Recursive Partitioning for Heterogeneous Causal Effects*

Susan Athey† Guido W. Imbens‡

First Draft: October 2013
This Draft: December 2015

Abstract

In this paper we propose methods for estimating heterogeneity in causal effects in experimental and observational studies, and for conducting hypothesis tests about the magnitude of the differences in treatment effects across subsets of the population. We provide a data-driven approach to partition the data into subpopulations which differ in the magnitude of their treatment effects. The approach enables the construction of valid confidence intervals for treatment effects, even in samples with many covariates relative to the sample size, and without "sparsity" assumptions. To accomplish this, we propose an "honest" approach to estimation, whereby one sample is used to construct the partition and another to estimate treatment effects for each subpopulation. Our approach builds on regression tree methods, modified to optimize for goodness of fit in *treatment effects* and to account for honest estimation. Our model selection criteria focus on improving the prediction of treatment effects conditional on covariates, anticipating that bias will be eliminated by honest estimation, but also accounting for the change in the variance of treatment effect estimates within each subpopulation as a result of the split. We also address the challenge that the "ground truth" for a causal effect is not observed for any individual unit, so that standard approaches to cross-validation must be modified. Through a simulation study, we show that honest estimation can result in substantial improvements in coverage of confidence intervals, where our method attains nominal coverage rates, without much sacrifice in terms of fitting treatment effects.

In this paper we study two closely related problems: first, estimating heterogeneity by covariates or 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 we use here ([19], [11], [14]), are comparisons between outcomes we observe and counterfactual outcomes we would have observed under a different regime or treatment. We introduce data-driven methods that select subpopulations to estimate more precisely average treatment effects

*We are grateful for comments provided at seminars at the National Academy of Science Sackler Colloquium, the Southern Economics Association, the Stanford Conference on Causality in the Social Sciences, the MIT Conference in Digital Experimentation, Harvard, University of Washington, Cornell, Microsoft Research, Facebook, KDD, the AAAI Embedded Machine Learning Conference, the University of Pennsylvania, the University of Arizona. 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

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. The supervised machine learning literature (e.g. [9]) has developed a variety of effective methods for a closely related problem, the problem of predicting outcomes as a function of covariates in similar environments. The most popular approaches (e.g. regression trees ([4]), random forests ([3]), LASSO ([24]), support vector machines ([26]), etc.) entail building a model of the relationship between attributes and outcomes, with a penalty parameter that penalizes model complexity. Cross-validation is often used to select the optimal level of complexity (the one that maximizes predictive power without "overfitting").

Within the prediction-based machine learning literature, regression trees differ from most other methods in that they produce a partition of the population according to covariates, whereby all units in a partition receive the same prediction. In this paper, we focus on the analogous goal of deriving a partition of the population according to treatment effect heterogeneity, building on standard regression trees ([4], [3]). Whether the ultimate goal in an application is to derive a partition or fully personalized treatment effect estimates depends on the setting; settings where partitions may be desirable include those where decision rules must be remembered, applied or interpreted by human beings or computers with limited processing power or memory. Examples include treatment guidelines to be used by physicians or even online personalization applications where having a simple lookup table reduces latency for the user. We show that an attractive feature of focusing on partitions is that we can achieve nominal converage of confidence intervals for estimated treatment effects even in settings with a modest number of observations and many covariates. Our approach has applicability even for settings such as clinical trials of drugs with only a few hundred patients, where the number of patient characteristics is potentially quite large. Our method may also be viewed as a complement to the use of "pre-analysis plans" where the researcher must commit in advance to the subgroups that will be considered for analysis. It enables researchers to let the data discover relevant subgroups without falling prey to concerns of multiple hypothesis testing that would invalidate p-values.

A first challenge for our goal of finding a partition and then testing hypotheses about treatment effects is that many existing machine learning methods cannot be used directly for constructing confidence intervals. This is because the methods are "adaptive": they use the training data for model selection, so that spurious correlations between covariates and outcomes affect the selected model, leading to biases that disappear only slowly as the sample size grows. In some contexts, additional assumptions such as "sparsity" (only a few covariates affect the outcomes) can be applied to guarantee consistency or asymptotic normality of predictions ([28]). In this paper, we use an alternative approach that places no restrictions on model complexity, which we refer to as "honesty." We say that a model is "honest" if it does not use the same information for selecting the model structure (in our case, the partition of the covariate space) and for estimation given a model structure. We accomplish this by splitting the training sample into two parts, one for constructing the tree (including the cross-validation step), and a second for estimating treatment effects within leaves of the tree, implying that the asymptotic

properties of treatment effect estimates within the partitions are the same as if the partition had been exogenously given. Although there is a loss of precision due to sample splitting (which reduces sample size in each step of estimation), there is a benefit for fit in terms of eliminating bias that offsets at least part of the cost.

A novel contribution of this paper is to show that criteria for both constructing the partition and cross-validation change when we anticipate honest estimation. In the first stage of estimation, the criteria is the expectation of the mean-squared error when treatment effects are re-estimated in the second stage. Crucially, we anticipate that second-stage estimates of treatment effects will be unbiased in each leaf, since they will be performed on a separate (and independent) sample. In that case, splitting and cross-validation criteria are adjusted to ignore systematic bias in estimation, and focus instead on the tradeoff between more tailored prediction (smaller leaf size) and the variance that will arise in the second (honest estimation) stage due to noisy estimation within small leaves.

A second and perhaps more fundamental challenge to applying machine learning methods such as regression trees [4] off-the-shelf to the problem of causal inference is that regularization approaches based on cross-validation typically rely on observing the "ground truth," that is, actual outcomes in a cross-validation sample. However, if our goal is to minimize the mean-squared error of treatment effects, we encounter what [11] calls the "fundamental problem of causal inference": the causal effect is not observed for any individual unit, and so we don't directly have a ground truth. We address this by proposing approaches for constructing unbiased estimates of the mean-squared error of the causal effect of the treatment.

Using theoretical arguments and a simulation exercise we evaluate the costs and benefits of honest estimation and compare our approach with previously proposed ones. We find in the settings we consider that honest estimation results in improvements in fit, in some cases very large improvements, over a more traditional "adaptive" estimation approach.

1 The Problem

1.1 The Set Up

We consider a setup where there are N units, indexed by $i = 1, \ldots, N$. 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 [19], [11], [14], with the unit-level causal effect defined as the difference in potential outcomes, $\tau_i = Y_i(1) - Y_i(0)$. 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. The realized outcome for unit i is the potential outcome corresponding to the treatment received:

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

Let X_i be a K-component vector of features, covariates or pretreatment variables, known not to be affected by the treatment. 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 a large population. Expectations and probabilities will refer to the distribution induced by the random sampling, or by the (conditional) random assignment of the treatment. We assume that observations are exchangeable, and that there is no interference (the stable unit treatment value assumption, or sutva [20]). This assumption may be violated in settings where some units are connected through networks. Let

$p = \text{pr}(W_i = 1)$ 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 [17]). In a randomized experiment with constant treatment assignment probabilities $e(x) = p$ for all values of x.

1.2 Unconfoundedness

Throughout the paper, we maintain the assumption of randomization conditional on the covariates, or "unconfoundedness" ([17]), formalized as:

Assumption 1. (Unconfoundedness)

$$W_i \perp\!\!\!\perp \Big(Y_i(0), Y_i(1)\Big) \;\Big|\; X_i.$$

This assumption, sometimes referred to as "selection on observables" in the econometrics literature, is satisfied in a randomized experiment without conditioning on covariates, but also may be justified in observational studies if the researcher is able to observe all the variables that affect the unit's receipt of treatment and are associated with the potential outcomes.

