Supporting Information: Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning

Künzel et al.

SI 1. Simulation Studies

In this section, we compare the S-, T-, and X-learners in several simulation studies. We examine prototypical situations where one learner is preferred to the others. In practice, we recommend choosing powerful machine-learning algorithms such as BART (1), Neural Networks, or RFs for the base learners, since such methods perform well for a large variety of data sets. In what follows, we choose all the base learners to be either BART or honest RF algorithms—as implemented in the `hte` R package (2)—and we refer to these meta-learners as S–RF, T–RF, X–RF, S–BART, T–BART, and X–BART, respectively. Using two machine-learning algorithms as base learners helps us to demonstrate that our conclusions about the performance of the different meta learners is often independent of the particular base learner. For example, for all our simulation results we observe that if X–RF outperforms T–RF, then X–BART also outperforms T–BART.

Remark SI 1 (BART and RF)**.** *BART and RF are regression tree-based algorithms that use all observations for each prediction, and they are in that sense global methods. However, BART seems to use global information more seriously than RF, and it performs particularly well when the data-generating process exhibits some global structures (e.g., global sparsity or linearity). RF, on the other hand, is relatively better when the data has some local structure that does not necessarily generalize to the entire space.*

Causal Forests. An estimator closely related to T–RF and S–RF is Causal Forests (CF) (3), because all three of these estimators can be defined as

$$\hat{\tau}(x) = \hat{\mu}(x, w = 1) - \hat{\mu}(x, w = 0),$$

where $\hat{\mu}(x, w)$ is a form of random forest with different constraints on the split on the treatment assignment, W. To be precise, in the S-learner the standard squared error loss function will decide where to split on W, and it can therefore happen anywhere in the tree. In the T-learner the split on W must occur at the very beginning.* For CF the split on W is always made to be the split right before the terminal leaves. To obtain such splits, the splitting criterion has to be changed, and we refer to (3) for a precise explanation of the algorithm. Figure SI 1 shows the differences between these learners for full trees with 16 leaves.

CF is not a meta-learner since the random forests algoirthm has to be changed. However, its similarity to T–RF and S–RF makes it interesting to evaluate its performance. Furthermore, one could conceivably generalize CF to other tree-based learners such as BART. However, this has not been done yet, and we will therefore compare CF in the following simulations to S–, T–, and X–RF.

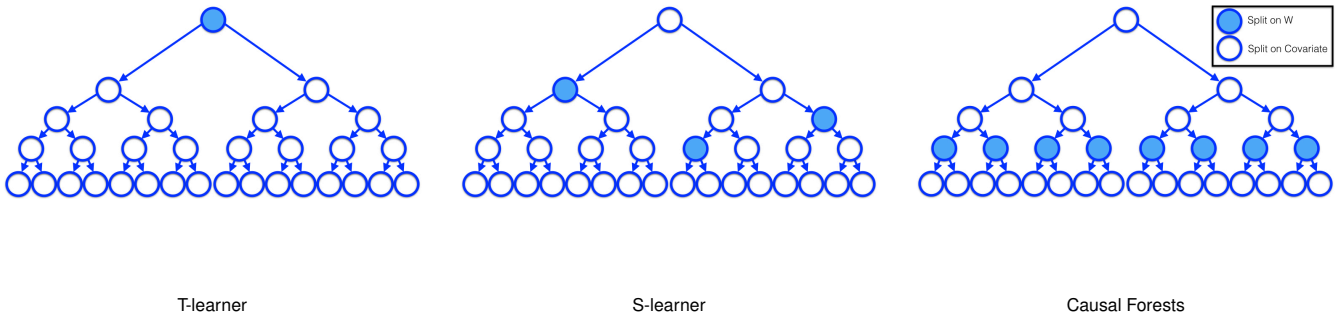

Fig. SI 1. Illustration of the structural form of the trees in T–RF, S–RF, and CF.

Simulation setup. Let us here introduce the general framework of the following simulations. For each simulation, we specify: the propensity score, e; the response functions, μ_0 and μ_1; the dimension, $d \in \mathbb{N}$, of the feature space; and a parameter, α, which specifies the amount of confounding between features. To simulate an observation, i, in the training set, we simulate its feature vector, X_i, its treatment assignment, W_i, and its observed outcome, Y_i, independently in the following way:

*In the original statement of the algorithm we train separate RF estimators for each of the treatment groups, but they are equivalent.

1. First, we simulate a d-dimensional feature vector,

$$X_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma), \qquad \text{[SI 1]}$$

where Σ is a correlation matrix that is created using the `vine` method (4).

2. Next, we create the potential outcomes according to

$$Y_i(1) = \mu_1(X_i) + \varepsilon_i(1),$$
$$Y_i(0) = \mu_0(X_i) + \varepsilon_i(0),$$

where $\varepsilon_i(1), \varepsilon_i(0) \overset{iid}{\sim} \mathcal{N}(0, 1)$ and independent of X_i.

3. Finally, we simulate the treatment assignment according to

$$W_i \sim \text{Bern}(e(X_i)),$$

we set $Y_i = Y(W_i)$, and we obtain (X_i, W_i, Y_i).[†]

We train each CATE estimator on a training set of N units, and we evaluate its performance against a test set of 10^5 units for which we know the true CATE. We repeat each experiment 30 times, and we report the averages.

SI 1.1. The unbalanced case with a simple CATE. We have already seen in Theorem 2 that the X-learner performs particularly well when the treatment group sizes are very unbalanced. We verify this effect as follows. We choose the propensity score to be constant and very small, $e(x) = 0.01$, such that on average only one percent of the units receive treatment. Furthermore, we choose the response functions in such a way that the CATE function is comparatively simple to estimate.

Simulation SI 1 (unbalanced treatment assignment)**.**

$$e(x) = 0.01, \quad d = 20,$$
$$\mu_0(x) = x^T\beta + 5\ \mathbb{I}(x_1 > 0.5), \quad \textit{with} \quad \beta \sim \textit{Unif}\left([-5, 5]^{20}\right),$$
$$\mu_1(x) = \mu_0(x) + 8\ \mathbb{I}(x_2 > 0.1).$$

The CATE function $\tau(x) = 8\ \mathbb{I}(x_2 > 0.1)$ is a one-dimensional indicator function, and thus simpler than the 20-dim function for the response functions $\mu_0(\cdot)$ and $\mu_1(\cdot)$. We can see in Figure SI 2 that the X-learner indeed performs much better in this unbalanced setting with both BART and RF as base learners.

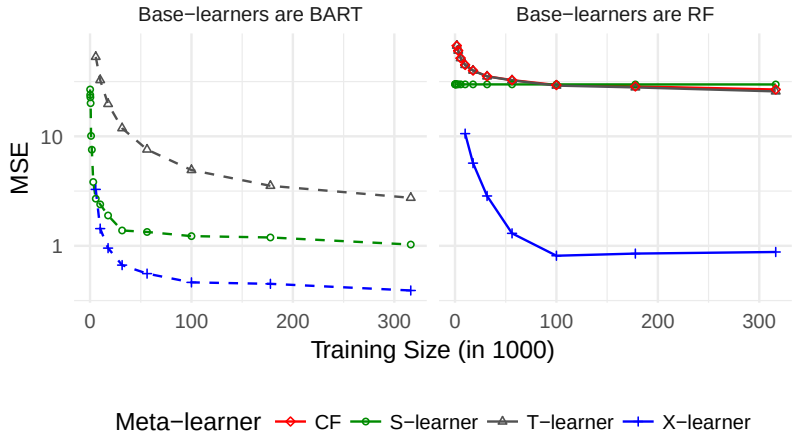

Fig. SI 2. Comparison of S–, T–, and X–BART (left) and S–, T–, and X–RF and CF (right) for Simulation SI 1.

SI 1.2. Balanced cases without confounding. Next, let us analyze two extreme cases: In one of them the CATE function is very complex and in the other one the CATE function is equal to zero. We will show that for the case of no treatment effect, the S-learner performs very well since it sometimes does not split on the treatment indicator at all and it tends to be biased toward zero. On the other hand, for the complex CATE case simulation we have chosen, there is nothing to be learned from the treatment group about the control group and vice versa. Here the T-learner performs very well, while the S-learner is often biased toward zero. Unlike the T-learner, the X-learner pools the data, and it therefore performs well in the simple CATE case. And unlike the S-learner, the X-learner is not biased toward zero. It therefore performs well in both cases.

[†] This is slightly different from the DGP we were considering for our theoretical results, because here m, the number of control units, and n, the number of treated units, are both random. The difference is, however, very small, since in our setups $N = m + n$ is very large.

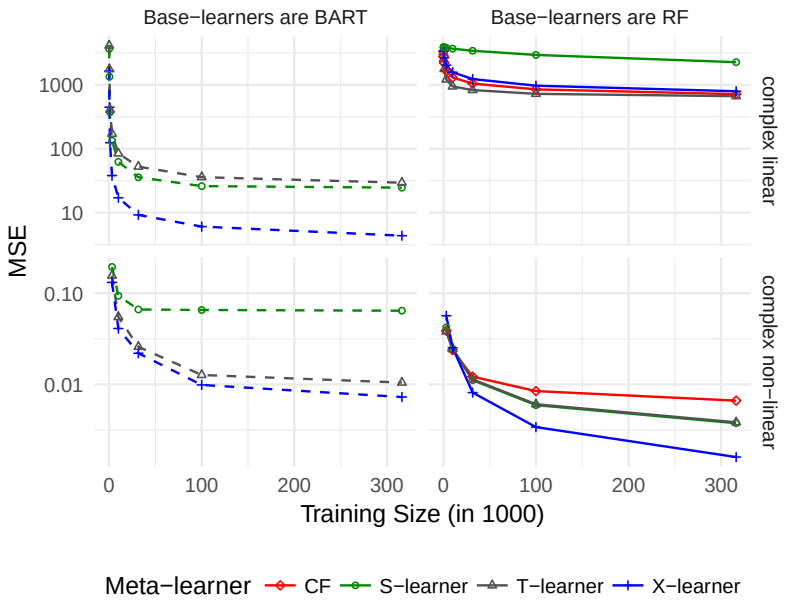

Fig. SI 3. Comparison of the S-, T-, and X-learners with BART (left) and RF (right) as base learners for Simulation SI 2 (top) and Simulation SI 3 (bottom).

SI 1.2.1. Complex CATE. Let us first consider the case where the treatment effect is as complex as the response functions in the sense that it does not satisfy regularity conditions (such as sparsity or linearity) that the response functions do not satisfy. We study two simulations here, and we choose for both the feature dimension to be $d = 20$, and the propensity score to be $e(x) = 0.5$. In the first setup (complex linear) the response functions are different linear functions of the entire feature space.

Simulation SI 2 (complex linear)**.**

$$\begin{aligned} e(x) &= 0.5, \quad d = 20, \\ \mu_1(x) &= x^T\beta_1, \ \textit{with } \beta_1 \sim \textit{Unif}([1,30]^{20}), \\ \mu_0(x) &= x^T\beta_0, \ \textit{with } \beta_0 \sim \textit{Unif}([1,30]^{20}). \end{aligned}$$

The second setup (complex non-linear) is motivated by (3). Here the response function are non-linear functions.

Simulation SI 3 (complex non-linear)**.**

$$\begin{aligned} e(x) &= 0.5, \quad d = 20, \\ \mu_1(x) &= \frac{1}{2}\varsigma(x_1)\varsigma(x_2), \\ \mu_0(x) &= -\frac{1}{2}\varsigma(x_1)\varsigma(x_2) \end{aligned}$$

with

$$\varsigma(x) = \frac{2}{1+e^{-12(x-1/2)}}.$$

Figure SI 3 shows the MSE performance of the different learners. In this case, it is best to separate the CATE estimation problem into the two problems of estimating μ_0 and μ_1 since there is nothing one can learn from the other assignment group. The T-learner follows exactly this strategy and should perform very well. The S-learner, on the other hand, pools the data and needs to learn that the response function for the treatment and the response function for the control group are very different. However, in the simulations we study here, the difference seems to matter only very little.

Another interesting insight is that choosing BART or RF as the base learner can matter a great deal. BART performs very well when the response surfaces satisfy global properties such as being globally linear, as in Simulation SI 2. However, in Simulation SI 3, the response surfaces do not satisfy such global properties. Here the optimal splitting policy differs throughout the space and this non-global behavior is harmful to BART. Thus, choosing RF as the base learners results in a better performance here. Researchers should use their subject knowledge when choosing the right base learner.

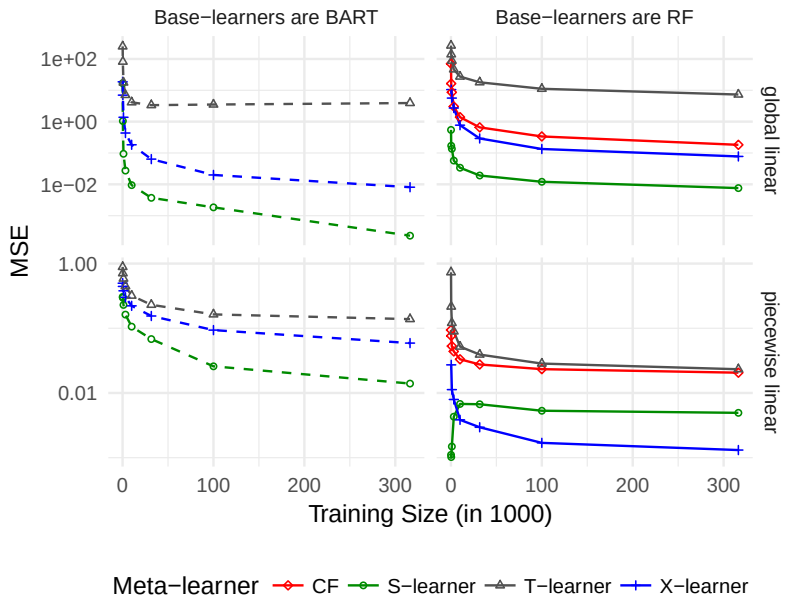

Fig. SI 4. Comparison of S-, T-, and X-learners with BART (left) and RF (right) as base learners for Simulation SI 4 (top) and Simulation SI 5 (bottom).

SI 1.2.2. No treatment effect. Let us now consider the other extreme where we choose the response functions to be equal. This leads to a zero treatment effect, which is very favorable for the S-learner. We will again consider two simulations where the feature dimension is 20, and the propensity score is constant and 0.5.

We start with a global linear model (Simulation SI 4) for both response functions. In Simulation SI 5, we simulate some interaction by slicing the space into three parts, $\{x : x_{20} < -0.4\}$, $\{x : -0.4 < x_{20} < 0.4\}$, and $\{x : 0.4 < x_{20}\}$, where for each of the three parts of the space a different linear response function holds. We do this because we believe that in many data sets there is a local structure, that appears only in some parts of the space.

