Meta-learners for Estimating Heterogeneous Treatment Effects using Machine Learning

Sören R. Künzel[a], Jasjeet S. Sekhon[a,b], Peter J. Bickel[a], and Bin Yu[a,c]

[a]Department of Statistics, University of California, Berkeley, CA 94720
[b]Department of Political Science, University of California, Berkeley, CA 94720
[c]Department of Electrical Engineering and Computer Science, University of California, Berkeley, CA 94720

April 25, 2019

Abstract

There is growing interest in estimating and analyzing heterogeneous treatment effects in experimental and observational studies. We describe a number of meta-algorithms that can take advantage of any supervised learning or regression method in machine learning and statistics to estimate the Conditional Average Treatment Effect (CATE) function. Meta-algorithms build on base algorithms—such as Random Forests (RF), Bayesian Additive Regression Trees (BART) or neural networks—to estimate the CATE, a function that the base algorithms are not designed to estimate directly. We introduce a new meta-algorithm, the X-learner, that is provably efficient when the number of units in one treatment group is much larger than in the other, and can exploit structural properties of the CATE function. For example, if the CATE function is linear and the response functions in treatment and control are Lipschitz continuous, the X-learner can still achieve the parametric rate under regularity conditions. We then introduce versions of the X-learner that use RF and BART as base learners. In extensive simulation studies, the X-learner performs favorably, although none of the meta-learners is uniformly the best. In two persuasion field experiments from political science, we demonstrate how our new X-learner can be used to target treatment regimes and to shed light on underlying mechanisms. A software package is provided that implements our methods.

With the rise of large data sets containing fine-grained information about humans and their behavior, researchers, businesses, and policymakers are increasingly interested in how treatment effects vary across individuals and contexts. They wish to go beyond the information provided by estimating the Average Treatment Effect (ATE) in randomized experiments and observational studies. Instead, they often seek to estimate the Conditional Average Treatment Effect (CATE) to personalize treatment regimes and to better understand causal mechanisms. We introduce a new estimator called the X-learner, and we characterize it and many other CATE estimators within a unified meta-learner framework. Their performance is compared using broad simulations, theory, and two data sets from randomized field experiments in political science.

In the first randomized experiment, we estimate the effect of a mailer on voter turnout (1) and, in the second, we measure the effect of door-to-door conversations on prejudice against gender-nonconforming individuals (2). In both experiments, the treatment effect is found to be non-constant, and we quantify this heterogeneity by estimating the CATE. We obtain insights into the underlying mechanisms, and the results allow us to better target the treatment.

To estimate the CATE, we build on regression or supervised learning methods in statistics and machine learning, which are successfully used in a wide range of applications. Specifically, we study meta-algorithms (or meta-learners) for estimating the CATE in a binary treatment setting. Meta-algorithms decompose estimating the CATE into several sub-regression problems that can be solved with any regression or supervised learning method.

The most common meta-algorithm for estimating heterogeneous treatment effects takes two steps. First, it uses so-called base learners to estimate the conditional expectations of the outcomes separately for units under control and those under treatment. Second, it takes the difference between these estimates. This approach has been analyzed when the base learners are linear regression (3) or tree-based methods (4). When used with trees, this has been called the *Two-Tree* estimator and we will

1

therefore refer to the general mechanism of estimating the response functions separately as the *T-learner*, "T" being short for "*two*."

Closely related to the T-learner is the idea of estimating the outcome using all of the features and the treatment indicator, without giving the treatment indicator a special role. The predicted CATE for an individual unit is then the difference between the predicted values when the treatment assignment indicator is changed from control to treatment, with all other features held fixed. This meta-algorithm has been studied with BART (5, 6) and regression trees (4) as the base learners. We refer to this meta-algorithm as the *S-learner*, since it uses a "*single*" estimator.

Not all methods that aim to capture the heterogeneity of treatment effects fall in the class of meta-algorithms. For example, some researchers analyze heterogeneity by estimating average treatment effects for meaningful subgroups (7). Another example is causal forests (8). Since causal forests are RF-based estimators, they can be compared to meta-learners with RFs in simulation studies. We will see that causal forests and the meta-learners used with RFs perform comparably well, but the meta-learners with other base learners can significantly outperform causal forests.

The main contribution of this paper is the introduction of a new meta-algorithm: the *X-learner*, which builds on the T-learner and uses each observation in the training set in an "X"–like shape. Suppose that we could observe the individual treatment effects directly. We could then estimate the CATE function by regressing the difference of individual treatment effects on the covariates. Structural knowledge about the CATE function (e.g., linearity, sparsity, or smoothness) could be taken into account by either picking a particular regression estimator for CATE or using an adaptive estimator that could learn these structural features. Obviously, we do not observe individual treatment effects because we observe the outcome either under control or under treatment, but never both. The X-learner uses the observed outcomes to estimate the unobserved individual treatment effects. It then estimates the CATE function in a second step as if the individual treatment effects were observed.

The X-learner has two key advantages over other estimators of the CATE. First, it can provably adapt to structural properties such as the sparsity or smoothness of the CATE. This is particularly useful since the CATE is often zero or approximately linear (9, 10). Secondly, it is particularly effective when the number of units in one treatment group (usually the control group) is much larger than in the other. This occurs because (control) outcomes and covariates are easy to obtain using data collected by administrative agencies, electronic medical record systems, or online platforms. This is the case in our first data example, where election turnout decisions in the U.S. are recorded by local election administrators for all registered individuals.

The rest of the paper is organized as follows. We start with a formal introduction of the meta-learners and provide intuitions for why we can expect the X-learner to perform well when the CATE is smoother than the response outcome functions and when the sample sizes between treatment and control are unequal. We then present the results of an extensive simulation study and provide advice for practitioners before we present theoretical results on the convergence rate for different meta-learners. Finally, we examine two field experiments using several meta-algorithms and illustrate how the X-learner can find useful heterogeneity with fewer observations.

Framework and Definitions

We employ the Neyman–Rubin potential outcome framework (11, 12), and assume a superpopulation or distribution $\mathcal{P}$ from which a realization of N independent random variables is given as the training data. That is, $(Y_i(0), Y_i(1), X_i, W_i) \sim \mathcal{P}$, where $X_i \in \mathbb{R}^d$ is a d-dimensional covariate or feature vector, $W_i \in \{0, 1\}$ is the treatment assignment indicator (to be defined precisely later), $Y_i(0) \in \mathbb{R}$ is the potential outcome of unit i when i is assigned to the control group, and $Y_i(1)$ is the potential outcome when i is assigned to the treatment group. With this definition, the Average Treatment Effect is defined as

$$\text{ATE} := \mathbb{E}[Y(1) - Y(0)].$$

It is also useful to define the response under control, μ_0, and the response under treatment, μ_1, as

$$\mu_0(x) := \mathbb{E}[Y(0)|X = x] \quad \text{and} \quad \mu_1(x) := \mathbb{E}[Y(1)|X = x].$$

Furthermore, we use the following representation of $\mathcal{P}$:

$$\begin{aligned} X &\sim \Lambda, \\ W &\sim \text{Bern}(e(X)), \\ Y(0) &= \mu_0(X) + \varepsilon(0), \\ Y(1) &= \mu_1(X) + \varepsilon(1), \end{aligned} \tag{1}$$

where Λ is the marginal distribution of X, $\varepsilon(0)$ and $\varepsilon(1)$ are zero-mean random variables and independent of X and W, and $e(x) = \mathbb{P}(W = 1|X = x)$ is the propensity score.

The fundamental problem of causal inference is that for each unit in the training data set, we observe either the potential outcome under control ($W_i = 0$), or the potential outcome under treatment ($W_i = 1$) but never both. Hence we denote the observed data as

$$\mathcal{D} = (Y_i, X_i, W_i)_{1 \leq i \leq N},$$

with $Y_i = Y_i(W_i)$. Note that the distribution of $\mathcal{D}$ is specified by $\mathcal{P}$. To avoid the problem that with a small but non-zero probability all units are under control or under treatment, we will analyze the behavior of different estimators conditional on the number of treated units. That is, for a fixed n with $0 < n < N$, we condition on the event that

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

This will enable us to state the performance of an estimator in terms of the number of treated units n and the number of control units $m = N - n$.

For a new unit i with covariate vector x_i, in order to decide whether to give the unit the treatment, we wish to estimate the Individual Treatment Effect (ITE) of unit i, D_i, which is defined as

$$D_i := Y_i(1) - Y_i(0).$$

However, we do not observe D_i for any unit, and D_i is not identifiable without strong additional assumptions in the sense that one can construct data-generating processes with the same distribution of the observed data, but a different D_i (Example 1). Instead, we will estimate the CATE function, which is defined as

$$\tau(x) := \mathbb{E}\Big[D\Big|X = x\Big] = \mathbb{E}\Big[Y(1) - Y(0)\Big|X = x\Big],$$

and we note that the best estimator for the CATE is also the best estimator for the ITE in terms of the MSE. To see that, let $\hat{\tau}_i$ be an estimator for D_i and decompose the MSE at x_i

$$\begin{aligned}
&\mathbb{E}\left[(D_i - \hat{\tau}_i)^2 | X_i = x_i\right] \\
&= \mathbb{E}\left[(D_i - \tau(x_i))^2 | X_i = x_i\right] + \mathbb{E}\left[(\tau(x_i) - \hat{\tau}_i)^2\right].
\end{aligned} \quad (2)$$

Since we cannot influence the first term in the last expression, the estimator that minimizes the MSE for the ITE of i also minimizes the MSE for the CATE at x_i.

In this paper, we are interested in estimators with a small Expected Mean Squared Error (EMSE) for estimating the CATE,

$$\text{EMSE}(\mathcal{P}, \hat{\tau}) = \mathbb{E}\left[(\tau(\mathcal{X}) - \hat{\tau}(\mathcal{X}))^2\right].$$

The expectation is here taken over $\hat{\tau}$ and $\mathcal{X} \sim \Lambda$, where $\mathcal{X}$ is independent of $\hat{\tau}$.

To aid our ability to estimate τ, we need to assume that there are no hidden confounders (13):

Condition 1

$$(\varepsilon(0), \varepsilon(1)) \perp W | X.$$

This assumption is, however, not sufficient to identify the CATE. One additional assumption that is often made to obtain identifiability of the CATE in the support of X is to assume that the propensity score is bounded away from 0 and 1:

Condition 2 *There exists $e_{\min}$ and $e_{\max}$ such that for all x in the support of X,*

$$0 < e_{\min} < e(x) < e_{\max} < 1.$$

Meta-algorithms

In this section, we formally define a meta-algorithm (or meta-learner) for the CATE as the result of combining supervised learning or regression estimators (i.e., base learners) in a specific manner while allowing the base learners to take any form. Meta-algorithms thus have the flexibility to appropriately leverage different sources of prior information in separate sub-problems of the CATE estimation problem: they can be chosen to fit a particular type of data, and they can directly take advantage of existing data analysis pipelines.

We first review both S- and T-learners, and we then propose the X-learner, which is a new meta-algorithm that can take advantage of unbalanced designs (i.e., the control or the treated group is much larger than the other group) and existing structures of the CATE (e.g., smoothness or sparsity). Obviously, flexibility is a gain only if the base learners in the meta-algorithm match the features of the data and the underlying model well.

The T-learner takes two steps. First, the control response function,

$$\mu_0(x) = \mathbb{E}[Y(0)|X = x],$$

is estimated by a base learner, which could be any supervised learning or regression estimator using the observations in the control group, $\{(X_i, Y_i)\}_{W_i=0}$. We denote the estimated function as $\hat{\mu}_0$. Second, we estimate the treatment response function,

$$\mu_1(x) = \mathbb{E}[Y(1)|X = x],$$

with a potentially different base learner, using the treated observations and denoting the estimator by $\hat{\mu}_1$. A T-learner is then obtained as

$$\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x). \quad (3)$$

Pseudocode for this T-learner can be found in Algorithm 1.

In the S-learner, the treatment indicator is included as a feature similar to all the other features without the

indicator being given any special role. We thus estimate the combined response function,

$$\mu(x, w) := \mathbb{E}[Y^{obs} | X = x, W = w],$$

using any base learner (supervised machine learning or regression algorithm) on the entire data set. We denote the estimator as $\hat{\mu}$. The CATE estimator is then given by

$$\hat{\tau}_S(x) = \hat{\mu}(x, 1) - \hat{\mu}(x, 0), \qquad (4)$$

and pseudocode is provided in Algorithm 2.

There are other meta-algorithms in the literature, but we do not discuss them here in detail because of limited space. For example, one may transform the outcomes so that any regression method can estimate the CATE directly (Algorithm 4) (4, 14, 15). In our simulations, this algorithm performs poorly, and we do not discuss it further, but it may do well in other settings.

X-learner

We propose the X-learner, and provide an illustrative example to highlight its motivations. The basic idea of the X-learner can be described in three stages:

1. Estimate the response functions

$$\mu_0(x) = \mathbb{E}[Y(0)|X = x], \text{ and} \qquad (5)$$
$$\mu_1(x) = \mathbb{E}[Y(1)|X = x], \qquad (6)$$

   using any supervised learning or regression algorithm and denote the estimated functions $\hat{\mu}_0$ and $\hat{\mu}_1$. The algorithms used are referred to as the base learners for the first stage.

2. Impute the treatment effects for the individuals in the treated group, based on the control outcome estimator, and the treatment effects for the individuals in the control group, based on the treatment outcome estimator, that is,

$$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_0(X_i^1), \text{ and} \qquad (7)$$
$$\tilde{D}_i^0 := \hat{\mu}_1(X_i^0) - Y_i^0, \qquad (8)$$

   and call these the imputed treatment effects. Note that if $\hat{\mu}_0 = \mu_0$ and $\hat{\mu}_1 = \mu_1$, then $\tau(x) = \mathbb{E}[\tilde{D}^1|X = x] = \mathbb{E}[\tilde{D}^0|X = x]$.

