z_{n+1}]$ and the event

$$\mathcal{E}_z = \big\{[Z_1,\dots,Z_{n+1}] = [z_1,\dots,z_{n+1}]\big\}.$$

Let Π_{n+1} be the set of all permutations of $\{1,\dots,n+1\}$. Writing $v_i = V(x_i,y_i) = V(z_i)$, for each $1\le i\le n+1$, it holds that

$$\mathbb{P}(V_{n+1}=v_i \,|\, \mathcal{E}_z) = \mathbb{P}(Z_{n+1}=z_i \,|\, \mathcal{E}_z) = \frac{\sum_{\sigma\in\Pi:\sigma(n+1)=i} f(z_{\sigma(1)},\dots,z_{\sigma(n+1)})}{\sum_{\sigma\in\Pi} f(z_{\sigma(1)},\dots,z_{\sigma(n+1)})},$$

where $\mathbb{P}$ is induced by the joint distribution of $\mathcal{D}_{\text{train}}\cup\mathcal{D}_{\text{calib}}\cup Z_{n+1}$. By the factorization [5], we have

$$\frac{\sum_{\sigma\in\Pi:\sigma(n+1)=i} f(z_{\sigma(1)},\dots,z_{\sigma(n+1)})}{\sum_{\sigma\in\Pi} f(z_{\sigma(1)},\dots,z_{\sigma(n+1)})} = \frac{\sum_{\sigma\in\Pi:\sigma(n+1)=i} w_{n+1}(z_i) g(z_{\sigma(1)},\dots,z_{\sigma(n+1)})}{\sum_{\sigma\in\Pi} w_{n+1}(z_{\sigma(n+1)}) g(z_{\sigma(1)},\dots,z_{\sigma(n+1)})} = \frac{w_{n+1}(z_i)}{\sum_{j=1}^{n+1} w_{n+1}(z_j)}.$$

Therefore, the distribution of V_{n+1} conditional on the event $\mathcal{E}_z$ is

$$V_{n+1}\,|\,\mathcal{E}_z \;\sim\; \sum_{i=1}^{n+1} \delta_{v_i} p_i^w, \quad \text{where } \; p_i^w = \frac{w_{n+1}(z_i)}{\sum_{j=1}^{n+1} w_{n+1}(z_j)} = \frac{w(x_i,y_i)}{\sum_{j=1}^{n+1} w(x_i,y_i)}.$$

Here δ_v denotes the point mass at $\{v\}$. For any unordered set $z=[z_1,\dots,z_n,z_{n+1}]$ and the corresponding k^* as defined in [B], it holds that

$$\mathbb{P}\big(Y_{n+1}\in\hat{C}(X_{n+1})\,\big|\,\mathcal{E}_z\big) = \mathbb{P}\big(V_{n+1}\le v_{[k^*]}\,\big|\,\mathcal{E}_z\big) = \sum_{i=1}^{n+1} p_i^w \mathbb{1}_{\{v_i\le v_{[k^*]}\}} = \frac{\sum_{i=1}^{n+1} w(x_i,y_i)\mathbb{1}_{\{v_i\le v_{[k^*]}\}}}{\sum_{j=1}^{n+1} w(x_i,y_i)}.$$

By the tower property of conditional expectations, we have

$$\mathbb{P}\big(Y_{n+1}\in\hat{C}(X_{n+1})\big) = \mathbb{E}\Big[\mathbb{P}\big(Y_{n+1}\in\hat{C}(X_{n+1})\,\big|\,\mathcal{E}_Z\big)\Big] = \mathbb{E}\left[\frac{\sum_{i=1}^{n+1} w(X_i,Y_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}}}{\sum_{j=1}^{n+1} w(X_i,Y_i)}\right]. \tag{6}$$

Equipped with the above preparations, we show the coverage guarantee in our setting. By definition [B], we know

$$V_{[k^*]} = \inf\left\{v\colon \frac{\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le v\}}}{\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le v\}} + \sum_{i=1}^n \hat{u}(X_i)\mathbb{1}_{\{V_i>v\}} + \hat{u}(X_{n+1})} \ge 1-\alpha\right\},$$

hence

$$\mathbb{E}\left[\frac{\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}}}{\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}} + \sum_{i=1}^n \hat{u}(X_i)\mathbb{1}_{\{V_i>V_{[k^*]}\}} + \hat{u}(X_{n+1})}\right] \ge 1-\alpha$$

since the inner random variable is always no smaller than $1-\alpha$. Combined with [6] and by the non-negativity of $w(X_i,Y_i)$, we have

$$\begin{aligned}
&\mathbb{P}\big(Y_{n+1}\in\hat{C}(X_{n+1})\big) - (1-\alpha) \\
&\ge \mathbb{E}\left[\frac{\sum_{i=1}^n w(X_i,Y_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}}}{\sum_{i=1}^{n+1} w(X_i,Y_i)}\right] - \mathbb{E}\left[\frac{\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}}}{\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}} + \sum_{i=1}^n \hat{u}(X_i)\mathbb{1}_{\{V_i>V_{[k^*]}\}} + \hat{u}(X_{n+1})}\right] := \mathbb{E}\left[\frac{\text{(i)}}{\text{(ii)}}\right],
\end{aligned}$$

where we denote

$$\begin{aligned}
\text{(i)} = &-w(X_{n+1},Y_{n+1})\cdot\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}} + \hat{u}(X_{n+1})\cdot\sum_{i=1}^n w(X_i,Y_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}} \\
&+ \left[\sum_{i=1}^n w(X_i,Y_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}}\right]\left[\sum_{i=1}^n \hat{u}(X_i)\mathbb{1}_{\{V_i>V_{[k^*]}\}}\right] - \left[\sum_{i=1}^n w(X_i,Y_i)\mathbb{1}_{\{V_i>V_{[k^*]}\}}\right]\left[\sum_{i=1}^n \hat{\ell}(X_i)\mathbb{1}_{\{V_i\le V_{[k^*]}\}}\right]
\end{aligned}$$

and

$$\text{(ii)} = \left[\sum_{i=1}^{n+1} w(X_i, Y_i)\right]\left[\sum_{i=1}^{n} \hat{\ell}(X_i)\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}} + \sum_{i=1}^{n} \hat{u}(X_i)\, \mathbb{1}_{\{V_i > V_{[\hat{k}^*]}\}} + \hat{u}(X_{n+1})\right].$$

We first establish a lower bound for the term (i). To this end, we define the random variables

$$\begin{aligned}
\widetilde{\ell}_i &= \max\left\{w(X_i, Y_i),\ \hat{\ell}(X_i)\right\} \quad \text{and} \\
\widetilde{u}_i &= \min\left\{w(X_i, Y_i),\ \hat{u}(X_i)\right\}, \quad i = 1, \ldots, n+1.
\end{aligned}$$

By the above definition, it holds that $0 \le \widetilde{u}_i \le \hat{u}(X_i)$ and $\widetilde{\ell}_i \ge \hat{\ell}(X_i) \ge 0$. We also define the differences

$$\begin{aligned}
\Delta\widetilde{\ell}_i &= \widetilde{\ell}_i - w(X_i, Y_i) = \left[\hat{\ell}(X_i) - w(X_i, Y_i)\right]_+ \quad \text{and} \\
\Delta\widetilde{u}_i &= \widetilde{u}_i - w(X_i, Y_i) = -\left[\hat{u}(X_i) - w(X_i, Y_i)\right]_-, \quad i = 1, \ldots, n+1.
\end{aligned}$$

Using the above notation, we have

$$\begin{aligned}
&\left[\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i \le V_{[k^*]}\}}\right]\left[\sum_{i=1}^{n} \hat{u}(X_i)\, \mathbb{1}_{\{V_i > V_{[k^*]}\}}\right] - \left[\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i > V_{[k^*]}\}}\right]\left[\sum_{i=1}^{n} \hat{\ell}(X_i)\, \mathbb{1}_{\{V_i \le V_{[k^*]}\}}\right] \\
&\ge \left[\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i \le V_{[k^*]}\}}\right]\left[\sum_{i=1}^{n} \left(w(X_i, Y_i) + \Delta\widetilde{u}_i\right) \mathbb{1}_{\{V_i > V_{[k^*]}\}}\right] \\
&\quad - \left[\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i > V_{[k^*]}\}}\right]\left[\sum_{i=1}^{n} \left(w(X_i, Y_i) + \Delta\widetilde{\ell}_i\right) \mathbb{1}_{\{V_i \le V_{[k^*]}\}}\right] \\
&= \left[\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i \le V_{[k^*]}\}}\right]\left[\sum_{i=1}^{n} \Delta\widetilde{u}_i\, \mathbb{1}_{\{V_i > V_{[k^*]}\}}\right] - \left[\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i > V_{[k^*]}\}}\right]\left[\sum_{i=1}^{n} \Delta\widetilde{\ell}_i\, \mathbb{1}_{\{V_i \le V_{[k^*]}\}}\right].
\end{aligned}$$

Following similar arguments, we have

$$\begin{aligned}
&- w(X_{n+1}, Y_{n+1}) \cdot \sum_{i=1}^{n} \hat{\ell}(X_i)\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}} + \hat{u}(X_{n+1}) \cdot \sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}} \\
&\ge -w(X_{n+1}, Y_{n+1}) \cdot \sum_{i=1}^{n} \left(w(X_i, Y_i) + \Delta\widetilde{\ell}_i\right) \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}} + \left(w(X_{n+1}, Y_{n+1}) + \Delta\widetilde{u}_{n+1}\right) \cdot \sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}} \\
&= -w(X_{n+1}, Y_{n+1}) \cdot \sum_{i=1}^{n} \Delta\widetilde{\ell}_i\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}} + \Delta\widetilde{u}_{n+1} \cdot \sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}}
\end{aligned}$$

By construction, we know that $\Delta\widetilde{\ell}_i \ge 0$ and $\Delta\widetilde{u}_i \le 0$ for $i = 1, \ldots, n+1$. Putting the lower bounds together, we obtain

$$\begin{aligned}
\text{(i)} &\ge \left[\sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}}\right]\left[\Delta\widetilde{u}_{n+1} + \sum_{i=1}^{n} \Delta\widetilde{u}_i\, \mathbb{1}_{\{V_i > V_{[\hat{k}^*]}\}}\right] \\
&\quad - \left[w(X_{n+1}, Y_{n+1}) + \sum_{i=1}^{n} w(X_i, Y_i)\, \mathbb{1}_{\{V_i > V_{[\hat{k}^*]}\}}\right]\left[\sum_{i=1}^{n} \Delta\widetilde{\ell}_i\, \mathbb{1}_{\{V_i \le V_{[\hat{k}^*]}\}}\right] \\
&\ge \left[\sum_{i=1}^{n} w(X_i, Y_i)\right]\left[\sum_{i=1}^{n+1} \Delta\widetilde{u}_i\right] - \left[\sum_{i=1}^{n+1} w(X_i, Y_i)\right]\left[\sum_{i=1}^{n} \Delta\widetilde{\ell}_i\right].
\end{aligned}$$