To simplify exposition, in the main body of the paper we maintain the stronger assumption of *complete randomization,* whereby $W_i \perp\!\!\!\perp (Y_i(0), Y_i(1), X_i)$. Later we show that by using propensity score weighting [19], we can adapt all of the methods to that case.

1.3 Conditional Average Treatment Effects and Partitioning

Define the conditional average treatment effect (CATE)

$$\tau(x) \equiv \mathbb{E}[Y_i(1) - Y_i(0)|X_i = x].$$

A large part of the causal inference literature (e.g. [14], [15]) is focused on estimating the population (marginal) average treatment effect $\mathbb{E}[Y_i(1) - Y_i(0)]$. The main focus of the current paper is on obtaining accurate estimates of and inferences for the conditional average treatment effect $\tau(x)$. We are interested in estimators $\hat{\tau}(\cdot)$ that are based on partitioning the feature space, and do not vary within the partitions.

2 Honest Inference for Population Averages

Our approach departs from conventional classification and regression trees (CART) in two fundamental ways. First, we focus on estimating conditional average treatment effects rather than predicting outcomes. This creates complications for conventional methods because we do not observe unit level causal effects for any unit. Second, we impose a separation between constructing the partition and estimating effects within leaves of the partition, using separate samples for the two tasks, in what we refer to as "honest" estimation. We contrast ''honest" estimation with "adaptive" estimation used in conventional CART, where the same data is used to build the partition and estimate leaf effects. In this section we introduce the changes induced by honest estimation in the context of the conventional prediction setting; in the next section we consider causal effects. In the discussion in this section we observe for each unit i a pair of variables (Y_i, X_i), with the interest in the conditional expectation $\mu(x) \equiv \mathbb{E}[Y_i|X_i = x]$.

2.1 Set Up

We begin by defining key concepts and functions. First, a tree or partitioning Π corresponds to a partitioning of the feature space $\mathbb{X}$, with $\#(\Pi)$ the number of elements in the partition. We write

$$\Pi = \{\ell_1, \ldots, \ell_{\#(\mathcal{T})}\}, \quad \text{with } \cup_{j=1}^{\#(\Pi)} \ell_j = \mathbb{X}.$$

Let $\mathbb{P}$ denote the space of partitions. Let $\ell(x;\Pi)$ denote the leaf $\ell \in \Pi$ such that $x \in \ell$. Let $\mathbb{S}$ be the space of data samples from a population. Let $\pi : \mathbb{S} \mapsto \mathbb{P}$ be an algorithm that on the basis of a sample $\mathcal{S} \in \mathbb{S}$ constructs a partition. As a very simple example, suppose the feature space is $\mathbb{X} = \{L, R\}$. In this case there are two possible partitions, $\Pi_N = \{L, R\}$ (no split), or $\Pi_S = \{\{L\}, \{R\}\}$ (full split), and so the space of trees is $\mathbb{P} = \{\Pi_N, \Pi_S\} = \underline{\{\{L, R\}, \{\{L\}, \{R, \}\}\}}$. Given a sample $\mathcal{S}$, the average outcomes in the two subsamples are $\overline{Y}_L$ and $\overline{Y}_R$. A simple example of an algorithm is one that splits if the difference in average outcomes exceeds a threshold c:

$$\pi(\mathcal{S}) = \begin{cases} \{\{L, R\}\} & \text{if } \overline{Y}_L - \overline{Y}_R \leq c, \\ \{\{L\}, \{R\}\} & \text{if } \overline{Y}_L - \overline{Y}_R > c. \end{cases}$$

The potential bias in leaf estimates from adaptive estimation can be seen in this simple example. While $\overline{Y}_L - \overline{Y}_R$ is in general an unbiased estimator for the difference in the population conditional means $\mu(L) - \mu(R)$, if we condition on finding that $\overline{Y}_L - \overline{Y}_R \geq c$ in a particular sample, we expect that $\overline{Y}_L - \overline{Y}_R$ is larger than the population analog.

Given a partition Π, define the conditional mean function $\mu(x;\Pi)$ as

$$\mu(x;\Pi) \equiv \mathbb{E}[Y_i | X_i \in \ell(x;\Pi)] = \mathbb{E}[\mu(X_i) | X_i \in \ell(x;\Pi)],$$

which can be viewed as a step-function approximation to $\mu(x)$. Given a sample $\mathcal{S}$ the estimated counterpart is

$$\hat{\mu}(x;\mathcal{S},\Pi) \equiv \frac{1}{\#(i \in \mathcal{S} : X_i \in \ell(x;\Pi))} \sum_{i \in \mathcal{S}: X_i \in \ell(x;\Pi)} Y_i,$$

which is unbiased for $\mu(x;\Pi)$. We index this estimator by the sample because we need to be precise which sample is used for estimation of the regression function.

2.2 The Honest Target

A central concern in this paper is the criterion used to compare alternative estimators; following much of the literature, we focus on Mean-squared error (MSE) criteria, but we will modify these criteria in a variety of ways. For the prediction case, we adjust the MSE by $\mathbb{E}[Y_i^2]$; since this does not depend on an estimator, subtracting it does not affect how the criterion ranks estimators. Given a partition Π, define the mean squared error, where we average over a test sample $\mathcal{S}^{\text{te}}$ and the conditional mean is estimated on an estimation sample $\mathcal{S}^{\text{est}}$, as

$$\text{MSE}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi) \equiv \frac{1}{\#(\mathcal{S}^{\text{te}})} \sum_{i \in \mathcal{S}^{\text{te}}} \left\{ \left(Y_i - \hat{\mu}(X_i; \mathcal{S}^{\text{est}}, \Pi) \right)^2 - Y_i^2 \right\}.$$

The (adjusted) expected mean squared error is the expectation of $\text{MSE}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)$ over the test sample and the estimation sample:

$$\text{EMSE}(\Pi) \equiv \mathbb{E}_{\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}} \left[\text{MSE}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi) \right],$$

where the test and estimation samples are independent. In the algorithms we consider, we will consider a variety of estimators for the (adjusted) EMSE, all of which take the form of MSE estimators $\text{MSE}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)$, evaluated at the units in sample $\mathcal{S}^{\text{te}}$, with the estimates based on sample $\mathcal{S}^{\text{est}}$ and the tree Π. For brevity in this paper we will henceforth omit the term "adjusted" and abuse terminology slightly by referring to these objects as MSE functions.

Our ultimate goal is to construct and assess algorithms $\pi(\cdot)$ that maximize the "honest" criterion

$$Q^H(\pi) \equiv -\mathbb{E}_{\mathcal{S}^{\text{est}}, \mathcal{S}^{\text{est}}, \mathcal{S}^{\text{tr}}} \left[\text{MSE}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi(\mathcal{S}^{\text{tr}})) \right].$$

Note that throughout the paper we focus on maximixing criterion functions, which typically involve the negative of mean-squared-error expressions.

2.3 The Adaptive Target

In the conventional CART approach the target is slightly different:

$$Q^C(\pi) \equiv -\mathbb{E}_{\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{tr}}} \left[\text{MSE}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{tr}}, \pi(\mathcal{S}^{\text{tr}})) \right],$$

where the same training sample is used to construct and estimate the tree. Compared to our target $Q^H(\pi)$ the difference is that in our approach different samples $\mathcal{S}^{\text{tr}}$ and $\mathcal{S}^{\text{est}}$ are used for construction of the tree and estimation of the conditional means respectively. We refer to the conventional CART approach as "adaptive," and our approach as "honest."

In practice there will be costs and benefits of the honest approach relative to the adaptive approach. The cost is sample size; given a data set, putting some data in the estimation sample leaves fewer units for the training data set. The advantage of honest estimation is that it avoids a problem of adaptive estimation, which is that spurious extreme values of Y_i are likely to be placed into the same leaf as other extreme values by the algorithm $\pi(\cdot)$, and thus the sample means (in sample $\mathcal{S}^{\text{tr}}$) of the elements of $\pi(\mathcal{S}^{\text{tr}})$ are more extreme than they would be in an independent sample.