   Employ any supervised learning or regression method(s) to estimate $\tau(x)$ in two ways: using the imputed treatment effects as the response variable in the treatment group to obtain $\hat{\tau}_1(x)$, and similarly in the control group to obtain $\hat{\tau}_0(x)$. Call the supervised learning or regression algorithms base learners of the second stage.
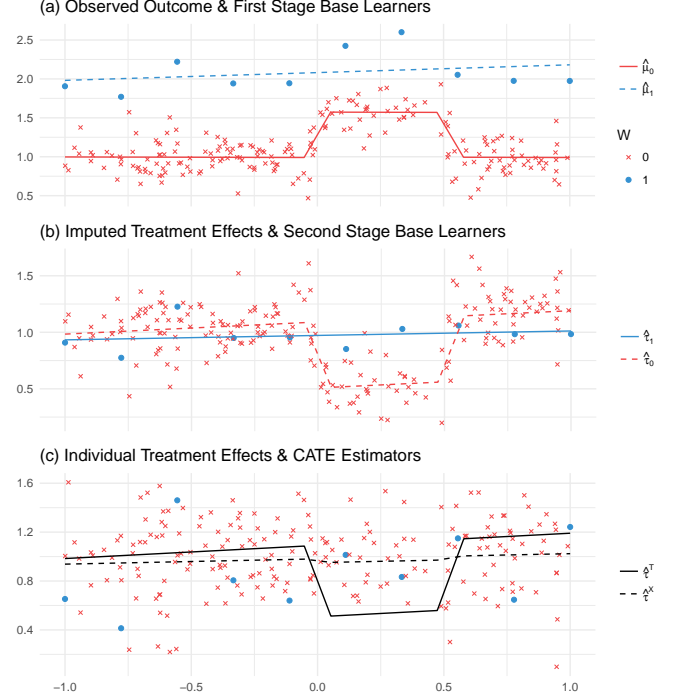

Figure 1: Intuition behind the X-learner with an unbalanced design.

3. Define the CATE estimate by a weighted average of the two estimates in Stage 2:

$$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x), \qquad (9)$$

   where $g \in [0, 1]$ is a weight function.

See Algorithm 3 for pseudocode.

Remark 1 *$\hat{\tau}_0$ and $\hat{\tau}_1$ are both estimators for τ, while g is chosen to combine these estimators to one improved estimator $\hat{\tau}$.*

Based on our experience, we observe that it is good to use an estimate of the propensity score for g, so that $g = \hat{e}$, but it also makes sense to choose $g = 1$ or 0, if the number of treated units is very large or small compared to the number of control units.

For some estimators, it might even be possible to estimate the covariance matrix of $\hat{\tau}_1$ and $\hat{\tau}_0$. One may then wish to choose g to minimize the variance of $\hat{\tau}$.

Intuition behind the meta-learners

The X-learner can use information from the control group to derive better estimators for the treatment group and vice versa. We will illustrate this using a simple example. Suppose that we want to study a treatment, and we are interested in estimating the CATE as a function of one covariate x. We observe, however, very few units in the

treatment group and many units in the control group. This situation often arises with the growth of administrative and online data sources: data on control units is often far more plentiful than data on treated units. Figure 1(a) shows the outcome for units in the treatment group (circles) and the outcome of unit in the untreated group (crosses). In this example, the CATE is constant and equal to one.

For the moment, let us look only at the treated outcome. When we estimate $\mu_1(x) = \mathbb{E}[Y(1)|X = x]$, we must be careful not to overfit the data since we observe only 10 data points. We might decide to use a linear model, $\hat{\mu}_1(x)$ (dashed line), to estimate μ_1. For the control group, we notice that observations with $x \in [0, 0.5]$ seem to be different, and we end up modeling $\mu_0(x) = \mathbb{E}[Y(0)|X = x]$ with a piecewise linear function with jumps at 0 and 0.5 (solid line). This is a relatively complex function, but we are not worried about overfitting since we observe many data points.

The *T-learner* would now use estimator $\hat{\tau}_T(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$ (see Figure 1(c), solid line), which is a relatively complicated function with jumps at 0 and 0.5, while the true $\tau(x)$ is a constant. This is, however, problematic because we are estimating a complex CATE function, based on ten observations in the treated group.

When choosing an estimator for the treatment group, we correctly avoided overfitting, and we found a good estimator for the treatment response function and, as a result, we chose a relatively complex estimator for the CATE, namely, the quantity of interest. We could have selected a piecewise linear function with jumps at 0 and 0.5, but this, of course, would have been unreasonable when looking only at the treated group. If, however, we were to also take the control group into account, this function would be a natural choice. In other words, we should change our objective for $\hat{\mu}_1$ and $\hat{\mu}_0$. We want to estimate $\hat{\mu}_1$ and $\hat{\mu}_0$ in such a way that their difference is a good estimator for τ.

The *X-learner* enables us to do exactly that. It allows us to use structural information about the CATE to make efficient use of an unbalanced design. The first stage of the X-learner is the same as the first stage of the T-learner, but in its second stage, the estimator for the controls is subtracted from the observed treated outcomes and similarly the observed control outcomes are subtracted from estimated treatment outcomes to obtain the imputed treatment effects,

$$\tilde{D}_i^1 := Y_i^1 - \hat{\mu}_0(X_i^1),$$
$$\tilde{D}_i^0 := \hat{\mu}_1(X_i^0) - Y_i^0.$$

Here we use the notation that Y_i^0 and Y_i^1 are the ith observed outcome of the control and the treated group, respectively. X_i^1, X_i^0 are the corresponding feature vectors. Figure 1(b) shows the imputed treatment effects, $\tilde{D}$.

By choosing a simple—here linear—function to estimate $\tau_1(x) = \mathbb{E}[\tilde{D}^1|X^1 = x]$ we effectively estimate a model for $\mu_1(x) = \mathbb{E}[Y^1|X^1 = x]$, which has a similar shape to $\hat{\mu}_0$. By choosing a relatively poor model for $\mu_1(x)$, $\tilde{D}^0$ (the red crosses in Figure 1(b)) are relatively far away from $\tau(x)$, which is constant and equal to 1. The model for $\tau_0(x) = \mathbb{E}[\tilde{D}^0|X = x]$ will thus be relatively poor. However, our final estimator combines these two estimators according to

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

If we choose $g(x) = \hat{e}(x)$, an estimator for the propensity score, $\hat{\tau}$ will be very similar to $\hat{\tau}_1(x)$, since we have many more observations in the control group; i.e., $\hat{e}(x)$ is small. Figure 1(c) shows the T-learner and the X-learner.

It is difficult to assess the general behavior of the S-learner in this example because we must choose a base learner. For example, when we use RF as the base learner for this data set, the S-learner's first split is on the treatment indicator in 97.5% of all trees in our simulations because the treatment assignment is very predictive of the observed outcome, Y (see also Figure 14). From there on, the S-learner and the T-learner are the same, and we observe them to perform similarly poorly in this example.

Simulation Results

In this section, we conduct a broad simulation study to compare the different meta-learners. In particular, we summarize our findings and provide general remarks on the strengths and weaknesses of the S-, T-, and X-learners, while deferring the details to the Supporting Information (SI). The simulations are key to providing an understanding of the performance of the methods we consider for model classes that are not covered by our theoretical results.

Our simulation study is designed to consider a range of situations. We include conditions under which the S-learner or the T-learner is likely to perform the best, as well as simulation setups proposed by previous researchers (8). We consider cases where the treatment effect is zero for all units (and so pooling the treatment and control groups would be beneficial), and cases where the treatment and control response functions are completely different (and so pooling would be harmful). We consider cases with and without confounding,[1] and cases with equal and unequal sample sizes across treatment conditions. All simulations discussed in this section are based on synthetic data. For details, please see Section A. We

---

[1] Confounding here refers to the existence of an unobserved covariate that influences both the treatment variable, W, and at least one of the portential outcomes $Y(0), Y(1)$.

provide additional simulations based on actual data when we discuss our applications.

We compare the S-, T-, and X-learners with RF and BART as base learners. We implement a version of RF for which the tree structure is independent of the leaf prediction given the observed features, the so-called honest RF in an R package called `hte` (16). This version of RF is particularly accessible from a theoretical point of view, it performs well in noisy settings, and it is better suited for inference (8, 17). For BART, our software uses the `dbarts` (18) implementation for the base learner.

Comparing different base learners enables us to demonstrate two things. On the one hand, it shows that the conclusions we draw about the S-, T-, and X-learner are not specific to a particular base learner and, on the other hand, it demonstrates that the choice of base learners can make a large difference in prediction accuracy. The latter is an important advantage of meta-learners since subject knowledge can be used to choose base learners that perform well. For example, in Simulations 2 and 4 the response functions are globally linear, and we observe that estimators that act globally such as BART have a significant advantage in these situations or when the data set is small. If, however, there is no global structure or when the data set is large, then more local estimators such as RF seem to have an advantage (Simulations 3 and 5).

We observe that the choice of meta-learner can make a large difference, and for each meta-learner there exist cases where it is the best-performing estimator.

The S-learner treats the treatment indicator like any other predictor. For some base learners such as k-nearest neighbors it is not a sensible estimator, but for others it can perform well. Since the treatment indicator is given no special role, algorithms such as the lasso and RFs can completely ignore the treatment assignment by not choosing/splitting on it. This is beneficial if the CATE is in many places 0 (Simulations 4 and 5), but—as we will see in our second data example—the S-learner can be biased toward 0.

The T-learner, on the other hand, does not combine the treated and control groups. This can be a disadvantage when the treatment effect is simple because by not pooling the data, it is more difficult for the T-learner to mimic a behavior that appears in both the control and treatment response functions (e.g., Simulation 4). If, however, the treatment effect is very complicated, and there are no common trends in μ_0 and μ_1, then the T-learner performs especially well (Simulations 2 and 3).

The X-learner performs particularly well when there are structural assumptions on the CATE or when one of the treatment groups is much larger than the other (Simulation 1 and 3). In the case where the CATE is 0, it usually does not perform as well as the S-learner, but it is significantly better than the T-learner (Simulations 4,

5, and 6), and in the case of a very complex CATE, it performs better than the S-learner and it often outperforms even the T-learner (Simulations 2 and 3). These simulation results lead us to the conclusion that unless one has a strong belief that the CATE is mostly 0, then, as a rule of thumb, one should use the X-learner with BART for small data sets and RF for bigger ones. In the sequel, we will further support these claims with additional theoretical results and empirical evidence from real data and data-inspired simulations.

Comparison of Convergence Rates

In this section, we provide conditions under which the X-learner can be proven to outperform the T-learner in terms of pointwise estimation rate. These results can be viewed as attempts to rigorously formulate intuitions regarding when the X-learner is desirable. They corroborate our intuition that the X-learner outperforms the T-learner when one group is much larger than the other group or when the CATE function has a simpler form than those of the underlying response functions themselves.

Let us start by reviewing some of the basic results in the field of minimax nonparametric regression estimation (19, 20, 21, 23). In the standard regression problem, one observes N independent and identically distributed tuples $(X_i, Y_i)_i \in \mathbb{R}^{d \times N} \times \mathbb{R}^N$ generated from some distribution $\mathcal{P}$ and one is interested in estimating the conditional expectation of Y given some feature vector x, $\mu(x) = \mathbb{E}[Y|X = x]$. The error of an estimator $\hat{\mu}_N$ can be evaluated by the Expected Mean Squared Error (EMSE),

$$\text{EMSE}(\mathcal{P}, \hat{\mu}_N) = \mathbb{E}[(\hat{\mu}_N(\mathcal{X}) - \mu(\mathcal{X}))^2].$$

For a fixed $\mathcal{P}$, there are always estimators which have a very small EMSE. For example, choosing $\hat{\mu}_N \equiv \mu$ would have no error. However, $\mathcal{P}$ and thus μ is unknown. Instead, one usually wants to find an estimator which achieves a small EMSE for a relevant set of distributions (such a set is relevant if it captures domain knowledge or prior information of the problem). To make this problem feasible, a typical approach is the minimax approach where one analyzes the worst performance of an estimator over a class or family, F, of distributions (22). The goal is to find an estimator which has a small EMSE for all distributions in this family. For example, if F_0 is the family of distributions $\mathcal{P}$ such that $X \sim \text{Unif}[0, 1]$, $Y = \beta X + \varepsilon$, $\varepsilon \sim N(0, 1)$, and $\beta \in \mathbb{R}$, then it is well known that the OLS estimator achieves the optimal parametric rate. That is, there exists a constant $C \in \mathbb{R}$ such that for all $\mathcal{P} \in F_0$,

$$\text{EMSE}(\mathcal{P}, \hat{\mu}_N^{\text{OLS}}) \le CN^{-1}.$$

If, however, F_1 is the family of all distributions $\mathcal{P}$ such that $X \sim \text{Unif}[0,1]$, $Y \sim \mu(X) + \varepsilon$ and μ is a Lipschitz continuous function with bounded Lipschitz constant, then there exists no estimator that achieves the parametric rate uniformly for all possible distributions in F_1. To be precise, we can at most expect to find an estimator that achieves a rate of $N^{-2/3}$ and there exists a constant C', such that

$$\liminf_{N \to \infty} \inf_{\hat{\mu}_N} \sup_{\mathcal{P} \in F_1} \frac{\text{EMSE}(\mathcal{P}, \hat{\mu}_N)}{N^{-2/3}} > C' > 0.$$

Estimators such as the Nadaraya–Watson and k-nearest neighbors can achieve this optimal rate (21, 23).

