The Honest Truth About Causal Trees: Accuracy Limits for Heterogeneous Treatment Effect Estimation

Matias D. Cattaneo[*] Jason M. Klusowski[*] Ruiqi (Rae) Yu[*]

September 14, 2025

Abstract

Recursive decision trees have emerged as a leading methodology for heterogeneous causal treatment effect estimation and inference in experimental and observational settings. These procedures are fitted using the celebrated CART (Classification And Regression Tree) algorithm [Breiman et al., 1984], or custom variants thereof, and hence are believed to be "adaptive" to high-dimensional data, sparsity, or other specific features of the underlying data generating process. Athey and Imbens [2016] proposed several "honest" causal decision tree estimators, which have become the standard in both academia and industry. We study their estimators, and variants thereof, and establish lower bounds on their estimation error. We demonstrate that these popular heterogeneous treatment effect estimators cannot achieve a polynomial-in-n convergence rate under basic conditions, where n denotes the sample size. Contrary to common belief, honesty does not resolve these limitations and at best delivers negligible logarithmic improvements in sample size or dimension. As a result, these commonly used estimators can exhibit poor performance in practice, and even be inconsistent in some settings. Our theoretical insights are empirically validated through simulations.

Keywords: recursive partitioning, decision trees, causal inference, heterogeneous treatment effects

[*]Department of Operations Research and Financial Engineering, Princeton University.

1 Introduction

Athey and Imbens [2016] proposed to use recursive decision trees to estimate (and later conduct inference about) heterogeneous causal effects in experimental and observational settings. Their methodology is often called "honest" causal trees. Due in part to its simple, interpretable structure, their causal inference methodology has been widely adopted in academic and industry empirical research over the last decade. For example, to advocate for their proposal, the authors wrote that "[i]t enables researchers to let the data discover relevant subgroups while preserving the validity of confidence intervals constructed on treatment effects within subgroups" [Athey and Imbens, 2016, page 7353].

Despite the widespread use of honest causal tree estimators, little is known about their theoretical properties for estimation and inference. Existing results typically require very strong assumptions on the tree-growing process [Wager and Athey, 2018], which we show are incompatible with canonical implementations of causal trees under basic conditions. Specifically, this paper establishes lower bounds on the estimation error of heterogeneous treatment effect estimators based on recursive adaptive partitioning. We demonstrate that such estimators cannot achieve a polynomial-in-n convergence rate under basic conditions, where n denotes the sample size. Instead, these popular estimators can exhibit arbitrarily slow convergence rates, if not become inconsistent in some cases. As a consequence, our theoretical insights demonstrate that honest causal tree estimators, and variant thereof, may be inaccurate for estimating heterogeneous causal effects, and invalid for constructing confidence intervals on treatment effects within subgroups.

Our work in the causal setting also complements the rich existing theoretical analyses of recursive adaptive partitioning estimators for regression [Scornet et al., 2015, Chi et al., 2022, Klusowski and Tian, 2024, Cattaneo et al., 2024, Mazumder and Wang, 2024] and contributes to the small but growing body of negative results. For example, Ishwaran [2015] showed that regression trees via CART methodology [Breiman et al., 1984] can create imbalanced cells containing a small number of samples. Tan et al. [2022] proved that regression trees are inefficient at estimating additive structure, regardless of the way in which they are optimized. Tan et al. [2024b] proved that mixing times for Bayesian Additive Regression Trees (BART) [Chipman et al., 2010] can increase with the training sample size. Finally, Tan et al. [2024a] established that adaptive regression trees with Boolean covariates can require exponentially many samples in the dimension and are high-dimensional inconsistent for learning ANOVA decompositions with certain interaction patterns.

The present paper supersedes the unpublished manuscript by Cattaneo, Klusowski, and Tian [2022], which showed that a one-dimensional regression stump (i.e., single-split regression trees with a single covariate) constructed via CART can suffer arbitrarily slow convergence rates, and furthermore conjectured (but did not prove) that causal trees might (i) exhibit the same pathology and (ii) fail to benefit from honesty. Our paper proves both conjectures, and goes further by establishing these results for arbitrary covariate dimension and for any causal tree structure with at least one split (i.e., allowing for an arbitrary number of splits or depth of the causal tree).

The supplemental appendix also reports analogous results for plain adaptive regression trees. As sketched in Section 4.1, with full details given in the supplemental appendix, our method of proof relies on new insights concerning non-asymptotic approximations for the suprema of partial sums and various Gaussian processes, which may be of independent theoretical interest. In particular, we correct an error in Eicker [1979].

2 Setup

The available data $\mathscr{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \ldots, n\}$ is a random sample, where y_i is an outcome variable, $\mathbf{x}_i = (x_{1,1}, \ldots, x_{1,p})^\top$ is a vector of (pre-treatment) covariates, and d_i is a binary treatment indicator. Employing standard potential outcomes notation [see, e.g., Hernán and Robins, 2020, for an introduction], we assume that

$$y_i = y_i(1)d_i + y_i(0)(1 - d_i),$$

where $y_i(1)$ is the potential outcome under treatment assignment ($d_i = 1$), and $y_i(0)$ is the potential outcome under control assignment ($d_i = 0$). In classical experimental settings, the treatment assignment mechanism is independent of both the potential outcomes and the covariates, that is, $(y_i(0), y_i(1), \mathbf{x}_i^\top) \perp\!\!\!\perp d_i$.

The parameter of interest is the conditional average treatment effect (CATE) function

$$\tau(\mathbf{x}) \equiv \mathbb{E}\big[y_i(1) - y_i(0)\big|\mathbf{x}_i = \mathbf{x}\big],$$

which captures average treatment effects for different values of observable (pre-treatment) covariates. In experimental settings, the CATE function is identifiable because

$$\begin{aligned} \tau(\mathbf{x}) &= \mathbb{E}\big[y_i\big|d_i = 1,\ \mathbf{x}_i = \mathbf{x}\big] - \mathbb{E}\big[y_i\big|d_i = 0,\ \mathbf{x}_i = \mathbf{x}\big] && (1) \\ &= \mathbb{E}\left[y_i \frac{d_i - \xi}{\xi(1 - \xi)}\middle|\mathbf{x}_i = \mathbf{x}\right], && (2) \end{aligned}$$

where the probability of treatment assignment $\xi = \mathbb{P}(d_i = 1)$ is known by virtue of the known randomization mechanism. The first equality (1) represents $\tau(\mathbf{x})$ as the difference of two conditional expectation functions based on observed data, while the second equality (2) represents $\tau(\mathbf{x})$ as a single conditional expectation of the "transformed" outcome $y_i \frac{d_i - \xi}{\xi(1-\xi)}$.

Traditional semiparametric methods would replace the unknown conditional expectations by estimators thereof to learn about heterogeneous treatment effects from experimental data. These methods do not cope well with high-dimensional data, sparsity, or other unknown specific features of the data generating process. Motivated by the recent success of modern (adaptive) machine learning methods, Athey and Imbens [2016] proposed to estimate $\tau(\mathbf{x})$ using recursive decision trees. While retaining the core ideas underlying the greedy recursive construction via standard

CART, their proposals customized the tree splitting criterion to the causal inference setting, and employed sample splitting (the so-called "honesty" property) to de-couple the tree construction from the estimation of $\tau(\mathbf{x})$ on the terminal nodes of the tree. This honesty modification has been viewed as a natural "fix," since separating model selection from estimation is believed to reduce overfitting and improve the validity of inference. Despite this prevailing view, we show that honesty cannot overcome the fundamental limitations of recursive partitioning for heterogeneous causal effect estimation (or for plain adaptive regression trees), offering only at best negligible logarithmic improvements in sample size or dimension.

