Machine Learning Methods for Estimating Heterogeneous Causal Effects

Article · April 2015

Source: arXiv

CITATIONS
59

READS
8,718

3 authors:

Susan Athey
Stanford University
156 PUBLICATIONS **5,831** CITATIONS
SEE PROFILE

Guido W. Imbens
Stanford University
165 PUBLICATIONS **24,635** CITATIONS
SEE PROFILE

Vikas Ramachandra
Stanford University
90 PUBLICATIONS **502** CITATIONS
SEE PROFILE

Machine Learning Methods for Estimating Heterogeneous Causal Effects*

Susan Athey† Guido W. Imbens‡

First Draft: October 2013
This Draft: April 2015

Abstract

In this paper we study the problems of estimating heterogeneity in causal effects in experimental or observational studies and conducting inference about the magnitude of the differences in treatment effects across subsets of the population. In applications, our method provides a data-driven approach to determine which subpopulations have large or small treatment effects and to test hypotheses about the differences in these effects. For experiments, our method allows researchers to identify heterogeneity in treatment effects that was not specified in a pre-analysis plan, without concern about invalidating inference due to multiple testing. In most of the literature on supervised machine learning (e.g. regression trees, random forests, LASSO, etc.), the goal is to build a model of the relationship between a unit's attributes and an observed outcome. A prominent role in these methods is played by cross-validation which compares predictions to actual outcomes in test samples, in order to select the level of complexity of the model that provides the best predictive power. Our method is closely related, but it differs in that it is tailored for predicting causal effects of a treatment rather than a unit's outcome. The challenge is that the "ground truth" for a causal effect is not observed for any individual unit: we observe the unit with the treatment, or without the treatment, but not both at the same time. Thus, it is not obvious how to use cross-validation to determine whether a causal effect has been accurately predicted. We propose several novel cross-validation criteria for this problem and demonstrate through simulations the conditions under which they perform better than standard methods for the problem of causal effects. We then apply the method to a large-scale field experiment re-ranking results on a search engine.

Keywords: Potential Outcomes, Heterogeneous Treatment Effects, Causal Inference, Supervised Machine Learning, Cross-Validation

*We are grateful for comments provided at seminars at the Southern Economics Association, Microsoft Research, the University of Pennsylvania, the University of Arizona, the Stanford Conference on Causality in the Social Sciences, and The MIT Conference in Digital Experimentation. Part of this research was conducted while the authors were visiting Microsoft Research.

†Graduate School of Business, Stanford University, and NBER. Electronic correspondence: athey@stanford.edu

‡Graduate School of Business, Stanford University, and NBER. Electronic correspondence: imbens@stanford.edu

1 Introduction

In this paper we study two closely related problems: first, estimating heterogeneity by features in causal effects in experimental or observational studies, and second, conducting inference about the magnitude of the differences in treatment effects across subsets of the population. Causal effects, in the Rubin Causal Model or potential outcome framework that we use here (Rubin, 1976, 1978; Imbens and Rubin, 2015), are comparisons between outcomes we observe and counterfactual outcomes we would have observed under a different regime or treatment. We introduce a method that provides a data-driven approach to select subpopulations with different average treatment effects and to test hypotheses about the differences between the effects in different subpopulations. For experiments, our method allows researchers to identify heterogeneity in treatment effects that was not specified in a pre-analysis plan, without concern about invalidating inference due to concerns about multiple testing.

Our approach is tailored for applications where there may be many attributes of a unit relative to the number of units observed, and where the functional form of the relationship between treatment effects and the attributes of units is not known. We build on methods from supervised machine learning (see Hastie, Tibshirani, and Friedman (2011) for an overview). This literature provides a variety of very effective methods for a closely related problem, the problem of predicting outcomes as a function of covariates in similar environments. The most popular methods (e.g. regression trees, random forests, LASSO, support vector machines, etc.) entail building a model of the relationship between attributes and outcomes, with a penalty parameter that penalizes model complexity. To select the optimal level of complexity (the one that maximizes predictive power without "overfitting"), the methods rely on cross-validation. The cross-validation approach compares a set of models with varying values of the complexity penalty, and selects the value of complexity parameter for which out-of-sample predictions best match the data using a criterion such as mean squared error (MSE). This method works well because in the test sample, the "ground truth" is known: we observe each unit's outcome, so that we can easily assess the performance of the model.

Our method is closely related to this approach, but it differs in that it is tailored for predicting causal effects of a treatment rather than a unit's outcome. We directly build the model that best predicts how treatment effects vary with the attributes of units. The challenge in applying the machine learning methods "off the shelf" is that the "ground truth" for a causal effect is not observed for any individual unit: we observe the unit with the treatment, or without the treatment, but not both at the same time, which is what Holland (1986) calls the "fundamental problem of causal inference." Thus, it is not obvious how to use cross-validation to determine whether a causal effect has been accurately predicted. We propose several novel cross-validation criteria for this problem and demonstrate through simulations the conditions under which they perform better than standard methods for the problem of causal effects. We then apply the method to applications, including a large-scale field experiment on voting and a large-scale field experiment re-ranking results on a search engine. Although we focus in the current paper mostly on regression tree methods (Breiman, Friedman, Olshen, and Stone, 1984), the methods extend to other approaches such as Lasso (Tibshirani, 1996), and support

vector machines (Vapnik, 1998, 2010).

2 The Problem

2.1 The Set Up

We consider a setup where either we have data from an experiment where a binary treatment is assigned randomly to units conditional on their observables, or where we have an observational study that satisfies the assumptions for "uncounfoundedness" or "selection on observables," (Rosenbaum and Rubin, 1983; Imbens and Rubin, 2015) There are N units, indexed by $i = 1, \ldots, N$. Let $W_i \in \{0, 1\}$ be the binary indicator for the treatment, with $W_i = 0$ indicating that unit i received the control treatment, and $W_i = 1$ indicating that unit i received the active treatment, and let X_i be a L-component vector of features, covariates or pretreatment variables, known not to be affected by the treatment. Let $p = \text{pr}(W_i = 1) = \mathbb{E}[W_i]$ be the marginal treatment probability, and let $e(x) = \text{pr}(W_i = 1|X_i = x)$ be the conditional treatment probability (the "propensity score" as defined by Rosenbaum and Rubin (1983)). In a randomized experiment with constant treatment assignment probabilities $e(x) = p$ for all values of x.

We assume that observations are exchangeable, and that there is no interference (the stable unit treatment value assumption, or sutva, Rubin, 1978). This assumption may be violated in network settings where some units are connected, or in settings where general equilibrium effects are important (Athey, Barrios, Eckles and Imbens, 2015). We postulate the existence of a pair of potential outcomes for each unit, $(Y_i(0), Y_i(1))$ (following the potential outcome or Rubin Causal Model, Rubin, 1977; Holland, 1986, Imbens and Rubin, 2015), with the unit-level causal effect defined as the difference in potential outcomes,

$$\tau_i = Y_i(1) - Y_i(0).$$

The realized and observed outcome for unit i is the potential outcome corresponding to the treatment received:

$$Y_i^{\text{obs}} = Y_i(W_i) = \begin{cases} Y_i(0) & \text{if } W_i = 0, \\ Y_i(1) & \text{if } W_i = 1. \end{cases}$$

Our data consist of the triple $(Y_i^{\text{obs}}, W_i, X_i)$, for $i = 1, \ldots, N$, which are regarded as an i.i.d sample drawn from an infinite superpopulation. Expectations and probabilities will refer to the distribution induced by the random sampling, or by the (conditional) random assignment of the treatment.

Define the conditional average treatment effect (CATE)

$$\tau(x) = \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x], \tag{2.1}$$

and the population average treatment effect

$$\tau^{\text{p}} = \mathbb{E}[Y_i(1) - Y_i(0)] = \mathbb{E}[\tau(X_i)].$$

A large part of the causal inference literature (Imbens and Rubin, 2015; Pearl, 2000; Hernán and Robins, 2015; Morgan and Winship, 2014) is focused on estimating the population average treatment effect τ^{p}. The main focus of the current paper is on obtaining accurate estimates of this conditional average treatment effect $\tau(x)$ for all values of x. The dimension of the covariates may be relatively large, and the covariates may be continuous or multi-valued, so we cannot simply partition the covariate space into subsets within which the covariates are constant.

There are a variety of reasons that researchers wish to conduct estimation and inference on the function $\tau(x)$. It may be used to assign future units to their optimal treatment, e.g., $W_i^{\text{opt}} = \mathbf{1}_{\tau(X_i)\geq 0}$. For example, if the treatment is a drug, we may only want to prescribe it to the subpopulation that benefits from it. Since the cost of the drug as well as the benefits of alternative drugs might be difficult to predict in advance, it is useful to know how the magnitude of the benefits vary with attributes of individuals in order to conduct cost-benefit analysis in the future. In addition, it might be that the population average effect of the drug is not positive, but that the drug is effective for particular categories of patients. Typically for clinical trials, researchers are required to pre-specify how they will analyze the data, to avoid concerns about multiple testing (whereby the researchers might conduct a large number of hypothesis tests for heterogeneous treatment effects, and we expect that some of those will show positive results even if the true effect is zero). A principled approach to estimating and conducting inference about $\tau(x)$ would allow such researchers to discover populations that do indeed benefit, even if they didn't have the foresight to specify this group as a subpopulation of interest in advance.