Crucially, the fastest rate of convergence that holds uniformly for a family F is a property of the family to which the underlying data-generating distribution belongs. It will be useful for us to define sets of families for which particular rates are achieved.

Definition 1 (Families with bounded minimax rate)
For $a \in (0,1]$, we define $S(a)$ to be the set of all families, F, with a minimax rate of at most N^{-a}.

Note that for any family $F \in S(a)$ there exists an estimator $\hat{\mu}$ and a constant C such that for all $N \geq 1$,

$$\sup_{\mathcal{P} \in F} \text{EMSE}(\mathcal{P}, \hat{\mu}_N) \leq CN^{-a}.$$

From the examples above, it is clear that $F_0 \in S(1)$ and $F_1 \in S(2/3)$.

Even though the minimax rate of the EMSE is not very practical since one rarely knows that the true data-generating process is in some reasonable family of distributions, it is nevertheless one of the very few useful theoretical tools to compare different nonparametric estimators. If for a big class of distributions, the worst EMSE of an estimator $\hat{\mu}^A$ is smaller than the worst EMSE of $\hat{\mu}^B$, then one might prefer estimator $\hat{\mu}^A$ over estimator $\hat{\mu}^B$. Furthermore, if the estimator of choice does not have a small error for a family that we believe based on domain information could be relevant in practice, then we might expect $\hat{\mu}$ to have a large EMSE in real data.

Implication for CATE estimation

Let us now employ the minimax approach to the problem of estimating the CATE. Recall that we assume a super–population of random variables $(Y(0), Y(1), X, W)$ according to some distribution $\mathcal{P}$. We observe n treated and m control units from this super-population, and our goal is to find an estimator $\hat{\tau}_{mn}$ which has a small EMSE,

$$\text{EMSE}(\mathcal{P}, \hat{\tau}_{mn}) = \mathbb{E}[(\tau(\mathcal{X}) - \hat{\tau}_{mn}(\mathcal{X}))^2].$$

Similar to the regression case, we can study the performance of estimators when $\mathcal{P}$ lies in some family of distributions. In the following we will introduce families for

which estimators based on the X-learner achieve provably a smaller EMSE than estimators based on the T-learner.

Similar to Definition 1, we define sets of families of super-populations.

Definition 2 (Superpopulations with given rates)
Recall the general characterization of a superpopulation given in 1. For $a_\mu, a_\tau \in (0,1]$, we define $S(a_\mu, a_\tau)$ to be the set of all families of distributions $\mathcal{P}$ of $(Y(0), Y(1), X, W)$ such that

1. *ignorability holds,*

2. *the distribution of $(X, Y(0))$ given $W = 0$ is in a class $F_0 \in S(a_\mu)$,*

3. *the distribution of $(X, Y(1))$ given $W = 1$ is in a class $F_1 \in S(a_\mu)$, and*

4. *the distribution of $(X, \mu_1(X) - Y(0))$ given $W = 0$ is in a class $F_{\tau 0} \in S(a_\tau)$.*

5. *the distribution of $(X, Y(1) - \mu_0(X))$ given $W = 1$ is in a class $F_{\tau 1} \in S(a_\tau)$.*

A simple example of a family in $S(2/3, 1)$, would be the set of distributions $\mathcal{P}$ for which $X \sim \text{Unif}[0,1]$, $W \sim \text{Bern}(1/2)$, μ_0 is any Lipschitz continuous function, τ is linear, and $\varepsilon(0), \varepsilon(1) \sim N(0,1)$.

The difference between the T and X learner is that the T-learner estimates the response functions separately, and does not benefit from the possible smoothness of the CATE. Hence, we can conclude the following theorem.

Theorem 1 (Minimax rates of the T-learner) *For $\mathcal{P} \in S(a_\mu, a_\tau)$, there exist base learners to be used in the T-learner so that the corresponding T-learner estimates the CATE at a rate of*

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

but, in general, we cannot expect it to be any faster.

The X-learner, on the other hand, can be seen as a weighted average of the two estimators, $\hat{\tau}_0$ and $\hat{\tau}_1$ (Eq. 9). Take for the moment, $\hat{\tau}_1$. It consists of an estimator for the outcome under control which achieves a rate of at most a_μ, and an estimator for the imputed treatment effects which should intuitively achieve a rate of at most a_τ. We therefore expect the following conjecture.

Conjecture 1 (Minimax rates of the X-learner)
Under some conditions on $\mathcal{P} \in S(a_\mu, a_\tau)$, there exist base learners such that $\hat{\tau}_0$ and $\hat{\tau}_1$ in the X-learner achieve the rates,

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

respectively.

It turns out to be mathematically very challenging to give a satisfying statement of the extra conditions needed on $\mathcal{P}$. We will therefore discuss two cases where we do not need any extra conditions on $\mathcal{P}$, and we emphasize that we believe the conjecture to hold in much greater generality. In the first case (Theorem 2), we discuss all families of distributions in $S(a_\mu, a_\tau)$ where the CATE is linear. This implies that $a_\tau = 1$, and we achieve the parametric rate in n. This is in particular important when the number of control units, m, is large. In Section I.3, we discuss the other extreme where we don't have any assumptions on the CATE. In this case, there is nothing to be inferred from the control units about the treated units and vice versa. Consequently, the T-learner is in some sense the best strategy and achieves the minimax optimal rate of $\mathcal{O}(m^{-a_\mu} + n^{-a_\mu})$ and we show that, for example, under Lipschitz continuity of the response functions, the X-learner will achieve the same rate and is therefore minimax optimal as well.

We also conduct an extensive simulation study (Section A) in which we compare the different meta-learners combined with Random Forests and BART for many different situations. We find that neither learner is uniformly the best, but the X-learner is never the worst, and it performs particularly well, when the group sizes are very unbalanced, or the CATE function satisfies some regularity conditions.

Smoothness conditions for the CATE

Even though it is theoretically possible that a_τ is similar to a_μ, our experience with real data suggests that it is often larger (or the treatment effect is *simpler* to estimate than the potential outcomes). Let us intuitively understand the difference between the T- and X-learners for a class $F \in S(a_\mu, a_\tau)$ with $a_\tau > a_\mu$. The T-learner splits the problem of estimating the CATE into the two subproblems of estimating μ_0 and μ_1 separately. By appropriately choosing the base learners, we can expect to achieve the minimax optimal rates of a_μ,

$$\sup_{\mathcal{P}_0 \in F_0} \text{EMSE}(\mathcal{P}_0, \hat{\mu}_0^m) \le Cm^{-a_\mu}, \quad \text{and}$$
$$\sup_{\mathcal{P}_1 \in F_1} \text{EMSE}(\mathcal{P}_1, \hat{\mu}_1^n) \le Cn^{-a_\mu}, \quad (10)$$

where C is some constant. Those rates translate immediately to rates for estimating τ, since

$$\sup_{\mathcal{P} \in F} \text{EMSE}(\mathcal{P}, \hat{\tau}_{nm}^T)$$
$$\le 2 \sup_{\mathcal{P}_0 \in F_0} \text{EMSE}(\mathcal{P}_0, \hat{\mu}_0^m) + 2 \sup_{\mathcal{P}_1 \in F_1} \text{EMSE}(\mathcal{P}_1, \hat{\mu}_1^n)$$
$$= 2C \left(m^{-a_\mu} + n^{-a_\mu} \right).$$

In general, we cannot expect to do better than this, when using an estimation strategy that falls in the class of T-learners, because the subproblems in Equation 10 are treated completely independently and there is nothing to be learned from the treatment group about the control group and vice versa.

In some cases, we observe that the number of control units is much larger than the number of treated units, $m \gg n$. This happens for example if we test a new treatment and we have a large number of previous (untreated) observations that can be used as the control group. In that case, the bound on the EMSE of the T-learner for the CATE will be dominated by the regression problem for the treated group,

$$\sup_{\mathcal{P} \in F} \text{EMSE}(\mathcal{P}, \hat{\tau}_{nm}^T) = \sup_{\mathcal{P}_1 \in F_1} \text{EMSE}(\mathcal{P}_1, \hat{\mu}_1^n) \le Cn^{-a_\mu}. \quad (11)$$

This is an improvement, but it still does not lead to the fast rate, a_τ. The X-learner, however, can achieve the fast rate a_τ. An expansion of the EMSE into two squared error terms and also a cross term involving biases can be used to show that the T-learner cannot achieve this fast rate in general in the unbalanced case of $m >> n$. To see the faster rate for the X-learner, recall that the number of control units is assumed so large that μ_0 can be predicted almost perfectly and choose the weighing function g equal to 0 in Equation 9. It follows that the error of the first stage of the X-learner is negligible and the imputed treatment effects for the treated group satisfy $D_i^1 = \tau(X_i(1)) + \varepsilon_i$. Per Assumption 5 in Definition 2, $\mathbb{E}[D^1 | X = x]$ can now be estimated using an estimator achieving the desired rate of a_τ,

$$\sup_{\mathcal{P} \in F} \text{EMSE}(\mathcal{P}, \hat{\tau}_{nm}^X) \le Cn^{-a_\tau}.$$

This is a substantial improvement over 11 when $a_\tau > a_\mu$ and intuitively demonstrates that, in contrast to the T-learner, the X-learner can exploit structural assumptions on the treatment effect. However, even for large m, we cannot expect to perfectly estimate μ_0. The following theorem deals carefully with this estimation error when τ is linear, but the response functions can be estimated at any nonparametric rate.

Theorem 2 *Assume we observe m control units and n treated units from some super population of independent and identically distributed observations $(Y(0), Y(1), X, W)$ coming from a distribution $\mathcal{P}$ given in 1 that satisfies the following assumptions:*

A1 The error terms ε_i are independent given X, with $\mathbb{E}[\varepsilon_i | X = x] = 0$ and $Var[\varepsilon_i | X = x] \le \sigma^2$.

A2 X has finite second moments,

$$\mathbb{E}[\|X\|_2^2] \le C_X.$$

A3 Ignorability holds.

A4 There exists an estimator $\hat{\mu}_0^m$ and $a > 0$ with

$$EMSE(\mathcal{P}, \hat{\mu}_0^m) = \mathbb{E}[(\mu_0(X) - \hat{\mu}_0^m(X))^2] \leq C_0 m^{-a}.$$

A5 The treatment effect is parametrically linear, $\tau(x) = x^T\beta$, with $\beta \in \mathbb{R}^d$.

A6 The eigenvalues of the sample covariance matrix of X^1 are well conditioned, in the sense that there exists an n_0, such that

$$\sup_{n > n_0} \gamma_{min}^{-1}(\hat{\Sigma}_n) \leq C_\Sigma. \tag{12}$$

Then the X-learner with $\hat{\mu}_0^m$ in the first stage, OLS in the second stage and weigh function $g \equiv 0$ has the following upper bound on its EMSE: for all $n > n_0$,

$$EMSE(\mathcal{P}, \hat{\tau}^{mn}) = \mathbb{E}\left[\|\tau(X) - \hat{\tau}^{mn}(X)\|^2\right] \leq C(m^{-a} + n^{-1})$$

with $C = \max(C_0, \sigma^2 d)C_X C_\Sigma$. In particular, if there are a lot of control units, such that $m \geq c_3 n^{1/a}$, then the X-learner achieves the parametric rate in n,

$$EMSE(\mathcal{P}, \hat{\tau}^{mn}) \leq (1 + c_3)Cn^{-1}.$$

It is symmetric that similar results hold if n (the size of the treatment group) is much larger than m (the size of the control group). Furthermore, we note that an equivalent statement also holds for the pointwise MSE (Theorem 4).

Applications

In this section, we consider two data examples. In the first example, we consider a large Get-Out-The-Vote (GOTV) experiment that explored if social pressure can be used to increase voter turnout in elections in the United States (1). In the second example, we consider an experiment that explored if door-to-door canvassing can be used to durably reduce transphobia in Miami (2). In both examples, the original authors failed to find evidence of heterogeneous treatment effects when using simple linear models without basis expansion, and subsequent researchers and policy makers have been acutely interested in treatment effect heterogeneity that could be used to better target the interventions. We use our honest random forest implementation (16) because of the importance of obtaining useful confidence intervals in these applications. Confidence intervals are obtained using a bootstrap procedure (Algorithm 6). We have evaluated several bootstrap procedures, and we have found that the results for all of them were very similar. We explain this particular bootstrap choice in detail in SI.3.

Social pressure and voter turnout

In a large field experiment, Gerber et al. show that substantially higher turnout was observed among registered voters who received a mailing promising to publicize their turnout to their neighbors (1). In the United States, whether someone is registered to vote and their past voting turnout are a matter of public record. Of course, *how* individuals voted is private. The experiment has been highly influential both in the scholarly literature and in political practice. In our reanalysis, we focus on two treatment conditions: the control group, which was assigned to 191,243 individuals, and the "neighbor's" treatment group, which was assigned to 38,218 individuals. Note the unequal sample sizes. The experiment was conducted in Michigan before the August 2006 primary election, which was a statewide election with a wide range of offices and proposals on the ballot. The authors randomly assigned households with registered voters to receive mailers. The outcome, whether someone voted, was observed in the primary election. The "neighbors" mailing opens with a message that states "DO YOUR CIVIC DUTY—VOTE!" It then continues by not only listing the household's voting records but also the voting records of those living nearby. The mailer informed individuals that "we intend to mail an updated chart" after the primary.

The study consists of seven key individual-level covariates, most of which are discrete: gender, age, and whether the registered individual voted in the primary elections in 2000, 2002, and 2004 or the general election in 2000 and 2002. The sample was restricted to voters who had voted in the 2004 general election. The outcome of interest is turnout in the 2006 primary election, which is an indicator variable. Because compliance is not observed, all estimates are of the Intention-to-Treat (ITT) effect, which is identified by the randomization. The average treatment effect estimated by the authors is 0.081 with a standard error of (0.003). Increasing voter turnout by 8.1% using a simple mailer is a substantive effect, especially considering that many individuals may never have seen the mailer.