We perform a comprehensive study of the estimation accuracy of *nine* distinct causal tree methods, which differ on how their three key underlying parts are implemented: (i) *CATE estimator*, (ii) *tree construction*, and (iii) *sample splitting*.

2.1 CATE Estimator

Leveraging the identification results in (1)–(2), Athey and Imbens [2016] considered the following two CATE estimators based on a tree T and a dataset $\mathscr{D}_\tau$. Sections 2.2 and 2.3 discuss specific choices of T and $\mathscr{D}_\tau$, respectively. Let $\mathbb{1}(\cdot)$ be the indicator function.

Definition 1 (CATE Estimators)**.** *Suppose* T *is the tree used, and* $\mathscr{D}_\tau = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n_\tau\}$*, with* $n_\tau \leq n$*, is the dataset used. Let* t *be the unique terminal node in* T *containing* $\mathbf{x} \in \mathscr{X}$*.*

- *The Difference-in-Means (DIM) estimator is*

$$\hat{\tau}_{\mathtt{DIM}}(\mathbf{x}; \mathsf{T}, \mathscr{D}_\tau) = \frac{1}{n_1(\mathsf{t})} \sum_{i:\mathbf{x}_i \in \mathsf{t}} d_i y_i - \frac{1}{n_0(\mathsf{t})} \sum_{i:\mathbf{x}_i \in \mathsf{t}} (1 - d_i) y_i,$$

  where $n_d(\mathsf{t}) = \sum_{i=1}^{n_\tau} \mathbb{1}(\mathbf{x}_i \in \mathsf{t},\ d_i = d)$*, for* $d = 0, 1$*, are the "local" sample sizes. We set* $\hat{\tau}_{\mathtt{DIM}}(\mathbf{x}; \mathsf{T}, \mathscr{D}_\tau) = 0$ *whenever* $n_0(\mathsf{t}) = 0$ *or* $n_1(\mathsf{t}) = 0$*.*

- *The Inverse Probability Weighting (IPW) estimator is*

$$\hat{\tau}_{\mathtt{IPW}}(\mathbf{x}; \mathsf{T}, \mathscr{D}_\tau) = \frac{1}{n(\mathsf{t})} \sum_{i:\mathbf{x}_i \in \mathsf{t}} \frac{d_i - \xi}{\xi(1 - \xi)} y_i,$$

  where $n(\mathsf{t}) = n_0(\mathsf{t}) + n_1(\mathsf{t}) = \sum_{i=1}^{n_\tau} \mathbb{1}(\mathbf{x}_i \in \mathsf{t})$ *is the "local" sample size. We set* $\hat{\tau}_{\mathtt{IPW}}(\mathbf{x}; \mathsf{T}, \mathscr{D}_\tau) = 0$ *whenever* $n(\mathsf{t}) = 0$*.*

Both estimators, $\hat{\tau}_{\mathtt{DIM}}(\mathbf{x}; \mathsf{T}, \mathscr{D}_\tau)$ and $\hat{\tau}_{\mathtt{IPW}}(\mathbf{x}; \mathsf{T}, \mathscr{D}_\tau)$, rely on localization near $\mathbf{x}$ via the tree construction: T forms a partition of the support of the covariates $\mathscr{X}$, and estimation of $\tau(\mathbf{x})$ uses only observations with covariates $\mathbf{x}_i$ belonging to the cell in the partition covering $\mathbf{x} \in \mathscr{X}$. Therefore, given a tree (or partition), both estimators can be represented as nonparametric partitioning-based estimates of $\tau(\mathbf{x})$. See Györfi et al. [2002], Cattaneo et al. [2020], Cattaneo et al. [2025], and references therein.

Since the estimators $\hat{\tau}_{\texttt{DIM}}(\mathbf{x};\mathsf{T},\mathscr{D}_{\tau})$ and $\hat{\tau}_{\texttt{IPW}}(\mathbf{x};\mathsf{T},\mathscr{D}_{\tau})$ output a constant fit for all $\mathbf{x}$ within each terminal node of T (or cell in the partition), we define

$$\hat{\tau}_l(\mathsf{t};\mathsf{T},\mathscr{D}_{\tau}) = \hat{\tau}_l(\mathbf{x};\mathsf{T},\mathscr{D}_{\tau}), \qquad l \in \{\texttt{DIM},\texttt{IPW}\}, \qquad \mathbf{x} \in \mathsf{t},$$

for all terminal nodes t of T.

2.2 Tree Construction

An axis-aligned recursive decision tree is a predictive model that makes decisions by repeatedly splitting the data into subsets based on both outcome and covariate values. At each node, the algorithm selects the feature and threshold that best separate the data according to some criterion (e.g., squared error, Gini impurity, or entropy), and this process continues recursively until a stopping condition is met (e.g., maximum depth or pure terminal nodes). See Berk [2020], Zhang and Singer [2010], and references therein.

The most popular implementation of recursive decision trees is via the CART algorithm, which proceeds in a top-down, greedy manner through recursive binary splitting. Given a dataset $\mathscr{D}_{\mathsf{T}} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n_{\mathsf{T}}\}$, with $n_{\mathsf{T}} \leq n$, a parent node t in the tree (i.e., a region in $\mathscr{X}$) is divided into two child nodes, t_{L} and t_{R}, by minimizing the sum-of-squares error (SSE),

$$\min_{1\leq j\leq p} \min_{\beta_{\mathsf{L}},\beta_{\mathsf{R}},\varsigma\in\mathbb{R}} \sum_{\mathbf{x}_i\in\mathsf{t}} \big(y_i - \beta_{\mathsf{L}}\mathbb{1}(x_{ij}\leq\varsigma) - \beta_{\mathsf{R}}\mathbb{1}(x_{ij}>\varsigma)\big)^2, \tag{3}$$

where the solution yields estimates $(\hat{\beta}_{\mathsf{L}}, \hat{\beta}_{\mathsf{R}}, \hat{\varsigma}, \hat{\jmath})$, being the two child nodes average output, split point and split direction, respectively. Because the splits occur along values of a single covariate, the induced partition of the input space $\mathscr{X}$ is a collection of hyper-rectangles, and hence the resulting refinement of t produces child nodes $\mathsf{t}_{\mathsf{L}} = \{\mathbf{x}\in\mathsf{t} : \mathbf{e}_{\hat{\jmath}}^\top\mathbf{x} \leq \hat{\varsigma}\}$ and $\mathsf{t}_{\mathsf{R}} = \{\mathbf{x}\in\mathsf{t} : \mathbf{e}_{\hat{\jmath}}^\top\mathbf{x} > \hat{\varsigma}\}$. More precisely, the normal equations imply that $\hat{\beta}_{\mathsf{L}} = \frac{1}{n(\mathsf{t}_{\mathsf{L}})}\sum_{\mathbf{x}_i\in\mathsf{t}_{\mathsf{L}}} y_i$ and $\hat{\beta}_{\mathsf{R}} = \frac{1}{n(\mathsf{t}_{\mathsf{R}})}\sum_{\mathbf{x}_i\in\mathsf{t}_{\mathsf{R}}} y_i$, the respective sample means after splitting the parent node at $\mathbf{e}_{\hat{\jmath}}^\top\mathbf{x} = \hat{\varsigma}$. These child nodes become new parent nodes at the next level of the tree construction, and can be further refined in the same manner, and so on and so forth, until a desired depth K is reached. While not every parent node needs to generate a new child node in a recursive tree construction, a maximal decision tree of depth K is a particular instance where the construction is iterated K times until (i) the node contains a single data point $(y_i, \mathbf{x}_i^\top)$ or (ii) all input values $\mathbf{x}_i$ and/or all response values y_i within the node are the same.