Since the term (ii) is non-negative, we have the lower bound

$$\begin{aligned}
\mathbb{E}\left[\frac{\text{(i)}}{\text{(ii)}}\right] =& \mathbb{E}\left[\frac{\sum_{i=1}^n w(X_i,Y_i)\,\mathbb{1}_{\{V_i\le V_{[\hat{k}^*]}\}}}{\sum_{i=1}^{n+1} w(X_i,Y_i)} - \frac{\sum_{i=1}^n \hat{\ell}(X_i)\,\mathbb{1}_{\{V_i\le V_{[\hat{k}^*]}\}}}{\sum_{i=1}^n \hat{\ell}(X_i)\,\mathbb{1}_{\{V_i\le V_{[\hat{k}^*]}\}} + \sum_{i=1}^n \hat{u}(X_i)\,\mathbb{1}_{\{V_i> V_{[\hat{k}^*]}\}} + \hat{u}(X_{n+1})}\right] \\
&\ge \mathbb{E}\left[\frac{\left[\sum_{i=1}^n w(X_i,Y_i)\right]\left[\sum_{i=1}^{n+1}\Delta\widetilde{u}_i\right] - \left[\sum_{i=1}^{n+1} w(X_i,Y_i)\right]\left[\sum_{i=1}^n \Delta\widetilde{\ell}_i\right]}{\left[\sum_{i=1}^{n+1} w(X_i,Y_i)\right]\left[\sum_{i=1}^n \hat{\ell}(X_i)\,\mathbb{1}_{\{V_i\le V_{[\hat{k}^*]}\}} + \sum_{i=1}^n \hat{u}(X_i)\,\mathbb{1}_{\{V_i> V_{[\hat{k}^*]}\}} + \hat{u}(X_{n+1})\right]}\right] \\
&\overset{\text{(a)}}{\ge} \mathbb{E}\left[\frac{\left[\sum_{i=1}^n w(X_i,Y_i)\right]\left[\sum_{i=1}^{n+1}\Delta\widetilde{u}_i\right] - \left[\sum_{i=1}^{n+1} w(X_i,Y_i)\right]\left[\sum_{i=1}^n \Delta\widetilde{\ell}_i\right]}{\left[\sum_{i=1}^{n+1} w(X_i,Y_i)\right]\left[\sum_{i=1}^n \hat{\ell}(X_i)\right]}\right] \\
&\overset{\text{(b)}}{\ge} -\mathbb{E}\left[\frac{\sum_{i=1}^{n+1}[\hat{u}(X_i)-w(X_i,Y_i)]_-}{\sum_{i=1}^n \hat{\ell}(X_i)}\right] - \mathbb{E}\left[\frac{\sum_{i=1}^{n}[\hat{\ell}(X_i)-w(X_i,Y_i)]_+}{\sum_{i=1}^n \hat{\ell}(X_i)}\right].
\end{aligned}$$

Above, step (a) follows from the fact that $\hat{u}(X_i) \ge \hat{\ell}(X_i) \ge 0$, and step (b) is due to the non-negativity of $w(X_i, Y_i)$. By Hölder's inequality,

$$\begin{aligned}
\mathbb{E}\left[\frac{\sum_{i=1}^n[\hat{\ell}(X_i)-w(X_i,Y_i)]_+}{\sum_{i=1}^n \hat{\ell}(X_i)}\right] &\le \left\|\frac{1}{n}\sum_{i=1}^n \big(\hat{\ell}(X_i)-w(X_i,Y_i)\big)_+\right\|_p \cdot \left\|\frac{n}{\sum_{i=1}^n \hat{\ell}(X_i)}\right\|_q \\
&\overset{\text{(a)}}{\le} \left\|\big(\hat{\ell}(X_i)-w(X_i,Y_i)\big)_+\right\|_p \cdot \left\|\frac{n}{\sum_{i=1}^n \hat{\ell}(X_i)}\right\|_q \\
&\overset{\text{(b)}}{\le} \left\|\big(\hat{\ell}(X_i)-w(X_i,Y_i)\big)_+\right\|_p \cdot \left\|\frac{1}{\hat{\ell}(X_i)}\right\|_q,
\end{aligned}$$

where step (a) follows from Minkowski's inequality, and the step (b) follows from

$$\frac{n}{\sum_{i=1}^n \hat{\ell}(X_i)} \le \frac{1}{n}\sum_{i=1}^n \frac{1}{\hat{\ell}(X_i)}$$

as implied by Cauchy-Schwarz inequality. Similarly,

$$\mathbb{E}\left[\frac{\sum_{i=1}^{n+1}[\hat{u}(X_i)-w(X_i,Y_i)]_-}{\sum_{i=1}^n \hat{\ell}(X_i)}\right] \le \left\|\big(\hat{u}(X_i)-u(X_i)\big)_-\right\|_p \cdot \left\|\frac{1}{\hat{\ell}(X_i)}\right\|_q + \frac{1}{n}\left\|\big(\hat{u}(X_{n+1})-u(X_{n+1})\big)_-\right\|_p \cdot \left\|\frac{1}{\hat{\ell}(X_i)}\right\|_q,$$

where the L_p norm for X_{n+1} is with respect to $\widetilde{\mathbb{P}}$, hence

$$\frac{1}{n}\left\|\big(\hat{u}(X_{n+1})-u(X_{n+1})\big)_-\right\|_p = \left\|\frac{w(X_i,Y_i)^{1/p}}{n}\cdot\big(\hat{u}(X_i)-u(X_i)\big)_-\right\|_p$$

Combining the above results, we have

$$\mathbb{P}\big(Y_{n+1}\in\hat{C}(X_{n+1})\big) \ge 1-\alpha-\hat{\Delta}\cdot\left\|1/\hat{\ell}(X_i)\right\|_q,$$

where

$$\hat{\Delta} = \left\|\big(\hat{\ell}(X_i)-\ell(X_i)\big)_+\right\|_p + \left\|\big(\hat{u}(X_i)-u(X_i)\big)_-\right\|_p + \left\|\frac{w(X_i,Y_i)^{1/p}}{n}\cdot\big(\hat{u}(X_i)-u(X_i)\big)_-\right\|_p,$$

which completes the proof of Theorem 1.2. □

C. Proof of Theorem 3.1.

Proof of Theorem 3.1. Throughout the proof, all statements are conditional on $\mathcal{D}_{\text{train}}$. By the independence of $\mathcal{D}_{\text{calib}} \cup \{(X_{n+1}, Y_{n+1})\}$, the scores $\{V(X_i, Y_i)\}_{i \in \mathcal{D}_{\text{calib}}}$ are i.i.d. and independent of $V(X_{n+1}, Y_{n+1})$.

Recall that $G(\cdot)$ is defined in [13]. To begin with, we define

$$\hat{q} = \inf\big\{t \colon G(t) \geq 1-\alpha\big\}, \quad \hat{q}_n = \inf\big\{t \colon \hat{G}_n(t) \geq 1-\alpha\big\}.$$

For any fixed $\epsilon > 0$, we have

$$\begin{aligned}\mathbb{P}(\hat{q}_n \leq \hat{q} - \epsilon) &= \mathbb{P}\big(\hat{G}_n(\hat{q}-\epsilon) \geq 1-\alpha\big) \\ &\leq \mathbb{P}\big(G(\hat{q}-\epsilon) \geq \hat{G}_n(\hat{q}-\epsilon) \geq 1-\alpha\big) + \mathbb{P}\big(G(\hat{q}-\epsilon) < \hat{G}_n(\hat{q}-\epsilon)\big) \leq \delta.\end{aligned}$$

Here the last inequality follows from the fact that $G(\hat{q}-\epsilon) < 1-\alpha$ for any fixed $\epsilon > 0$ and $\mathbb{P}\big(G(\hat{q}-\epsilon) < \hat{G}_n(\hat{q}-\epsilon)\big) \leq \delta$ by [14] with $t = \hat{q} - \epsilon$. Therefore, by the continuity of probability measures, we have

$$\mathbb{P}(\hat{q}_n \geq \hat{q}\,) = 1 - \lim_{\epsilon \to 0^+} \mathbb{P}(\hat{q}_n \leq \hat{q} - \epsilon) \geq 1 - \delta.$$

Moreover, on the event $\{\hat{q}_n \geq \hat{q}\,\}$, by the definition of $\hat{C}(X_{n+1})$, it holds for any $\widetilde{\mathbb{P}} \in \mathcal{P}(\mathbb{P}, \ell, u)$ that

$$\begin{aligned}\widetilde{\mathbb{P}}\big(Y_{n+1} \in \hat{C}(X_{n+1}) \,\big|\, \mathcal{D}_{\text{calib}}\big) &= \widetilde{\mathbb{P}}\big(V(X_{n+1}, Y_{n+1}) \leq \hat{q}_n \,\big|\, \mathcal{D}_{\text{calib}}\big) \\ &\geq \widetilde{\mathbb{P}}\big(V(X_{n+1}, Y_{n+1}) \leq \hat{q} \,\big|\, \mathcal{D}_{\text{calib}}\big) \\ &= \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}\,\}}\, w(X,Y) \,\big|\, \mathcal{D}_{\text{calib}}\big],\end{aligned}$$

where the expectation is with respect to $(X, Y) \sim \mathbb{P}$ independent of $\mathcal{D}_{\text{calib}}$. By the definition of $G(t)$

$$\hat{q} = \inf\left\{t \colon \max\left\{\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq t\}}\, \hat{\ell}(X)\big],\ 1 - \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) > t\}}\, \hat{u}(X)\big]\right\} \geq 1-\alpha\right\} = \min\{\hat{q}_1, \hat{q}_2\},$$

where $(X, Y) \sim \mathbb{P}$ is an independent copy and we define

$$\begin{aligned}\hat{q}_1 &= \inf\left\{t \colon \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq t\}}\, \hat{\ell}(X)\big] \geq 1-\alpha\right\}, \\ \hat{q}_2 &= \inf\left\{t \colon 1 - \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) > t\}}\, \hat{u}(X)\big] \geq 1-\alpha\right\}.\end{aligned}$$

For constants $\hat{q}_1, \hat{q}_2$, we have

$$\begin{aligned}&\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}\,\}}\, w(X,Y) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &= \min\left\{\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_1\}}\, w(X,Y) \,\big|\, \mathcal{D}_{\text{calib}}\big],\ \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_2\}}\, w(X,Y) \,\big|\, \mathcal{D}_{\text{calib}}\big]\right\}.\end{aligned}$$

We analyze the two terms separately. Firstly,

$$\begin{aligned}&\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_1\}}\, w(X,Y) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &= \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_1\}}\, \hat{\ell}(X) \,\big|\, \mathcal{D}_{\text{calib}}\big] - \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_1\}}\, \big(\hat{\ell}(X) - w(X,Y)\big) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &\overset{\text{(a)}}{\geq} 1 - \alpha - \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_1\}}\, \big(\hat{\ell}(X) - w(X,Y)\big) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &\geq 1 - \alpha - \mathbb{E}\big[\big(\hat{\ell}(X) - w(X,Y)\big)_+\big],\end{aligned}$$

where the step (a) follows from the definition of $\hat{q}_1$. Similarly,

$$\begin{aligned}&\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_2\}}\, w(X,Y) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &= 1 - \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) > \hat{q}_2\}}\, w(X,Y) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &= 1 - \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_2\}}\, \hat{u}(X) \,\big|\, \mathcal{D}_{\text{calib}}\big] + \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_2\}}\, \big(\hat{u}(X) - w(X,Y)\big) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &\geq 1 - \alpha + \mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y) \leq \hat{q}_2\}}\, \big(\hat{u}(X) - w(X,Y)\big) \,\big|\, \mathcal{D}_{\text{calib}}\big] \\ &\geq 1 - \alpha - \mathbb{E}\big[\big(\hat{u}(X) - w(X,Y)\big)_-\big],\end{aligned}$$

where the first equality follows from the fact that $w(x,y)$ is a likelihood ratio, and the first inequality follows from the definition of $\hat{q}_2$. Putting them together, on the event $\{\hat{q}_n \geq \hat{q}\}$ which happens with probability at least $1-\delta$ with repsect to $\mathcal{D}_{\text{calib}}$, it holds that

$$\widetilde{\mathbb{P}}\big(Y_{n+1} \in \hat{C}(X_{n+1}) \,\big|\, \mathcal{D}_{\text{calib}}\big) \geq 1 - \alpha - \hat{\Delta},$$

where the gap is

$$\hat{\Delta} = \max\Big\{\mathbb{E}\big[\big(\hat{\ell}(X) - w(X,Y)\big)_+\big],\ \mathbb{E}\big[\big(\hat{u}(X) - w(X,Y)\big)_-\big]\Big\},$$

and the expectations are with respect to an independent copy $(X,Y) \sim \mathbb{P}$. Therefore, we conclude the proof of Theorem 3.1. □

D. Proof of Proposition 2.1.

Proof of Proposition 2.1. Let $t \in \mathbb{R}$ be any fixed constant or any random variable in $\sigma(\mathcal{D}_{\text{train}})$, and fix any $\delta \in (0,1)$. We condition on $\mathcal{D}_{\text{train}}$ throughout the proof, so that V, $\hat{\ell}$ and $\hat{u}$ can be viewed as fixed and t can be viewed as constant.