We may also be interested in the relative value of observing and basing decisions on different sets of covariates. For example, consider the policy of giving the treatment to those units associated with the covariates x that have the highest estimates of $\tau(x)$. The gain, in terms of expected outcomees, of such optimal assignment compared to uniform assignment is what we call the return to optimal treatment assignment for a particular set of covariates. Let $\mu_c(x) = \mathbb{E}[Y_i(0)|X_i = x]$ and $\mu_t(x) = \mathbb{E}[Y_i(1)|X_i = x]$. Now let $\theta_0 = \max(\mathbb{E}[\mu_c(X_i), \mathbb{E}[\mu_t(X_i)])$ be the maximum achievable average outcome given uniform assignment, and let $\theta_x = \mathbb{E}[\max(\mu_c(X_i), \mu_t(X_i)]$ be the maximum average outcome given optimal assignment based on the covariate x.

The return to optimal treatment assignment based on x, relative to using no covariates, is

$$r_x = \theta_x - \theta_0 = \mathbf{1}_{\tau^{\text{p}}\geq 0} \cdot \mathbb{E}\left[\mathbf{1}_{\tau(X_i)<0} \cdot |\tau(X_i)|\right] + \mathbf{1}_{\tau^{\text{p}}<0} \cdot \mathbb{E}\left[\mathbf{1}_{\tau(X_i)>0} \cdot |\tau(X_i)|\right].$$

Now compare this return to optimal treatment assignment based on x alone to that based on a richer set of covariates, say (x, x'). We may be interested in comparing r_x with $r_{(x,x')}$ to assess whether we should invest in estimating a system that assigns units to treatment based on x and x' rather than based on x alone.

3 Estimating Conditional Average Treatment Effects

3.1 The Problem

The goal is to develop an algorithm that generally leads to an accurate approximation $\hat{\tau}(x)$ to the conditional average treatment effect $\tau(x)$. Ideally we would measure the quality of the

approximation in terms of the goodness of fit, e.g., minus the expected squared error

$$-\mathbb{E}\left[\Big(\hat{\tau}(X_i)-\tau(X_i)\Big)^2\right].$$

The problem, of course, is that we do not know $\tau(\cdot)$, and so cannot directly compare different estimators $\hat{\tau}(\cdot)$ based on this criterion.

The general class of algorithms we consider has the following structure, common in the supervised learning literature. We consider a sequence of models of increasing complexity. We choose a method for estimating or training any model in that sequence given a training sample. We then compare the in-sample goodness-of-fit of the model with others in the sequence of models, adding to the in-sample goodness-of-fit measure a penalty term that increases with the complexity of the model. The penalty term involves a free parameter that determines how much increased complexity of the model is penalized. This parameter is chosen through out-of-sample cross-validation. Finally, different algorithms may be compared through out-of-sample goodness-of-fit measures on a test sample.

For the case where the goal is to construct an algorithm for a conditional expectation, $\mu(x)=\mathbb{E}[Y_i^{\rm obs}|X_i=x]$, which is the focus of conventional supervised-learning literature, many such algorithms have been proposed. See, for example, Hastie, Tibshirani and Friedman (2011). Such algorithms may involve splitting the sample based on feature values (building regression trees), where given a set of splits the conditional expectation within each leaf of the tree is estimated by the sample average within that leaf. The in-sample goodness-of-fit measure is the negative of the sum of squared deviations of the outcomes from these within-leaf averages. A conventional penalty term is a constant times the number of splits/leaves, with the constant chosen through out-of-sample cross-validation. The out-of-sample goodness-of-fit measure is the sum of squared deviations from the predicted values in the test sample.

The principal problem addressed in the current paper is that this approach does not extend directly to the case where the object of interest is the conditional average treatment effect because the conditional average treatment effect is not a conditional expectation of a variable we observe for all units in the sample. Specifically, given a candidate estimator for the conditional expectation $\tau(x)$, say $\hat{\tau}(x)$, we cannot measure the out-of-sample goodness-of-fit as minus the sum of deviations in the test sample,

$$Q^{\rm infeas}=-\frac{1}{N^{\rm te}}\sum_{i=1}^{N^{\rm te}}\Big(Y_i(1)-Y_i(0)-\hat{\tau}(X_i)\Big)^2.$$

This criteria is infeasible because we do not observe the values of the unit-level causal effects $\tau_i=Y_i(1)-Y_i(0)$ for any unit in the population.

3.2 The Structure of Supervised Learning Algorithms Using Cross-Validation for Model Selection

Our method follows the general structure of supervised learning algorithms that rely on cross-validation for model selection. We decompose this structure, which is common across a wide

range of conventional machine learning methods, into five components: (i) a sequence of possible models of greater complexity that will be considered; (ii) the method for estimation or training of a given model on the training sample; (iii) the in-sample goodness-of-fit measure to rank the models in the set of models considered on the training sample; (iv) the out-of-sample goodness-of-fit measure that is used to rank multiple candidate models for estimating conditional average treatment effects on the test sample; (v) the form of the penalty function that operates on each model and the choice of the tuning parameters in that penalty function.

Consider initially what these five components look like in the conventional case where the goal is to estimate a conditional expectation, $\mu(x) = \mathbb{E}[Y_i^{\text{obs}}|X_i = x]$ on the basis of information on features and outcomes $(\mathbf{X}^{\text{tr}}, \mathbf{Y}^{\text{tr,obs}})$ for units in a training sample, and compare different estimators in a test sample. For concreteness, we focus here on simple regression tree methods, but the components are similar for other supervised learning methods.

First, the sequence of regression tree models entails alternative partitions of the sample on the basis of feature values. Second, within each model in the sequence, the prediction function $\hat{\mu}(x)$ is constructed by taking the average value of the outcome within each member of the partition (each leaf of the tree). Third, the in-sample goodness-of-fit measure is minus the within-training-sample average of squared deviations from the estimated conditional expectation:

$$Q^{\text{is}}(\hat{\mu}; \mathbf{X}^{\text{tr}}, \mathbf{Y}^{\text{tr,obs}}) = -\frac{1}{N^{\text{tr}}} \sum_{i=1}^{N^{\text{tr}}} \Big(Y_i^{\text{tr,obs}} - \hat{\mu}(X_i^{\text{tr}}) \Big)^2.$$

Fourth, the penalty term is chosen to be proportional to the number of leaves in the tree, so that we choose the model by minimizing the criterion function

$$Q^{\text{crit}}(\hat{\mu}; \alpha, \mathbf{X}^{\text{tr}}, \mathbf{Y}^{\text{tr,obs}}) = Q^{\text{is}}(\hat{\mu}; \mathbf{X}^{\text{tr}}, \mathbf{Y}^{\text{tr,obs}}) - \alpha \cdot K = -\frac{1}{N^{\text{tr}}} \sum_{i=1}^{N^{\text{tr}}} \Big(Y_i^{\text{tr,obs}} - \hat{\mu}(X_i^{\text{tr}}) \Big)^2 - \alpha \cdot K,$$

where K is the number of leaves in the tree, measuring the complexity of the model. Fifth, the out-of-sample goodness-of-fit measure is minus the average of squared deviations from the candidate conditional expectation, over the units in the test sample,

$$Q^{\text{os}}(\hat{\mu}; \mathbf{X}^{\text{te}}, \mathbf{Y}^{\text{te,obs}}) = -\frac{1}{N^{\text{te}}} \sum_{i=1}^{N^{\text{te}}} \Big(Y_i^{\text{te,obs}} - \hat{\mu}(X_i^{\text{te}}) \Big)^2.$$

Thus the in-sample goodness-of-fit measure has the same functional form as the out-of-sample goodness-of-fit measure, and the two measures differ from the criterion function solely by the absence of the penalty term. The tuning parameter in the penalty term, α, is chosen by minimizing the out-of-sample goodness-of-fit measure over a number of cross-validation samples, often ten.

We propose two methods for estimating heterogeneous treatment effects where estimation follows the standard framework, and the results are applied to the problem of heterogeneous treatment effects. We also propose three additional methods that have the same structure as the conventional machine learning algorithm, but they differ in the implementation of some of

the components. The primary differences are in the two goodness-of-fit measures, both out-of-sample and in-sample, to address the problem that we do not observe the unit-level causal effects $\tau_i = Y_i(1) - Y_i(0)$ whose conditional expectation we attempt to estimate. There are also minor modifications of the estimation method and the sequence of models considered. The form of the penalty term we consider is the same for the regression tree case, linear in the number of leaves of the tree, with the tuning parameter again chosen by out-of-sample cross-validation.

Given choices for the five components of our method, the steps of the tree algorithm given the value for the penalty parameter α can be described as follows, where the tree is updated by splitting a terminal node in two on each iteration u.