Figure 2 presents the estimated treatment effects, using X–RF where the potential voters are grouped by their voting history. The upper panel of the figure shows the proportion of voters with a significant positive (blue) and a significant negative (red) CATE estimate. We can see that there is evidence of a negative backlash among a small number of people who voted only once in the past five elections prior to the general election in 2004. Applied researchers have observed a backlash from these mailers; e.g., some recipients called their Secretary of States office or local election registrar to complain (26, 27). The lower panel shows the distribution of CATE estimates for
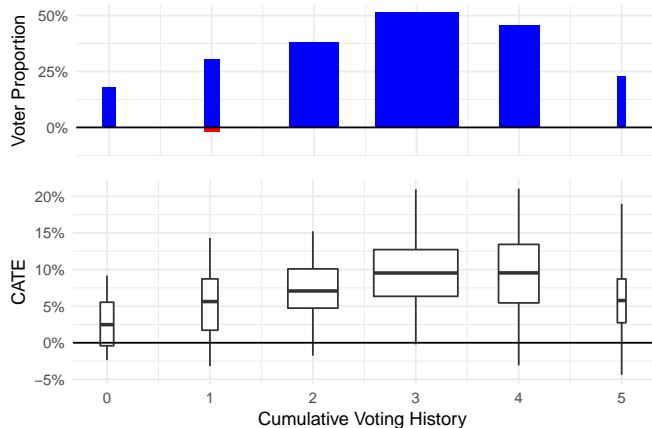
Figure 2: Social pressure and voter turnout. Potential voters are grouped by the number of elections they participated in, ranging from 0 (potential voters who did not vote during the past five elections) to 5 (voters who participated in all five past elections). The width of each group is proportional to the size of the group. Positive values in the first plot correspond to the percentage of voters for which the predicted CATE is significantly positive, while negative values correspond to the percentage of voters for which the predicted CATE is significantly negative. The second plot shows the CATE estimate distribution for each bin.

each of the subgroups. Having estimates of the heterogeneity enables campaigns to better target the mailers in the future. For example, if the number of mailers is limited, one should target potential voters who voted three times during the past five elections, since this group has the highest average treatment effect and it is a very big group of potential voters.[2]

S–RF, T–RF, and X–RF all provide similar CATE estimates. This is unsurprising given the very large sample size, the small number of covariates, and their distributions. For example, the correlation between the CATE estimates of S–RF and T–RF is 0.99 (results for S–RF and T–RF can be found in Figure 13).

We conducted a data-inspired simulation study to see how these estimators would behave in smaller samples. We take the CATE estimates produced by T–RF, and we assume that they are the truth. We can then impute the potential outcomes under both treatment and control for every observation. We then sample training data from

---

[2]In praxis, it is not necessary to identify a particular subgroup. Instead, one can simply target units for which the predicted CATE is large. If the goal of our analysis were to find subgroups with different treatment effects, one should validate those subgroup estimates. We suggest either splitting the data and letting the X-learner use part of the data to find subgroups and the other part to validate the subgroup estimates, or to use the suggested subgroups to conduct further experiments.
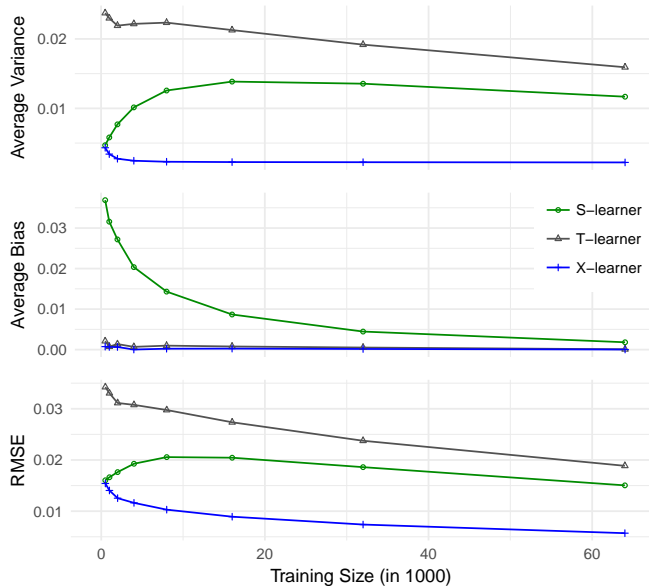

Figure 3: RMSE, bias, and variance for a simulation based on the social pressure and voter turnout experiment.

the complete data and predict the CATE estimates for the test data using, S–, T–, and X–RF. We keep the unequal treatment proportion observed in the full data fixed, i.e., $\mathbb{P}(W = 1) = 0.167$. Figure 3 presents the results of this simulation. They show that in small samples both X–RF and S–RF outperform T–RF, with X–RF performing the best, as one may conjecture given the unequal sample sizes.

Reducing transphobia: A field experiment on door-to-door canvassing

In an experiment that received widespread media attention, Broockman et al. show that brief (10 minutes) but high-quality door-to-door conversations can markedly reduce prejudice against gender-nonconforming individuals for at least three months (2). This experiment was published in *Science* after the journal retracted an earlier article claiming to show the same in an experiment about gay rights (28). Broockman et al. showed that the earlier published study was fraudulent, and they conducted the new one to determine if the pioneering behavioral intervention of encouraging people to actively take the perspective of others was effective in decreasing prejudice (29).

There are important methodological differences between this example and our previous one. The experiment is a placebo-controlled experiment with a parallel survey that measures attitudes, which are the outcomes of interest. The authors follow the design of (30). The authors

first recruited registered voters ($n = 68,378$) via mail for an unrelated online survey to measure baseline outcomes. They then randomly assigned respondents of the baseline survey to either the treatment group ($n = 913$) or the placebo group that was targeted with a conversation about recycling ($n = 912$). Randomization was conducted at the household level ($n = 1295$), and because the design employs a placebo control, the estimand of interest is the complier-average-treatment effect. Outcomes were measured by the online survey three days, three weeks, six weeks, and three months after the door-to-door conversations. We analyze results for the first follow-up.

The final experimental sample consists of only 501 observations. The experiment was well powered despite its small sample size because it includes a baseline survey of respondents as well as post-treatment surveys. The survey questions were designed to have high over-time stability. The R^2 of regressing the outcomes of the placebo-control group on baseline covariates using OLS is 0.77. Therefore, covariate adjustment greatly reduces sampling variation. There are 26 baseline covariates that include basic demographics (gender, age, ethnicity) and baseline measures of political and social attitudes and opinions about prejudice in general and Miami's nondiscrimination law in particular.

The authors find an average treatment effect of 0.22 (SE: 0.072, t-stat: 3.1) on their transgender tolerance scale.[3] The scale is coded so that a larger number implies greater tolerance. The variance of the scale is 1.14, with a minimum observed value of -2.3 and a maximum of 2. This is a large effect given the scale. For example, the estimated decrease in transgender prejudice is greater than Americans' average decrease in homophobia from 1998 to 2012, when both are measured as changes in standard deviations of their respective scales.

The authors report finding no evidence of heterogeneity in the treatment effect that can be explained by the observed covariates. Their analysis is based on linear models (OLS, lasso, and elastic net) without basis expansions.[4] Figure 4(a) presents our results for estimating the CATE, using X–RF. We find that there is strong evidence that the positive effect that the authors find is only found among a subset of respondents that can be targeted based on observed covariates. The average of our CATE estimates is within half a standard deviation of the ATE that the authors report.

Unlike in our previous data example, there are marked differences in the treatment effects estimated by our three learners. Figure 4(b) presents the estimates from T–RF.

---

[3]The authors' transgender tolerance scale is the first principal component of combining five -3 to $+3$ Likert scales. See (2) for details.

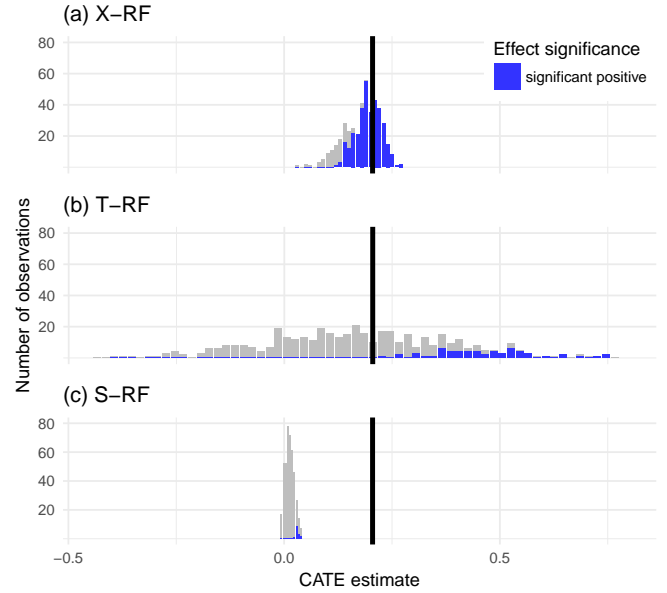[4](2) estimates the CATE using Algorithm 4.

Figure 4: Histograms for the distribution of the CATE estimates in the Reducing Transphobia study. The horizontal line shows the position of the estimated ATE.

These estimates are similar to those of X–RF, but with a larger spread. Figure 4(c) presents the estimates from S–RF. Note that the average CATE estimate of S–RF is much lower than the ATE reported by the original authors and the average CATE estimates of the other two learners. Almost none of the CATE estimates are significantly different from zero. Recall that the ATE in the experiment was estimated with precision, and was large both substantively and statistically (t-stat=3.1).

In this data, S–RF shrinks the treatment estimates toward zero. The ordering of the estimates we see in this data application is what we have often observed in simulations: the S-learner has the least spread around zero, the T-learner has the largest spread, and the X-learner is somewhere in between. Unlike in the previous example, the covariates are strongly predictive of the outcomes, and the splits in the S–RF are mostly on the features rather than the treatment indicator, because they are more predictive of the observed outcomes than the treatment assignment (cf., Figure 14).

Conclusion

This paper reviewed meta-algorithms for CATE estimation including the S- and T-learners. It then introduced a new meta-algorithm, the X-learner, that can translate any supervised learning or regression algorithm or a combination of such algorithms into a CATE estimator. The X-learner is adaptive to various settings. For example,

both theory and data examples show that it performs particularly well when one of the treatment groups is much larger than the other or when the separate parts of the X-learner are able to exploit the structural properties of the response and treatment effect functions. Specifically, if the CATE function is linear, but the response functions in the treatment and control group satisfy only the Lipschitz-continuity condition, the X-learner can still achieve the parametric rate if one of the groups is much larger than the other (Theorem 2). If there are no regularity conditions on the CATE function and the response functions are Lipschitz continuous, then both the X-learner and the T-learner obtain the same minimax optimal rate (Theorem 7). We conjecture that these results hold for more general model classes than those in our theorems.

We have presented a broad set of simulations to understand the finite sample behaviors of different implementations of these learners, especially for model classes that are not covered by our theoretical results. We have also examined two data applications. Although none of the meta-algorithms is always the best, the X-learner performs well overall, especially in the real-data examples. In practice, in finite samples, there will always be gains to be had if one accurately judges the underlying data-generating process. For example, if the treatment effect is simple, or even zero, then pooling the data across treatment and control conditions will be beneficial when estimating the response model (i.e., the S-learner will perform well). However, if the treatment effect is strongly heterogeneous and the response surfaces of the outcomes under treatment and control are very different, pooling the data will lead to worse finite sample performance (i.e., the T-learner will perform well). Other situations are possible and lead to different preferred estimators. For example, one could slightly change the S-learner so that it shrinks to the estimated ATE instead of zero, and it would then be preferred when the treatment effect is constant and non-zero. One hopes that the X-learner can adapt to these different settings. The simulations and real-data studies presented have demonstrated the X-learner's adaptivity. However, further studies and experience with more real data sets are necessary. To enable practitioners to benchmark these learners on their own data sets, we have created an easy-to-use software library called `hte`. It implements several methods of selecting the best CATE estimator for a particular data set, and it implements confidence interval estimators for the CATE.

In ongoing research, we are investigating using other supervised learning algorithms. For example, we are creating a deep learning architecture for estimating the CATE that is based on the X-learner with a particular focus on transferring information between different data sets and treatment groups. Furthermore, we are concerned with finding better confidence intervals for the CATE. This might enable practitioners to better design experiments, and determine the required sample size before an experiment is conducted.

Acknowledgement

References

[1] Gerber AS, Green DP, Larimer CW (2008) Social pressure and voter turnout: Evidence from a large-scale field experiment. *American Political Science Review* 102(1):33–48.

[2] Broockman D, Kalla J (2016) Durably reducing transphobia: A field experiment on door-to-door canvassing. *Science* 352(6282):220–224.

[3] Foster JC (2013) Ph.D. thesis (The University of Michigan).

[4] Athey S, Imbens GW (2015) Machine learning methods for estimating heterogeneous causal effects. *stat* 1050(5).

[5] Hill JL (2011) Bayesian nonparametric modeling for causal inference. *Journal of Computational and Graphical Statistics* 20(1):217–240.

[6] Green DP, Kern HL (2012) Modeling heterogeneous treatment effects in survey experiments with Bayesian additive regression trees. *Public Opinion Quarterly* 76(3):491–511.

[7] Hansen BB, Bowers J (2009) Attributing effects to a cluster-randomized get-out-the-vote campaign. *Journal of the American Statistical Association* 104(487):873–885.