Since $\hat{G}_n(t) = \max\big\{\hat{G}_n^{\ell}(t), \hat{G}_n^{U}(t)\big\}$, by the definition [13], it suffices to show that

$$\mathbb{P}_{\mathcal{D}_{\text{calib}}}\Big(\hat{G}_n^{\ell}(t) \leq \mathbb{E}\big[\mathbb{1}_{\{V_i \leq t\}}\,\hat{\ell}(X_i)\big]\Big) \geq 1 - \delta/2$$
$$\text{and}\quad \mathbb{P}_{\mathcal{D}_{\text{calib}}}\Big(\hat{G}_n^{U}(t) \leq 1 - \mathbb{E}\big[\mathbb{1}_{\{V_i > t\}}\,\hat{u}(X_i)\big]\Big) \geq 1 - \delta/2,$$

Now let

$$f(X_i) = \mathbb{1}_{\{V_i \leq t\}}\,\hat{\ell}(X_i)/M, \quad h(X_i) = 1 - \mathbb{1}_{\{V_i > t\}}\,\hat{u}(X_i)/M,$$

so that $f(X_i)$ and $h(X_i)$, $1 \leq i \leq n$, are i.i.d. random variables in $[0,1]$. We rescale the bounds into $\hat{f}_n = \hat{G}_n^{\ell}(t)/M$ and $\hat{h}_n = (\hat{G}_n^{U}(t) - 1 + M)/M$, hence by the tower property of conditional expectations, it suffices to show that

$$\mathbb{P}_{\mathcal{D}_{\text{calib}}}\Big(\hat{f}_n \leq \mathbb{E}\big[f(X_i)\big]\Big) \geq 1 - \delta/2 \quad \text{and} \quad \mathbb{P}_{\mathcal{D}_{\text{calib}}}\Big(\hat{h}_n \leq \mathbb{E}\big[h(X_i)\big]\Big) \geq 1 - \delta/2.$$

We show the desired result for $f(\cdot)$ and that for $h(\cdot)$ naturally applies.

Consider the filtration $\{\mathcal{F}_i\}_{i\geq 1}$, where the σ-algebra $\mathcal{F}_i = \sigma\{X_1, Y_1, \ldots, X_i, Y_i\}$. Then by definition, $\hat{\sigma}_i^L$ and $\hat{\mu}_i^L$ are measurable with respect to $\mathcal{F}_i$, and $\{\nu_i^L\}_{i\geq 1}$ is a predictable sequence with respect to $\{\mathcal{F}_i\}_{i\geq 1}$. Thus letting

$$f_0 = \mathbb{E}\big[f(X_i)\big] = \mathbb{E}\big[\mathbb{1}_{\{V_i \leq t\}}\,\ell(X_i)\big]/M,$$

we have

$$\mathbb{E}\big[\mathcal{K}_i^{\ell}(f_0) \,\big|\, \mathcal{F}_{i-1}\big] = \mathcal{K}_{i-1}^{\ell}(f_0) \cdot \Big(1 + \nu_j^L \cdot \mathbb{E}\big[f(X_i) - f_0 \,\big|\, \mathcal{F}_{i-1}\big]\Big) = \mathcal{K}_{i-1}^{\ell}(f_0).$$

Thus $\{K_i^{\ell}(f_0)\}_{i=1}^n$ is a martingale. Meanwhile, since $f(X_i) \in [0,1]$, we know $f(X_i) - f_0 \geq -1$. Hence $\mathcal{K}_i^{\ell}(f_0) \geq 0$ for all $i \in [n]$. Therefore by Ville's inequality,

$$\mathbb{P}_{\mathcal{D}_{\text{calib}}}\Big(\max_{1\leq i\leq n} \mathcal{K}_i^L(f_0) > \frac{2}{\delta}\Big) \leq \frac{\delta}{2}.$$

Also note that $\mathcal{K}_i(g)$ is a decreasing function in $g \in \mathbb{R}$, hence

$$\mathbb{P}_{\mathcal{D}_{\text{calib}}}\Big(\hat{f}_n > \mathbb{E}\big[f(X_i)\big]\Big) \leq \mathbb{P}\Big(\max_{1\leq i\leq n} \mathcal{K}_i(f_0) > \frac{2}{\delta}\Big) \leq \delta/2,$$

Therefore, we conclude the proof of the desired results. □

E. Proof of Proposition 3.2.

Proof of Proposition 3.2. Throughout the proof, we denote the generic random variables $(X, Y) \sim \mathbb{P}$ and $V = V(X, Y)$. Denoting the right-hand side of [18] as $F^*(t)$, We are to show that

- (i) $F^*(\cdot)\colon \mathbb{R} \to [0, 1]$ is a distribution function.
- (ii) $F^*(t)$ is a lower bound for $F(t; \mathcal{P}(\mathbb{P}, \ell, u))$ for all $t \in \mathbb{R}$.
- (iii) $F^*(t)$ can be achieved by some element in $\mathcal{P}(\mathbb{P}, \ell, u)$.

First of all, (i) is straightforward by noting that $F^*(\cdot)$ is right-continuous due to the continuity of probability measures and $\lim_{t\to-\infty} F^*(t) = 0$, $\lim_{t\to+\infty} F^*(t) = 1$. Also, similar to the arguments in the proof of Theorem 3.1, for any $\widetilde{\mathbb{P}} \in \mathcal{P}(\mathbb{P}, \ell, u)$, we have

$$F(t\,; V, \widetilde{\mathbb{P}}) = \mathbb{E}\left[\mathbb{1}_{\{V(X,Y)\le t\}} \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(X, Y) \right] \ge \mathbb{E}\big[\mathbb{1}_{\{V(X,Y)\le t\}}\, \ell(X) \big],$$
$$F(t\,; V, \widetilde{\mathbb{P}}) = 1 - \mathbb{E}\left[\mathbb{1}_{\{V(X,Y)> t\}} \frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(X, Y) \right] \ge 1 - \mathbb{E}\big[\mathbb{1}_{\{V(X,Y)> t\}}\, u(X) \big],$$

hence (ii) follows. For (iii), we are to construct one distribution $\mathbb{P}^* \in \mathcal{P}(\mathbb{P}, \ell, u)$ so that $F(\cdot\,; V, \mathbb{P}^*) = F^*(\cdot)$.

If $\mathbb{E}[u(X)] = 1$, then since $\mathbb{E}[w(X, Y)] = 1$ for the likelihood ratio function, the collection $\mathcal{P}(\mathbb{P}, \ell, u) = \{\mathbb{P}^*\}$ is a singleton with $\mathrm{d}\mathbb{P}^*/\,\mathrm{d}\mathbb{P}(x, y) = u(x)$, and

$$F\big(t\,; \mathcal{P}(\mathbb{P}, \ell, u)\big) = F(t\,; \mathbb{P}^*) = \mathbb{E}\big[\mathbb{1}_{\{V(X,Y)\le t\}}\, u(X) \big] = F^*(t),$$

where the last equality follows from $\ell(x) \le u(x)$. Then $\mathbb{P}^*$ satisfies (iii). Similarly, for the case where $\mathbb{E}[\ell(X)] = 1$, the collection $\mathcal{P}(\mathbb{P}, \ell, u) = \{\mathbb{P}^*\}$ is a singleton with $\mathrm{d}\mathbb{P}^*/\,\mathrm{d}\mathbb{P}(x, y) = \ell(x)$, and

$$F\big(t\,; \mathcal{P}(\mathbb{P}, \ell, u)\big) = F(t\,; \mathbb{P}^*) = \mathbb{E}\big[\mathbb{1}_{\{V(X,Y)\le t\}}\, l(X) \big] = F^*(t),$$

hence $\mathbb{P}^*$ satisfies (iii). In the sequel, we consider the case where $\mathbb{E}[\ell(X)] < 1 < \mathbb{E}[u(X)]$. We define

$$H(t) \;:=\; \mathbb{E}\big[\ell(X) \cdot \mathbb{1}_{\{V\le t\}} + u(X) \cdot \mathbb{1}_{\{V> t\}} \big].$$

By the construction and the continuity of probability measures, we have

$$\lim_{t\to\infty} H(t) = \mathbb{E}\big[\ell(X)\big] < 1 < \mathbb{E}\big[u(X)\big] = \lim_{t\to-\infty} H(t).$$

We additionally define $t^* = \inf\{t \in \mathbb{R} : H(t) \le 1\}$. Note that $t^* < \infty$ and $H(t^*) \le 1$ by the right continuity of $H(t)$. We also define the left limit of $H(t)$ at t^* as

$$H_-(t^*) = \lim_{t\uparrow t^*} H(t),$$

and define the weight as

$$\gamma = \frac{1 - H(t^*)}{H_-(t^*) - H(t^*)}\, \mathbb{1}_{\{H_-(t^*)>1\}}$$

Here by the definition of t^*, we have $H_-(t^*) \ge 1 \ge H(t^*)$, and the left limit takes the form

$$H_-(t^*) = \mathbb{E}\big[\ell(X) \cdot \mathbb{1}_{\{V< t^*\}} + u(X) \cdot \mathbb{1}_{\{V\ge t^*\}} \big].$$

We now construct the worst-case distribution $\mathbb{P}^*$ by

$$\frac{\mathrm{d}\mathbb{P}^*}{\mathrm{d}\mathbb{P}}(x, y) = \gamma \cdot \big[\ell(x)\, \mathbb{1}_{\{V(x,y)< t^*\}} + u(x)\, \mathbb{1}_{\{V(x,y)\ge t^*\}} \big] + (1 - \gamma) \cdot \big[\ell(x)\, \mathbb{1}_{\{V(x,y)\le t^*\}} + u(x)\, \mathbb{1}_{\{V(x,y)> t^*\}} \big].$$

We denote the likelihood ratio $w^*(x, y) = \mathrm{d}\mathbb{P}^*/\,\mathrm{d}\mathbb{P}(x, y)$ as constructed above. Note that

$$\mathbb{P}^*(\mathcal{X} \times \mathcal{Y}) = \mathbb{E}\big[w^*(X, Y)\big] = \gamma \cdot H_-(t^*) + (1 - \gamma) \cdot H(t^*) = 1,$$

and also $\ell(x) \le w^*(x,y) \le u(x)$ for all $x \in \mathcal{X}$. Therefore $\mathbb{P}^*$ is a probability measure and is an element of $\mathcal{P}(\mathbb{P}, \ell, u)$. In the following, we check that $F(t\,; V, \mathbb{P}^*) = F^*(t)$ for all $t \in \mathbb{R}$. Recall that we work with the case $\mathbb{E}[\ell(X)] < 1 < \mathbb{E}[u(X)]$. For any constant $t < t^*$, by the construction of w^*,

$$\begin{aligned} F(t\,; V, \mathbb{P}^*) &= \mathbb{E}\big[w^*(X,Y)\,\mathbb{1}_{\{V(X,Y)\le t\}}\big] \\ &= \gamma \cdot \mathbb{E}\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)\le t\}}\big] + (1-\gamma)\cdot\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)\le t\}}\big] \\ &= \mathbb{E}\big[\ell(X)\cdot\mathbb{1}_{\{V\le t\}}\big] = F^*(t), \end{aligned}$$

where the last equality follows from the fact that

$$\mathbb{E}\big[\ell(X)\cdot\mathbb{1}_{\{V\le t\}}\big] > 1 - \mathbb{E}\big[u(X)\cdot\mathbb{1}_{\{V>t\}}\big]$$

since $H(t) > 1$ for $t < t^*$. When $t = t^*$, note that

$$\mathbb{E}\big[\ell(X)\cdot\mathbb{1}_{\{V\le t^*\}}\big] - 1 + \mathbb{E}\big[u(X)\cdot\mathbb{1}_{\{V>t^*\}}\big] = H(t^*) - 1 \le 0,$$

hence the right-hand side of [18] admits the form

$$F^*(t^*) = 1 - \mathbb{E}\big[u(X)\cdot\mathbb{1}_{\{V>t^*\}}\big].$$

Meanwhile, by the construction of $w^*(x,y)$, we have

$$\begin{aligned} F(t^*\,; V, \mathbb{P}^*) &= \mathbb{E}\big[w^*(X,Y)\,\mathbb{1}_{\{V(X,Y)\le t^*\}}\big] \\ &= \gamma\cdot\mathbb{E}\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)<t^*\}} + u(X)\,\mathbb{1}_{\{V(X,Y)=t^*\}}\big] + (1-\gamma)\cdot\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)\le t^*\}}\big] \\ &= \mathbb{E}\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)\le t^*\}}\big] - \gamma\cdot\Big(\mathbb{E}\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)=t^*\}} - u(X)\cdot\mathbb{1}_{\{V(X,Y)=t^*\}}\big]\Big) \\ &= \mathbb{E}\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)\le t^*\}}\big] - \gamma\cdot\big(H_-(t^*) - H(t^*)\big) \\ &= \mathbb{E}\big[\ell(X)\,\mathbb{1}_{\{V(X,Y)\le t^*\}}\big] + 1 - H(t^*) = 1 - \mathbb{E}\big[u(X)\cdot\mathbb{1}_{\{V>t^*\}}\big] = F^*(t^*). \end{aligned}$$