Simulation SI 4 (global linear)**.**

$$\begin{aligned} e(x) &= 0.5, \quad d = 5, \\ \mu_0(x) &= x^T\beta, \ \textit{with } \beta \sim \textit{Unif}([1, 30]^5), \\ \mu_1(x) &= \mu_0(x). \end{aligned}$$

Simulation SI 5 (piecewise linear)**.**

$$\begin{aligned} e(x) &= 0.5, \quad d = 20, \\ \mu_0(x) &= \begin{cases} x^T\beta_l & \textit{if } \ x_{20} < -0.4 \\ x^T\beta_m & \textit{if } \ -0.4 \leq x_{20} \leq 0.4 \\ x^T\beta_u & \textit{if } \ 0.4 < x_{20}, \end{cases} \\ \mu_1(x) &= \mu_0(x), \end{aligned}$$

with

$$\beta_l(i) = \begin{cases} \beta(i) & \textit{if } \ i \leq 5 \\ 0 & \textit{otherwise} \end{cases} \qquad \beta_m(i) = \begin{cases} \beta(i) & \textit{if } \ 6 \leq i \leq 10 \\ 0 & \textit{otherwise} \end{cases} \qquad \beta_u(i) = \begin{cases} \beta(i) & \textit{if } \ 11 \leq i \leq 15 \\ 0 & \textit{otherwise} \end{cases}$$

and

$$\beta \sim \textit{Unif}([-15, 15]^d).$$

Figure SI 4 shows the outcome of these simulations. For both simulations, the CATE is globally 0. As expected, the S-learner performs very well, since the treatment assignment has no predictive power for the combined response surface. The S-learner thus often ignores the variable encoding the treatment assignment, and the S-learner correctly predicts a zero treatment effect. We can again see that the global property of the BART harms its performance in the piecewise linear case since here the importance of the features is different in different parts of the space.

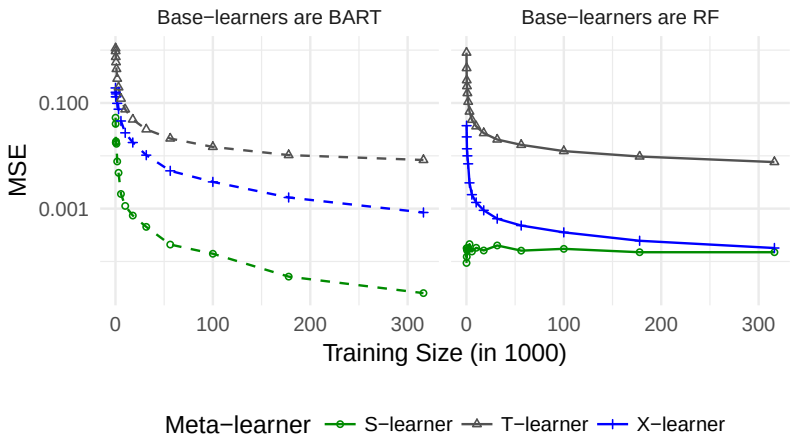

Fig. SI 5. Comparison of S–, T–, and X–BART (left) and S–, T–, and X–RF (right) for Simulation SI 6.

SI 1.3. Confounding. In the preceding examples, the propensity score was globally equal to some constant. This is a special case, and in many observational studies, we cannot assume this to be true. All of the meta-learners we discuss can handle confounding, as long as the ignorability assumption holds. We test this in a setting that has also been studied in (3). For this setting we choose $x \sim Unif([0,1]^{n\times 20})$ and we use the notation that $\beta(x_1, 2, 4)$ is the β distribution with parameters 2 and 4.

Simulation SI 6 (beta confounded)**.**

$$\begin{aligned} e(x) &= \frac{1}{4}(1 + \beta(x_1, 2, 4)), \\ \mu_0(x) &= 2x_1 - 1, \\ \mu_1(x) &= \mu_0(x). \end{aligned}$$

Figure SI 5 shows that none of the algorithms performs significantly worse under confounding. We do not show the performance of causal forests, because—as noted by the authors—it is not designed for observational studies with only conditional unconfoundedness and it would not be fair to compare it here (3).

SI 2. Notes on the ITE

We provide an example that demonstrates that the ITE is not identifiable without further assumptions. Similar arguments and examples have been given before (5), and we list it here only for completeness.

Example SI 1 (D_i is not identifiable)**.** *Assume that we observe a one-dimensional and uniformly distributed feature between 0 and 1, $X \sim Unif([0,1])$, a treatment assignment that is independent of the feature and Bernoulli distributed, $W \sim Bern(0.5)$, and a Rademacher-distributed outcome under control that is independent of the features and the treatment assignment,*

$$P(Y(0) = 1) = P(Y(0) = -1) = 0.5.$$

Now consider two Data-Generating Processes (DGP) identified by the distribution of the outcomes under treatment:

1. *In the first DGP, the outcome under treatment is equal to the outcome under control:*

$$Y(1) = Y(0).$$

2. *In the second DGP, the outcome under treatment is the negative of the outcome under control:*

$$Y(1) = -Y(0).$$

Note that the observed data, $\mathcal{D} = (Y_j, X_j, W_j)_{1\leq j\leq N}$, has the same distribution for both DGPs, but $D_i = 0$ for all i in DGP 1, and $D_i \in \{-2, 2\}$ for all i in DGP 2. Thus, no estimator based on the observed data $\mathcal{D}$ can be consistent for the ITEs, $(D_i)_{1\leq i\leq n}$. The CATE, $\tau(X_i)$, is, however, equal to 0 in both DGPs. $\hat{\tau} \equiv 0$, for example, is a consistent estimator for the CATE.

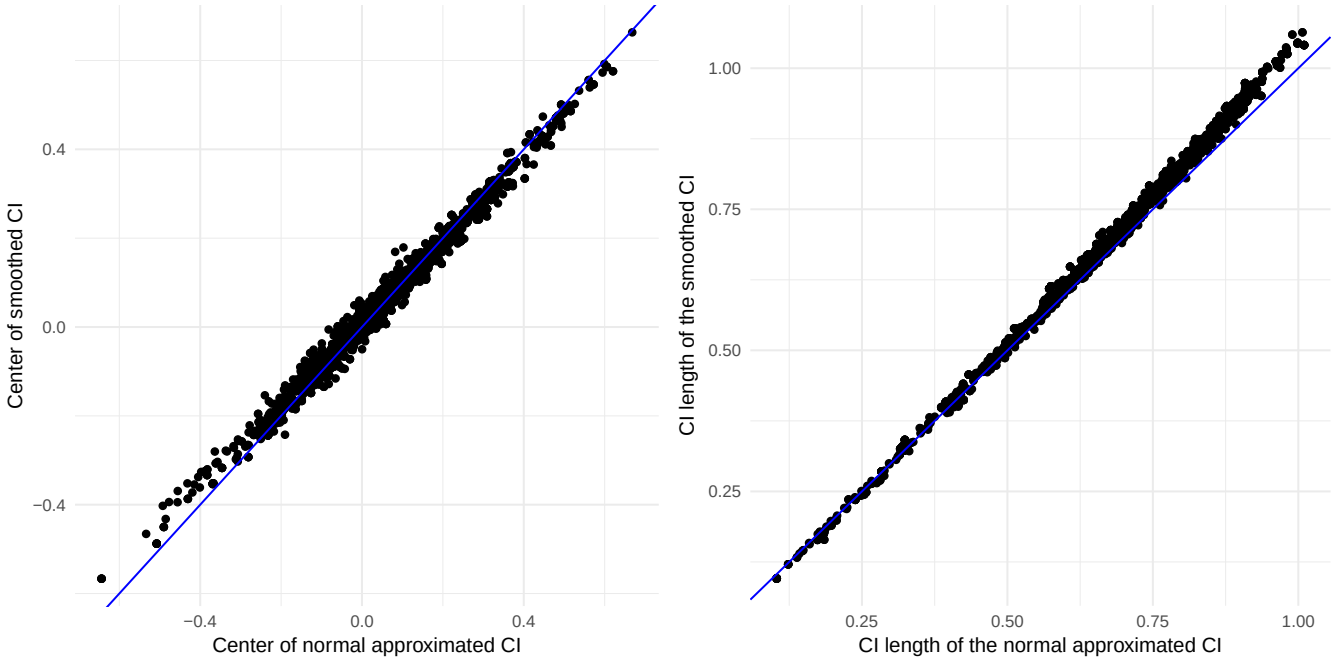

Fig. SI 6. Comparison of normal approximated CI (Algorithm SI 6) and smoothed CI (Algorithm SI 7). The blue line is the identity function.

SI 3. Confidence Intervals for the Social Pressure Analysis

In this paper, we study general meta-learners without making any parametric assumptions on the CATE. This generality makes it very difficult to provide confidence intervals with formal guarantees. In the GOTV section of the main paper, we used bootstrap confidence intervals; in this section, we explain why we choose the bootstrap and details of the variant of the bootstrap, we selected.

The bootstrap has been proven to perform well in many situations (6) and it is straightforward to apply to any estimator that can be written as a function of iid data. There are, however, many ways to obtain bootstrap confidence intervals. We have decided to use Algorithm SI 6, because it performed well for X–RF in the Atlantic Causal Inference Conference (ACIC) challenge (7), where one of the goals was to create confidence intervals for a wide variety of CATE estimation problems. We refer to these confidence intervals as normal approximated CIs.

It was seen in the ACIC challenge that constructing confidence intervals for the CATE that achieve their nominal coverage is extremely difficult, and no method always provides the correct coverage. To argue that the conclusions we draw in this paper are not specific to a single bootstrap method, we implement another version of the bootstrap to estimate confidence intervals due to (8) and (9). We refer to it as the smoothed bootstrap, and we call the corresponding confidence intervals smoothed CIs. Pseudocode for this method can be found in Algorithm SI 7.

There are many other versions of the bootstrap that could have been chosen, but we focus on two that performed well in the ACIC challenge. To compare these methods, we use the GOTV data, and we estimate confidence intervals for $2,000$ test points based on $50,000$ training points. We have to use this much smaller subset of the data for computational reasons.

For both methods, we use $B = 10,000$ bootstrap samples. This is a large number of replications, but it is necessary because the smoothed CIs (Algorithm SI 7) are unstable for a smaller B. Figure SI 6 compares the center and the length of the confidence intervals of the two methods for T–RF. We can see that the two methods lead to almost the same confidence intervals. The normal approximated CIs are slightly larger, but the difference is not substantial. This is not surprising given the size of the data, and it confirms that our analysis of the GOTV data would have come to the same conclusion had we used smoothed CIs (Algorithm SI 7). However, normal approximated CIs (Algorithm SI 6) are computationally much less expensive and they are therefore our default method.

SI 3.1. CI-Simulation 1: Comparison of the coverage of the CI estimation methods. To analyze the coverage of the different bootstrap methods, we use a simulation study informed by the GOTV data. We generate the data in the following way:

CI-Simulation 1.

1. We start by training the T-learner with random forests on the entire GOTV data set to receive CATE estimates. We take this estimate as the ground truth and call it $\tau(x)$.

2. We then compute for each unit i the missing potential outcome. That is, for a unit in the control group, we add $\tau(x_i)$ to the observed outcome to obtain the outcome under treatment, and for each unit in the treatment group, we subtract

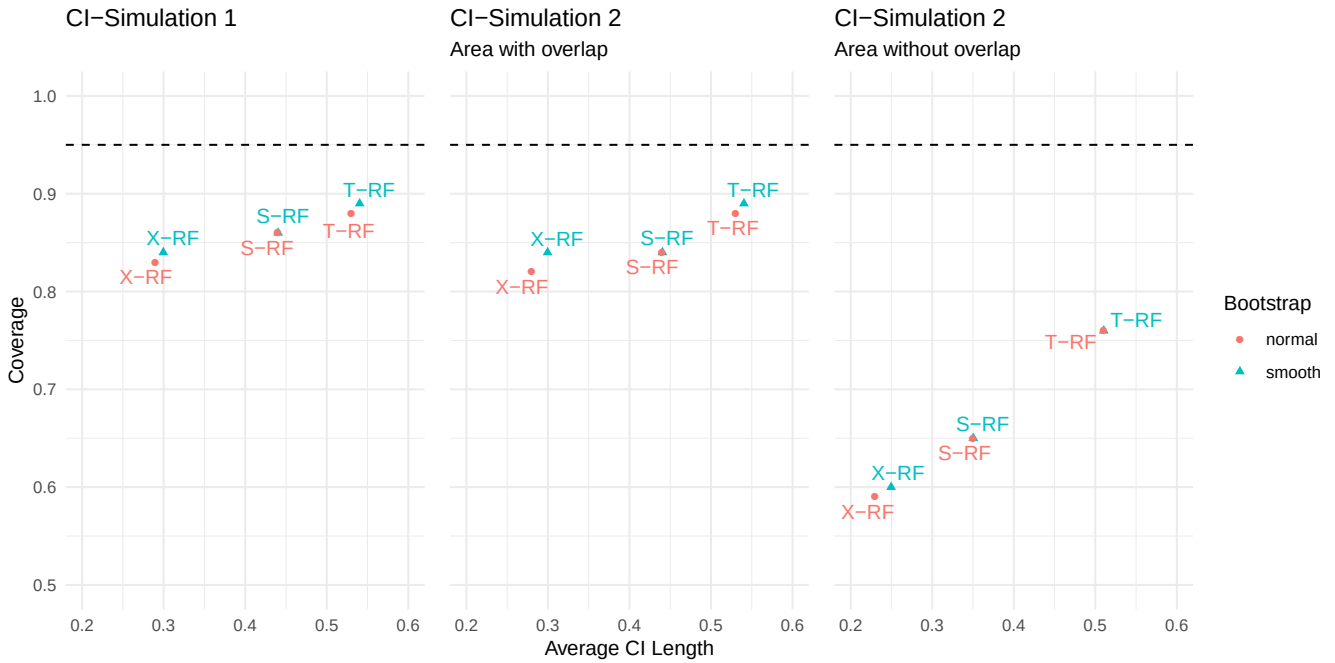

Fig. SI 7. Coverage and average confidence interval length of the three meta-learners for normal approximated CI (Algorithm SI 6) and smoothed CI (Algorithm SI 7). The left figure corresponds to Simulation 3.1; the middle figure corresponds to units in an area with overlap in Simulation SI 3.2, and the right figure corresponds to units in an area without overlap in Simulation SI 3.2. The dotted line corresponds to the target 95% confidence interval.

$\tau(x_i)$ from the observed outcome to obtain the outcome under control.

3. Next, we create a new treatment assignment by permuting the original one. This also determines our new observed outcome.

4. Finally, we sample uniformly and without replacement a test set of $2,000$ observations and a training set of $50,000$ observations.