[8] Wager S, Athey S (2017) Estimation and inference of heterogeneous treatment effects using random

forests. *Journal of the American Statistical Association.*

[9] Kalla JL, Broockman DE (2018) The minimal persuasive effects of campaign contact in general elections: Evidence from 49 field experiments. *American Political Science Review* 112(1):148–166.

[10] Sekhon JS, Shem-Tov Y (2017) Inference on a new class of sample average treatment effects. *arXiv preprint arXiv:1708.02140.*

[11] Rubin DB (1974) Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology* 66(5):688.

[12] Splawa-Neyman J, Dabrowska DM, Speed T (1990) On the application of probability theory to agricultural experiments. *Statistical Science* 5(4):465–472.

[13] Rosenbaum PR, Rubin DB (1983) The central role of the propensity score in observational studies for causal effects. *Biometrika* 70(1):41–55.

[14] Tian L, Alizadeh AA, Gentles AJ, Tibshirani R (2014) A simple method for estimating interactions between a treatment and a large number of covariates. *Journal of the American Statistical Association* 109(508):1517–1532.

[15] Powers S, et al. (2017) Some methods for heterogeneous treatment effect estimation in high-dimensions. *arXiv preprint arXiv:1707.00102.*

[16] 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.*

[17] Scornet E, Biau G, Vert JP (2015) Consistency of random forests. *The Annals of Statistics* 43(4):1716–1741.

[18] Chipman HA, George EI, E. McCulloch R (2010) Bart: Bayesian additive regression trees. *The Annals of Applied Statistics* 4(1):266–298.

[19] Stone CJ (1982) Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics* 10(4):1040–1053.

[20] Birgé L (1983) Approximation dans les espaces métriques et théorie de l'estimation. *Probability Theory and Related Fields* 65(2):181–237.

[21] Györfi L, Kohler M, Krzyzak A, Walk H (2006) *A distribution-free theory of nonparametric regression.* (Springer Science & Business Media).

[22] Tsybakov AB (2009) *Introduction to nonparametric estimation.* (Springer Series in Statistics).

[23] Bickel PJ, Doksum KA (2015) *Mathematical statistics: Basic ideas and selected topics.* (CRC Press) Vol. 2.

[24] Hájek J (1967) On basic concepts of statistics in *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probabilities.* Vol. 1, pp. 139–162.

[25] Le Cam L (1956) On the asymptotic theory of estimation and testing hypotheses in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability.* Vol. 1.

[26] Mann CB (2010) Is there backlash to social pressure? A large-scale field experiment on voter mobilization. *Political Behavior* 32(3):387–407.

[27] Michelson MR (2016) The risk of over-reliance on the institutional review board: An approved project is not always an ethical project. *PS: Political Science & Politics* 49(02):299–303.

[28] Bohannon J (2016) For real this time: Talking to people about gay and transgender issues can change their prejudices. *Science.*

[29] Broockman D, Kalla J, Aronow P (2015) Irregularities in LaCour (2014). *Work. pap., Stanford Univ. http://stanford. edu/˜ dbroock/broockman_kalla_aronow_lg_irregularities. pdf.*

[30] Broockman DE, Kalla JL, Sekhon JS (2017) The design of field experiments with survey outcomes: A framework for selecting more efficient, robust, and ethical designs. *Political Analysis* 25:435–464.

[31] Breiman L (2001) Random forests. *Machine Learning* 45(1):5–32.

[32] 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.

[33] 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.

[34] 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.*

[35] 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.

[36] Efron B (2014) Estimation and accuracy after model selection. *Journal of the American Statistical Association* 109(507):991–1007.

[37] Putter H, Van Zwet WR (2012) Resampling: consistency of substitution estimators in *Selected Works of Willem van Zwet.* (Springer), pp. 245–266.

A. 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 (5), Neural Networks, or RFs (31) 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 (16)—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 2 (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) (8), 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.[5] 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 (8) for a precise explanation of the algorithm.

Figure 5 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.
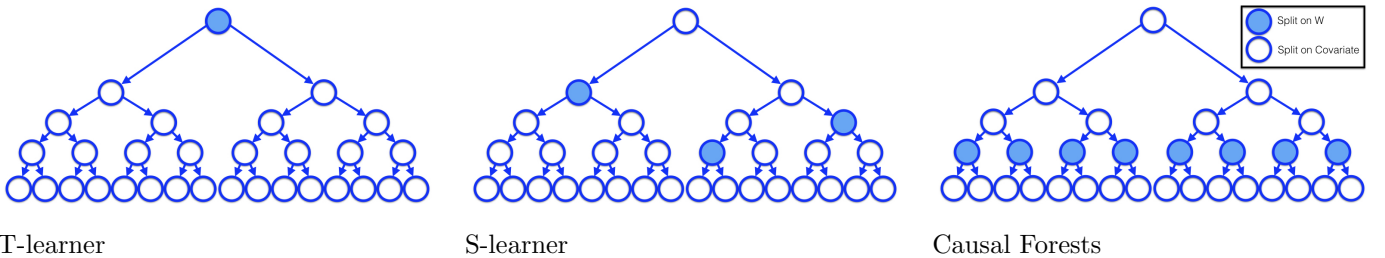

T-learner   S-learner   Causal Forests

Figure 5: 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:

---

[5]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,

$$X_i \overset{iid}{\sim} \mathcal{N}(0, \Sigma), \tag{13}$$

   where Σ is a correlation matrix that is created using the `vine` method (33).

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).[6]

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.

A.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 1 (unbalanced treatment assignment)

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

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 6 that the X-learner indeed performs much better in this unbalanced setting with both BART and RF as base learners.

A.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.

---

[6]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.
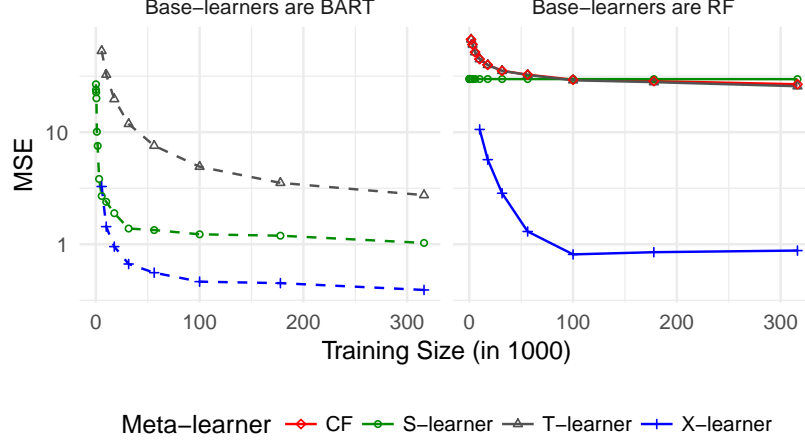
Figure 6: Comparison of S–, T–, and X–BART (left) and S–, T–, and X–RF and CF (right) for Simulation 1.

A.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 2 (complex linear)

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

The second setup (complex non-linear) is motivated by (8). Here the response function are non-linear functions.

Simulation 3 (complex non-linear)

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

with

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

Figure 7 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 2. However, in Simulation 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.
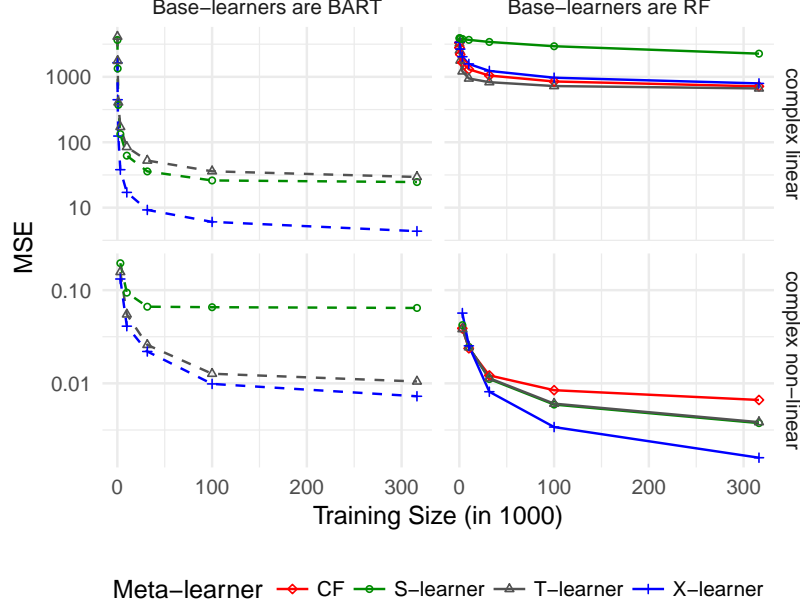
17

Figure 7: Comparison of the S-, T-, and X-learners with BART (left) and RF (right) as base learners for Simulation 2 (top) and Simulation 3 (bottom).

A.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 4) for both response functions. In Simulation 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 4 (global linear)

$$e(x) = 0.5, \quad d = 5,$$
$$\mu_0(x) = x^T \beta, \; with \; \beta \sim Unif([1, 30]^5),$$
$$\mu_1(x) = \mu_0(x).$$

Simulation 5 (piecewise linear)

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

with

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

and
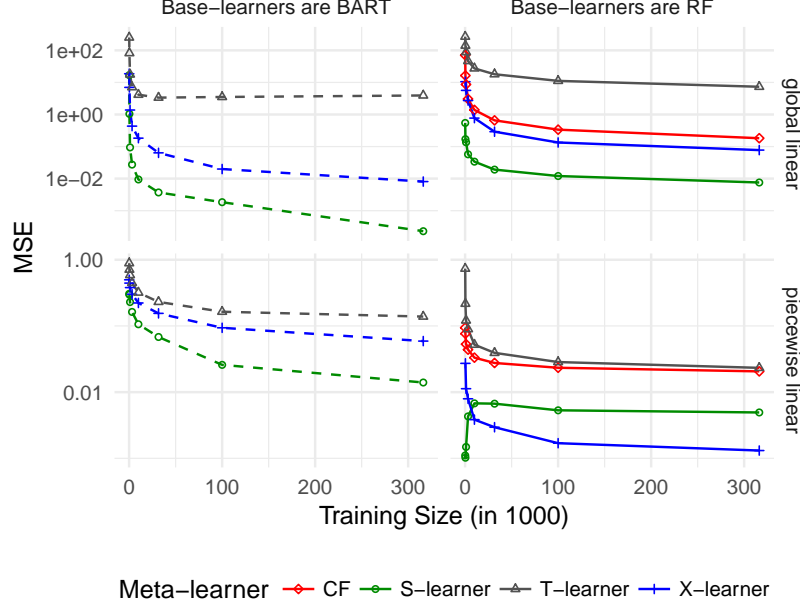
$$\beta \sim Unif([-15, 15]^d).$$

Figure 8: Comparison of S-, T-, and X-learners with BART (left) and RF (right) as base learners for Simulation 4 (top) and Simulation 5 (bottom).

Figure 8 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.

A.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 (8). 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 6 (beta confounded)

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

Figure 9 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 (8).

B. 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 (32), and we list it here only for completeness.
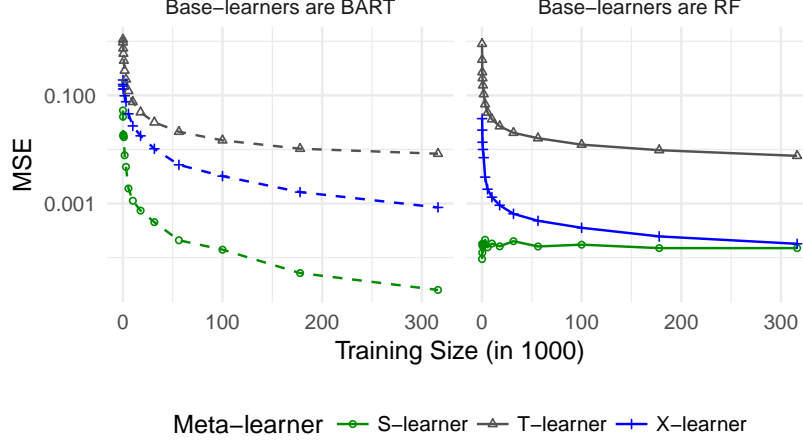
Figure 9: Comparison of S–, T–, and X–BART (left) and S–, T–, and X–RF (right) for Simulation 6.

Example 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.