Building on the CART algorithm, Athey and Imbens [2016] proposed the following two custom criteria for constructing a tree T to implement their causal tree estimators.

Definition 2 (Tree Construction). *Suppose $\mathscr{D}_{\mathsf{T}} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n_{\mathsf{T}}\}$, with $n_{\mathsf{T}} \leq n$, is the dataset used to construct the tree T. There is a unique node $\mathsf{t}_0 = \mathscr{X}$ at initialization, and child nodes are generated by iterative axis-aligned splitting of the parent node based on either of the following two rules.*

- *Variance Maximization: A parent node* t *(i.e., a terminal node partitioning* $\mathscr{X}$*) in a previous tree* T' *is divided into two child nodes,* $\mathsf{t}_{\mathtt{L}}$ *and* $\mathsf{t}_{\mathtt{R}}$*, forming the new tree* T*, by maximizing*

$$\frac{n(\mathsf{t}_{\mathtt{L}})n(\mathsf{t}_{\mathtt{R}})}{n(\mathsf{t})}\Big(\hat{\tau}_l(\mathsf{t}_{\mathtt{L}};\mathsf{T},\mathscr{D}_{\mathsf{T}}) - \hat{\tau}_l(\mathsf{t}_{\mathtt{R}};\mathsf{T},\mathscr{D}_{\mathsf{T}})\Big)^2, \qquad l \in \{\mathtt{DIM}, \mathtt{IPW}\}. \tag{4}$$

  Assuming at least one split, the two final causal trees are denoted by $\mathsf{T}^{\mathtt{DIM}}(\mathscr{D}_{\mathsf{T}})$ *and* $\mathsf{T}^{\mathtt{IPW}}(\mathscr{D}_{\mathsf{T}})$*, respectively.*

- *SSE Minimization: A parent node* t *(i.e., a terminal node partitioning* $\mathscr{X}$*) in the previous tree* T' *is divided into two child nodes,* $\mathsf{t}_{\mathtt{L}}$ *and* $\mathsf{t}_{\mathtt{R}}$*, forming the next tree* T*, by solving*

$$\min_{a_{\mathtt{L}},b_{\mathtt{L}},a_{\mathtt{R}},b_{\mathtt{R}}\in\mathbb{R}} \sum_{\mathbf{x}_i\in\mathsf{t}_{\mathtt{L}}} (y_i - a_{\mathtt{L}} - b_{\mathtt{L}}d_i)^2 + \sum_{\mathbf{x}_i\in\mathsf{t}_{\mathtt{R}}} (y_i - a_{\mathtt{R}} - b_{\mathtt{R}}d_i)^2, \tag{5}$$

  where only the data $\mathscr{D}_{\mathsf{T}}$ *is used. Assuming at least one split, the final causal tree is denoted by* $\mathsf{T}^{\mathtt{SSE}}(\mathscr{D}_{\mathsf{T}})$*.*

The variance maximization splitting criterion is somewhat different than the original CART criteria (3), since it explicitly selects splits based on maximizing the squared difference of the child treatment effect estimates. For the IPW estimator, this rule is equivalent to applying the CART criterion in (3) to the transformed outcome $\tilde{y}_i = y_i\dfrac{d_i-\xi}{\xi(1-\xi)}$. This transformation satisfies $\mathbb{E}[\tilde{y}_i \mid \mathbf{x}_i = \mathbf{x}] = \tau(\mathbf{x})$ for all $\mathbf{x} \in \mathscr{X}$, and thus CART operates on an outcome whose conditional mean equals the CATE. The DIM estimator follows the same idea of predicting the within-node average treatment effect, but it constructs these predictions somewhat differently.

The SSE Minimization criterion resembles the original CART criteria (3), but its formulation still targets treatment effect heterogeneity as the splitting criteria: in Section SA-3.3 of the supplemental appendix we show that the objective function (5) can be recast as maximization of the sum of variances of treatment and control group outcomes given by

$$\begin{aligned}&\frac{n_1(\mathsf{t}_{\mathtt{L}})n_1(\mathsf{t}_{\mathtt{R}})}{n_1(\mathsf{t})}\Big(\frac{1}{n_1(\mathsf{t}_{\mathtt{L}})}\sum_{i:\mathbf{x}_i\in\mathsf{t}_{\mathtt{L}}} d_iy_i - \frac{1}{n_1(\mathsf{t}_{\mathtt{R}})}\sum_{i:\mathbf{x}_i\in\mathsf{t}_{\mathtt{R}}} d_iy_i\Big)^2\\ &\qquad + \frac{n_0(\mathsf{t}_{\mathtt{L}})n_0(\mathsf{t}_{\mathtt{R}})}{n_0(\mathsf{t})}\Big(\frac{1}{n_0(\mathsf{t}_{\mathtt{L}})}\sum_{i:\mathbf{x}_i\in\mathsf{t}_{\mathtt{L}}} (1-d_i)y_i - \frac{1}{n_0(\mathsf{t}_{\mathtt{R}})}\sum_{i:\mathbf{x}_i\in\mathsf{t}_{\mathtt{R}}} (1-d_i)y_i\Big)^2.\end{aligned}$$

Each of the causal recursive tree constructions leads to a distinct data-driven partition of $\mathscr{X}$. A key observation in this paper is that they do not generate quasi-uniform partitions, and thus known results in the nonparametric partitioning-based estimation literature [Györfi et al., 2002, Cattaneo et al., 2020, 2025] are not applicable. The supplemental appendix considers other recursive partitioning constructions, including the standard CART algorithm and variants thereof.

2.3 Sample Splitting

The final ingredient of the causal tree estimators concerns the data used at each stage of their construction. It is believed that de-coupling the CATE estimation (Definition 1) from the tree implementation (Definition 2) can lead to better performance of the final estimator. In practice, this approach corresponds to sample splitting, and Athey and Imbens [2016] and others referred to it as "honesty." To avoid confusion, we emphasize that procedures without sample splitting are not "dishonest" in any formal sense; they are simply harder to analyze formally.

To elucidate the relative merits of sample splitting, we consider two distinct scenarios: (i) no sample splitting, where the same data is used throughout (as the original CART procedure is often implemented); and (ii) honesty, where two independent datasets are used, one for tree construction and the other for CATE estimation (these are the procedures proposed by Athey and Imbens [2016] and many others). Formally, we consider the following data usages and resulting treatment effect estimators.