Similarly, when $t > t^*$, by the cosntruction of $w^*(x,y)$ we have

$$\begin{aligned} F(t^*\,; V, \mathbb{P}^*) &= 1 - \mathbb{E}\big[w^*(X,Y)\,\mathbb{1}_{\{V(X,Y)>t\}}\big] \\ &= 1 - \gamma\cdot\mathbb{E}\big[u(X)\,\mathbb{1}_{\{V(X,Y)>t\}}\big] + (1-\gamma)\cdot\mathbb{E}\big[u(X)\,\mathbb{1}_{\{V(X,Y)>t\}}\big] \\ &= 1 - \mathbb{E}\big[u(X)\,\mathbb{1}_{\{V(X,Y)>t\}}\big] = F^*(t), \end{aligned}$$

where the last equality follows from the fact that $H(t) \le 1$ thus $1 - \mathbb{E}[u(X)\,\mathbb{1}_{\{V(X,Y)>t\}}] \ge \mathbb{E}[\ell(X)\,\mathbb{1}_{\{V(X,Y)\le t\}}]$. Combining the three cases, we arrive at $F^*(\cdot) = F(\cdot\,; V, \mathbb{P}^*)$, hence [18] follows and we conclude the proof of Proposition 3.2. □

F. Proof of Proposition 3.3.

Proof of Proposition 3.3. The proof proceeds by showing that $\mathcal{P} \subset \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$ and $\mathcal{P}(\mathbb{P}, f, \ell_0, u_0) \subset \mathcal{P}$, which together lead to the desired result.

Step 1: $\mathcal{P} \subset \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$. Let $\mathbb{P}^{\text{sup}}$ be any super-population that satisfies [3] and [19]. Due to the partial observation of potential outcomes, the observed distribution adimits the decomposition

$$\mathbb{P}^{\text{obs}}_{X,Y,T} = \mathbb{P}^{\text{obs}}_{T=1} \times \mathbb{P}^{\text{obs}}_{X,Y(1)\,|\,T=1} + \mathbb{P}^{\text{obs}}_{T=0} \times \mathbb{P}^{\text{obs}}_{X,Y(0)\,|\,T=0}.$$

Therefore, the data-compatibility condition [19] is equivalent to

$$\mathbb{P}^{\text{sup}}_{T} = \mathbb{P}^{\text{obs}}_{T}, \quad \mathbb{P}^{\text{sup}}_{X,Y(1)\,|\,T=1} = \mathbb{P}^{\text{obs}}_{X,Y(1)\,|\,T=1}, \quad \mathbb{P}^{\text{sup}}_{X,Y(0)\,|\,T=0} = \mathbb{P}^{\text{obs}}_{X,Y(0)\,|\,T=0}.$$

where the latter two are further equivalent to

$$\mathbb{P}^{\text{sup}}_{X\,|\,T=w} = \mathbb{P}^{\text{obs}}_{X\,|\,T=w}, \quad \mathbb{P}^{\text{sup}}_{Y(1)\,|\,X,T=w} = \mathbb{P}^{\text{obs}}_{Y(1)\,|\,X,T=w}, \quad w \in \{0,1\}. \tag{7}$$

Recall that $\widetilde{\mathbb{P}} = \mathbb{P}^{\text{sup}}_{X,Y(1)\,|\,T=0}$ and $\mathbb{P} = \mathbb{P}^{\text{obs}}_{X,Y(1)\,|\,T=1}$. Then we have

$$\frac{\mathrm{d}\widetilde{\mathbb{P}}_X}{\mathrm{d}\mathbb{P}_X} = \frac{\mathrm{d}\mathbb{P}^{\text{sup}}_{X\,|\,T=0}}{\mathrm{d}\mathbb{P}^{\text{obs}}_{X\,|\,T=1}} = \frac{\mathrm{d}\mathbb{P}^{\text{obs}}_{X\,|\,T=0}}{\mathrm{d}\mathbb{P}^{\text{obs}}_{X\,|\,T=1}} = \frac{\mathbb{P}^{\text{obs}}(T=1)\cdot\mathbb{P}^{\text{obs}}(T=0)}{\mathbb{P}^{\text{obs}}(T=0)\cdot\mathbb{P}^{\text{obs}}(T=1)} = f(X),$$

where the second equality follows from [7] and the third equality follows from the Bayes rule. On the other hand, the shift of conditional distribution is

$$\frac{\mathrm{d}\widetilde{\mathbb{P}}_{Y(1)\,|\,X}}{\mathrm{d}\mathbb{P}_{Y(1)\,|\,X}} = \frac{\mathrm{d}\mathbb{P}^{\mathrm{sup}}_{Y(1)\,|\,X,T=0}}{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{Y(1)\,|\,X,T=1}} \in [1/\Gamma, \Gamma]$$

according to Lemma 2.1. Therefore, $\mathbb{P}^{\mathrm{sup}} \in \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$ by the definition. Hence we have $\mathcal{P} \subset \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$.

Step 2: $\mathcal{P}(\mathbb{P}, f, \ell_0, u_0) \subset \mathcal{P}$**.** For this part, we are to show that for any $\widetilde{\mathbb{P}} \in \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$, there exists some $\mathbb{P}^{\mathrm{sup}}$ satisfying [3] and [19] such that $\widetilde{\mathbb{P}} = \mathbb{P}^{\mathrm{sup}}_{X,Y(1)\,|\,T=0}$. Fixing an arbitrary probability distribution $\widetilde{\mathbb{P}} \in \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$, we define the function

$$w(y\,|\,x) = \frac{\mathrm{d}\widetilde{\mathbb{P}}_{Y(1)\,|\,X}}{\mathrm{d}\mathbb{P}_{Y(1)\,|\,X}}(y\,|\,x),$$

so that $w(y\,|\,x) \in [1/\Gamma, \Gamma]$ for $\mathbb{P}^{\mathrm{obs}}$-almost all $(x, y) \in \mathcal{X} \times \mathcal{Y}$. Also, since $\widetilde{\mathbb{P}}$ is a distribution, we have

$$\mathbb{E}\big[w(Y(1)\,|\,x)\,|\,X = x\big] = \widetilde{\mathbb{P}}\big(Y(1) \in \mathcal{Y}\,\big|\,X = x\big) = 1$$

for $\mathbb{P}$-almost all $x \in \mathcal{X}$, and the conditional expectation is induced by $(X, Y(1)) \sim \mathbb{P} = \mathbb{P}^{\mathrm{obs}}_{X,Y(1)\,|\,T=1}$. Applying Lemma 3.1 with $r(x, y) = w(y\,|\,x)$ and $t = 1$, we know there exists a distribution $\mathbb{P}^{\mathrm{sup}}$ over $(X, Y(0), Y(1), U, T)$ for some confounder U that satisfies [19] and [3], and

$$\frac{\mathrm{d}\mathbb{P}^{\mathrm{sup}}_{Y(1),X\,|\,T=0}}{\mathrm{d}\mathbb{P}^{\mathrm{sup}}_{Y(1),X\,|\,T=1}}(y\,|\,x) = w(y\,|\,x).$$

Since $\mathbb{P}^{\mathrm{sup}}$ satisfies [19], we have $\mathbb{P}^{\mathrm{sup}}_{Y(1),X\,|\,T=1} = \mathbb{P}^{\mathrm{obs}}_{Y(1),X\,|\,T=1} = \mathbb{P}_{Y(1)\,|\,X}$ where we recall the definition of $\mathbb{P}$, the distribution at hand. Hence

$$w(y\,|\,x) = \frac{\mathrm{d}\mathbb{P}^{\mathrm{sup}}_{Y(1),X\,|\,T=0}}{\mathrm{d}\mathbb{P}_{Y(1)\,|\,X}}(y\,|\,x) = \frac{\mathrm{d}\widetilde{\mathbb{P}}_{Y(1)\,|\,X}}{\mathrm{d}\mathbb{P}_{Y(1)\,|\,X}}(y\,|\,x).$$

Therefore, we have $\widetilde{\mathbb{P}}_{Y(1)\,|\,X} = \mathbb{P}^{\mathrm{sup}}_{Y(1),X\,|\,T=0}$. Furthermore, since $\mathbb{P}^{\mathrm{sup}}$ satisfies [19], we know

$$\frac{\mathrm{d}\mathbb{P}^{\mathrm{sup}}_{X\,|\,T=0}}{\mathrm{d}\mathbb{P}_X} = \frac{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{X\,|\,T=0}}{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{X\,|\,T=1}} = f(X) = \frac{\mathrm{d}\widetilde{\mathbb{P}}_X}{\mathrm{d}\mathbb{P}_X},$$

where the second equality follows from the Bayes rule and the last equality follows from the fact that $\widetilde{\mathbb{P}} \in \mathcal{P}(\mathbb{P}, f, \ell_0, u_0)$. Thus, we have $\widetilde{\mathbb{P}}_X = \mathbb{P}^{\mathrm{sup}}_{X\,|\,T=0}$. Putting the two parts together, we have $\widetilde{\mathbb{P}} = \mathbb{P}^{\mathrm{sup}}_{X,Y(1)\,|\,T=0}$. By the arbitrariness of $\widetilde{\mathbb{P}}$, we arrive at $\mathcal{P}(\mathbb{P}, f, \ell_0, u_0) \subset \mathcal{P}$.

Combining the two steps, we conclude the proof of Proposition 3.3. □

G. Sharpness of the identification set.

Lemma 3.1 (Sharpness of Lemma 2.1)**.** *Given $t \in \{0, 1\}$, a marginal distribution $\mathbb{P}^{\mathrm{obs}}$ over (X, Y, T) and a function $r(x, y) \in [1/\Gamma, \Gamma]$ such that*

$$\mathbb{E}^{\mathrm{obs}}\big[r\big(X, Y(t)\big)\,\big|\,X, T = t\big] = 1, \quad \mathbb{P}^{\mathrm{obs}}\text{-almost surely,}$$

there exists a distribution $\mathbb{P}^{\mathrm{sup}}$ over $(X, Y(0), Y(1), U, T)$ for some confounder U such that

- *(i) $\mathbb{P}^{\mathrm{sup}}_{X,Y,T} = \mathbb{P}^{\mathrm{obs}}_{X,Y,T}$ for $Y = Y(T)$;*
- *(ii) $\mathbb{P}^{\mathrm{sup}}$ satisfies the marginal Γ-selection condition;*
- *(iii) the likelihood ratio is exactly $r(x, y)$, so that $r(x, y) = \frac{\mathrm{d}\mathbb{P}^{\mathrm{sup}}_{Y(t)\,|\,X,T=1-t}}{\mathrm{d}\mathbb{P}^{\mathrm{sup}}_{Y(t)\,|\,X,T=t}}(x, y)$ for $\mathbb{P}^{\mathrm{sup}}$-almost all x, y.*

hold simulatenously.

Proof of Lemma 3.1. Fix any marginal distribution $\mathbb{P}^{\mathrm{obs}}$ over (X, Y, T) for $Y = Y(T)$, and a function $r(x, y) \in [1/\Gamma, \Gamma]$ satisfying the given condition. We show the result for $t = 1$, while that for $t = 0$ follows exactly the same arguments.

