

Outlier Resistant Inference for Conditional Average Treatment Effect

Indiana University Indianapolis

Supplementary Material

Technical proofs, more simulation studies, and the full DR-learner and machine-learning meta-learner comparison studies, together with their implementation details, are provided in the supplementary material.

S1 The Proof of Theorem 1

$$\begin{aligned}
& E \left[\frac{T}{\pi(X)} \psi(Y - \mu_1(x_1)) \middle| X_1 = x_1 \right] \\
&= E \left[E \left[\frac{T}{\pi(X)} \psi(Y - \mu_1(x_1)) \middle| X \right] \middle| X_1 = x_1 \right] \\
&= E \left[E \left[\frac{T}{\pi(X)} \psi(Y^{(1)} - \mu_1(x_1)) \middle| X \right] \middle| X_1 = x_1 \right] \\
&= E \left[\frac{E[T|X]}{\pi(X)} E \left[\psi(Y^{(1)} - \mu_1(x_1)) \middle| X \right] \middle| X_1 = x_1 \right] \\
&= E \left[E \left[\psi(Y^{(1)} - \mu_1(x_1)) \middle| X \right] \middle| X_1 = x_1 \right] \\
&= E \left[\psi(Y^{(1)} - \mu_1(x_1)) \middle| X_1 = x_1 \right] \\
&= \int \psi(y^{(1)} - \mu_1(x_1)) f(y^{(1)} | x_1) dy^{(1)}
\end{aligned}$$

Since $\psi(y^{(1)} - \mu_1(x_1))$ is antisymmetric and $f(y^{(1)} | x_1)$ is symmetric with respect to $\mu_1(x_1)$, we have:

$$\int \psi(y^{(1)} - \mu_1(x_1)) f(y^{(1)} | x_1) dy^{(1)} = 0.$$

■

S2 The Proof of Theorem 2(1)

When $\sum_{Y_i \in S_{Y_c}} W_i < A(\sum_{Y_i \in S_{Y_c}} W_i + \sum_{Y_i \in S_{Y_n}} W_i)$, there exists $\delta > 0$ s.t. $\sum_{Y_i \in S_{Y_c}} W_i + \sum_{Y_i \in S_{Y_n}} W_i \delta < A(\sum_{Y_i \in S_{Y_c}} W_i + \sum_{Y_i \in S_{Y_n}} W_i)$.

Let C be such that $\rho(x) \geq -\delta$ for $|x| \geq C$, let t be any real number such that $|y - t| \geq C$ for all $y \in Y$, then we have

$$\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - \mu) W_{Y_i} \leq -A \left(\sum_{Y_i \in S_{Y_c}} W_i + \sum_{Y_i \in S_{Y_n}} W_i \right), \quad (\text{S2.1})$$

$$\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - t) W_{Y_i} \geq -\delta \sum_{Y_i \in S_{Y_n}} W_i - \sum_{Y_i \in S_{Y_c}} W_i. \quad (\text{S2.2})$$

Hence we have

$$\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - \mu) W_i < \sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - t) W_i. \quad (\text{S2.3})$$

Then μ must fall within distance C from a point in Y

When $\sum_{Y_i \in S_{Y_c}} W_i > A(\sum_{Y_i \in S_{Y_c}} W_i + \sum_{Y_i \in S_{Y_n}} W_i)$.

Let $\delta > 0$ be such that $\sum_{Y_i \in S_{Y_c}} W_i - \delta \sum_{Y_i \in S_{Y_c}} W_i > A(\sum_{Y_i \in S_{Y_c}} W_i + \sum_{Y_i \in S_{Y_n}} W_i)$

Let C be such that $\rho(x) \geq -\delta$ for $|x| \geq C$.

Let k be any real number and assume that all points in S_{Y_c} are equal to k .

Assume all points in S_{Y_c} are equal to y^* . Then, for all t with $|Y_i - t| \geq C$, we obtain

$$\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - t) W_i \geq -A \left(\sum_{Y_i \in S_{Y_c}} W_i + \sum_{Y_i \in S_{Y_n}} W_i \right) - \delta \sum_{Y_i \in S_{Y_c}} W_i \quad (\text{S2.4})$$

$$\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - y^*) W_i \leq - \sum_{Y_i \in S_{Y_c}} W_i \quad (\text{S2.5})$$

Then we have

$$\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - y^*) W_i < \sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - t) W_i \quad (\text{S2.6})$$

Hence μ must in distance C from y^* . Let $y^* \rightarrow +\infty$ we have $\mu \rightarrow +\infty$.

■

S3 The Proof of Theorem 2(2)

By Lemma 4.2 in Huber (1984), we can modify their Lemma 4.3 by:

Lemma A1. *Put*

$$\Delta_{Y \cup S_{Y_c}}(t) = \sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} \rho(Y_i - t) - \rho(Y_i)$$

Then there is a constant C which depends on Y and on $\sum_{Y_i \in S_{Y_c}} W_i$ but not on the actual values in S_{Y_c} , such that for all t

$$\rho(t) \left(\sum_{Y_i \in S_{Y_n}} W_i - \sum_{Y_i \in S_{Y_c}} W_i \right) - C \leq \Delta_{Y \cup S_{Y_c}}(t) \leq \rho(t) \left(\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} W_i \right) + C$$

Proof:

$$\begin{aligned} \Delta_Y(t) &= \sum_{Y_i \in S_{Y_n}} [\rho(Y_i - t) - \rho(Y_i)] W_i \\ &= \rho(t) \sum_{Y_i \in S_{Y_n}} W_i + \sum_{Y_i \in S_{Y_n}} [\rho(Y_i - t) - \rho(t)] W_i - \sum_{Y_i \in S_{Y_n}} \rho(Y_i) W_i. \end{aligned}$$

Since $|\rho(y - t) - \rho(t)| = |\rho(t) - \rho(t - y)| \leq |y| \psi(y_0)$, we have $|\Delta_Y(t) -$

$$\rho(t) \sum_{Y_i \in S_{Y_n}} W_i| \leq C_1 \text{ with } C_1 = \sum_{Y_i \in S_{Y_n}} \rho(Y_i) W_i + \sum_{Y_i \in S_{Y_n}} |Y_i| \psi(y_0) W_i.$$

On the other hand

$$\begin{aligned} |\Delta_{S_{Y_c}}(t)| &= \left| \sum_{Y_i \in S_{Y_c}} [\rho(Y_i - t) - \rho(Y_i)] W_i \right| \leq M(t) \sum_{Y_i \in S_{Y_c}} W_i \\ &= (\rho(t) + \eta(t)) \sum_{Y_i \in S_{Y_c}} W_i. \end{aligned}$$

Since $\eta(t)$ is bounded $|\Delta_{S_{Y_c}}(t)| \leq \rho(t) \sum_{Y_i \in S_{Y_c}} W_i + C_2 \sum_{Y_i \in S_{Y_c}} W_i$.

Let $C = C_1 + C_2 \sum_{Y_i \in S_{Y_c}} W_i$ we have

$$\rho(t) \left(\sum_{Y_i \in S_{Y_n}} W_i - \sum_{Y_i \in S_{Y_c}} W_i \right) - C \leq \Delta_{Y \cup S_{Y_c}}(t) \leq \rho(t) \left(\sum_{Y_i \in S_{Y_n} \cup S_{Y_c}} W_i \right) + C.$$

■

With the modified Lemma 4.3, we prove Theorem 3.2.

Proof:

When

$$\sum_{Y_i \in S_{Y_c}} W_i < \sum_{Y_i \in S_{Y_n}} W_i,$$

we have $\Delta_{Y \cup S_{Y_c}}(t)$ be bounded away from 0 for sufficiently large t , uniformly in Y .

Since $\Delta_{Y \cup S_{Y_c}}(0) = 0$ and since $\Delta_{Y \cup S_{Y_c}}(t)$ reaches its absolute minimum at μ , it follows that μ cannot be outside a certain bounded neighborhood of 0.

Hence we have μ bounded when

$$\sum_{Y_i \in S_{Y_c}} W_i < \sum_{Y_i \in S_{Y_n}} W_i.$$

Now we consider the opposite case, when $\frac{\sum_{Y_i \in S_{Y_n}} W_i}{\sum_{Y_i \in S_{Y_c}} W_i} < 1$.

Let $l(\mu) = \sum_{S_{Y_n}} \rho(Y_i - \mu) W_i + \sum_{S_{Y_c}} \rho(Y_i - \mu) W_i$ denote the loss function.

Let all the Y_i in S_{Y_c} equals s_{Y_c} and solve the inequality $l(t) > l(s_{Y_c})$, that is

$$\sum_{S_{Y_n}} \rho(Y_i - t) W_i + \sum_{S_{Y_c}} \rho(Y_i - t) W_i > \sum_{S_{Y_n}} \rho(Y_i - s_{Y_c}) W_i + \sum_{S_{Y_c}} \rho(Y_i - s_{Y_c}) W_i,$$

we have

$$\frac{\rho(s_{Y_c} - t)}{\rho(Y_i - s_{Y_c})} > \frac{\sum_{Y_i \in S_{Y_n}} W_i}{\sum_{Y_i \in S_{Y_c}} W_i}.$$

By property of ρ , we have $\frac{\rho(s_{Y_c} - t)}{\rho(Y_i - s_{Y_c})} \rightarrow 1$ as $s_{Y_c} \rightarrow \infty$. Then for any fixed t the inequality holds under sufficiently large s_{Y_c} . Then we have for any fixed t , the loss function l does not attain its minimum at t .

Hence we have the estimator unbounded.

■

To completely show the breakdown property of μ under condition $2(U\rho)$ we need to show the estimator μ is going to be unbounded for some $Y_i \in S_{Y_c}$ when

$$\sum_{Y_i \in S_{Y_c}} W_i \geq \sum_{Y_i \in S_{Y_n}} W_i.$$

To do this, we try to reverse the set of observations S_Y and S_{Y_c} by the following settings:

Let $Y_i = s_{Y_c}$ for all $Y_i \in s_{Y_c}$,

let $\tilde{y}_i = Y_i - s_{Y_c} \in \tilde{S}_Y$ for $Y_i \in S_{Y_n}$,

and let $\tilde{y}_i = Y_i - s_{Y_c} \in \tilde{S}_Y^*$ for $Y_i \in S_{Y_c}$.

For the estimator, we let $\tilde{\mu} = \mu - s_{Y_c}$.

When $\tilde{\mu}$ is bounded, we have $\mu \rightarrow \infty$ as $s_{Y_c} \rightarrow \infty$.

Then by the Theorem 2, we have $\tilde{\mu}$ bounded for any choice of $\tilde{y}_i \in \tilde{Y}$ when

$$\sum_{Y_i \in \tilde{Y}^*} W_i > \sum_{Y_i \in \tilde{Y}} W_i.$$

This is the same as we have μ unbounded for some choice of s_{Y_c} when

$$\sum_{Y_i \in S_{Y_c}} W_i > \sum_{Y_i \in S_{Y_n}} W_i.$$

■

S4 The Proof of Theorem 2(3)

For the estimating equation:

$$0 = \sum_{Y_i \in S_{Y_n}} W_i \psi(Y_i - \mu) + \sum_{Y_i \in S_{Y_c}} W_i \psi(Y_i - \mu)$$

First, we assume μ is finite. Then for some C we have $|\mu| \leq C$ for all $Y_i \in S_{Y_c}$.

Hence we can take $Y_i = S_{Y_c}$ for $\forall Y_i \in S_{Y_c}$.

So that

$$0 = \sum_{Y_i \in S_{Y_n}} W_i \psi(Y_i - \mu) + \sum_{Y_i \in S_{Y_c}} W_i \psi(S_{Y_c} - \mu)$$

Let $S_{Y_c} \rightarrow \infty$. Since μ is bounded, we have $\psi(S_{Y_c} - \mu) \rightarrow k_2$.