C. 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 (35) 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 6, because it performed well for X–RF in the Atlantic Causal Inference Conference (ACIC) challenge (34), 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
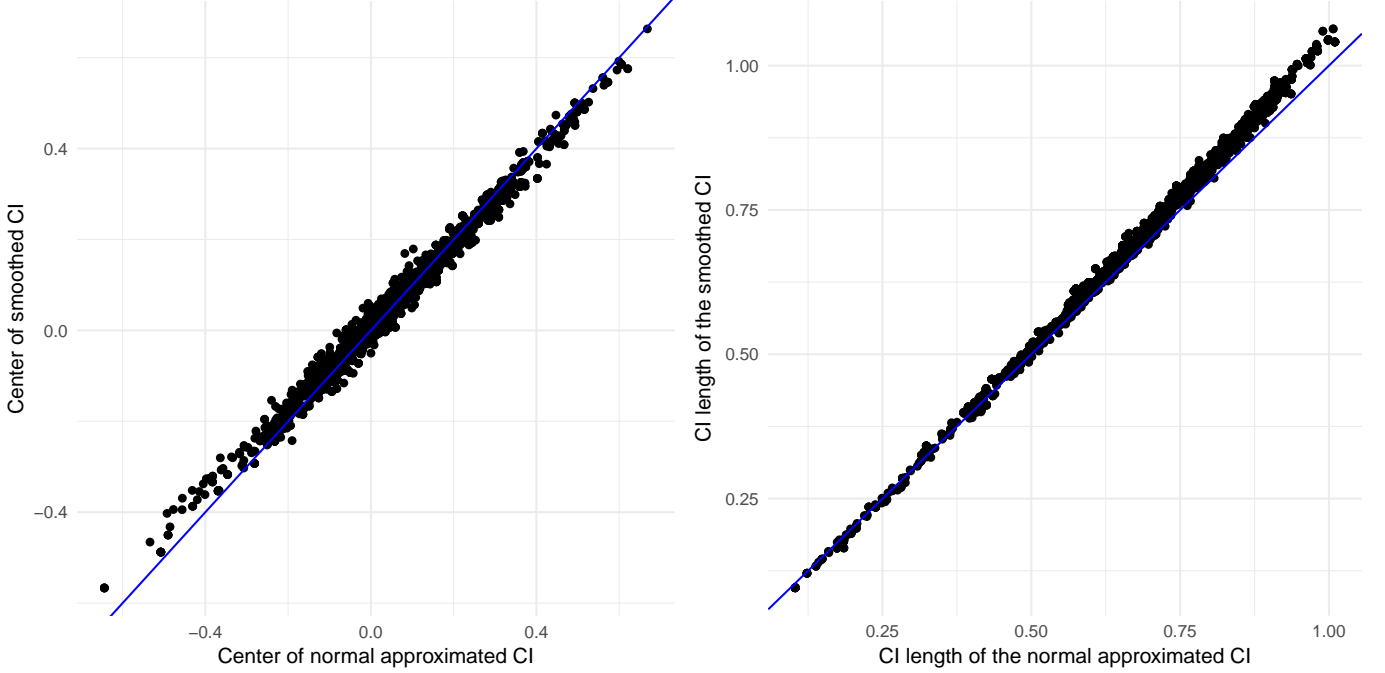
Figure 10: Comparison of normal approximated CI (Algorithm 6) and smoothed CI (Algorithm 7). The blue line is the identity function.

to estimate confidence intervals due to (37) and (36). 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 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 7) are unstable for a smaller B. Figure 10 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 7). However, normal approximated CIs (Algorithm 6) are computationally much less expensive and they are therefore our default method.

C.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 $\tau(x_i)$ from the observed outcome to obtain the outcome under control.
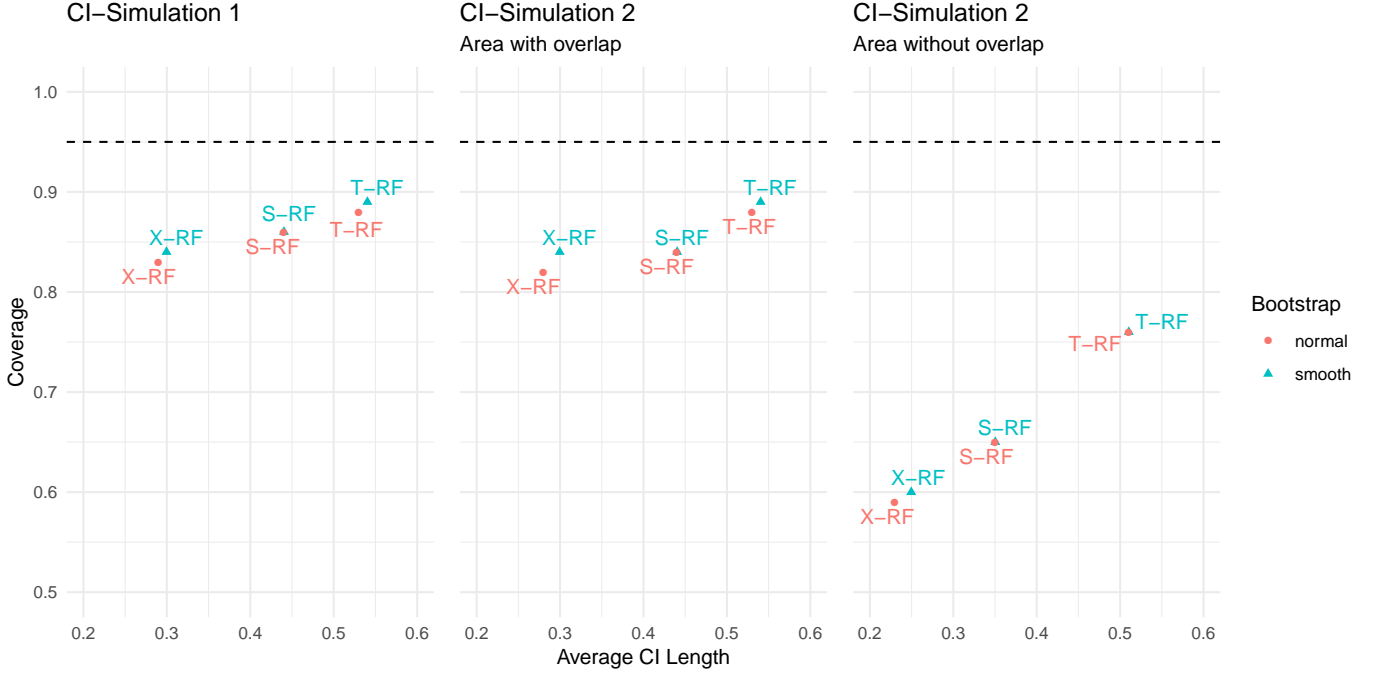
Figure 11: Coverage and average confidence interval length of the three meta-learners for normal approximated CI (Algorithm 6) and smoothed CI (Algorithm 7). The left figure corresponds to Simulation 3.1; the middle figure corresponds to units in an area with overlap in Simulation SI 3.2, and the right figure corresponds to units in an area without overlap in Simulation SI 3.2. The dotted line corresponds to the target 95% confidence interval.

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 11 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 8. The density plot in Figure 12 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 9 to estimate it. The upper subfigure in Figure 12 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.
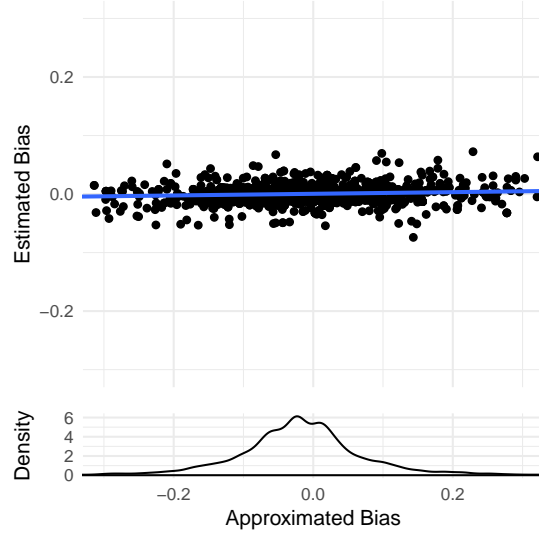
Figure 12: Approximated bias using Algorithm 8 versus estimated bias using Algorithm 9 and X–RF.

C.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 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 C.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.
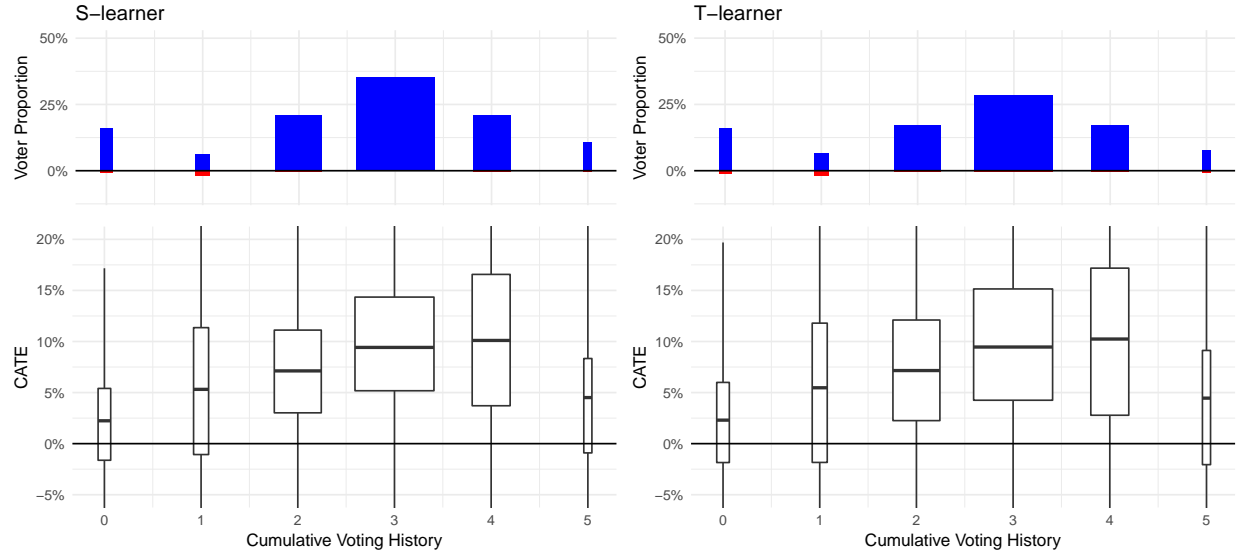
Figure 13: Results for the S-learner (left) and the T-learner (right) for the get-out-the-vote experiment.

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 11 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 is, however, very different. For this subgroup, we do not have overlap. We should be very cautious when estimating the CATE or confidence intervals of the CATE, and we would hope to see this reflected by very wide confidence intervals. This is unfortunately not the case. We observe that the confidence intervals are tighter and the coverage is much lower than on the data where we have overlap. 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.

D. 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 repeated our analysis with the S and T learner (cf. Figure 13). 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.

E. 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 14 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.
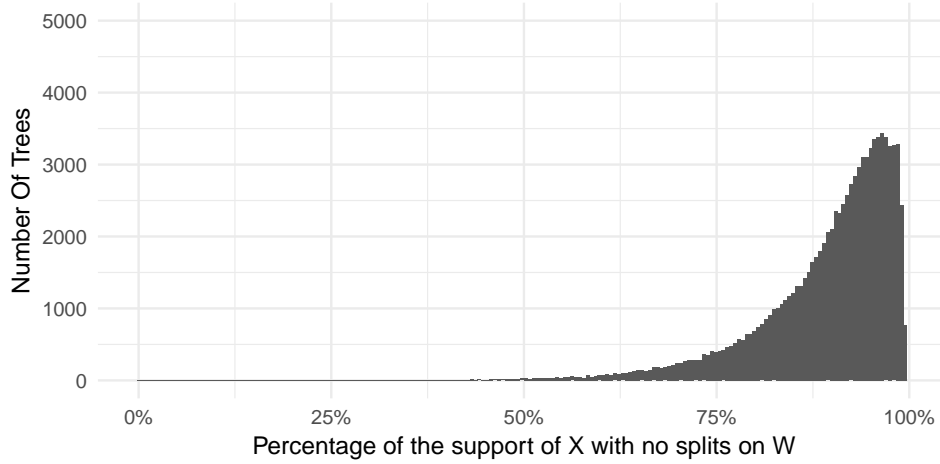
24

Figure 14: 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.

F. 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 4 and 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 4 and an MSE for Simulation 2.

Figure 15 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.

F.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 (34), 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.
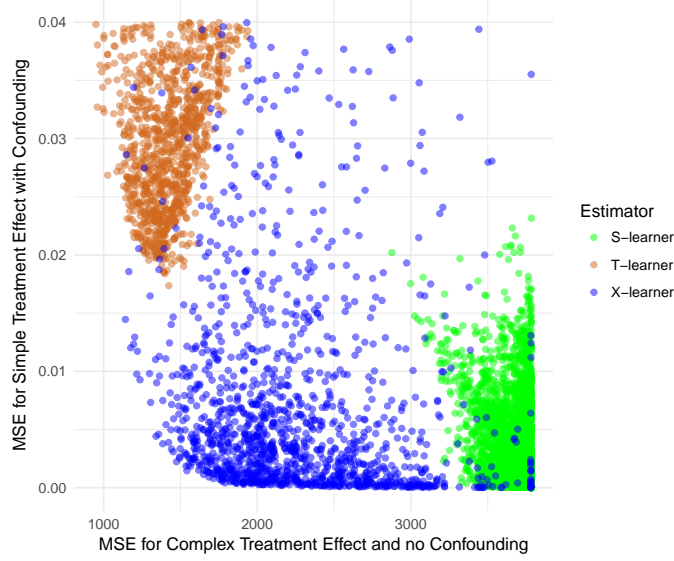
Figure 15: Each point corresponds to a different hyperparameter setting in random forests as the base learner in one of the S-, T-, or X-learners. The y-axis value is the MSE of Simulation 4 and the x-axis value is the MSE in Simulation 2. A perfect estimator that gets an MSE error of 0 in both simulations 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.

G. 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,

26

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 3 *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 1, we prove that the distributions are independent and in Lemma 2 we prove that they are identical.