We then compute 95% confidence intervals for each point in the test set using the the normal and smoothed bootstrap combined with the S, T, and X-learner. The left part of Figure SI 7 shows a comparison of the six methods. We find that none of the methods provide the correct coverage. The coverage of the smooth bootstrap intervals is slightly higher than the coverage of the normal approximated confidence intervals, but the difference is within 1%. It also appears that the T-learner provides the best coverage, but it also has the largest confidence interval length.

Based on this simulation, we believe that the smooth CIs have a slightly higher coverage but the intervals are also slightly longer. However, the smooth CIs are computationally much more expensive and need a lot of bootstrap samples to be stable. They are therefore unfeasible for our data. Hence we prefer the normal approximated CIs.

In general, we observe that none of the methods achieve the anticipated 95% coverage and we suspect that this is the case, because the CATE estimators are biased and the bootstrap is not adjusting for the bias terms. To analyze this, we approximated the bias using a Monte Carlos simulation for each of the 2,000 test points using Algorithm SI 8. The density plot in Figure SI 8 shows that the bias of X–RF in our sample is substantial and in particular of the same order as the size of the confidence intervals of X–RF. For example, more than 11% of all units had bias bigger than 0.15.

This raises the question whether it is possible to correct for the bias. We tried to use the bootstrap again to estimate the bias. Specifically, we used Algorithm SI 9 to estimate it. The upper subfigure in Figure SI 8 is a scatter plot of the Monte-Carlo-approximated bias versus the bootstrap-estimated bias. We can see that the bootstrap does not correctly estimate the bias.

SI 3.2. CI-Simulation 2: Confounding without overlap. In observational studies, researchers have no control over the treatment assignment process and, in some cases, even the overlap condition may be violated. That is, there exists a subgroup of units that is identifiable by observed features for which the propensity score is 0 or 1. Consequently, all units are either treated or not and estimating the CATE is impossible without very strong assumptions. We generally advise researchers to be very cautious when using these methods on observational data. In this section, we want to study how well one can estimate confidence intervals in observational studies where the overlap condition is violated. Ideally, we would hope that the confidence intervals in areas with no overlap are extremely wide.

To test the behavior of the different confidence interval estimation methods, we set up another simulation based on real data. In this simulation we intentionally violate the overlap condition by assigning all units between 30 and 40 years to the

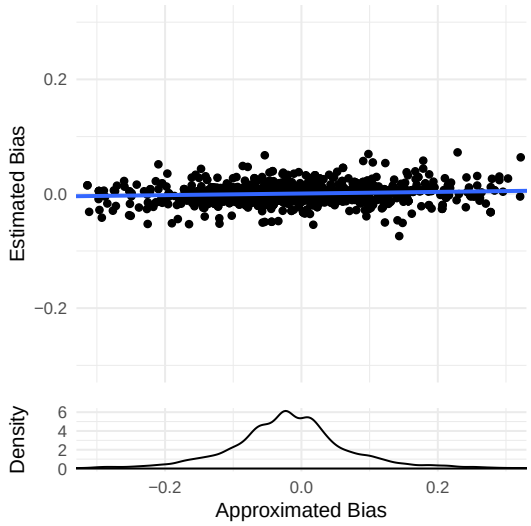

Fig. SI 8. Approximated bias using Algorithm SI 8 versus estimated bias using Algorithm SI 9 and X–RF.

control group. We then compared the confidence intervals for this subgroup with the other units where the overlap condition is not violated. For our simulation, we follow the same steps as in Section SI 3.1, but we modified Step 3 to ensure that all units between 30 and 40 years of age are in the control group. Specifically, we construct the data in the following way:

1. We start by training the T-learner with random forests on the entire GOTV data set to construct CATE estimates. We take this estimate as the ground truth and call it $\tau(x)$.

2. We then use $\tau(x)$ to impute the missing potential outcomes. That is, for a unit in the control group, we add $\tau(x_i)$ to the observed outcome to obtain the outcome under treatment, and for each unit in the treatment group, we subtract $\tau(x_i)$ from the observed outcome to obtain the outcome under control.

3. Next, we create a new treatment assignment by permuting the original treatment assignment vector and assigning all entries for units between 30 and 40 years old to the control group. This also determines our new observed outcome.

4. Finally, we sample uniformly and without replacement two test sets and one training set. We first sample the training set of $50,000$ observations. Next, we sample the first test set of $20,000$ units out of all units that are not in the 30 to 40-year-old age group. This test set is called the **overlap test set**. Finally, we sample the second test set of $20,000$ units out of all units in the 30 to 40-year-old age group and we call this test set the **non-overlap test set**.

Note that by construction the overlap condition is violated for the subgroup of units between 30 and 40 years and satisfied for units outside of that age group.

We trained each method on the training set and estimated the confidence intervals for the CATE in both test sets. The middle and the right part of Figure SI 7 shows the results for the overlap test set and the non-overlap test set, respectively. We find that the coverage and the average confidence interval length for the overlap test set is very similar to that of the previous simulation study, CI-Simulation 1. This is not surprising, because the two setups are very similar and the overlap condition is satisfied in both.

The coverage and the average length of the confidence intervals for the non-overlap test set are, however, very different. For this subgroup, we do not have overlap. We should be cautious when estimating the CATE or confidence intervals of the CATE when there is no overlap, and we hope to see this reflected in very wide confidence intervals. Unfortunately, this is not the case. We observe that for all methods the confidence intervals are tighter and the coverage is much lower than on the data where we have overlap because they try to extrapolate into regions of the covariate space without information on the treatment group. This is a problematic finding and suggests that confidence interval estimation in observational data is extremely difficult and that a violation of the overlap condition can lead to invalid inferences. We believe that this is an artifact of how random forests deal with the predictions for units outside of the support of the training data. We are currently working on an improved version of random forests that better captures this uncertainty.

SI 4. Stability of the Social Pressure Analysis across Meta-learners

In Figure 2, we present how the CATE varies with the observed covariates. We find a very interesting behavior in the fact that the largest treatment effect can be observed for potential voters who voted three or four times before the 2004 general election. The treatment effect for potential voters who voted in none or all five of the observed elections was much smaller. We concluded this based on the output of the X-learner. To show that a similar conclusion can be drawn using different meta-learners, we

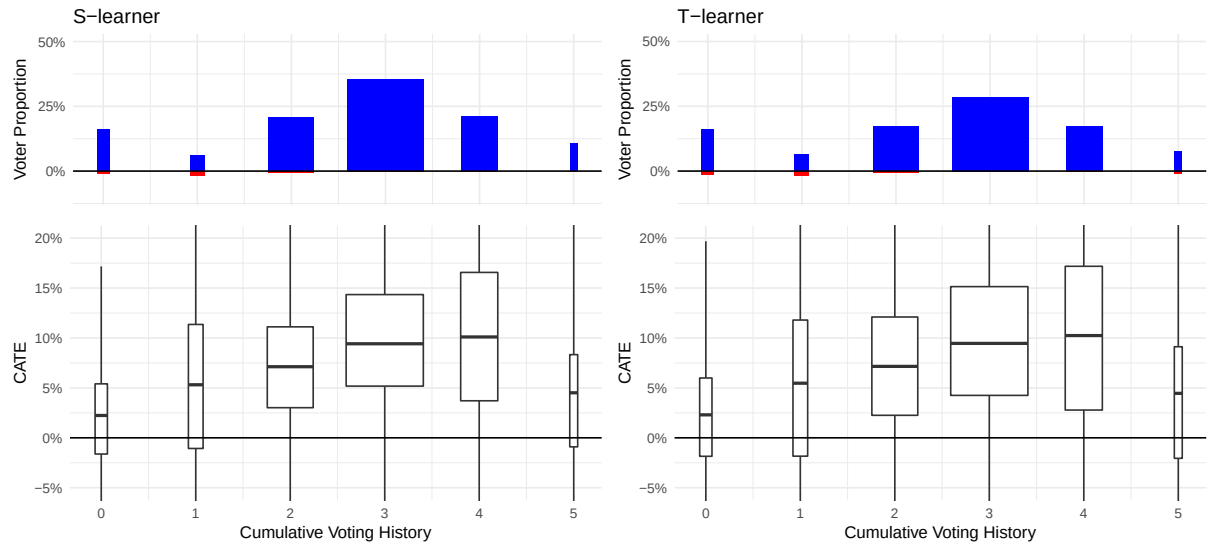

Fig. SI 9. Results for the S-learner (left) and the T-learner (right) for the get-out-the-vote experiment.

repeated our analysis with the S and T learner (cf. Figure SI 9). We find that the output is almost identical to the output of the X-learner. This is not surprising since the data set is very large and most of the covariates are discrete.

SI 5. The Bias of the S-learner in the Reducing Transphobia Study

For many base learners, the S-learner can completely ignore the treatment assignment and thus predict a 0 treatment effect. This often leads to a bias toward 0, as we can see in Figure 4. To further analyze this behavior, we trained a random forest estimator on the transphobia data set with 100,000 trees, and we explored how often the individual trees predict a 0 treatment effect by not splitting on the treatment assignment. Figure SI 10 shows that the trees very rarely split on the treatment assignment. This is not surprising for this data set since the covariates are very predictive of the control response function and the treatment assignment is a relatively weak predictor.

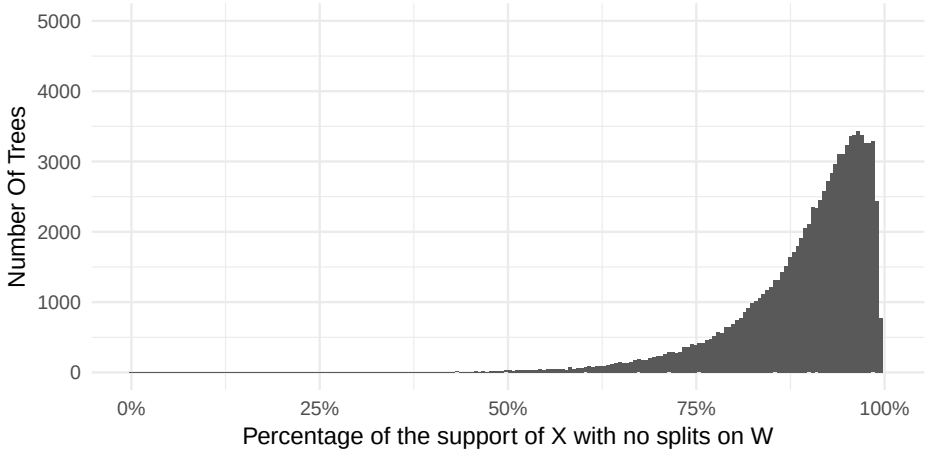

Fig. SI 10. This figure is created from an S–RF learner to show that the S-learner often ignores the treatment effect entirely. It is based on 100,000 trees and it shows the histogram of trees by what percentage of the support of X is not split on W.

SI 6. Adaptivity to Different Settings and Tuning

Tuning the base learners to receive better CATE estimators or even selecting the best CATE estimator from a finite set of CATE estimators is very difficult, and our recent R package, `hte`, attempts to implement some tuning and selection methods. This is, however, very difficult and in the preceding sections, we did not tune our random forest algorithm or our BART estimators on the given data sets. Instead, we used fixed hyperparameters that were chosen in a different simulation study. In the sequel, we show that tuning the base learners and being able to select the best meta-learner can be very beneficial to constructing a good CATE estimator.

We conduct a simple experiment showing the potential benefits of hyperparameter tuning of the base learners. Specifically, we evaluate S–RF, T–RF, and X–RF in Simulations SI 4 and SI 2. We sample 1,000 hyperparameter settings for each of the learners and evaluate them in both simulations. In other words, for each hyperparameter setting, we obtain an MSE for Simulation SI 4 and an MSE for Simulation SI 2.

Figure SI 11 shows the MSE pairs. As expected, we observe that the T-learner generally does very well when the treatment effect is complex, while it does rather poorly when the treatment effect is simple. This was expected as the T-learner generally performs poorly compared to the S-learner when the treatment effect is simple or close to 0. Also as expected, the S-learner performs well when the treatment effect is simple, but it performs relatively poorly compared to the T-learner when the treatment effect is complex. The X-learner, on the other hand, is extremely adaptive. In fact, depending on the set of hyperparameters, the X-learner can perform as well as the T-learner or the S-learner. However, there is not a single set of parameters that is optimal for both settings. In fact, the optimal settings almost describe a utility curve.

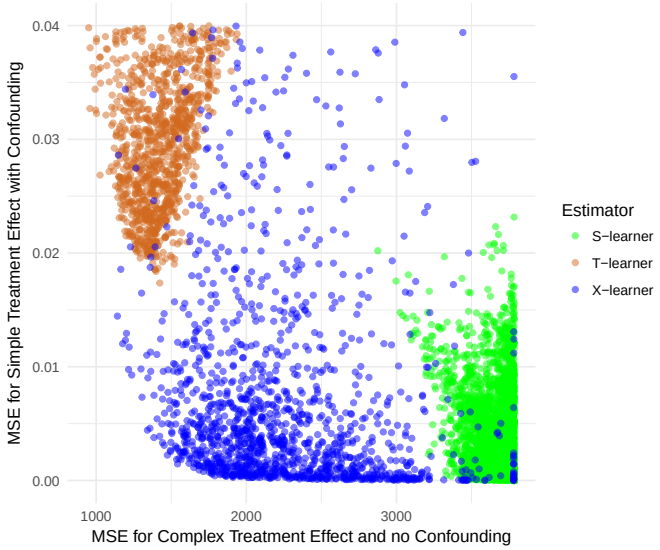

Fig. SI 11. Each point corresponds to a different hyperparameter setting in random forests as the base learner in one of the S-, T-, or X-learners. The y-axis value is the MSE of Simulation SI 4 and the x-axis value is the MSE in Simulation SI 2. A perfect estimator that gets an MSE error of 0 in both simulations would thus correspond to a point at the origin (0,0). The training set size had 1,000 units and the test set that was used to estimate the MSE had 10,000 units.

SI 6.1. Setting the Tuning Parameters. Since tuning each algorithm for each data set separately turns out to be very challenging, we decided to hold the hyperparameters fix for each algorithm. To chose those preset hyperparameters, we used the 2016 Atlantic Causal Inference Conference competition (7), and we chose the parameters in such a way that the algorithms perform very well in this competition. Specifically, we randomly generated for each algorithm 10,000 hyperparameters. We then evaluated the performance of these 10,000 hyperparameter settings on the 20 data sets of the "Do it yourself!"-challenge, and we chose the hyperparameter combination which did best for that challenge.

SI 7. Conditioning on the Number of Treated Units

In our theoretical analysis, we assume a superpopulation and we condition on the number of treated units both to avoid the problem that with a small but non-zero probability all units are in the treatment group or the control group and to be able to state the performance of different estimators in terms of n, the number of treated units, and m, the number of control units. This conditioning, however, leads to nonindependent samples. The crucial step in dealing with this dependent structure is to condition on the treatment assignment, W.