2.4 The Implementation of CART

There are two distinct parts of the conventional CART algorithm, initial tree building and cross-validation to select a complexity parameter used for pruing. Each part of the algorithm relies on a criterion function based on mean-squared error. In this paper we will take as given the overall structure of the CART algorithm (e.g., [4], [9]), and our focus will be on modifying the criteria.

In the tree-building phase, CART recursively partitions the observations of the training sample. For each leaf, the algorithm evaluates all candidate splits of that leaf (which induce alternative partitions Π) using a "splitting" criterion that we refer to as the "in-sample" goodness of fit criterion $-\text{MSE}(\mathcal{S}^{\text{tr}}, \mathcal{S}^{\text{tr}}, \Pi)$. It is well-understood that the conventional criterion leads to "over-fitting," a problem that is solved by cross-validation to select a penalty on tree depth. The in-sample goodness of fit criterion will always improve with additional splits, even though

additional refinements of a partition Π might in fact increase the expected mean squared error, especially when the leaf sizes become small. The reason is that the criterion ignores the fact that smaller leaves lead to higher-variance estimates of leaf means.

To account for this factor, the conventional approach to avoiding "overfitting" is to add a penalty term to the criterion that is equal to a constant times the number of splits, so that essentially we only consider splits where the improvement in a goodness-of-fit criterion is above some threshold. The penalty term is choosen to maximize a goodness of fit criterion in cross-validation samples. In the conventional cross-validation the training sample is repeatedly split into two subsamples, the $\mathcal{S}^{\text{tr,tr}}$ sample that is used to build a new tree as well as estimate the conditional means and the $\mathcal{S}^{\text{tr,cv}}$ sample that is used to evaluate the estimates. We "prune" the tree using a penalty parameter that represents the cost of a leaf. We choose the optimal penalty parameter by evaluating the trees associated with each value of the penalty parameter. The goodness of fit criterion for cross-validation can be written as $-\text{MSE}(\mathcal{S}^{\text{tr,cv}}, \mathcal{S}^{\text{tr,tr}}, \Pi)$. Note that the cross-validation criterion directly addresses the issue we highlighted with the in-sample goodness of fit criterion, since $\mathcal{S}^{\text{tr,cv}}$ is independent of $\mathcal{S}^{\text{tr,tr}}$, and thus too-extreme estimates of leaf means will be penalized. The issue that smaller leaves lead to noisier estimates of leaf means is implicitly incorporated by the fact that a smaller leaf penalty will lead to deeper trees and thus smaller leaves, and the noisier estimates will lead to larger average MSE across the cross-validation samples.

2.5 Honest Splitting

In our honest estimation algorithm, we modify CART in two ways. First, we use an independent sample $\mathcal{S}^{\text{est}}$ instead of $\mathcal{S}^{\text{tr}}$ to estimate leaf means. Second (and closely related), we modify our splitting and cross-validation criteria to incorporate the fact that we will generate unbiased estimates using $\mathcal{S}^{\text{est}}$ for leaf estimation (eliminating one aspect of over-fitting), where $\mathcal{S}^{\text{est}}$ is treated as a random variable in the tree building phase. In addition, we explicitly incorporate the fact that finer partitions generate greater variance in leaf estimates.

To begin developing our criteria, let us expand $\text{EMSE}(\Pi)$:

$$
\begin{aligned}
&-\text{EMSE}(\Pi) = -\mathbb{E}_{(Y_i,X_i),\mathcal{S}^{\text{est}}}[(Y_i - \mu(X_i;\Pi))^2 - Y_i^2] \\
&-\mathbb{E}_{X_i,\mathcal{S}^{\text{est}}}\left[\left(\hat{\mu}(X_i;\mathcal{S}^{\text{est}},\Pi) - \mu(X_i;\Pi)\right)^2\right] = \\
&\mathbb{E}_{X_i}[\mu^2(X_i;\Pi)] - \mathbb{E}_{\mathcal{S}^{\text{est}},X_i}\left[\mathbb{V}(\hat{\mu}^2(X_i;\mathcal{S}^{\text{est}},\Pi)\right],
\end{aligned}
$$

where we exploit the equality $\mathbb{E}_{\mathcal{S}}[\hat{\mu}(x;\mathcal{S},\Pi)] = \mu(x;\Pi)$.

We wish to estimate $-\text{EMSE}(\Pi)$ on the basis of the training sample $\mathcal{S}^{\text{tr}}$ and knowledge of the sample size of the estimation sample N^{est}. To construct an estimator for the second term, observe that within each leaf of the tree there is an unbiased estimator for the variance of the estimated mean in that leaf. Specifically, to estimate the variance of $\hat{\mu}(x;\mathcal{S}^{\text{est}},\Pi)$ on the training sample we can use

$$
\widehat{\mathbb{V}}(\hat{\mu}(x;\mathcal{S}^{\text{est}},\Pi)) \equiv \frac{S^2_{\mathcal{S}^{\text{tr}}}(\ell(x;\Pi))}{N^{\text{est}}(\ell(x;\Pi))},
$$

where $S^2_{\mathcal{S}^{\text{tr}}}(\ell)$ is the within-leaf variance, to estimate the variance. We then weight this by the leaf shares p_ℓ to estimate the expected variance. Assuming the leaf shares are approximately

the same in the estimation and training sample, we can approximate this variance estimator by

$$\widehat{\mathbb{E}}\left[\mathbb{V}(\hat{\mu}^2(X_i;\mathcal{S}^{\text{est}},\Pi)|i\in\mathcal{S}^{\text{te}}\right] \equiv \frac{1}{N^{\text{est}}}\cdot\sum_{\ell\in\Pi} S^2_{\mathcal{S}^{\text{tr}}}(\ell).$$

To estimate the average of the squared outcome $\mu^2(x;\Pi)$ (the first term of the target criterion), we can use the square of the estimated means in the training sample $\hat{\mu}^2(x;\Pi)$, minus an estimate of its variance,

$$\widehat{\mathbb{E}}[\mu^2(x;\Pi)] = \hat{\mu}^2(x;\mathcal{S}^{\text{tr}},\Pi) - \frac{S^2_{\mathcal{S}^{\text{tr}}}(\ell(x;\Pi))}{N^{\text{tr}}(\ell(x;\Pi))}.$$

Combining these estimators leads to the following unbiased estimator for EMSE(Π), denoted $\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr}},N^{\text{est}},\Pi)$:

$$\frac{1}{N^{\text{tr}}}\sum_{i\in\mathcal{S}^{\text{tr}}}\hat{\mu}^2(X_i;\mathcal{S}^{\text{tr}},\Pi) - \left(\frac{1}{N^{\text{tr}}}+\frac{1}{N^{\text{est}}}\right)\cdot\sum_{\ell\in\Pi} S^2_{\mathcal{S}^{\text{tr}}}(\ell).$$

In practice we use the same sample size for the estimation sample and the training sample, so we use as the estimator

$$\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr}},\Pi) \equiv \frac{1}{N^{\text{tr}}}\sum_{i\in\mathcal{S}^{\text{tr}}}\hat{\mu}^2(X_i;\mathcal{S}^{\text{tr}},\Pi) - \frac{2}{N^{\text{tr}}}\cdot\sum_{\ell\in\Pi} S^2_{\mathcal{S}^{\text{tr}}}(\ell).$$

Comparing this to the criterion used in the conventional CART algorithm, which can be written as

$$\text{MSE}(\mathcal{S}^{\text{tr}},\mathcal{S}^{\text{tr}},\Pi) = \frac{1}{N^{\text{tr}}}\sum_{i\in\mathcal{S}^{\text{tr}}}\hat{\mu}^2(X_i;\mathcal{S}^{\text{tr}},\Pi),$$

the difference comes from the terms involving the variance. In the prediction setting the adjustment makes very little difference. Because of the form of the within-leaf sample variances, it follows that the gain from a particular split according to the unadjusted criterion $\text{MSE}(\mathcal{S}^{\text{tr}},\mathcal{S}^{\text{tr}},\Pi)$ is proportional to the gain based on $\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr}},\Pi)$, with the constant of proportionality a function of the leaf size. Thus, in contrast to the treatment effect case discussed below, the variance adjustment does matter much here.