Lemma 1 (independence) *Let n, N, and w be as in Theorem 3 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 $B_i \subset \mathbb{R}^p \times \mathbb{R}$,*

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{I}} \{(X_i, Y_i) \in B_i\} \,\Big|\, W = w\right) = \prod_{i \in \mathcal{I}} \mathbb{P}\left((X_i, Y_i) \in B_i \,\Big|\, W = w\right). \tag{14}$$

Note that another way of writing 14 is

$$\mathbb{P}^w\left(\bigcap_{i \in \mathcal{I}} \{(X_i, Y_i) \in B_i\}\right) = \prod_{i \in \mathcal{I}} \mathbb{P}^w\left((X_i, Y_i) \in B_i\right). \tag{15}$$

Proof. [Proof of Lemma 1]

$$\mathbb{P}\left(\bigcap_{i \in \mathcal{I}} \{(X_i, Y_i) \in B_i\} \,\Big|\, W = w\right)$$

$$= \mathbb{P}\left(\left(\bigcap_{i \in \mathcal{I}} \{(X_i, Y_i) \in B_i\}\right) \cap \left(\bigcap_{j \in S} \{W_j = 1\} \cap \bigcap_{k \in S^c} \{W_k = 0\}\right)\right) \,/\, \mathbb{P}\left(W = w\right)$$

$$= \mathbb{P}\left(\left(\bigcap_{i \in \mathcal{I}} \{(X_i, Y_i, W_i) \in B_i \times \{1\}\}\right) \cap \left(\bigcap_{j \in S \setminus \mathcal{I}} \{W_j = 1\} \cap \bigcap_{k \in S^c} \{W_k = 0\}\right)\right) \,/\, \mathbb{P}\left(W = w\right)$$

$$= \prod_{i \in \mathcal{I}} \mathbb{P}\left((X_i, Y_i, W_i) \in B_i \times \{1\}\right) \frac{\mathbb{P}\left(\bigcap_{j \in S \setminus \mathcal{I}} \{W_j = 1\} \cap \bigcap_{k \in S^c} \{W_k = 0\}\right)}{\mathbb{P}\left(W = w\right)} = (*).$$

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:

$$\frac{\mathbb{P}\left(\bigcap_{j \in S \setminus \mathcal{I}} \{W_j = 1\} \cap \bigcap_{k \in S^c} \{W_k = 0\}\right)}{\mathbb{P}\left(W = w\right)} = \frac{\prod_{j \in S \setminus \mathcal{I}} \mathbb{P}(W_j = 1) \prod_{k \in S^c} \mathbb{P}(W_k = 0)}{\prod_{j \in S} \mathbb{P}(W_j = 1) \prod_{k \in S^c} \mathbb{P}(W_k = 0)}$$

$$= \prod_{j \in J} \frac{1}{\mathbb{P}(W_j = 1)}$$

$$= \prod_{j \in J} \frac{\prod_{j \in S \setminus \{j\}} \mathbb{P}(W_j = 1) \prod_{k \in S^c} \mathbb{P}(W_k = 0)}{\prod_{j \in S} \mathbb{P}(W_j = 1) \prod_{k \in S^c} \mathbb{P}(W_k = 0)}$$

$$= \prod_{i \in \mathcal{I}} \frac{\mathbb{P}\left[\bigcap_{j \in S \setminus \{i\}} \{W_j = 1\} \cap \bigcap_{k \in S^c} \{W_k = 0\}\right]}{\mathbb{P}\left[W = w\right]}.$$

27

Thus,

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

which completes the proof.

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

Lemma 2 (identical distribution) *Assume the same assumptions as in Lemma 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

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

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 $\mathcal{P}$.

H. Convergence Rate Results for the T-learner

In this section, we want to prove that

$$
\mathrm{EMSE}(\hat{\tau}_T^{mn}, \mathcal{P}) \leq C(n^{-a_\mu} + m^{-a_\mu}).
$$

We start with a short lemma that will be useful for the proof of the theorem.

Lemma 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*

$$
\frac{e_{\min}}{e_{\max}}\mathbb{E}[g(X)] \leq \mathbb{E}[g(X)|W = 1] \leq \frac{e_{\max}}{e_{\min}}\mathbb{E}[g(X)], \tag{16}
$$

$$
\frac{1 - e_{\max}}{1 - e_{\min}}\mathbb{E}[g(X)] \leq \mathbb{E}[g(X)|W = 0] \leq \frac{1 - e_{\min}}{1 - e_{\max}}\mathbb{E}[g(X)]. \tag{17}
$$

Proof. [Proof of Lemma 3] Let us prove 16 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

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

17 follows from a symmetrical argument.

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,

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

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. [Conferegence Rate of the T-learner] The EMSE can be upper bounded by the errors of the single base learners:

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

$$\leq 2\underbrace{\mathbb{E}\left[(\hat{\mu}_1^n(\mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n\right]}_{A} + 2\underbrace{\mathbb{E}\left[(\hat{\mu}_0^m(\mathcal{X}) - \mu_0(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n\right]}_{B}.$$

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

$$A = \mathbb{E}\left[(\hat{\mu}_1^n(\mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| \sum_{i=1}^N W_i = n\right]$$

$$= \mathbb{E}\left[\mathbb{E}\left[(\hat{\mu}_1^n(\mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| W, \sum_{i=1}^N W_i = n\right] \middle| \sum_{i=1}^N W_i = n\right]. \tag{18}$$

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 ith 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 3 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 3 and use the definition of $S(a_\mu, a_\tau)$ to conclude that the inner expectation in 18 is in $\mathcal{O}(n^{-a_\mu})$:

$$\mathbb{E}\left[\hat{\mu}_1^n((X_{q_i}, Y_{q_i})_{i=1}^n, \mathcal{X}) - \mu_1(\mathcal{X}))^2 \middle| W, \sum_{i=1}^N W_i = n\right]$$

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

$$\leq \frac{e_{\max}}{e_{\min}} C n^{-a_\mu}.$$

Hence, it follows that

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

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

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

I. 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 4) 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 7). 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.

I.1. MSE and EMSE convergence rate for the linear CATE

Theorem 4 (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,

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

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

$$\sup_{n > n_0} \mathbb{E}\left[\gamma_{min}^{-1}(\hat{\Sigma}_n) \Big| \sum_{i=1}^N W_i = n\right] < c_1 \qquad and \qquad \sup_{n > n_0} \mathbb{E}\left[\gamma_{max}(\hat{\Sigma}_n)/\gamma_{min}^2(\hat{\Sigma}_n) \Big| \sum_{i=1}^N W_i = n\right] < c_2, \qquad (19)$$

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 $n > n_0$,

$$\mathbb{E}\left[(\tau(x) - \hat{\tau}_X(x))^2 \middle| \sum_{i=1}^N W_i = n\right] \leq C_x\left(m^{-a} + n^{-1}\right) \tag{20}$$

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

Proof. [Proof of Theorem 4] 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:

$$\mathbb{E}\left[(\tau(x) - \hat{\tau}_X(x))^2 \middle| \sum_{i=1}^N W_i = n\right] = \mathbb{E}\left[(x'(\beta - \hat{\beta}))^2 \middle| E_n\right] \leq \|x\|^2 \mathbb{E}\left[\|(X'X)^{-1}X'\delta\|^2 + \|(X'X)^{-1}X'\varepsilon\|^2 \middle| E_n\right]. \tag{21}$$

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}:

$$\begin{aligned}\mathbb{E}\left[\|(X'X)^{-1}X'\varepsilon\|^2 \middle| E_n\right] &= \mathbb{E}\left[tr\left(X(X'X)^{-1}(X'X)^{-1}X'\mathbb{E}\left[\varepsilon\varepsilon' \middle| X, E_n\right]\right) \middle| E_n\right] \\ &\leq \sigma^2 d\mathbb{E}\left[\gamma_{\min}^{-1}(\hat{\Sigma}_n) \middle| E_n\right] n^{-1} \\ &\leq \sigma^2 d c_1 n^{-1}.\end{aligned} \tag{22}$$

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

$$\begin{aligned}\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|_2^2 \middle| E_n\right] &\leq \mathbb{E}\left[\gamma_{\max}(\hat{\Sigma}_n)/\gamma_{\min}^2(\hat{\Sigma}_n)\|\delta\|_2^2 \middle| E_n\right] n^{-1} \\ &\leq \mathbb{E}\left[\gamma_{\max}(\hat{\Sigma}_n)/\gamma_{\min}^2(\hat{\Sigma}_n) \middle| E_n\right] C^0 m^{-a} \\ &\leq c_2 C^0 m^{-a}.\end{aligned} \tag{23}$$

Here we used that $\gamma_{\max}(\hat{\Sigma}_n^{-2}) = \gamma_{\min}^{-2}(\hat{\Sigma}_n)$, and $\mathbb{E}\left[\|\delta\|_2^2 \middle| X, E_n\right] = \mathbb{E}\left[\sum_{i=1}^n \delta^2(X_i) \middle| X, E_n\right] \leq nC^0 m^{-a}$. For the last statement, we used assumption B5. This leads to [20].

Bounding the EMSE

Proof. [Proof of Theorem 2] This proof is very similar to the proof of Theorem 4. 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:

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

31

We start with a similar decomposition as in [21]:

$$\mathbb{E}\left[(\tau(\mathcal{X}) - \hat{\tau}_X(\mathcal{X}))^2 \big| E_n\right] \leq \mathbb{E}\left[\|\mathcal{X}\|^2\right] \mathbb{E}\left[\|\beta - \hat{\beta}\|^2 \big| E_n\right]$$
$$= \mathbb{E}\left[\|\mathcal{X}\|^2\right] \mathbb{E}\left[\|(X'X)^{-1}X'\delta\|^2 + \|(X'X)^{-1}X'\varepsilon\|^2 \big| E_n\right]. \quad (24)$$

Following exactly the same steps as in [22], we get

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

Bounding $\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|_2^2 \big| E_n\right]$ is now slightly different than in [23]:

$$\begin{aligned}
\mathbb{E}\left[\|(X'X)^{-1}X'\delta\|_2^2 \big| E_n\right] &\leq \mathbb{E}\left[\gamma_{\min}^{-1}(X'X)\|X(X'X)^{-1}X'\delta\|_2^2 \big| E_n\right] \\
&\leq \mathbb{E}\left[\gamma_{\min}^{-1}(X'X)\|\delta\|_2^2 \big| E_n\right] \\
&\leq \mathbb{E}\left[\gamma_{\min}^{-1}(\Sigma_n)\frac{1}{n}\|\delta\|_2^2 \big| E_n\right] \\
&\leq C_\Sigma \mathbb{E}\left[\|\delta_1\|_2^2 \big| E_n\right].
\end{aligned} \quad (25)$$

Here the last inequality follows from Condition A6.

We now apply 16, 17, and Condition A4 to conclude that

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

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

$$\mathbb{E}\left[(\tau(\mathcal{X}) - \hat{\tau}_X(\mathcal{X}))^2 \big| E_n\right] \leq C_\mathcal{X}\left(\frac{e_{\max} - e_{\max}e_{\min}}{e_{\min} - e_{\max}e_{\min}} C_\Sigma C_0 m^{-a} + \sigma^2 d C_\Sigma n^{-1}\right). \quad (26)$$

I.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, $S, \bar{S} \subset \{1, \ldots, d\}$ 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 β 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 $m > m_0$,*

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

Here $\hat{\mu}_0^m$ is defined as in Condition A4.

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 5 *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 $n > 1$,*

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

We will prove the following lemma first, because it will be useful for the proof of Theorem 5.

Lemma 4 *Under the assmuption of Theorem 5, there exists a constant C such that for all $n > n_0$, $m > m_0$, and $s > 0$,*

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

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

Proof. [Proof of Lemma 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 A6 and then Chebyshev's inequality to conclude that for all $n > n_0$ (n_0 is determined by Condition A6),

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

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

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

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

Let us now bound the summands:

$$
\begin{aligned}
&\mathbb{E}\left[(f(X) - f(X^{(i)}))^2 \middle| E_n, \delta\right] \\
&= \mathbb{E}\left[\left(\|X'\delta\|_2^2 - \|X'\delta - (X_i - \tilde{X}_i)\delta_i\|_2^2\right)^2 \middle| E_n, \delta\right]
\end{aligned}
$$

33

$$= \mathbb{E}\left[\underbrace{\left(2\delta' X(X_i - \tilde{X}_i)\delta_i\right)^2}_{A} + \underbrace{\|(X_i - \tilde{X}_i)\delta_i\|_2^4}_{B} - \underbrace{4\delta' X(X_i - \tilde{X}_i)\delta_i\|(X_i - \tilde{X}_i)\delta_i\|_2^2}_{C} \Big| E_n, \delta\right].$$

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

$$\mathbb{E}\left[\left(2\delta' X(X_i - \tilde{X}_i)\delta_i\right)^2 \Big| E_n, \delta\right] = \mathbb{E}\left[4 \sum_{j,k=1}^{n} \delta_j X'_j(X_i - \tilde{X}_i)\delta_i \delta_k X'_k (X_i - \tilde{X}_i)\delta_i \Big| E_n, \delta\right]$$

$$\overset{(a)}{=} \mathbb{E}\left[4 \sum_{j=1}^{n} (\delta_j X'_j(X_i - \tilde{X}_i)\delta_i)^2 \Big| E_n, \delta\right]$$

$$\leq 4\delta_i^4 (n-1)\mathbb{E}\left[(X'_1(X_2 - \tilde{X}_2))^2 \Big| E_n, \delta\right] + 4\delta_i^4 \mathbb{E}\left[(X'_1(X_1 - \tilde{X}_1))^2 \Big| E_n, \delta\right]$$

$$\leq C_A \delta_i^4 n.$$

Here

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

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

$$E\left[\delta_j X'_j(X_i - \tilde{X}_i)\delta_i \delta_k X'_k(X_i - \tilde{X}_i)\delta_i \Big| E_n, \delta\right]$$
$$= \delta_j \mathbb{E}\left[\mathbb{E}\left[X'_j \Big| W, E_n, \delta\right] \mathbb{E}\left[(X_i - \tilde{X}_i)\delta_i \delta_k X'_k (X_i - \tilde{X}_i)\delta_i \Big| W, E_n, \delta\right] \Big| E_n, \delta\right] \qquad (27)$$
$$= 0,$$

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

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

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

$$\mathbb{E}\left[4\delta' X(X_i - \tilde{X}_i)\delta_i\|(X_i - \tilde{X}_i)\delta_i\|_2^2 \Big| E_n, \delta\right] = \mathbb{E}\left[\sum_{j=1}^{n} \delta_j X'_j (X_i - \tilde{X}_i)\delta_i\|(X_i - \tilde{X}_i)\delta_i\|_2^2 \Big| E_n, \delta\right]$$

$$= \mathbb{E}\left[\delta_i^4 X'_i(X_i - \tilde{X}_i)\|X_i - \tilde{X}_i\|_2^2 \Big| E_n, \delta \Big| E_n, \delta\right]$$

$$= C_C \delta_i^4,$$

where the second equality follows from the same argument as in 27, and the last equality is implied by Condition 8.