And since $\psi \geq -k_1$, we have

$$0 \geq -k_1 \sum_{Y_i \in S_{Y_n}} W_i + k_2 \sum_{Y_i \in S_{Y_c}} W_i,$$

$$k_1 \sum_{Y_i \in S_{Y_n}} W_i \geq k_2 \sum_{Y_i \in S_{Y_c}} W_i.$$

And let $y^{*} \rightarrow -\infty$, we have $\psi(S_{Y_c} - \mu) = -K$ and since $\psi \leq k_2$, we have

$$0 \leq k_2 \sum_{Y_i \in S_{Y_n}} W_i - k_1 \sum_{Y_i \in S_{Y_c}} W_i,$$

$$k_1 \sum_{Y_i \in S_{Y_c}} W_i \leq k_2 \sum_{Y_i \in S_{Y_n}} W_i.$$

Under condition 3, we have $k_1, k_2 > 0$ then we have: when μ is bounded, we have

$$\frac{\sum_{Y_i \in S_{Y_n}} W_i}{\sum_{Y_i \in S_{Y_c}} W_i} \geq \max \left\{ \frac{k_2}{k_1}, \frac{k_1}{k_2} \right\}.$$

On the other hand, suppose there exists a sequence y_n such that μ_n is unbounded as a solution of

$$0 = \sum_{Y_i \in S_{Y_n}} W_i \psi(Y_i - \mu) + \sum_{Y_i \in S_{Y_c}} W_i \psi(y_n - \mu).$$

Suppose there is a subsequence of $\{\mu_n\}$ tending to $+\infty$.

Then for this subsequence, $Y_i - \mu_n \rightarrow -\infty$ for $Y_i \in S_{Y_n}$, and since $\psi \leq k_2$ we have

$$0 \leq -k_1 \sum_{Y_i \in S_{Y_n}} W_i + k_2 \sum_{Y_i \in S_{Y_c}} W_i,$$

$$k_1 \sum_{Y_i \in S_{Y_n}} W_i \leq k_2 \sum_{Y_i \in S_{Y_c}} W_i.$$

Or if the subsequence $\mu_n \rightarrow +\infty$ for $Y_i \in S_{Y_n}$, and since $\psi \geq k_1$ we have

$$0 \geq k_2 \sum_{Y_i \in S_{Y_n}} W_i - k_1 \sum_{Y_i \in S_{Y_c}} W_i,$$

$$k_1 \sum_{Y_i \in S_{Y_c}} W_i \geq k_2 \sum_{Y_i \in S_{Y_n}} W_i.$$

Then we have

$$\frac{\sum_{Y_i \in S_{Y_n}} W_i}{\sum_{Y_i \in S_{Y_c}} W_i} \leq \frac{k_1}{k_2},$$

or

$$\frac{\sum_{Y_i \in S_{Y_n}} W_i}{\sum_{Y_i \in S_{Y_c}} W_i} \leq \frac{k_2}{k_1}.$$

That is

$$\frac{\sum_{Y_i \in S_{Y_n}} W_i}{\sum_{Y_i \in S_{Y_c}} W_i} \leq \max \left\{ \frac{k_1}{k_2}, \frac{k_2}{k_1} \right\}.$$

■

S5 The Proof of Lemma 1

For simplicity, we will demonstrate the case when $d = 1$ and $s = 2$, as the proof can be easily extended to higher orders and dimensions.

Proof:

$$E[h(x_1, s)] = \int \delta(x_1 - X_{1i}) E \left[\psi(Y_i - s) \frac{T_i}{\hat{\pi}(X_i)} \middle| X_{1i} \right] g(X_{1i}) dX_{1i}$$

Since

$$\frac{1}{\hat{\pi}(X_i)} = \frac{1}{\pi(X_i)} + \frac{-1}{\hat{\pi}(X_i)\pi(X_i)}(\hat{\pi}(X_i) - \pi(X_i))$$

We can rewrite $E[h(x_1, s)] = \text{term1} - \text{term2}$

Where

$$\text{term1} = \int \delta_n(x_1 - X_{1i}) E \left[\psi(Y_i - s) \frac{T_i}{\pi(X_i)} \middle| X_{1i} \right] g(X_{1i}) dX_{1i}$$

and

$$term2 = \int \delta_n(x_1 - X_{1i}) E \left[\psi(Y_i - s) \frac{T_i}{\hat{\pi}(X_i)\pi(X_i)} (\hat{\pi}(X_i) - \pi(X_i)) | X_i \right] g(X_{1i}) dX_{1i}$$

Since in term2, $\psi(Y_i - s)$, T_i , $\frac{1}{\hat{\pi}(X_i)\pi(X_i)}$, $\int \delta_n(x_1 - X_{1i}) dX_{1i}$ are all bounded,

and by Assumption (A3), we have $\sup_{x \in \mathcal{X}} |\hat{\pi}(X) - \pi(X)| = O_p(\sqrt{n})$, we have

$$term2 \xrightarrow{P} 0,$$

For term1, by Härdle (1984) Lemma 2.2, we have

$$term1 \rightarrow E \left[\psi(Y_i - s) \frac{T_i}{\pi(X_i)} \middle| X_{1i} = x_1 \right] g(x_1)$$

$$\text{Hence } E[h(x_1, s)] \xrightarrow{n \rightarrow \infty} H(x_1, s).$$

Let $H_{in}(X_i, s) = \psi(Y_i - s) \frac{T_i}{\hat{\pi}(X_i)} \delta_n(x_1 - X_{1i})$, then

$$Var(h(x_1, s)) = \frac{Var(H_{in}(X_i, s))}{n} = \frac{E(H_{in}(X_i, s)^2)}{n} - \frac{E(H_{in}(X_i, s))^2}{n}$$

where

$$\begin{aligned} E[H_{in}(x_1, s)^2] &= E [E [H_{in}(X_i, s)^2 | X_{1i}]] \\ &= E \left[\left(\psi(Y_i - s) \frac{T_i}{\hat{\pi}(X_i)} \delta_n(x_1 - X_{1i}) \right)^2 \middle| X_{1i} \right] \\ &= \delta_n(x_1 - X_{1i})^2 E \left[\left(\psi(Y_i - s) \frac{T_i}{\hat{\pi}(X_i)} \right)^2 \middle| X_{1i} \right] \end{aligned}$$

Again, we can expand the square of inverse propensity score as:

$$\frac{1}{\hat{\pi}(X_i)^2} = \frac{1}{\pi(X_i)^2} + \frac{-2}{\xi^3(X_i)} (\hat{\pi}(X_i) - \pi(X_i))$$

where $\xi(X_i)$ is a value between $\pi(X_i)$, $\hat{\pi}(X_i)$.

Then $E[H_{in}(x_1, s)^2]/n = \text{term3} + \text{term4}$

where

$$\text{term3} = \frac{1}{n} \int \delta_n(x_1 - X_{1i})^2 E \left[\psi^2(Y_i - s) \frac{T_i}{\pi(X_i)^2} \middle| X_{1i} = x \right] g(x_1) dx$$

$$\begin{aligned} \text{term4} = & \frac{1}{n} \int \delta_n(x_1 - X_{1i})^2 E \left[\psi^2(Y_i - s) T_i \left(-\frac{2}{\xi(X_i)^3} \right) \right. \\ & \left. \cdot (\hat{\pi}(X_i) - \pi(X_i)) \middle| X_{1i} = x \right] g(x_1) dx \end{aligned}$$

Similar to term 1 and term 2, for term 4, we have ψ^2, T_i bounded, and hence

we just need to show that

$$\begin{aligned} & (\hat{\pi}(X_i) - \pi(X_i)) \delta_n(x_1 - X_{1i})^2 / n \\ & \leq |\hat{\pi}(X_i) - E[\hat{\pi}(X_i)]| \delta_n(x_1 - X_{1i})^2 / n + |E[\hat{\pi}(X_i)] - \pi(X_i)| \delta_n(x_1 - X_{1i})^2 / n \\ & \xrightarrow{P} 0 \end{aligned}$$

By the definition of δ_n , we have $\delta_n^2 \leq Ch^{-2}$

For the first part, by chebyshev's inequality

$$\begin{aligned} P \left(|\hat{\pi}(X_i) - E[\hat{\pi}(X_i)]| \frac{C}{nh^2} \geq \epsilon \right) &= P \left(|\hat{\pi}(X_i) - E[\hat{\pi}(X_i)]| \geq \frac{\epsilon h^2 n}{C} \right) \\ &\leq \frac{\sigma_n^2 C^2}{\epsilon^2 h^2 n} \rightarrow 0 \end{aligned}$$

Under assumption $nh^2 \rightarrow \infty$ and $\sigma_n^2 = O(1/n)$, we have $\frac{\sigma_n^2 C^2}{\epsilon^2 h^2 n} \rightarrow 0$

We have $|\hat{\pi}(X_i) - E[\hat{\pi}(X_i)]| \frac{C}{nh^2} \xrightarrow{P} 0$.

Hence the first part converge to 0 in probability.

And for the second part, we assumed $|E[\hat{\pi}(X_i)] - \pi(X_i)| = O(1/n)$, we

have

$$|E[\hat{\pi}(X_i)] - \pi(X_i)| \frac{C}{nh^2} \rightarrow 0$$

Hence, we have $term4 \xrightarrow{P} 0$

For term3, define $\delta^*(u) = \frac{\delta_n^2}{\alpha_n(2)}$ where $\alpha_n(p) = \int |\delta_n(u)|^p du$, by Lemma

2.1 of (Hardle 1984) $\delta^*(u)$ is again a DFS.

Then

$$term3 = \alpha_n(2) \int \delta_n^*(x_1 - X_{1i}) E \left[\psi^2(Y_i - s) \frac{T_i}{\pi(X_i)^2} \middle| X_{1i} = x \right] g(x_1) dx$$

Then

$$\frac{term3}{\alpha_n(2)} \xrightarrow{P} E \left[\psi^2(Y_i - s) \frac{T_i}{\pi(X_i)^2} \middle| X_{1i} = x \right] g(x_1)$$

and hence

$$\begin{aligned} \frac{\text{Var}(H_{ni}(x, s))}{\alpha_n(2)} &\rightarrow E \left[\psi^2(Y_i - s) \frac{T_i}{\pi(X_i)^2} \middle| X_{1i} = x_1 \right] g(x_1) \\ &\quad - \frac{1}{\alpha_\infty(2)} \left(E \left[\psi(Y_i - s) \frac{T_i}{\pi(X_i)} \middle| X_{1i} = x_1 \right] g(x_1) \right)^2, \end{aligned}$$

where $\alpha_\infty(2) = \lim_{n \rightarrow \infty} \alpha_n(2)$ and by (Hardle 1984 Lemma 2.1) we have

$$\alpha_\infty(2) = \infty.$$

Hence

$$\frac{n \text{Var}(H_n(x, s))}{\alpha_n(2)} \rightarrow E \left[\psi^2(Y_i - s) \frac{T_i}{\pi(X_i)^2} \middle| X_{1i} = x \right] g(x_1).$$

Then by Chebyshev Inequality

$$P(|h(x_1, s) - E(h(x_1, s))| > \epsilon) \leq \frac{\text{Var}(h(x_1, s))}{\epsilon^2}.$$

Assume $\lim_{n \rightarrow \infty} \frac{\alpha_n(2)}{n} = 0$,

we have $\frac{n \text{Var}(H_n(x, s))}{\alpha_n(2)} \rightarrow C(x) < \infty$

$\text{Var}(H_n(x, s)) \rightarrow 0$

Then we have

$$\lim_{n \rightarrow \infty} P(|h(x_1, s) - E(h(x_1, s))| > \epsilon) = 0$$

Hence $h(x_1, s) \xrightarrow{P} E[h(x_1, s)]$

and since $E[h(x_1, s)] \rightarrow H(x_1, s)$

We have $h(x_1, s) \xrightarrow{P} H(x_1, s)$.

■

S6 The Proof of Theorem 3

Based on the continuous and strictly increased assumption of $\psi(\cdot)$, we have

$H(x_1, s)$ and $h(x_1, s)$ also monotone and continuous.

By intermediate theorem, for any $\epsilon > 0$ we have the event

$$h(x_1, c - \epsilon) < 0 \cap h(x_1, c + \epsilon) > 0 \text{ implies } c - \epsilon < \hat{\mu}_1(x_1) < c + \epsilon. \quad (\text{S6.7})$$

Then

$$\begin{aligned} P(h(x_1, c + \epsilon) < 0 \cap h(x_1, c - \epsilon) > 0) &\leq P(c - \epsilon < \hat{\mu}_1(x_1) < c + \epsilon) \\ &= P(|\hat{\mu}_1(x_1) - c| < \epsilon) \end{aligned}$$

Since we've showed $h(x_1, s) \xrightarrow{P} H(x_1, s)$ we have

$$\begin{aligned} P(h(x_1, c - \epsilon) < 0 \cap h(x_1, c + \epsilon) > 0) &\rightarrow \\ P(H(x_1, c - \epsilon) < 0 \cap H(x_1, c + \epsilon) > 0) &= 1 \end{aligned}$$

Hence $\lim_{n \rightarrow \infty} P(|\hat{\mu}_1(x_1) - c| < \epsilon) = 1 \Rightarrow \hat{\mu}_1(x_1) \xrightarrow{P} c$.

