Supplementary Information *for*

Generalizing Trimming Bounds for Endogenously Missing Outcome Data Using Random Forests

Cyrus Samii (NYU) Ye Wang (UNC) Junlong Aaron Zhou (Independent Researcher)

Contents

A Algorithm

Algorithm 1 summarizes the estimation strategy. We first randomly split the sample into K folds for cross-fitting. Next, we employ the "probability forest" and "quantile forest"—both are variants of *grf*—to estimate the nuisance parameters, using $(K-1)$ folds from the sample. Then, we rely on the remaining fold to estimate either the conditional or the aggregated bounds, with the previously estimated nuisance parameters plugged in. The aggregated bounds can be directly estimated by the sample analogues of their orthogonalized moment conditions. For the conditional ones, we first estimate $\theta(\mathbf{x}) = (\theta_1^L(\mathbf{x}), \theta_1^U(\mathbf{x}), \theta_0(\mathbf{x}))$ by applying the "regression forest," another variant of *grf*, to their orthogonalized moment conditions, and then calculate the bounds according to their definition.

Algorithm 1: Honest inference for the covariate-tightened trimming bounds

1 Define the set of covariates $\mathbf{X} = (X_1, \ldots, X_P)$ and evaluation points $\mathcal{X}$ (all sample values for aggregate bounds or specified evaluation points for conditional bounds).
2 Randomly split the sample into K sets $(\mathcal{I}_1, \mathcal{I}_2, \ldots, \mathcal{I}_K)$ with approximately equal size.
3 **for** *each* $k \in \{1, 2, \ldots, K\}$ **do**
4 Fit the $q(\mathbf{X})$ model using the probability forests on $\mathcal{I}_{-k} = \cup_{j \neq k} \mathcal{I}_j$.
5 For any $\mathbf{x} \in \mathcal{X}$, treat $q(\mathbf{x})$ as fixed and fit the models $(y_{q(\mathbf{x})}(\mathbf{X}),\ y_{1-q(\mathbf{x})}(\mathbf{X}))$ using the quantile forests on $\mathcal{I}_{-k}$.
6 Treat $q(\mathbf{x})$, $y_{q(\mathbf{x})}(\mathbf{x})$ and $y_{1-q(\mathbf{x})}(\mathbf{x})$ as pre-fixed, construct orthogonalized moment conditions, and fit them on $\mathcal{I}_k$.
7 Calculate estimates for the conditional bounds $(\hat{\tau}^U_{CTB,\mathbf{x}}(1,1), \hat{\tau}^L_{CTB,\mathbf{x}}(1,1))$ or aggregated bounds $(\hat{\tau}^U_{CTB}(1,1), \hat{\tau}^L_{CTB}(1,1))$ using the orthogonalized moment conditions.
8 Take the average over the K estimates.

In the next section of the Appendix, we show formally that following Algorithm 1 yields estimates of either the conditional bounds, $(\hat{\tau}^U_{CTB,\mathbf{x}}(1,1), \hat{\tau}^L_{CTB,\mathbf{x}}(1,1))$, or the aggregated bounds, $(\hat{\tau}^U_{CTB}(1,1), \hat{\tau}^L_{CTB}(1,1))$, that are consistent and asymptotically Normal. Inference based on this algorithm is honest by the definition of Athey and Imbens (2016), because it accounts for the uncertainties caused by estimating nuisance parameters properly. For the conditional bounds, the standard error estimates are available from the regression forest output. For the aggregated bounds, we can calculate their standard errors from the variance of the orthogonalized moment conditions. We can then use normal approximation to construct asymptotically valid confidence intervals for either the conditional or the aggregated bounds. Note that they are not confidence intervals for the average always-responder effect *per se*, but rather for the bounds, and hence tend to be conservative in covering the true always-responder effect (Imbens and Manski, 2004; Stoye, 2009).

B Proofs

B.1 Improvement from incorporating covariates

Here we show how incorporating the information from covariates reduces the length of the partially identified set (i.e., the length of the interval between the bounds). We first re-state a lemma from Lee (2002):

Lemma B.1. *Suppose* $f_Y(y) = qf_Y^\dagger(y) + (1-q)f_Y^\ddagger(y)$, *then*

$$\frac{1}{q}\int_{-\infty}^{y_q} y f_Y(y)dy \leq \int y f_Y^\dagger(y)dy,$$
$$\frac{1}{1-q}\int_{y_{1-q}}^{\infty} y f_Y(y)dy \geq \int y f_Y^\ddagger(y)dy,$$

where $\int_{-\infty}^{y_q} f_Y(y)dy = q$.

Proof. Let's denote $\int_{y_q}^{\infty} f_Y^\dagger(y)dy$ as κ, then

$$\frac{1}{\kappa}\left[\frac{1}{q}\int_{-\infty}^{y_q} y f_Y(y)dy - \int y f_Y^\dagger(y)dy\right]$$
$$= \int_{-\infty}^{y_q} y\frac{f_Y(y)/q - f_Y^\dagger(y)}{\kappa}dy - \int_{y_q}^{\infty} y\frac{f_Y^\dagger(y)}{\kappa}dy.$$

Note that $\int_{-\infty}^{y_q}\frac{f_Y(y)/q - f_Y^\dagger(y)}{\kappa}dy = \frac{1}{\kappa} - \frac{1-\kappa}{\kappa} = 1$ and $\int_{y_q}^{\infty}\frac{f_Y^\dagger(y)}{\kappa}dy = 1$, hence the first term is the expectation of Y over $[-\infty, y_q]$ and the second is the expectation of Y over $[y_q, \infty]$. As the first support is below the second one, the first expectation is always smaller than the second as long as $f_Y^\dagger(y)$ is not degenerate. The other inequality can be similarly proved.

□

For the lower bound, we define

$$f^\dagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y) = f_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)\mathbf{1}\{y \leq y_{q(\mathbf{x})}(\mathbf{x})\}/q(\mathbf{x})$$
$$f^\ddagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y) = f_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)\mathbf{1}\{y \geq y_{q(\mathbf{x})}(\mathbf{x})\}/(1-q(\mathbf{x}))$$
$$f^\dagger_{Y|D=0,S=1}(y) = \int f^\dagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)\frac{q(\mathbf{x})}{q}f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}$$
$$f^\ddagger_{Y|D=0,S=1}(y) = \int f^\ddagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)\frac{1-q(\mathbf{x})}{1-q}f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}.$$

Then, by definition,

$$\begin{aligned}
f_{Y|D=0,S=1}(y) =& \int f_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}\\
=& \int f^\dagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)q(\mathbf{x})f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}\\
&+ \int f^\ddagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)(1-q(\mathbf{x}))f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}\\
=& q\int f^\dagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)\frac{q(\mathbf{x})}{q}f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}\\
&+ (1-q)\int f^\ddagger_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)\frac{1-q(\mathbf{x})}{1-q}f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}\\
=& qf^\dagger_{Y|D=0,S=1}(y) + (1-q)f^\ddagger_{Y|D=0,S=1}(y)
\end{aligned}$$

From Lemma B.1, we know that

$$
\begin{aligned}
&E[Y_i|D_i=0,S_i=1,Y_i\leq y_q] = \frac{1}{q}\int_{-\infty}^{y_q} y f_{Y|D=0,S=1}(y)dy \\
\leq &\int y f^{\dagger}_{Y|D=0,S=1}(y)dy = \int\int y f^{\dagger}_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)\frac{q(\mathbf{x})}{q} f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}dy \\
= &\int \theta_0^L(\mathbf{x})\frac{q(\mathbf{x})}{q} f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x} \\
= &\int \theta_0^L(\mathbf{x}) f_{\mathbf{X}|D=1,S=1}(\mathbf{x})d\mathbf{x} \\
= &E[E[Y_i|D_i=0,\mathbf{X}_i=\mathbf{x},Y_i\leq y_{q(\mathbf{x})}(\mathbf{x})]]
\end{aligned}
$$

The second to the last equality holds since by definition

$$
\begin{aligned}
q(\mathbf{x}) = &\frac{P[S=1,D=1|\mathbf{X}=\mathbf{x}]}{p(\mathbf{x})} = \frac{P[S=1,D=1,\mathbf{X}=\mathbf{x}]}{p(\mathbf{x})P[\mathbf{X}=\mathbf{x}]} \\
= &\frac{P[\mathbf{X}=\mathbf{x}|S=1,D=1]P[S=1,D=1]}{p(\mathbf{x})P[\mathbf{X}=\mathbf{x}]} \\
= &\frac{f_{\mathbf{X}|D=1,S=1}(\mathbf{x})P[S=1,D=1]}{p(\mathbf{x})f_{\mathbf{X}|D=0,S=1}(\mathbf{x})} \\
= &\frac{q f_{\mathbf{X}|D=1,S=1}(\mathbf{x})}{f_{\mathbf{X}|D=0,S=1}(\mathbf{x})}.
\end{aligned}
$$

From the derivation, we see that

$$
\begin{aligned}
&E[Y_i|D_i=0,S_i=1,Y_i\leq y_q] - E[E[Y_i|D_i=0,\mathbf{X}_i=\mathbf{x},Y_i\leq y_{q(\mathbf{x})}(\mathbf{x})]] \\
= &\frac{1}{q}\int_{-\infty}^{y_q} y f_{Y|D=0,S=1}(y)dy - \int \theta_0^L(\mathbf{x})\frac{q(\mathbf{x})}{q} f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x} \\
= &\int \theta_0^L f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x} - \int \theta_0^L(\mathbf{x})\frac{q(\mathbf{x})}{q} f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x} \\
= &\int \frac{\theta_0^L q - \theta_0^L(\mathbf{x})q(\mathbf{x})}{q} f_{\mathbf{X}|D=0,S=1}(\mathbf{x})d\mathbf{x}
\end{aligned}
$$

Therefore, the improvement is more pronounced when the $\theta_0^L q - \theta_0^L(\mathbf{x})q(\mathbf{x})$ is small where $f_{\mathbf{X}|D=0,S=1}(\mathbf{x})$ is small.

Let's consider a simple case where X takes two values $\{1,2\}$ with equal probability (0.5). Suppose $q(1)=0.2$, $q(2)=0.4$, then $q=0.3$. Note that

$$
f_{Y|D=0,S=1}(y) = 0.5 * f_{Y|D=0,S=1,X=1}(y) + 0.5 * f_{Y|D=0,S=1,X=2}(y).
$$

We illustrate why the CTB lead to a smaller identified set using Figure 1. The top-left plot shows how we construct the lower bound using the basic method, where the trimming probability equals 0.3 for both values of X. We take expectation of $Y_i(1)$ over the shaded region under each condition distribution and their average equals the lower bound of the treated outcome. The top-right plot shows the idea behind the CTB, where the trimming probability equals 0.2 when $X=1$ and 0.4 when $X=2$. We follow the same steps, taking conditional expectations first and then calculating their average. Note that in the two plots, the total area of the shaded region is the same. But the value of $Y_i(1)$ is strictly larger in the top-right plot, as illustrated by the plot at the bottom. This explains where the improvement comes from under the CTB.