2.6 Honest Crossvalidation

Even though $\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr}},\Pi)$ is approximately unbiased as an estimator of our ideal criterion EMSE(Π) for a fixed Π, it is not unbiased when we use it repeatedly to evaluate splits using recursive partitioning on the training data $\mathcal{S}^{\text{tr}}$. The reason is that initial splits tend to group together observations with similar, extreme outcomes. So after the training data has been divided once, the sample variance of observations in the training data within a given leaf is on average lower than the sample variance would be in a new, independent sample. Thus, $\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr}},\Pi)$ is likely to overstate goodness of fit as we grow a deeper and deeper tree, implying that cross-validation can still play an important role with our honest estimation approach, though perhaps less so than in the conventional CART.

Because the conventional CART cross-validation criterion does not account for honest estimation we consider the analogue of our unbiased estimate of the criterion, which accounts for honest estimation by evaluating a partition Π using only outcomes for units from the cross-validation sample $\mathcal{S}^{\text{tr,cv}}$:

$$-\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr,cv}}, \Pi)$$

This estimator for the honest criterion is unbiased, although it may have higher variance than $\text{MSE}(\mathcal{S}^{\text{tr,cv}}, \mathcal{S}^{\text{tr,tr}}, \Pi)$ due to the small sample size of the cross-validation sample.

3 Honest Inference for Treatment Effects

In this section we change the focus to estimating conditional average treatment effects instead of estimating conditional population means. We refer to the estimators developed in this section as "Causal Tree" (CT) estimators. The setting with treatment effects creates some specific problems because we do not observe the value of the treatment effect whose conditional mean we wish to estimate. This complicates the calculation of the criteria we introduced in the previous section. However, a key point of this paper is that we can estimate these criteria and use those estimates for splitting and cross-validation.

We now observe in each sample the triple $(Y_i^{\text{obs}}, X_i, W_i)$. For a sample $\mathcal{S}$ let $\mathcal{S}_{\text{treat}}$ and $\mathcal{S}_{\text{control}}$ denote the subsamples of treated and control units respectively, with cardinality N_{treat} and N_{control} respectively, and let $p = N_{\text{treat}}/N$ be the share of treated units. The concept of a tree remains the same as in the previous section. Given a tree Π, define for all x and both treatment levels w the population average outcome

$$\mu(w, x; \Pi) \equiv \mathbb{E}\left[Y_i(w) \middle| X_i \in \ell(x; \Pi)\right],$$

and the average causal effect

$$\tau(x; \Pi) \equiv \mathbb{E}\left[Y_i(1) - Y_i(0) \middle| X_i \in \ell(x; \Pi)\right].$$

The estimated counter parts are

$$\hat{\mu}(w, x; \mathcal{S}, \Pi) \equiv \frac{1}{\#(\{i \in \mathcal{S}_w : X_i \in \ell(x; \Pi)\})} \sum_{i \in \mathcal{S}_w : X_i \in \ell(x;\Pi)} Y_i^{\text{obs}},$$

and

$$\hat{\tau}(x; \mathcal{S}, \Pi) \equiv \hat{\mu}(1, x; \mathcal{S}, \Pi) - \hat{\mu}(0, x; \mathcal{S}, \Pi).$$

Define the mean-squared error for treatment effects as

$$\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi) \equiv \frac{1}{\#(\mathcal{S}^{\text{te}})} \sum_{i \in \mathcal{S}^{\text{te}}} \left\{ \left(\tau_i - \hat{\tau}(X_i; \mathcal{S}^{\text{est}}, \Pi)\right)^2 - \tau_i^2 \right\},$$

and define $\text{EMSE}_\tau(\Pi)$ to be its expectation over the estimation and test samples,

$$\text{EMSE}_\tau(\Pi) \equiv \mathbb{E}_{\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}}\left[\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)\right].$$

A key challenge is that the workhorse mean-squared error function $\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \Pi)$ is *infeasible*, because we do not observe the τ_i. However, we show below that we can estimate it.

3.1 Modifying Conventional CART for Treatment Effects

Consider first modifying conventional (adaptive) CART to estimate heterogeneous treatment effects. Note that in the prediction case, using the fact that $\hat{\mu}$ is constant within each leaf, we can write

$$\text{MSE}_\mu(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{tr}}, \Pi) = -\frac{2}{N^{\text{tr}}} \sum_{i \in \mathcal{S}^{\text{te}}} \hat{\mu}(X_i; \mathcal{S}^{\text{te}}, \Pi) \cdot \hat{\mu}(X_i; \mathcal{S}^{\text{tr}}, \Pi)$$

$$+\frac{1}{N^{\text{tr}}} \sum_{i \in \mathcal{S}^{\text{te}}} \hat{\mu}^2(X_i; \mathcal{S}^{\text{tr}}, \Pi).$$

In the treatment effect case we can use the fact that

$$\mathbb{E}_{\mathcal{S}^{\text{te}}}\left[\tau_i | i \in \mathcal{S}^{\text{te}} : i \in \ell(x, \Pi)\right] = \mathbb{E}_{\mathcal{S}^{\text{te}}}\left[\hat{\tau}(x; \mathcal{S}^{\text{te}}, \Pi)\right]$$

to construct an unbiased estimator of $\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{tr}}, \Pi)$:

$$\widehat{\text{MSE}}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{tr}}, \Pi) \equiv -\frac{2}{N^{\text{tr}}} \sum_{i \in \mathcal{S}^{\text{te}}} \hat{\tau}(X_i; \mathcal{S}^{\text{te}}, \Pi) \cdot \hat{\tau}(X_i; \mathcal{S}^{\text{tr}}, \Pi)$$

$$+\frac{1}{N^{\text{tr}}} \sum_{i \in \mathcal{S}^{\text{te}}} \hat{\tau}^2(X_i; \mathcal{S}^{\text{tr}}, \Pi).$$

This leads us to propose, by analogy to CART's in-sample mean-squared error criterion $-\text{MSE}_\mu(\mathcal{S}^{\text{tr}}, \mathcal{S}^{\text{tr}}, \Pi)$,

$$-\widehat{\text{MSE}}_\tau(\mathcal{S}^{\text{tr}}, \mathcal{S}^{\text{tr}}, \Pi) = \frac{1}{N^{\text{tr}}} \sum_{i \in \mathcal{S}^{\text{tr}}} \hat{\tau}^2(X_i; \mathcal{S}^{\text{tr}}, \Pi),$$

as an estimator for the infeasible in-sample goodness of fit criterion.

For cross-validation we used in the prediction case $-\text{MSE}_\mu(\mathcal{S}^{\text{tr,cv}}, \mathcal{S}^{\text{tr,tr}}, \Pi)$. Again, the treatment effect analog is infeasible, but we can use an unbiased estimate of it, which leads to $-\widehat{\text{MSE}}_\tau(\mathcal{S}^{\text{tr,cv}}, \mathcal{S}^{\text{tr,tr}}, \Pi)$.