In Theorem 2.1, we've shown that $\mu_1(x_1)$ is a solution of $H(x_1, s)$, and there's only one solution for monotone continuous function, we have $\mu_1(x_1) = c$.

■

S7 The Proof of Theorem 4

Proof: By the mean value theorem, we have

$$\hat{\mu}_1(x_1) - \mu(x_1) = h(x_1, \mu_1(x_1)) / D_n(x_1).$$

Where

$$\begin{aligned} D_n(x_1) = \frac{1}{n} \sum_{i=1}^n \delta_n(x_1 - X_{1i}) \frac{T_i}{\hat{\pi}(X_i)} \psi' \left(Y_i - \mu_1(x_1) \right. \\ \left. - W_i(\hat{\mu}_1(x_1) - \mu_1(x_1)) \right), \quad W_i \in (0, 1) \end{aligned}$$

For the denominator $D_n(x)$, we can expand $D_n(x) = term5 + term6$

Where

$$\begin{aligned}
term5 &= \frac{1}{n} \sum_{i=1}^n \delta_n(x_1 - X_{1i}) \frac{T_i}{\pi(X_i)} \psi'(Y_i - \mu_1(x_1) - W_i(\hat{\mu}_1(x_1) - \mu_1(x_1))) \\
term6 &= \frac{1}{n} \sum_{i=1}^n \delta_n(x_1 - X_{1i}) \frac{-T_i}{\xi^2(X_i)} (\hat{\pi}(X_i) - \pi(X_i)) \\
&\quad \cdot \psi'(Y_i - \mu_1(x_1) - W_i(\hat{\mu}_1(x_1) - \mu_1(x_1)))
\end{aligned}$$

with $\epsilon(X_i) \in (\pi(X_i), \hat{\pi}(X_i))$.

For term 6, similar to Lemma 4.1, we have $term6 \xrightarrow{P} 0$.

For term 5, by $\hat{\mu}_1(x_1) \xrightarrow{P} \mu_1(x_1)$ and by WLLN, we have

$$E[term5] = E \left[\delta_n(x_1 - X_{1i}) \frac{T_i}{\pi(X_i)} \psi'(Y_i - \mu_1(x_1)) \right] (1 + o(1))$$

Then similar to what we did when proving consistency, by Härdle (1984)

Lemma 2.2, we have

$$term5 \xrightarrow{P} E \left[\frac{T_i}{\pi(X_i)} \psi'(Y_i - \mu_1(x_1)) \middle| X_{1i} = x_1 \right] g(x_1)$$

Then

$$D_n(x_1) \xrightarrow{P} E \left[\frac{T_i}{\pi(X_i)} \psi'(Y_i - \mu_1(x_1)) \middle| X_{1i} = x_1 \right] g(x_1) = C(x_1)g(x_1)$$

For the numerator, from our proof of consistency, we have

$$h(x_1, \mu) = \frac{1}{n} \sum_{i=1}^n \left[\delta_n(X_{1i} - x_1) \psi(Y_i - s) \frac{T_i}{\pi(X_i)} \right] (1 + o_p(1))$$

and

$$\frac{n}{\alpha_n(2)} \text{Var}(h(x_1, \mu)) \rightarrow E \left[\psi(Y_i - s)^2 \frac{T_i}{\pi(X_i)^2} \middle| X_{1i} = x_1 \right] g(x_1) = \sigma^2(x_1)g(x_1)$$

Hence, let $h^*(x_1) = \frac{1}{n} \sum_{i=1}^n \delta_n(X_{1i} - x_1) \psi(Y_i - s) \frac{T_i}{\pi(X_i)}$

$$H_{in}^*(x_1) = \delta_n(X_{1i} - x_1) \psi(Y_i - s) \frac{T_i}{\pi(X_i)}$$

$$B_n(x_1) = E(h^*(x_1))$$

$$W_n(x_1) = \frac{(H_n^*(x_1) - B_n(x_1))}{(\frac{\alpha n(2)}{n} \sigma^2(x_1) g(x_1))^{1/2}}$$

To verify the condition of Lyapunov CLT, we let $s_n^2 = \sum_{i=1}^n Var(h^*(x_1))_i^2$

And for some $\eta > 0$, we have

$$\begin{aligned} & \frac{1}{s_n^{2+\eta}} \sum_{i=1}^n E [|H_{in}^*(x_1) - B_n(x_1)|^{2+\eta}] \\ &= \frac{n}{s_n^{2+\eta}} E [|H_{in}^*(x_1) - B_n(x_1)|^{2+\eta}] \\ &= \gamma_n \frac{n}{s_n^{2+\eta}} E \left[\delta_n^*(X_{1i} - x_1) E \left[\left| \psi(Y_i - s) \frac{T_i}{\pi(X_i)} - B_n(x_1) \right|^{2+\eta} \middle| X_{1i} \right] \right] \end{aligned}$$

Where $\delta_n^*(\cdot) = |\delta_n(\cdot)|^{2+\eta} / \gamma_n$.

Similar to how we did in the proof of Lemma 1, under Assumption (2)

of this theorem:

$$\frac{1}{s_n^{2+\eta}} \sum_{i=1}^n E [|H_{in}^*(x_1) - B_n(x_1)|^{2+\eta}] = O \left(\frac{\gamma_n n}{s_n^{2+\eta}} \right) \rightarrow 0$$

Then by Lyapunov CLT, we have $Z_n \xrightarrow{D} N(0, 1)$.

■

S8 The Proof of $\hat{\tau}(x_1)$

For the normality, by our proof of the asymptotic properties of $h(x_1)$ we have

$$\begin{aligned}\hat{\tau}(x_1) - \tau(x_1) &= (\hat{\mu}_1(x_1) - \mu_1(x_1)) - (\hat{\mu}_0(x_1) - \mu_0(x_1)) \\ &= \left(\frac{H_n^{*1}(x_1)}{D^1(x_1)g(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1)g(x_1)} \right) (1 + o_p(1))\end{aligned}$$

Where

$$\begin{aligned}H_n^{*1} &= \frac{1}{nh} \sum_{i=1}^n K \left(\frac{X_{1i} - x_1}{h} \right) \left(\psi(Y_i - \mu_1(x_1)) \frac{T_i}{\pi(X_i)} \right) \\ H_n^{*0} &= \frac{1}{nh} \sum_{i=1}^n K \left(\frac{X_{1i} - x_1}{h} \right) \left(\psi(Y_i - \mu_0(x_1)) \frac{1 - T_i}{1 - \pi(X_i)} \right) \\ D^1(x_1) &= E \left[\frac{T_i}{\pi(X_i)} \psi'(Y_i - \mu_1(x_1)) \middle| X_{1i} = x_1 \right] \\ D^0(x_1) &= E \left[\frac{1 - T_i}{1 - \pi(X_i)} \psi'(Y_i - \mu_0(x_1)) \middle| X_{1i} = x_1 \right]\end{aligned}$$

Similar to the proof of the variance of $h(x_1)$ we have the variance

$$\frac{n}{\alpha_n(2)} \text{Var}(\hat{\tau}(x_1)) \xrightarrow{n \rightarrow \infty} E \left[\left(\frac{H_n^{*1}(x_1)}{D^1(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1)} \right)^2 \middle| X_{1i} = x_1 \right] g(x_1)^{-1}$$

Where

$$\begin{aligned}H^{*1}(x_1) &= \psi(Y_i - \mu_1(x_1)) \frac{T_i}{\pi(X_i)}, \\ H^{*0}(x_1) &= \psi(Y_i - \mu_0(x_1)) \frac{1 - T_i}{1 - \pi(X_i)}.\end{aligned}$$

Verify the condition similarly to Theorem 4.5, by Lyapunov's CLT, we have the normality:

$$\sqrt{\frac{ng(x_1)}{\alpha_n(2)\sigma^2(x_1)}}(\hat{\tau}(x_1) - \tau(x_1)) \xrightarrow{D} N(0, 1),$$

where

$$\sigma^2(x_1) = E \left[\left(\frac{H_n^{*1}(x_1)}{D^1(x_1)} - \frac{H_n^{*0}(x_1)}{D^0(x_1)} \right)^2 \middle| X_{1i} = x_1 \right].$$

S8.1 Simulation Result Verify the Asymptotic normality

For sample sizes $n \in \{500, 5000\}$ we estimate the CATE over 5 fixed points $x_1 \in \{-0.4, -0.2, 0, 0.2, 0.4\}$ with 3 types of loss functions:

1. L_2 : $\rho(x) = x^2$;
2. Huber loss function;
3. Tukey's loss function.

Where we use the L_2 loss function as an example of an unbounded loss function, Huber's loss function as a bounded loss function, and Tukey's loss function as an example of redescending loss function.

For the choice of bandwidths, we use fixed bandwidths $h = 0.04$ for $n = 500$ and $h = 0.025$ for $n = 5000$. These values are obtained by averaging the bandwidths selected through cross-validation across repetitions.

With 1000 repetitions, we output the simulation bias, simulation variance, simulation MSE, mean of the estimated variance, and the probability

of the 90% and 95% confidence interval constructed with estimated variances cover the truth.

Comparing $n = 500$ and $n = 5000$, we find that a larger sample size results in reduced bias, variance, and MSE, and improves the probability of achieving coverage rates closer to 95% and 90%. This indicates that the estimator approaches its theoretical normal distribution as the sample size increases.

S9 Comparison with DR-Learner Variants

This section and Section S10 present the full comparison studies summarised in Section 5.3 of the main paper. To further contextualise the proposed methods, we compare against three variants of the DR-learner, which represents a natural class of doubly-robust competitors. The three variants share the same first step (propensity score estimation via logistic regression) but differ in how they construct and smooth the pseudo-outcome.

Following Kennedy (2023), the **DR-learner** estimates the propensity score $\hat{\pi}(X)$ by logistic regression (truncated to $[0.01, 0.99]$) and the conditional mean outcomes $\hat{m}_1(X)$, $\hat{m}_0(X)$ by ordinary least squares (OLS)

Table S1: Simulation results with $n = 500$ and outlier distribution $\mathcal{N}(0, 10)$.