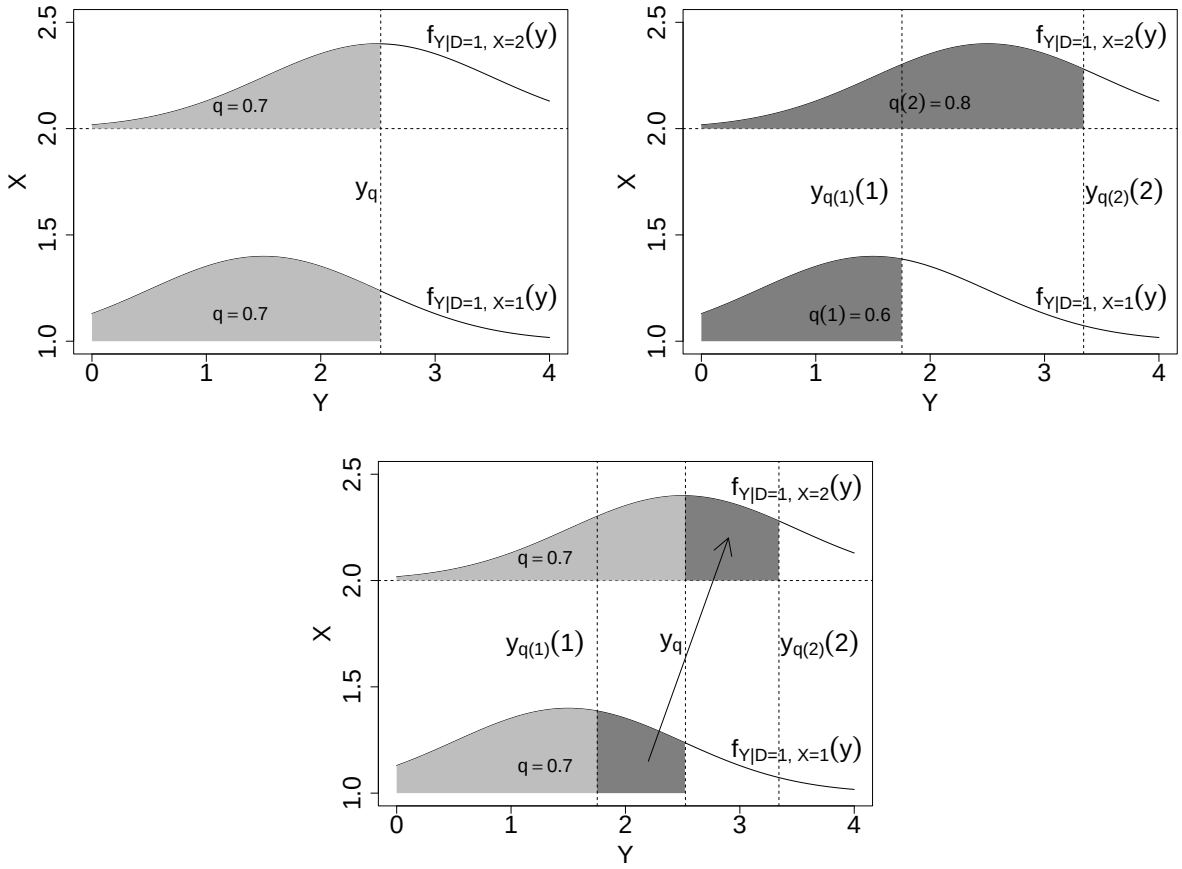


Figure 1: An illustration of the CTB

B.2 Moment conditions

We start from the scenario where Assumptions 1 and 2 are satisfied. It is equivalent to conditioning on the set $\mathcal{X}^+$ under Assumption 3. Following the terminology in the generalized method of moments, we denote $\nu(\mathbf{x}) = c(q_0(\mathbf{x}), q_1(\mathbf{x}), y_{q(\mathbf{x})}(\mathbf{x}), y_{1-q(\mathbf{x})}(\mathbf{x}))$ as the nuisance parameters and $\theta(\mathbf{x}) = c(\theta_0(\mathbf{x}), \theta_1^L(\mathbf{x}), \theta_1^U(\mathbf{x}))$ as the target parameters. Note that $q(\mathbf{x}) = \frac{q_0(\mathbf{x})}{q_1(\mathbf{x})}$ and $p(\mathbf{x}) = P(D=1|\mathbf{X}=\mathbf{x})$. To facilitate illustration, we focus on the modified target parameters $\tilde{\theta}(\mathbf{x}) = c(\tilde{\theta}_0(\mathbf{x}), \tilde{\theta}_1^L(\mathbf{x}), \tilde{\theta}_1^U(\mathbf{x})) = c(q_0(\mathbf{x})\theta_0(\mathbf{x}), q_0(\mathbf{x})\theta_1^L(\mathbf{x}), q_0(\mathbf{x})\theta_1^U(\mathbf{x}))$. We use the superscript ^{0} to denote the true value of a parameter. From their definition, we can see that the nuisance parameters satisfy the following local moment conditions:

$$
\begin{aligned}
E\left[m^{(1)}(\mathbf{O}_i;\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right] &= E\left[S_i(1-D_i) - q_0(\mathbf{x})(1-p(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right], \\
E\left[m^{(2)}(\mathbf{O}_i;\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right] &= E\left[S_iD_i - q_1(\mathbf{x})p(\mathbf{x})|\mathbf{X}_i=\mathbf{x}\right], \\
E\left[m^{(3)}(\mathbf{O}_i;\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right] &= E\left[S_iD_i\mathbf{1}\{Y_i \le y_{q(\mathbf{x})}(\mathbf{x})\} - q_0(\mathbf{x})p(\mathbf{x})|\mathbf{X}_i=\mathbf{x}\right], \\
E\left[m^{(4)}(\mathbf{O}_i;\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right] &= E\left[S_iD_i\mathbf{1}\{Y_i \ge y_{1-q(\mathbf{x})}(\mathbf{x})\} - q_0(\mathbf{x})p(\mathbf{x})|\mathbf{X}_i=\mathbf{x}\right].
\end{aligned}
$$

For the target parameters, we can show that under Assumptions 1 and 2, they satisfy the following local moment conditions:

$$
\begin{aligned}
&E\left[\psi^{(1)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right] = E\left[S_i(1-D_i)Y_i - (1-p(\mathbf{x}))\tilde{\theta}_0(\mathbf{x})|\mathbf{X}_i=\mathbf{x}\right] \\
=&(1-p(\mathbf{x}))\left[q_0^0(\mathbf{x})\int yf_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)dy - \tilde{\theta}_0(\mathbf{x})\right], \\
&E\left[\psi^{(2)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right] = E\left[S_iD_iY_i\mathbf{1}\{Y_i \le y_{q(\mathbf{x})}(\mathbf{x})\} - p(\mathbf{x})\tilde{\theta}_1^L(\mathbf{x})|\mathbf{X}_i=\mathbf{x}\right] \\
=&p(\mathbf{x})\left[q_1^0(\mathbf{x})\int_{-\infty}^{y_{q(\mathbf{x})}(\mathbf{x})} yf_{Y|D=1,S=1,\mathbf{X}=\mathbf{x}}(y)dy - \tilde{\theta}_1^L(\mathbf{x})\right], \\
&E\left[\psi^{(3)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right] = E\left[S_iD_iY_i\mathbf{1}\{Y_i \ge y_{1-q(\mathbf{x})}(\mathbf{x})\} - p(\mathbf{x})\tilde{\theta}_1^U(\mathbf{x})|\mathbf{X}_i=\mathbf{x}\right] \\
=&p(\mathbf{x})\left[q_1^0(\mathbf{x})\int_{y_{1-q(\mathbf{x})}(\mathbf{x})}^{\infty} yf_{Y|D=1,S=1,\mathbf{X}=\mathbf{x}}(y)dy - \tilde{\theta}_1^U(\mathbf{x})\right].
\end{aligned}
$$

When $\mathbf{X}$ is discrete or low-dimensional, we can approximate each of the moment conditions with stratification or kernel regression (Lee, 2009; Olma, 2020). We then solve $\nu(\mathbf{x})$ from their moment conditions and plug their values into the moment conditions above to solve $\theta(\mathbf{x})$. Nevertheless, with a large number of covariates, this approach suffers from the "curse of dimensionality" (Robins and Ritov, 1997). Therefore, we adopt the generalized random forest algorithm (*grf*) (Wager and Athey, 2018; Athey et al., 2019) to approximate each moment condition with the following expression:

$$
E[\psi_{\theta(\mathbf{x}),\nu(\mathbf{x})}(\mathbf{O}_i)|\mathbf{X}_i=\mathbf{x}] \approx \sum_{i=1}^{N}\alpha_i(\mathbf{x})\psi_{\theta(\mathbf{x}),\nu(\mathbf{x})}(\mathbf{O}_i)
$$

where $\psi(\cdot)$ represents the moment function, $\mathbf{O}_i = (Y_i, S_i, D_i)$, and $\alpha_i(\mathbf{x})$ is an adaptive kernel that weights the contribution of each observation i to the moment condition evaluated at $\mathbf{x}$. The *grf* estimation of $\alpha_i(\mathbf{x})$ is "honest" in the sense that only half of each sub-sample is used to train the tree and the other half is reserved for making predictions. We can estimate any parameter by minimizing its approximated local moment condition's L_2 norm, $||\sum_{i=1}^{N}\alpha_i(\mathbf{x})\psi_{\theta(\mathbf{x}),\hat{\nu}(\mathbf{x})}(\mathbf{O}_i)||_2$.

Nevertheless, as illustrated in the main text, using the same sample to estimate both $\nu(\mathbf{x})$ and $\theta(\mathbf{x})$ leads to the regularization bias. It can be avoided if we implement both cross-fitting and Neyman orthogonalization (Belloni et al., 2017; Chernozhukov et al., 2018). The idea behind Neyman orthogonalization is to adjust the local moment conditions for $\theta(\mathbf{x})$ such that they are invariant to small perturbations of the estimated $\nu(\mathbf{x})$. To be more precise, If a moment condition $\psi(\cdot)$ satisfies Neyman orthogonality, then the following two equations must hold:

$$E[\psi_{\theta^0(\mathbf{x}),\nu^0(\mathbf{x})}(\mathbf{O}_i)|\mathbf{X}_i = \mathbf{x}] = 0,$$

and

$$\frac{\partial}{\partial \nu} E[\psi_{\theta^0(\mathbf{x}),\nu(\mathbf{x})}(\mathbf{O}_i)|\mathbf{X}_i = \mathbf{x}]|_{\nu(\mathbf{x})=\nu^0(\mathbf{x})} = 0.$$

The process of modifying a moment condition such that it satisfies the two equations is termed as Neyman orthogonalization.

Note that our algorithm is essentially a three-step process: we first estimate $q(\mathbf{x})$, then $(y_{q(\mathbf{x})}(\mathbf{x}), y_{1-q(\mathbf{x})}(\mathbf{x}))$, finally the conditional or aggregated bounds, with the estimates of nuisance parameters plugged in. Therefore, to conduct Neyman orthogonalization, we first orthogonalize $m^{(3)}(\mathbf{O}_i;\nu(\mathbf{x}))$ and $m^{(4)}(\mathbf{O}_i;\nu(\mathbf{x}))$, and then $\psi^{(2)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))$ and $\psi^{(3)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))$ with regards to the nuisance parameters. We use $\tilde{m}^{(3)}(\mathbf{O}_i;\nu(\mathbf{x}))$, $\tilde{m}^{(4)}(\mathbf{O}_i;\nu(\mathbf{x}))$, $\tilde{\psi}^{(2)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))$, and $\tilde{\psi}^{(3)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))$ to denote the orthogonalized moments. We have

$$\begin{aligned}
&\tilde{m}^{(3)}(\mathbf{O}_i;\nu(\mathbf{x})) = m^{(3)}(\mathbf{O}_i;\nu(\mathbf{x})) - \frac{p(\mathbf{x})}{1-p(\mathbf{x})} m^{(1)}(\mathbf{O}_i;\nu(\mathbf{x}))\\
=&S_i D_i \mathbf{1}\{Y_i \le y_{q(\mathbf{x})}(\mathbf{x})\} - \frac{S_i(1-D_i)p(\mathbf{x})}{1-p(\mathbf{x})}\\
&\tilde{m}^{(4)}(\mathbf{O}_i;\nu(\mathbf{x})) = m^{(4)}(\mathbf{O}_i;\nu(\mathbf{x})) - \frac{p(\mathbf{x})}{1-p(\mathbf{x})} m^{(1)}(\mathbf{O}_i;\nu(\mathbf{x}))\\
=&S_i D_i \mathbf{1}\{Y_i \ge y_{1-q(\mathbf{x})}(\mathbf{x})\} - \frac{S_i(1-D_i)p(\mathbf{x})}{1-p(\mathbf{x})}\\
&\tilde{\psi}^{(2)}(\mathbf{O};\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x})) = \psi^{(2)}(\mathbf{O};\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x})) - y^0_{q(\mathbf{x})}(\mathbf{x})\tilde{m}^{(3)}_i(\mathbf{O}_i;\nu(\mathbf{x}))\\
=&S_i D_i (Y_i - y^0_{q(\mathbf{x})}(\mathbf{x}))\mathbf{1}\{Y_i \le y_{q(\mathbf{x})}(\mathbf{x})\} - p(\mathbf{x})\tilde{\theta}^L_1(\mathbf{x}) + \frac{y^0_{q(\mathbf{x})}(\mathbf{x})p(\mathbf{x})S_i(1-D_i)}{1-p(\mathbf{x})}\\
&\tilde{\psi}^{(3)}(\mathbf{O};\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x})) = \psi^{(3)}(\mathbf{O};\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x})) - y^0_{1-q(\mathbf{x})}(\mathbf{x})\tilde{m}^{(4)}_i(\mathbf{O}_i;\nu(\mathbf{x}))\\
=&S_i D_i (Y_i - y^0_{1-q(\mathbf{x})}(\mathbf{x}))\mathbf{1}\{Y_i \ge y_{1-q(\mathbf{x})}(\mathbf{x})\} - p(\mathbf{x})\tilde{\theta}^U_1(\mathbf{x}) + \frac{y^0_{1-q(\mathbf{x})}(\mathbf{x})p(\mathbf{x})S_i(1-D_i)}{1-p(\mathbf{x})}.
\end{aligned}$$

As each of the orthogonalized moments is a linear combination of two original moments, the first requirement is satisfied. To verify that the second requirement is also satisfied, first note that