Specifically, there are three models to be considered.

1. The first one is defined by [1]. It specifies a distribution, $\mathcal{P}$, of (X, W, Y), and we assume to observe N independent samples from this distribution,
$$(X_i, W_i, Y_i)_{i=1}^N \overset{iid}{\sim} \mathcal{P}.$$
We denote the joint distribution of $(X_i, W_i, Y_i)_{i=1}^N$ by $\mathcal{P}^N$.

2. We state our technical results in terms of a conditional distribution. For a fixed n with $0 < n < N$, we consider the distribution of $(X_i, W_i, Y_i)_{i=1}^N$ given that we observe n treated units and $m = N - n$ control units. We denote this distribution by $\mathcal{P}^{nm}$.
$$\left[(X_i, W_i, Y_i)_{i=1}^N \,\middle|\, \sum_{i=1}^N W_i = n \right] \sim \mathcal{P}^{nm}.$$

Note that under $\mathcal{P}^{nm}$ the (X_i, W_i, Y_i) are identical in distribution, but not independent.

3. For technical reasons, we also introduce a third distribution, which we will use only in some of the proofs. Here, we condition on the vector of treatment assignments, W.

$$\left[(X_i, W_i, Y_i)_{i=1}^N \middle| W = w\right] \sim \mathcal{P}^w.$$

Under this distribution W is non-random and (X_i, Y_i) are not identical in distribution. However, within each treatment group the (X_i, Y_i) tuples are independent and identical in distribution. To make this more precise, define $\mathcal{P}_1$ to be the conditional distribution of (X, Y) given $W = 1$; then, under $\mathcal{P}^w$, we have

$$(X_i, Y_i)_{W_i=1} \overset{iid}{\sim} \mathcal{P}_1.$$

We prove these facts as follows.

Theorem SI 1. *Let n and N be such that $0 < n < N$ and let $w \in \{0,1\}^N$ with $\sum_{i=1}^N w_i = n$. Then, under the distribution $\mathcal{P}^w$,*

$$(X_k, Y_k)_{W_k=1} \overset{iid}{\sim} \mathcal{P}_1.$$

We prove this in two steps. In Lemma SI 1, we prove that the distributions are independent and in Lemma SI 2 we prove that they are identical.

Lemma SI 1 (independence)**.** *Let n, N, and w be as in Theorem SI 1 and define $S = \{j \in \mathbb{N} : w_j = 1\}$. Then for all $\emptyset \neq \mathcal{I} \subset S$, and all $(B_i)_{i\in\mathcal{I}}$ with $B_i \subset \mathbb{R}^p \times \mathbb{R}$,*

$$\mathcal{P}\left(\bigcap_{i\in\mathcal{I}}\{(X_i, Y_i) \in B_i\}\middle| W = w\right) = \prod_{i\in\mathcal{I}} \mathcal{P}\left((X_i, Y_i) \in B_i \middle| W = w\right). \tag{SI 2}$$

Note that another way of writing [SI 2] is

$$\mathcal{P}^w\left(\bigcap_{i\in\mathcal{I}}\{(X_i, Y_i) \in B_i\}\right) = \prod_{i\in\mathcal{I}} \mathcal{P}^w\left((X_i, Y_i) \in B_i\right). \tag{SI 3}$$

Proof of Lemma SI 1.

$$\begin{aligned}
&\mathcal{P}\left(\bigcap_{i\in\mathcal{I}}\{(X_i, Y_i) \in B_i\}\middle| W = w\right) \\
&= \mathcal{P}\left(\left(\bigcap_{i\in\mathcal{I}}\{(X_i, Y_i) \in B_i\}\right) \cap \left(\bigcap_{j\in S}\{W_j = 1\} \cap \bigcap_{k\in S^c}\{W_k = 0\}\right)\right) / \mathcal{P}\Big(W = w\Big) \\
&= \mathcal{P}\left(\left(\bigcap_{i\in\mathcal{I}}\big\{(X_i, Y_i, W_i) \in B_i \times \{1\}\big\}\right) \cap \left(\bigcap_{j\in S\setminus\mathcal{I}}\{W_j = 1\} \cap \bigcap_{k\in S^c}\{W_k = 0\}\right)\right) / \mathcal{P}\Big(W = w\Big) \\
&= \prod_{i\in\mathcal{I}} \mathcal{P}\Big((X_i, Y_i, W_i) \in B_i \times \{1\}\Big) \frac{\mathcal{P}\Big(\bigcap_{j\in S\setminus\mathcal{I}}\{W_j = 1\} \cap \bigcap_{k\in S^c}\{W_k = 0\}\Big)}{\mathcal{P}\Big(W = w\Big)} = (*).
\end{aligned}$$

The last equality holds because $(X_i, Y_i, W_i)_{i=1}^N$ are mutually independent. The second term can be rewritten in the following way:

$$\begin{aligned}
\frac{\mathcal{P}\Big(\bigcap_{j\in S\setminus\mathcal{I}}\{W_j = 1\} \cap \bigcap_{k\in S^c}\{W_k = 0\}\Big)}{\mathcal{P}\Big(W = w\Big)} &= \frac{\prod_{j\in S\setminus\mathcal{I}} \mathcal{P}(W_j = 1) \prod_{k\in S^c} \mathcal{P}(W_k = 0)}{\prod_{j\in S} \mathcal{P}(W_j = 1) \prod_{k\in S^c} \mathcal{P}(W_k = 0)} \\
&= \prod_{j\in J} \frac{1}{\mathcal{P}(W_j = 1)} \\
&= \prod_{j\in J} \frac{\prod_{j\in S\setminus\{j\}} \mathcal{P}(W_j = 1) \prod_{k\in S^c} \mathcal{P}(W_k = 0)}{\prod_{j\in S} \mathcal{P}(W_j = 1) \prod_{k\in S^c} \mathcal{P}(W_k = 0)} \\
&= \prod_{i\in\mathcal{I}} \frac{\mathcal{P}\Big[\bigcap_{j\in S\setminus\{i\}}\{W_j = 1\} \cap \bigcap_{k\in S^c}\{W_k = 0\}\Big]}{\mathcal{P}\Big[W = w\Big]}.
\end{aligned}$$

Thus,

$$
\begin{aligned}
(*) &= \prod_{i\in\mathcal{I}} \mathcal{P}\Big[(X_i,Y_i,W_i)\in B_i\times\{1\}\Big] \prod_{i\in\mathcal{I}} \frac{\mathcal{P}\Big[\bigcap_{j\in S\setminus\{i\}}\{W_j=1\}\cap\bigcap_{k\in S^c}\{W_k=0\}\Big]}{\mathcal{P}\Big[W=w\Big]} \\
&= \prod_{i\in\mathcal{I}} \left(\mathcal{P}\left[(X_i,Y_i,W_i)\in B_i\times\{1\}\ \cap\ \Big(\bigcap_{j\in S\setminus\{i\}}\{W_j=1\}\cap\bigcap_{k\in S^c}\{W_k=0\}\Big)\right] / \mathcal{P}\Big[W=w\Big]\right) \\
&= \prod_{i\in\mathcal{I}} \Big(\mathcal{P}\left((X_i,Y_i)\in B_i\ \cap\ \Big\{W=w\Big\}\right) / \mathcal{P}\Big(W=w\Big)\Big) \\
&\prod_{i\in\mathcal{I}} \mathcal{P}\left((X_i,Y_i)\in B_i\Big|W=w\right),
\end{aligned}
$$

which completes the proof. □

Next, we are concerned with showing that all treated units have the same distribution.

Lemma SI 2 (identical distribution)**.** *Assume the same assumptions as in Lemma SI 1 and let $i\neq j\in S$. Under the conditional distribution of $W=w$, (X_i,Y_i) and (X_j,Y_j) have the same distribution, $\mathcal{P}_1$.*

Proof. Let $B\subset\mathbb{R}^p\times\mathbb{R}$; then

$$
\begin{aligned}
\mathcal{P}\left((X_i,Y_i)\in B\Big|W=w\right) &\overset{*}{=} \mathcal{P}\left((X_i,Y_i)\in B\Big|W_i=1\right) \\
&= \frac{\mathcal{P}\left((X_i,Y_i,W_i)\in B\times\{1\}\right)}{\mathcal{P}(W_i=1)} \\
&\overset{a}{=} \frac{\mathcal{P}\left((X_j,Y_j,W_j)\in B\times\{1\}\right)}{\mathcal{P}(W_j=1)} \\
&= \mathcal{P}\left((X_j,Y_j)\in B\Big|W_j=1\right) \\
&\overset{*}{=} \mathcal{P}\left((X_j,Y_j)\in B\Big|W=w\right).
\end{aligned}
$$

Here * follows from $(X_i,Y_i,W_i)_{i=1}^N$ being mutually independent, and ^{a} follows from $(X_i,Y_i,W_i)_{i=1}^N$ being identically distributed under $\mathcal{P}$. □

SI 8. Convergence Rate Results for the T-learner

In this section, we want to prove Theorem 1 of the main paper. We start with a short lemma that will be useful for the proof of the theorem.

Lemma SI 3. *Let $\mathcal{P}$ be defined as in [1] with $0<e_{\min}<e(x)<e_{\min}<1$. Furthermore, let X,W be distributed according to $\mathcal{P}$, and let g be a positive function such that the expectations below exist; then*

$$
\frac{e_{\min}}{e_{\max}}\mathbb{E}[g(X)] \ \leq \mathbb{E}[g(X)|W=1] \ \leq\ \frac{e_{\max}}{e_{\min}}\mathbb{E}[g(X)], \tag{SI 4}
$$

$$
\frac{1-e_{\max}}{1-e_{\min}}\mathbb{E}[g(X)] \ \leq \mathbb{E}[g(X)|W=0] \ \leq\ \frac{1-e_{\min}}{1-e_{\max}}\mathbb{E}[g(X)]. \tag{SI 5}
$$

Proof of Lemma SI 3. Let us prove [SI 4] first. The lower bound follows from

$$
\mathbb{E}[g(X)|W=1]\geq\mathbb{E}[g(X)]\frac{\inf_x e(x)}{E[W]}\geq\frac{e_{\min}}{E[W]}\mathbb{E}[g(X)]\geq\frac{e_{\min}}{e_{\max}}\mathbb{E}[g(X)],
$$

and the upper bound from

$$
\mathbb{E}[g(X)|W=1]\leq\mathbb{E}[g(X)]\frac{\sup_x e(x)}{E[W]}\leq\frac{e_{\max}}{e_{\min}}\mathbb{E}[g(X)].
$$

[SI 5] follows from a symmetrical argument. □

Let us now restate Theorem 1. Let $m, n \in \mathbb{N}^+$ and $N = m + n$ and let $\mathcal{P}$ be a distribution of (X, W, Y) according to [1] with the propensity score bounded away from 0 and 1. That is, there exists $e_{\min}$ and $e_{\max}$ such that $0 < e_{\min} < e(x) < e_{\max} < 1$. Furthermore, let $(X_i, W_i, Y_i)_{i=1}^N$ be i.i.d. from $\mathcal{P}$ and define $\mathcal{P}^{nm}$ to be the conditional distribution of $(X_i, W_i, Y_i)_{i=1}^N$ given that we observe n treated units, $\sum_{i=1}^N W_i = n$.

Note that n and m are not random under $\mathcal{P}^{nm}$. We are interested in the performance of the T-learner, $\hat{\tau}_T^{mn}$, under $\mathcal{P}^{nm}$ as measured by the EMSE,

$$\text{EMSE}(\hat{\tau}_T^{mn}, \mathcal{P}^{nm}) \stackrel{\text{def}}{=} \mathbb{E}\left[(\hat{\tau}_T^{mn}(\mathcal{X}) - \tau(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n \right].$$

The expectation is here taken over the training data set $(X_i, W_i, Y_i)_{i=1}^N$, which is distributed according to $\mathcal{P}^{nm}$, and $\mathcal{X}$, which is distributed according to the marginal distribution of X in $\mathcal{P}$.

For a family of superpopulations, $F \in S(a_\mu, a_\tau)$, we want to show that the T-learner with an optimal choice of base learners achieves a rate of

$$\mathcal{O}(m^{-a_\mu} + n^{-a_\mu}).$$

An optimal choice of base learners is estimators that achieve the minimax rate of n^{-a_μ} and m^{-a_μ} in F.

Proof of Theorem 1. The EMSE can be upper bounded by the errors of the single base learners:

$$\begin{aligned}
\text{EMSE}(\hat{\tau}_T^{mn}, \mathcal{P}^{nm}) &= \mathbb{E}\left[(\hat{\tau}_T^{mn}(\mathcal{X}) - \tau(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n \right] \\
&\leq 2\underbrace{\mathbb{E}\left[(\hat{\mu}_1^n(\mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n \right]}_{A} + 2\underbrace{\mathbb{E}\left[(\hat{\mu}_0^m(\mathcal{X}) - \mu_0(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n \right]}_{B}.
\end{aligned}$$

Here we use the following inequality:

$$(\hat{\tau}_T^{mn}(\mathcal{X}) - \tau(\mathcal{X}))^2 \leq 2(\hat{\mu}_1^n(\mathcal{X}) - \mu_1(\mathcal{X}))^2 + 2(\hat{\mu}_0^m(\mathcal{X}) - \mu_0(\mathcal{X}))^2.$$

Let us look only at the first term. We can write

$$\begin{aligned}
A &= \mathbb{E}\left[(\hat{\mu}_1^n(\mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n \right] \\
&= \mathbb{E}\left[\mathbb{E}\left[(\hat{\mu}_1^n(\mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| W, \sum_{i=1}^N W_i = n \right] \middle| \sum_{i=1}^N W_i = n \right]. \qquad \text{[SI 6]}
\end{aligned}$$

It is of course not necessary to condition on $\sum_{i=1}^N W_i = n$ in the inner expectation, and we only do so as a reminder that there are n treated units.