x1		-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0					0.1				
Bias	L2	0.046	0.048	-0.075	0.041	0.019	0.047	0.083	-0.105	0.045	0.005
	Huber	0.046	0.048	-0.075	0.041	0.019	0.050	0.072	-0.096	0.036	0.008
	Tukey	0.046	0.047	-0.074	0.041	0.018	0.055	0.060	-0.086	0.028	0.009
Var	L2	0.049	0.040	0.031	0.033	0.045	0.648	0.491	0.425	0.406	0.509
	Huber	0.049	0.040	0.031	0.033	0.045	0.262	0.210	0.176	0.175	0.226
	Tukey	0.049	0.040	0.031	0.033	0.046	0.106	0.088	0.069	0.069	0.096
MSE	L2	0.051	0.042	0.036	0.035	0.045	0.650	0.498	0.436	0.408	0.509
	Huber	0.051	0.042	0.036	0.035	0.045	0.265	0.216	0.185	0.176	0.226
	Tukey	0.051	0.043	0.036	0.035	0.046	0.109	0.092	0.076	0.069	0.096
EstVar	L2	0.052	0.051	0.035	0.034	0.051	0.678	0.538	0.439	0.454	0.552
	Huber	0.052	0.051	0.035	0.034	0.051	0.285	0.232	0.183	0.191	0.249
	Tukey	0.053	0.052	0.035	0.035	0.051	0.116	0.103	0.076	0.078	0.107
Rate	Huber	0.967	0.978	0.968	0.947	0.963	0.982	0.979	0.970	0.978	0.983
	Tukey	0.968	0.977	0.968	0.947	0.964	0.972	0.971	0.966	0.964	0.973
Outlier Ratio		0.2					0.3				
Bias	L2	0.056	-0.004	-0.076	0.045	0.071	-0.025	0.033	-0.117	0.002	-0.024
	Huber	0.049	0.008	-0.087	0.033	0.043	-0.018	0.041	-0.105	0.024	-0.007
	Tukey	0.051	0.031	-0.083	0.025	0.019	0.008	0.052	-0.085	0.046	0.015
Var	L2	1.212	0.909	0.820	0.790	1.072	1.789	1.376	1.186	1.204	1.567
	Huber	0.623	0.449	0.397	0.389	0.534	1.114	0.828	0.686	0.704	0.930
	Tukey	0.247	0.186	0.155	0.142	0.209	0.543	0.383	0.295	0.325	0.396
MSE	L2	1.215	0.909	0.826	0.792	1.077	1.790	1.377	1.200	1.204	1.568
	Huber	0.625	0.449	0.404	0.390	0.535	1.115	0.829	0.697	0.705	0.930
	Tukey	0.250	0.187	0.162	0.143	0.210	0.543	0.386	0.303	0.327	0.397
EstVar	L2	1.298	0.971	0.824	0.859	1.039	1.857	1.446	1.222	1.238	1.515
	Huber	0.678	0.491	0.411	0.432	0.528	1.156	0.880	0.730	0.746	0.929
	Tukey	0.262	0.202	0.161	0.166	0.217	0.535	0.412	0.330	0.336	0.424
Rate	Huber	0.983	0.969	0.973	0.974	0.967	0.970	0.975	0.970	0.969	0.971
	Tukey	0.981	0.962	0.966	0.971	0.967	0.966	0.974	0.983	0.970	0.964
Outlier Ratio		0.4					0.45				
Bias	L2	0.029	0.028	-0.166	0.029	0.035	-0.029	0.023	-0.027	-0.009	0.018
	Huber	0.064	0.034	-0.140	0.022	0.047	-0.005	0.038	-0.029	0.009	0.013
	Tukey	0.101	0.048	-0.099	0.022	0.048	0.024	0.045	-0.043	0.025	-0.005
Var	L2	2.656	1.820	1.790	1.519	1.738	2.721	2.033	1.720	1.765	2.384
	Huber	1.966	1.303	1.190	1.046	1.250	2.158	1.512	1.271	1.332	1.826
	Tukey	1.142	0.742	0.587	0.584	0.779	1.409	0.925	0.793	0.780	1.124
MSE	L2	2.657	1.820	1.817	1.520	1.739	2.721	2.034	1.721	1.765	2.385
	Huber	1.970	1.304	1.210	1.046	1.252	2.158	1.514	1.272	1.332	1.826
	Tukey	1.152	0.745	0.597	0.584	0.782	1.410	0.927	0.795	0.781	1.124
EstVar	L2	2.483	1.898	1.674	1.647	2.019	2.737	2.144	1.807	1.898	2.283
	Huber	1.902	1.398	1.177	1.174	1.499	2.260	1.690	1.385	1.470	1.800
	Tukey	1.328	0.803	0.643	0.663	0.894	1.569	1.087	0.866	0.900	1.175
Rate	Huber	0.963	0.969	0.961	0.962	0.980	0.968	0.962	0.971	0.970	0.962
	Tukey	0.963	0.969	0.966	0.976	0.976	0.970	0.974	0.966	0.966	0.961

S9. COMPARISON WITH DR-LEARNER VARIANTS

Table S2: Simulation results with $n = 500$ and Cauchy(0, 0.5).

x1		-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0					0.1				
Bias	L2	0.046	0.048	-0.075	0.041	0.019	0.038	0.032	-0.051	0.422	-0.055
	Huber	0.046	0.048	-0.075	0.041	0.019	0.049	0.060	-0.088	0.032	0.016
	Tukey	0.046	0.047	-0.074	0.041	0.018	0.047	0.058	-0.086	0.032	0.012
Var	L2	0.049	0.040	0.031	0.033	0.045	1.429	2.171	0.696	1.83e2	2.995
	Huber	0.049	0.040	0.031	0.033	0.045	0.074	0.062	0.046	0.047	0.062
	Tukey	0.049	0.040	0.031	0.033	0.046	0.059	0.053	0.037	0.038	0.053
MSE	L2	0.051	0.042	0.036	0.035	0.045	1.430	2.172	0.699	1.83e2	2.998
	Huber	0.051	0.042	0.036	0.035	0.045	0.076	0.065	0.053	0.048	0.062
	Tukey	0.051	0.043	0.036	0.035	0.046	0.061	0.057	0.045	0.039	0.053
EstVar	L2	0.052	0.051	0.035	0.034	0.051	1.725	1.801	0.706	1.28e2	5.013
	Huber	0.052	0.051	0.035	0.034	0.051	0.074	0.069	0.051	0.050	0.072
	Tukey	0.053	0.052	0.035	0.035	0.051	0.061	0.059	0.041	0.041	0.061
Rate	Huber	0.967	0.978	0.968	0.947	0.963	0.967	0.967	0.972	0.966	0.965
	Tukey	0.968	0.977	0.968	0.947	0.964	0.964	0.966	0.969	0.964	0.965
Outlier Ratio		0.2					0.3				
Bias	L2	1.575	0.062	1.732	0.078	-0.037	-0.073	-2.139	0.411	0.161	0.105
	Huber	0.056	0.047	-0.072	0.037	0.043	0.054	0.063	-0.092	0.039	0.033
	Tukey	0.048	0.046	-0.069	0.036	0.043	0.054	0.062	-0.090	0.038	0.029
Var	L2	1.14e3	1.772	2.63e3	1.77e1	6.276	8.669	2.41e3	1.14e2	1.57e1	3.94e1
	Huber	0.097	0.080	0.067	0.061	0.093	0.121	0.091	0.077	0.080	0.104
	Tukey	0.069	0.058	0.045	0.046	0.066	0.077	0.063	0.048	0.055	0.072
MSE	L2	1.14e3	1.776	2.63e3	1.77e1	6.277	8.674	2.42e3	1.14e2	1.57e1	3.94e1
	Huber	0.100	0.082	0.072	0.062	0.095	0.124	0.095	0.085	0.081	0.105
	Tukey	0.071	0.060	0.049	0.047	0.068	0.080	0.067	0.056	0.057	0.073
EstVar	L2	9.89e2	3.537	2.21e3	1.26e1	6.351	4.60e1	3.08e3	3.24e2	7.04e1	4.98e1
	Huber	0.096	0.092	0.071	0.066	0.096	0.122	0.105	0.088	0.083	0.114
	Tukey	0.069	0.069	0.049	0.048	0.070	0.082	0.075	0.057	0.056	0.078
Rate	Huber	0.976	0.974	0.968	0.963	0.976	0.973	0.979	0.982	0.964	0.976
	Tukey	0.970	0.967	0.962	0.957	0.973	0.973	0.975	0.971	0.955	0.967
Outlier Ratio		0.4					0.45				
Bias	L2	-0.143	0.048	-0.219	-0.739	-0.406	-0.581	-0.278	-0.059	0.196	-0.226
	Huber	0.057	0.047	-0.070	0.042	0.026	0.044	0.049	-0.074	0.037	0.011
	Tukey	0.058	0.051	-0.072	0.036	0.026	0.048	0.047	-0.077	0.035	0.005
Var	L2	1.39e1	1.46e2	2.75e1	5.78e2	8.67e2	2.02e2	7.05e2	4.36e1	4.89e1	6.35e1
	Huber	0.163	0.134	0.108	0.101	0.123	0.165	0.124	0.107	0.104	0.145
	Tukey	0.092	0.079	0.067	0.063	0.071	0.092	0.080	0.061	0.062	0.087
MSE	L2	1.39e1	1.46e2	2.75e1	5.79e2	8.67e2	2.02e2	7.05e2	4.36e1	4.89e1	6.36e1
	Huber	0.166	0.137	0.113	0.103	0.124	0.167	0.126	0.112	0.105	0.145
	Tukey	0.095	0.082	0.072	0.064	0.072	0.094	0.082	0.067	0.064	0.087
EstVar	L2	1.34e1	3.19e2	5.20e1	8.55e3	1.10e4	2.57e2	8.71e2	5.44e1	5.93e1	5.16e1
	Huber	0.159	0.135	0.103	0.105	0.135	0.177	0.141	0.114	0.114	0.155
	Tukey	0.096	0.087	0.064	0.064	0.088	0.103	0.092	0.069	0.070	0.096
Rate	Huber	0.974	0.968	0.963	0.959	0.972	0.975	0.977	0.972	0.976	0.970
	Tukey	0.962	0.969	0.956	0.954	0.969	0.963	0.970	0.966	0.969	0.966

Table S3: Simulation results with $n = 500$ and $\log \mathcal{N}(0, 1) - \exp(1/2)$.

x1		-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0					0.1				
Bias	L2	0.046	0.048	-0.075	0.041	0.019	0.044	0.067	-0.105	0.026	0.026
	Huber	0.046	0.048	-0.075	0.041	0.019	-0.087	-0.082	-0.224	-0.111	-0.107
	Tukey	0.046	0.047	-0.074	0.041	0.018	-0.118	-0.114	-0.252	-0.144	-0.146
Var	L2	0.049	0.040	0.031	0.033	0.045	0.618	0.524	0.361	0.436	0.404
	Huber	0.049	0.040	0.031	0.033	0.045	0.164	0.125	0.098	0.103	0.124
	Tukey	0.049	0.040	0.031	0.033	0.046	0.108	0.090	0.067	0.072	0.087
MSE	L2	0.051	0.042	0.036	0.035	0.045	0.620	0.528	0.372	0.437	0.404
	Huber	0.051	0.042	0.036	0.035	0.045	0.171	0.132	0.148	0.116	0.135
	Tukey	0.051	0.043	0.036	0.035	0.046	0.122	0.103	0.131	0.093	0.108
Est Var	L2	0.052	0.051	0.035	0.034	0.051	0.680	0.550	0.349	0.433	0.435
	Huber	0.052	0.051	0.035	0.034	0.051	0.159	0.135	0.105	0.108	0.138
	Tukey	0.053	0.052	0.035	0.035	0.051	0.113	0.099	0.075	0.075	0.101
Rate	Huber	0.967	0.978	0.968	0.947	0.963	0.958	0.973	0.972	0.963	0.985
	Tukey	0.968	0.977	0.968	0.947	0.964	0.939	0.957	0.964	0.951	0.975
Outlier Ratio		0.2					0.3				
Bias	L2	0.011	0.048	-0.041	0.013	-0.012	-0.061	0.115	-0.133	0.020	0.049
	Huber	-0.252	-0.220	-0.334	-0.228	-0.259	-0.376	-0.283	-0.486	-0.359	-0.350
	Tukey	-0.332	-0.307	-0.428	-0.311	-0.337	-0.508	-0.461	-0.627	-0.509	-0.515
Var	L2	1.151	0.899	0.970	0.601	0.804	1.335	1.318	1.071	1.051	1.387
	Huber	0.259	0.196	0.176	0.182	0.214	0.372	0.295	0.232	0.236	0.313
	Tukey	0.179	0.139	0.117	0.121	0.152	0.246	0.199	0.157	0.162	0.214
MSE	L2	1.151	0.901	0.972	0.601	0.804	1.338	1.332	1.089	1.051	1.389
	Huber	0.322	0.244	0.288	0.234	0.281	0.513	0.376	0.469	0.364	0.436
	Tukey	0.290	0.234	0.301	0.217	0.265	0.504	0.411	0.549	0.420	0.480
Est Var	L2	1.195	0.996	0.981	0.662	0.887	1.350	1.343	1.063	1.117	1.420
	Huber	0.275	0.223	0.188	0.184	0.233	0.392	0.328	0.267	0.270	0.339
	Tukey	0.194	0.158	0.127	0.127	0.167	0.289	0.231	0.193	0.191	0.241
Rate	Huber	0.968	0.973	0.967	0.968	0.977	0.971	0.969	0.975	0.966	0.967
	Tukey	0.941	0.952	0.949	0.953	0.963	0.952	0.946	0.963	0.955	0.947
Outlier Ratio		0.4					0.45				
Bias	L2	0.107	0.065	-0.084	0.003	0.015	0.044	0.123	-0.092	-0.019	0.049
	Huber	-0.422	-0.476	-0.586	-0.478	-0.485	-0.523	-0.493	-0.646	-0.529	-0.553
	Tukey	-0.669	-0.720	-0.822	-0.705	-0.731	-0.805	-0.799	-0.934	-0.796	-0.836
Var	L2	2.914	1.655	1.512	1.585	1.763	2.442	2.348	1.576	1.428	3.850
	Huber	0.580	0.376	0.343	0.300	0.466	0.568	0.427	0.357	0.372	0.572
	Tukey	0.360	0.259	0.234	0.213	0.299	0.404	0.287	0.241	0.258	0.380
MSE	L2	2.925	1.659	1.519	1.585	1.763	2.444	2.363	1.585	1.429	3.852
	Huber	0.758	0.602	0.686	0.529	0.702	0.842	0.670	0.775	0.652	0.878
	Tukey	0.807	0.777	0.909	0.710	0.834	1.052	0.925	1.113	0.891	1.078
Est Var	L2	2.744	1.872	1.596	1.621	1.760	2.724	2.356	1.624	1.558	3.490
	Huber	0.579	0.426	0.369	0.366	0.463	0.633	0.497	0.427	0.414	0.523
	Tukey	0.430	0.318	0.273	0.270	0.342	0.480	0.365	0.320	0.314	0.399
Rate	Huber	0.976	0.978	0.970	0.975	0.965	0.975	0.967	0.973	0.973	0.951
	Tukey	0.960	0.977	0.961	0.963	0.961	0.965	0.966	0.967	0.963	0.947