3.2 Modifying the Honest Approach

The honest approach described in the previous section for prediction problems also needs to be modified for the treatment effect setting. Using the same expansion as before, now applied to the treatment effect setting, we find

$$-\text{EMSE}_\tau(\Pi) = \mathbb{E}_{X_i}[\tau^2(X_i; \Pi)] - \mathbb{E}_{\mathcal{S}^{\text{est}}, X_i}\left[\mathbb{V}(\hat{\tau}^2(X_i; \mathcal{S}^{\text{est}}, \Pi)\right].$$

For splitting we can estimate both components of this expectation using only the training sample. This leads to an estimator for the infeasible criterion that depends only on $\mathcal{S}^{\text{tr}}$:

$$\widehat{\text{EMSE}}_\tau(\mathcal{S}^{\text{tr}}, \Pi) \equiv \frac{1}{N^{\text{tr}}} \sum_{i \in \mathcal{S}^{\text{tr}}} \hat{\tau}^2(X_i; \mathcal{S}^{\text{tr}}, \Pi)$$

$$-\frac{2}{N^{\text{tr}}} \cdot \sum_{\ell \in \Pi} \left(\frac{S^2_{\mathcal{S}^{\text{tr}}_{\text{treat}}}(\ell)}{p} + \frac{S^2_{\mathcal{S}^{\text{tr}}_{\text{control}}}(\ell)}{1-p} \right).$$

For cross-validation we use the same expression, now with the cross-validation sample: $\widehat{\text{EMSE}}_\tau(\mathcal{S}^{\text{tr,cv}}, \Pi)$.

These expressions are directly analogous to the criteria we proposed for the honest version of CART in the prediction case. The criteria reward a partition for finding strong heterogeneity in treatment effects, and penalize a partition that creates variance in leaf estimates. One difference with the prediction case, however, is that in the prediction case, the two terms are proportional; whereas for the treatment effect case they are not. It is possible to reduce the variance of a treatment effect estimator by introducing a split, even if both child leaves have the same average treatment effect, if a covariate affects the mean outcome but not treatment effects. In such a case, the split results in more homogenous leaves, and thus lower-variance estimates of the means of the treatment group and control group outcomes. Thus, the distinction between adaptive and honest splitting criterion will be more pronounced in this case.

The cross-validation criterion estimates treatment effects within leaves using the $\mathcal{S}^{\text{tr,cv}}$ sample rather than $\mathcal{S}^{\text{tr,tr}}$, to account for the fact that leaf estimates will subsequently be constructed using an estimation sample that is independent of the training sample.

4 Four Partitioning Estimators for Causal Effects

In this section we briefly summarize our CT estimator, and then describe three alternative types of estimators. We compare CT to the alternatives theoretically and through simulations. For each of the four types there is an adaptive version and an honest version, where the latter takes into account that estimation will be done on a sample separate from the sample used for constructing the partition, leading to a total of eight estimators. Note that further variations are possible; for example, one could use adaptive splitting and cross-validation methods to construct a tree, but still perform honest estimation on a separate sample. We do not consider those variations in this paper.

4.1 Causal Trees (CT)

The discussion above developed our preferred estimator, Causal Trees. To summarize, for the adaptive version of causal trees, denoted CT-A, we use for splitting the objective $-\widehat{\text{MSE}}(\mathcal{S}^{\text{tr}}, \mathcal{S}^{\text{tr}}, \Pi)$. For cross-validation we use the same objective function, but evaluated at the samples $\mathcal{S}^{\text{tr,cv}}$ and $\mathcal{S}^{\text{tr,tr}}$, namely $-\widehat{\text{MSE}}(\mathcal{S}^{\text{tr,cv}}, \mathcal{S}^{\text{tr,tr}}, \Pi)$. For the honest version, CT-H, the splitting objective function is $-\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr}}, \Pi)$. For cross-validation we use the same objective function, but now evaluated at the cross validation sample, $-\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr,cv}}, \Pi)$.

4.2 Transformed Outcome Trees (TOT)

Our first alternative method is based on the insight that by using a transformed version of the outcome $Y_i^* = (Y_i - W_i)/(p \cdot (1-p))$, it is possible to use off-the-shelf regression tree methods to focus splitting and cross-validation on treatment effects rather than outcomes. Similar approaches are used in [2], [6], [22], and [29]. Because $\mathbb{E}[Y_i^* | X_i = x] = \tau(x)$, off-the-shelf CART methods can be used directly, where estimates of the sample average of Y_i^* within each leaf can be interpreted as estimates of treatment effects. This ease of application is the key attraction of this method. The main drawback (relative to CT-A) is that in general it is not efficient because it does not use the information in the treatment indicator beyond the construction of the transformed outcome. For example, the sample average in $\mathcal{S}$ of Y_i^* within a

given leaf $\ell(x;\Pi)$ will only be equal to $\hat{\tau}(x;\Pi,\mathcal{S})$ if the fraction of treated observations within the leaf is exactly equal to p. Since this method is primarily considered as a benchmark, in simulations we focus only on an adaptive version that can use existing learning methods entirely off-the-shelf. The adaptive version of the transformed outcome tree estimator we consider, TOT-A, uses the conventional CART algorithm with the transformed outcome replacing the original outcome. The honest version, TOT-H, uses the same splitting and cross-validation criteria, so that it builds the same trees; it differs only in that a separate estimation sample is used to construct the leaf estimates. The treatment effect estimator within a leaf is the same as the adaptive method, that is, the sample mean of Y_i^* within the leaf.

4.3 Fit-based Trees (F)

We consider two additional alternative methods for constructing trees, based on suggestions in the literature. In the first of these alternatives the choice of which feature to split on, and at what value of the feature to split, is based on comparisons of the goodness–of–fit of the outcome rather than the treatment effect. In standard CART of course goodness–of–fit of outcomes is also the split criterion, but here we estimate a model for treatment effects within each leaf. Specifically, we have a linear model with an intercept and an indicator for the treatment as the regressors, rather only an intercept as in standard CART. This approach is used in [30], who consider building general models at the leaves of the trees. Treatment effect estimation is a special case of their framework. [30] propose using statistical tests based on improvements in goodness-of-fit to determine when to stop growing the tree, rather than relying on cross-validation, but for ease of comparison to CART, in this paper we will stay closer to traditional CART in terms of growing deep trees and pruning them. We modify the mean-squared error function:

$$\text{MSE}_{\mu,W}(\mathcal{S}^{\text{te}},\mathcal{S}^{\text{est}},\Pi) \equiv \sum_{i\in\mathcal{S}^{\text{te}}} ((Y_i^{\text{obs}} - \hat{\mu}_w(W_i,X_i;\mathcal{S}^{\text{est}},\Pi))^2 - Y_i^2).$$

For the adaptive version F-A we follow conventional CART, using the criterion $-\text{MSE}_{\mu,W}$ in place of $-\text{MSE}$ for splitting, and the analog of $-\widehat{\text{MSE}}(\mathcal{S}^{\text{tr,cv}},\mathcal{S}^{\text{tr,tr}},\Pi)$ with with $\hat{\mu}_w$ in place of $\hat{\mu}$ for cross-validation. For the honest version we use the analogs of $-\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr}},\Pi)$ and $-\widehat{\text{EMSE}}(\mathcal{S}^{\text{tr,cv}},\Pi)$, with $\hat{\mu}_w$ in place of $\hat{\mu}$, for splitting and cross-validation. Similar to the prediction case, the variance term in the honest splitting criterion does not make much of a difference for the choice of splits. An advantage of the fit-based tree approach is that it is a straightforward extension of conventional CART methods. In particular, the mean-squared error criterion is feasible, since Y_i is observed. To highlight the disadvantages of the F approach, consider a case where two splits improve the fit to an equal degree. In one case, the split leads to variation in average treatment effects, and in the other case it does not. The first split would be better from the perspective of estimating heterogeneous treatment effects, but the fit criterion would view the two splits as equally attractive.