Let $\mathcal{T}$ denote a tree, where each parent node has at most two children. The initial node 1 corresponds to a leaf containing all observations in the dataset. The children of node t are labeled t and $2t+1$, and each child is associated with a subset of the covariate space $\mathbb{X}$, so that a tree is a set of pairs $(t, \mathbb{X}_t)$. Terminal nodes are nodes with no children. Let $\mathcal{T}^{\text{term}}$ denote the set of terminal nodes of tree, where $\cup_{t \in \mathcal{T}^{\text{term}}} \mathbb{X}_t = \mathbb{X}$.

- Fix α. Initialize a tree $\mathcal{T}$ to be $\{(1, \mathbb{X})\}$. Initialize the set of completed nodes $\mathcal{C} = \{\}$.
- Until all terminal nodes $\mathcal{T}^{\text{term}}$ are in the set $\mathcal{C}$ of completed nodes, do the following:
  - Construct an estimator $\hat{\tau}(\cdot; \mathcal{T})$, as follows. For each terminal node t of $\mathcal{T}^{\text{term}}$, denoted $\mathbb{X}_t$, estimate $\hat{\tau}(\cdot; \mathcal{T})$ as a constant for all $x \in \mathbb{X}_t$ using the approach selected for component (ii) of the method.[1]
  - For each terminal node t of $\mathcal{T}^{\text{term}}$ not in the set of completed nodes $\mathcal{C}$:
    - For each feature $l = 1, .., L$:
      - For each potential threshold x_l^{thr} in the support of the l-th feature X_l, construct a new candidate tree $\mathcal{T}_{x_l^{\text{thr}}}$ by splitting $\mathbb{X}_t$ into two new nodes $2t$ and $2t+1$ based on the l-th feature and threshold x_l^{thr}: $\{x \in \mathbb{X}_t : x_l \leq x_l^{\text{thr}}\}$ and $\{x \in \mathbb{X}_t : x_l > x_l^{\text{thr}}\}$. Create a new estimate $\hat{\tau}(\cdot; \mathcal{T}_{x_l^{\text{thr}}})$ on this candidate tree as above.
      - Find the value $x_l^{t,*}$ that maximizes $Q^{\text{crit}}(\hat{\tau}(\cdot; \mathcal{T}_{x_l^{\text{thr}}}); \alpha, \mathbf{X}^{\text{tr}}, \mathbf{Y}^{\text{tr,obs}})$ over the threshold x_l^{thr}, where the form of Q^{crit} is selected as component (iii) of a given method.
    - If $\max_{l=1}^{L} Q^{\text{crit}}(\hat{\tau}(\cdot; \mathcal{T}_{x_l^{t,*}}); \alpha, \mathbf{X}^{\text{tr}}, \mathbf{Y}^{\text{tr,obs}}) \leq Q^{\text{crit}}(\hat{\tau}(\cdot; \mathcal{T}); \alpha, \mathbf{X}^{\text{tr}}, \mathbf{Y}^{\text{tr,obs}})$, add leaf t to the set of completed terminal nodes $\mathcal{C}$. Otherwise, let l^* be the feature with the highest gain, and update $\mathcal{T}$ to $\mathcal{T}_{x_{l^*}^{t,*}}$.
- Define $\mathcal{T}^{\alpha}$ to be the tree from the final iteration, and let $\hat{\tau}^{\alpha}(x)$ be the associated estimator.

To choose the penalty parameter α we do the following. Consider a compact set of potential values of α, denoted A. Let the lowest considered value of α be denoted α_0. We use R-fold cross-validation, where in the literature often $R = 10$.

[1] For some approaches, $\hat{\tau}(x)$ may be the mean of the transformed outcome within the set $\mathbb{X}_t$ that contains x; for other approaches, $\hat{\tau}(x)$ may be the difference between the average outcome for the treatment group and that of the control group, weighted by the inverse estimated propensity score.

- Partition the training sample into R subsamples, where the r-th subsample is denoted $(\mathbf{X}_r^{\text{tr}}, \mathbf{Y}_r^{\text{tr,obs}})$, and where its complement is denoted $(\mathbf{X}_{(r)}^{\text{tr}}, \mathbf{Y}_{(r)}^{\text{tr,obs}})$. For $r = 1, .., R$, we define $\hat{\tau}_{(r)}^{\text{pru}}(\cdot; \alpha)$ iteratively as follows:

  - Build a large tree $\mathcal{T}^{\alpha_0, r}$ using $(\mathbf{X}_{(r)}^{\text{tr}}, \mathbf{Y}_{(r)}^{\text{tr,obs}})$, following the procedure described above. Initialize $\mathcal{T}^{\text{pru},r} = \mathcal{T}^{\alpha_0, r}$ and $u = 1$.
  - Until $\mathcal{T}^{\text{pru},r} = \{(1, \mathbb{X})\}$:
    * For each node t in $\mathcal{T}^{\text{pru},r}$, define the subtree $\mathcal{T}_{(-t)}^{\text{pruned},r} \subset \mathcal{T}^{\underline{\alpha},r}$ that deletes all of the children of node t. Define

$$\Delta(t, \mathcal{T}^{\text{pru},r}) = \frac{Q^{\text{is}}(\hat{\tau}(\cdot; \mathcal{T}^{\text{pru},r}); \mathbf{X}_{(r)}^{\text{tr}}, \mathbf{Y}_{(r)}^{\text{tr,obs}}) - Q^{\text{is}}(\hat{\tau}(\cdot; \mathcal{T}_{(-t)}^{\text{pru},r}); \mathbf{X}_{(r)}^{\text{tr}}, \mathbf{Y}_{(r)}^{\text{tr,obs}})}{|\mathcal{T}^{\text{pru},r}| - |\mathcal{T}_{(-t^*)}^{\text{pru},r}|}$$

    * Find the "weakest link" which is the node t^* that maximizes $\Delta(t, \mathcal{T}^{\text{pru},r})$.
    * For α in $[\alpha_{u-1}, \Delta(t^*, \mathcal{T}^{\text{pru},r}))$, define $\hat{\tau}_{(r)}^{\text{pru}}(\cdot; \alpha) = \hat{\tau}(\cdot; \mathcal{T}_{(-t)}^{\text{pru},r})$.
    * Let $u = u + 1$.
  - For $\alpha \in A$ such that $\alpha > \alpha_{u-1}$, let $\hat{\tau}_{(r)}^{\text{pru}}(\cdot; \alpha) = \hat{\tau}(\cdot; \{(1, \mathbb{X})\})$.

- We evaluate the goodness-of-fit of an estimator on the r-th subsample using the method's choice of Q^{os}. We average these goodness-of-fit measures over the r subsamples to get

$$\overline{Q}^{\text{os}}(\alpha) = \frac{1}{R} \sum_{r=1}^{R} Q^{\text{os}}(\hat{\tau}_{(r)}^{\text{pru}}(\cdot; \alpha); \mathbf{X}_r^{\text{tr}}, \mathbf{Y}_r^{\text{tr,obs}}).$$

  We choose the value of α that maximizes this criterion function:

$$\alpha^* = \arg\max_{\alpha \in A} \overline{Q}^{\text{os}}(\alpha).$$

  $\mathcal{T}^*$ is then defined to be the optimal tree $\mathcal{T}^{\alpha^*}$ estimated using the approach above using the full training sample with $\alpha = \alpha^*$, and let the final estimator be $\hat{\tau}(x) = \hat{\tau}^{\alpha^*}(x)$.

3.3 The CATE-generating Transformation of the Outcome

Now let us return to the problem of estimating the conditional average treatment effect. A key insight is that we can characterize the conditional average treatment effect as a conditional expectation of an observed variable by transforming the outcome using the treatment indicator and the assignment probability. Recall that we maintain the assumption of randomization conditional on the covariates, or unconfoundedness (Rosenbaum and Rubin, 1983), formalized as follows:

Assumption 1. (Unconfoundedness)

$$W_i \perp \Big(Y_i(0), Y_i(1)\Big) \Big| X_i. \tag{3.1}$$

Then define the CATE-generating transformation of the outcome,

$$Y_i^* = Y_i^{\text{obs}} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))}, \tag{3.2}$$

where $e(x) = \Pr(W_i = 1 | X_i = x)$ is the conditional treatment probability, or the propensity score (Rosenbaum and Rubin, 1983). In the case with complete randomization the propensity score is constant $e(x) = p$ for all x, and the transformation simplifies to

$$Y_i^* = Y_i^{\text{obs}} \cdot \frac{W_i - p}{p \cdot (1 - p)}, \tag{3.3}$$

where $p = \mathbb{E}[e(X_i)] = \mathbb{E}[W_i] = \text{pr}(W_i = 1)$ is the common probability of assignment to the treatment. This transformation of the outcome has a key property.