S9. COMPARISON WITH DR-LEARNER VARIANTS

Table S4: Simulation results with $n = 500$ and Dirac Delta at $\epsilon_{2i} = 20$.

x1		-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0					0.1				
Bias	L2	0.046	0.048	-0.075	0.041	0.019	2.053	2.135	1.948	2.064	2.053
	Huber	0.046	0.048	-0.075	0.041	0.019	0.943	0.973	0.797	0.918	0.908
	Tukey	0.046	0.047	-0.074	0.041	0.018	0.051	0.062	-0.082	0.031	0.014
Var	L2	0.049	0.040	0.031	0.033	0.045	2.496	1.749	1.400	1.469	1.819
	Huber	0.049	0.040	0.031	0.033	0.045	0.696	0.510	0.388	0.430	0.525
	Tukey	0.049	0.040	0.031	0.033	0.046	0.054	0.049	0.035	0.034	0.047
MSE	L2	0.051	0.042	0.036	0.035	0.045	6.710	6.305	5.194	5.728	6.033
	Huber	0.051	0.042	0.036	0.035	0.045	1.586	1.458	1.023	1.272	1.349
	Tukey	0.051	0.043	0.036	0.035	0.046	0.056	0.053	0.041	0.035	0.047
Est Var	L2	0.052	0.051	0.035	0.034	0.051	2.486	1.979	1.609	1.686	2.045
	Huber	0.052	0.051	0.035	0.034	0.051	0.550	0.443	0.346	0.362	0.452
	Tukey	0.053	0.052	0.035	0.035	0.051	0.057	0.058	0.039	0.038	0.058
Rate	Huber	0.967	0.978	0.968	0.947	0.963	0.799	0.877	0.869	0.874	0.865
	Tukey	0.968	0.977	0.968	0.947	0.964	0.965	0.972	0.974	0.961	0.973
Outlier Ratio		0.2					0.3				
Bias	L2	4.117	4.036	3.937	3.999	4.054	6.035	6.073	5.826	6.018	5.995
	Huber	2.576	2.438	2.349	2.422	2.471	5.245	5.213	4.962	5.137	5.142
	Tukey	0.050	0.056	-0.066	0.037	0.038	1.442	1.265	1.123	1.203	1.245
Var	L2	4.075	2.738	2.704	2.681	2.935	5.169	3.933	3.544	3.169	3.979
	Huber	3.039	1.796	1.888	1.899	2.056	6.813	5.354	5.081	4.550	5.626
	Tukey	0.061	0.055	0.041	0.042	0.060	8.242	5.096	5.033	4.452	5.338
MSE	L2	2.10e1	1.90e1	1.82e1	1.87e1	1.94e1	4.16e1	4.08e1	3.75e1	3.94e1	3.99e1
	Huber	9.675	7.740	7.406	7.764	8.161	3.43e1	3.25e1	2.97e1	3.09e1	3.21e1
	Tukey	0.063	0.058	0.045	0.043	0.061	1.03e1	6.697	6.293	5.899	6.888
Est Var	L2	4.949	3.712	3.183	3.289	3.969	7.251	5.603	4.709	4.921	5.842
	Huber	2.017	1.422	1.252	1.270	1.556	6.651	4.897	4.184	4.229	5.161
	Tukey	0.064	0.067	0.045	0.044	0.068	1.30e2	5.759	2.891	1.901	7.83e1
Rate	Huber	0.833	0.853	0.835	0.829	0.874	0.829	0.848	0.834	0.871	0.849
	Tukey	0.971	0.979	0.970	0.958	0.966	0.380	0.405	0.406	0.396	0.411
Outlier Ratio		0.4					0.45				
Bias	L2	8.047	8.110	7.904	8.091	8.100	9.060	9.118	8.870	8.997	9.046
	Huber	8.021	8.082	7.882	8.065	8.071	9.059	9.117	8.868	8.996	9.045
	Tukey	6.892	6.828	6.554	6.762	6.877	8.767	8.806	8.491	8.638	8.720
Var	L2	6.017	4.314	3.766	3.814	4.770	6.187	4.642	3.757	3.940	4.723
	Huber	6.210	4.496	3.893	3.984	4.976	6.196	4.644	3.772	3.946	4.731
	Tukey	1.37e1	1.11e1	9.704	9.741	1.15e1	9.471	7.449	6.317	6.477	7.535
MSE	L2	7.08e1	7.01e1	6.62e1	6.93e1	7.04e1	8.83e1	8.78e1	8.24e1	8.49e1	8.66e1
	Huber	7.05e1	6.98e1	6.60e1	6.90e1	7.01e1	8.83e1	8.78e1	8.24e1	8.49e1	8.65e1
	Tukey	6.12e1	5.77e1	5.27e1	5.55e1	5.88e1	8.63e1	8.50e1	7.84e1	8.11e1	8.36e1
Est Var	L2	9.938	7.378	6.340	6.571	7.907	1.09e1	8.343	7.122	7.409	8.841
	Huber	1.97e1	1.26e1	1.05e1	1.10e1	1.38e1	1.37e1	1.00e1	8.213	8.864	1.06e1
	Tukey	1.03e2	1.40e2	2.22e6	6.62e2	1.44e3	1.37e2	5.53e1	1.21e1	1.75e3	1.27e2
Rate	Huber	0.959	0.965	0.965	0.970	0.958	0.963	0.976	0.986	0.981	0.979
	Tukey	0.766	0.793	0.765	0.801	0.803	0.910	0.928	0.923	0.929	0.931

Table S5: Simulation results with $n = 5000$ and $\mathcal{N}(0, 10)$

		x1	-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0					0.1					
Bias	L2	0.024	0.016	-0.033	0.015	0.008	-0.005	0.031	-0.037	-0.008	0.021	
	Huber	0.024	0.016	-0.033	0.015	0.008	0.005	0.025	-0.034	-0.002	0.012	
	Tukey	0.024	0.016	-0.033	0.015	0.008	0.013	0.017	-0.033	0.007	0.005	
Var	L2	0.007	0.006	0.005	0.005	0.006	0.273	0.219	0.172	0.191	0.220	
	Huber	0.007	0.006	0.005	0.005	0.006	0.156	0.121	0.099	0.116	0.125	
	Tukey	0.007	0.006	0.005	0.005	0.006	0.065	0.049	0.042	0.050	0.053	
MSE	L2	0.008	0.006	0.006	0.006	0.006	0.273	0.220	0.173	0.191	0.221	
	Huber	0.008	0.006	0.006	0.006	0.006	0.156	0.122	0.101	0.116	0.125	
	Tukey	0.008	0.006	0.006	0.006	0.007	0.065	0.049	0.043	0.050	0.053	
EstVar	L2	0.008	0.006	0.005	0.005	0.007	0.288	0.214	0.184	0.191	0.231	
	Huber	0.008	0.006	0.005	0.005	0.007	0.162	0.122	0.105	0.109	0.132	
	Tukey	0.008	0.006	0.005	0.005	0.007	0.068	0.052	0.044	0.046	0.056	
Rate	Huber	0.954	0.962	0.952	0.943	0.956	0.960	0.964	0.955	0.962	0.956	
	Tukey	0.952	0.962	0.952	0.945	0.956	0.950	0.948	0.950	0.961	0.961	
Outlier Ratio		0.2					0.3					
Bias	L2	0.022	0.001	-0.042	0.031	0.000	-0.011	0.035	-0.061	0.038	0.010	
	Huber	0.025	0.010	-0.039	0.027	0.004	0.002	0.031	-0.055	0.038	0.012	
	Tukey	0.024	0.018	-0.034	0.020	0.011	0.018	0.024	-0.043	0.031	0.014	
Var	L2	0.386	0.283	0.258	0.275	0.287	0.392	0.313	0.270	0.283	0.338	
	Huber	0.265	0.191	0.168	0.177	0.194	0.284	0.225	0.192	0.208	0.244	
	Tukey	0.132	0.098	0.081	0.084	0.100	0.175	0.129	0.107	0.125	0.137	
MSE	L2	0.386	0.283	0.260	0.276	0.287	0.392	0.314	0.273	0.284	0.339	
	Huber	0.266	0.191	0.170	0.178	0.194	0.284	0.226	0.195	0.210	0.244	
	Tukey	0.132	0.098	0.082	0.085	0.100	0.175	0.130	0.109	0.126	0.137	
EstVar	L2	0.375	0.281	0.247	0.253	0.307	0.423	0.317	0.279	0.281	0.341	
	Huber	0.255	0.191	0.168	0.171	0.209	0.311	0.232	0.204	0.205	0.252	
	Tukey	0.131	0.098	0.086	0.088	0.108	0.178	0.134	0.117	0.118	0.146	
Rate	Huber	0.945	0.957	0.957	0.947	0.956	0.967	0.957	0.953	0.940	0.956	
	Tukey	0.958	0.954	0.961	0.955	0.962	0.957	0.957	0.950	0.940	0.959	
Outlier Ratio		0.4					0.45					
Bias	L2	0.015	0.030	-0.046	0.027	0.013	0.012	0.036	-0.050	0.034	0.020	
	Huber	0.018	0.032	-0.048	0.028	0.015	0.014	0.038	-0.054	0.036	0.022	
	Tukey	0.020	0.036	-0.051	0.030	0.018	0.016	0.040	-0.056	0.038	0.024	
Var	L2	0.482	0.389	0.341	0.365	0.432	0.511	0.432	0.374	0.392	0.478	
	Huber	0.352	0.285	0.252	0.271	0.328	0.390	0.325	0.281	0.294	0.361	
	Tukey	0.210	0.174	0.151	0.165	0.199	0.237	0.195	0.166	0.177	0.217	
MSE	L2	0.482	0.390	0.343	0.366	0.432	0.511	0.432	0.376	0.394	0.478	
	Huber	0.352	0.286	0.254	0.272	0.329	0.391	0.326	0.283	0.296	0.361	
	Tukey	0.210	0.174	0.152	0.166	0.200	0.237	0.196	0.168	0.179	0.218	
EstVar	L2	0.472	0.380	0.333	0.356	0.421	0.500	0.422	0.364	0.382	0.466	
	Huber	0.342	0.276	0.242	0.262	0.318	0.380	0.315	0.271	0.284	0.349	
	Tukey	0.200	0.164	0.141	0.155	0.189	0.227	0.185	0.156	0.167	0.207	
Rate	Huber	0.940	0.945	0.937	0.945	0.953	0.960	0.948	0.938	0.947	0.950	
	Tukey	0.935	0.940	0.932	0.942	0.951	0.955	0.943	0.933	0.942	0.946	

S9. COMPARISON WITH DR-LEARNER VARIANTS

Table S6: Simulation results with $n = 5000$ and Cauchy(0,0.5).