Definition 3 (Sample Splitting and Estimators)**.** *Recall Definition 1 and Definition 2, and that* $\mathscr{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \dots, n\}$ *is the available random sample.*

- *No Sample Splitting (NSS): The dataset* $\mathscr{D}$ *is used for both the tree construction and the treatment effect estimation, that is,* $\mathscr{D}_{\mathsf{T}} = \mathscr{D}$ *and* $\mathscr{D}_\tau = \mathscr{D}$. *The causal tree estimators are*

$$\begin{aligned} \hat{\tau}^{\texttt{NSS}}_{\texttt{DIM}}(\mathbf{x}) &= \hat{\tau}_{\texttt{DIM}}(\mathbf{x}; \mathsf{T}^{\texttt{DIM}}(\mathscr{D}), \mathscr{D}), \\ \hat{\tau}^{\texttt{NSS}}_{\texttt{IPW}}(\mathbf{x}) &= \hat{\tau}_{\texttt{IPW}}(\mathbf{x}; \mathsf{T}^{\texttt{IPW}}(\mathscr{D}), \mathscr{D}), \quad \text{and} \\ \hat{\tau}^{\texttt{NSS}}_{\texttt{SSE}}(\mathbf{x}) &= \hat{\tau}_{\texttt{DIM}}(\mathbf{x}; \mathsf{T}^{\texttt{SSE}}(\mathscr{D}), \mathscr{D}). \end{aligned}$$

- *Honesty (HON): The dataset* $\mathscr{D}$ *is divided in two independent datasets* $\mathscr{D}_{\mathsf{T}}$ *and* $\mathscr{D}_\tau$ *with sample sizes* n_{T} *and* n_τ, *respectively, and satisfying* $n \lesssim n_{\mathsf{T}}, n_\tau \lesssim n$. *The causal tree estimators are*

$$\begin{aligned} \hat{\tau}^{\texttt{HON}}_{\texttt{DIM}}(\mathbf{x}) &= \hat{\tau}_{\texttt{DIM}}(\mathbf{x}; \mathsf{T}^{\texttt{DIM}}(\mathscr{D}_{\mathsf{T}}), \mathscr{D}_\tau), \\ \hat{\tau}^{\texttt{HON}}_{\texttt{IPW}}(\mathbf{x}) &= \hat{\tau}_{\texttt{IPW}}(\mathbf{x}; \mathsf{T}^{\texttt{IPW}}(\mathscr{D}_{\mathsf{T}}), \mathscr{D}_\tau), \quad \text{and} \\ \hat{\tau}^{\texttt{HON}}_{\texttt{SSE}}(\mathbf{x}) &= \hat{\tau}_{\texttt{DIM}}(\mathbf{x}; \mathsf{T}^{\texttt{SSE}}(\mathscr{D}_{\mathsf{T}}), \mathscr{D}_\tau). \end{aligned}$$

The no-sample-splitting and honesty data usages are commonly encountered in the literature, and thus our results will speak directly to theoretical, methodological and empirical work relying on these sample splitting designs. While the estimators $\hat{\tau}^{\texttt{NSS}}_l(\mathbf{x})$ and $\hat{\tau}^{\texttt{HON}}_l(\mathbf{x})$, $l \in \{\texttt{DIM}, \texttt{IPW}, \texttt{SSE}\}$, depend on the depth of the tree construction used, our notation does not make this dependence explicit because our results apply whenever at least one split takes place. See Section 5 for more discussion, and a setting where the number of splits is assumed to increase with the sample size.

3 Assumptions

We impose the following assumption throughout the paper.

Assumption 1 (Data Generating Process). *$\mathscr{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : 1 \leq i \leq n\}$ is a random sample, where $y_i = d_i y_i(1) + (1 - d_i) y_i(0)$, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^\top$, and the following conditions hold for all $d = 0, 1$ and $i = 1, 2, \ldots, n$.*

(i) *$(y_i(0), y_i(1), \mathbf{x}_i) \perp\!\!\!\perp d_i$, and $\xi = \mathbb{P}(d_i = 1) \in (0, 1)$.*

(ii) *$y_i(d) = \mu_d(\mathbf{x}_i) + \varepsilon_i(d)$, with $\mathbb{E}[\varepsilon_i(d)|\mathbf{x}_i] = 0$ and $\mathbf{x}_i \perp\!\!\!\perp \varepsilon_i(d)$.*

(iii) *$\mu_d(\mathbf{x}) = c_d$ for all $\mathbf{x} \in \mathscr{X}$, where c_d is some constant and $\mathscr{X}$ is the support of $\mathbf{x}_i$.*

(iv) *$x_{i,1}, \ldots, x_{i,p}$ are independent and continuously distributed.*

(v) *There exists $\alpha > 0$ such that $\mathbb{E}[\exp(\lambda \varepsilon_i(d))] < \infty$ for all $|\lambda| < 1/\alpha$ and $\mathbb{E}[\varepsilon_i^2(d)] > 0$.*

Assumption 1(i) corresponds to simple randomized experiments. Assumption 1(ii) further assumes a canonical homoskedastic causal regression model, while Assumption 1(iii) implies that there is no heterogeneity in the causal treatment effect $\tau = c_1 - c_0$. Because trees are invariant with respect to monotone transformations of the coordinates of $\mathbf{x}_i$, without loss of generality, Assumption 1(iv) can be replaced by the assumption that covariates are uniformly distributed on $\mathscr{X} = [0, 1]^p$, i.e., $x_{i,j} \overset{\text{i.i.d.}}{\sim} \text{Uniform}([0, 1])$ for $j = 1, 2, \ldots, p$. Finally, Assumption 1(v) means that potential outcome errors are sub-exponential, or equivalently, they satisfy a Bernstein moment condition.

Since we are interested in establishing lower bounds on the estimation accuracy of the causal tree estimators in Definition 3, it is sufficient to consider the constant treatment effect model in Assumption 1 for several reasons. First, this statistical model is a canonical member of any interesting class of data generating processes because the constant function belongs to all classical smoothness function classes, as well as to the set of functions with bounded total variation. It follows that our results will shed light in settings where uniformity over any of the aforementioned classes of functions is of interest: our lower bounds can be applied directly in those cases because for any estimator $\hat{\tau}(\mathbf{x})$ of the parameter $\tau(\mathbf{x})$,

$$\sup_{\mathbb{P} \in \mathscr{P}} \mathbb{P}\Big(\sup_{\mathbf{x} \in \mathscr{X}} |\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| > \epsilon \Big) \geq \mathbb{P}_1\Big(\sup_{\mathbf{x} \in \mathscr{X}} |\hat{\tau}(\mathbf{x}) - \tau(\mathbf{x})| > \epsilon \Big),$$

for all $\epsilon > 0$, and for any data generating class $\mathscr{P}$ that includes the distribution $\mathbb{P}_1$ satisfying Assumption 1. In fact, the constant treatment effect model is a canonical case to consider in causal inference.