$\tilde{m}^{(3)}(\mathbf{O}_i;\nu(\mathbf{x}))$ and $\tilde{m}^{(4)}(\mathbf{O}_i;\nu(\mathbf{x}))$ no longer depend on $q(\mathbf{x})$. Furthermore,

$$
\begin{aligned}
&\frac{\partial E\left[\tilde{\psi}^{(2)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right]}{\partial y_{q(\mathbf{x})}(\mathbf{x})}\Big|_{\nu(\mathbf{x})=\nu^0(\mathbf{x})}\\
=&\frac{\partial E\left[\psi^{(2)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right]}{\partial y_{q(\mathbf{x})}(\mathbf{x})}\Big|_{\nu(\mathbf{x})=\nu^0(\mathbf{x})}-y^0_{q(\mathbf{x})}(\mathbf{x})\frac{\partial E\left[\tilde{m}^{(3)}(\mathbf{O}_i;\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right]}{\partial y_{q(\mathbf{x})}(\mathbf{x})}\Big|_{\nu(\mathbf{x})=\nu^0(\mathbf{x})}\\
=&p(\mathbf{x})q^0_1(\mathbf{x})y^0_{q(\mathbf{x})}(\mathbf{x})f_{Y|D=1,S=1,\mathbf{X}=\mathbf{x}}(y^0_{q(\mathbf{x})}(\mathbf{x}))-y^0_{q(\mathbf{x})}(\mathbf{x})p(\mathbf{x})q^0_1(\mathbf{x})f_{Y|D=1,S=1,\mathbf{X}=\mathbf{x}}(y^0_{q(\mathbf{x})}(\mathbf{x}))=0\\
&\frac{\partial E\left[\psi^{(3)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right]}{\partial y_{1-q(\mathbf{x})}(\mathbf{x})}\Big|_{\nu(\mathbf{x})=\nu^0(\mathbf{x})}\\
=&\frac{\partial E\left[\psi^{(3)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right]}{\partial y_{1-q(\mathbf{x})}(\mathbf{x})}\Big|_{\nu(\mathbf{x})=\nu^0(\mathbf{x})}-y^0_{1-q(\mathbf{x})}(\mathbf{x})\frac{\partial E\left[\tilde{m}^{(4)}(\mathbf{O}_i;\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right]}{\partial y_{1-q(\mathbf{x})}(\mathbf{x})}\Big|_{\nu(\mathbf{x})=\nu^0(\mathbf{x})}\\
=&-p(\mathbf{x})q^0_1(\mathbf{x})y^0_{1-q(\mathbf{x})}(\mathbf{x})f_{Y|D=1,S=1,\mathbf{X}=\mathbf{x}}(y^0_{1-q(\mathbf{x})}(\mathbf{x}))\\
&+y^0_{1-q(\mathbf{x})}(\mathbf{x})p(\mathbf{x})q^0_1(\mathbf{x})f_{Y|D=1,S=1,\mathbf{X}=\mathbf{x}}(y^0_{1-q(\mathbf{x})}(\mathbf{x}))=0.
\end{aligned}
$$

Hence, $\tilde{m}^{(3)}(\mathbf{O}_i;\nu(\mathbf{x}))$, $\tilde{m}^{(4)}(\mathbf{O}_i;\nu(\mathbf{x}))$, $\tilde{\psi}^{(2)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))$, and $\tilde{\psi}^{(3)}(\mathbf{O}_i;\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))$ all satisfy Neyman orthogonality, which justifies our algorithm in the main text. It is worth noting that $E\left[\frac{S_i(1-D_i)}{1-p(\mathbf{x})}|\mathbf{X}_i=\mathbf{x}\right]=q^0_0(\mathbf{x})$. Thus, Neyman orthogonalization does not affect the estimation of $y_{q(\mathbf{x})}(\mathbf{x})$ or $y_{1-q(\mathbf{x})}(\mathbf{x})$. For the conditional lower bound $\tilde{\theta}^L_1(\mathbf{x})$, we know that

$$
\begin{aligned}
&E\left[\tilde{\psi}^{(2)}(\mathbf{O};\tilde{\theta}(\mathbf{x}),\nu(\mathbf{x}))|\mathbf{X}_i=\mathbf{x}\right]\\
=&E\left[S_iD_i(Y_i-y^0_{q(\mathbf{x})}(\mathbf{x}))\mathbf{1}\{Y_i\le y_{q(\mathbf{x})}(\mathbf{x})\}-p(\mathbf{x})\tilde{\theta}^L_1(\mathbf{x})+y^0_{q(\mathbf{x})}(\mathbf{x})p(\mathbf{x})q^0_0(\mathbf{x})|\mathbf{X}_i=\mathbf{x}\right]\\
=&E\left[(Y_i-y^0_{q(\mathbf{x})}(\mathbf{x}))\mathbf{1}\{Y_i\le y_{q(\mathbf{x})}(\mathbf{x})\}|S_i=1,D_i=1,\mathbf{X}_i=\mathbf{x}\right]p(\mathbf{x})q^0_1(\mathbf{x})\\
&-p(\mathbf{x})\tilde{\theta}^L_1(\mathbf{x})+y^0_{q(\mathbf{x})}(\mathbf{x})p(\mathbf{x})q^0_0(\mathbf{x}).
\end{aligned}
$$

Therefore,

$$
\theta^L_1(\mathbf{x})=\frac{E\left[(Y_i-y^0_{q(\mathbf{x})}(\mathbf{x}))\mathbf{1}\{Y_i\le y_{q(\mathbf{x})}(\mathbf{x})\}|S_i=1,D_i=1,\mathbf{X}_i=\mathbf{x}\right]}{q^0(\mathbf{x})}+y^0_{q(\mathbf{x})},
$$

and a similar formula holds for $\theta^U_1(\mathbf{x})$.

For the aggregated bounds, we conduct Neyman orthogonalization on the unconditional moments, which leads to the following score functions:

$$
\begin{aligned}
s^L(\mathbf{X}_i)&=\frac{S_iD_i(Y_i-\hat{y}_{\hat{q}(\mathbf{X}_i)}(\mathbf{X}_i))\mathbf{1}\{Y_i\le\hat{y}_{\hat{q}(\mathbf{X}_i)}(\mathbf{X}_i))\}}{p(\mathbf{X}_i)}-\frac{S_i(1-D_i)(Y_i-\hat{y}_{\hat{q}(\mathbf{X}_i)}(\mathbf{X}_i))}{1-p(\mathbf{X}_i)}\\
s^U(\mathbf{X}_i)&=\frac{S_iD_i(Y_i-\hat{y}_{1-\hat{q}(\mathbf{X}_i)}(\mathbf{X}_i))\mathbf{1}\{Y_i\ge\hat{y}_{1-\hat{q}(\mathbf{X}_i)}(\mathbf{X}_i)\}}{p(\mathbf{X}_i)}-\frac{S_i(1-D_i)(Y_i-\hat{y}_{1-\hat{q}(\mathbf{X}_i)}(\mathbf{X}_i))}{1-p(\mathbf{X}_i)}.
\end{aligned}
$$

Then,

$$
\hat{\tau}^L_{CTB}(1,1)=\frac{\frac{1}{N}\sum_{i=1}^N s^L(\mathbf{X}_i)}{\hat{q}_0},\hat{\tau}^U_{CTB}(1,1)=\frac{\frac{1}{N}\sum_{i=1}^N s^U(\mathbf{X}_i)}{\hat{q}_0},
$$

where $\hat{q}_0 = \frac{1}{N}\sum_{i=1}^{N} \frac{S_i(1-D_i)}{1-p(\mathbf{X}_i)}$.

When the propensity scores need to be estimated, we have an extra moment condition

$$E\left[m^{(0)}(\mathbf{O}_i; \nu(\mathbf{x}))|\mathbf{X}_i = \mathbf{x}\right] = E[D_i - p(\mathbf{x})|\mathbf{X}_i = \mathbf{x}] = 0.$$

We need to orthogonalize our moment conditions to account for the estimation error from $m^{(0)}(\cdot)$. It turns out that there will be an extra correction term in $\tilde{\psi}^{(1)}(\cdot)$, $\tilde{\psi}^{(2)}(\cdot)$, and $\tilde{\psi}^{(3)}(\cdot)$, which are $\tilde{\theta}_0(\mathbf{x})[D_i - p(\mathbf{x})]$, $y_{q(\mathbf{x})}q_0(\mathbf{x})\left[\frac{D_i}{p(\mathbf{x})} - \frac{1-D_i}{1-p(\mathbf{x})}\right] - \tilde{\theta}_1^L(\mathbf{x})[D_i - p(\mathbf{x})]$ and $y_{1-q(\mathbf{x})}q_0(\mathbf{x})\left[\frac{D_i}{p(\mathbf{x})} - \frac{1-D_i}{1-p(\mathbf{x})}\right] - \tilde{\theta}_1^U(\mathbf{x})[D_i - p(\mathbf{x})]$, respectively.

If the assumption of monotonic selection holds for the other direction: $S_i(0) \geq S_i(1)$, then the moment conditions $m^{(3)}(\mathbf{O}_i; \nu(\mathbf{x}))$, $m^{(4)}(\mathbf{O}_i; \nu(\mathbf{x}))$, and $\psi(\mathbf{O}; \tilde{\theta}(\mathbf{x}), \nu(\mathbf{x}))$ become:

$$\begin{aligned}
E\left[m_i^{(3)}(\mathbf{O}_i; \nu(\mathbf{x}))\right] &= E\left[S_i(1-D_i)\mathbf{1}\{Y_i \leq y_{q(\mathbf{x})}(\mathbf{x})\} - q_1^0(\mathbf{x})(1-p(\mathbf{x}))|\mathbf{X}_i = \mathbf{x}\right], \\
E\left[m_i^{(4)}(\mathbf{O}_i; \nu(\mathbf{x}))\right] &= E\left[S_i(1-D_i\mathbf{1})\{Y_i \geq y_{1-q(\mathbf{x})}(\mathbf{x})\} - q_1^0(\mathbf{x})(1-p(\mathbf{x}))|\mathbf{X}_i = \mathbf{x}\right] \\
E\left[\psi^{(1)}(\mathbf{O}; \tilde{\theta}(\mathbf{x}), \nu(\mathbf{x}))\right] &= E\left[S_i D_i Y_i - p(\mathbf{x})\tilde{\theta}_1(\mathbf{x})|\mathbf{X}_i = \mathbf{x}\right] \\
&= p(\mathbf{x})\left[q_1^0(\mathbf{x})\int y f_{Y|D=1,S=1,\mathbf{X}=\mathbf{x}}(y)dy - \tilde{\theta}_1(\mathbf{x})\right], \\
E\left[\psi^{(2)}(\mathbf{O}; \tilde{\theta}(\mathbf{x}), \nu(\mathbf{x}))\right] &= E\left[S_i(1-D_i)Y_i\mathbf{1}\{Y_i \leq y_{q(\mathbf{x})}(\mathbf{x})\} - (1-p(\mathbf{x}))\tilde{\theta}_0^L(\mathbf{x})|\mathbf{X}_i = \mathbf{x}\right] \\
&= (1-p(\mathbf{x}))\left[q_0^0(\mathbf{x})\int_{-\infty}^{y_{q(\mathbf{x})}(\mathbf{x})} y f_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)dy - \tilde{\theta}_0^L(\mathbf{x})\right], \\
E\left[\psi^{(3)}(\mathbf{O}; \tilde{\theta}(\mathbf{x}), \nu(\mathbf{x}))\right] &= E\left[S_i(1-D_i)Y_i\mathbf{1}\{Y_i \geq y_{1-q(\mathbf{x})}(\mathbf{x})\} - (1-p(\mathbf{x}))\tilde{\theta}_0^U(\mathbf{x})|\mathbf{X}_i = \mathbf{x}\right] \\
&= (1-p(\mathbf{x}))\left[q_0^0(\mathbf{x})\int_{y_{1-q(\mathbf{x})}(\mathbf{x})}^{\infty} y f_{Y|D=0,S=1,\mathbf{X}=\mathbf{x}}(y)dy - \tilde{\theta}_0^U(\mathbf{x})\right].
\end{aligned}$$

We can similarly obtain the orthogonalized moment conditions:

$$\begin{aligned}
&\tilde{\psi}^{(2)}(\mathbf{O}_i; \tilde{\theta}(\mathbf{x}), \nu(\mathbf{x})) \\
=&S_i(1-D_i)(Y_i - y_{q(\mathbf{x})}^0(\mathbf{x}))\mathbf{1}\{Y_i \leq y_{q(\mathbf{x})}(\mathbf{x})\} - (1-p(\mathbf{x}))\tilde{\theta}_1^L(\mathbf{x}) + \frac{(1-p(\mathbf{x}))y_{q(\mathbf{x})}^0(\mathbf{x})S_i D_i}{p(\mathbf{x})} \\
&\tilde{\psi}^{(3)}(\mathbf{O}; \tilde{\theta}(\mathbf{x}), \nu(\mathbf{x})) \\
=&S_i(1-D_i)(Y_i - y_{1-q(\mathbf{x})}^0(\mathbf{x}))\mathbf{1}\{Y_i \geq y_{1-q(\mathbf{x})}(\mathbf{x})\} - (1-p(\mathbf{x}))\tilde{\theta}_1^U(\mathbf{x}) \\
&+ \frac{(1-p(\mathbf{x}))y_{1-q(\mathbf{x})}^0(\mathbf{x})S_i D_i}{p(\mathbf{x})}.
\end{aligned}$$

When we have conditionally monotonic selection, $S_i(0) \leq S_i(1)$ for certain units and $S_i(0) \geq S_i(1)$ for the others. We define a binary variable $direction_i := \mathbf{1}\{\mathbf{X}_i \in \mathcal{X}^+\}$. When $\mathbf{X}_i \in \mathcal{X}^+$, we have score functions $s^{L,help}(\mathbf{X}_i)$ and $s_i^{U,help}(\mathbf{X}_i)$. When $\mathbf{X}_i \in \mathcal{X}^-$, we have $s^{L,hurt}(\mathbf{X}_i)$ and $s_i^{U,hurt}(\mathbf{X}_i)$. Finally, $s^L(\mathbf{X}_i) = direction_i * s^{L,help}(\mathbf{X}_i) + (1 - direction_i) * s^{L,hurt}(\mathbf{X}_i)$ and $s^U(\mathbf{X}_i) = direction_i * s^{U,help}(\mathbf{X}_i) + (1 - direction_i) * s^{U,hurt}(\mathbf{X}_i)$.