		x1	-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0						0.1				
Bias	L2	0.024	0.016	-0.033	0.015	0.008		-0.234	0.965	0.053	-0.007	-0.096
	Huber	0.024	0.016	-0.033	0.015	0.008		0.024	0.025	-0.024	0.015	0.007
	Tukey	0.024	0.016	-0.033	0.015	0.008		0.022	0.022	-0.025	0.015	0.005
Var	L2	0.007	0.006	0.005	0.005	0.006		6.49e1	9.03e2	8.618	1.171	2.888
	Huber	0.007	0.006	0.005	0.005	0.006		0.011	0.009	0.008	0.007	0.009
	Tukey	0.007	0.006	0.005	0.005	0.006		0.009	0.007	0.006	0.006	0.007
MSE	L2	0.008	0.006	0.006	0.006	0.006		6.50e1	9.03e2	8.621	1.171	2.90e0
	Huber	0.008	0.006	0.006	0.006	0.006		0.012	0.009	0.008	0.007	0.009
	Tukey	0.008	0.006	0.006	0.006	0.007		0.010	0.007	0.006	0.006	0.007
EstVar	L2	0.008	0.006	0.005	0.005	0.007		6.59e1	2.65e3	9.27e0	1.80e0	5.36e0
	Huber	0.008	0.006	0.005	0.005	0.007		0.011	0.009	0.007	0.007	0.010
	Tukey	0.008	0.006	0.005	0.005	0.007		0.009	0.008	0.006	0.006	0.008
Rate95	Huber	0.954	0.962	0.952	0.943	0.956		0.957	0.967	0.953	0.959	0.973
	Tukey	0.952	0.962	0.952	0.945	0.956		0.952	0.967	0.952	0.960	0.971
Outlier Ratio		0.2						0.3				
Bias	L2	0.040	0.117	-0.131	0.072	0.217		0.048	0.076	-0.056	0.095	-0.022
	Huber	0.021	0.013	-0.031	0.014	0.012		0.016	0.025	-0.031	0.012	0.006
	Tukey	0.021	0.015	-0.029	0.015	0.010		0.014	0.022	-0.031	0.014	0.005
Var	L2	1.57e0	5.64e0	1.10e1	5.81e0	2.43e1		4.03e1	1.59e1	2.60e0	7.23e0	5.04e1
	Huber	0.015	0.011	0.010	0.010	0.012		0.019	0.014	0.012	0.012	0.015
	Tukey	0.011	0.008	0.007	0.007	0.009		0.012	0.009	0.008	0.008	0.010
MSE	L2	1.57e0	5.66e0	1.10e1	5.82e0	2.44e1		4.03e1	1.59e1	2.60e0	7.24e0	5.04e1
	Huber	0.016	0.011	0.011	0.010	0.013		0.019	0.015	0.013	0.012	0.015
	Tukey	0.011	0.008	0.008	0.007	0.009		0.012	0.010	0.009	0.008	0.010
EstVar	L2	2.71e0	3.13e2	1.28e1	6.16e0	1.95e1		6.37e1	2.89e1	5.28e0	1.18e1	5.55e1
	Huber	0.015	0.012	0.010	0.010	0.013		0.018	0.015	0.012	0.013	0.016
	Tukey	0.010	0.009	0.007	0.007	0.009		0.012	0.010	0.008	0.008	0.011
Rate95	Huber	0.947	0.973	0.952	0.954	0.961		0.958	0.957	0.947	0.960	0.953
	Tukey	0.946	0.963	0.951	0.950	0.952		0.949	0.957	0.955	0.955	0.955
Outlier Ratio		0.4						0.45				
Bias	L2	0.211	0.074	0.180	1.57e0	-0.042		0.250	0.182	0.258	-0.619	0.039
	Huber	0.015	0.029	-0.034	0.020	0.005		0.019	0.023	-0.036	0.013	0.008
	Tukey	0.015	0.024	-0.032	0.020	0.008		0.018	0.024	-0.034	0.015	0.009
Var	L2	3.33e1	1.20e1	5.95e1	2.71e3	7.86e1		3.80e1	1.21e1	7.06e1	2.60e3	8.04e1
	Huber	0.023	0.017	0.014	0.016	0.019		0.025	0.019	0.016	0.016	0.020
	Tukey	0.013	0.010	0.008	0.010	0.012		0.015	0.011	0.009	0.010	0.012
MSE	L2	3.34e1	1.20e1	5.95e1	2.71e3	7.86e1		3.80e1	1.21e1	7.06e1	2.60e3	8.04e1
	Huber	0.024	0.017	0.015	0.016	0.019		0.026	0.019	0.017	0.017	0.021
	Tukey	0.013	0.011	0.009	0.010	0.012		0.015	0.012	0.010	0.010	0.013
EstVar	L2	3.82e1	3.35e1	5.11e1	2.61e3	1.12e2		4.01e1	3.50e1	5.52e1	2.61e3	1.20e2
	Huber	0.023	0.018	0.015	0.015	0.019		0.025	0.019	0.017	0.017	0.021
	Tukey	0.014	0.011	0.009	0.010	0.012		0.015	0.012	0.010	0.010	0.013
Rate95	Huber	0.953	0.957	0.958	0.951	0.957		0.974	0.969	0.956	0.960	0.960
	Tukey	0.959	0.959	0.964	0.952	0.952		0.964	0.973	0.959	0.952	0.960

Table S7: Simulation results with $n = 5000$ and $\log \mathcal{N}(0, 1) - \exp(1/2)$.

		x1	-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0						0.1				
Bias	L2	0.024	0.016	-0.033	0.015	0.008		0.015	0.012	-0.027	0.014	0.010
	Huber	0.024	0.016	-0.033	0.015	0.008		-0.121	-0.121	-0.166	-0.125	-0.131
	Tukey	0.024	0.016	-0.033	0.015	0.008		-0.149	-0.147	-0.193	-0.152	-0.160
Var	L2	0.007	0.006	0.005	0.005	0.006		0.087	0.073	0.068	0.066	0.073
	Huber	0.007	0.006	0.005	0.005	0.006		0.022	0.017	0.015	0.016	0.017
	Tukey	0.007	0.006	0.005	0.005	0.006		0.016	0.012	0.011	0.010	0.012
MSE	L2	0.008	0.006	0.006	0.006	0.006		0.088	0.073	0.069	0.066	0.073
	Huber	0.008	0.006	0.006	0.006	0.006		0.037	0.032	0.043	0.031	0.034
	Tukey	0.008	0.006	0.006	0.006	0.007		0.038	0.033	0.048	0.034	0.037
EstVar	L2	0.008	0.006	0.005	0.005	0.007		0.091	0.069	0.068	0.068	0.076
	Huber	0.008	0.006	0.005	0.005	0.007		0.024	0.018	0.016	0.016	0.020
	Tukey	0.008	0.006	0.005	0.005	0.007		0.016	0.013	0.011	0.011	0.014
Rate	Huber	0.954	0.962	0.952	0.943	0.956		0.958	0.958	0.950	0.950	0.966
	Tukey	0.952	0.962	0.952	0.945	0.956		0.950	0.962	0.950	0.961	0.961
Outlier Ratio		0.2						0.3				
Bias	L2	0.040	0.117	-0.131	0.072	0.217		0.048	0.076	-0.056	0.095	-0.022
	Huber	0.021	0.013	-0.031	0.014	0.012		0.016	0.025	-0.031	0.012	0.006
	Tukey	0.021	0.015	-0.029	0.015	0.010		0.014	0.022	-0.031	0.014	0.005
Var	L2	1.57	5.64	1.10e1	5.81	2.43e1		4.03e1	15.9	2.60	7.23	50.4
	Huber	0.015	0.011	0.010	0.010	0.012		0.019	0.014	0.012	0.012	0.015
	Tukey	0.011	0.008	0.007	0.007	0.009		0.012	0.009	0.008	0.008	0.010
MSE	L2	1.57	5.66	1.10e1	5.82	2.44e1		4.03e1	15.9	2.60	7.24	50.4
	Huber	0.016	0.011	0.011	0.010	0.013		0.019	0.015	0.013	0.012	0.015
	Tukey	0.011	0.008	0.008	0.007	0.009		0.012	0.010	0.009	0.008	0.010
EstVar	L2	2.71	3.13e2	12.8	6.16	19.5		6.37e1	28.9	5.28	11.8	55.5
	Huber	0.015	0.012	0.010	0.010	0.013		0.018	0.015	0.012	0.013	0.016
	Tukey	0.010	0.009	0.007	0.007	0.009		0.012	0.010	0.008	0.008	0.011
Rate	Huber	0.947	0.973	0.952	0.954	0.961		0.958	0.957	0.947	0.960	0.953
	Tukey	0.946	0.963	0.951	0.950	0.952		0.949	0.957	0.955	0.955	0.955
Outlier Ratio		0.4						0.45				
Bias	L2	0.211	0.074	0.180	1.57	-0.042		0.250	0.182	0.258	-0.619	0.039
	Huber	0.015	0.029	-0.034	0.020	0.005		0.019	0.023	-0.036	0.013	0.008
	Tukey	0.015	0.024	-0.032	0.020	0.008		0.018	0.024	-0.034	0.015	0.009
Var	L2	3.33e1	12.0	59.5	2.71e3	7.86e1		38.0	12.1	70.6	2.60e3	8.04e1
	Huber	0.023	0.017	0.014	0.016	0.019		0.025	0.019	0.016	0.016	0.020
	Tukey	0.013	0.010	0.008	0.010	0.012		0.015	0.011	0.009	0.010	0.012
MSE	L2	3.34e1	12.0	59.5	2.71e3	7.86e1		38.0	12.1	70.6	2.60e3	8.04e1
	Huber	0.024	0.017	0.015	0.016	0.019		0.026	0.019	0.017	0.017	0.021
	Tukey	0.013	0.011	0.009	0.010	0.012		0.015	0.012	0.010	0.010	0.013
EstVar	L2	3.82e1	33.5	51.1	2.61e3	1.12e2		40.1	35.0	55.2	2.61e3	1.20e2
	Huber	0.023	0.018	0.015	0.015	0.019		0.025	0.019	0.017	0.017	0.021
	Tukey	0.014	0.011	0.009	0.010	0.012		0.015	0.012	0.010	0.010	0.013
Rate	Huber	0.953	0.957	0.958	0.951	0.957		0.974	0.969	0.956	0.960	0.960
	Tukey	0.959	0.959	0.964	0.952	0.952		0.964	0.973	0.959	0.952	0.960

S9. COMPARISON WITH DR-LEARNER VARIANTS

Table S8: Simulation results with $n = 5000$ and Dirac Delta at $\epsilon_{2i} = 20$.

		x1	-0.4	-0.2	0	0.2	0.4	-0.4	-0.2	0	0.2	0.4
Outlier Ratio		0						0.1				
Bias	L2	0.024	0.016	-0.033	0.015	0.008		2.02	2.01	1.98	2.02	2.02
	Huber	0.024	0.016	-0.033	0.015	0.008		0.856	0.853	0.809	0.854	0.847
	Tukey	0.024	0.016	-0.033	0.015	0.008		0.019	0.020	-0.028	0.014	0.005
Var	L2	0.007	0.006	0.005	0.005	0.006		0.31	0.26	0.21	0.20	0.28
	Huber	0.007	0.006	0.005	0.005	0.006		0.076	0.062	0.053	0.051	0.070
	Tukey	0.007	0.006	0.005	0.005	0.006		0.009	0.006	0.006	0.005	0.006
MSE	L2	0.008	0.006	0.006	0.006	0.006		4.39	4.30	4.12	4.30	4.34
	Huber	0.008	0.006	0.006	0.006	0.006		0.809	0.790	0.707	0.781	0.788
	Tukey	0.008	0.006	0.006	0.006	0.007		0.009	0.007	0.006	0.005	0.006
Est Var	L2	0.008	0.006	0.005	0.005	0.007		0.38	0.28	0.25	0.26	0.31
	Huber	0.008	0.006	0.005	0.005	0.007		0.073	0.055	0.048	0.050	0.060
	Tukey	0.008	0.006	0.005	0.005	0.007		0.008	0.007	0.006	0.006	0.008
Rate	Huber	0.954	0.962	0.952	0.943	0.956		0.940	0.915	0.927	0.945	0.911
	Tukey	0.952	0.962	0.952	0.945	0.956		0.950	0.964	0.956	0.960	0.974
Outlier Ratio		0.2						0.3				
Bias	L2	4.03	4.00	3.99	3.99	3.99		6.00	6.03	5.96	6.05	5.98
	Huber	2.32	2.29	2.27	2.29	2.28		5.01	5.02	4.94	5.05	4.97
	Tukey	0.022	0.017	-0.027	0.015	0.006		0.47	0.48	0.42	0.48	0.47
Var	L2	0.60	0.46	0.38	0.42	0.49		0.83	0.54	0.49	0.52	0.65
	Huber	0.33	0.26	0.22	0.24	0.28		1.25	0.84	0.79	0.88	1.03
	Tukey	0.009	0.007	0.006	0.006	0.008		0.14	0.12	0.11	0.13	0.13
MSE	L2	1.69e1	1.64e1	1.63e1	1.64e1	1.64e1		3.69e1	3.69e1	3.60e1	3.72e1	3.64e1
	Huber	5.72	5.50	5.35	5.46	5.49		2.63e1	2.60e1	2.52e1	2.64e1	2.58e1
	Tukey	0.010	0.007	0.007	0.006	0.008		0.36	0.34	0.28	0.36	0.35
Est Var	L2	0.75	0.56	0.49	0.50	0.60		1.11	0.84	0.73	0.76	0.90
	Huber	0.25	0.19	0.17	0.17	0.20		0.79	0.59	0.51	0.53	0.63
	Tukey	0.009	0.008	0.006	0.006	0.009		0.030	0.024	0.020	0.020	0.026
Rate	Huber	0.908	0.888	0.913	0.896	0.892		0.869	0.896	0.880	0.870	0.863
	Tukey	0.952	0.965	0.956	0.960	0.949		0.688	0.642	0.587	0.557	0.597
Outlier Ratio		0.4						0.45				
Bias	L2	8.011	8.002	7.949	8.021	8.052		9.031	9.030	8.980	9.009	9.023
	Huber	8.011	8.002	7.949	8.021	8.052		9.031	9.030	8.980	9.009	9.023
	Tukey	6.648	6.612	6.550	6.633	6.698		8.776	8.768	8.720	8.743	8.765
Var	L2	0.904	0.662	0.615	0.574	0.731		0.930	0.668	0.586	0.652	0.742
	Huber	0.904	0.662	0.615	0.574	0.731		0.930	0.668	0.586	0.652	0.742
	Tukey	2.368	1.770	1.705	1.623	1.982		1.467	1.066	0.940	1.045	1.182
MSE	L2	6.51e1	6.47e1	6.38e1	6.49e1	6.56e1		8.25e1	8.22e1	8.12e1	8.18e1	8.22e1
	Huber	6.51e1	6.47e1	6.38e1	6.49e1	6.56e1		8.25e1	8.22e1	8.12e1	8.18e1	8.22e1
	Tukey	4.66e1	4.55e1	4.46e1	4.56e1	4.68e1		7.85e1	7.80e1	7.70e1	7.75e1	7.80e1
Est Var	L2	1.481	1.111	0.969	0.999	1.208		1.672	1.253	1.093	1.125	1.352
	Huber	2.980	2.192	1.904	1.987	2.408		1.673	1.255	1.094	1.126	1.354
	Tukey	2.005	1.293	1.105	1.147	1.468		2.337	1.727	1.497	1.539	1.869
Rate95	Huber	0.996	0.997	0.995	0.996	0.993		0.991	0.991	0.993	0.990	0.988
	Tukey	0.875	0.859	0.862	0.872	0.882		0.970	0.967	0.976	0.972	0.969