Second, Assumption 1 also removes issues related to smoothing (or misspecification) bias, heteroskedasticity, and heavy tail distributions. In particular, since the CATE function $\tau(\mathbf{x})$ is constant for all $\mathbf{x} \in \mathscr{X}$, our results will not be driven by standard (boundary or other smoothing) bias in nonparametrics. For example, if the distributions of $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are symmetric about zero,

$$\mathbb{E}[\hat{\tau}_l^{q}(\mathbf{x})] = \tau, \qquad q \in \{\mathtt{NSS}\}, \qquad \text{and} \qquad \mathbb{E}[\hat{\tau}_l^{\mathtt{HON}}(\mathbf{x})] = \tau - \tau\mathbb{P}(n(\mathtt{t}) = 0),$$

for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$ and $\mathbf{x} \in \mathtt{t}$ where $\mathtt{t}$ is a terminal node in the tree. Unbiasedness of $\hat{\tau}_l^{\mathtt{NSS}}(\mathbf{x})$ follows from the fact that the split points are symmetric functions of the residuals. In the case of $\hat{\tau}_l^{\mathtt{HON}}(\mathbf{x})$, sample splitting can generate empty cells with positive probability, which is captured by the term $\tau\mathbb{P}(n(\mathtt{t}) = 0)$; see Lemma SA-37 in the supplemental appendix. It follows that, in particular, $\hat{\tau}_l^{\mathtt{HON}}(\mathbf{x})$ is unbiased when $\tau = 0$ (or for any other known treatment effect value), as well as in tree constructions ensuring that $\mathbb{P}(n(\mathtt{t}) = 0) = 0$; otherwise, $\hat{\tau}_l^{\mathtt{HON}}(\mathbf{x})$ is asymptotically unbiased whenever $\mathbb{P}(n(\mathtt{t}) = 0) \to 0$ as $n \to \infty$. Our results will be driven by the fact that canonical adaptive decision tree constructions can generate small cells containing only a handful of observations, thereby making the estimator highly inaccurate in some regions of $\mathscr{X}$, regardless of bias. In other words, inconsistency is due to a large variance problem, not a large bias problem.

Third, the local constant treatment effect model could also be interpreted as a first-order approximation of the smooth function $\tau(\mathbf{x})$. Because the recursive partitioning schemes lead to a partitioning-based estimator of the CATE function, it follows that $\tau(\mathbf{x})$ is approximated locally by a Haar basis (piecewise constant functions). In fact, our results can be extended to hold uniformly over appropriate shrinking neighborhoods of smooth functions local to the constant function, provided that the signal to noise ratio (bias-variance trade-off) is small.

4 Main Results

The following theorem summarizes our first main result. Let e denote Euler's constant.

Theorem 1 (Uniform Accuracy)**.** *Suppose Assumption 1 holds, and the underlying causal tree has at least one split (i.e., at least two terminal nodes). Then, for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$ and all $b \in (0, 1)$,*

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{\mathbf{x}\in\mathscr{X}} \big|\hat{\tau}_l^{\mathtt{NSS}}(\mathbf{x}) - \tau(\mathbf{x})\big| \geq C_1 n^{-b/2}\sqrt{\log\log n}\Big) \geq b/e,$$

where the positive constant C_1 only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, and

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{\mathbf{x}\in\mathscr{X}} \big|\hat{\tau}_l^{\mathtt{HON}}(\mathbf{x}) - \tau(\mathbf{x})\big| \geq C_2 n^{-b/2}\Big) \geq C_3 b,$$

where the positive constants C_2 and C_3 only depend on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, and the sample splitting scheme via $\liminf_{n\to\infty} \frac{n_\mathtt{T}}{n_\tau}$ and $\limsup_{n\to\infty} \frac{n_\mathtt{T}}{n_\tau}$. The precise definitions of the constants are given in the supplemental appendix.

Section 4.1 gives an overview of the proof strategy of Theorem 1, with all omitted technical details given in the supplemental appendix (see Section SA-1.2 for details). Our proof relies on several non-asymptotic approximation steps for the suprema of partial sums and various Gaussian processes leveraging key technical results from Chernozhukov et al. [2017], Chernozhuokov et al. [2022], Csörgö and Révész [1981], Csörgö and Horváth [1997], Eicker [1979], El-Yaniv and Pechyony [2009], Göing-Jaeschke and Yor [2003], Horváth [1993], Latała and Matlak [2017], Petrov [2007], Shorack and Smythe [1976], and Skorski [2023]. As a technical by-product, we correct a mistake in Eicker [1979]: see Remark SA-1 in the supplemental appendix.

Theorem 1 presents precise lower bounds on the uniform convergence rate of the six causal tree estimators introduced in Section 2. Starting with procedures that do not employ sample splitting, Theorem 1 demonstrates that the three estimators $\hat{\tau}^{\mathtt{NSS}}_{\mathtt{DIM}}(\mathbf{x})$, $\hat{\tau}^{\mathtt{NSS}}_{\mathtt{IPW}}(\mathbf{x})$ and $\hat{\tau}^{\mathtt{NSS}}_{\mathtt{SSE}}(\mathbf{x})$ cannot achieve a uniform convergence rate of $n^{-b/2}\sqrt{\log\log n}$, for any $b > 0$. That is, they must have a worse than polynomial-in-n uniform convergence rate, and thus suffer from low accuracy in estimating heterogeneous treatment effects in certain regions of the support $\mathscr{X}$.

Athey and Imbens [2016], and many others, argue that sample splitting (the so-called "honesty" property) can improve the performance of machine learning estimators, and in particular their proposed causal tree estimators, because such sample usage de-couples the causal tree construction and the CATE estimation steps. The second result in Theorem 1 considers exactly their honest causal tree estimators, $\hat{\tau}^{\mathtt{HON}}_{\mathtt{DIM}}(\mathbf{x})$, $\hat{\tau}^{\mathtt{HON}}_{\mathtt{IPW}}(\mathbf{x})$ and $\hat{\tau}^{\mathtt{HON}}_{\mathtt{SSE}}(\mathbf{x})$. It follows from the theorem that these estimators cannot achieve a uniform convergence rate that is polynomial-in-n either. Notably, our results show that sample splitting (or honesty) improves the best achievable uniform convergence rate of the estimators, but this improvement is quite modest: the penalty term $\sqrt{\log\log n}$ is removed, thereby improving the uniform convergence rate by a very slow factor.

The results in Theorem 1 offer a pessimistic outlook on the utility of adaptive decision tree methods in causal inference when the goal is to learn about heterogeneous treatment effects: the estimators cannot perform well pointwise (and hence uniformly) over the entire support of the covariates; see Section 4.1 for more formal details. As a point of contrast, the same procedures considered in Theorem 1 can achieve near-optimal convergence rates "on average" over $\mathscr{X}$, as the following theorem establishes. Here again, honesty delivers only negligible improvements of order $\log(p)$.