For $i \in \{1, \ldots, n\}$, let q_i be the i^{th} smallest number in $\{k : W_k = 1\}$. That is, $\{q_i : i \in \{1, \ldots, n\}\}$ are the indexes of the treated units. To emphasize that $\hat{\mu}_1^n(\mathcal{X})$ depends only on the treated observations, $(X_{q_i}, Y_{q_i})_{i=1}^n$, we write $\hat{\mu}_1^n((X_{q_i}, Y_{q_i})_{i=1}^n, \mathcal{X})$. Furthermore, we define $\mathcal{P}_1$ to be the conditional distribution of (X, Y) given $W = 1$. Conditioning on W, Theorem SI 1 implies that $(X_{q_i}, Y_{q_i})_{i=1}^n$ is i.i.d. from $\mathcal{P}_1$. Let us define $\tilde{\mathcal{X}}$ to be distributed according to $\mathcal{P}_1$. Then we can apply Lemma SI 3 and use the definition of $S(a_\mu, a_\tau)$ to conclude that the inner expectation in [SI 6] is in $\mathcal{O}(n^{-a_\mu})$:

$$\begin{aligned}
&\mathbb{E}\left[\hat{\mu}_1^n((X_{q_i}, Y_{q_i})_{i=1}^n, \mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| W, \sum_{i=1}^N W_i = n \right] \\
&\leq \frac{e_{\max}}{e_{\min}} \mathbb{E}\left[(\hat{\mu}_1^n((X_{q_i}, Y_{q_i})_{i=1}^n, \tilde{\mathcal{X}}) - \mu_1(\tilde{\mathcal{X}}))^2 \middle| W, \sum_{i=1}^n W_i = n \right] \\
&\leq \frac{e_{\max}}{e_{\min}} C n^{-a_\mu}.
\end{aligned}$$

Hence, it follows that

$$A \leq 2\mathbb{E}\left[\frac{e_{\max}}{e_{\min}} C n^{-a_\mu} \middle| \sum_{i=1}^n W_i = n \right] \leq 2\frac{e_{\max}}{e_{\min}} C n^{-a_\mu}.$$

By a symmetrical argument, it also holds that

$$B \leq 2\frac{1 - e_{\min}}{1 - e_{\max}} C m^{-a_\mu},$$

and we can conclude that

$$\text{EMSE}(\hat{\tau}_T^{mn}, \mathcal{P}) \leq 2C\left[\frac{1 - e_{\min}}{1 - e_{\max}} + \frac{e_{\max}}{e_{\min}} \right](n^{-a_\mu} + m^{-a_\mu}).$$

□

SI 9. Convergence Rate Results for the X-learner

In this section, we are concerned with the convergence rate of the X-learner. Given our motivation of the X-learner in the main paper, we believe that $\hat{\tau}_0$ of the X-learner should achieve a rate of $\mathcal{O}(m^{-a_\tau} + n^{-a_\mu})$ and $\hat{\tau}_1$ should achieve a rate of $\mathcal{O}(m^{-a_\mu} + n^{-a_\tau})$. In what follows, we prove this for two cases, and we show that for those cases the rate is optimal. In the first case, we assume that the CATE is linear and thus $a_\tau = 1$. We don't assume any regularity conditions on the response functions, and we show that the X-learner with an OLS estimator in the second stage and an appropriate estimator in the first stage achieves the optimal convergence rate. We show this first for the MSE (Theorem SI 2) and then for the EMSE (Theorem 2). We then focus on the case where we don't impose any additional regularity conditions on the CATE, but the response functions are Lipschitz continuous (Theorem SI 5). The optimal convergence rate is here not obvious, and we will first prove a minimax lower bound for the EMSE, and we will then show that the X-learner with the KNN estimates achieves this optimal performance.

SI 9.1. MSE and EMSE convergence rate for the linear CATE.

Theorem SI 2 (rate for the pointwise MSE)**.** *Assume that we observe m control units and n treated units from some superpopulation of independent and identically distributed observations $(Y(0), Y(1), X, W)$ coming from a distribution $\mathcal{P}$ given in equation [1] and assume that the following assumptions are satisfied:*

B1 Ignorability holds.

B2 The treatment effect is linear, $\tau(x) = x^T\beta$, with $\beta \in \mathbb{R}^d$.

B3 There exists an estimator $\hat{\mu}_0$ such that for all x,

$$\mathbb{E}\left[\left(\mu_0(x) - \hat{\mu}_0^m(x)\right)^2 \middle| \sum_{i=1}^N W_i = n\right] \leq C^0 m^{-a}.$$

B4 The error terms ε_i are independent given X, with $\mathbb{E}[\varepsilon_i | X = x] = 0$ and $Var[\varepsilon_i | X = x] \leq \sigma^2 < \infty$.

B5 The eigenvalues of the sample covariance matrix of the features of the treated units are well conditioned, in the sense that there exists an n_0, such that

$$\sup_{n>n_0} \mathbb{E}\left[\gamma_{min}^{-1}(\hat{\Sigma}_n) \middle| \sum_{i=1}^N W_i = n\right] < c_1 \qquad \text{and} \qquad \sup_{n>n_0} \mathbb{E}\left[\gamma_{max}(\hat{\Sigma}_n)/\gamma_{min}^2(\hat{\Sigma}_n) \middle| \sum_{i=1}^N W_i = n\right] < c_2, \tag{SI 7}$$

where $\hat{\Sigma}_n = \frac{1}{n}(X^1)'X^1$ and X^1 is the matrix consisting of the features of the treated units.

Then the X-learner with $\hat{\mu}_0$ in the first stage, OLS in the second stage, and weighting function $g \equiv 0$ has the following upper bound: for all $x \in \mathbb{R}^d$ and all $n > n_0$,

$$\mathbb{E}\left[\left(\tau(x) - \hat{\tau}_X(x)\right)^2 \middle| \sum_{i=1}^N W_i = n\right] \leq C_x \left(m^{-a} + n^{-1}\right) \tag{SI 8}$$

with $C_x = max(c_2 C^0, \sigma^2 d c_1)\|x\|^2$.

Proof of Theorem SI 2. To simplify the notation, we write X instead of X^1 for the observed features of the treated units. Furthermore, we denote that when $g \equiv 0$ in [9] in the main paper, the X-learner is equal to $\hat{\tau}_1$ and we only have to analyze the performance of $\hat{\tau}_1$.

The imputed treatment effects for the treatment group can be written as

$$D_i^1 = Y_i - \hat{\mu}_0(X_i) = X_i\beta + \delta_i + \epsilon_i,$$

with $\delta_i = \mu_0(X_i) - \hat{\mu}_0(X_i)$. In the second stage we estimate β using an OLS estimator,

$$\hat{\beta} = (X'X)^{-1}X'D^1.$$

To simplify the notation, we define the event of observering n treated units as $E_n = \{\sum_{i=1}^N W_i = n\}$. We decompose the MSE of $\hat{\tau}(x)$ into two orthogonal error terms:

$$\mathbb{E}\left[\left(\tau(x) - \hat{\tau}_X(x)\right)^2 \middle| \sum_{i=1}^N W_i = n\right] = \mathbb{E}\left[\left(x'(\beta - \hat{\beta})\right)^2 \middle| E_n\right] \leq \|x\|^2 \mathbb{E}\left[\|(X'X)^{-1}X'\delta\|^2 + \|(X'X)^{-1}X'\varepsilon\|^2 \middle| E_n\right]. \tag{SI 9}$$

Throughout the proof, we assume that $n > n_0$ such assumption B5 can be used. We will show that the second term decreases at the parametric rate, n^{-1}, while the first term decreases at a rate of m^{-a}:

$$\begin{aligned}\mathbb{E}\left[\|(X'X)^{-1}X'\varepsilon\|^2\middle|E_n\right] &= \mathbb{E}\left[tr\left(X(X'X)^{-1}(X'X)^{-1}X'\mathbb{E}\left[\varepsilon\varepsilon'\middle|X,E_n\right]\right)\middle|E_n\right] \\ &\leq \sigma^2 d\mathbb{E}\left[\gamma_{\min}^{-1}(\hat{\Sigma}_n)\middle|E_n\right]n^{-1} \\ &\leq \sigma^2 d c_1 n^{-1}.\end{aligned} \tag{SI 10}$$

For the last inequality we used assumption B5. Next, we are concerned with bounding the error coming from not perfectly predicting μ_0:

$$\begin{aligned}\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|_2^2\middle|E_n\right] &\leq \mathbb{E}\left[\gamma_{\max}(\hat{\Sigma}_n)/\gamma_{\min}^2(\hat{\Sigma}_n)\|\delta\|_2^2\middle|E_n\right]n^{-1} \\ &\leq \mathbb{E}\left[\gamma_{\max}(\hat{\Sigma}_n)/\gamma_{\min}^2(\hat{\Sigma}_n)\middle|E_n\right]C^0 m^{-a} \\ &\leq c_2 C^0 m^{-a}.\end{aligned} \tag{SI 11}$$

Here we used that $\gamma_{\max}(\hat{\Sigma}_n^{-2}) = \gamma_{\min}^{-2}(\hat{\Sigma}_n)$, and $\mathbb{E}\left[\|\delta\|_2^2\middle|X,E_n\right] = \mathbb{E}\left[\sum_{i=1}^n \delta^2(X_i)\middle|X,E_n\right] \leq nC^0m^{-a}$. For the last statement, we used assumption B5. This leads to [SI 8]. □

Bounding the EMSE.

Proof of Theorem 2. This proof is very similar to the proof of Theorem SI 2. The difference is that here we bound the EMSE instead of the pointwise MSE, and we have a somewhat weaker assumption, because $\hat{\mu}_0$ only satisfies that its EMSE converges at a rate of a, but not necessarily the MSE at every x. We introduce $\mathcal{X}$ here to be a random variable with the same distribution as the feature distribution such that the EMSE can be written as $\mathbb{E}[(\tau(\mathcal{X}) - \hat{\tau}_X(\mathcal{X}))^2|E_n]$. Recall that we use the notation that E_n is the event that we observe exactly n treated units and $m = N - n$ control units:

$$E_n = \left\{\sum_{i=1}^N W_i = n\right\}.$$

We start with a similar decomposition as in [SI 9]:

$$\begin{aligned}\mathbb{E}\left[(\tau(\mathcal{X}) - \hat{\tau}_X(\mathcal{X}))^2\middle|E_n\right] &\leq \mathbb{E}\left[\|\mathcal{X}\|^2\right]\mathbb{E}\left[\|\beta - \hat{\beta}\|^2\middle|E_n\right] \\ &= \mathbb{E}\left[\|\mathcal{X}\|^2\right]\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|^2 + \|(X'X)^{-1}X'\varepsilon\|^2\middle|E_n\right].\end{aligned} \tag{SI 12}$$

Following exactly the same steps as in [SI 10], we get

$$\mathbb{E}\left[\|(X'X)^{-1}X'\varepsilon\|^2\middle|E_n\right] \leq \sigma^2 d C_\Sigma n^{-1}.$$

Bounding $\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|_2^2\middle|E_n\right]$ is now slightly different than in [SI 11]:

$$\begin{aligned}\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|_2^2\middle|E_n\right] &\leq \mathbb{E}\left[\gamma_{\min}^{-1}(X'X)\|X(X'X)^{-1}X'\delta\|_2^2\middle|E_n\right] \\ &\leq \mathbb{E}\left[\gamma_{\min}^{-1}(X'X)\|\delta\|_2^2\middle|E_n\right] \\ &\leq \mathbb{E}\left[\gamma_{\min}^{-1}(\Sigma_n)\frac{1}{n}\|\delta\|_2^2\middle|E_n\right] \\ &\leq C_\Sigma\mathbb{E}\left[\|\delta_1\|_2^2\middle|E_n\right].\end{aligned} \tag{SI 13}$$

Here the last inequality follows from Condition 6.

We now apply [SI 4], [SI 5], and Condition 4 to conclude that

$$\begin{aligned}\mathbb{E}\left[\|\delta_1\|_2^2\middle|E_n\right] &= \mathbb{E}\left[\|\mu_0(X_1) - \hat{\mu}_0(X_1)\|_2^2\middle|E_n, W_1 = 1\right] \\ &\leq \frac{e_{\max} - e_{\max}e_{\min}}{e_{\min} - e_{\max}e_{\min}}\mathbb{E}\left[\|\mu_0(X_1) - \hat{\mu}_0(X_1)\|_2^2\middle|E_n, W_1 = 0\right] \\ &\leq \frac{e_{\max} - e_{\max}e_{\min}}{e_{\min} - e_{\max}e_{\min}}C_0 m^{-a_\mu}.\end{aligned}$$

Lastly, we use the assumption that $\mathbb{E}\left[\|\mathcal{X}\|^2\middle|E_n\right] \leq C_\mathcal{X}$ and conclude that

$$\mathbb{E}\left[(\tau(\mathcal{X}) - \hat{\tau}_X(\mathcal{X}))^2\middle|E_n\right] \leq C_\mathcal{X}\left(\frac{e_{\max} - e_{\max}e_{\min}}{e_{\min} - e_{\max}e_{\min}}C_\Sigma C_0 m^{-a} + \sigma^2 d C_\Sigma n^{-1}\right). \tag{SI 14}$$

□

SI 9.2. Achieving the parametric rate. When there are a lot of control units, such that $m \geq n^{1/a}$, then we have seen that the X-learner achieves the parametric rate. However, in some situations the X-learner also achieves the parametric rate even if the number of control units is of the same order as the number of treated units. To illustrate this, we consider an example in which the conditional average treatment effect and the response functions depend on disjoint and independent subsets of the features.

Specifically, we assume that we observe m control units and n treated units according to Model 1. We assume the same setup and the same conditions as in Theorem 2. In particular, we assume that there exists an estimator $\hat{\mu}_0^m$ that depends only on the control observations and estimates the control response function at a rate of at most m^{-a}. In addition to these conditions we also assume the following independence condition.

Condition 7. *There exists subsets, $S, \bar{S} \subset \{1, \dots, d\}$ with $S \cap \bar{S} = \emptyset$, such that*

- *$(X_i)_{i \in S}$ and $(X_i)_{i \in \bar{S}}$ are independent.*
- *For all $i \in S$, $E[X_i | W_i = 1] = 0$.*
- *There exist a function $\tilde{\mu}_0$, and a vector $\tilde{\beta}$ with $\mu_0(x) = \tilde{\mu}_0(x_{\bar{S}})$ and $\tau(x) = x_S^T \tilde{\beta}$.*

For technical reasons, we also need bounds on the fourth moments of the feature vector and the error of the estimator for the control response.

Condition 8. *The fourth moments of the feature vector X are bounded:*

$$\mathbb{E}[\|X\|_2^4 | W = 1] \leq C_X.$$

Condition 9. *There exists an m_0 such that for all $m > m_0$,*

$$\mathbb{E}\left[(\mu_0(X) - \hat{\mu}_0^m(X))^4 \middle| W = 1 \right] \leq C_\delta.$$

Here $\hat{\mu}_0^m$ is defined as in Condition 4.

This condition is satisfied, for example, when μ_0 is bounded.

Under these additional assumptions, the EMSE of the X-learner achieves the parametric rate in n, given that $m > m_0$.

Theorem SI 3. *Assume that Conditions 1–9 hold. Then the X-learner with $\hat{\mu}_0^m$ in the first stage and OLS in the second stage achieves the parametric rate in n. That is, there exists a constant C such that for all $m > m_0$ and $n > 1$,*

$$\mathbb{E}\left[(\tau(\mathcal{X}) - \hat{\tau}_{\mathcal{X}}^{mn}(\mathcal{X}))^2 \middle| \sum_i W_i = n \right] \leq C n^{-1}.$$

We will prove the following lemma first, because it will be useful for the proof of Theorem SI 3.