B.3 Statistical theory

We first derive the asymptotic behavior of the estimator for the aggregated bounds, using Theorem 3.1 in Chernozhukov et al. (2018). Our moment conditions are linear in the target parameters

and satisfy Neyman orthogonality, hence Assumption 3.1 required by the theorem holds. Another required assumption, Assumption 3.2, is ensured by the convergence rate of the *grf* algorithm. Therefore, under monotonic selection, we know that

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[s^L(\mathbf{X}_i) - q_0\tau^L_{CTB}(1,1)\right] \to \mathcal{N}(0, V_s^L)$$
$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left[s^U(\mathbf{X}_i) - q_0\tau^U_{CTB}(1,1)\right] \to \mathcal{N}(0, V_s^U)$$
$$\sqrt{N}(\hat{q}_0 - q_0) \to \mathcal{N}(0, V_{q_0}),$$

Using the Delta method, we can obtain

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\hat{\tau}^L_{CTB}(1,1) - \tau^L_{CTB}(1,1)\right) \to \mathcal{N}(0, V^L)$$
$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N}\left(\hat{\tau}^U_{CTB}(1,1) - \tau^U_{CTB}(1,1)\right) \to \mathcal{N}(0, V^U),$$

We can estimate V^L via its sample analogue:

$$\hat{V}^L = \begin{pmatrix} \frac{1}{\hat{q}_0} & -\frac{\hat{\tau}^L_{CTB}(1,1)}{\hat{q}_0} \end{pmatrix} \begin{pmatrix} \hat{V}_s^L & \widehat{Cov}_{(s^L,q_0)} \\ \widehat{Cov}_{(s^L,q_0)} & \hat{V}_{q_0} \end{pmatrix} \begin{pmatrix} \frac{1}{\hat{q}_0} \\ -\frac{\hat{\tau}^L_{CTB}(1,1)}{\hat{q}_0} \end{pmatrix},$$

where

$$\hat{V}_s^L = \frac{1}{N^2}\sum_{i=1}^{N}\left[s^L(\mathbf{X}_i) - \frac{1}{N}\sum_{i=1}^{N}s^L(\mathbf{X}_i)\right]^2, \hat{V}_{q_0} = \frac{1}{N^2}\sum_{i=1}^{N}\left[\frac{S_i(1-D_i)}{1-p(\mathbf{X}_i)} - \hat{q}_0\right]^2,$$
$$\widehat{Cov}_{(s^L,q_0)} = \frac{1}{N^2}\sum_{i=1}^{N}\left[s^L(\mathbf{X}_i) - \frac{1}{N}\sum_{i=1}^{N}s^L(\mathbf{X}_i)\right]\left[\frac{S_i(1-D_i)}{1-p(\mathbf{X}_i)} - \hat{q}_0\right].$$

Similar results hold under conditionally monotonic selection, as the estimate is a weighted average of two estimates under monotonic selection.

For the conditional bounds, it is easy to verify that the moment conditions for $\nu(\mathbf{x})$ satisfy all the assumptions imposed in Athey et al. (2019). These assumptions require that the moment conditions are sufficiently smooth in the parameters and the parameters are smooth functions of the data. Example 2 in Athey et al. (2019) verified each of the assumptions for conditional quantiles. Therefore, $\hat{\nu}(\mathbf{x})$ are consistent and asymptotically Normal, according to Theorem 5 in the paper. In addition, the theorem guarantees that these estimates converge to their true values at a sufficiently fast rate (higher than $1/N^{\frac{1}{4}}$). Since our algorithm incorporates both Neyman orthogonalization and cross-fitting, it is honest for the estimation of the target parameters. Consequently, the consistency and asymptotically Normality of $\hat{\theta}(\mathbf{x})$ and the conditional bounds, $(\hat{\tau}^U_{CTB,\mathbf{x}}(1,1), \hat{\tau}^L_{CTB,\mathbf{x}}(1,1))$, can be similarly derived from Theorem 5 in Athey et al. (2019). To summarize, we have the following formal results:

Theorem B.1. *Under regularity conditions required by Theorem 5 in Athey et al. (2019), estimates of* $\tau^L_{CTB,\mathbf{x}}(1,1)$, $\tau^U_{CTB,\mathbf{x}}(1,1)$, $\tau^L_{CTB}(1,1)$, *and* $\tau^U_{CTB}(1,1)$ *from Algorithm 1 are consistent and asymptotically Normal:*

$$\frac{\hat{\tau}^L_{CTB,\mathbf{x}}(1,1) - \tau^L_{CTB,\mathbf{x}}(1,1)}{\sqrt{Var(\hat{\tau}^L_{CTB,\mathbf{x}}(1,1))}} \to \mathcal{N}(0,1), \frac{\hat{\tau}^U_{CTB,\mathbf{x}}(1,1) - \tau^U_{CTB,\mathbf{x}}(1,1)}{\sqrt{Var(\hat{\tau}^U_{CTB,\mathbf{x}}(1,1))}} \to \mathcal{N}(0,1),$$

$$\frac{\hat{\tau}^L_{CTB}(1,1) - \tau^L_{CTB}(1,1)}{\sqrt{Var(\hat{\tau}^L_{CTB}(1,1))}} \to \mathcal{N}(0,1) \ , \frac{\hat{\tau}^U_{CTB}(1,1) - \tau^U_{CTB}(1,1)}{\sqrt{Var(\hat{\tau}^U_{CTB}(1,1))}} \to \mathcal{N}(0,1).$$

Theorem B.1 allows us to construct valid confidence intervals for either the conditional bounds or the aggregated bounds using estimates from our algorithm. One may also construct confidence intervals for the ATE on the always-responders following the approach developed by Imbens and Manski (2004) and Stoye (2009). Intuitively, this approach utilizes the idea that the ATE on the always-responders cannot appear on both ends of the identified set simultaneously. Therefore, as the width of the identified set increases, the critical value for a two-sided 95% confidence interval surrounding $\tau(1,1)$ will approach that of a one-sided 95% confidence interval surrounding the bounds. For the conditional bounds, the standard error estimates are available in the output of the regression forest. This is due to the fact that our moment conditions for the target parameters have been orthogonalized and accounted for the extra uncertainties caused by the estimation of $\nu(\mathbf{x})$ in the first stage. Finally, if we only have conditionally monotonic selection, some extra conditions are needed to ensure that we can predict $\mathcal{X}^+$ and $\mathcal{X}^-$ accurately. Details of these conditions are discussed in Semenova (2020).

C Extensions

In this section, we discuss extensions to unit-level missing data, binary outcomes, and estimated propensity scores (e.g., for non-experimental data).

C.1 Unit-level missing data

In practice, we sometimes do not even have covariate information on subjects who do not respond. This will happen, for instance, if we collect the information on $(Y_i, D_i, \mathbf{X}_i)$ only in a post-treatment survey. If so, it will be no longer feasible to estimate $q(\mathbf{x})$ directly, since we do not know the value of $\mathbf{X}_i$ when $S_i = 0$. However, under Assumption 1 from the main text, we can use Bayes' rule to show

$$\begin{aligned}
q(\mathbf{x}) =& P(S_i(1) = S_i(0) = 1 | \mathbf{X}_i = \mathbf{x}, S = 1, D = 1) \\
=& \frac{P(S_i(0) = 1 | D = 1, \mathbf{X}_i = \mathbf{x})}{P(S_i(1) = 1 | D = 1, \mathbf{X}_i = \mathbf{x})} \\
=& \frac{P(S_i(0) = 1 | \mathbf{X}_i = \mathbf{x})}{P(S_i(1) = 1 | \mathbf{X}_i = \mathbf{x})} \\
=& \frac{p(x) P(D = 0 | S = 1, \mathbf{X}_i = \mathbf{x})}{(1 - p(x)) P(D = 1 | S = 1, \mathbf{X}_i = \mathbf{x})} \\
=& \frac{p(x) P(D = 0 | S = 1, \mathbf{X}_i = \mathbf{x})}{(1 - p(x))(1 - P(D = 0 | S = 1, \mathbf{X}_i = \mathbf{x}))}
\end{aligned}$$

where the last equality comes from

$$\begin{aligned}
1 - P(D = 0 | S = 1, \mathbf{X}_i = \mathbf{x}) =& P(D = 1 | S = 1, \mathbf{X}_i = \mathbf{x}) \\
=& P(S = 1 | D = 1, \mathbf{X}_i = \mathbf{x}) \frac{P(D = 1 | \mathbf{X}_i = \mathbf{x})}{P(S = 1 | \mathbf{X}_i = \mathbf{x})} \\
=& P(S_i(1) = 1 | D = 1, \mathbf{X}_i = \mathbf{x}) \frac{p(x)}{P(S = 1 | \mathbf{X}_i = \mathbf{x})} \\
=& P(S_i(1) = 1 | \mathbf{X}_i = \mathbf{x}) \frac{p(x)}{P(S = 1 | \mathbf{X}_i = \mathbf{x})}, \\
P(D = 0 | S = 1, \mathbf{X}_i = \mathbf{x}) =& P(S = 1 | D = 0, \mathbf{X}_i = \mathbf{x}) \frac{P(D = 0 | \mathbf{X}_i = \mathbf{x})}{P(S = 1 | \mathbf{X}_i = \mathbf{x})} \\
=& P(S_i(0) = 1 | D = 0, \mathbf{X}_i = \mathbf{x}) \frac{1 - p(x)}{P(S = 1 | \mathbf{X}_i = \mathbf{x})} \\
=& P(S_i(0) = 1 | \mathbf{X}_i = \mathbf{x}) \frac{1 - p(x)}{P(S = 1 | \mathbf{X}_i = \mathbf{x})}.
\end{aligned}$$

$$\begin{aligned}
1 - p(\mathbf{x}) =& P(D_i = 0 | \mathbf{X}_i = \mathbf{x}) \\
=& P(D_i = 0 | S_i(1) = 1, S_i = 1, \mathbf{X}_i = \mathbf{x}) P(S_i(1) = 1 | S_i = 1, \mathbf{X}_i = \mathbf{x}) + \\
& P(D_i = 0 | S_i(1) = 0, S_i = 1, \mathbf{X}_i = \mathbf{x}) P(S_i(1) = 0 | S_i = 1, \mathbf{X}_i = \mathbf{x}) \\
=& P(D_i = 0 | A, \mathbf{X}_i = \mathbf{x}, S_i = 1) P(A | S_i = 1, \mathbf{X}_i = \mathbf{x}) + P(S_i(1) = 0 | S_i = 1, \mathbf{X}_i = \mathbf{x}) \\
=& \frac{P(D_i = 0, A | \mathbf{X}_i = \mathbf{x}, S_i = 1)}{P(A | S_i = 1, \mathbf{X}_i = \mathbf{x})} P(A | S_i = 1, \mathbf{X}_i = \mathbf{x}) + P(S_i(1) = 0 | S_i = 1, \mathbf{X}_i = \mathbf{x}) \\
=& P(D_i = 0 | \mathbf{X}_i = \mathbf{x}, S_i = 1) + P(S_i(1) = 0 | S_i = 1, \mathbf{X}_i = \mathbf{x}).
\end{aligned}$$

Therefore, we can infer the value of $q(\mathbf{x})$ from the estimate of $P(D_i = 0 | \mathbf{X}_i = \mathbf{x}, S_i = 1)$, which only involves observations with $S_i = 1$ and can be estimated with the probability forest.