Proposition 1. *Suppose that Assumption 3.1 holds. Then:*

$$\mathbb{E}\left[Y_i^* \middle| X_i = x\right] = \tau(x).$$

Proof: Because this property plays a crucial role in our discussion, let us expand on this equality. By definition,

$$\begin{aligned}
\mathbb{E}\left[Y_i^* \middle| X_i = x\right] &= \mathbb{E}\left[Y_i^{\text{obs}} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \middle| X_i = x \right] \\
&= \mathbb{E}\left[W_i \cdot Y_i^{\text{obs}} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} + (1 - W_i) \cdot Y_i^{\text{obs}} \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \middle| X_i = x \right].
\end{aligned}$$

Because $W_i \cdot Y_i^{\text{obs}} = W_i \cdot Y_i(1)$ and $(1 - W_i) \cdot Y_i^{\text{obs}} = (1 - W_i) \cdot Y_i(0)$, we can re-write this as

$$\begin{aligned}
&\mathbb{E}\left[W_i \cdot Y_i(1) \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} + (1 - W_i) \cdot Y_i(0) \cdot \frac{W_i - e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \middle| X_i = x \right] \\
&\quad = \mathbb{E}\left[Y_i(1) \cdot \frac{W_i \cdot (1 - e(X_i))}{e(X_i) \cdot (1 - e(X_i))} \middle| X_i = x \right] - \mathbb{E}\left[Y_i(0) \cdot \frac{(1 - W_i) \cdot e(X_i)}{e(X_i) \cdot (1 - e(X_i))} \middle| X_i = x \right] \\
&\quad = \mathbb{E}\left[Y_i(1) \cdot W_i \middle| X_i = x \right] \cdot \frac{1}{e(X_i)} - \mathbb{E}\left[Y_i(0) \cdot (1 - W_i) \middle| X_i = x \right] \cdot \frac{1}{1 - e(X_i)}.
\end{aligned}$$

Because of unconfoundedness this is equal to

$$\begin{aligned}
\mathbb{E}\left[Y_i^* \middle| X_i = x\right] &= \mathbb{E}\left[Y_i(1) \middle| X_i = x\right] \cdot \mathbb{E}\left[W_i \middle| X_i = x\right] \cdot \frac{1}{e(X_i)} \\
&\qquad - \mathbb{E}\left[Y_i(0) \middle| X_i = x\right] \cdot \mathbb{E}\left[1 - W_i \middle| X_i = x\right] \cdot \frac{1}{1 - e(X_i)} \\
&= \mu_1(x) - \mu_0(x) = \tau(x).
\end{aligned}$$

□

Remark I: At first sight it may appear that this transformation solves all the issues in applying conventional supervised learning methods to the problem of estimating the conditional average treatment effect. Using Y_i^* as the pseudo outcome allows one to directly use the conventional

algorithms for estimating conditional expectations without modification. This is in fact the basis of one of the algorithms we consider for estimation $\tau(\cdot)$. However, doing so need not be optimal. We essentially discard information by using only the sample values of the pairs (Y_i^*, X_i) rather than the sample values of the triples $(Y_i^{\rm obs}, W_i, X_i)$. It is possible that one can estimate $\tau(x)$ more efficiently by exploiting the information in observing the triple $(Y_i^{\rm obs}, W_i, X_i)$ beyond the information contained in the pair (Y_i^*, X_i). In fact, it is easy to see that this is the case. Suppose that the variance $\mathbb{V}(Y_i^{\rm obs}|W_i, X_i)$ is zero, so that $\mathbb{E}[Y_i|W_i = w, X_i = x]$ can be estimated without error for all x and w. Then it is also feasible to estimate the difference $\tau(x) = \mathbb{E}[Y_i^{\rm obs}|W_i = 1, X_i = x] - \mathbb{E}[Y_i^{\rm obs}|W_i = 0, X_i = x]$ without error. However, if there is variation in the treatment effect the variance $\mathbb{V}(Y_i^*|X_i)$ will be positive, and as a result there will be estimation error in estimates of $\mathbb{E}[Y_i^*|X_i = x]$ based on the values of the pairs (Y_i^*, X_i). Hence, using this transformation is not necessarily an efficient solution to the problem of estimating the conditional average treatment effect $\tau(x)$. Nevertheless, this CATE-generating transformation will play an important role in the discussion. □

Remark II: This transformation is related to a well-studied approach based on the inverse propensity score (Rosenbaum and Rubin (1983), Hirano, Imbens and Ridder (2003)). Building on weighting approaches for analysis of surveys developed by Horvitz and Thompson (1952), inverse propensity score methods correct for having a propensity for actions in the observed data that differs from the policy under consideration; for example, if our goal is to estimate the average outcome if all observations in the sample were treated ($\mu_t = \mathbb{E}[Y_i(1)]$), then under the assumption of unconfoundedness, we can use $\frac{1}{N}\sum_{i=1}^N Y_i^{\rm obs} W_i/e(X_i)$ as an estimator for μ_t, following arguments similar to those in Proposition 1. Beygelzimer and Langford (2009) consider the problem of assigning the optimal treatment to each unit, and they use inverse propensity score methods to evaluate the returns for alternative policies that map from attributes to treatments. They transform the problem to a conventional classification problem, and they use the outcome weighted by the inverse propensity score as importance weights. Given the transformation, the classifier predicts the optimal policy as a function of unit attributes, and the importance-weighted regret of the classifier is then equal to the loss from using a suboptimal policy. The loss from the classifier is equal to zero if the optimal policy is chosen, and so the approach is tailored towards finding values of the attributes where the optimal policy varies. Our approach differs in that we directly estimate the difference in mean outcomes and provide inference for those differences in means. Our approach is tailored to finding differences in the effect of the policy, even within regions where a single policy is optimal. In addition, this paper shows that performance can be improved by the approaches to estimation and goodness of fit that we propose. □

3.4 Two Out-of-sample Goodness-of-fit Measures

As discussed above, the ideal goodness of fit measure for the problem of estimating heterogeneous treatment effects, $Q^{\rm infeas}$, is infeasible. This motivates an analysis of alternative goodness of fit measures that rank models $\hat{\tau}(x)$ in the same way as the infeasible criterion.

More formally, given a test sample of size $N^{\rm te}$, we are looking for a function $Q^{\rm os} : \mathbb{C} \times \mathbb{R}^{(2+L)\cdot N^{\rm te}} \mapsto \mathbb{R}$ (where $\mathbb{C}$ is the space of functions from $\mathbb{R}^L$ to $\mathbb{R}$), that takes as input an

estimate $\hat{\tau}(\cdot)$ of the conditional average treatment effect and a test sample $(\mathbf{Y}^{\text{te,obs}}, \mathbf{W}^{\text{te}}, \mathbf{X}^{\text{te}})$ and gives a measure of fit such that as the test sample gets large, the function is minimized at the true conditional average treatment effect $\tau(\cdot)$.

3.4.1 A Measure Based on the Transformed Outcome

The first out-of-sample goodness-of-fit measure we propose exploits the CATE-generating transformation. We define

$$Q^{\text{os},TO}(\hat{\tau}; \mathbf{Y}^{\text{te,obs}}, \mathbf{W}^{\text{te}}, \mathbf{X}^{\text{te}}) = -\frac{1}{N^{\text{te}}} \sum_{i=1}^{N^{\text{te}}} \Big(Y_i^{\text{te},*} - \hat{\tau}(X_i^{\text{te}}) \Big)^2. \tag{3.4}$$

Holding fixed a particular training sample and associated estimator $\hat{\tau}(\cdot)$, we can take the expectation (over the realizations of the test sample) of the goodness of fit measure:

$$\mathbb{E}\left[Q^{\text{os},TO}(\hat{\tau} | \mathbf{Y}^{\text{te,obs}}, \mathbf{W}^{\text{te}}, \mathbf{X}^{\text{te}}) \right] = \mathbb{E}\left[\Big(\tau(X_i^{\text{te}}) - \hat{\tau}(X_i^{\text{te}}) \Big)^2 \right] + \mathbb{E}\left[\Big(Y_i^{\text{te},*} - \tau(X_i^{\text{te}}) \Big)^2 \right].$$

Because the second term, $\mathbb{E}[(Y_i^{\text{te},*} - \tau(X_i^{\text{te}}))^2]$, does not depend on the estimator $\hat{\tau}(\cdot)$, the sum of the two terms is minimized over $\hat{\tau}(\cdot)$ by minimizing the first term, $\mathbb{E}[(\tau(X_i^{\text{te}}) - \hat{\tau}(X_i^{\text{te}}))^2]$, which is uniquely minimized at $\hat{\tau}(x) = \tau(x)$ for all x. Thus, this criterion is likely to select the optimal estimator among a set of estimators if the test sample is sufficiently large so that the expectation is well approximated by the average over the test sample.

3.4.2 A Measure Based on Matching

As discussed in Remark I above, the transformed outcome approach introduces variance, since the sample average of $Y^{\text{te},*}$ may not be the lowest-variance estimator of $\tau(x)$, although it is unbiased. This motivates a search for alternative goodness-of-fit measures. Although there are many possible approaches to estimating treatment effects that we could in principle draw on, in this setting we are guided by a few desiderata. First, to preserve the spirit of conventional cross-validation approaches, it is desirable to have an individualized "target" treatment effect for each test observation, or at least based on a very small set of test units. Second, we wish to prioritize bias reduction over variance reduction to the extent possible. Thus motivated, we propose an approach based on matching, where we estimate the causal effect $\tau(x)$ by finding pairs of units, one treated and one control, with X_i close to x, and differencing their outcomes. As discussed in Imbens and Rubin (2015), matching approaches perform well in practice, and although they introduce bias (by comparing treatment observations to observations that are nearby but still have distinctly different attributes), the bias can be minimized over the set of all matching estimators by using a single match.