4.4 Squared T-statistic Trees (TS)

For the last estimator we look for splits with the largest value for the square of the t-statistic for testing the null hypothesis that the average treatment effect is the same in the two potential leaves. This estimator was proposed by [21]. If the two leaves are denoted L (Left) and R

(Right), the square of the t-statistic is

$$T^2 \equiv N \cdot \frac{(\overline{Y}_L - \overline{Y}_R)^2}{S^2/N_L + S^2/N_R},$$

where S^2 is the conditional sample variance given the split. At each leaf, successive splits are determined by selecting the split that maximizes T^2. The concern with this criterion is that it places no value on splits that improve the fit. While such splits do not deserve as much weight as the fit criterion puts on them, they do have some value.

Both the adaptive and honest versions of the TS approach use T^2 as the splitting criterion. For cross-validation and pruning, it is less obvious how to proceed. [30] suggests that when using a statistical test for splitting, if it is desirable in an application to grow deep trees and then cross-validate to determine depth, then one can use a standard goodness of fit measure for pruning and cross-validation. However, this could undermine the key advantage of TS, to focus on heterogeneous treatment effects. For this reason, we instead propose to use the CT-A and CT-H criteria for cross-validation for TS-A and TS-H, respectively.

4.5 Comparison of the Causal Trees, the Fit Criterion, and the Squared t-statistic Criterion

It is useful to compare our proposed criterion to the F and TS criteria in a simple setting to gain insight into the relative merits of the three approaches. We do so here focusing on a decision whether to proceed with a single possible split, based on a binary covariate $X_i \in \{L, R\}$. Let Π_N and Π_S denote the trees without and with the split, and let $\overline{Y}_w$, $\overline{Y}_{Lw}$ and $\overline{Y}_{Rw}$ denote the average outcomes for units with treatment status $W_i = w$. Let N_w, N_{Lw}, and N_{Rw} be the sample sizes for the corresponding subsamples. Let S^2 be the sample variance of the outcomes given a split, and let $\tilde{S}^2$ be the sample variance without a split. Define the squared t-statistics for testing that the average outcomes for control (treated) units in both leaves are identical,

$$T_0^2 \equiv \frac{(\overline{Y}_{L0} - \overline{Y}_{R0})^2}{S^2/N_{L0} + S^2/N_{R0}}, \quad T_1^2 \equiv \frac{(\overline{Y}_{L1} - \overline{Y}_{R1})^2}{S^2/N_{L1} + S^2/N_{R1}}.$$

Then we can write the improvement in goodness of fit from splitting the single leaf into two leaves as

$$F = \tilde{S}^2 \cdot \frac{2 \cdot (T_0^2 + T_1^2)}{1 + 2 \cdot (T_0^2 + T_1^2)/N}.$$

Ignoring degrees-of-freedom correctcions, the change in our proposed criterion for the honest version of the causal tree in this simple setting can be written as a combination of the F and TS criteria:

$$\widehat{\mathrm{EMSE}}_\tau(\mathcal{S}, \Pi_N) - \widehat{\mathrm{EMSE}}_\tau(\mathcal{S}, \Pi_S) = \frac{(T^2 - 4)(\tilde{S}^2 - F/N) + 2\tilde{S}^2}{p \cdot (1-p)}.$$

Our criterion focuses primarily on T^2. Unlike the TS approach, however, it incorporates the benefits of splits due to improvement in the fit.

5 Inference

Given the estimated conditional average treatment effect we also would like to do inference. Once constructed, the tree is a function of covariates, and if we use a distinct sample 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 therefore valid for the estimates obtained via honest estimation, and in particular, no assumptions about model complexity are required. For the adaptive methods standard approaches to confidence intervals are not generally valid for the reasons discussed above, and below we document through simulations that this can be important in practice.

6 A Simulation Study

To assess the relative performance of the proposed algorithms we carried out a small simulation study with three distinct designs. In Table 1 we report a number of summary statistics from the simulations. We report averages; results for medians are similar. We report results for $N^{\text{tr}} = N^{\text{est}}$ with either 500 or 1000 observations. When comparing adaptive to honest approaches, we report the ratio of the MSE_τ for adaptive estimation with $N^{\text{tr}} = 1000$ to MSE_τ for honest estimation with $N^{\text{tr}} = N^{\text{est}} = 500$, in order to highlight the tradeoff between sample size and bias reduction that arises with honest estimation. We evaluate MSE_τ using a test sample with $N^{\text{te}} = 6000$ observations to test the methods in order to minimize the sampling variance in our simulation results.

In all designs, the marginal treatment probability is $p = 0.5$. K denotes the number of features. In each design, we have a model $\eta(x)$ for the mean effect and $\kappa^{\text{sim}}(x)$ for the treatment effect. Then, the potential outcomes are written

$$Y_i(w) = \eta^{\text{sim}}(X_i) + \frac{1}{2} \cdot (2w - 1) \cdot \kappa(X_i) + \epsilon_i,$$

where $\epsilon_i \sim \mathcal{N}(0, .01)$, and the X_i are independent of ϵ_i and one another, and $X_i \sim \mathcal{N}(0, 1)$. The designs are summarized as follows:

$$1\text{: } K = 2; \eta(x) = \frac{1}{2}x_1 + x_2; \kappa(x) = \frac{1}{2}x_1.$$

$$2\text{: } K = 10; \eta(x) = \frac{1}{2}\sum_{k=1}^{2} x_k + \sum_{k=3}^{6} x_k; \kappa(x) = \sum_{k=1}^{2} 1\{x_k > 0\} \cdot x_k$$

$$3\text{: } K = 20; \eta(x) = \frac{1}{2}\sum_{k=1}^{4} x_k + \sum_{k=5}^{8} x_k; \kappa(x) = \sum_{k=1}^{4} 1\{x_k > 0\} \cdot x_k$$

In each design, there are some covariates that affect treatment effects (κ) and mean outcomes (η); some covariates that enter η but not κ; and some covariates that do not affect outcomes at all ("noise" covariates). Design 1 does not have noise covariates. In Designs 2 and 3, the first few covariates enter κ, but only when their signs are positive, while they affect η throughout their range. Different criterion will thus lead to different optimal splits, even within a covariate; F will focus more on splits when the covariates are negative.

The first panel of Table 1 compares the number of leaves in different designs and different values of $N^{\text{tr}} = N^{\text{est}}$. Recalling that TOT-A and TOT-H have the same splitting method,

we see that it tends to build shallow trees. The failure to control for the realized value of W_i leads to additional noise in estimates, which tends to lead to aggressive pruning. For the other estimators, the adaptive versions lead to shallower trees than the honest versions, as the honest versions correct for overfitting, and the main cost of small leaf size is high variance in leaf estimates. F-A and F-H are very similar; as discussed above, the splitting criterion are very similar, and further, the F estimators are less prone to overfitting treatment effects, because they split based upon overall model fit. We also observe that the F estimators build the deepest trees; they reward splitting on covariates that affect mean outcomes as well as treatment effects.

The second panel of Table 1 examines the performance of the alternative honest estimators, as evaluated by the infeasible criterion MSE_τ. We report the average of the ratio of MSE_τ for a given estimator to MSE_τ for our preferred estimtor, CT-H. The TOT-H estimator performs well in Designs 2 and 3, but suffers in Design 1. In Design 1, the variance of Y_i conditional on (W_i, X_i) is very low at .01, and so the failure of TOT to account for the realization of W_i results in a noticeable loss of performance. The F-H estimator suffers in all 3 designs; all designs give the F-H criterion attractive opportunities to split based on covariates that do not enter κ. F-H would perform better in alternative designs where $\eta(x) = \kappa(x)$; F-H also does well at avoiding splits on noise covariates. The TS-H estimator performs well in Design 1, where x_1 affects η and κ the same way, so that the CT-H criterion is aligned with TS-H. Design 3 is more complex, and the ideal splits from the perspective of balancing overall mean-squared error of treatment effects (including variance reduction) are different from those favored by TS-H. Thus, TS performs worse, and the difference is exacerbated with larger sample size, where there are more opportunities for the estimators to build deeper trees and thus to make different choices. We also calculate comparisons based on a feasible criterion, the average squared difference between the transformed outcome Y_i^* and the estimated treatment effect $\hat{\tau}_i$. For details for this comparison see the SI Appendix. In general the results are consistent with those from the infeasible criterion.