C.2 Binary outcome

The outcome variable is often binary rather than continuous, in which case the conditional quantile is no longer continuous in $\mathbf{X}$, as it can only be either 0 or 1. Nevertheless, we can still estimate the bounds based on similar ideas as for continuous outcomes. Supposing that the observed mean is between 0 and 1, the upper bound would be based, essentially, on trimming the appropriate share of 0 outcomes, and the lower bound on trimming the appropriate share of 1 outcomes. Figure 2 provides a visual illustration. Formally, define $\xi(\mathbf{x}) = Pr(Y_i = 0|D_i = 1, S_i = 1, \mathbf{X}_i = \mathbf{x})$. In the case where $1 - q(\mathbf{x}) \leq \xi(\mathbf{x}) \leq q(\mathbf{x})$, for instance, the lower bound equals $\frac{q(\mathbf{x})-\xi(\mathbf{x})}{q(\mathbf{x})}$ and the upper bound equals $\frac{1-\xi(\mathbf{x})}{q(\mathbf{x})}$. In general, we have the following expressions for the lower and upper bound at point $\mathbf{x}$:

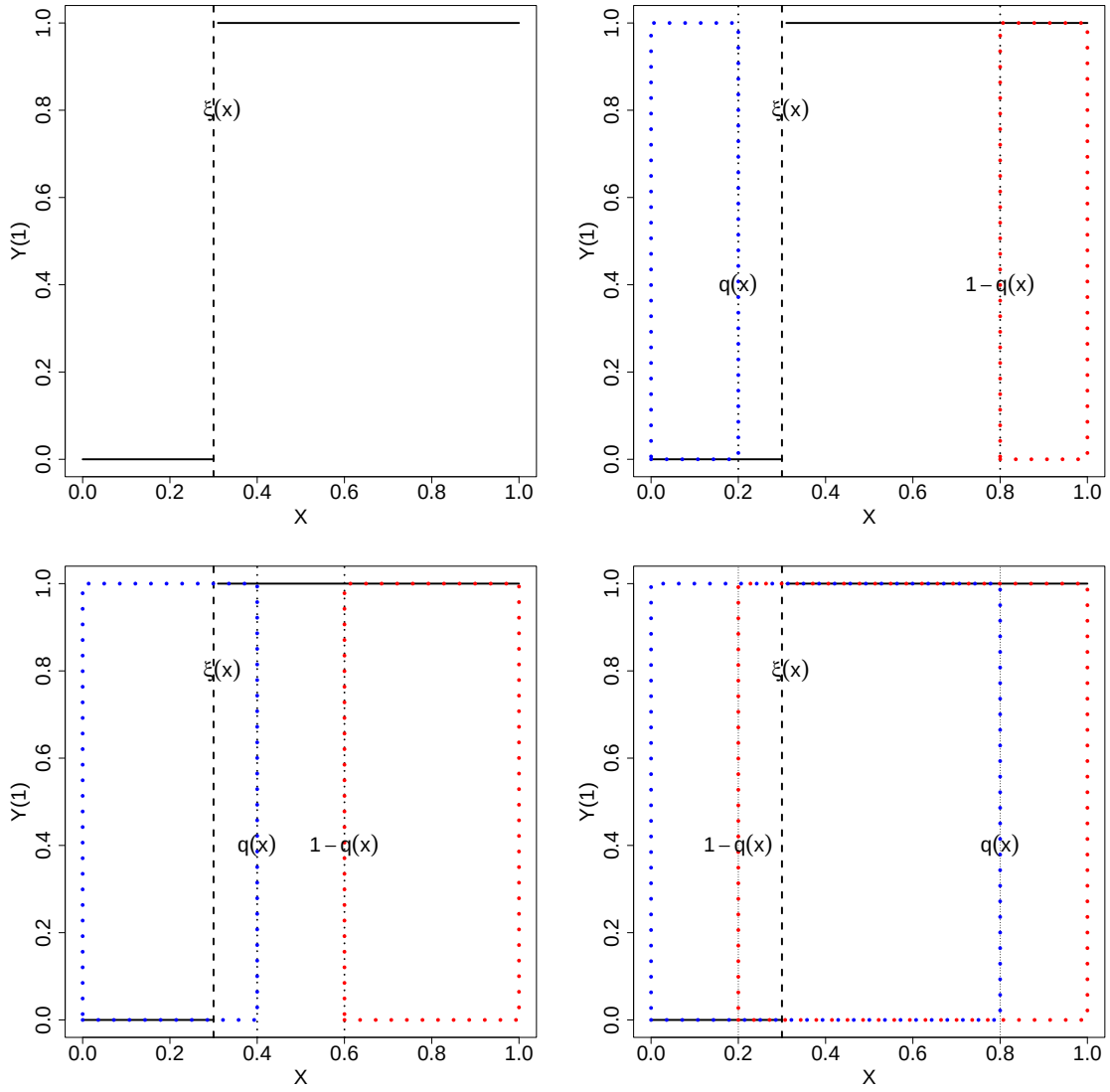
$$\hat{\tau}^{L}_{CTB,\mathbf{x}}(1,1) = \frac{(\hat{q}(\mathbf{x}) - \hat{\xi}(\mathbf{x}))\mathbf{1}\{\hat{\xi}(\mathbf{x}) \leq \hat{q}(\mathbf{x})\}}{\hat{q}(\mathbf{x})} - \hat{\theta}_0(\mathbf{x}),$$
$$\hat{\tau}^{U}_{CTB,\mathbf{x}}(1,1) = \frac{\hat{q}(\mathbf{x}) + (1 - \hat{q}(\mathbf{x}) - \hat{\xi}(\mathbf{x}))\mathbf{1}\{\hat{\xi}(\mathbf{x}) \geq 1 - \hat{q}(\mathbf{x})\}}{\hat{q}(\mathbf{x})} - \hat{\theta}_0(\mathbf{x}).$$

We can use the probability forest to estimate $\xi(\mathbf{x})$ and obtain estimates for the conditional bounds. Estimates for the aggregated bounds can be constructed by aggregating these conditional bounds.

C.3 Estimated propensity scores

In observational studies, the treatment propensity score $p(\mathbf{X})$ is unknown and has to be estimated from data. Since $p(\mathbf{X})$ affects all the moment conditions derived above, we need to orthogonalize them with regard to this extra nuisance parameter. This generalization is illustrated at the end of the analysis in section B.2 of the previous section. We also generate an extra split of data for the estimation of $p(\mathbf{X})$.

Figure 2: Trimming Bounds with Binary Outcome



Note: These plots demonstrate how to construct the trimming bounds when the outcome variable is binary. The top-left plot shows the distribution of $Y_i(1)$. The dashed line indicates the value of $\xi(\mathbf{x})$. In the rest three plots, the dotted lines mark the values of $q(\mathbf{x})$ and $1 - q(\mathbf{x})$. The rectangle with a blue border represents $\theta_1^L(\mathbf{x})$ and the rectangle with a red border represents $\theta_1^U(\mathbf{x})$. From top-right to bottom-right, the plots show cases where $q(\mathbf{x}) \leq \xi(\mathbf{x}) \leq 1 - q(\mathbf{x})$, $\xi(\mathbf{x}) \leq q(\mathbf{x}) \leq 1 - q(\mathbf{x})$, and $1 - q(\mathbf{x}) \leq \xi(\mathbf{x}) \leq q(\mathbf{x})$.

D Extra results

D.1 Extra results from simulation

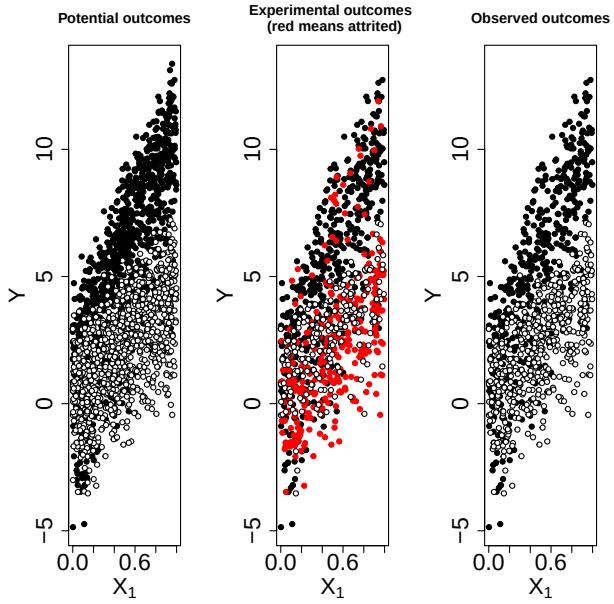
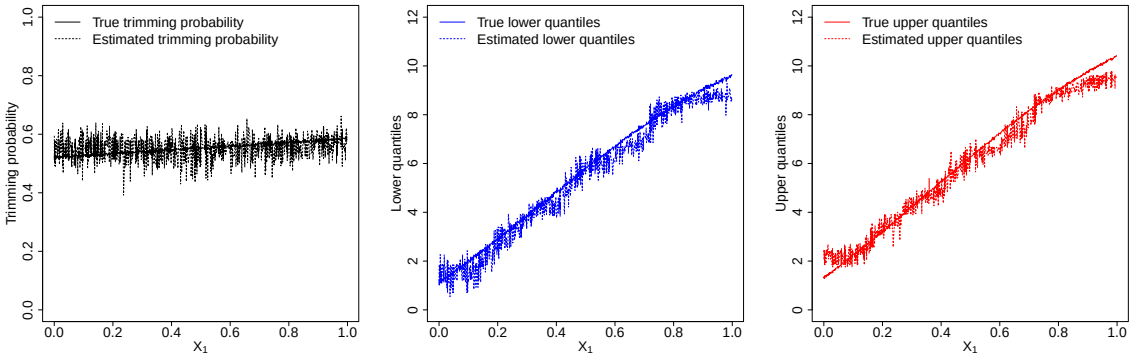


Figure 3: Simulated Data

Note: The left plot presents how $Y_i(0)$ (in white) and $Y_i(1)$ (in black) vary across the value of the first covariate for all the units in the simulated sample. The middle plot shows realized outcome for all the units under one assignment, with missing outcome values marked by red spots. The right plots shows outcome values that can be observed by the researcher.

Figure 4: Estimates of Nuisance Parameters



Note: These plots compare the estimated nuisance parameters with their true values. The x-axis is X_1, the only observable covariate that affects the response rates. Left: trimming probability $q_0(\mathbf{x})/q_1(\mathbf{x})$; Middle: lower quantile $y_{q\mathbf{x}}(\mathbf{x})$; Right: upper quantile $y_{1-q\mathbf{x}}(\mathbf{x})$.

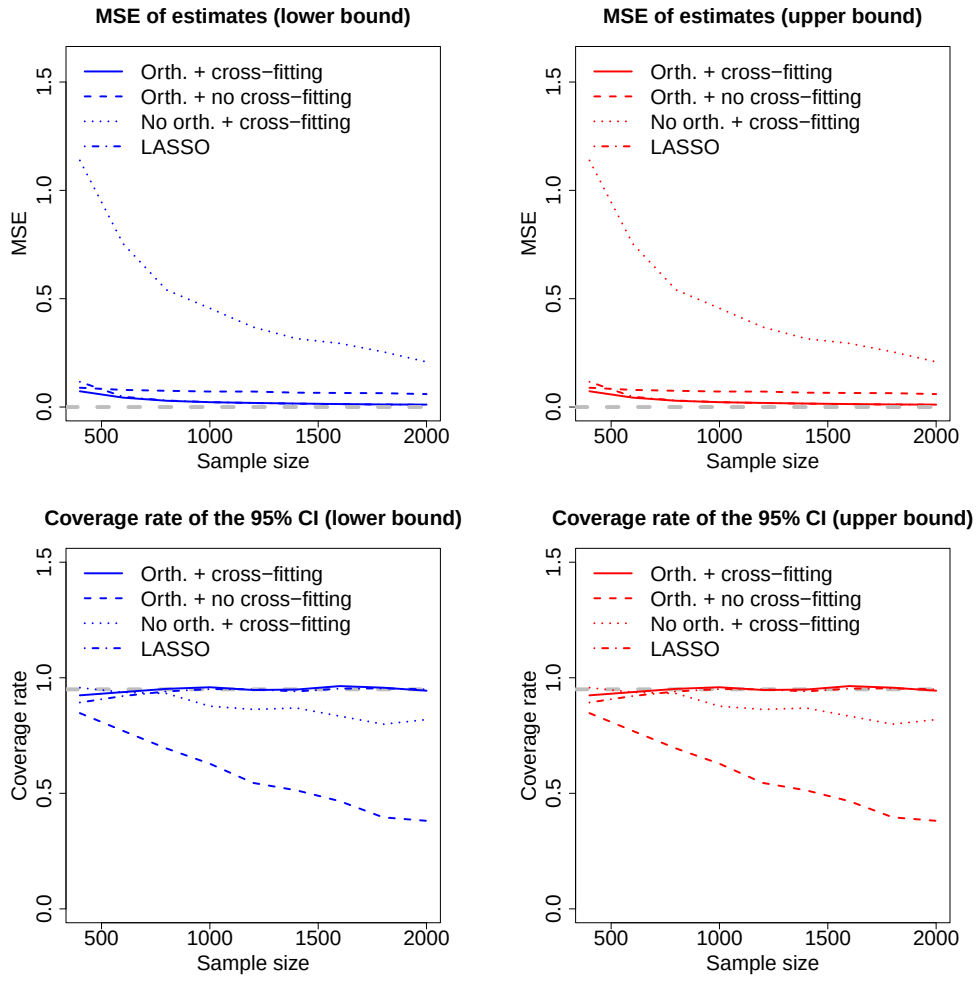


Figure 5: Asymptotic Performance of Methods (Increasing Sample Size)

Note: Plots on the left show how the MSE and coverage rate of the lower bound estimates vary with sample sizes. Plots on the right show the same for the upper bound estimates. The solid lines represent the performance of the proposed method, with both Neyman orthogonalization and cross-fitting. The dashed lines represent the performance of the method with Neyman orthogonalization but without cross-fitting. The dotted lines represent the performance of the method with cross-fitting but without Neyman orthogonalization. The dash-dotted lines represent the performance of the LASSO method proposed by Semenova (2020). The gray lines on the bottom mark the nominal level of coverage, 95%. We can see that the proposed method's MSE declines to zero as the sample size grows, and its coverage rate remains at the level of 95%. It outperforms the LASSO-based approach under small samples, although the difference diminishes gradually. Without either Neyman orthogonalization or cross-fitting, the results are affected by the regularization bias hence do not lead to accurate inference.