The construction of $\mathbb{P}^{\sup}$ To begin with, we let the confounder be the counterfactual itself, so that $U = Y(1)$. The joint distribution is thus

$$\mathbb{P}^{\sup}_{(X,Y(1),Y(0),T} = \mathbb{P}^{\sup}_{X,T} \times \mathbb{P}^{\sup}_{(Y(1),Y(0))\,|\,X,T}.$$

We set the two parts separately. Firstly, we set $\mathbb{P}^{\sup}_{X,T} = \mathbb{P}^{\mathrm{obs}}_{X,T}$ for the distribution on (X,T). The joint distribution of potential outcomes given (X,T) admits

$$\mathbb{P}^{\sup}_{(Y(1),Y(0))\,|\,X,T} = \mathbb{P}^{\sup}_{Y(1)\,|\,X,T} \times \mathbb{P}^{\sup}_{Y(0)\,|\,X,T,Y(1)},$$

Since our target is for $Y(1)$, we take a simple coupling where $Y(0)$ is independent of $(T, Y(1))$ conditional on X, so that we set

$$\mathbb{P}^{\sup}_{Y(0)\,|\,X,T,Y(1)} = \mathbb{P}^{\sup}_{Y(0)\,|\,X} = \mathbb{P}^{\sup}_{Y(0)\,|\,X,T=0} = \mathbb{P}^{\mathrm{obs}}_{Y(0)\,|\,X,T=0}. \tag{8}$$

On the other hand, we set $\mathbb{P}^{\sup}_{Y(1)\,|\,X,T}$ for $T = 0, 1$ by

$$\mathbb{P}^{\sup}_{Y(1)\,|\,X,T=1} = \mathbb{P}^{\mathrm{obs}}_{Y(1)\,|\,X,T=1} \quad \text{and} \quad \frac{\mathrm{d}\mathbb{P}^{\sup}_{Y(1)\,|\,X,T=0}}{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{Y(1)\,|\,X,T=1}}(y\,|\,x) = r(x,y). \tag{9}$$

So far we've completed the pieces of constructing $\mathbb{P}^{\sup}$. It remains to check that it is indeed a probability measure. By construction, $\mathbb{P}^{\sup}_{X,T} = \mathbb{P}^{\mathrm{obs}}_{X,T}$ is a probability measure on (X,T). Also, by the construction,

$$\mathbb{P}^{\sup}\big(Y(1) \in \mathcal{Y}\,|\,X, T=1\big) = \mathbb{P}^{\mathrm{obs}}\big(Y(1) \in \mathcal{Y}\,|\,X, T=1\big) = 1,$$

and

$$\begin{aligned}\mathbb{P}^{\sup}\big(Y(1) \in \mathcal{Y}\,|\,X, T=0\big) &= \mathbb{E}^{\mathrm{obs}}\left[\mathbb{1}_{\{Y(1)\in\mathcal{Y}\}}\,\frac{\mathrm{d}\mathbb{P}^{\sup}_{Y(1)\,|\,X,T=0}}{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{Y(1)\,|\,X,T=1}}\,\middle|\,X, T=1\right] \\ &= \mathbb{E}^{\mathrm{obs}}\left[\frac{\mathrm{d}\mathbb{P}^{\sup}_{Y(1)\,|\,X,T=0}}{\mathrm{d}\mathbb{P}^{\mathrm{obs}}_{Y(1)\,|\,X,T=1}}\,\middle|\,X, T=1\right] = \mathbb{E}^{\mathrm{obs}}\big[r\big(X, Y(1)\big)\,\big|\,X, T=1\big] = 1,\end{aligned}$$

where the first equality is by the change-of-measure formula, the second equality follows from $1 = \mathbb{1}_{\{Y(1)\in\mathcal{Y}\}}$, the third equality follows from the construction, and the last equality is the given condition on $r(x,y)$. Thus, $\mathbb{P}^{\sup}_{Y(1)\,|\,X,T}$ is a probability measure. Also, by the construction of $\mathbb{P}_{Y(0)\,|\,X,T,Y(1)}$, we have

$$\mathbb{P}^{\sup}\big(Y(0) \in \mathcal{Y}\,|\,X, T, Y(1)\big) = \mathbb{P}^{\mathrm{obs}}\big(Y(0) \in \mathcal{Y}\,|\,X, T=0\big) = 1,$$

hence $\mathbb{P}^{\sup}_{Y(0)\,|\,X,T,Y(1)}$ is also a probability measure. Putting them together, we know that

$$\mathbb{P}^{\sup}_{X,T,Y(1),Y(0)} = \mathbb{P}^{\sup}_{X,T} \times \mathbb{P}^{\sup}_{Y(1)\,|\,X,T} \times \mathbb{P}^{\sup}_{Y(0)\,|\,X,T,Y(1)}$$

is indeed a probability measure over $(X, T, Y(1), Y(0))$.

Verify the properties We now proceed to verify the three stated properties. For (i), due to partial obervability, the observed distribution admits the decomposition

$$\mathbb{P}^{\mathrm{obs}}_{X,Y,T} = \mathbb{P}^{\mathrm{obs}}_{T} \times \mathbb{P}^{\mathrm{obs}}_{X,Y\,|\,T} = \mathbb{P}^{\mathrm{obs}}_{T=1} \times \mathbb{P}^{\mathrm{obs}}_{X,Y(1)\,|\,T=1} + \mathbb{P}^{\mathrm{obs}}_{T=0} \times \mathbb{P}^{\mathrm{obs}}_{X,Y(0)\,|\,T=0}.$$

Similarly, the projection on the observable of $\mathbb{P}^{\sup}$ is

$$\mathbb{P}^{\sup}_{X,Y,T} = \mathbb{P}^{\sup}_{T} \times \mathbb{P}^{\sup}_{X,Y\,|\,T} = \mathbb{P}^{\sup}_{T=1} \times \mathbb{P}^{\sup}_{X,Y(1)\,|\,T=1} + \mathbb{P}^{\sup}_{T=0} \times \mathbb{P}^{\sup}_{X,Y(0)\,|\,T=0}.$$

Here by the construction of $\mathbb{P}^{\sup}_{X,T} = \mathbb{P}^{\mathrm{obs}}_{X,T}$, we have $\mathbb{P}^{\mathrm{obs}}_{T=w} = \mathbb{P}^{\sup}_{T=w}$ and $\mathbb{P}^{\sup}_{X\,|\,T=w} = \mathbb{P}^{\mathrm{obs}}_{X\,|\,T=w}$ for $w \in \{0,1\}$. Also, $\mathbb{P}^{\sup}_{Y(w)\,|\,X,T=w} = \mathbb{P}^{\mathrm{obs}}_{Y(w)\,|\,X,T=w}$ holds for $w \in \{0,1\}$ by [8] and [9]. The equivalences altogether leads to $\mathbb{P}^{\sup}_{X,Y,T} = \mathbb{P}^{\mathrm{obs}}_{X,Y,T}$. For (ii), by the Bayes rule, we have

$$r(x,y) = \frac{\mathrm{d}\mathbb{P}^{\sup}_{Y(1)\,|\,X,T=0}}{\mathrm{d}\mathbb{P}^{\sup}_{Y(1)\,|\,X,T=1}}(y\,|\,x) = \frac{\mathbb{P}^{\sup}(T=0\,|\,X=x, Y(1)=y)}{\mathbb{P}^{\sup}(T=1\,|\,X=x, Y(1)=y)} \cdot \frac{\mathbb{P}^{\sup}(T=1\,|\,X=x)}{\mathbb{P}^{\sup}(T=0\,|\,X=x)} \in [1/\Gamma, \Gamma],$$

so $\mathbb{P}^{\sup}$ satisfies the marginal Γ-selection condition [3] hence (ii) is verified. Property (iii) has also been verified as above. So far, we've constructed $\mathbb{P}^{\sup}$ that satisfies all stated conditions and we conclude the proof of Lemma 3.1. □

H. Proof of Proposition 3.4.

Proof of Proposition 3.4. For simplicity, we denote [21] as $F^*(t)$, and aim to show that

- (i) $F^*(t)$ is a distribution function;
- (ii) $F^*(t)$ is a lower bound for $F(t\,;V,\widetilde{\mathbb{P}})$ for all $t\in\mathbb{R}$ and all $\widetilde{\mathbb{P}}\in\mathcal{P}(\mathbb{P},f,\ell_0,u_0)$;
- (iii) $F^*(t)$ can be achieved by $\mathbb{P}^*\in\mathcal{P}(\mathbb{P},f,\ell_0,u_0))$ where $\mathrm{d}\mathbb{P}^*/\,\mathrm{d}\mathbb{P}(x,y)=w^*(x,y)$ as defined in [3.4].

To verify (i), we note that $F^*(t)$ is right continuous by the continuity of probability measures, as well as $\lim_{t\to-\infty}F^*(t)=0$ and $\lim_{t\to+\infty}F^*(t)=1$. To show (ii), we are to show that for any $\widetilde{\mathbb{P}}\in\mathcal{P}(\mathbb{P},f,\ell_0,u_0)$, it holds $\mathbb{P}$-almost surely that

$$\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,w^*(X,Y)\,\big|\,X\big]\le\widetilde{\mathbb{P}}\big(V(X,Y)\le t\,\big|\,X\big)=\mathbb{E}\left[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,\frac{\mathrm{d}\widetilde{\mathbb{P}}}{\mathrm{d}\mathbb{P}}(X,Y)\,\middle|\,X\right].$$

Fixing any $\widetilde{\mathbb{P}}\in\mathcal{P}(\mathbb{P},f,\ell_0,u_0)$, we denote the conditional likelihood as $w_0(y\,|\,x)=\mathrm{d}\widetilde{\mathbb{P}}_{Y\,|\,X}/\,\mathrm{d}\mathbb{P}_{Y\,|\,X}(y\,|\,x)$, so that $\ell_0(x)\le w_0(y\,|\,x)\le u_0(x)$ for $\mathbb{P}$-almost all $x\in\mathcal{X}$. Then the marginal likelihood ratio is $w(x,y):=\mathrm{d}\widetilde{\mathbb{P}}/\,\mathrm{d}\mathbb{P}(x,y)=f(x)\cdot w_0(y\,|\,x)$. Hence for any $t\in\mathbb{R}$,

$$\begin{aligned}
&\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,w^*(X,Y)\,\big|\,X\big]-\widetilde{\mathbb{P}}\big(V(X,Y)\le t\,\big|\,X\big)\\
&=\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,w^*(X,Y)\,\big|\,X\big]-\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,w(X,Y)\,\big|\,X\big]\\
&=\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,\big(w^*(X,Y)-w(X,Y)\big)\,\big|\,X\big]\cdot\mathbb{1}_{\{t<q(\tau(X)\,;X,\mathbb{P})\}}\\
&\quad+\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,\big(w^*(X,Y)-w(X,Y)\big)\,\big|\,X\big]\cdot\mathbb{1}_{\{t\ge q(\tau(X)\,;X,\mathbb{P})\}}\,.
\end{aligned}$$

We treat the two terms in the last summation separately. By the definition of $w^*(x,y)$,

$$\begin{aligned}
&\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,\big(w^*(X,Y)-w(X,Y)\big)\,\big|\,X\big]\cdot\mathbb{1}_{\{t<q(\tau(X)\,;X,\mathbb{P})\}}\\
&=f(X)\cdot\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,\big(\ell_0(X)-w_0(Y\,|\,X)\big)\,\big|\,X\big]\cdot\mathbb{1}_{\{t<q(\tau(X)\,;X,\mathbb{P})\}}\le 0.
\end{aligned}$$

Meanwhile, since $1=\mathbb{E}[w_0(Y\,|\,X)\,|\,X]=\mathbb{E}[w^*(X,Y)/f(X)\,|\,X]=1$ holds $\mathbb{P}$-almost surely, we have

$$\begin{aligned}
&\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,\big(w^*(X,Y)-w(X,Y)\big)\,\big|\,X\big]\cdot\mathbb{1}_{\{t\ge q(\tau(X)\,;X,\mathbb{P})\}}\\
&=f(X)\cdot\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)>t\}}\,\big(w_0(Y\,|\,X)-w^*(X,Y)/f(X)\big)\,\big|\,X\big]\cdot\mathbb{1}_{\{t\ge q(\tau(X)\,;X,\mathbb{P})\}}\\
&=f(X)\cdot\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)>t\}}\,\big(w_0(Y\,|\,X)-u_0(X)\big)\,\big|\,X\big]\cdot\mathbb{1}_{\{t\ge q(\tau(X)\,;X,\mathbb{P})\}}\le 0.
\end{aligned}$$