Lemma SI 4. *Under the assmuption of Theorem SI 3, there exists a constant C such that for all $n > n_0$, $m > m_0$, and $s > 0$,*

$$\mathcal{P}\left(n \|(X^{1\prime} X^1)^{-1} X^{1\prime} \delta\|_2^2 \geq s \middle| \sum_i W_i = n \right) \leq C \frac{1}{s^2},$$

where $\delta_i = \mu_0(X_i^1) - \hat{\mu}_0^m(X_i^1)$.

Proof of Lemma SI 4. To simplify the notation, we write X instead of X^1 for the feature matrix of the treated units, and we define the event of observing exactly n treated units as

$$E_n = \left\{ \sum_{i=1}^n W_i = n \right\}.$$

We use Condition 6 and then Chebyshev's inequality to conclude that for all $n > n_0$ (n_0 is determined by Condition 6),

$$\begin{aligned}
\mathcal{P}\left(n \|(X'X)^{-1} X'\delta\|_2^2 \geq s \middle| E_n \right) &= \mathcal{P}\left(\frac{1}{n} \|\Sigma_n^{-1} X'\delta\|_2^2 \geq s \middle| E_n \right) \\
&\leq \mathcal{P}\left(\frac{1}{n} \gamma_{\min}^{-2}(\Sigma_n) \|X'\delta\|_2^2 \geq s \middle| E_n \right) \\
&\leq \mathbb{E}\left[\mathcal{P}\left(\frac{1}{n} C_\Sigma^2 \|X'\delta\|_2^2 \geq s \middle| E_n, \delta \right) \middle| E_n \right] \\
&\leq \mathbb{E}\left[\frac{C_\Sigma^4}{s^2 n^2} \mathrm{Var}\left(\|X'\delta\|_2^2 \middle| E_n, \delta \right) \middle| E_n \right].
\end{aligned}$$

Next we apply the Efron–Stein inequality to bound the variance term:

$$\mathrm{Var}\left(\|X'\delta\|_2^2 \middle| E_n, \delta\right) \leq \frac{1}{2}\sum_{i=1}^{n} \mathbb{E}\left[(f(X) - f(X^{(i)}))^2 \middle| E_n, \delta\right].$$

Here $f(x) = \|x'\delta\|_2^2$, $X^{(i)} = (X_1, \ldots, X_{i-1}, \tilde{X}_i, X_{i+1}, \ldots, X_n)$, and $\tilde{X}$ is an independent copy of X.

Let us now bound the summands:

$$\begin{aligned}
&\mathbb{E}\left[(f(X) - f(X^{(i)}))^2 \middle| E_n, \delta\right] \\
=&\mathbb{E}\left[\left(\|X'\delta\|_2^2 - \|X'\delta - (X_i - \tilde{X}_i)\delta_i\|_2^2\right)^2 \middle| E_n, \delta\right] \\
=&\mathbb{E}\Bigg[\underbrace{\left(2\delta' X(X_i - \tilde{X}_i)\delta_i\right)^2}_{A} + \underbrace{\|(X_i - \tilde{X}_i)\delta_i\|_2^4}_{B} - \underbrace{4\delta' X(X_i - \tilde{X}_i)\delta_i \|(X_i - \tilde{X}_i)\delta_i\|_2^2}_{C} \Bigg| E_n, \delta\Bigg].
\end{aligned}$$

Let us first bound $\mathbb{E}[A|E_n, \delta]$:

$$\begin{aligned}
\mathbb{E}\left[\left(2\delta' X(X_i - \tilde{X}_i)\delta_i\right)^2 \middle| E_n, \delta\right] &= \mathbb{E}\left[4\sum_{j,k=1}^{n} \delta_j X_j'(X_i - \tilde{X}_i)\delta_i\delta_k X_k'(X_i - \tilde{X}_i)\delta_i \middle| E_n, \delta\right] \\
&\overset{(a)}{=} \mathbb{E}\left[4\sum_{j=1}^{n}(\delta_j X_j'(X_i - \tilde{X}_i)\delta_i)^2 \middle| E_n, \delta\right] \\
&\leq 4\delta_i^4(n-1)\mathbb{E}\left[(X_1'(X_2 - \tilde{X}_2))^2 \middle| E_n, \delta\right] + 4\delta_i^4\mathbb{E}\left[(X_1'(X_1 - \tilde{X}_1))^2 \middle| E_n, \delta\right] \\
&\leq C_A\delta_i^4 n.
\end{aligned}$$

Here

$$C_A = 4\max\left(\mathbb{E}\left[(X_1'(X_2 - \tilde{X}_2))^2 \middle| E_n\right], \mathbb{E}\left[(X_1'(X_1 - \tilde{X}_1))^2 \middle| E_n\right]\right),$$

which is bounded by Condition 8. For equation (a) we used that for $k \neq j$; therefore, we have that either k or j is not equal to i. Without loss of generality let $j \neq i$. Then

$$\begin{aligned}
&E\left[\delta_j X_j'(X_i - \tilde{X}_i)\delta_i\delta_k X_k'(X_i - \tilde{X}_i)\delta_i \middle| E_n, \delta\right] \\
&= \delta_j\mathbb{E}\left[\mathbb{E}\left[X_j' \middle| W, E_n, \delta\right]\mathbb{E}\left[(X_i - \tilde{X}_i)\delta_i\delta_k X_k'(X_i - \tilde{X}_i)\delta_i \middle| W, E_n, \delta\right] \middle| E_n, \delta\right] \\
&= 0,
\end{aligned} \qquad \text{[SI 15]}$$

because $\mathbb{E}\left[X_j' | W, E_n, \delta\right] = 0$ as per the assumption.

In order to bound $\mathbb{E}[B|E_n, \delta]$, note that all the fourth moments of X are bounded and thus

$$\mathbb{E}\left[\|(X_i - \tilde{X}_i)\delta_i\|_2^4 \middle| E_n, \delta\right] \leq C_B\delta_i^4.$$

Finally, we bound $\mathbb{E}[C|E_n, \delta]$:

$$\begin{aligned}
\mathbb{E}\left[4\delta' X(X_i - \tilde{X}_i)\delta_i\|(X_i - \tilde{X}_i)\delta_i\|_2^2 \middle| E_n, \delta\right] &= \mathbb{E}\left[\sum_{j=1}^{n}\delta_j X_j'(X_i - \tilde{X}_i)\delta_i\|(X_i - \tilde{X}_i)\delta_i\|_2^2 \middle| E_n, \delta\right] \\
&= \mathbb{E}\left[\delta_i^4 X_i'(X_i - \tilde{X}_i)\|X_i - \tilde{X}_i\|_2^2 \middle| E_n, \delta \middle| E_n, \delta\right] \\
&= C_C\delta_i^4,
\end{aligned}$$

where the second equality follows from the same argument as in [SI 15], and the last equality is implied by Condition 8.

Plugging in terms A, B, and C, we have that for all $n > n_0$,

$$\mathrm{Var}\left(\|X'\delta\|_2^2 \middle| E_n, \delta\right) \leq \frac{1}{2}\sum_{i=1}^{n} E[(f(X, \delta) - f(X^{(i)}, \delta^{(i)}))^2] \leq C\delta^4 n^2,$$

with $C = C_A + C_B + C_C$. Thus for $n > n_0$,

$$\mathcal{P}\left(n\|(X'X)^{-1}X'\delta\|_2^2 \geq s \middle| E_n\right) \leq \mathbb{E}\left[\frac{CC_\Sigma^4}{s^2}\delta^4 \middle| E_n\right] \leq CC_\Sigma^4 C_\delta \frac{1}{s^2}.$$

□

Proof of Theorem SI 3. We start with the same decomposition as in [SI 12]:

$$\mathbb{E}\left[(\tau(\mathcal{X}) - \hat{\tau}_X^{mn}(\mathcal{X}))^2 \middle| E_n\right] \leq \mathbb{E}\left[\|\mathcal{X}\|^2\right] \mathbb{E}\left[\|(X'X)^{-1}X'\delta\|^2 + \|(X'X)^{-1}X'\varepsilon\|^2 \middle| E_n\right],$$

and we follow the same steps to conclude that

$$\mathbb{E}\left[\|(X'X)^{-1}X'\varepsilon\|^2 \middle| E_n\right] \leq \sigma^2 d C_\Sigma n^{-1} \qquad \text{and} \qquad \mathbb{E}\left[\|\mathcal{X}\|^2\right] \leq C_{\mathcal{X}}.$$

From Lemma SI 4, we can conclude that there exists a constant C such that

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{E}\left[n\|(X'X)^{-1}X'\delta\|_2^2 \middle| E_n\right] &= \lim_{n\to\infty, n>n_0} \int_0^\infty \mathcal{P}\left(n\|(X'X)^{-1}X'\delta\|_2^2 \geq s \middle| E_n\right) ds \\
&\leq \lim_{n\to\infty, n>n_0} \int_0^\infty \max(1, C\frac{1}{s^2}) ds \\
&\leq 1 + C.
\end{aligned}$$

Thus there exists a $\tilde{C}$ such that for all $n > 1$,

$$\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|_2^2 \middle| E_n\right] \leq \tilde{C} n^{-1}.$$

□

SI 9.3. EMSE convergence rate for Lipschitz continuous response functions. In Section SI 9.1, we considered an example where the distribution of $(Y(0), Y(1), W, X)$ was assumed to be in some family $F \in S(a_\mu, a_\tau)$ with $a_\tau > a_\mu$, and we showed that one can expect the X-learner to outperform the T-learner in this case. Now we want to explore the case where $a_\tau \leq a_\mu$.

Let us first consider the case, where $a_\tau < a_\mu$. This is a somewhat artificial case, since having response functions that can be estimated at a rate of N^{-a_μ} implies that the CATE cannot be too complicated. For example, if μ_0 and μ_1 are Lipschitz continuous, then the CATE is Lipschitz continuous as well, and we would expect $a_\tau \approx a_\mu$. Even though it is hard to construct a case with $a_\tau < a_\mu$, we cannot exclude such a situation, and we would expect that in such a case the T-learner performs better than the X-learner.

We therefore believe that the case where $a_\tau \approx a_\mu$ is a more reasonable assumption than the case where $a_\tau < a_\mu$. In particular, we would expect the T- and X-learners to perform similarly when compared to their worst-case convergence rate. Let us try to back up this intuition with a specific example. Theorem 2 already confirms that $\hat{\tau}_1$ achieves the expected rate,

$$\mathcal{O}\left(m^{-a_\mu} + n^{-a_\tau}\right),$$

for the case where the CATE is linear. Below, we consider another example, where the CATE is of the same order as the response functions. We assume some noise level σ that is fixed, and we start by introducing a family F^L of distributions with Lipschitz continuous regression functions.

Definition SI 1 (Lipschitz continuous regression functions)**.** *Let F^L be the class of distributions on $(X, Y) \in [0,1]^d \times \mathbb{R}$ such that:*

1. *The features, X_i, are i.i.d. uniformly distributed in $[0,1]^d$.*
2. *The observed outcomes are given by*
$$Y_i = \mu(X_i) + \varepsilon_i,$$
where the ε_i is independent and normally distributed with mean 0 and variance σ^2.
3. *X_i and ε_i are independent.*
4. *The regression function μ is Lipschitz continuous with parameter L.*

Remark SI 2. *The optimal rate of convergence for the regression problem of estimating $x \mapsto \mathbb{E}[Y|X = x]$ in Definition SI 1 is $N^{-2/(2+d)}$. Furthermore, the KNN algorithm with the right choice of the number of neighbors and the Nadaraya–Watson estimator with the right kernels achieve this rate, and they are thus minimax optimal for this regression problem.*

Now let's define a related distribution on $(Y(0), Y(1), W, X)$.

Definition SI 2. *Let $\mathcal{D}_{mn}^L$ be the family of distributions of $(Y(0), Y(1), W, X) \in \mathbb{R}^N \times \mathbb{R}^N \times \{0,1\}^N \times [0,1]^{d\times N}$ such that:*

1. *$N = m + n$.*
2. *The features, X_i, are i.i.d. uniformly distributed in $[0,1]^d$.*
3. *There are exactly n treated units,*
$$\sum_i W_i = n.$$

4. *The observed outcomes are given by*

$$Y_i(w) = \mu_w(X_i) + \varepsilon_{wi},$$

where $(\varepsilon_{0i}, \varepsilon_{1i})$ *is independent normally distributed with mean 0 and marginal variances* σ^2.[‡]

5. X, W *and* $\varepsilon = (\varepsilon_{0i}, \varepsilon_{1i})$ *are independent.*

6. *The response functions* μ_0, μ_1 *are Lipschitz continuous with parameter* L.

Note that if $(Y(0), Y(1), W, X)$ is distributed according to a distribution in D^L_{mn}, then $(Y(0), X)$ given $W = 0$ and $(Y(1), X)$ given $W = 1$ have marginal distributions in F^L, and $(X, \mu_1(X) - Y(0))$ given $W = 0$ and $(X, Y(1) - \mu_0(X))$ given $W = 1$ have distributions in F^{2L}, and we therefore conclude that $D^L_{mn} \in S\left(\frac{2}{2+d}, \frac{2}{2+d}\right)$.

We will first prove in Theorem SI 4 that the best possible rate that can be uniformly achieved for distributions in this family is

$$\mathcal{O}(n^{2/(2+d)} + m^{2/(2+d)}).$$

This is precisely the rate the T-learner with the right base learners achieves (Theorem 1). We will then show in Theorem SI 5 that the X-learner with the KNN estimator for both stages achieves this optimal rate as well, and conclude that both the T- and X-learners achieve the optimal minimax rate for this class of distributions.

Minimax lower bound. In this section, we will derive a lower bound on the best possible rate for $\mathcal{D}^L_{mn}$.

Theorem SI 4 (Minimax Lower Bound)**.** *Let* $\hat{\tau}$ *be an arbitrary estimator, let* $a_1, a_2 > 0$*, and let* c *be such that for all* $n, m \geq 1$*,*

$$\sup_{\mathcal{P} \in \mathcal{D}^L_{mn}} EMSE(\mathcal{P}, \hat{\tau}^{mn}) \leq c(m^{-a_0} + n^{-a_1}); \qquad \text{[SI 16]}$$

then a_1 *and* a_2 *are at most* $2/(2+d)$*:*

$$a_0, a_1 \leq 2/(2+d).$$

Proof of Theorem SI 4. To simplify the notation, we define $a = 2/(2+d)$. We will show by contradiction that $a_1 \leq a$. The proof of a_0 is mathematically symmetric. We assume that a_1 is bigger than a, and we show that this implies that there exists a sequence of estimators $\hat{\mu}^n_1$, such that

$$\sup_{\mathcal{P}_1 \in F^L} \mathbb{E}_{\mathcal{D}^n_1 \sim \mathcal{P}^n_1}\left[(\mu_1(\mathcal{X}) - \hat{\mu}^n_1(\mathcal{X}; \mathcal{D}^n_1))^2\right] \leq 2cn^{-a_1},$$

which is a contradiction, since by the definition of D^{mn}_L, μ_1 cannot be estimated at a rate faster than n^{-a} (cf., (10)). Note that we write here $\hat{\mu}^n_1(\mathcal{X}; \mathcal{D}^n_1)$, because we want to be explicit that $\hat{\mu}^n_1$ depends only on the treated observations.