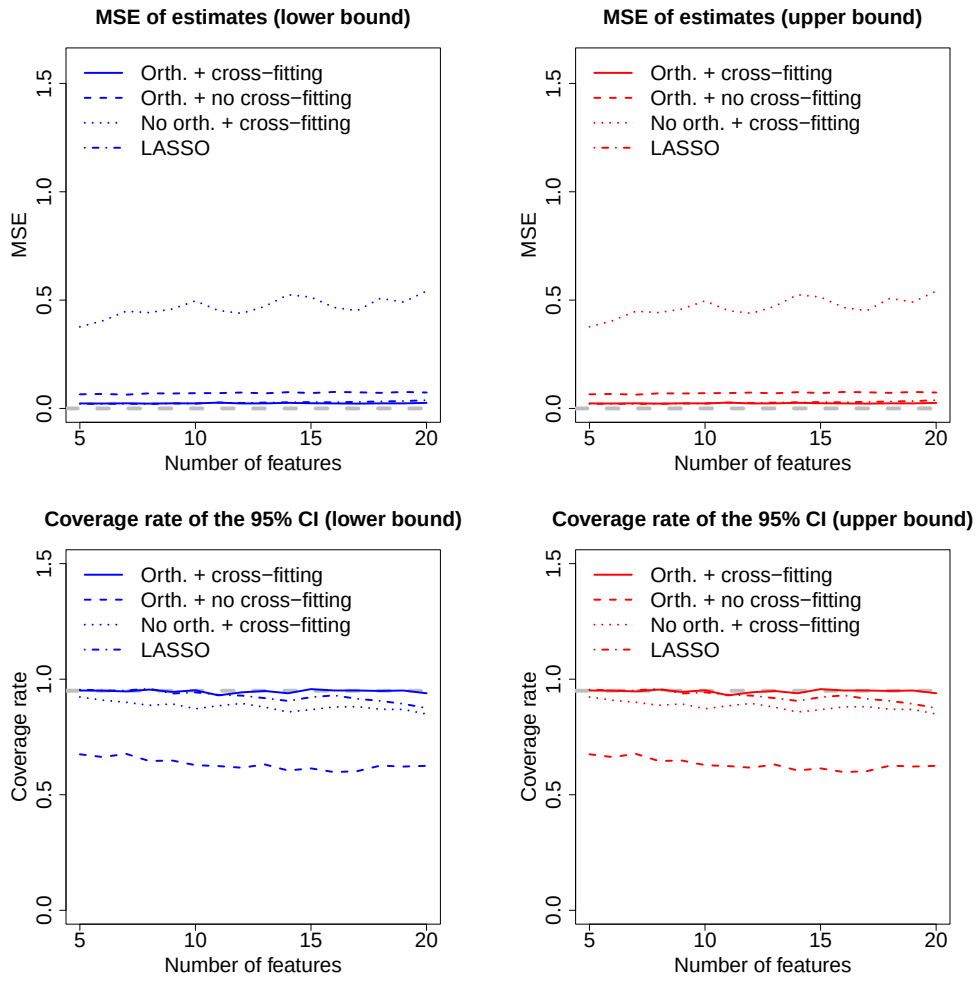
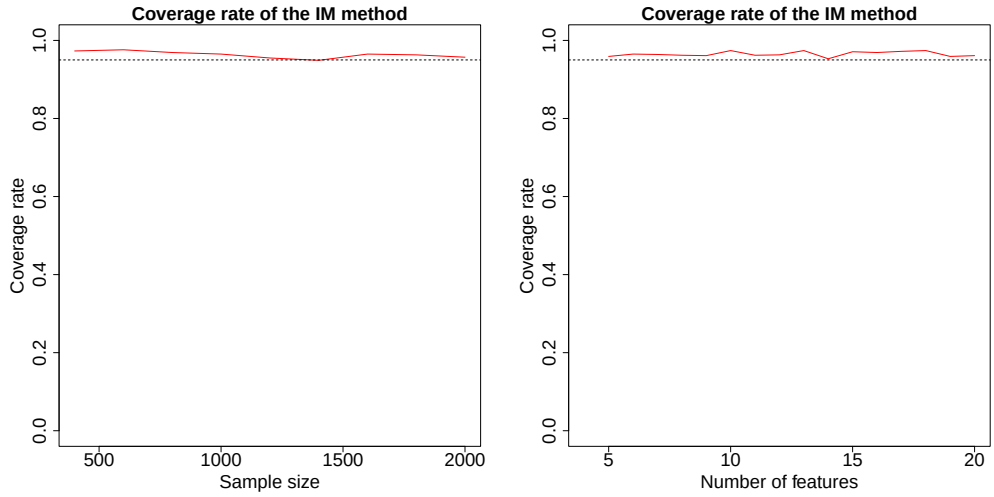


Figure 6: Asymptotic Performance of Methods (Increasing Number of Covariates)

Note: Plots on the left show how the MSE and coverage rate of the lower bound estimates vary with the number of irrelevant covariates. Plots on the right show the same for the upper bound estimates. The solid lines represent the performance of the proposed method, with both Neyman orthogonalization and cross-fitting. The dashed lines represent the performance of the method with Neyman orthogonalization but without cross-fitting. The dotted lines represent the performance of the method with cross-fitting but without Neyman orthogonalization. The dash-dotted lines represent the performance of the LASSO method proposed by Semenova (2020). The gray lines on the bottom mark the nominal level of coverage, 95%. We can see that the proposed method's MSE declines to zero and its coverage rate remains at the level of 95%, even when the number of covariates is large. In contrast, the coverage rate of the LASSO-based approach falls below 95% when there are more than 15 covariates. Without either Neyman orthogonalization or cross-fitting, the results are affected by the regularization bias hence do not lead to accurate inference.

Figure 7: Coverage Rate of the Imbens-Manski Confidence Interval



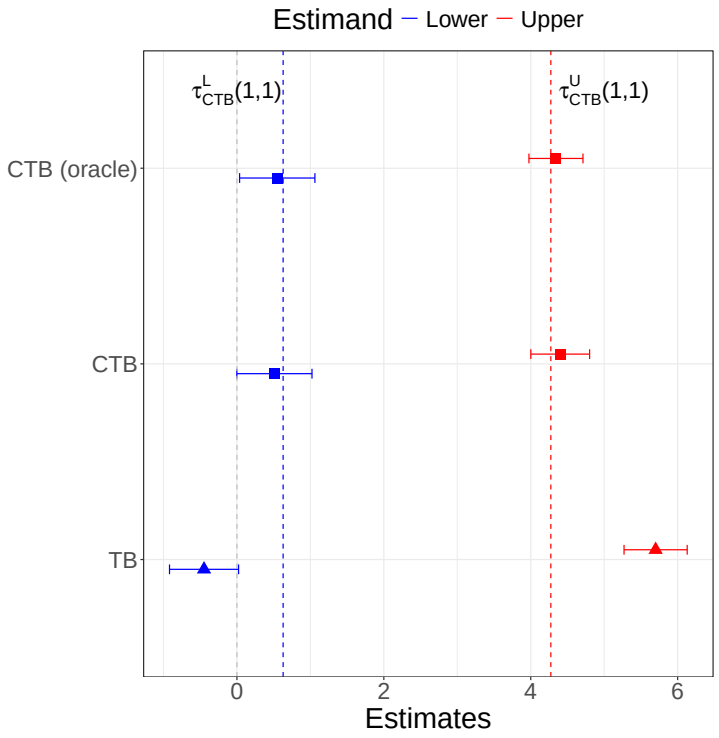
Note: These two plots demonstrate how the coverage rates of confidence intervals based on the Imbens-Manski (IM) method (Imbens and Manski, 2004) vary over both sample size and the number of irrelevant covariates. We can see that they remain above the nominal level of 95%.

Table 1: CTB Running Time (seconds)

N \ P	5	10	20	50	100
500	2.3	2.6	4	5.1	5.9
1000	4.7	7.1	12.1	17.6	19.6
2000	15.2	25.2	45.9	67	77.8
5000	90.5	167.4	318.8	476.2	554.7
10000	393.6	771	1516.6	2314.4	2739.6

Notes: The table above shows running time for the CTB algorithm using the *grf* package. Each row represents the number of observations (N) in the simulation, while each column represents the number of features (P). We set cross-validation to a default value of 5 folds. We also set "regression spliting" to be false, which is the default choice of *grf* and uses specialized splits for quantile estimation. We run the simulation with version 2.3.0 of the *grf* package and version 4.3.1 of R on a Macbook Air laptop with Apple M2 processor and 16 GB RAM.

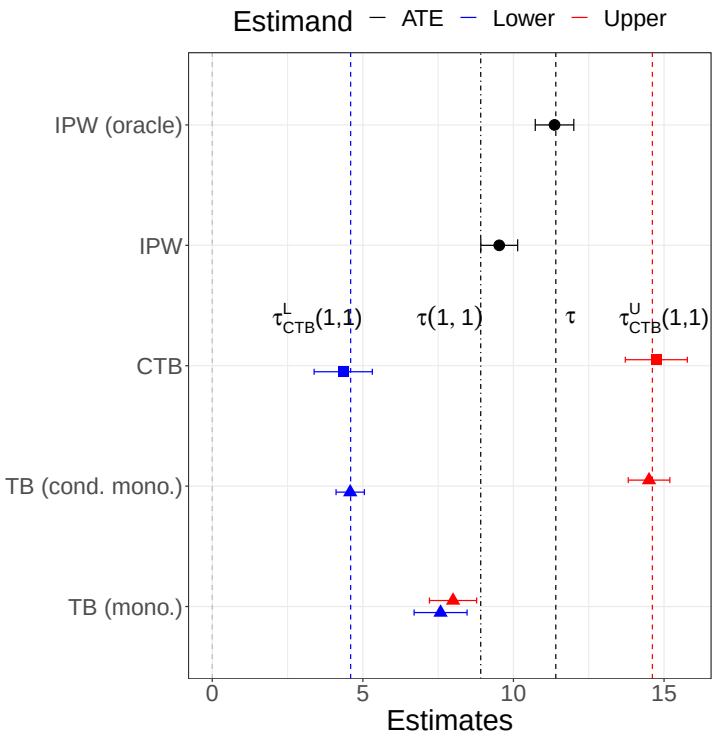
Figure 8: Performance with Unknown Propensity Scores



Note: This plot shows the estimated covariates-tightened trimming bounds (CTB, in squares) and basic trimming bounds (TB, in triangles) in simulation, when the propensity scores have to be estimated from data. It also compares the results with the estimated covariates-tightened trimming bounds under the true propensity scores (CTB (oracle), in squares). Lower bound estimates are in blue and upper bounds estimates are in red. The segments represent the 95% confidence intervals for the estimates. The results are averaged across 1,000 treatment assignments. The red and blue dotted lines mark the true values of the CTB. The black dotted line represents the true ATE and the black dash-dotted line represents the true ATE for the always-responders. We can see that the averages of the CTB estimates are close to their true values even when the propensity scores are unknown.