Theorem 2 (Mean Square Accuracy)**.** *Suppose Assumption 1 holds and the underlying causal tree has depth at most $K \geq 1$, and let $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{x}_i \leq \mathbf{x})$. Then, for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$,*

$$\mathbb{E}\Big[\int_{\mathscr{X}} \big|\hat{\tau}^{\mathtt{NSS}}_{l}(\mathbf{x}) - \tau(\mathbf{x})\big|^2 dF_{\mathbf{X}}(\mathbf{x})\Big] \leq C_1 \frac{2^K \log^4(n)\log(np)}{n},$$

where the constant C_1 only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$, and

$$\mathbb{E}\Big[\int_{\mathscr{X}} \big|\hat{\tau}^{\mathtt{HON}}_{l}(\mathbf{x}) - \tau(\mathbf{x})\big|^2 dF_{\mathbf{X}}(\mathbf{x})\Big] \leq C_2 \frac{2^K \log^5(n)}{n},$$

provided that $\rho \leq n_{\mathsf{T}}/n_{\tau} \leq 1 - \rho$ for some $\rho \in (0,1)$, and the constant C_2 only depends on ρ and the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

The proof of this theorem is given in the supplemental appendix (see Section SA-1.2 for details). It leverages ideas and technical results from Györfi et al. [2002] and Klusowski and Tian [2024]. Crucially, the result applies only when Assumption 1 holds, that is, when $\tau(\mathbf{x})$ is constant. The main purpose of Theorem 2 is to demonstrate that in the same basic setting when uniform convergence fails, causal decision trees nonetheless achieve favorable performance on average in an integrated mean-squared sense. A natural way to interpret the juxtaposition between Theorem 1 and Theorem 2 is related to the often claimed tension between causal inference and prediction in machine learning settings: adaptive causal trees can perform poorly pointwise (hence uniformly), but excellently on average, over the feature space.

From a technical perspective, the results in Theorem 2 are new in the context of causal tree estimation and, notably, for the formal comparison between no-sample-splitting and honest implementations. Furthermore, our theoretical work in the supplemental appendix establishes the integrated mean-squared error bounds with high-probability, enabling a sharper comparison with Theorem 1. For example, for the case of no-sample-splitting, we show that

$$\limsup_{n\to\infty} \mathbb{P}\Big(\int_{\mathcal{X}} \big|\hat{\tau}_l^{\texttt{NSS}}(\mathbf{x}) - \tau(\mathbf{x})\big|^2 dF_{\mathbf{X}}(\mathbf{x}) \geq C_1 \frac{2^K \log^4(n)\log(np)}{n}\Big) = 0,$$

where C_1 is the constant in Theorem 2.

4.1 Proof Strategy of Theorem 1

Underlying our theoretical insights are a collection of technical results concerning a decision stump, and hence a decision tree of depth one. For each tree splitting criteria and sample splitting design, we first study the probabilistic properties of the split location at the root node, and thus characterize the regions of the support $\mathcal{X}$ where the first split index is most likely to realize. These theoretical results also characterize the effective sample size of the resulting child nodes. We establish that with non-vanishing probability, the first split will concentrate near a region of the boundary of the parent node (a cell in the partition of $\mathcal{X}$), from the beginning of any tree construction. More precisely, let $\hat{\imath} = n(\mathtt{t}_L)$ and $\hat{\jmath}$ be the CART split index and split variable at the root node, respectively, with $l \in \{\texttt{DIM}, \texttt{IPW}, \texttt{SSE}\}$, noticing that the first split coincide for no-sample-splitting and honest constructions. For each $a, b \in (0,1)$ with $a < b$ and $j \in \{1, 2, \ldots, p\}$, and $l \in \{\texttt{DIM}, \texttt{IPW}, \texttt{SSE}\}$, we have

$$\liminf_{n\to\infty} \mathbb{P}\big(n^a \leq \hat{\imath} \leq n^b,\ \ \hat{\jmath} = j\big) = \liminf_{n\to\infty} \mathbb{P}\big(n - n^b \leq \hat{\imath} \leq n - n^a,\ \ \hat{\jmath} = j\big) \geq \frac{b-a}{2pe}. \tag{6}$$

The slow uniform convergence rate of a decision stump estimator occurs because the optimal split point concentrates near the boundary of the support, causing the two nodes in the stump to be imbalanced, with one containing a much smaller number of samples, and therefore rendering a situation where local averaging is less accurate. This can be deduced from (6): for each coordinate $j = 1, 2, \ldots, p$ and $b \in (0, 1)$, there is non-vanishing $b/(pe)$ probability that the child cells $\{\mathbf{x} \in \mathcal{X} :$

$x_j \leq \hat{\varsigma}\}$ or $\{\mathbf{x} \in \mathscr{X} : x_j > \hat{\varsigma}\}$ are highly anisotropic and will contain at most n^b samples. Thus, with non-vanishing probability, the causal tree procedures will exhibit arbitrarily slow convergence rate in a region of $\mathscr{X}$. These results are then carefully recycled to characterize the properties of the deeper trees: due to their recursive nature, and since $p > 1$, the problematic regions take the form of many hyper-rectangles, and will realize anywhere in $\mathscr{X}$, with non-vanishing probability.

The core of proof strategy is to study the tree construction as the maximizer of the split criterion from (4) and (5), as indexed by the optimal split location and covariate coordinate. We leverage non-asymptotic high-dimensional central limit theorems, Gaussian comparison inequalities, Gaussian process embeddings, the Darling-Erdös theorem, and empirical process techniques [El-Yaniv and Pechyony, 2009, Petrov, 2007, Shorack and Smythe, 1976, Skorski, 2023], as explained in the following four main steps.

Step 1: Split Criterion Approximation. Using empirical process theory techniques, we establish an asymptotic equivalence between the split criterion underlying each of the causal tree estimators and the split criterion of a standard (non-causal) decision regression tree employing CART. For $l = \texttt{DIM}$ and $l = \texttt{IPW}$, the latter can be viewed as a standard regression tree with transformed outcomes $y_i \frac{d_i - \xi}{\xi(1-\xi)}$. For $l = \texttt{SSE}$, approximating process is the sum of two independent split criterion processes, one with transformed outcome $\frac{d_i}{\xi} y_i$ for treated units, and the other with transformed outcome $\frac{1-d_i}{1-\xi} y_i$ for control units. We employ a careful truncation argument to remove extremely small or large split indices [Csörgö and Horváth, 1997, Theorem A.4.1], where empirical process techniques are hard to apply.

Step 2: Conditional Gaussian Approximation. We show that, conditional on the covariates ordering, the square root of the split criterion processes from step 1 can be approximated by Gaussian processes with the same conditional covariance structure. For $l = \texttt{DIM}$ and $l = \texttt{IPW}$, we view the split criterion process as a summation of i.i.d. high-dimensional random vectors, each entry corresponding to one pair of split index and coordinate. The high-dimensional central limit theorem of [Chernozhukov et al., 2017, Theorem 2.1] implies that the split criterion process in high-dimensional vector form is close to a high-dimensional Gaussian random vector with the same covariance matrix conditional the ordering, the latter can then be interpreted as a Gaussian process conditional on the ordering. Due to the structure of the splitting criteria, a high-dimensional CLT for hyper-rectangles is sufficient. For $l = \texttt{SSE}$, we stack the control and treatment groups process in a twice as long high-dimensional vector. However, due to the structure the splitting criteria in this case, we employ instead Chernozhukov et al. [2017, Proposition 3.1], which gives a high-dimensional CLT for convex sets.

Step 3: Unconditional Gaussian Approximation. For the special case of $p = 1$, this step is not necessary because there is only one ordering possible. However, for $p > 1$, recursive decision trees find the best split along each dimension of $\mathbf{x}_i$, which implies a different ordering of the vector. Nevertheless, we show that the conditional Gaussian process from step 2 is close to an unconditional Gaussian process with zero correlation for different split coordinate indexes. Zero correlation between splits of different coordinates implies that the (sub)-processes corresponding to splitting different coordinates are asymptotically independent, reducing the problem to studying the arg max of the split criterion over one coordinate. The result is proven by applying a Gaussian-to-Gaussian comparison inequality [Chernozhuokov et al., 2022, Proposition 2.1], after establishing an upper bound on the matrix max norm of the difference between the conditional covariance matrix (which depends on the ordering) and the unconditional covariance matrix (which does not depend on the ordering). For $l = \texttt{DIM}$ and $l = \texttt{IPW}$, the results is immediate because the high-dimensional CLT was established over hyper-rectangles. For $l = \texttt{SSE}$, the additional error induced by considering a simple convex sets approximation is be controlled using Nazarov's inequality [Nazarov, 2003].