Similiarly to $\hat{\mu}^n_1(\mathcal{X}; \mathcal{D}^n_1)$, we will use the notation $\hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}^m_0, \mathcal{D}^n_1)$ to be explicit about the dependence of the estimator $\hat{\tau}^{mn}$ on the data in the control group, $\mathcal{D}^m_0$, and on the data in the treatment group, $\mathcal{D}^n_1$. Furthermore, note that in Definition SI 2 each distribution in $\mathcal{D}^L_{mn}$ is fully specified by the distribution of W, ε, and the functions μ_1 and μ_2. Define C_L to be the set of all functions $f : [0,1]^d \longrightarrow \mathbb{R}$ that are L-Lipschitz continuous. For $f_1 \in C_L$, define $\mathbb{D}(f_1)$ to be the distribution in $\mathcal{D}^L_{mn}$ with $\mu_0 = 0$, $\mu_1 = f_1$, $\varepsilon_0 \perp \varepsilon_1$, and W defined componentwise by

$$W_i = \begin{cases} 1 \text{ if } i \leq n \\ 0 \text{ otherwise.} \end{cases}$$

Then [SI 16] implies that

$$\begin{aligned} c(m^{-a_0} + n^{-a_1}) &\geq \sup_{\mathcal{P} \in \mathcal{D}^L_{mn}} \mathbb{E}_{(\mathcal{D}^m_0 \times \mathcal{D}^n_1) \sim \mathcal{P}}\left[(\tau^{\mathcal{P}}(\mathcal{X}) - \hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}^m_0, \mathcal{D}^n_1))^2\right] \\ &\geq \sup_{f_1 \in C_L} \mathbb{E}_{(\mathcal{D}^m_0 \times \mathcal{D}^n_1) \sim \mathbb{D}(f_1)}\left[(\mu_1^{\mathbb{D}(f_1)}(\mathcal{X}) - \hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}^m_0, \mathcal{D}^n_1))^2\right]. \end{aligned}$$

This follows, because in $\mathbb{D}(f_1)$, $\tau^{\mathbb{D}(f_1)} = \mu_1^{\mathbb{D}(f_1)} = f_1$. We use here the notation $\tau^{\mathcal{P}}$, $\tau^{\mathbb{D}(f_1)}$, and $\mu_1^{\mathbb{D}(f_1)}$ to emphasize that those terms depend on the distribution of $\mathcal{P}$ and $\mathbb{D}(f_1)$, respectively.

Let $\mathcal{P}_0$ be the distribution of $\mathcal{D}^m_0 = (X^0_i, Y^0_i)^N_{i=1}$ under $\mathbb{D}(f_1)$. Note that under $\mathcal{P}_0$, $X_i \overset{iid}{\sim} [0,1]$, and $Y^0 \overset{iid}{\sim} \mathbb{N}(0, \sigma^2)$, and X^0 and Y^0 are independent. In particular, $\mathcal{P}_0$ does not depend on f_1. We can thus write

$$\begin{aligned} c(m^{-a_0} + n^{-a_1}) &\geq \sup_{f_1 \in C_L} \mathbb{E}_{(\mathcal{D}^m_0 \times \mathcal{D}^n_1) \sim \mathbb{D}(f_1)}\left[\left(\mu_1^{\mathbb{D}(f_1)}(\mathcal{X}) - \hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}^m_0, \mathcal{D}^n_1)\right)^2\right] \\ &= \sup_{f_1 \in C_L} \mathbb{E}_{\mathcal{D}^n_1 \sim \mathbb{D}_1(f_1)} \mathbb{E}_{\mathcal{D}^n_0 \sim \mathcal{P}_0}\left[\left(\mu_1^{\mathbb{D}_1(f_1)}(\mathcal{X}) - \hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}^m_0, \mathcal{D}^n_1)\right)^2\right] \end{aligned}$$

[‡] We do not assume that $\varepsilon_{0i} \perp \varepsilon_{1i}$.

$$\geq \sup_{f_1 \in C_L} \mathbb{E}_{\mathcal{D}_1^n \sim \mathbb{D}_1(f_1)} \left[\left(\mu_1^{\mathbb{D}_1(f_1)}(\mathcal{X}) - \mathbb{E}_{\mathcal{D}_0^n \sim \mathcal{P}_0} \hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}_0^m, \mathcal{D}_1^n) \right)^2 \right].$$

$\mathbb{D}_1(f_1)$ is here the distribution of $\mathcal{D}_1^n$ under $\mathbb{D}(f_1)$. For the last step we used Jensen's inequality.

Now choose a sequence m_n in such a way that $m_n^{-a_1} + n^{-a_2} \leq 2n^{-a_1}$, and define

$$\hat{\mu}_1^n(x; \mathcal{D}_1^n) = \mathbb{E}_{\mathcal{D}_0^{m_n} \sim \mathcal{P}_0^{m_n}} \left[\hat{\tau}^{mn}(x; \mathcal{D}_0^{m_n}, \mathcal{D}_1^n) \right].$$

Furthermore, note that

$$\{\mathbb{D}_1(f_1) \ : \ f_1 \in C_L\} = \left\{\mathcal{P}_1 \in F^L\right\}$$

in order to conclude that

$$\begin{aligned} 2cn^{-a_1} \geq c(m_n^{-a_0} + n^{-a_1}) &\geq \sup_{f_1 \in C_L} \mathbb{E}_{\mathcal{D}_1^n \sim \mathbb{D}_1(f_1)} \left[\left(\mu_1^{\mathbb{D}_1(f_1)}(\mathcal{X}) - \hat{\mu}_1^{nm}(D_1^n; \mathcal{X}) \right)^2 \right] \\ &\geq \sup_{\mathcal{P}_1 \in F^L} \mathbb{E}_{D_1^n \sim \mathcal{P}_1^n} \left[\left(\mu_1^{\mathcal{P}_1^n}(\mathcal{X}) - \hat{\mu}_1^{nm}(D_1^n; \mathcal{X}) \right)^2 \right]. \end{aligned}$$

This is, however, a contradiction, because we assumed $a_1 > a$. □

EMSE convergence of the X-learner. Finally, we can show that the X-learner with the right choice of base learners achieves this minimax lower bound.

Theorem SI 5. *Let $d > 2$ and assume $(X, W, Y(0), Y(1)) \sim \mathcal{P} \in \mathcal{D}_{mn}^L$. In particular, μ_0 and μ_1 are Lipschitz continuous with constant L,*

$$|\mu_w(x) - \mu_w(z)| \leq L\|x - z\| \quad \text{for } w \in \{0, 1\},$$

and $X \sim Unif([0,1]^d)$.

Furthermore, let $\hat{\tau}^{mn}$ be the X-learner with

- *$g \equiv 0$,*
- *the base learner of the first stage for the control group $\hat{\mu}_0$, is a KNN estimator with constant $k_0 = \left\lceil (\sigma^2/L^2)^{\frac{d}{2+d}} m^{\frac{2}{d+2}} \right\rceil$,*
- *the base learner of the second stage for the treatment group, $\hat{\tau}_1$, is a KNN estimator with constant $k_1 = \left\lceil (\sigma^2/L^2)^{\frac{d}{2+d}} n^{\frac{2}{d+2}} \right\rceil$.*

Then $\hat{\tau}^{mn}$ achieves the optimal rate as given in Theorem SI 4. That is, there exists a constant C such that

$$\mathbb{E}\|\tau - \hat{\tau}^{mn}\|^2 \leq C\sigma^{\frac{4}{d+2}} L^{\frac{2d}{2+d}} \left(m^{-2/(2+d)} + n^{-2/(2+d)} \right). \qquad \text{[SI 17]}$$

Note that in the third step of the X-learner, Equation [9], $\hat{\tau}_0$ and $\hat{\tau}_1$ are averaged:

$$\hat{\tau}^{mn}(x) = g(x)\hat{\tau}_0^{mn}(x) + (1 - g(x))\hat{\tau}_1^{mn}(x).$$

By choosing $g \equiv 0$, we are analyzing $\hat{\tau}_1^{mn}$. By a symmetry argument it is straightforward to show that with the right choice of base learners, $\hat{\tau}_0^{mn}$ also achieves a rate of $\mathcal{O}\left(m^{-2/(2+d)} + n^{-2/(2+d)}\right)$. With this choice of base learners the X-learner achieves this optimal rate for every choice of g.

We first state two useful lemmata that we will need in the proof of this theorem.

Lemma SI 5. *Let $\hat{\mu}_0^m$ be a KNN estimator based only on the control group with constant k_0, and let $\hat{\mu}_1^n$ be a KNN estinator based on the treatment group with constant k_1; then, by the assumption of Theorem SI 5,*

$$\begin{aligned} \mathbb{E}[\|\hat{\mu}_0^m - \mu_0\|^2] &\leq \frac{\sigma^2}{k_0} + cL^2 \left(\frac{k_0}{m}\right)^{2/d}, \\ \mathbb{E}[\|\hat{\mu}_1^n - \mu_1\|^2] &\leq \frac{\sigma^2}{k_1} + cL^2 \left(\frac{k_1}{n}\right)^{2/d}, \end{aligned}$$

for some constant c.

Proof of Lemma SI 5. This is a direct implication of Theorem 6.2 in (10). □

Lemma SI 6. *Let $x \in [0,1]^d$, $X_1, \ldots, X_n \overset{iid}{\sim} Unif([0,1]^d)$ and $d > 2$. Define $\tilde{X}(x)$ to be the nearest neighbor of x; then there exists a constant c such that for all $n > 0$,*

$$\mathbb{E}\|\tilde{X}(x) - x\|^2 \leq \frac{c}{n^{2/d}}.$$

Proof of Lemma SI 6. First of all we consider

$$\mathcal{P}(\|\tilde{X}(x) - x\| \geq \delta) = (1 - \mathcal{P}(\|X_1 - x\| \leq \delta))^n \leq (1 - \tilde{c}\delta^d)^n \leq e^{-\tilde{c}\delta^d n}.$$

Now we can compute the expectation:

$$\mathbb{E}\|\tilde{X}(x) - x\|^2 = \int_0^\infty \mathcal{P}(\|\tilde{X}(x) - x\| \geq \sqrt{\delta})d\delta \leq \int_0^d e^{-\tilde{c}\delta^{d/2}n} d\delta \leq \frac{1 - \frac{1}{-d/2+1}}{(\tilde{c}n)^{2/d}}.$$

□

Proof of Theorem SI 5. Many ideas in this proof are motivated by (10) and (11). Furthermore, note that we restrict our analysis here only to $\hat{\tau}_1^{mn}$, but the analysis of $\hat{\tau}_0^{mn}$ follows the same steps.

We decompose $\hat{\tau}_1^{mn}$ into

$$\hat{\tau}_1^{mn}(x) = \frac{1}{k_1}\sum_{i=1}^{k_1}\left[Y^1_{(i,n)}(x) - \hat{\mu}_0^m\left(X^1_{(i,n)}(x)\right)\right] = \hat{\mu}_1^n(x) - \frac{1}{k_1}\sum_{i=1}^{k_1}\hat{\mu}_0^m\left(X^1_{(i,n)}(x)\right),$$

where the notation that $\left(\left(X^w_{(1,n_w)}(x), Y^w_{(1,n_w)}(x)\right), \ldots, \left(X^w_{(n_w,n_w)}(x), Y^w_{(n_w,n_w)}(x)\right)\right)$ is a reordering of the tuples $\left(X^w_j(x), Y^w_j(x)\right)$ such that $\|X^w_{(i,n_w)}(x) - x\|$ is increasing in i. With this notation we can write the estimators of the first stage as

$$\hat{\mu}_0^m(x) = \frac{1}{k_0}\sum_{i=1}^{k_0} Y^0_{(i,m)}(x), \qquad \text{and} \qquad \hat{\mu}_1^n(x) = \frac{1}{k_1}\sum_{i=1}^{k_1} Y^1_{(i,n)}(x),$$

and we can upper bound the EMSE with the following sum:

$$\begin{aligned}
&\mathbb{E}[|\tau(\mathcal{X}) - \hat{\tau}_1^{mn}(\mathcal{X})|^2] \\
&= \mathbb{E}\left[\left|\mu_1(\mathcal{X}) - \mu_0(\mathcal{X}) - \hat{\mu}_1^n(\mathcal{X}) + \frac{1}{k_1}\sum_{i=1}^{k_1}\hat{\mu}_0^m(X^1_{(i,n)}(\mathcal{X}))\right|^2\right] \\
&\leq 2\mathbb{E}\left[|\mu_1(\mathcal{X}) - \hat{\mu}_1^n(\mathcal{X})|^2\right] + 2\mathbb{E}\left[\left|\mu_0(\mathcal{X}) - \frac{1}{k_1}\sum_{i=1}^{k_1}\hat{\mu}_0^m(X^1_{(i,n)}(\mathcal{X}))\right|^2\right].
\end{aligned}$$

The first term corresponds to the regression problem of estimating the treatment response function in the first step of the X-learner and we can control this term with Lemma SI 5:

$$\mathbb{E}[\|\mu_1 - \hat{\mu}_1^n\|^2] \leq \frac{\sigma^2}{k_1} + c_1 L^2\left(\frac{k_1}{n}\right)^{2/d}.$$

The second term is more challenging:

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}\left[\left|\mu_0(\mathcal{X}) - \frac{1}{k_1}\sum_{i=1}^{k_1}\hat{\mu}_0^m(X^1_{(i,n)}(\mathcal{X}))\right|^2\right] \\
&\leq \mathbb{E}\left[\left|\mu_0(\mathcal{X}) - \frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mu_0\left(X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right)\right|^2\right] && \text{[SI 18]} \\
&+ \mathbb{E}\left[\left|\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mu_0\left(X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right) - \frac{1}{k_1}\sum_{i=1}^{k_1}\hat{\mu}_0^m(X^1_{(i,n)}(\mathcal{X}))\right|^2\right]. && \text{[SI 19]}
\end{aligned}$$

[SI 19] can be bound as follows:

$$\begin{aligned}
[SI19] &= \mathbb{E}\left(\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mu_0\left(X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right) - Y^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right)^2 \\
&\leq \max_i \frac{1}{k_m^2}\sum_{j=1}^{k_0}\mathbb{E}\left(\mu_0\left(X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right) - Y^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right)^2 \\
&= \max_i \frac{1}{k_m^2}\sum_{j=1}^{k_0}\mathbb{E}\left[\mathbb{E}\left[\left(\mu_0\left(X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right) - Y^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right)^2 \middle\| \mathcal{D}, \mathcal{X}\right]\right] \leq \frac{\sigma^2}{k_0}.
\end{aligned}$$

The last inequality follows from the assumption that, conditional on $\mathcal{D}$,

$$Y^0_{(j,m)}(x) \sim \mathcal{N}\left(\mu_0\left(X^0_{(j,m)}(x)\right), \sigma^2\right).$$

Next we find an upper bound for [SI 18]:

$$\begin{aligned}
[SI18] &\leq \mathbb{E}\left(\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\left\|\mu_0(\mathcal{X}) - \mu_0\left(X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right)\right\|\right)^2 \\
&\leq \mathbb{E}\left(\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0} L\left\|\mathcal{X} - X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right\|\right)^2 \\
&\leq L^2\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mathbb{E}\left\|\mathcal{X} - X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right\|^2 && \text{[SI 20]} \\
&\leq L^2\frac{1}{k_1}\sum_{i=1}^{k_1}\mathbb{E}\left\|\mathcal{X} - X^1_{(i,n)}(\mathcal{X})\right\|^2 && \text{[SI 21]} \\
&+ L^2\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mathbb{E}\left\|X^1_{(i,n)}(\mathcal{X}) - X^0_{(j,m)}(X^1_{(i,n)}(\mathcal{X}))\right\|^2 && \text{[SI 22]}
\end{aligned}$$

where [SI 20] follows from Jensen's inequality.