We demonstrate the utility of the conditionally monotonic selection assumption using a simulation exercise. There is one covariate $X_i \in \{0, 1\}$. When $X_i = 0$, $\tau_i = 5 + 1.8 * U_i + 3 * \sin(-0.7)$, and $S_i = \mathbf{1}\{1 - U_i + 2.2 * D_i * U_i + \nu_i \geq 0\}$. When $X_i = 1$, $\tau_i = 10 + 6 * U_i + 3 * \sin(-0.7)$, and $S_i = \mathbf{1}\{1 + 1.2 * U_i - 2.2 * D_i * U_i + \nu_i \geq 0\}$. $U_i \sim Unif[0, 3]$, $\nu \sim Normal(0, 1)$, and $D_i \sim Bernoulli(0.5)$. Clearly, $S_i(1) \geq S_i(0)$ if $X_i = 0$ and $S_i(1) \leq S_i(0)$ otherwise. Monotonic selection is violated yet conditionally monotonic selection still holds. From Figure 9, we can see that basic trimming bounds that assume monotonic selection do not cover the ATE on the always-responders, while both basic trimming bounds and CTB under the assumption of conditionally monotonic selection generate valid results. We also compare results from methods based on missing at random with results from our approach. When U_i is observable to the researcher, MAR is satisfied, and the IPW estimator (which we denote as IPW (oracle) on Figure 9) results in a consistent estimate of the ATE. But in reality, U_i is unobservable hence MAR does not hold. In this case, the IPW estimator is inconsistent yet our method remains informative.

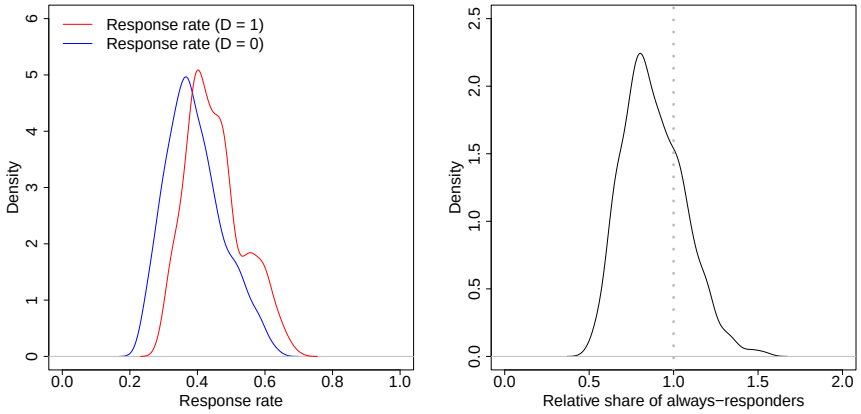
Figure 9: Performance under Conditionally Monotonic Selection



Note: This plot shows the estimated covariates-tightened trimming bounds (CTB, in squares) and basic trimming bounds (TB (cond. mono.), in triangles) under conditionally monotonic selection, as well as basic trimming bounds under monotonic selection (TB (mono.), in triangles), when only conditionally monotonic selection is satisfied. It also compares the results with those from the IPW estimator with and without U_i being observable to the researcher (IPW (oracle) and IPW, in circles). Lower bound estimates are in blue and upper bounds estimates are in red. The segments represent the 95% confidence intervals for the estimates. The results are averaged across 1,000 treatment assignments. The red and blue dotted lines mark the true values of the CTB. The black dotted line represents the true ATE and the black dash-dotted line represents the true ATE for the always-responders.

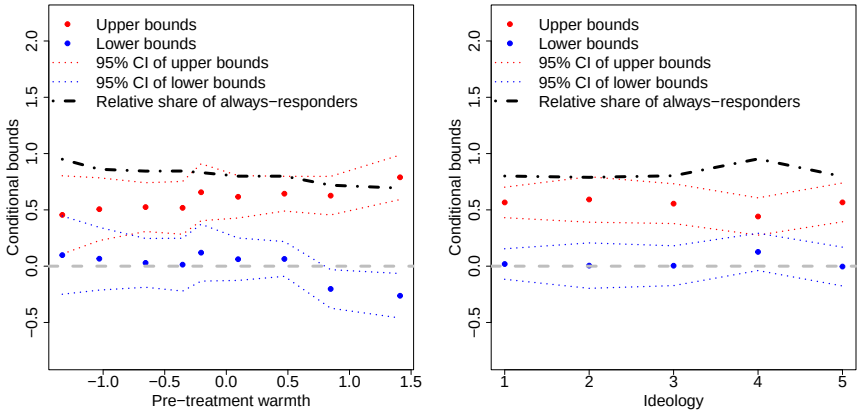
D.2 Extra results from application in main text

Figure 10: Trimming Probability Estimates in Santoro and Broockman (2022)



Note: The left plot presents the distribution of the estimated conditional response rate under treatment ($q_1(\mathbf{x})$, in red) or under control ($q_0(\mathbf{x})$, in blue). The right plot presents the distribution of the estimated conditional trimming probability ($q(\mathbf{x}) = q_0(\mathbf{x})/q_1(\mathbf{x})$).

Figure 11: Conditional Bounds in Santoro and Broockman (2022)



Note: These plots show the estimates of the conditional covariates-tightened trimming bounds across 9 observations whose demographic attributes are fixed at the sample mean or mode while their pre-treatment warmth toward outpartisan voters or ideology varies across its quantiles. The red and blue dots represent upper and lower bound estimate, respectively. The dotted lines around them are the 95% confidence intervals. The dash-dotted curve depicts the conditional trimming probability.

D.3 Replication results of Blattman and Annan (2010) (observational study)

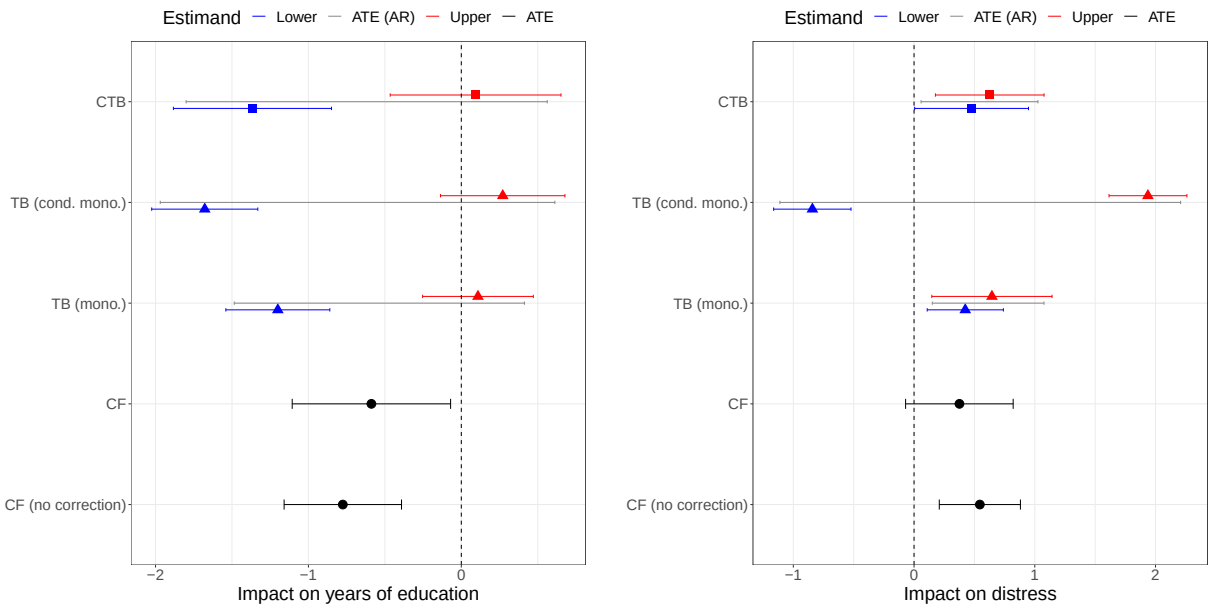
Blattman and Annan (2010) is an observational study investigating the consequences of being abducted into a rebel organization as a youth in Northern Uganda. The authors argue that conditional on personal characteristics, whether an individual was abducted (D_i) was a random event. This motivates an identification strategy based on covariate conditioning. To conduct the study, they constructed a roster of household members in 1996 from 1,100 households across eight rural sub-counties of Uganda. Then, a random sample of 1,216 males born between 1975 and 1991 was drawn from the roster. For each subject in the sample, the authors collaborated with the household head to determine his treatment status—whether they were abducted by the rebels—as well as their relevant demographic characteristics. Among these 1,216 males, 346 had died or not returned from abduction when the study was conducted. The survey enumerators succeeded in tracking 741 males among the remaining 870 and asked them to complete the survey questionnaire. Analysis shows that while abductees exhibited "resilience" on certain dimensions, abduction affected these subjects negatively along various dimensions: they have fewer years of education, worse labor market performance, and a higher level of distress.

Blattman and Annan note that selection rates differed by treatment status: "abductees are half as likely to be unfound migrants, twice as likely to have perished, and comprise all of those who did not return from abduction" (p. 888). The authors applied several techniques to gauge the potential influence of sample selection on their findings, including re-weighting the subjects with the estimated probability of attrition, a sensitivity analysis, and basic trimming bounds. We use our proposed covariate-adjusted bounds to evaluate the robustness of their findings.

We focus on two outcome variables: years of education and the level of psychological distress. We observe years of education for all 870 males that returned from abduction (household heads were able to provide information in case the subject himself was unreachable for interview), but psychological distress is measured only for the 741 males that were reachable for interview. We use the covariates from Blattman and Annan's original analysis that do not contain any missing values, such as the age and location of each male, as well as the wealth level of his household. We estimate the treatment propensity score using the probability forest on the entire sample. We adjust our moment conditions accordingly following the discussion in Section 6 of the main text.

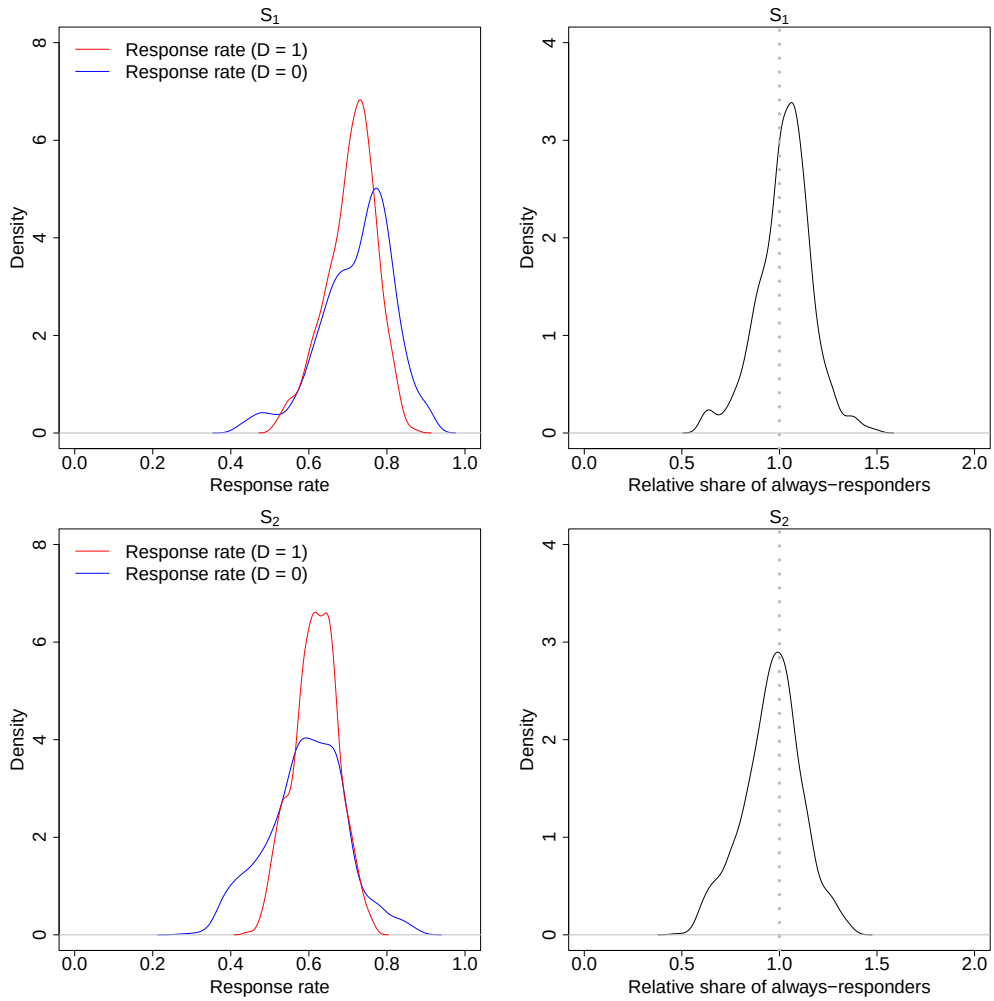
Estimates are presented in Figure 12. At the very bottom are "causal forest" estimates that model outcomes in the treated and control group as a function of covariates using generalized random forest ("CF (no correction)"). Above those are causal forest estimates that are also weighted by the inverse probability of attrition ("CF"). These are more elaborate specifications than in the original paper, which only used linear regression specifications for the covariates. Nonetheless, the findings are similar to what the original paper found. Above the causal forest estimates are the basic trimming bounds, assuming monotonic selection. When we check for monotonicity using the approach discussed in Section 6 of the main text, we find evidence against it. Thus, we also show basic trimming bounds that allow for the direction of the monotonicity to vary by covariate-defined subgroups ("TB (cond. mono.)"). Doing so yields extremely wide bounds. When we implement the full covariate-adjusted bound ("CTB"), which allows for conditional monotonicity, we obtain results that are more informative, and much more so for psychological distress.