The third panel of Table 1 explores the costs and benefits to honest estimation. The Table reports the ratio of $\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}, \pi^{\text{Estimator-A}}(\mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}))$ to $\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator-H}}(\mathcal{S}^{\text{tr}})$ for each estimator. The adaptive version uses the union of the training and estimation samples for tree-building, cross-validation, and leaf estimation. Thus it has double the sample size (1000 observations) at each step, while the honest version uses 500 of the observations in training and cross-validation, with the complement used for estimating treatment effects within leaves. The results show that there is a cost to honest estimation in terms of MSE_τ, varying by design and estimator.

The final two panels of Table 1 show the coverage rate for 90% confidence intervals. We achieve nominal coverage rates for honest methods in all designs, where, in contrast, the adaptive methods have coverage rates substantially below nominal rates. Thus, our simulations bear out the tradeoff that honest estimation sacrifices some goodness of fit (of treatment effects) in exchange for valid confidence intervals.

7 Observational Studies with Unconfoundedness

The discussion so far has focused on the setting where the assignment to treatment is randomized. The proposed methods can be adapted to observational studies under the assumption of unconfoundedness. In that case we need to modify the estimates within leaves to remove the bias from simple comparisons of treated and control units. There is a large literature on meth-

ods for doing so, e.g., [14]. For example, as in [10] we can do so by propensity score weighting. Efficiency will improve if we renormalize the weights within each leaf and and within the treatment and control group when estimating treatment effects. [5] propose approaches to trimming observations with extreme values for the propensity score to improve robustnesses. Note that there are some additional conditions required to establish asymptotic normality of treatment effect estimates when propensity score weighting is used (see, e.g., [10]); these results apply without modification to the estimation phase of honest partitioning algorithms.

8 The Literature

A small but growing literature seeks to apply supervised machine learning techniques to the problem of estimating heterogeneous treatment effects. Beyond those previously discussed, [23] transform the features rather than the outcomes and then apply LASSO to the model with the original outcome and the transformed features. [7] estimate $\mu(w,x) = \mathbb{E}[Y_i(w)|X_i = x]$ for $w = 0, 1$ 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. [13] use LASSO to estimate the effects of both treatments and attributes, but with different penalty terms for the two types of features 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. [25] consider a model with the outcome linear in the covariates and the interaction with the treatment variable. Using Bayesian nonparametric methods with Dirichlet priors, they project their estimates of heterogeneous treatment effects down onto the feature space using LASSO-type regularization methods to get low-dimensional summaries of the heterogeneity. [6] and [2] 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. Also related is the work on Targeted Learning [27], which modifies the loss function to increase the weight on the parts of the likelihood that concern the parameters of interest.

9 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 in randomized experiments and in observational studies satisfying unconfoundedness, without restrictions on the number of covariates or the complexity of the data generating process. 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.

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

[1] A. Abadie and G. Imbens, Large Sample Properties of Matching Estimators for Average Treatment Effects, Econometrica, 74(1), 235-267.

[2] A. Beygelzimer and J. Langford, The Offset Tree for Learning with Partial Labels, http://arxiv.org/pdf/0812.4044v2.pdf, (2009).

[3] L. Breiman, Random forests, Machine Learning, 45, (2001), 5-32.

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

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

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

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

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

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

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

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

[12] D. Horvitz, and D. Thompson, A generalization of sampling without replacement from a finite universe, Journal of the American Statistical Association, Vol. 47, (1952), 663685.

[13] K. Imai and M. Ratkovic, Estimating Treatment Effect Heterogeneity in Randomized Program Evaluation, Annals of Applied Statistics, 7(1), (2013), 443-470.

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

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

[16] P. Rosenbaum, Observational Studies, (2002), Springer.

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

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

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

[20] D. Rubin, Bayesian inference for causaleffects: The Role of Randomization, Annals of Statistics, 6, (1978), 34-58.

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

[22] J. Signovitch, J., Identifying informative biological markers in high-dimensional genomic data and clinical trials, PhD Thesis, Department of Biostatistics, Harvard University, (2007).

[23] L. Tian, A. Alizadeh, A. Gentles, and R. Tibshirani, A Simple Method for Estimating Interactions Between a Treatment and a Large Number of Covariates, Journal of the American Statistical Association, 109(508), (2014) 1517-1532.

[24] R. Tibshirani, Regression shrinkage and selection via the lasso, Journal of the Royal Statistical Society. Series B (Methodological), Volume 58, Issue 1. (1996), 267-288.

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

[26] V. Vapnik, Statistical Learning Theory, Wiley, (1998).

[27] M. Van Der Laan, and S. Rose, Targeted Learning: Causal Inference for Observational and Experimental Data, Springer, (2011).

[28] S. Wager, and S. Athey, Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, http://arxiv.org/pdf/1510.04342v2.pdf, (2015).

[29] H. Weisburg, H. and V. Pontes, Post hoc subgroups in Clinical Trials: Anathema or Analytics? Clinical Trials, June, 2015.

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

Table 1: Simulation Study

Design	1		2		3	
$N^{\text{tr}} = N^{\text{est}}$	500	1000	500	1000	500	1000
Estimator	Number of Leaves					
TOT	2.8	3.4	2.1	2.7	4.7	6.1
F-A	6.1	13.2	6.3	13.1	6.1	13.2
TS-A	4.0	5.6	2.5	3.3	4.4	8.9
CT-A	4.0	5.7	2.3	2.5	4.5	6.2
F-H	6.1	13.2	6.4	13.3	6.3	13.4
TS-H	4.4	7.7	5.3	11.0	6.0	12.3
CT-H	4.2	7.5	5.3	11.2	6.2	12.3
Infeasible MSE Divided by Infeasible MSE for CT-H*						
TOT-H	1.77	2.12	1.03	1.04	1.03	1.05
F-A	1.93	1.54	1.69	2.07	1.63	2.08
TS-H	1.01	1.02	1.06	0.99	1.24	1.38
CT-H	1.00	1.00	1.00	1.00	1.00	1.00
Ratio of Infeasible MSE: Honest to Adaptive**						
TOT-H/TOT-A		0.99		0.86		0.76
F-H/F-A		0.50		0.98		0.91
TS-H/TS-A		0.92		0.90		0.85
CT-H/CT-A		0.91		0.93		0.76
Coverage of 90% Confidence Intervals - Adaptive						
TOT-A	0.83	0.86	0.83	0.83	0.74	0.79
F-A	0.89	0.89	0.86	0.86	0.82	0.82
TS-A	0.85	0.85	0.80	0.83	0.77	0.80
CT-A	0.85	0.85	0.81	0.83	0.80	0.81
Coverage of 90% Confidence Intervals - Honest						
TOT-H	0.90	0.89	0.90	0.92	0.89	0.89
F-H	0.91	0.90	0.90	0.90	0.90	0.89
TS-H	0.89	0.90	0.90	0.90	0.90	0.90
CT-H	0.90	0.90	0.90	0.89	0.90	0.90

*$\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator}}(\mathcal{S}^{\text{tr}}))/\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{CT-H}}(\mathcal{S}^{\text{tr}}))$

**$\text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}, \pi^{\text{Estimator-A}}(\mathcal{S}^{\text{est}} \cup \mathcal{S}^{\text{tr}}))/\quad \text{MSE}_\tau(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator-H}}(\mathcal{S}^{\text{tr}})$

Additional Simulation Details (Online Appendix)

This Appendix describes some additional details of the simulation study, and also presents additional simulation results in Appendix Table A1.