Step 4: Lower bound on imbalanced split probability. The unconditional Gaussian approximation processes from Step 3 take the form of the square Euclidean norm of a univariate (for $l \in \{\texttt{DIM}, \texttt{IPW}\}$) or bivariate (for $l = \texttt{SSE}$) Ornstein-Uhlenbeck process, where the split and time of Ornstein-Uhlenbeck process satisfies a one-to-one transformation [Csörgö and Révész, 1981, Göing-Jaeschke and Yor, 2003]. Since Darling-Erdös [Eicker, 1979, Horváth, 1993] allows for calculation of the maximum of norm of an O-U process within any time interval, we can find the lower bound on the probability of split occurs with a small or large index from (6) with the help of Gaussian correlation inequality [Latała and Matlak, 2017, Remark 3 (i)]. In turn, this characterizes precisely the effective sample sizes of each child node.

The remaining of our proofs leverage the technical insights above, applying then recursively to understand deeper tree constructions and the concentration in probability properties of the resulting CATE estimates.

5 X-Adaptivity and Inconsistency

The estimators considered in Theorem 1 either employ the full sample in their entire construction, or they rely on a two-sample independent split (honesty), where one subsample is use for training the tree, and the other is used for estimation of the conditional average treatment effects. As discussed in Devroye et al. [2013], and references therein, **X**-adaptivity offers a middle ground between the two sample usage designs considered in Definition 2: the tree construction and the final estimation step share the same covariates but each step employs different outcomes variables, that is, the two subsamples are independence conditional on the covariates.

We leverage the idea of **X**-Adaptivity, and study causal tree estimators where the outcome variable and treatment indicator are independent across all levels of the tree construction and the final CATE estimation step, but the same covariates are used throughout. This **X**-adaptive data design is of theoretical interest because it offers a bridge between no-sample-splitting and honesty. The following definition formalizes the construction of the **X**-adaptive causal tree estimators.

Definition 4 (**X**-Adaptive Estimation). *Recall Definition 1 and Definition 2, and that* $\mathscr{D} = \{(y_i, \mathbf{x}_i^\top, d_i) : i = 1, 2, \dots, n\}$ *is the available random sample.*

1. *The dataset* $\mathscr{D}$ *is divided into* $K + 1$ *datasets* $(\mathscr{D}_{\mathsf{T}_1}, \dots, \mathscr{D}_{\mathsf{T}_K}, \mathscr{D}_\tau)$*, with sample sizes given by* $(n_{\mathsf{T}_1}, \dots, n_{\mathsf{T}_K}, n_\tau)$*, respectively, and satisfying* $n_{\mathsf{T}_1} = \cdots = n_{\mathsf{T}_K} = n_\tau$ *(possibly after*

dropping $n \mod K$ data points at random). For each of the datasets $\mathscr{D}_j = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, \ldots, n_{\mathsf{T}_j}\}$, $j = 1, \ldots, K$, replace $\{(y_i, d_i) : i = 1, \ldots, n_{\mathsf{T}_j}\}$ with independent copies $\{(\tilde{y}_i, \tilde{d}_i) : i = 1, \ldots, n_{\mathsf{T}_j}\}$, while keeping the same $\{\mathbf{x}_i : i = 1, \ldots, n_{\mathsf{T}_j}\}$.

2. *The maximal decision tree of depth K, $\mathsf{T}^l_K(\mathscr{D}_{\mathsf{T}_1}, \cdots, \mathscr{D}_{\mathsf{T}_K})$, is obtained by iterating K times the $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$ splitting procedures in Definition 2, each time splitting* all *terminal nodes until (i) the node contains a single data point $(y_i, d_i, \mathbf{x}_i^\top)$, or (ii) the input values $\mathbf{x}_i$ and/or all (d_i, y_i) within the node are the same.*

3. *The* **X***-adaptive estimators are*

$$\hat{\tau}^{\mathtt{X}}_{\mathtt{DIM}}(\mathbf{x}; K) = \hat{\tau}_{\mathtt{DIM}}(\mathbf{x}; \mathsf{T}^{\mathtt{DIM}}_K(\mathscr{D}_{\mathsf{T}_1}, \ldots, \mathscr{D}_{\mathsf{T}_K}), \mathscr{D}_\tau),$$
$$\hat{\tau}^{\mathtt{X}}_{\mathtt{IPW}}(\mathbf{x}; K) = \hat{\tau}_{\mathtt{IPW}}(\mathbf{x}; \mathsf{T}^{\mathtt{IPW}}_K(\mathscr{D}_{\mathsf{T}_1}, \ldots, \mathscr{D}_{\mathsf{T}_K}), \mathscr{D}_\tau), \quad \textit{and}$$
$$\hat{\tau}^{\mathtt{X}}_{\mathtt{SSE}}(\mathbf{x}; K) = \hat{\tau}_{\mathtt{DIM}}(\mathbf{x}; \mathsf{T}^{\mathtt{SSE}}_K(\mathscr{D}_{\mathsf{T}_1}, \ldots, \mathscr{D}_{\mathsf{T}_K}), \mathscr{D}_\tau).$$

As in the previous cases, if the distributions of $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are symmetric about zero, then the **X**-adaptive estimators are unbiased: $\mathbb{E}[\hat{\tau}^{\mathtt{X}}_l(\mathbf{x}; K)] = \tau$, for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$.

Theorem 3 (Accuracy of **X**-Adaptive Causal Tree Estimators)**.** *Suppose Assumption 1 holds and additionally that $\mathbb{E}[\varepsilon_i^2(0)] = \mathbb{E}[\varepsilon_i^2(1)]$. Then, for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$,*

$$\liminf_{n \to \infty} \mathbb{P}\Big(\sup_{\mathbf{x} \in \mathscr{X}} \big| \hat{\tau}^{\mathtt{X}}_l(\mathbf{x}; K_n) - \tau(\mathbf{x}) \big| \geq C_1 \Big) \geq C_2,$$

provided that $\liminf_{n \to \infty} \frac{K_n}{\log \log n} = \kappa > 0$, and where the positive constants C_1 and C_2 only depend on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$ and κ.

Furthermore, for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$ and any $K \geq 1$,

$$\mathbb{E}\Big[\int_{\mathscr{X}} \big(\hat{\tau}^{\mathtt{X}}_l(\mathbf{x}, K) - \tau(\mathbf{x}) \big)^2 dF_{\mathbf{X}}(\mathbf{x}) \Big] \leq C_3 \frac{K 2^K}{n},$$

where the positive constant C_3 only depends on the distribution of $(\varepsilon_i(0), \varepsilon_i(1), d_i)$.