Summing them up and by the tower property of conditional expectations, it holds for any $t\in\mathbb{R}$ that

$$F^*(t)=\mathbb{E}\big[\,\mathbb{1}_{\{V(X,Y)\le t\}}\,w^*(X,Y)\big]\le\widetilde{\mathbb{P}}\big(V(X,Y)\le t\big),$$

which verifies property (ii). Finally, we define $\mathbb{P}^*$ by $\mathrm{d}\mathbb{P}^*/\,\mathrm{d}\mathbb{P}(x,y)=w^*(x,y)$. Then since $\mathbb{E}[w^*(X,Y)/f(X)\,|\,X=x]=1$, the marginal likelihood ratio satisfies $\mathrm{d}\mathbb{P}^*_X/\,\mathrm{d}\mathbb{P}_X=f(x)$, hence $\mathrm{d}\mathbb{P}^*_{Y\,|\,X}/\,\mathrm{d}\mathbb{P}_{Y\,|\,X}(x,y)=w^*(x,y)/f(x)$. To verify $\mathbb{P}^*\in\mathcal{P}(\mathbb{P},f,\ell_0,u_0)$, it remains to show $\ell_0(x)\le\gamma_0(x)\le u_0(x)$ when it is nonzero, i.e., $\mathbb{P}(V(x,Y)=q(\tau(x)\,;x,\mathbb{P})\,|\,X=x)>0$. In this case,

$$\gamma_0(x)=\frac{1-\ell_0(x)\cdot\mathbb{P}(V(x,Y)<q(\tau(x)\,;x,\mathbb{P})\,|\,X=x)-u_0(x)\cdot\mathbb{P}(V(x,Y)>q(\tau(x)\,;x,\mathbb{P})\,|\,X=x)}{\mathbb{P}(V(x,Y)=q(\tau(x)\,;x,\mathbb{P})\,|\,X=x)}.$$

Note that $\mathbb{P}(V(x,Y)<q(\tau(x)\,;x,\mathbb{P})\,|\,X=x)\le\tau(x)=(u_0(x)-1)/(u_0(x)-\ell_0(x))$, hence

$$\begin{aligned}
&1-\ell_0(x)\cdot\mathbb{P}\big(V(x,Y)<q(\tau(x)\,;x,\mathbb{P})\,\big|\,X=x\big)\\
&\le u_0(x)-u_0(x)\cdot\mathbb{P}\big(V(x,Y)<q(\tau(x)\,;x,\mathbb{P})\,\big|\,X=x\big)\\
&=u_0(x)\cdot\mathbb{P}\big(V(x,Y)\ge q(\tau(x)\,;x,\mathbb{P})\,\big|\,X=x\big)\\
&=u_0(x)\cdot\mathbb{P}\big(V(x,Y)>q(\tau(x)\,;x,\mathbb{P})\,\big|\,X=x\big)+u_0(x)\cdot\mathbb{P}\big(V(x,Y)=q(\tau(x)\,;x,\mathbb{P})\,\big|\,X=x\big),
\end{aligned}$$

which leads to $\gamma_0(x) \le u_0(x)$. On the other hand, by the definition of quantiles, we have $\mathbb{P}(V(x,Y) \le q(\tau(x)\,; x, \mathbb{P}) \,|\, X = x) \ge \tau(x) = (u_0(x) - 1)/(u_0(x) - \ell_0(x))$. Hence

$$\begin{aligned}
&\ell_0(x) \cdot \mathbb{P}\big(V(x,Y) < q(\tau(x)\,; x, \mathbb{P}) \,\big|\, X = x\big) + \ell_0(x) \cdot \mathbb{P}\big(V(x,Y) = q(\tau(x)\,; x, \mathbb{P}) \,\big|\, X = x\big) \\
&= \ell_0(x) \cdot \mathbb{P}\big(V(x,Y) \le q(\tau(x)\,; x, \mathbb{P}) \,\big|\, X = x\big) \\
&\le 1 - u_0(x) + u_0(x) \cdot \mathbb{P}\big(V(x,Y) \le q(\tau(x)\,; x, \mathbb{P}) \,\big|\, X = x\big) \\
&= 1 - u_0(x) \cdot \mathbb{P}\big(V(x,Y) > q(\tau(x)\,; x, \mathbb{P}) \,\big|\, X = x\big),
\end{aligned}$$

which leads to $\gamma_0(x) \ge \ell_0(x)$. Therefore, we conclude the proof of Proposition 3.4. □

I. Proofs of Proposition 4.1.

Proof of Proposition 4.1. Recall that Γ^* is the smallest sensitivity parameter such that $\mathbb{P}^{\text{sup}} \in \mathcal{P}(\Gamma^*)$. By the definition of $\mathcal{R}$ in [B] and the nested property of the prediction sets, we know $\mathcal{R} = [1, \hat{\Gamma})$ if $C \cap \hat{C}(X_{n+1}, 1) = \varnothing$, where

$$\hat{\Gamma} = \sup\{\Gamma \ge 1 : C \cap \hat{C}(X_{n+1}, \gamma) = \varnothing\},$$

and $\mathcal{R} = \varnothing$ otherwise. Also, by the nested nature of $H_0(\Gamma)$, we know $\mathcal{H}_0 = [\Gamma^*, \infty)$ when $Y_{n+1}(1) - Y_{n+1}(0) \in C$ and $\mathcal{H}_0 = \varnothing$ otherwise. Hence

$$\begin{aligned}
\text{mErr} &= \mathbb{P}(\mathcal{R} \cap \mathcal{H}_0 \ne \varnothing) \\
&= \mathbb{P}\Big(\big\{Y_{n+1}(1) - Y_{n+1}(0) \in C\big\} \cap \big\{\exists\, \Gamma \ge \Gamma^*,\ C \cap \hat{C}(X_{n+1}, \Gamma) = \varnothing\big\}\Big) \\
&\le \mathbb{P}\Big(\big\{Y_{n+1}(1) - Y_{n+1}(0) \in C\big\} \cap \big\{C \cap \hat{C}(X_{n+1}, \Gamma^*) = \varnothing\big\}\Big) \\
&\le \mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \notin \hat{C}(X_{n+1}, \Gamma^*)\big).
\end{aligned}$$

Following exactly the same arguments, we have

$$\begin{aligned}
\text{dErr} &= \mathbb{P}(\mathcal{R} \cap \mathcal{H}_0 \ne \varnothing \,|\, \mathcal{D}_{\text{calib}}) \\
&= \mathbb{P}\Big(\big\{Y_{n+1}(1) - Y_{n+1}(0) \in C\big\} \cap \big\{\exists\, \Gamma \ge \Gamma^*,\ C \cap \hat{C}(X_{n+1}, \Gamma) = \varnothing\big\} \,\Big|\, \mathcal{D}_{\text{calib}}\Big) \\
&\le \mathbb{P}\Big(\big\{Y_{n+1}(1) - Y_{n+1}(0) \in C\big\} \cap \big\{C \cap \hat{C}(X_{n+1}, \Gamma^*) = \varnothing\big\} \,\Big|\, \mathcal{D}_{\text{calib}}\Big) \\
&\le \mathbb{P}\big(Y_{n+1}(1) - Y_{n+1}(0) \notin \hat{C}(X_{n+1}, \Gamma^*) \,\big|\, \mathcal{D}_{\text{calib}}\big),
\end{aligned}$$

completing the proof of Proposition 4.1. □

4. Additional simulation results

A. Additional results for Section C. In this part, we provide additional simulation results on the counterfactual prediction task in Section C.

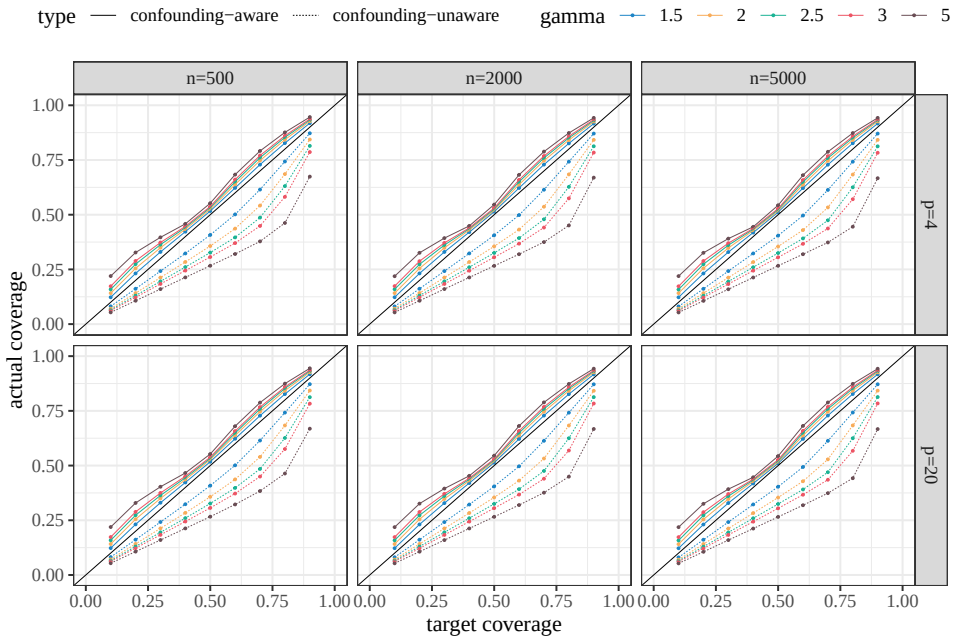

Fig. S1. Empirical (average) coverage of Algorithm 1 when $\ell(\cdot)$ and $u(\cdot)$ are known. The details are otherwise the same as in Figure 3.

B. Additional simulation results for Section D. In this part, we provide additional simulation results on the counterfactual prediction task in Section D.

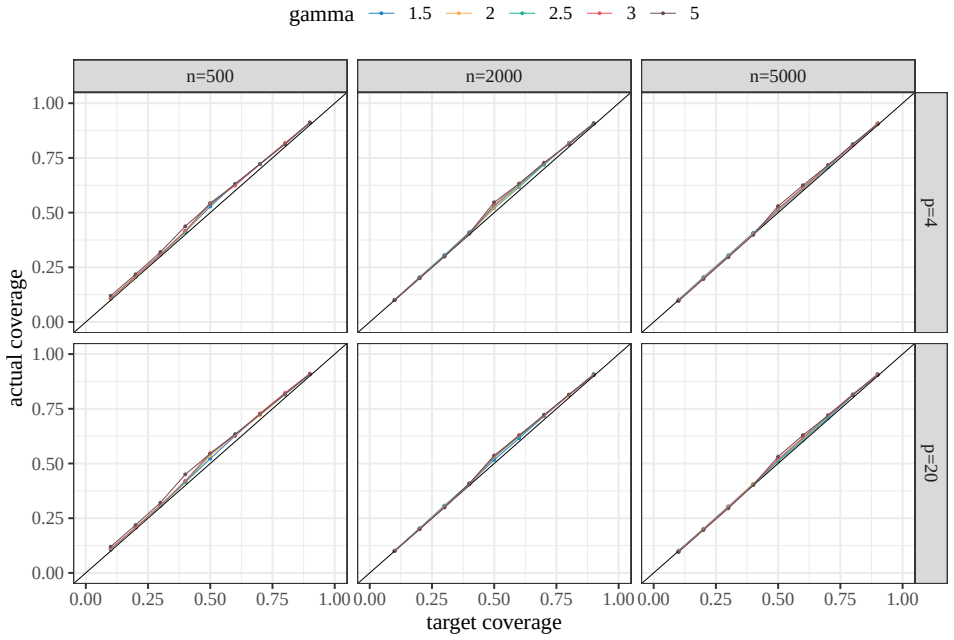

Fig. S2. 0.05-th quantile of empirical coverage on test samples in Algorithm 2 when $\ell(\cdot)$, $u(\cdot)$ are known. The details are otherwise the same as in Figure 5.