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

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

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

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

Proof. [Proof of Theorem 5] We start with the same decomposition as in 24:

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

and we follow the same steps to conclude that

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

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

$$\lim_{n\to\infty} \mathbb{E}\left[n\|(X'X)^{-1}X'\delta\|_2^2\big|E_n\right] = \lim_{n\to\infty, n>n_0} \int_0^\infty \mathbb{P}\left(n\|(X'X)^{-1}X'\delta\|_2^2 \geq s\Big|E_n\right) ds$$

$$\leq \lim_{n\to\infty, n>n_0} \int_0^\infty \max(1, C\frac{1}{s^2}) ds$$

$$\leq 1 + C.$$

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

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

I.3. EMSE convergence rate for Lipschitz continuous response functions

In Section I.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 3 (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 3 *The optimal rate of convergence for the regression problem of estimating $x \mapsto \mathbb{E}[Y|X = x]$ in Definition 3 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 4 *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.[7]

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_{mn}^L, 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 $D_{mn}^L \in S\left(\frac{2}{2+d}, \frac{2}{2+d}\right)$.

We will first prove in Theorem 6 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. We will then show in Theorem 7 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 $\mathcal{D}_{mn}^L$.

Theorem 6 (Minimax Lower Bound) *Let $\hat{\tau}$ be an arbitrary estimator, let $a_1, a_2 > 0$, and let c be such that for all $n, m \geq 1$,*
$$\sup_{\mathcal{P} \in \mathcal{D}_{mn}^L} EMSE(\mathcal{P}, \hat{\tau}^{mn}) \leq c(m^{-a_0} + n^{-a_1}); \tag{28}$$

then a_1 and a_2 are at most $2/(2+d)$:
$$a_0, a_1 \leq 2/(2+d).$$

Proof. [Proof of Theorem 6] 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 $\hat{\mu}_1^n$, such that
$$\sup_{\mathcal{P}_1 \in F^L} \mathbb{E}_{D_1^n \sim \mathcal{P}_1^n}\left[(\mu_1(\mathcal{X}) - \hat{\mu}_1^n(\mathcal{X}; \mathcal{D}_1^n))^2\right] \leq 2cn^{-a_1},$$

which is a contradiction, since by the definition of D_L^{mn}, μ_1 cannot be estimated at a rate faster than n^{-a} (cf., (21)). Note that we write here $\hat{\mu}_1^n(\mathcal{X}; \mathcal{D}_1^n)$, because we want to be explicit that $\hat{\mu}_1^n$ depends only on the treated observations.

Similiarly to $\hat{\mu}_1^n(\mathcal{X}; \mathcal{D}_1^n)$, we will use the notation $\hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}_0^m, \mathcal{D}_1^n)$ to be explicit about the dependence of the estimator $\hat{\tau}^{mn}$ on the data in the control group, $\mathcal{D}_0^m$, and on the data in the treatment group, $\mathcal{D}_1^n$. Furthermore, note that in Definition 4 each distribution in $\mathcal{D}_{mn}^L$ 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 $\mathcal{D}_{mn}^L$ with $\mu_0 = 0$, $\mu_1 = f_1$, $\varepsilon_0 \perp \varepsilon_1$, and W defined componentwise by

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

---

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

Then 28 implies that

$$c(m^{-a_0} + n^{-a_1}) \geq \sup_{\mathcal{P} \in \mathcal{D}_{mn}^L} \mathbb{E}_{(\mathcal{D}_0^m \times \mathcal{D}_1^n) \sim \mathcal{P}} \left[(\tau^{\mathcal{P}}(\mathcal{X}) - \hat{\tau}^{mn}(\mathcal{X}; \mathcal{D}_0^m, \mathcal{D}_1^n))^2 \right]$$

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

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 $\mathcal{D}_0^m = (X_i^0, Y_i^0)_{i=1}^N$ 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

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

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

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

$\mathbb{D}_1(f_1)$ 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

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

in order to conclude that

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

$$\geq \sup_{\mathcal{P}_1 \in F^L} \mathbb{E}_{D_1^n \sim \mathcal{P}_1^n} \left[\left(\mu_1^{\mathcal{P}_1^n}(\mathcal{X}) - \hat{\mu}_1^{nm}(D_1^n; \mathcal{X}) \right)^2 \right].$$

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 7 *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 6. That is, there exists a constant C such that

$$\mathbb{E}\|\tau - \hat{\tau}^{mn}\|^2 \leq C\sigma^{\frac{4}{d+2}} L^{\frac{2d}{2+d}} \left(m^{-2/(2+d)} + n^{-2/(2+d)} \right). \tag{29}$$

37

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 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 7,*

$$\mathbb{E}[\|\hat{\mu}_0^m - \mu_0\|^2] \leq \frac{\sigma^2}{k_0} + cL^2 \left(\frac{k_0}{m}\right)^{2/d},$$

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

for some constant c.

Proof. [Proof of Lemma 5] This is a direct implication of Theorem 6.2 in (21).

Lemma 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. [Proof of Lemma 6] First of all we consider

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

Now we can compute the expectation:

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

Proof. [Proof of Theorem 7] Many ideas in this proof are motivated by (21) and (23). 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

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

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

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

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

$$\mathbb{E}[|\tau(\mathcal{X}) - \hat{\tau}_1^{mn}(\mathcal{X})|^2]$$

$$= \mathbb{E}\left[\left|\mu_1(\mathcal{X}) - \mu_0(\mathcal{X}) - \hat{\mu}_1^n(\mathcal{X}) + \frac{1}{k_1} \sum_{i=1}^{k_1} \hat{\mu}_0^m(X_{(i,n)}^1(\mathcal{X}))\right|^2\right]$$

38

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

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 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:

$$\frac{1}{2}\mathbb{E}\left[\left|\mu_0(\mathcal{X}) - \frac{1}{k_1}\sum_{i=1}^{k_1}\hat{\mu}_0^m(X_{(i,n)}^1(\mathcal{X}))\right|^2\right]$$

$$\leq \mathbb{E}\left[\left|\mu_0(\mathcal{X}) - \frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mu_0\left(X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right)\right|^2\right] \tag{30}$$

$$+ \mathbb{E}\left[\left|\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mu_0\left(X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right) - \frac{1}{k_1}\sum_{i=1}^{k_1}\hat{\mu}_0^m(X_{(i,n)}^1(\mathcal{X}))\right|^2\right]. \tag{31}$$

31 can be bound as follows:

$$[31] = \mathbb{E}\left(\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mu_0\left(X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right) - Y_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right)^2$$

$$\leq \max_i \frac{1}{k_m^2}\sum_{j=1}^{k_0}\mathbb{E}\left(\mu_0\left(X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right) - Y_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right)^2$$

$$= \max_i \frac{1}{k_m^2}\sum_{j=1}^{k_0}\mathbb{E}\left[\mathbb{E}\left[\left(\mu_0\left(X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right) - Y_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right)^2\bigg\|\mathcal{D},\mathcal{X}\right]\right] \leq \frac{\sigma^2}{k_0}.$$

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

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

Next we find an upper bound for [30]:

$$[30] \leq \mathbb{E}\left(\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\left\|\mu_0(\mathcal{X}) - \mu_0\left(X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right)\right\|\right)^2$$

$$\leq \mathbb{E}\left(\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}L\left\|\mathcal{X} - X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right\|\right)^2$$

$$\leq L^2\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mathbb{E}\left\|\mathcal{X} - X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right\|^2 \tag{32}$$

$$\leq L^2\frac{1}{k_1}\sum_{i=1}^{k_1}\mathbb{E}\left\|\mathcal{X} - X_{(i,n)}^1(\mathcal{X})\right\|^2 \tag{33}$$

$$+ L^2\frac{1}{k_1 k_0}\sum_{i=1}^{k_1}\sum_{j=1}^{k_0}\mathbb{E}\left\|X_{(i,n)}^1(\mathcal{X}) - X_{(j,m)}^0(X_{(i,n)}^1(\mathcal{X}))\right\|^2 \tag{34}$$

where [32] follows from Jensen's inequality.

Let's consider [33]. 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

$$\frac{1}{k_1} \sum_{i=1}^{k_1} \mathbb{E}\left\|\mathcal{X} - X^1_{(i,n)}(\mathcal{X})\right\|^2 \leq \frac{1}{k_1} \sum_{i=1}^{k_1} \mathbb{E}\left\|\mathcal{X} - \tilde{X}_{i,1}(\mathcal{X})\right\|^2$$

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

Here the last inequality follows from Lemma 6. With exactly the same argument, we can bound [34] and we thus have

$$[30] \leq L^2 \tilde{c} * \left(\frac{1}{\lfloor \frac{n}{k_1} \rfloor^{2/d}} + \frac{1}{\lfloor \frac{n_2}{k_2} \rfloor^{2/d}}\right) \leq 2\tilde{c}L^2 * \left(\left(\frac{k_1}{n}\right)^{2/d} + \left(\frac{k_0}{m}\right)^{2/d}\right).$$

Plugging everything in, we have

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

$$\leq C\left(\frac{\sigma^2}{k_1} + L^2 \left(\frac{k_1}{n}\right)^{2/d} + \frac{\sigma^2}{k_0} + \left(\frac{k_0}{m}\right)^{2/d}\right)$$

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

J. 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_i^1 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_i^1 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 1 T-learner

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

4:     $\hat{\tau}(x) = \hat{\mu}_1(x) - \hat{\mu}_0(x)$

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

Algorithm 2 S-learner

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

$M(Y \sim (X, W))$ 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 3 X-learner

1: **procedure** X-LEARNER(X, Y, W, g)

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

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

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

8:　　$\hat{\tau}(x) = g(x)\hat{\tau}_0(x) + (1 - g(x))\hat{\tau}_1(x)$　　　　　　　　　　　　$\triangleright$ 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 $\text{Cov}(\tau_0(x), \tau_1(x))$, 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 4 F-learner

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

Algorithm 5 U-learner

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

Algorithm 6 Bootstrap Confidence Intervals 1

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)
2:　　$S_0 = \{i \; : \; w_i = 0\}$
3:　　$S_1 = \{i \; : \; w_i = 1\}$
4:　　$n_0 = \#S_0$
5:　　$n_1 = \#S_1$
6:　　**for** b in $\{1, \ldots, B\}$ **do**
7:　　　　$s_b^* = c(\text{sample}(S_0, \text{replace} = \text{T}, \text{size} = n_0), \text{sample}(S_1, \text{replace} = \text{T}, \text{size} = n_1))$
8:　　　　$x_b^* = x[s_b^*]$
9:　　　　$w_b^* = w[s_b^*]$
10:　　　$y_b^* = y[s_b^*]$
11:　　　$\hat{\tau}_b^*(p) = \text{learner}(x_b^*, w_b^*, y_b^*)(p)$
12:　　$\hat{\tau}(p) = \text{learner}(x, w, y)(p)$
13:　　$\sigma = sd(\{\hat{\tau}_b^*(p)\}_{b=1}^B)$
14:　　**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 7 Bootstrap Confidence Intervals 2

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)
2:     $S_0 = \{i \ : \ w_i = 0\}$
3:     $S_1 = \{i \ : \ w_i = 1\}$
4:     $n_0 = \#S_0$
5:     $n_1 = \#S_1$
6:     **for** b in $\{1, \ldots, B\}$ **do**
7:         $s_b^* = c(\text{sample}(S_0, \text{replace} = \text{T}, \text{size} = n_0), \text{sample}(S_1, \text{replace} = \text{T}, \text{size} = n_1))$
8:         $x_b^* = x[s_b^*]$
9:         $w_b^* = w[s_b^*]$
10:       $y_b^* = y[s_b^*]$
11:       $\hat{\tau}_b^*(p) = \text{learner}(x_b^*, w_b^*, y_b^*)(p)$
12:     $\tilde{\tau}(p) = \frac{1}{B} \sum_{b=1}^{B} \hat{\tau}_b^*(p)$
13:     For all b in $\{1, \ldots, B\}$ and j in $\{1, \ldots, n\}$ define

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

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

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

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

This version of the bootstrap was proposed in (36).

Algorithm 8 Monte Carlos Bias Approximation

1: **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)

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

3:          Create a new treatment assignment by permuting the original one,

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

4:          Define the observed outcome,
$$y_i = y(1)w_i + y(0)(1 - w_i).$$

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

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

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

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

8:      **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 9 Bootstrap Bias

1: **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)

2:      $S_0 = \{i \ : \ w_i = 0\}$

3:      $S_1 = \{i \ : \ w_i = 1\}$

4:      $n_0 = \#S_0$

5:      $n_1 = \#S_1$

6:      **for** b in $\{1, \ldots, B\}$ **do**

7:          $s_b^* = c(\text{sample}(S_0, \text{replace} = T, \text{size} = n_0), \text{sample}(S_1, \text{replace} = T, \text{size} = n_1))$

8:          $x_b^* = x[s_b^*]$

9:          $w_b^* = w[s_b^*]$

10:         $y_b^* = y[s_b^*]$

11:         $\hat{\tau}_b^*(p) = \text{learner}(x_b^*, w_b^*, y_b^*)(p)$

12:      $\hat{\tau}(p) = \text{learner}(x, w, y)(p)$

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

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