Figure 12: Covariate-tightened Trimming Bounds in Blattman and Annan (2010)



Note: In the left plot, the outcome variable is years of education, while in the right one, it is psychological distress. From top to bottom, the plots show the estimated covariates-tightened trimming bounds (CTB, in crosses), basic trimming bounds under monotonicity (TB (mono.), in squares) and under conditional monotonicity (TB (cond. mono.), in triangles), the ATE estimate using causal forest on the non-missing sample weighted by the inverse probability of attrition (CF, in circles), and the same estimate without re-weighting (CF (no correction), in circles). Lower bound estimates are in blue and upper bound estimates are in red. The ATE estimates are in black. The segments represent the 95% confidence intervals for the estimates. Black segments between the red and blue ones represent the 95% confidence intervals for the ATE for always-responders, using the Imbens-Manski approach.

Figure 13: Trimming Probability Estimates in Blattman and Annan (2010)



Note: The left plots present the distribution of the estimated conditional response rate under treatment ($q_1(\mathbf{x})$, in red) or under control ($q_0(\mathbf{x})$, in blue). The right plots present the distribution of the estimated conditional trimming probability ($q(\mathbf{x}) = q_0(\mathbf{x})/q_1(\mathbf{x})$).

Table 2: Summary of Results

	CTB	TB (mono.)	TB (cond. mono.)	OLS	IPW
			Panel A: Simulation (ATE for the always-responders = 3.114)		
$\hat{\tau}^L(1,1)$	0.559	-0.482	-	-	-
	[0.050, 1.068]	[-0.960, -0.005]	-	-	-
$\hat{\tau}^U(1,1)$	4.346	5.724	-	-	-
	[3.985, 4.707]	[5.287, 6.162]	-	-	-
$\hat{\tau}$	-	-	-	2.620	2.606
	-	-	-	[2.135, 3.105]	[2.200, 3.012]
N	693/1,000	693/1,000	693/1,000	693/1,000	693/1,000
			Panel B: Santoro and Broockman (2022)		
$\hat{\tau}^L(1,1)$	0.272	0.022	-0.758	-	-
	[0.060, 0.484]	[-0.213, 0.257]	[-0.996, -0.521]	-	-
$\hat{\tau}^U(1,1)$	0.355	0.466	1.086	-	-
	[0.156, 0.554]	[0.227, 0.706]	[0.913, 1.258]	-	-
$\hat{\tau}$	-	-	-	0.340	-
	-	-	-	[0.218, 0.463]	-
N	469/986	469/986	469/986	469/986	469/986
			Panel C: Blattman and Annan (2010), Education		
$\hat{\tau}^L(1,1)$	-1.366	-1.200	-1.678	-	-
	[-1.883 , -0.849]	[-1.540, -0.860]	[-2.025, -1.331]	-	-
$\hat{\tau}^U(1,1)$	0.093	0.109	0.271	-	-
	[-0.465, 0.652]	[-0.253, 0.471]	[-0.136, 0.678]	-	-
$\hat{\tau}$	-	-	-	-0.799	-0.588
	-	-	-	[-1.229, -0.369]	[-1.106, -0.070]
N	870/1,216	870/1,216	870/1,216	870/1,216	870/1,216
			Panel D: Blattman and Annan (2010), Distress		
$\hat{\tau}^L(1,1)$	0.477	0.424	-0.842	-	-
	[0.005, 0.948]	[0.109, 0.740]	[-1.163, -0.522]	-	-
$\hat{\tau}^U(1,1)$	0.627	0.645	1.937	-	-
	[0.177, 1.077]	[0.147, 1.143]	[1.615, 2.260]	-	-
$\hat{\tau}$	-	-	-	0.376	0.545
	-	-	-	[-0.070, 0.821]	[0.209, 0.881]
N	741/1,216	741/1,216	741/1,216	741/1,216	741/1,216

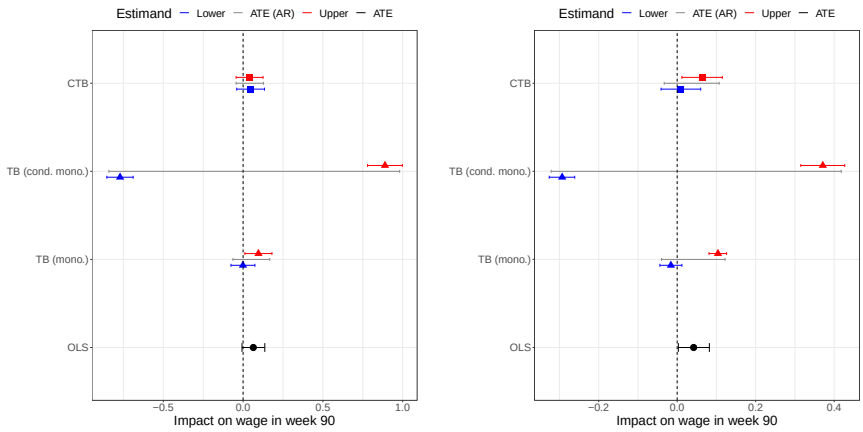
Note: This table summarizes the results from our simulation study (panel A), replication of Santoro and Broockman (2022) (Panel B), and replication of Blattman and Annan (2010) (Panels C and D). Numbers in the brackets are the endpoints of the 95% confidence intervals. The first number in the last row of each panel is the sample size with observed outcome values and the second number is the total sample size.

D.4 Revisiting the Job Corps experiment

We apply our method to estimate the intensive margin effect of job training on wages using data from the Job Corps program, the original example investigated by Lee (2009). It is one of the largest federal-funded job training programs in the U.S., in which disadvantaged youth aged 16-24 years were offered residence at a Job Corps center and about 1100 hours of vocational and academic training. In the mid-1990s, the program conducted lottery-based admission to assess its effect on the basis of randomization. The sample consists of 9,145 Job Corps applicants. Among them, 5,977 applicants in the control group were embargoed from the program for 3 years, while the remaining ones (the treatment group) could enroll in Job Corps as usual. The data includes the lottery outcome, hours worked, and wages for 208 consecutive weeks after randomization of each applicant, as well as their background information, such as educational attainment, employment record, recruiting experience, household composition, income, drug use, and arrest record.

To be consistent with Lee (2009), we focus the intensive margin effect generated by the program, which is defined as the effect on wages among those who would be employed regardless of the access to Job Corps.[1] Such an effect is not identified by randomization alone, since whether applicants are employed may result from other unobservable factors. Lee (2009) constructed bounds for the intensive margin effect on the logarithm of wage in week 208 under the assumption of monotonic selection: treated applicants are more likely to be employed. Yet as pointed out by Semenova (2020), this assumption may not always hold in the data. We replicate Lee's analysis using all the available covariates in the dataset and focus on the wage level in both week 90 and week 208.

Figure 14: Covariate-tightened Trimming Bounds in JobCorps



Note: In the left plot, the outcome variable is weekly wage in week 90, while in the right one it is weekly wage in week 208. From top to bottom, the plots show the estimated covariates-tightened trimming bounds (CTB, in squares), basic trimming bounds under monotonic selection (TB (mono.), in triangles) and under conditionally monotonic selection (TB (cond. mono.), in triangles), and the ATE estimate using OLS on the non-missing sample (OLS, in circles). Lower bound estimates are in blue and upper bounds estimates are in red. The ATE estimates are in black. The segments represent the 95% confidence intervals for the estimates. Black segments between the red and blue ones represent the 95% confidence intervals for the ATE for always-responders, using the Imbens-Manski approach.

[1] This is distinct from the extensive margin effect, which applies to those whose employment status is affected by the program.

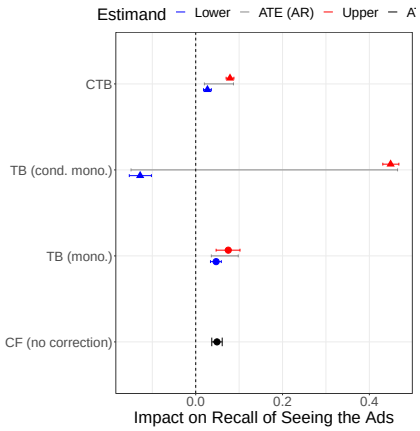
The replication results are presented in Figure 14. Again, we find that the assumption of monotonic selection is violated and has to be replaced by conditionally monotonic selection. It turns out that $\hat{\mathcal{X}}^-$ includes about one third of all the observations. Under conditionally monotonic selection, the basic trimming bounds become much wider and cover zero. However, under the same assumption, our bounds are still informative. They are even narrower than the Lee bounds under monotonic selection. Both the lower bound and the upper bound are positive for wage level in week 90, although the former is not statistically significant at the level of 5%.[2] Overall, the replication confirms the necessity of allowing for conditionally monotonic selection and the superiority of the proposed method to the basic trimming bounds.

[2] Semenova (2020) reports a significantly positive estimate for the lower bound. But her analysis relies on covariates that are unavailable in the original dataset.

D.5 Replication results of Kalla and Broockman (2022) (binary outcome)

To demonstrate how to apply our method when the outcome variable is binary, we replicate the results in Kalla and Broockman (2022), a field experiment that aims to test the impacts of televised issue advertisements on American voters. We focus on one of the treatments in the experiment ("Prosperous Future" and "MariCruz/Kandy"), which consists of two immigration ads, one about how a white woman changed her view on the immigration system and the other showing that undocumented workers also pay taxes. The outcome we are interested in is whether subjects recall the content of the ads at the end of the experiment. In the baseline survey, there were 20,937 subjects in the control group and the treatment group of interest. But the number dropped to 10,650 in the endline survey. We calculate the basic trimming bounds and the aggregated covariate-tightened trimming bounds using the moment conditions presented in Section 4 of the main text. The standard error estimates of the basic trimming bounds are obtained from bootstrap.[3]

Figure 15: Covariate-tightened Trimming Bounds in Kalla and Broockman (2022)



Note: From top to bottom, we have the estimated covariates-tightened trimming bounds (CTB, in squares), basic trimming bounds under monotonicity (TB (mono.), in triangles) and under conditionally monotonic selection (TB (cond. mono.), in triangles), and the ATE estimate using OLS on the non-missing sample (OLS, in circles). Lower bound estimates are in blue and upper bounds estimates are in red. The ATE estimate is in black. The segments represent the 95% confidence intervals for the estimates. Black segments between the red and blue ones represent the 95% confidence intervals for the ATE for always-responders, using the Imbens-Manski approach.

[3]It is straightforward to show that the nuisance parameters, $\nu(\mathbf{x}) = (q_0(\mathbf{x}), q_1(\mathbf{x}), \xi(\mathbf{x}))$, converge to a joint normal distribution, using the same arguments in Lee (2009). Hence, the target parameters are also asymptotically normal, which justifies using bootstrap for inference.

References

Athey, S. and G. Imbens (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences 113*(27), 7353–7360.

Athey, S., J. Tibshirani, and S. Wager (2019). Generalized random forests. *The Annals of Statistics 47*(2), 1148–1178.

Belloni, A., V. Chernozhukov, I. Fernández-Val, and C. Hansen (2017). Program evaluation and causal inference with high-dimensional data. *Econometrica 85*(1), 233–298.

Blattman, C. and J. Annan (2010). The consequences of child soldiering. *The review of economics and statistics 92*(4), 882–898.

Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, and J. Robins (2018). Double/debiased machine learning for treatment and structural parameters. *The Econometrics Journal 21*(1), C1–C68.

Imbens, G. W. and C. F. Manski (2004). Confidence intervals for partially identified parameters. *Econometrica 72*(6), 1845–1857.

Kalla, J. L. and D. E. Broockman (2022). "outside lobbying" over the airwaves: A randomized field experiment on televised issue ads. *American Political Science Review 116*(3), 1126–1132.

Lee, D. S. (2002). Trimming for bounds on treatment effects with missing outcomes. *National Bureau of Economic Research Cambridge, Mass., USA*.

Lee, D. S. (2009). Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *The Review of Economic Studies 76*(3), 1071–1102.

Olma, T. (2020). Nonparametric estimation of truncated conditional expectation functions. Technical report, University of Bonn and University of Mannheim, Germany.

Robins, J. M. and Y. Ritov (1997). Toward a curse of dimensionality appropriate (coda) asymptotic theory for semi-parametric models. *Statistics in medicine 16*(3), 285–319.

Santoro, E. and D. E. Broockman (2022). The promise and pitfalls of cross-partisan conversations for reducing affective polarization: Evidence from randomized experiments. *Science advances 8*(25), eabn5515.

Semenova, V. (2020). Better lee bounds. *arXiv preprint arXiv:2008.12720*.

Stoye, J. (2009). More on confidence intervals for partially identified parameters. *Econometrica 77*(4), 1299–1315.

Wager, S. and S. Athey (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association 113*(523), 1228–1242.