To be specific, fix an integer M smaller than the minimum of the number of control and treated units in the test sample size. We will construct M pairs of test units that will be used to estimate unit-level causal effects τ_i to be used in turn to evaluate estimates of the conditional average treatment effect. Randomly draw M units from the subsample of treated units in the test sample. Let the feature values for the m-th unit in this sample be $X_{\text{t},m}^{\text{te}}$ and the outcome

$Y_{\mathrm{t},m}^{\mathrm{te,obs}}$. For each of these M treated units find its closest match among the controls in the test sample, in terms of the feature values by minimizing

$$\min_{i:W_i=1} \left\| X_i^{\mathrm{te}} - X_{\mathrm{t},m}^{\mathrm{te}} \right\|.$$

Let $X_{\mathrm{c},m}^{\mathrm{te}}$ be the features for the control match, and let $Y_{\mathrm{c},m}^{\mathrm{te,obs}}$ be its outcome. For this pair of test units we estimate the average causal effect as

$$\tilde{\tau}_m^{\mathrm{te}} = Y_{\mathrm{t},m}^{\mathrm{te,obs}} - Y_{\mathrm{c},m}^{\mathrm{te,obs}}.$$

If the match is perfect so that the feature values for the match unit and its match are equal, $X_{\mathrm{c},m}^{\mathrm{te}} = X_{\mathrm{t},m}^{\mathrm{te}}$, then $\hat{\tau}_m^{\mathrm{te}}$ is unbiased for the conditional average treatment effect in the pair, defined as

$$\tau_m^{\mathrm{te}} = \frac{1}{2} \cdot \Big(\tau(X_{\mathrm{t},m}^{\mathrm{te}}) + \tau(X_{\mathrm{c},m}^{\mathrm{te}}) \Big).$$

Then, compare $\hat{\tau}_m^{\mathrm{te}}$ to the predicted average treatment effect based on the estimator for CATE, $\hat{\tau}(x)$:

$$\hat{\tau}_m^{\mathrm{te}} = \frac{1}{2} \cdot \Big(\hat{\tau}(X_{\mathrm{t},m}^{\mathrm{te}}) + \hat{\tau}(X_{\mathrm{c},m}^{\mathrm{te}}) \Big).$$

Finally, we average the difference between these two over M randomly drawn matched pairs from the test sample:

$$Q^{\mathrm{os,match}}(\hat{\tau}; \mathbf{Y}^{\mathrm{tr,obs}}, \mathbf{W}^{\mathrm{tr}}, \mathbf{X}^{\mathrm{tr}}) = \frac{1}{N^{\mathrm{te}}} \sum_m \Big(\hat{\tau}_m^{\mathrm{te}} - \tau_m^{\mathrm{te}} \Big)^2 = -\frac{1}{M} \sum_{m=1}^{M} \Big(\hat{\tau}_m^{\mathrm{te}} - \tilde{\tau}_m^{\mathrm{te}} \Big)^2. \tag{3.5}$$

If the test sample is large the matched pairs will be very close in terms of feature values, $X_{\mathrm{t},m}^{\mathrm{te}} \approx X_{\mathrm{c},m}^{\mathrm{te}}$. If in fact the matching were exact, and $X_{\mathrm{t},m}^{\mathrm{te}} = X_{\mathrm{c},m}^{\mathrm{te}}$ for all m, then $\mathbb{E}[\tilde{\tau}_m^{\mathrm{te}}] = \tau_m^{\mathrm{te}}$, and the goodness-of-fit measure satisfies

$$\mathbb{E}\left[Q^{\mathrm{os,match}}(\hat{\tau}; \mathbf{Y}^{\mathrm{tr,obs}}, \mathbf{W}^{\mathrm{tr}}, \mathbf{X}^{\mathrm{tr}}) \middle| \mathbf{X}^{\mathrm{te}} \right] = \frac{1}{N^{\mathrm{te}}} \sum_m \Big(\hat{\tau}_m^{\mathrm{te}} - \tau_m^{\mathrm{te}} \Big)^2,$$

plus a constant, with the expectation of the right hand side minimized at the true value $\tau(\cdot)$ for the CATE.

A disadvantage of this approach is that matching estimators are computationally costly, since it is necessary to find the closest observation with the opposite treatment status for each test observation. Thus, in this paper we focus on using $Q^{\mathrm{os,match}}$ primarily to compare alternative algorithms, but not for cross-validation. We also do not suggest using it for measuring in-sample goodness of fit. In applications, however, the matching estimator could also be used in these ways if computational time was not a concern.

3.5 In-Sample Goodness-Of-Fit Measures

The second component of the algorithm that differs from the corresponding component in conventional supervised learning methods is the in-sample goodness-of-fit measure. In conventional supervised learning algorithms for predicting continuous outcomes, these measures are typically identical to the out-of-sample goodness-of-fit measures, with the only difference that now they are evaluated in sample (however, for classification methods, it is common to use different metrics in-sample and out-of-sample). Here we consider two such measures, one which is the analog of its out-of-sample counterpart and one which is distinct.

3.5.1 An In-sample Goodness-of-fit Measure Based on the Transformed Outcome

The first measure is the direct analogue of the first out-of-sample measure using the transformed outcome. Now using the training sample, we calculate

$$Q^{\text{is},TO}(\hat{\tau}|\mathbf{Y}^{\text{tr,obs}},\mathbf{W}^{\text{tr}},\mathbf{X}^{\text{tr}}) = -\frac{1}{N^{\text{tr}}}\sum_{i=1}^{N^{\text{tr}}}\Big(Y_i^{\text{tr},*} - \hat{\tau}(X_i^{\text{tr}})\Big)^2. \tag{3.6}$$

3.5.2 An Alternative In-sample-goodness-of-fit Measure

The second goodness-of-fit measure is based on an alternative characterization of the in-sample goodness-of-fit measure in the conventional supervised learning case. In that case the goodness-of-fit measure takes the form

$$-\frac{1}{N^{\text{tr}}}\sum_{i=1}^{N^{\text{tr}}}\Big(Y_i^{\text{tr,obs}} - \hat{\mu}(X_i^{\text{tr}})\Big)^2 = -\frac{1}{N^{\text{tr}}}\sum_{i=1}^{N^{\text{tr}}}\Big((Y_i^{\text{tr,obs}})^2 - 2\cdot Y_i^{\text{tr,obs}}\cdot\hat{\mu}(X_i^{\text{tr}}) + \hat{\mu}^2(X_i^{\text{tr}})\Big)$$

$$= -\frac{1}{N^{\text{tr}}}\sum_{i=1}^{N^{\text{tr}}}\Big((Y_i^{\text{tr,obs}})^2 - 2(Y_i^{\text{tr,obs}} - \hat{\mu}(X_i^{\text{tr}}))\cdot\hat{\mu}(X_i^{\text{tr}}) - \hat{\mu}^2(X_i^{\text{tr}})\Big).$$

If the models include an intercept, as they usually do, most estimation methods would ensure that the average of $(Y_i^{\text{tr,obs}} - \hat{\mu}(X_i^{\text{tr}}))\cdot\hat{\mu}(X_i^{\text{tr}})$ would be equal to zero, so that the goodness-of-fit measure is equivalent to

$$-\frac{1}{N^{\text{tr}}}\sum_{i=1}^{N^{\text{tr}}}\Big((Y_i^{\text{tr,obs}})^2 - \hat{\mu}^2(X_i^{\text{tr}})\Big).$$

To interpret this, because the first component does not depend on the estimator being used, a model fits better according to this criteria if it yields higher variance predictions. This criteria makes sense because the estimation forces the predictions to be unbiased and the estimator is efficient given the model. Thus, additional variance corresponds to more refined discrimination among units in terms of their outcomes. In this regard, it is analogous to using a Gini coefficient to evaluate the performance of a classification algorithm in sample. For classification, more inequality among predicted probabilities corresponds to more accurate predictions for an unbiased classifier.

Since our approaches to estimating conditional average treatment effects will also be unbiased, we propose a second in-sample goodness of fit measure based on the sum of squared predicted treatment effects. When comparing two unbiased estimators, the higher-variance estimator will fit better according to this criteria. Formally, we define:

$$Q^{\text{is,sq}}(\hat{\tau}) = \frac{1}{N^{\text{tr}}} \sum_{i=1}^{N^{\text{tr}}} \hat{\tau}(X_i^{\text{tr}})^2. \tag{3.7}$$

In the conventional setting where the goal is to estimate a conditional expectation, $Q^{\text{is,sq}}(\hat{\tau})$ and $Q^{\text{is},TO}$ are identical up to a constant, but in our setting where we attempt to estimate causal effect their difference is

$$Q^{\text{is},TO}(\hat{\tau}) - Q^{\text{is,sq}}(\hat{\tau}) = -\frac{2}{N^{\text{tr}}} \sum_{i=1}^{N^{\text{tr}}} \Big(Y_i^{\text{tr},*} - \hat{\tau}(X_i) \Big) \cdot \hat{\tau}(X_i) + \frac{1}{N^{\text{tr}}} \sum_{i=1}^{N^{\text{tr}}} (Y_i^{\text{tr},*})^2,$$

which need not be constant as a function of $\hat{\tau}(\cdot)$. Since the two approaches will not yield the same results, comparing the two requires further analysis. We explore their relative performance in simulations and applications. Some intuition can be gained, however, from the example in Remark I: if both the treatment and control outcomes can be predicted very accurately with features x, then the transformed outcome approach is likely to yield higher variance.