The code for our simulations was written as an R software package that is in preparation for public release. It is based on the 'rpart' R package, available at https://cran.r-project.org/web/packages/rpart/index.html. For TOT, we directly use rpart applied to the transformed outcome Y_i^*, and we use 10-fold cross-validation for pruning the tree. For each of our other estimators, we modified several components of the package. In the remainder of this Appendix, discussions of modifications apply to F, CT, and TS estimators. For these estimators, we create new versions of the "anova" functions, functions that in the standard rpart package are used for calculating the total goodness of fit for a node of the tree, evaluating the quality of alternative splits, and estimating the goodness of fit for pruned trees using cross-validation samples. We maintain the overall structure of the rpart package. The rpart package has an important tuning parameter, which is the minimum number of observations per leaf, denoted n_m. We modify the rpart routine to insist on at least n_m treated *and* n_m control observations per leaf, to ensure that we can calculate a treatment effect within each leaf. In the simulations reported in Table 1 of the paper and in this Appendix, we use $n_m = 25$ for all models except TOT, while for the TOT model the minimum leaf size is 50 (without restrictions on treated and control observations).

We make one additional modification to the way the standard rpart splitting function works. We restrict the set of potential split points considered, and further, in the splitting process we rescale the covariate values within each leaf and each treatment group in order to ensure that when moving from one potential split point to the next, we move the same number of treatment and control observations from the right leaf to the left leaf. We begin by describing the motivation for these modifications, and then we give details.

The rpart algorithm considers every value of $X_{i,k}$ in $\mathcal{S}^{\text{tr}}$ as a possible split point for covariate X_k. An obvious disadvantage of this approach is that computation time can grow prohibitively large in models with many observations and covariates. But there are some more subtle disadvantages as well. The first is that there will naturally be sampling variation in estimates of $\hat{\tau}$ as we vary the possible split points. A problem akin to a multiple hypothesis testing problem arises: since we are looking for the maximum value of an estimated criterion across a large number of possible split points, as the number of split points tested grows, it becomes more and more likely that one of the splits for a given covariate appears to improve the fit criterion even if the true value of the criterion would indicate that it is better not to split. One way to mitigate both the computation time problem and the multiple-testing problem is to consider only a limited number of split points.

A third problem is specific to considering treatment effect heterogeneity. To see the problem, suppose that a covariate strongly affects the mean of outcomes, but not treatment effects. Within a leaf, some observations are treated and some are control. If we consider every level of the covariate in the leaf as a possible split point, then shifting from one split point to the next shifts a single observation from the right leaf to the left leaf. This observation is in the treatment or the control group, but not both; suppose it is in the treatment group. If the covariate has a strong effect on the level of outcomes, the observation that is shifted will be likely have an outcome more extreme than average. It will change the sample average of the treatment group, but not the control group, leading to a large change in the estimated treatment

effect difference. We expect the estimated difference in treatment effects across the left and right leaves to fluctuate greatly with the split point in this scenario. This variability around the true mean difference in treatment effects occurs more often when covariates affect mean outcomes, and thus it can lead the estimators to split too much on such covariates, and also to find spurious opportunities to split.

To address this problem, we propose the following modifications to the splitting rule. We include a parameter b, the target number of observations per "bucket." For each leaf, before testing possible splits for a particular covariate, we order the observations by the covariate value in the treatment and control group separately. Within each group, we place the observations into buckets with b observations per bucket. If this results in less than n_m buckets, then we use fewer observations per bucket (to attain n_m buckets). We number the buckets, and considering splitting by bucket number rather than the raw values of the covariates. This guarantees that when we shift from one split point to the next, we add both treatment and control observations, leading to a smoother estimate of the goodness of fit function as a function of the split point. After the best bucket number to split on is selected, we translate that into a split point by averaging the maximum covariate value in the corresponding treatment and control buckets. In the simulations presented in this paper, we do not constrain the maximum number of buckets, and we let $b = 4$. We found that this discretization approach improved goodness of fit on average for the simulations we considered, although it can in principle make things worse.

In the simulations reported in Table 1 of ths paper, we used the infeasible MSE_τ to evalute alternative estimators. In practice, we must estimate the infeasible criterion. In the paper, we propose estimators that rely on the tree structure of our estimator, but we may also wish to compare our performance to estimators that don't rely on partitions. One alternative is the MSE^{TOT} criterion. Given an estimator $\hat{\tau}_i$, it is equal to

$$MSE^{\text{TOT}} = \frac{1}{N^{\text{te}}} \sum_{i=1}^{N^{\text{te}}} (Y_i^* - \hat{\tau}_i)^2.$$

Because $\mathbb{E}[Y_i^*|X_i] = \tau(X_i)$, this is an unbiased (but noisy) estimator for MSE_τ. In Appendix Table A1, we present rankings of estimators using this criterion. We see that it ranks estimators in the same way as MSE_τ except in one case (Design 3 with 500 observations), but the percentage differences between estimators are smaller than with the infeasible criterion.

Another tuning parameter for standard CART as well as the methods proposed here is the number of cross-validation samples. A common convention is to use 10 samples. We deviate from that convention and use 5 cross-validation samples. The reason is that our methods require various quantities to be estimated within leaves. Given a minimum leaf restriction of 25 treated and control units, if we take a cross-validation sample of one-tenth of the original training sample, we might end up with no treated or no control observations in a leaf in a cross-validation sample. In addition, it may be difficult to estimate a sample variance within a leaf. Rather than require larger leaf sizes, we simply use fewer cross-validation samples.

Appendix Table A1 also includes the full set of estimates for the infeasible criterion MSE_τ, to illustrate how sample size and honest versus adaptive estimation affects the criterion. Note that for purposes of comparison to the simulation results from Table 1, Table 1 reports the average over simulations of the ratio of the goodness of fit measures; the second panel of this table shows the average of goodness of fit measures, but the ratio of the averages shown here is not exactly equal to the average of the ratios shown in Table 1.

Appendix Table A1: Infeasible and Feasible MSE Estimates for Simulation Study

Design	1		2		3	
$N^{\text{tr}} = N^{\text{est}}$	500	1000	500	1000	500	1000
Estimator	MSE_τ^{TOT} Divided by MSE_τ^{TOT} for CT-H*					
TOT-H	1.010	1.009	1.001	1.001	1.004	1.008
F-H	1.013	1.004	1.039	1.049	1.121	1.161
TS-H	0.999	1.000	1.003	0.999	1.046	1.056
CT-H	1.000	1.000	1.000	1.000	1.000	1.000
	Infeasible MSE_τ					
TOT-A	0.171	0.139	1.267	1.010	3.235	2.344
F-A	0.150	0.075	1.914	1.861	4.914	4.443
TS-A	0.103	0.077	1.509	1.080	4.068	3.169
CT-A	0.104	0.079	1.418	1.071	3.324	2.314
TOT-H	0.140	0.104	1.176	0.932	3.102	2.252
F-H	0.151	0.075	1.898	1.850	4.860	4.420
TS-H	0.083	0.053	1.201	0.887	3.732	2.935
CT-H	0.087	0.054	1.149	0.910	3.033	2.143

*$\text{MSE}_\tau^{\text{TOT}}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{Estimator}}(\mathcal{S}^{\text{tr}}))/\text{MSE}_\tau^{\text{TOT}}(\mathcal{S}^{\text{te}}, \mathcal{S}^{\text{est}}, \pi^{\text{CT-H}}(\mathcal{S}^{\text{tr}}))$