within each treatment arm, constructs the AIPW pseudo-outcome

$$\varphi_i = \frac{T_i(Y_i - \hat{m}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1 - T_i)(Y_i - \hat{m}_0(X_i))}{1 - \hat{\pi}(X_i)} + \hat{m}_1(X_i) - \hat{m}_0(X_i),$$

and applies Nadaraya–Watson (NW) kernel regression of φ_i on X_{1i} , with the plug-in bandwidth of Ruppert et al. (1995), to obtain $\hat{\tau}^{\text{DR}}(x_1)$. The **Robust DR-learner** replaces the NW smoother with Huber IRLS kernel regression of φ_i on X_{1i} , and the **Fully Robust DR-learner** additionally replaces the OLS outcome regressions with robust linear regression via Huber M-estimation. Pointwise confidence intervals are constructed from sandwich variance estimators appropriate to each smoother; the complete step-by-step implementation, including the variance formulas, is provided in Section S11.

The results for all eight methods are summarised in Figure S1. The key findings are as follows.

Symmetric contamination (Normal and Cauchy). Under Normal contamination, the DR-learner achieves competitive MISE relative to AIPW but undercoverage (88–92%). The Robust DR-learner improves MISE substantially—approaching the proposed Tukey estimator and surpassing it at contamination ratios ≥ 0.4 —but exhibits severe undercoverage at low-to-moderate contamination levels (71% at ratio = 0.1, recovering to 92% at ratio = 0.45). The Fully Robust DR-learner provides marginal MISE gains over the Ro-

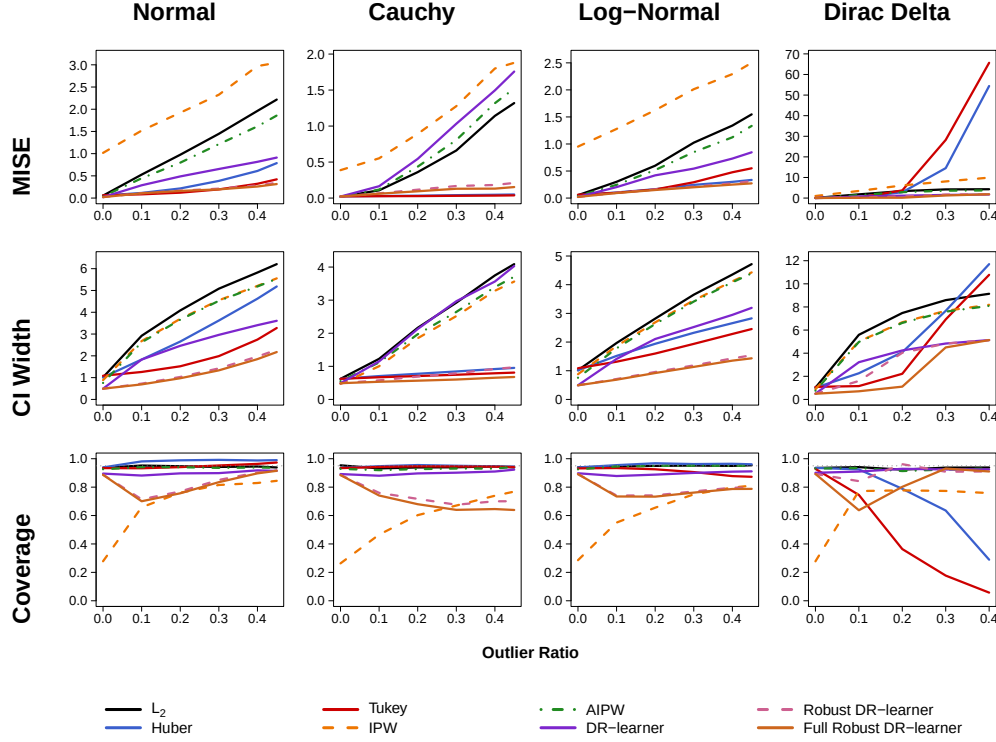


Figure S1: Comparison of MISE, confidence interval width, and coverage probability for all eight estimators across four outlier distributions and six contamination levels ($n = 500, 1000$ repetitions). Proposed methods (L_2 , Huber, Tukey) use LOO-CV bandwidth; IPW, AIPW, and DR-learner variants use the plug-in bandwidth of Ruppert et al. (1995).

bust DR-learner but shows essentially the same undercoverage pattern.

The undercoverage of the Robust DR-learner under Normal contamination is driven by *bias*, not variance: the outcome regression—even a robust one—is pulled toward the Normal outliers in Y , producing a systematically shifted pseudo-outcome φ that the downstream Huber smoother cannot correct (a bias decomposition quantifying this shift is given in Section S11). Coverage improves at higher contamination ratios not because the bias shrinks—it remains approximately constant—but because the growing variance of $\hat{\tau}^{\text{RobustDR}}$ widens the confidence intervals enough to cover the bias.

Under Cauchy contamination, the Robust and Fully Robust DR-learners achieve the lowest MISE among all methods except the proposed Huber and Tukey estimators, but their coverage collapses to 64–76%, *worse* than the plain DR-learner. The mechanism is the propensity weighting: a Cauchy-distributed Y_i with extreme value y_0 generates a pseudo-outcome $\varphi_i \approx y_0/\hat{\pi}(X_i)$, even more extreme than Y_i itself due to the $1/\hat{\pi}$ factor, so the Huber smoother acts on already-amplified heavy-tailed values. The M-estimator sandwich variance formula, derived under finite-moment assumptions, underestimates the true variability of $\hat{\tau}^{\text{RobustDR}}$ in this regime, leading to confidence intervals that are too narrow.

In contrast, the proposed Huber and Tukey estimators apply robustness

directly to the potential outcome Y_i in the original IPW-weighted regression, before any propensity amplification occurs. This upstream robustification maintains near-nominal coverage (93–96%) across all contamination levels and keeps MISE nearly flat under Cauchy contamination, where, across contamination ratios 0.1–0.45, the proposed estimators achieve MISE of 0.026–0.050 compared to 0.060–0.210 for the robustified DR variants and up to 1.76 for the plain DR-learner.

Asymmetric contamination (Log-Normal and Dirac Delta). Under log-normal contamination, the two robustified DR variants undercover (73–80%), for the same reason as under Cauchy: the asymmetric log-normal outliers propagate through the propensity weighting into φ , and the downstream Huber smoother cannot fully recover; the plain DR-learner attains 88–91% coverage. The proposed Huber estimator, operating directly on Y , achieves 95–97% coverage under contamination with substantially lower MISE.

Under Dirac Delta contamination (a point mass at 20), the Fully Robust DR-learner exhibits a striking pattern. At low contamination (ratio = 0.1–0.2), it achieves the *lowest MISE of all methods* (0.159–0.168), because the robust outcome regression successfully rejects the discrete spike at 20, producing a nearly uncontaminated pseudo-outcome φ on which the

Huber smoother operates cleanly. However, at higher contamination ratios (0.3–0.4), even the robust outcome regression is overwhelmed by the growing proportion of observations concentrated at 20, and MISE grows to 1.3–1.8 with coverage declining to 91%. For comparison, the proposed L_2 estimator maintains 92–94% coverage across all Dirac Delta contamination levels by consistently estimating the contaminated conditional mean, whose variability is correctly captured by the variance estimator.

Overall assessment. The DR-learner variants demonstrate that naively inserting robustness into the smoothing step of the DR pipeline—or even into both the outcome-regression and smoothing steps—is insufficient to achieve reliable inference under outlier contamination. The fundamental limitation is structural: the AIPW pseudo-outcome φ channels contamination from two sources—biased outcome regression and propensity-weight amplification—neither of which is remedied by downstream M-estimation alone. The proposed estimators circumvent both pathways by applying robust local regression directly to Y within each treatment arm, achieving proper coverage and low MISE simultaneously under symmetric contamination. This comparison reinforces the paper’s central contribution: robustness must be applied at the level of the potential outcome rather than at the level of the derived pseudo-outcome.

S10 Comparison with Machine-Learning Meta-Learners

We further compare the proposed estimators with three state-of-the-art machine-learning CATE learners: the causal forest (Athey et al., 2019), the X-learner (Künzel et al., 2019), and the R-learner (Nie and Wager, 2021). The causal forest is fitted with the `causal_forest` function of the `grf` package. Since `grf` does not provide the X-learner or the R-learner, we construct these two learners ourselves, following Künzel et al. (2019) and Nie and Wager (2021), with `grf` regression forests as the base learner. All forests use 500 trees. The X- and R-learners use the same logistic-regression propensity estimate truncated to $[0.01, 0.99]$ as the methods in Section S9, whereas the causal forest uses its default internal regression-forest estimate of the propensity score; the full constructions and implementation details are provided in Section S11.

These learners estimate $\tau(x)$ on the full covariate vector, whereas our estimand is the partially conditional CATE $\tau(x_1) = E\{\tau(X) \mid X_1 = x_1\}$. We therefore marginalize each learner’s individual-level predictions by Nadaraya–Watson kernel smoothing of $\hat{\tau}(X_i)$ on X_{1i} , with the plug-in bandwidth of Ruppert et al. (1995), and evaluate at the same five grid points as in Section 5.2 of the main paper. We report MISE only for these learners: to the best of our knowledge, no valid pointwise inference the-

ory is available for the marginalized CATE obtained from these pipelines, which itself underscores that inference—the focus of this paper—is the difficult part of the problem. Figure S2 displays the MISE (on a log scale) of the three meta-learners alongside the proposed L_2 , Huber, and Tukey estimators and the DR-learner of Section S9, across the four contamination types and six contamination ratios ($n = 500, 1000$ repetitions). The main findings are as follows.

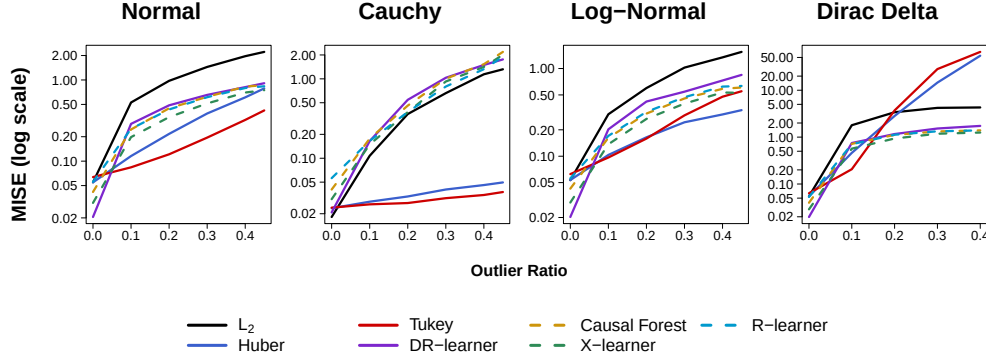


Figure S2: MISE (log scale) of the proposed estimators (L_2 , Huber, Tukey), the DR-learner, and three machine-learning meta-learners (causal forest, X-learner, R-learner) across four outlier distributions and increasing contamination levels ($n = 500, 1000$ repetitions). Meta-learner predictions are marginalized to the partially conditional CATE $\tau(x_1)$ by kernel smoothing on X_1 . For the Dirac delta setting, results are shown up to contamination ratio 0.4, as in Figure S1.