Let's consider [SI 21]. We partition the data into $A_1, \ldots, A_{k_1}$ sets, where the first $k_1 - 1$ sets have $\lfloor \frac{n}{k_1} \rfloor$ elements and we define $\tilde{X}_{i,1}(x)$ to be the nearest neighbor of x in A_i. Then we can conclude that

$$\begin{aligned}
\frac{1}{k_1}\sum_{i=1}^{k_1}\mathbb{E}\left\|\mathcal{X} - X^1_{(i,n)}(\mathcal{X})\right\|^2 &\leq \frac{1}{k_1}\sum_{i=1}^{k_1}\mathbb{E}\left\|\mathcal{X} - \tilde{X}_{i,1}(\mathcal{X})\right\|^2 \\
&= \frac{1}{k_1}\sum_{i=1}^{k_1}\mathbb{E}\left[\mathbb{E}\left[\left\|\mathcal{X} - \tilde{X}_{i,1}(\mathcal{X})\right\|^2 \middle| \mathcal{X}\right]\right] \leq \frac{\tilde{c}}{\lfloor \frac{n}{k_1} \rfloor^{2/d}}.
\end{aligned}$$

Here the last inequality follows from Lemma SI 6. With exactly the same argument, we can bound [SI 22] and we thus have

$$[SI18] \leq L^2\tilde{c} * \left(\frac{1}{\lfloor \frac{n}{k_1} \rfloor^{2/d}} + \frac{1}{\lfloor \frac{n_2}{k_2} \rfloor^{2/d}}\right) \leq 2\tilde{c}L^2 * \left(\left(\frac{k_1}{n}\right)^{2/d} + \left(\frac{k_0}{m}\right)^{2/d}\right).$$

Plugging everything in, we have

$$\begin{aligned}
\mathbb{E}[|\tau(\mathcal{X}) - \hat{\tau}_1^{mn}(\mathcal{X})|^2] &\leq 2\frac{\sigma^2}{k_1} + 2(c_2 + 2\tilde{c})L^2\left(\frac{k_1}{n}\right)^{2/d} + 2\frac{\sigma^2}{k_0} + 4\tilde{c}L^2\left(\frac{k_0}{m}\right)^{2/d} \\
&\leq C\left(\frac{\sigma^2}{k_1} + L^2\left(\frac{k_1}{n}\right)^{2/d} + \frac{\sigma^2}{k_0} + \left(\frac{k_0}{m}\right)^{2/d}\right)
\end{aligned}$$

with $C = 2\max(1, c_2 + 2\tilde{c}, 2\tilde{c})$. □

SI 10. Pseudocode

In this section, we present pseudocode for the algorithms in this paper. We denote by Y^0 and Y^1 the observed outcomes for the control group and the treatment group, respectively. For example, Y^1_i is the observed outcome of the ith unit in the treatment group. X^0 and X^1 are the features of the control units and the treated units, and hence X^1_i corresponds to the feature vector of the ith unit in the treatment group. $M_k(Y \sim X)$ is the notation for a regression estimator, which estimates $x \mapsto \mathbb{E}[Y|X = x]$. It can be any regression/machine learning estimator. In particular, it can be a black box algorithm.

Algorithm SI 1 T-learner

procedure T-LEARNER(X, Y, W)
  $\hat{\mu}_0 = M_0(Y^0 \sim X^0)$
  $\hat{\mu}_1 = M_1(Y^1 \sim X^1)$
  $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

M_0 and M_1 are here some, possibly different, machine-learning/regression algorithms.

Algorithm SI 2 S-learner

procedure S-LEARNER(X, Y, W)
  $\hat{\mu} = M(Y \sim (X, W))$
  $\hat{\tau}(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0)$

$M(Y \sim (X, W))$ is the notation for estimating $(x, w) \mapsto \mathbb{E}[Y|X = x, W = w]$ while treating W as a 0,1–valued feature.

Algorithm SI 3 X-learner

procedure X-LEARNER(X, Y, W, g)

  $\hat{\mu}_0 = M_1(Y^0 \sim X^0)$ ▷ Estimate response function
  $\hat{\mu}_1 = M_2(Y^1 \sim X^1)$

  $\tilde{D}_i^1 = Y_i^1 - \hat{\mu}_0(X_i^1)$ ▷ Compute imputed treatment effects
  $\tilde{D}_i^0 = \hat{\mu}_1(X_i^0) - Y_i^0$

  $\hat{\tau}_1 = M_3(\tilde{D}^1 \sim X^1)$ ▷ Estimate CATE in two ways
  $\hat{\tau}_0 = M_4(\tilde{D}^0 \sim X^0)$

  $\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$ ▷ Average the estimates

$g(x) \in [0, 1]$ is a weighting function that is chosen to minimize the variance of $\hat{\tau}(x)$. It is sometimes possible to estimate $\mathrm{Cov}(\tau_0(x), \tau_1(x))$, and compute the best g based on this estimate. However, we have made good experiences by choosing g to be an estimate of the propensity score.

Algorithm SI 4 F-learner

procedure F-LEARNER(X, Y, W)
  $\hat{e} = M_e[W \sim X]$
  $Y_i^* = Y_i \frac{W_i - \hat{e}(X_i)}{\hat{e}(X_i)(1-\hat{e}(X_i))}$
  $\hat{\tau} = M_\tau(Y^* \sim X)$

Algorithm SI 5 U-learner

procedure U-LEARNER(X, Y, W)
  $\hat{\mu}_{obs} = M_{obs}(Y^{obs} \sim X)$
  $\hat{e} = M_e[W \sim X]$
  $R_i = (Y_i - \hat{\mu}_{obs}(X_i))/(W_i - \hat{e}(X_i))$
  $\hat{\tau} = M_\tau(R \sim X)$

Algorithm SI 6 Bootstrap Confidence Intervals 1

procedure COMPUTECI(
x: features of the training data,
w: treatment assignments of the training data,
y: observed outcomes of the training data,
p: point of interest)
  $S_0 = \{i \ : \ w_i = 0\}$
  $S_1 = \{i \ : \ w_i = 1\}$
  $n_0 = \#S_0$
  $n_1 = \#S_1$
  for b in $\{1, \ldots, B\}$ **do**
    $s_b^* = c(\text{sample}(S_0, \text{replace} = \text{T}, \text{size} = n_0), \text{sample}(S_1, \text{replace} = \text{T}, \text{size} = n_1))$
    $x_b^* = x[s_b^*]$
    $w_b^* = w[s_b^*]$
    $y_b^* = y[s_b^*]$
    $\hat{\tau}_b^*(p) = \text{learner}(x_b^*, w_b^*, y_b^*)(p)$
  $\hat{\tau}(p) = \text{learner}(x, w, y)(p)$
  $\sigma = sd(\{\hat{\tau}_b^*(p)\}_{b=1}^B)$
  return $(\hat{\tau}(p) - q_{\alpha/2}\sigma, \hat{\tau}(p) + q_{1-\alpha/2}\sigma)$

For this pseudo code we use `R` notation. For example, $c()$ is here a function that combines its arguments to form a vector.

Algorithm SI 7 Bootstrap Confidence Intervals 2

procedure COMPUTECI(
x: features of the training data,
w: treatment assignments of the training data,
y: observed outcomes of the training data,
p: point of interest)
  $S_0 = \{i \ : \ w_i = 0\}$
  $S_1 = \{i \ : \ w_i = 1\}$
  $n_0 = \#S_0$
  $n_1 = \#S_1$
  for b in $\{1, \ldots, B\}$ **do**
    $s_b^* = c(\text{sample}(S_0, \text{replace} = \text{T}, \text{size} = n_0), \text{sample}(S_1, \text{replace} = \text{T}, \text{size} = n_1))$
    $x_b^* = x[s_b^*]$
    $w_b^* = w[s_b^*]$
    $y_b^* = y[s_b^*]$
    $\hat{\tau}_b^*(p) = \text{learner}(x_b^*, w_b^*, y_b^*)(p)$
  $\tilde{\tau}(p) = \frac{1}{B}\sum_{b=1}^B \hat{\tau}_b^*(p)$
  For all b in $\{1, \ldots, B\}$ and j in $\{1, \ldots, n\}$ define

$$S_{bj}^* = \#\{k : s_b^*[k] = j\}$$

  For all j in $\{1, \ldots, n\}$ define $\overline{S_{\cdot j}^*} = \frac{1}{B}\sum_{b=1}^B S_{bj}^*$ and

$$\text{Cov}_j = \frac{1}{B}\sum_{b=1}^{B}(\hat{\tau}_b^*(p) - \tilde{\tau}(p))(S_{bj}^* - \overline{S_{\cdot j}^*})$$

  $\sigma = \left(\sum_{j=1}^n Cov_j^2\right)^{0.5}$
  return $(\tilde{\tau}(p) - q_{\alpha/2}\sigma, \tilde{\tau}(p) + q_{1-\alpha/2}\sigma)$

This version of the bootstrap was proposed in (9).

Algorithm SI 8 Monte Carlos Bias Approximation

procedure APPROXIMATEBIAS(
x: features of the full data set,
w: treatment assignments of the full data set,
$y(0)$: potential outcome under control of the full data set,
$y(1)$: potential outcome under treatment of the full data set,
S: indices of observations that are not in the test set,
S_T: indices of the training set,
p: point of interest,
$\tau(p)$: the true CATE at p)

for i in $\{1, \ldots, 1000\}$ **do**

Create a new treatment assignment by permuting the original one,

$$w_i = \text{sample}(w, \text{replace} = F).$$

Define the observed outcome,

$$y_i = y(1)w_i + y(0)(1 - w_i).$$

Sample uniformly a training set of $50,000$ observations,

$$\begin{aligned} s_i^* &= \text{sample}(S, \text{replace} = F, \text{size} = 50,000), \\ w_i^* &= w_i[s_i^*], \\ x_i^* &= x[s_i^*], \\ y_i^* &= y_i[s_i^*]. \end{aligned}$$

Estimate the CATE,

$$\hat{\tau}_i^*(p) = \text{learner}(x_i^*, w_i^*, y_i^*)(p).$$

$\bar{\tau}^*(p) = \frac{1}{1000} \sum_{i=1}^{1000} \hat{\tau}_i^*(p)$

return $\bar{\tau}^*(p) - \tau(p)$

This algorithm is used to compute the bias in a simulation study where the potential outcomes and the CATE function are known. S, the indices of the units that are not in the test set and S_T, the indices of the units in the training set are not the same, because the training set is in this case a subset of 50,000 units of the full data set.

Algorithm SI 9 Bootstrap Bias

procedure ESTIMATEBIAS(
x: features of the training data,
w: treatment assignments of the training data,
y: observed outcomes of the training data,
p: point of interest)

$S_0 = \{i \ : \ w_i = 0\}$
$S_1 = \{i \ : \ w_i = 1\}$
$n_0 = \#S_0$
$n_1 = \#S_1$
for b in $\{1, \ldots, B\}$ **do**
$s_b^* = c(\text{sample}(S_0, \text{replace} = \text{T}, \text{size} = n_0), \text{sample}(S_1, \text{replace} = \text{T}, \text{size} = n_1))$
$x_b^* = x[s_b^*]$
$w_b^* = w[s_b^*]$
$y_b^* = y[s_b^*]$
$\hat{\tau}_b^*(p) = \text{learner}(x_b^*, w_b^*, y_b^*)(p)$
$\hat{\tau}(p) = \text{learner}(x, w, y)(p)$
$\bar{\tau}^*(p) = \frac{1}{B} \sum_{i=1}^{B} \hat{\tau}_i^*(p)$
return $\bar{\tau}^*(p) - \hat{\tau}(p)$

References

1. Hill JL (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1):217–240.
2. Künzel S, Tang A, Bickel P, Yu B, Sekhon J (2017) hte: An implementation of heterogeneous treatment effect estimators and honest random forests in c++ and r. *https://github.com/soerenkuenzel/hte (10/18/2017)*.
3. Wager S, Athey S (2017) Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* 113:1228–1242.
4. Lewandowski D, Kurowicka D, Joe H (2009) Generating random correlation matrices based on vines and extended onion method. *Journal of Multivariate Analysis* 100(9):1989–2001.
5. Heckman JJ, Smith J, Clements N (1997) Making the most out of programme evaluations and social experiments: Accounting for heterogeneity in programme impacts. *The Review of Economic Studies* 64(4):487–535.
6. Liu H, Yu B (2013) Asymptotic properties of lasso+mls and lasso+ridge in sparse high-dimensional linear regression. *Electronic Journal of Statistics* 7:3124–3169.
7. Dorie V, Hill J, Shalit U, Scott M, Cervone D (2017) Automated versus do-it-yourself methods for causal inference: Lessons learned from a data analysis competition. *arXiv preprint arXiv:1707.02641 (11/10/2017)*.
8. Putter H, Van Zwet WR (2012) Resampling: consistency of substitution estimators in *Selected Works of Willem van Zwet*. (Springer), pp. 245–266.
9. Efron B (2014) Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109(507):991–1007.
10. Györfi L, Kohler M, Krzyzak A, Walk H (2006) *A distribution-free theory of nonparametric regression*. (Springer Science & Business Media).
11. Bickel PJ, Doksum KA (2015) *Mathematical statistics: Basic ideas and selected topics*. (CRC Press) Vol. 2.