4 Five Algorithms for Estimating Conditional Average Treatment Effects

This section describes the five algorithms that we consider for estimating the CATE function. Each of them makes different choices about the five different components of supervised learning algorithms. All of our algorithms share the same form for the penalty function, namely a fixed penalty per leaf in the tree. They all have a similar approach for component (i) the sequence of models, namely tree-based models that sequentially partition the covariate space. The following table summarizes the dimensions on which the algorithms differ:

4.1 A Single Tree Based on the Observed Outcome

In the first algorithm, which we refer to as the Single Tree (ST) algorithm, we use conventional methods for trees to estimate the conditional expectation $\mu(w,x) = \mathbb{E}[Y_i^{\text{obs}} | W_i = w, X_i = x]$, with the observed outcome Y_i^{obs} as the outcome and both the treatment W_i and X_i as the features. Given the estimate $\hat{\mu}(w,x) = \hat{\mathbb{E}}[Y_i^{\text{obs}} | W_i = w, X_i = x]$, we estimate the CATE $\tau(x)$ as $\hat{\tau}_{\text{ST}}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x)$.

4.2 Separate Trees for the Observed Outcome by Treatment Groups

One simple modification of the first algorithm is to split the problem of estimating $\tau(x)$ into two separate supervised learning problems. We refer to this algorithm as the Two Tree (TT)

Model Structure	Estimation Approach	In-sample fit for Obs. i	Out-of-sample fit for Obs. i
ST: Single Tree Outcome Y_i Features X_i, W_i	$\hat{Y}_i^{\text{obs}}$ = sample ave. of Y_i in leaf	MSE $-(Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}})^2$	MSE $-(Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}})^2$
TT: Two Trees (by W_i) Outcome Y_i Features X_i	$\hat{Y}_i^{\text{obs}}$ = sample ave. of Y_i in leaf	MSE $-(Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}})^2$	MSE $-(Y_i^{\text{obs}} - \hat{Y}_i^{\text{obs}})^2$
TOT: Transformed Outcome Y_i^* Features X_i	$\hat{\tau}(X_i)$ = sample ave. of Y_i^* in leaf	$Q^{\text{is},TO}$ $-(Y_i^* - \hat{\tau}(X_i))^2$	$Q^{\text{os},TO}$ $-(Y_i^* - \hat{\tau}(X_i))^2$
CT-TO: Causal Tree-Trans. Outcome for $\tau(x)$ Features X_i	$\hat{\tau}(X_i)$ = difference betw. (weighted) sample averages Treat - Control in leaf	$Q^{\text{is},TO}$ $-(Y_i^* - \hat{\tau}(X_i))^2$	$Q^{\text{os},TO}$ $-(Y_i^* - \hat{\tau}(X_i))^2$
CT: Causal Tree for $\tau(x)$ Features X_i	$\hat{\tau}(X_i)$ = difference betw. (weighted) sample averages Treat - Control in leaf	$Q^{\text{is,sq}}$ $(\hat{\tau}(X_i))^2$	$Q^{\text{os,TO}}$ $-(Y_i^* - \hat{\tau}(X_i))^2$

algorithm. First we can use the conventional supervised learning algorithms to estimate the two conditional expectations $\mu(0,x)$ and $\mu(1,x)$ separately. For estimating $\mu(0,x)$ we use the subsample with $W_i = 0$, and construct a tree with Y_i^{obs} as the outcome and X_i as the features. For estimating $\mu(1,x)$ we use the subsample with $W_i = 1$, and construct a tree with Y_i^{obs} as the outcome and X_i as the features. Next, we estimate $\tau(x)$ as the difference between the two estimates, $\hat{\tau}_{\text{TT}}(x) = \hat{\mu}(1,x) - \hat{\mu}(0,x)$.

For the problem of estimating average treatment effects, Hahn (1998) showed that estimating treatment effects using unweighted nonparametric regression achieves the efficiency bound asymptotically, which motivates the above procedure as well as the ST algorithm. On the other hand, Hirano, Imbens and Ridder (2003) established that taking the difference of the sample averages within each of the two groups, weighted by the inverse of the nonparametrically estimated propensity score is also asymptotically efficient. This suggests that in the case where the propensity score $e(x)$ varies with x, there may be some gain to modifying the TT approach by weighting outcomes by the inverse of the estimated propensity score. That is, form an estimate of the propensity score $\hat{e}(x)$ and estimate the expected average outcome in leaf l corresponding to $\mathbb{X}_l$ as $\sum_{i:X_i\in\mathbb{X}_l} Y_i^{\text{obs}}/\hat{e}(X_i) / \sum_{i:X_i\in\mathbb{X}_l} 1/\hat{e}(X_i)$ in the treatment tree and $\sum_{i:X_i\in\mathbb{X}_l} Y_i^{\text{obs}}/(1-\hat{e}(X_i)) / \sum_{i:X_i\in\mathbb{X}_l} 1/(1-\hat{e}(X_i))$ in the control tree. Another variant is to include $\hat{e}(x)$ as a feature in the estimation of each of the trees so that the tree may directly split by values of the estimated propensity score

4.3 The Transformed Outcome Tree Method

In the third algorithm we apply conventional supervised learning methods to the transformed outcome Y_i^*, without any further modifications to the algorithms, treating Y_i^* as the outcome and X_i as the features, ignoring the presence of the treatment indicator W_i. More precisely,

we build a regression tree using the MSE for both in-sample and out-of-sample goodness of fit, comparing the actual transformed outcome to the predicted transformed outcome. This Transformed Outcome Tree (TOT) algorithm gives us an estimate of $\mathbb{E}[Y_i^*|X_i = x]$, which is equal to $\tau(x)$ given our assumption of unconfoundedness. We use this as an estimate of the CATE: $\hat{\tau}_{\text{TOT}}(x) = \hat{\mathbb{E}}[Y_i^*|X_i = x]$.

Note that in practice, unless there is a constant treatment probability for all x, it is necessary to estimate $e(x)$ in order to implement this method, because the definition of Y_i^* involves the propensity score. There is a large literature on methods to do this (see Imbens and Rubin, 2015), but for applications where the dimensionality of x is large, a supervised learning method such as a random forest would be a practical.

To set the stage for the fourth and fifth algorithms, an observation about the TOT method is that within a leaf, the estimate $\hat{\tau}(x)$ is *not* equal to the difference in average outcomes for the treatment group and the control group. To see this, suppose that $e(X)$ is constant and equal to p. If the fraction treated within a particular leaf in the training sample is equal to $p' \neq p$, then using the TOT method, the estimated treatment effect within the leaf will be equal to the treatment sample average within the leaf weighted by p'/p minus the control sample average within the leaf weighted by $(1-p')/(1-p)$. This introduces some variance in the estimated treatment effects that can be avoided. In fact, the TOT algorithm may as a result end up with leaves that contain only a single unit, so that estimates of the average treatment effect within such leaves are questionable. Our fourth and fifth methods will address this issue. In our view the primary reason to use the TOT method rather than those that follow is that it can be implemented using existing supervised algorithms "off-the-shelf" without need for additional coding.

4.4 The Causal Tree-Transformed Outcome Algorithm

The fourth algorithm, the Causal Tree-Transformed Outcome (CT-TO) algorithm, differs from the third in one key aspect: the method of estimating average treatment effects within a leaf uses the difference of average outcomes, weighted by the inverse propensity score, between treatment and control observations in a given leaf as the estimator of treatment effects. This is an efficient estimator following Hirano, Imbens and Ridder (2003).

To be more precise, if the tree defines with partition of the covariate space with leaf l corresponding to $\mathbb{X}_l$, the estimated treatment effect for all units with $X_i \in \mathbb{X}_l$ is calculated as

$$\hat{\tau}(X_i) = \frac{\sum_{j:X_j \in \mathbb{X}_l} Y_i^{\text{obs}} \cdot W_i/\hat{e}(X_i)}{\sum_{j:X_j \in \mathbb{X}_l} W_i/\hat{e}(X_i)} - \frac{\sum_{j:X_j \in \mathbb{X}_l} Y_i^{\text{obs}} \cdot (1-W_i)/(1-\hat{e}(X_i))}{\sum_{j:X_j \in \mathbb{X}_l} (1-W_i)/(1-\hat{e}(X_i))}.$$

The difference with the third algorithm is that the average effect within the leafs is not estimated as the average of $Y_i^* = Y_i^{\text{obs}}(W_i - \hat{e}(X_i))/(\hat{e}(X_i)(1-\hat{e}(X_i))$ within the leafs, but as the difference in average outcomes by treatment status. The answers will be different when there is variation in $\hat{e}(X_i)$ within the leaf. In the special case where we have a randomized experiment, and estimate the propensity score as $\hat{e}(X_i) = \sum_{i=1}^N W_i/N$, the estimator reduces to taking the difference in sample averages. In that simple case the conceptual advantage over

the TOT algorithm is clear: the difference in average outcomes by treatment status is a more credible estimator for the average treatment effect than the average of Y_i^*.