First, in the absence of contamination all three meta-learners are competitive: their MISE ranges over 0.029–0.056 across the four outcome dis-

tributions, comparable to the proposed estimators (0.018–0.064), with the X-learner consistently the best of the three.

Second, under contamination with finite variance the meta-learners degrade gracefully. Under Normal contamination their MISE at ratio 0.45 (0.75–0.86) is similar to the Huber estimator (0.79), well below L_2 (2.22) but roughly twice the Tukey estimator (0.42). Under log-normal contamination they fall between Huber (0.34) and L_2 (1.55). This moderate resistance is expected: the contaminated errors have finite variance, so the within-leaf averaging and honest subsampling of the forests attenuate—but do not bound—the influence of outliers.

Third, under heavy-tailed Cauchy contamination all three meta-learners break down: their MISE grows to 1.90–2.19 at ratio 0.45, exceeding even the non-robust L_2 estimator (1.32), while the proposed Huber (0.050) and Tukey (0.038) estimators remain essentially flat—a roughly forty- to sixty-fold difference. The mechanism parallels that of Section S9: regression forests rest on squared-error splitting and within-leaf averaging, and averages of Cauchy-contaminated outcomes do not concentrate, so the forest predictions inherit the heavy tails at every stage of the pipeline. No component of these learners bounds the influence of an individual outcome.

Fourth, under Dirac delta contamination the pattern reverses, consis-

tent with Section 5.2 of the main paper: the point mass at 20 enters both treatment arms additively, so it largely cancels in the treatment-effect contrast that the causal forest and R-learner target directly and that the X-learner forms by differencing fitted arms. The meta-learners consequently degrade gracefully (X-learner MISE 1.25 at ratio 0.4, the best of all methods), whereas the Huber and Tukey estimators, which downweight the systematic mass at 20 within each arm separately, break down at high contamination (54 and 66, respectively), and L_2 reaches 4.3.

In summary, the forest-based meta-learners offer moderate, incidental robustness under bounded or finite-variance contamination but no protection against heavy tails, and none of them provides valid pointwise confidence intervals for the partially conditional CATE. The proposed estimators are the only methods in the comparison that combine low MISE under symmetric contamination with theoretically justified inference.

S11 Implementation Details for the Comparison Studies in Sections S9 and S10

This section provides the full implementation details of the DR-learner variants (Section S9) and the machine-learning meta-learners (Section S10),

together with the bias decomposition referenced in Section S9.

S11.1 DR-Learner Variants

DR-learner. Following Kennedy (2023), the estimator proceeds in four steps. (1) Estimate the propensity score $\hat{\pi}(X)$ via logistic regression, truncated to $[0.01, 0.99]$. (2) Estimate the conditional mean outcomes $\hat{m}_1(X)$ and $\hat{m}_0(X)$ separately for treated and control units via ordinary least squares (OLS). (3) Construct the AIPW pseudo-outcome

$$\varphi_i = \frac{T_i(Y_i - \hat{m}_1(X_i))}{\hat{\pi}(X_i)} - \frac{(1 - T_i)(Y_i - \hat{m}_0(X_i))}{1 - \hat{\pi}(X_i)} + \hat{m}_1(X_i) - \hat{m}_0(X_i).$$

(4) Apply Nadaraya–Watson (NW) kernel regression of φ_i on X_{1i} , with bandwidth selected via the plug-in method of Ruppert et al. (1995), to obtain $\hat{\tau}^{\text{DR}}(x_1)$. Pointwise confidence intervals are constructed from a sandwich variance estimator for the NW smoother.

Robust DR-learner. Steps (1)–(3) are identical to the DR-learner. Step (4) replaces NW with Huber IRLS kernel regression of φ_i on X_{1i} . The variance is estimated using the M-estimator sandwich formula $\hat{A}(x_1)/\hat{B}(x_1)^2 \cdot R(K)/(nh\hat{f}(x_1))$, where $\hat{A}(x_1) = \mathbb{E}_n[\psi(\varphi - \hat{\tau})^2 \mid X_1 \approx x_1]$ and $\hat{B}(x_1) = \mathbb{E}_n[\psi'(\varphi - \hat{\tau}) \mid X_1 \approx x_1]$ are kernel-weighted averages of the Huber influence function and its derivative.

Fully Robust DR-learner. Step (2) replaces OLS with robust linear

regression via Huber M-estimation (`r1m`), producing robust estimates $\hat{m}_1^{\text{rob}}$ and $\hat{m}_0^{\text{rob}}$. Steps (1), (3), and (4) use the Huber IRLS smoother as in the Robust DR-learner.

S11.2 Machine-Learning Meta-Learners

Only the causal forest is available as a ready-made function in the `grf` package (Athey et al., 2019); `grf` does not implement the X-learner or the R-learner. Both of these are meta-algorithms that can be combined with any regression method as the base learner (Künzel et al., 2019; Nie and Wager, 2021), and we implement them ourselves, as described below, using the `regression_forest` function of `grf` as the base learner. All forests use 500 trees and the default `grf` tuning parameters.

Causal forest. The causal forest (`grf::causal_forest`) is applied to (X, Y, T) directly, and individual-level effect estimates $\hat{\tau}(X_i)$ are obtained as out-of-bag predictions. We use its default nuisance estimates: $E(Y \mid X)$ and the propensity score are estimated internally by regression forests, rather than by the logistic-regression propensity estimate used by the other methods. The causal forest centers Y and T at these estimates before growing the trees, which is the same residual-on-residual orthogonalization that underlies the R-learner.

X-learner. The X-learner (Künzel et al., 2019) fits regression forests $\hat{m}_1$ and $\hat{m}_0$ within the treated and control arms, forms the imputed individual effects $Y_i - \hat{m}_0(X_i)$ for treated units and $\hat{m}_1(X_i) - Y_i$ for control units, fits regression forests $\hat{\tau}_1$ and $\hat{\tau}_0$ to these imputed effects, and combines them as $\hat{\tau}(x) = \hat{\pi}(x)\hat{\tau}_0(x) + \{1 - \hat{\pi}(x)\}\hat{\tau}_1(x)$, where $\hat{\pi}$ is the truncated logistic-regression propensity estimate. All four regression forests are fitted with `grf::regression_forest`.

R-learner. The R-learner (Nie and Wager, 2021) regresses the pseudo-outcome $\{Y_i - \hat{m}(X_i)\} / \{T_i - \hat{\pi}(X_i)\}$ on X_i with sample weights $\{T_i - \hat{\pi}(X_i)\}^2$ via a regression forest, where $\hat{m}$ is a regression forest estimate of $E(Y \mid X)$ (evaluated out-of-bag) and $\hat{\pi}$ is the truncated logistic-regression propensity estimate. This is the R-learner with squared-error loss, because its objective satisfies $\sum_i [Y_i - \hat{m}(X_i) - \{T_i - \hat{\pi}(X_i)\}\tau(X_i)]^2 = \sum_i \{T_i - \hat{\pi}(X_i)\}^2 [Y_i - \hat{m}(X_i) / \{T_i - \hat{\pi}(X_i)\} - \tau(X_i)]^2$; the truncation of $\hat{\pi}$ keeps $|T_i - \hat{\pi}(X_i)| \geq 0.01$. Both forests are fitted with `grf::regression_forest`, the second with the `sample.weights` argument.

Marginalization to the partially conditional CATE. Each learner produces individual-level predictions $\hat{\tau}(X_i)$ on the full covariate vector, whereas the estimand of the main paper is $\tau(x_1) = E\{\tau(X) \mid X_1 = x_1\}$. The predictions are therefore marginalized by Nadaraya–Watson kernel re-

gression of $\hat{\tau}(X_i)$ on X_{1i} , with the plug-in bandwidth of Ruppert et al. (1995) (floored at 0.01), and evaluated at the grid $x_1 \in \{-0.4, -0.2, 0, 0.2, 0.4\}$.

S11.3 Bias Decomposition for the Robust DR-Learner under Normal Contamination

A bias decomposition of the Robust DR-learner under Normal contamination reveals that the Huber smoother applied to φ introduces a persistent negative bias of approximately -0.29 to -0.36 at $x_1 = 0$ across all contaminated settings (contamination ratios 0.1–0.45). This bias originates in the outcome-regression step: even with robust `rlm`, the global parametric outcome regression is pulled toward the Normal outliers in Y , producing a systematically biased pseudo-outcome φ . The Huber smoother in the final step cannot correct a mean shift already present in φ . Coverage improves at higher contamination ratios not because the bias shrinks—it remains approximately constant—but because the variance of $\hat{\tau}^{\text{RobustDR}}$ grows as more outliers inflate φ , widening the confidence interval enough to cover the bias.

S12 Additional Remarks

Remark 1 (Illustration of the breakdown point). To illustrate the breakdown point of Theorem 2 of the main paper concretely, consider a Gaussian

kernel with bandwidth h and suppose propensity scores are approximately equal ($\pi(X_i) \approx \pi_0$ for all i). Then $W_i(x_1) \propto K((X_{i1} - x_1)/h)$, and observations far from x_1 contribute negligible weight. In the worst case, an adversary contaminates the m observations with the highest kernel weights (i.e., those closest to x_1), achieving $\sum_{S_{Y_c}} W_i \approx m \cdot K(0)/h$. The estimator breaks down under $(U\rho)$ when this exceeds $\sum_{S_{Y_n}} W_i$, roughly when m exceeds half the effective local sample size $n_{\text{eff}} = \sum_i K((X_{i1} - x_1)/h)$. Increasing h spreads the weights more evenly, raising the effective local sample size and increasing resilience to concentrated contamination.

Remark 2 (Bandwidth selection for the variance estimator). For the bandwidths h_1 , h_2 , and h_3 used in the variance estimator of Section 4.4 of the main paper, we recommend selecting $h_1 = h$ (the same cross-validated bandwidth used for the primary estimator), since $\hat{\sigma}_\tau^2(x_1)$ targets the conditional variance of plug-in quantities that share the same local neighbourhood structure. For h_2 and h_3 , which target the derivative quantities D^1 and D^0 , a separate cross-validation on the corresponding weighted residuals $T_i\psi'(Y_i - \hat{\mu}_1(x_1))/\hat{\pi}(X_i)$ and $(1 - T_i)\psi'(Y_i - \hat{\mu}_0(x_1))/(1 - \hat{\pi}(X_i))$ can be used. Alternatively, a simple rule-of-thumb is $h_2 = h_3 = c \cdot h$ for some constant $c > 1$ (e.g., $c = 1.5$), reflecting that derivative estimation tolerates somewhat more smoothing. In our simulations and application, we set

$h_1 = h_2 = h_3 = h$ and find adequate coverage; a sensitivity analysis varying these bandwidths separately showed minimal impact on the coverage probabilities.

Bibliography

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests.

Ann. Statist. *47*(2), 1148–1178.

Härdle, W. (1984). Robust regression function estimation. *Journal of Mul-*

tivariate Analysis *14*(2), 169–180.

Huber, P. J. (1984). Finite sample breakdown of m - and p -estimators. *The*

Annals of Statistics *12*(1), 119–126.

Kennedy, E. H. (2023). Towards optimal doubly robust estimation of hetero-

geneous causal effects. *Electronic Journal of Statistics* *17*(2), 3008–3049.

Künzel, S. R., J. S. Sekhon, P. J. Bickel, and B. Yu (2019). Metalearners

for estimating heterogeneous treatment effects using machine learning.

Proceedings of the national academy of sciences *116*(10), 4156–4165.

Nie, X. and S. Wager (2021). Quasi-oracle estimation of heterogeneous

treatment effects. *Biometrika* *108*(2), 299–319.

Ruppert, D., S. J. Sheather, and M. P. Wand (1995). An effective bandwidth selector for local least squares regression. *Journal of the American Statistical Association* 90(432), 1257–1270.