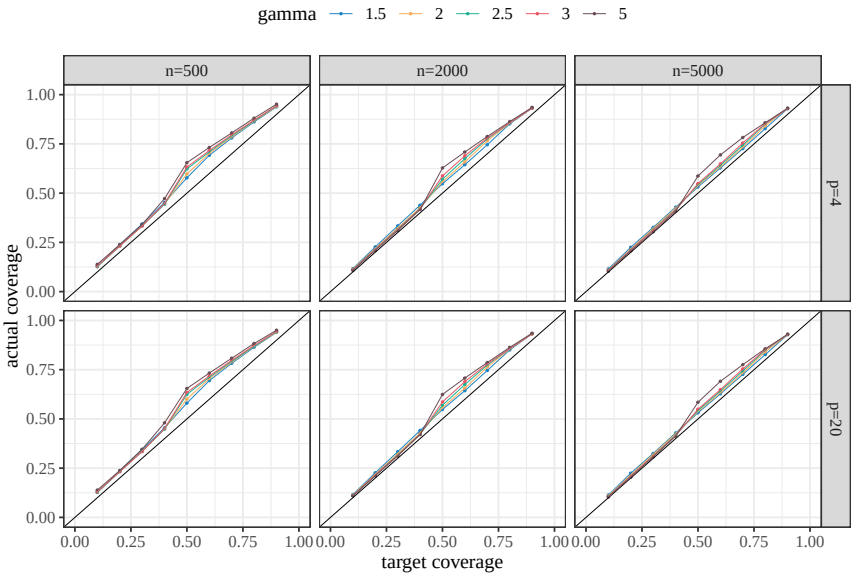

Fig. S3. Empirical (average) coverage of Algorithm 2 when $\ell(\cdot)$ and $u(\cdot)$ are known. The details are otherwise the same as in Figure 5.

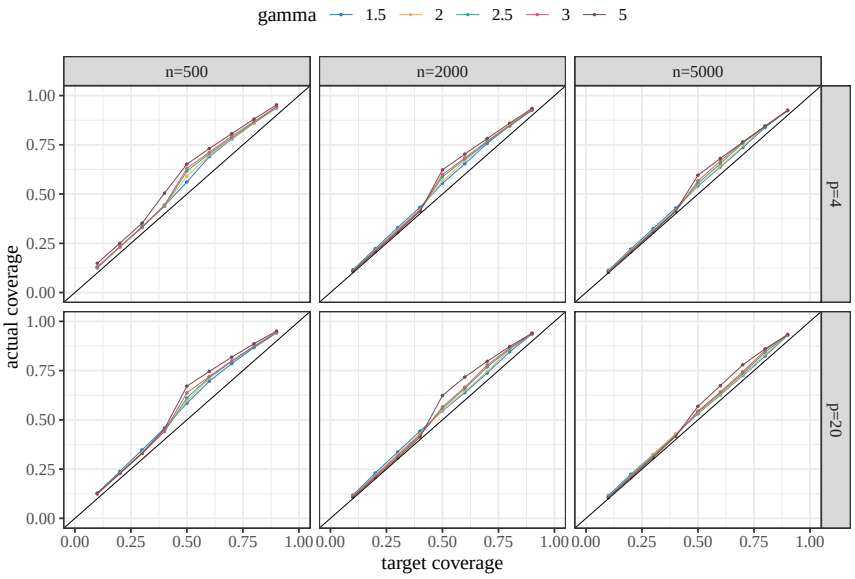

Fig. S4. Empirical (average) coverage of Algorithm 2 when $\hat{\ell}(\cdot)$ and $\hat{u}(\cdot)$ are estimated. The details are otherwise the same as in Figure 5.

C. Additional results for Section E. In this part, we collect the results in Section E of procedures with known bound functions.

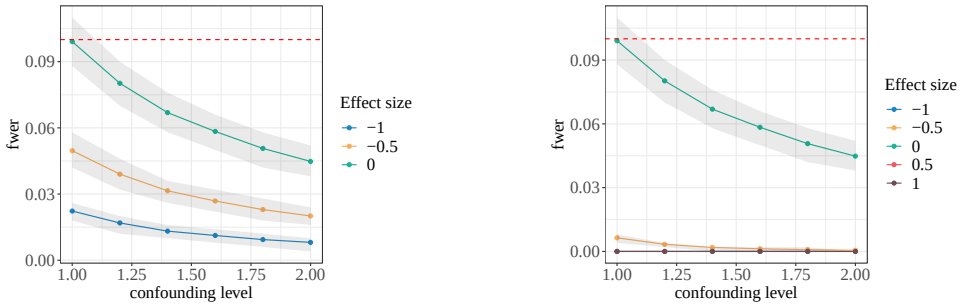

Fig. S5. Empirical FWER for fixed ITE (left) and random ITE (right) with Algorithm 1. The details are otherwise the same as in Figure S9.

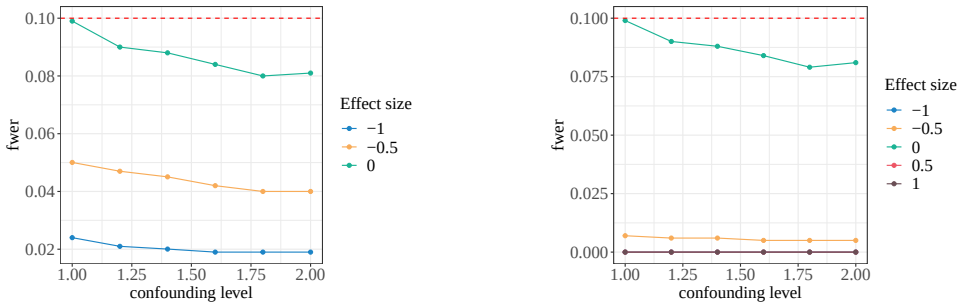

Fig. S6. Empirical FWER for fixed ITE (left) and random ITE (right) with Algorithm 2. The details are otherwise the same as in Figure S9.

D. Counterfactual prediction on a semi-real dataset. We consider an observational study dataset from (4), based on which we generate confounded synthetic potential outcomes, and conduct the ATE-type prediction of $Y(1)$.

Randomly splitting the original dataset into two folds of sizes $|\mathcal{Z}_1| = 2078$ and $|\mathcal{Z}_2| = 8313$ respectively, we use the samples in $\mathcal{Z}_1$ to fit a regression function $\hat{\mu}_0(x)$ for $\mathbb{E}[Y(0) \,|\, X = x, T = 0]$ and a propensity score function $\hat{e}(x)$ for $\mathbb{P}(T = 1 \,|\, X = x)$. We sample $n = 10000$ covariates $\{X_i\}_{1 \le i \le n}$ from $\mathcal{Z}_2$ with replacement. Then i.i.d. counterfactuals are generated via

$$Y_i(1) = \hat{\mu}_0(X_i) + \tau(X_i) + U_i, \quad Y_i(0) = \hat{\mu}_0(X_i) - U_i, \quad U_i \sim N(0, 0.2^2), \quad i = 1, \ldots, n,$$

where the conditional treatment effect function $\tau(x)$ is specified the same way as in equation (1) of (4). The propensity scores are specified as $e(X_i) := \hat{e}(X_i)$. For each confounding level $\Gamma \in \{1.5, 2, 2.5, 3, 5\}$, both $e(X_i, U_i)$ and T_i are generated the same way as in Section C.

We then conduct counterfactual inference on the synthetic dataset $\mathcal{D}_{\text{obs}} = \{Y_i, X_i, T_i\}_{1 \le i \le n}$. We randomly split $\mathcal{D}_{\text{obs}}$ into three folds with $1 : 2 : 1$ sizes. The treated samples in the first fold are used as $\mathcal{D}_{\text{train}}$. The treated samples in the second fold are used as $\mathcal{D}_{\text{calib}}$, so that $\mathcal{D} = \mathcal{D}_{\text{train}} \cup \mathcal{D}_{\text{calib}}$. All samples in the third fold (where we have ground truth even for those in the control group) are used as test samples. The process is repeated $N = 1000$ times, where there are approximately $|\mathcal{D}_{\text{calib}}| = 1900$ calibration samples fed into the procedures and 2500 test samples.

Figure S7 summarizes the empirical coverage using Algorithm 1 and estimated $\hat{\ell}(\cdot)$, $\hat{u}(\cdot)$. The output of Algorithm 1 always achieves valid coverage. The solid lines are close to the 45°-line, showing the tightness of our procedure in this setting.

The empirical coverage of Algorithm 2 with estimated $\hat{\ell}(\cdot)$, $\hat{u}(\cdot)$ is summarized in Figure S8, which validate the PAC-type guarantee (referring to the lower boundary of the shaded area in each plot, which is the 0.05-th quantile of empirical coverage). Due to the conservativeness of the confidence lower bound constructed by the WSR inequality, the average coverage of Algorithm 2 is a bit higher than the target for targets around 0.5. However, the 0.05-th quantile (lower boundary of the shaded area) is still very close to the target.

It is also worth pointing out that the shaded bands indicate the quantiles of the empirical coverage on test samples, which could be understood as an estimate of $\hat{c}(\mathcal{D}) := \mathbb{P}(Y_{n+1} \in \hat{C}(X_{n+1}) \,|\, \mathcal{D})$. Although PAC-type guarantee is not theoretically provided by Algorithm 1, most of the time, $\hat{c}(\mathcal{D})$ is above the target in this example (referring to the lower boundary of the shaded area, which is the 0.05-th quantile of $\hat{c}(\mathcal{D})$). The widths of the 0.05-th and the 0.95-th

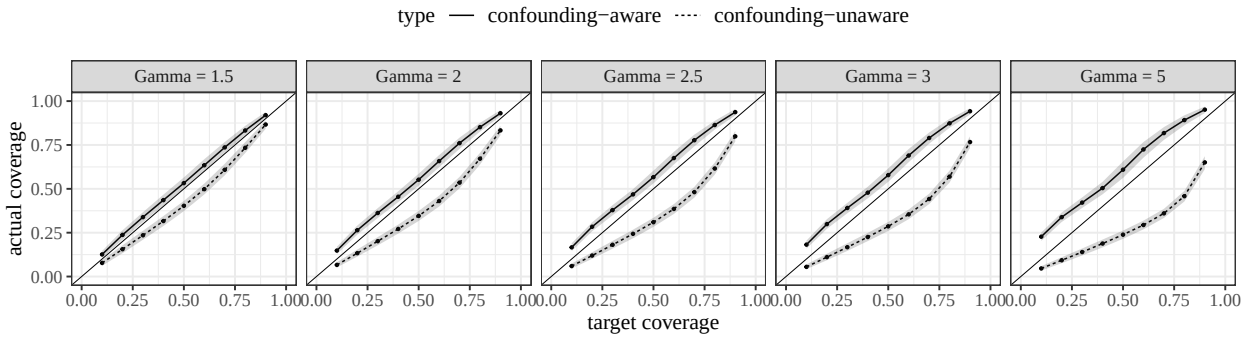

Fig. S7. Empirical coverage on the test sample. Each plot corresponds to a confounding level Γ. The points are average empirical coverage. The shaded bands corresponds to the 0.05-th and 0.95-th quantiles of coverage on test samples. The solid lines correspond to Algorithm 1. The dashed lines assume no confounding and are shown for comparison. In this case, counterfactual prediction intervals are invalid.

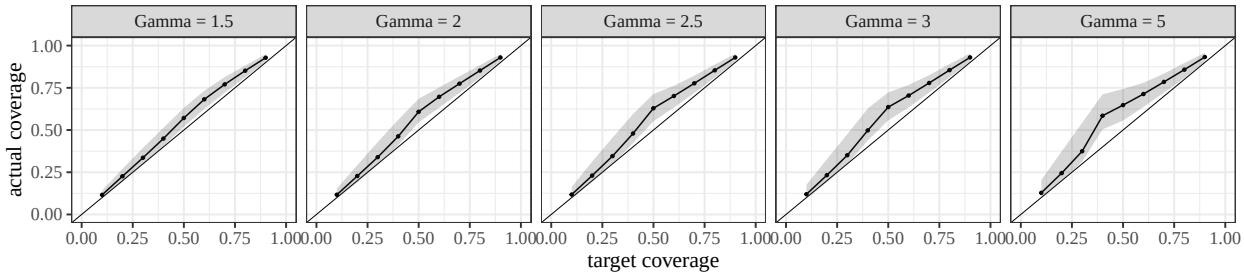

Fig. S8. Empirical coverage of Algorithm 2. Each plot corresponds to a confounding level Γ. The shaded bands corresponds to the 0.05-th and 0.95-th quantiles of coverage on test samples.

quantiles also provide empirical evidence that $\hat{c}(\mathcal{D})$ from Algorithm 1 might be less variable than Algorithm 2. The theoretical analysis of this phenomenon might deserve future investigation.