4.5 The Causal Tree Method

The fifth algorithm, the Causal Tree (CT) algorithm, is identical to the CT-TO algorithm except that it modifies the in-sample goodness-of-fit measure to be $Q^{\text{is,sq}}$, so that the transformed outcome is only used in the step where cross-validation is used to select the tuning parameter (the penalty for leaves).

5 Inference

Given the estimated conditional average treatment effect we also would like to do inference. For the last three algorithms, the TOT, CT-TO, and CT trees, this is particularly straightforward, because the construction of the tree can be easily decoupled from the problem of conducting inference on treatment effects. Once constructed, the tree is a function of covariates, and if we use a distinct sample (the test sample, for example) to conduct inference, then the problem reduces to that of estimating treatment effects in each member of a partition of the covariate space. For this problem, standard approaches are valid.

Begin by considering the simpler case with complete randomization where the propensity score is constant ($e(x) = p$). Conditional on the tree $\mathcal{T}$, consider the leaf corresponding to subset $\mathbb{X}_m$. Within this leaf the average treatment effect is

$$\tau_{\mathbb{X}_m} = \mathbb{E}[Y_i(1) - Y_i(0) | X_i \in \mathbb{X}_m].$$

Because of the randomization, we can view the the data for the subset of the test sample with features in this subset of the feature space as arising from a completely randomized experiment. Hence the difference in average outcomes by treatment status is unbiased for the average effect in this subset of the feature space, and we can estimate the variance without bias using the sample variance of the treated and control units in this subset.

To be specific, let $\overline{Y}_{t,m}^{\text{te,obs}}$ and $\overline{Y}_{c,m}^{\text{te,obs}}$ be the average outcome for treated and control units in leaf m in the test sample, let $N_{t,m}^{\text{te}}$ and $N_{c,m}^{\text{te}}$ be the number of treated and control units in this leaf, and let $S_{t,m}^{\text{te,2}} = \sum_{S_i=m}(Y_i^{\text{obs}} - \overline{Y}_{t,m})^2/(N_{t,m}-1)$ and $S_{c,m}^{\text{te,2}} = \sum_{S_i=m}(Y_i^{\text{obs}} - \overline{Y}_{c,m})^2/(N_{c,m}-1)$ be the sample variances. Then conditional on the training sample, the variance of $\hat{\tau}_{\mathbb{X}_m}$ is $S_{t,m}^{\text{te,2}}/N_{t,m}^{\text{te}} + S_{c,m}^{\text{te,2}}/N_{c,m}^{\text{te}}$. For the estimates of the treatment effects within the leafs based on the training sample these variances are not necessarily valid because the tree is constructed on the training sample; in particular, the construction of the tree will be sensitive to sampling variation in treatment effects, so that the algorithm is likely to create a leaf where the sample average treatment effect is more extreme than its population value.

Outside of the setting of simple randomized assignment, under the unconfoundedness assumption, the TOT, CT-TO and CT methods estimate treatment effects using inverse propensity score methods. Hirano, Imbens, and Ridder (2003) establish conditions under which the estimated treatment effects are asymptotically normal. For those conditions to be satisfied in

this application, it is necessary to restrict the size of the leaves of the tree relative to the size of the sample. One could also use other methods for estimating treatment effects within the leaves on the test sample, including matching (Abadie and Imbens, 2006; for a general survey see Imbens and Rubin (2015)). Matching is computationally costly, which is particularly problematic during training, but may be less of a concern for a single application in the test sample.

6 An Application

To compare the methods discussed in the paper we first apply them to data from a randomized experiment designed to evaluate the placement of search results on the Bing search engine, and conducted by Microsoft.[2] The unit of analysis is a search query. As search queries arrive, the search engine, Bing here, but the same applies to other search engines, produces a number of ranked search results. The status quo is that the top search results are reported on the screen in the order they were ranked by the search engine. The treatment studied in the original experiments consists of moving the highest ranked search result from the top place where it would normally be displayed to the third place on the screen. This is expected to reduce the rate at which the search result gets clicked on, which is the outcome we focus on. Queries that were classified as being highly navigational in nature (that is, the historical click share on a single link is very high) were omitted from the study to preserve the user experience.

The substantive question is whether this reduction in click-through rate varies by characteristics of the search query. Many of the features of the search query used in our analysis were created through a series of classifiers developed for a variety of purposes by the researchers at the search engine. These include scores of whether the query is about consumer electronics, celebrities, clothes and shoes, whether it is about movie times, whether there are related images available, whether the query is a good match for a Wikipedia entry, etc. Other features pertain to the source of the query, such as the operating system and type of device that accessed the search engine.

We analyze data for 499,486 search queries, with 60 pre-treatment variables or features. The simple experimental estimate of overall average effect on click-through rate based on differencing average outcomes in treatment and control groups is -0.13, which is very precisely estimated with a standard error of 0.0015.

In Table 1 we present the values of the out-of-sample goodness-of-fit measures, evaluated on the test samples. The modified causal tree performs best according to both criteria, although the differences are modest. Part of this is because a common component of the first goodness-of-fit measure for all algorithm includes the variance of Y_i^*, which is substantial. The TOT, CT-TO, and CT trees that only split based on estimates of the average treatment effect lead to relatively small number of leaves, making them easier to display and interpret. In Table 2 we present for each of the 24 leaves of the modified causal tree the estimates of the average causal effect, it standard error, and the size of the leaf, both in the training sample, where the

[2]We do not use the full data set here for confidentiality reasons. The subset of the data used here is non-representative, selected by dropping observations with a probability that is a function of the covariates so that the randomization is valid on the subset used here.

Table 1: Goodness of Fit Results for Search Data

	Single Tree	Two Trees	Transf. Outc. Tree	Causal Tree-TO	Causal Tree
$Q^{\text{os,TO}}$	0.8053	0.8053	0.8046	0.8048	**0.8044**
$Q^{\text{os,match}}$	0.3111	0.3111	0.3107	0.3106	**0.3105**
# Leaves	52	36 26	21	21	24

inferences are not necessarily valid because the tree is constructed on that sample, and for the test sample, where the inferences are valid. The results are also shown in Figure 1.

It is interesting to look at some of the specific results. For Leaf 3 the estimated effect is -0.0099 (s.e. 0.0044), close to zero, with the proportion of search queries in this leaf equal to 0.0129. This leaf corresponds to search queries that are likely to pertain to celebrities and match with images, where it is not surprising that the actual placement of the first ranked answer does not matter much. This type of query is likely to trigger a box with celebrity images and other information, and so it is not surprising that the first link is unimportant.

For Leaf 4 the estimated effect is much bigger: -0.2148 (s.e. 0.0128), with the proportion of search queries in this leaf equal to 0.0214. These are searches that do not have good image matches, not likely to be a search bot, are more "navigational" in nature, match well with Wikipedia articles. Here the placement of the first ranked result is very important. In this case it is likely that the search query represents a genuine request for information, in which case the ranking may be viewed by the searcher as informative about the relevance of the result.

Also observe that the standard deviation of estimated treatment effects is substantially higher in the training sample (0.0281) than in the test sample (0.0178). As discussed in the section on inference, this is expected, because the construction of the tree makes it more likely to create a leaf on a subsample where, due to sampling variation, the treatment effect is unusually extreme.

7 The Literature

A small but growing literature seeks to apply supervised machine learning techniques to the problem of estimating heterogeneous treatment effects.

Foster, Taylor and Ruberg (2010) estimate $\mu(0,x) = \mathbb{E}[Y_i(0)|X_i = x]$ and $\mu(1,x) = \mathbb{E}[Y_i(1)|X_i = x]$ (both using random forests), then calculate $\hat{\tau}_i = \hat{\mu}(1,X_i) - \hat{\mu}(0,X_i)$. They then use machine learning algorithms to estimate $\hat{\tau}_i$ as a function of the units' attributes, X_i. Our approach differs in that we apply machine learning methods directly to the treatment effect in a single stage procedure.