E. Simulations for sensitivity analysis. We focus on the ATT-type inference for the directional null

$$H_0(\Gamma)\colon Y(1) - Y(0) \le 0 \quad \text{and} \quad \mathbb{P}^{\text{sup}} \in \mathcal{P}(\Gamma).$$

The test sample is from $\mathbb{P}_{X,Y(0),Y(1)\,|\,T=1}$, for which we observe $(X_{n+1}, Y_{n+1}(1))$ and would like to predict $Y_{n+1}(0)$.

We fix $n_{\text{train}} = n_{\text{calib}} = 2000$ and $p = 4$. The covariates X, unobserved confounders U and counterfactual $Y(0)$ (instead of $Y(1)$) are generated in the same way as in Section C. The treatment mechanism $e(x, u)$ is also the same as in [12] with confounding level $\Gamma \in \{1.2, 1.4, \ldots, 2\}$. The training data are $(X_i, Y_i(0))$ for those $T_i = 0$. We generate $Y(1)$ in two ways: 1) $Y(1) - Y(0) \equiv a$ (fixed ITE) and 2) $Y(1) - Y(0) = a \cdot U$ (random ITE). Here a ranges in $\{-1, -0.5, 0, 0.5, 1\}$.

Fixing level $\alpha = 0.1$ and $\delta = 0.05$ (for Algorithm 2), for each fixed $\Gamma \ge 1$, we construct a one-sided ATT-type prediction interval for $Y(0)$, which takes the form $\hat{C}(X_{n+1}, \Gamma) = (-\infty, \hat{Y}(X_{n+1}, \Gamma)]$. The prediction interval is obtained via the non-conformity score

$$V(x, y) = y - \hat{q}(x, 1 - \alpha).$$

We reject no hypotheses if $Y_{n+1}(1) \le \hat{Y}(X_{n+1}, 1)$; otherwise, we reject all $H_0(\Gamma)$ such that $Y_{n+1}(1) > \hat{Y}(X_{n+1}, \Gamma)$, hence the rejection set is $\mathcal{R} = [1, \hat{\Gamma})$ for some $\hat{\Gamma} > 1$, which we define as the Γ-value.

E.1. FWER control. We evaluate the empirical FWER, which is the proportion of making a false rejection among $\{H_0(\Gamma)\}_{\Gamma \ge 1}$, averaged over all test samples in all $N = 1000$ independent runs. The results with estimated $\hat{\ell}(\cdot)$ and $\hat{u}(\cdot)$ are presented in Figures S9 and S10, showing control of the FWER even when the likelihood ratio bounds are estimated; we omit the case of fixed ITE at $a > 0$ since the hypotheses $H_0(\Gamma)$ are always false. The results with the ground truth $\ell(\cdot)$ and $u(\cdot)$ are in SI Appendix C.

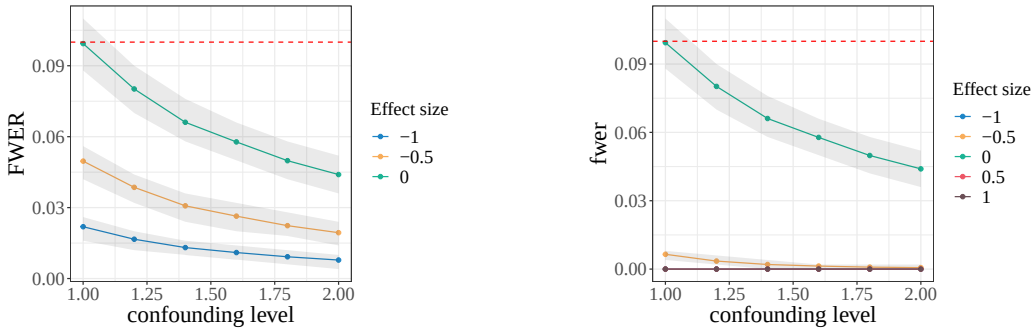

Fig. S9. Empirical FWER of Algorithm 1. The effect size a ranges in $\{-1, -0.5, 0, 0.5, 1\}$. The solid lines are averaged over $N = 1000$ runs, while the 0.25-th and 0.75-th quantiles form the shaded area.

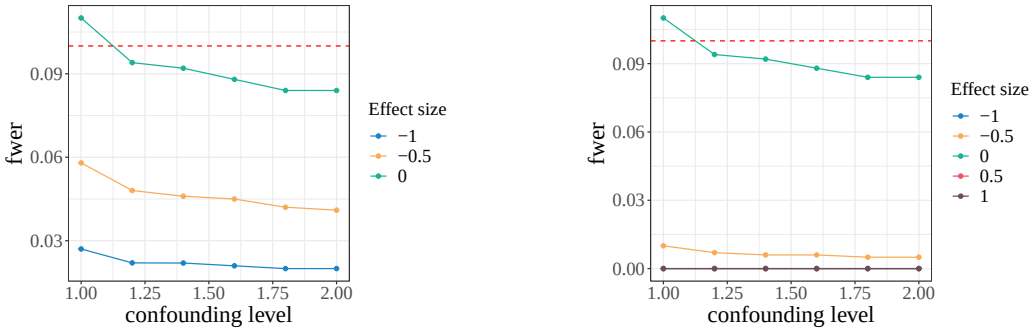

Fig. S10. 0.05-th quantile of empirical FWER using Algorithm 2 with estimated $\hat{\ell}(\cdot)$ and $\hat{u}(\cdot)$. with the effect size a ranging in $\{-1, -0.5, 0, 0.5, 1\}$ for fixed ITE (left) and random ITE (right).

E.2. **Γ-values .** We plot the estimated survival function $\hat{\Gamma}$ defined as $S(\Gamma) = \mathbb{P}(\hat{\Gamma} > \Gamma)$, which characterizes the proportion of test units that are identified as positive ITE with each confounding level Γ. Figure S11 presents the results from one run of the procedure with Algorithm 1, where we focus on the random ITE: $Y(1) - Y(0) = a \cdot U$.

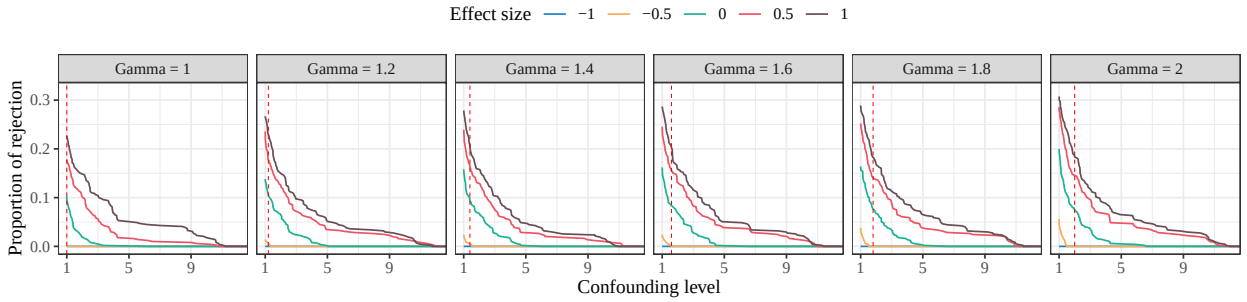

Fig. S11. Empirical evaluation of $S(\Gamma)$ reported by (one run of) the sensitivity analysis procedure with Algorithm 1. The red dashed vertical lines are the true confounding levels.

To see how to interpret these plots, let us consider the example where $a = 1$ and the true confounding level is 1.6. We can see that around 10% of the samples have a Γ-value greater than 2.5, and 5% have a Γ-value greater than 5, showing strong evidence for positive ITEs. Note also that the ITE is always zero when $a = 0$, and the Γ-value should be a 90% lower confidence bound for the true confounding level. In Figure S11, we indeed observe that for around 90% of the samples, the Γ-value is below the true confounding level (see the green curves).

With random ITEs, both the magnitude of actual ITE and the gap between observed outcomes in treated and control groups increase with the effects size a. Within each subplot, the reported Γ-values become larger as the effect size increases. Thresholding at the true confounding level Γ, we see that larger magnitude of true effects also makes it easier to detect positive ITEs at the correct confounding level Γ.

The results from one run of Algorithm 2 are in Figure S12. The patterns are similar to Figure S11 in general, except that it is sometimes sharper than Algorithm 1 and provides slightly stronger evidence against unmeasured confounding.

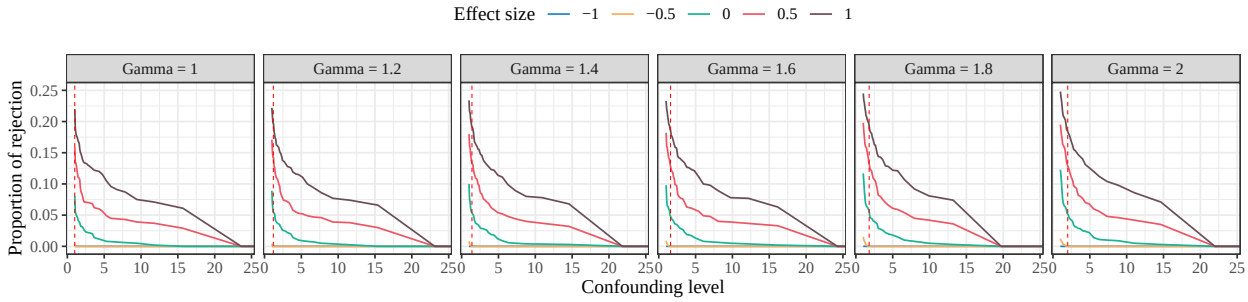

Fig. S12. Empirical evaluation of $S(\Gamma)$ reported by (one run of) the sensitivity analysis procedure with Algorithm 2. The red dashed vertical lines are the true confounding levels.

In Figure S12, some test units have large Γ-values especially when the effect size is positive. Such strong evidence would happen if there is a large gap between the observed $Y(1)$ and the typical behavior of $Y(0)$ predicted with the training data—thus, the only way our procedures can output a prediction interval that overlaps with $Y(1)$ is $\hat{C}(X, \Gamma) = \{V(X, y) \leq \hat{v}\} = \mathbb{R}$, where $\hat{v} = \infty$. In that case, there is a certain value $\Gamma_\infty(X)$ such that $\hat{v} = \infty$ once $\Gamma \geq \Gamma_\infty(X)$. Since Algorithm 2 outputs a same $\hat{v}$ for all test samples, the value Γ_∞ is the same for all such extreme individuals, leading to steep tails in Figure S12. In contrast, Algorithm 1 returns different $\Gamma_\infty(X)$ for different individuals, and produces a smoother curve (Figure S11).

E.3. ***False discovery proportions.*** We also track the empirical false discovery proportion $\mathrm{FDP}(\Gamma) = \frac{|\{j \in \mathcal{D}_{\mathrm{test}} : \hat{\Gamma} > \Gamma,\ Y_j(1) \leq Y_j(0)\}|}{|\{j \in \mathcal{D}_{\mathrm{test}} : \hat{\Gamma} > \Gamma\}|}$ for random ITE with $a \neq 0$, which is the proportion of false rejections among test units that are rejected at confounding level Γ. With both the two procedures, we find that $\mathrm{FDP}(\Gamma)$ is always zero—the units that survive certain levels

of adjustment for confounding all have positive ITE. It could be explained by the conservativeness of the procedure: to survive the adjustment, the observed $Y(1)$ needs to be larger than the whole $1-\alpha$ prediction interval for $Y(0)$. Therefore, a unit that survives the adjustment is much more likely to have a positive ITE, leading to vanishing FDPs.

References

1. I Waudby-Smith, A Ramdas, Estimating means of bounded random variables by betting (2021).
2. Y Romano, E Patterson, E Candes, Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* **32**, 3543–3553 (2019).
3. RJ Tibshirani, RF Barber, EJ Candès, A Ramdas, Conformal prediction under covariate shift in *Advances in Neural Information Processing Systems 32*. (2019).
4. C Carvalho, A Feller, J Murray, S Woody, D Yeager, Assessing treatment effect variation in observational studies: Results from a data challenge. *Obs. Stud.* **5**, 21–35 (2019).