Imai and Ratkovic (2013) consider multi-valued treatments but focus on the purely experi-

Table 2: Estimated Trees on Search Data

leaf	test sample			training sample		
	est	se	share	est	se	share
1	-0.1235	0.0036	0.2022	-0.1236	0.0036	0.2018
2	-0.1339	0.0099	0.0247	-0.1349	0.0102	0.0240
3	-0.0099	0.0044	0.0129	-0.0073	0.0044	0.0132
4	-0.2148	0.0128	0.0214	-0.2467	0.0126	0.0216
5	-0.1453	0.0030	0.3049	-0.1480	0.0030	0.3044
6	-0.1109	0.0056	0.0628	-0.1100	0.0055	0.0635
7	-0.2303	0.0283	0.0036	-0.2675	0.0284	0.0037
8	-0.0575	0.0096	0.0165	-0.0324	0.0095	0.0168
9	-0.0868	0.0307	0.0026	-0.0559	0.0294	0.0025
10	-0.1505	0.0048	0.1191	-0.1693	0.0047	0.1191
11	-0.1741	0.0236	0.0045	-0.1682	0.0239	0.0046
12	0.0255	0.1267	0.0003	0.2857	0.1235	0.0002
13	-0.0297	0.0264	0.0019	-0.0085	0.0250	0.0022
14	-0.1352	0.0142	0.0106	-0.1139	0.0147	0.0100
15	-0.1591	0.0552	0.0010	-0.1432	0.0526	0.0011
16	-0.0135	0.0260	0.0005	0.0080	0.0502	0.0004
17	-0.0809	0.0118	0.0131	-0.0498	0.0124	0.0132
18	-0.0453	0.0231	0.0014	-0.0454	0.0208	0.0014
19	-0.1694	0.0158	0.0105	-0.1997	0.0162	0.0106
20	-0.2072	0.0304	0.0031	-0.2790	0.0305	0.0030
21	-0.0955	0.0106	0.0226	-0.0834	0.0108	0.0223
22	-0.0955	0.0053	0.0690	-0.0956	0.0052	0.0699
23	-0.1392	0.0126	0.0131	-0.1587	0.0129	0.0131
24	-0.1309	0.0056	0.0777	-0.1281	0.0057	0.0776

mental setting. In a simplified case they focus on

$$Y_i^{\text{obs}} = \beta_0 + \tau \cdot W_i + \sum_{k=1}^{K} Z_i \cdot \beta_k + \sum_{k=1}^{K} Z_i \cdot W_i \cdot \gamma_k.$$

They then use Lasso on both sets of coefficients, β and γ, but with different penalty terms to allow for the possibility that the treatment effects are present but the magnitudes of the interactions are small. Their approach is similar to ours in that they distinguish between the estimation of treatment effects and the estimation of the impact of other attributes of units.

Taddy, Gardner, Chen, and Draper (2015) consider a model with the outcome linear in the covariates and the interaction. Initially unconstrained, this leads to an estimate of $\tau(x) = \hat{\gamma}'x$. They then project this down onto the feature space, by minimizing $|\hat{\gamma}'x - \delta'x|$ with a penalty term that is linear in the number of non-zero δ's.

Zeileis, Hothorn, and Hornik (2005) develop a procedure they call "model-based recursive

partitioning" whereby they develop a tree-based method for estimating parametric models on subsets of the data. At each leaf of the tree, they propose to estimate a model such as a linear regression or a maximum-likelihood based model. Leaves are split further using a test for parameter stability; the feature with the highest instability is chosen for a split. In terms of a method for building a tree, this approach is similar to several of our proposed methods, in that in our methods, we estimate a simple model (estimate treatment effects) within leaves of the tree, and we split the leaves when we find covariates that lead to different parameter estimates within the splits (heterogeneous treatment effects). Zeileis, Hothorn, and Hornick (2008) differ in that they base the split on model fit, and in that they do not consider cross-validation for selecting a complexity tuning parameter, so the issue of selecting an out-of-sample goodness of fit metric does not arise. In addition, our in-sample goodness-of-fit measure differs.

Su, Tsai, Wang, Nickerson, and Li (2009) construct a tree to estimate treatment effects. The splitting is based on the t-statistic for the test of no difference between the two groups. The cross-validation of the overall tree is based on the sum of the squares of the split t-statistics, with a penalty term that is linear in the number of splits. Our splitting criteria is conceptually similar, but our approach considers alternative loss functions for cross-validation that directly assess goodness of fit of estimated treatment effects.

We discussed Beygelzimer and Langford (2009) above in Remark II. Dudick, Langford, and Li (2011) propose a related appoach for finding the optimal treatment policy that combines inverse propensity score methods with "direct methods" (e.g. the "single tree" approach considered above) that predict the outcome as a function of the treatment and the unit attributes. The methods can be used to evaluate the average difference in outcomes from any two policies that map attributes to treatments, as well as to select the optimal policy function. They do not focus on hypothesis testing for heterogeneous treatment effects, and they use conventional approaches for cross-validation.

Another line of work that is similar in motivation to ours is the work on Targeted Maximum Likelihood (van der Laan and Rose, 2011). This approach is also motivated by the idea that estimation approaches should be tailored towards the parameters of interest, so that bias/variance tradeoffs are optimized for the most important parameters. Targeted maximum likelihood approaches modify the loss function to increase the weight on the parts of the likelihood that concern the parameters of interest. The methods are semi-parametric, and importantly they rely on a parametric model for the parameters of interest. Our approach is focused on nonparametric estimates of $\tau(x)$ in an environment where there may be a large number of covariates x; thus, a parametric model is likely to be restrictive and it may not be possible to incorporate all covariates in "big data" settings where the covariate space is large relative to the sample size.

8 Conclusion

In this paper we introduce new methods for constructing trees for causal effects that allow us to do valid inference for the causal effects. Our methods partition the feature space into subspaces. The output of our method is a set of treatment effects and confidence intervals for each subspace.

We apply these methods to data from an experiment in which the placement of results for search queries was altered, and find that the algorithms lead to plausible and interpretable results about the heterogeneity of the effect of placement.

A potentially important application of the techniques is to "data-mining" in randomized experiments. Our method can be used to explore any previously conducted randomized controlled trial, for example, medical studies or field experiments in developed economics. A researcher can apply our methods and discover subpopulations with lower-than-average or higher-than-average treatment effects, and can report confidence intervals for these estimates without concern about multiple testing.

References

Beygelzimer, A., and J. Langford, (2009), "The Offset Tree for Learning with Partial Labels," http://arxiv.org/pdf/08

Breiman, L. (2001), "Random forests," *Machine Learning*, 45, 532.

Breiman, L., J. Friedman, R. Olshen, and C. Stone, (1984), *Classification and Regression Trees*, Wadsworth.

Crump, R., J. Hotz, G. Imbens, and O. Mitnik, (2008), "Nonparametric Tests for Treatment Effect Heterogeneity", *Review of Economics and Statistics*, 90(3):389-405.

Dudik, M., J. Langford and L. Li, (2011), "Doubly Robust Policy Evaluation and Learning ", *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*.

Foster, J., J. Taylor and S. Ruberg, (2010), "Subgroup Identification from Randomized Clinical Data ", *Statistics in Medicine*, 30, 2867-2880.

Green, D., and H. Kern, (2010), "Detecting Heterogeneous Treatment Effects in Large-Scale Experimetns Using Bayesian Additive Regression Trees", Unpublished Manuscript, Yale University.

Hahn, J., (1998) "On the Role of the Propensity Score in Efficient Semiparametric Estimation of Average Treatment Effects," *Econometrica* 66 (2), 315-331.

Hastie, T., R. Tibshirani, and J. Friedman, (2011), *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*, Second Edition, Springer.

Hernán, M., and J. Robins, (2015), *Causal Inference*, Chapman and Hall.

Hirano, K., G. Imbens and G. Ridder, (2003), "Efficient Estimation of Average Treatment Effects Using the Estimated Propensity Score", *Econometrica*, 71 (4), 1161-1189.

Holland, P., (1986), "Statistics and Causal Inference"(with discussion), *Journal of the American Statistical Association*, 81, 945-970.

Imai, K., and M. Ratkovic, (2013), "Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation", *Annals of Applied Statistics*, Vol. 7(1): 443-470.

Imbens, G., and D. Rubin, (2015), *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*, Cambridge University Press.

Kehl, V., Ulm, K., (2006), "Responder identification in clinical trials with censored data", *Computational Statistics and Data Analysis* 50(5): 1338-1355.

Morgan, S., and C. Winship, (2002), *Counterfactuals and Causal Inference: Methods and Principles for Social Research*, Cambridge University Press.

Pearl, J., (2000), *Causality: Models, Reasoning and Inference*, Cambridge University Press.

Rosenbaum, P., (2002), *Observational Studies*, Springer.

Rosenbaum, P., (2009), *Design of Observational Studies*, Springer.

Rosenbaum, P., and D. Rubin, (1983), "The Central Role of the Propensity Score in Observational Studies for Causal Effects", *Biometrika*, 70, 41-55.

Rosenblum, M, and M. Van Der Laan., (2011), "Optimizing Randomized Trial Designs to Distinguish which Subpopulations Benefit from Treatment ", *Biometrika*, 98(4): 845-860.

Rubin, D. (1974), "Estimating Causal Effects of Treatments in Randomized and Non-randomized Studies," *Journal of Educational Psychology*, 66, 688-701.

Rubin, D. B., (1978), "Bayesian inference for causal effects: The Role of Randomization," *Annals of Statistics*, 6:34–58.

Su, X., C. Tsai, H. Wang, D. Nickerson, and B. Li, (2009), "Subgroup Analysis via Recursive Partitioning", *Journal of Machine Learning Research*, 10, 141-158.

Taddy, M., M. Gardner, L. Chen, and D. Draper,, (2015), "Heterogeneous Treatment Effects in Digital Experimentation," Unpublished Manuscript, arXiv:1412.8563.

Vapnik, V., (1998), *Statistical Learning Theory*, Wiley.

Vapnik, V., (2010), *The Nature of Statistical Learning Theory*, Springer.

Zeileis, A., T. Hothorn, and K. Hornik (2008) "Model-based recursive partitioning." *Journal of Computational and Graphical Statistics*, 17(2), 492-514.