The theorem establishes uniform inconsistency of the **X**-adaptive causal tree estimator so long as $K_n \gtrsim \log \log n$. To put this side rate restriction in perspective, if $n/K_n \approx 1$ billion then $\log \log(10^9) \approx 3$. Therefore, the inconsistency of the estimator will manifest as soon as $K_n \approx 3$, a shallow tree when compared to those commonly encountered in practice (even in settings with much more moderate sample sizes, that is, with n much smaller than K_n billions). This result also shows that the integrated mean square error (IMSE) of a uniformly inconsistent $\mathscr{X}$-adaptive causal tree estimator can nonetheless decay at the optimal $\sqrt{n}$ rate, up to a poly-logarithmic-n factor. As demonstrated before, the performance of the causal tree estimators can vary widely depending on whether the input $\mathbf{x}$ is average or worst case.

6 Discussion

6.1 Decision Stumps

The phenomenon of generating unbalanced cells in adaptive recursive partitioning schemes has been observed in various forms since the inception of CART. Historically, this phenomenon has been called the *end-cut preference*, where splits along noisy directions tend to concentrate along the boundary of the parent node. More specifically, considering the standard CART for regression estimation without sample splitting, Breiman et al. [1984, Theorem 11.1] and Ishwaran [2015, Theorem 4] showed that in one-dimension ($p = 1$), for each $\delta \in (0, 1)$, $\mathbb{P}(n(\mathtt{t}_\mathtt{L}) \leq \delta n \text{ or } n(\mathtt{t}_\mathtt{R}) \geq (1-\delta)n) \to 1$ as $n \to \infty$. If applicable to the context of this paper, their result would only imply rates in uniform norm slower than any *constant multiple* of the already nearly optimal rate $\sqrt{n/\log\log(n)}$, i.e., for any $C > 0$,

$$\liminf_{n\to\infty} \mathbb{P}\Big(\sup_{x\in\mathcal{X}} \big|\hat{\tau}_l^{\mathtt{NSS}}(x) - \tau(x)\big| \geq C\sigma n^{-1/2}\sqrt{\log\log(n)} \Big) = 1.$$

In contrast, our results hold for all $p \geq 1$ and characterize precisely the regions of the support $\mathcal{X}$ where the pointwise rates of estimation are slower than any polynomial-in-n (see Corollary SA-7, Theorem SA-14, Corollary SA-21 in the supplemental appendix). Thus, past theoretical work is not strong enough to illustrate the weaknesses of causal trees for pointwise estimation (i.e., prior lower bounds in the literature would be too loose to be informative). Furthermore, our results also study settings where sample splitting (honesty) is used, and demonstrate that they cannot mitigate the low convergence rate of adaptive causal trees under Assumption 1. Last but not least, our results apply to the causal tree constructions which are different (and more complicated) than those in plain vanilla CART regression (Definition 2).

6.2 Deeper Trees, Multivariate Covariates, and the Location of Small Cells

Our theoretical results show that, under Assumption 1, the first split of any decision tree construction will generate a small child cell with non-vanishing probability. As a result, and due to their recursive nature, deeper tree constructions will have multiple regions with too small sample sizes (with non-vanishing probability). This problem is exacerbated in multiple dimensions ($p > 1$), which is exactly the setting where causal tree estimators would be potentially more useful to uncover treatment effect heterogeneity.

The small regions of the support $\mathcal{X}$, and hence the slower than any polynomial-in-n convergence rate (or inconsistency) of causal tree estimators, need not occur near a region of the boundary of $\mathcal{X}$. At each stage in the tree construction, a parent node $\mathtt{t}$ will generate two child nodes, one small and the other large, but the splitting may realize anywhere on $\mathtt{t}$ (parent cell) and along any individual covariate (in $\mathbf{x}_i$, or axis), thereby generating problematic hyper-rectangle cells all over the support $\mathcal{X}$ with non-vanishing probability.

6.3 Regularization and Bias

It is tempting to try to regularize the decision tree estimator in order to eliminate the small cell problem, and thus improve its convergence rate. For instance, the tree construction algorithm may not split a parent node if the effective sample size is to small, or it may include a penalty term for overfitting. However, it is also important to note that adaptive decision tree constructions purposely select small cells for two opposing reasons: misspecification bias vs. low signal-to-noise ratio. More precisely, on the one hand, if the unknown conditional expectation function exhibits high curvature (bias) in a certain region of $\mathscr{X}$, then the tree construction will tend to generate a small child cell (node) in that region to reduce misspecification bias, which is precisely a celebrated feature of an "adaptive" procedure. On the other hand, as shown in this paper, small cells also emerge with non-vanishing probability when there is no misspecification bias in that region, that is, when the unknown conditional expectation function is locally constant. In practice, it is impossible to distinguish between the two equality possible scenarios.

Our theoretcal results purposely remove misspecification bias by considering data generating processes with constant conditional expectation functions. In real application settings, however, the conditional expectation functions may exhibit heterogeneity (even if locally constant), in which case regularization to remove small cells may led to large bias in the causal decision tree estimators, also affecting their convergence rate.

6.4 α-Regularity and Causal Random Forests

Under specific assumptions, Wager and Athey [2018] and others established polynomial-in-n convergence rates for honest causal trees and forests. The slow convergence rates establish in Theorem 1 do not contradict, but are rather precluded by existing polynomial-in-n convergence guarantees in the literature because they assume that each split generates two child nodes that contain a constant fraction of the number of observations in the parent node, i.e., $n(\mathtt{t}_{\mathtt{L}}) \gtrsim n(\mathtt{t})$ and $n(\mathtt{t}_{\mathtt{R}}) \gtrsim n(\mathtt{t})$. The key assumption is often called α-regularity, because it assumes that the tree construction generates an $\alpha > 0$ proportion of the data in each terminal node (cell).

Our theoretical results imply that assumptions such as α-regularity, or variants thereof, which require *balanced* cells almost surely, are incompatible with standard decision tree constructions employing causal trees [Athey and Imbens, 2016] or any other conventional CART methodology [e.g., Behr et al., 2022, and references therein]. By implication, results for causal random forests relying on α-regularity, or variants thereof, do not apply to standard recursive partitioning using CART-type algorithms. Some form of (algorithmic and/or statistical) regularization is needed, thereby introducing a bias in the estimation as well as additional tuning parameters that would need to chosen in practice.

6.5 Decision Tree Regression

The supplemental appendix also studies standard adaptive decision tree regression via CART for nonparametric estimation of the conditional expectation of an output given a collection of features. Section SA-2 in the supplemental appendix establishes an analogue of Theorem 1, demonstrating that adaptive decision tree regression exhibits slow convergence rate or inconsistency, as causal trees do, depending on the sample splitting design used.

Our results are connected to Bühlmann and Yu [2002] and Banerjee and McKeague [2007], and subsequent work in the statistical literature. They study large sample properties of the decision stump without sample splitting with a univariate covariate ($p = 1$ and $K = 1$), and show that the minimizers $(\hat{\beta}_{\mathtt{L}}, \hat{\beta}_{\mathtt{R}}, \hat{\varsigma})$ in (3) at the root node converge to well-defined population minimizers $(\beta^*_{\mathtt{L}}, \beta^*_{\mathtt{R}}, \varsigma^*)$ at a cube-root rate $n^{1/3}$ when the population minimizers are unique and the population conditional expectation function is continuously differentiable and has nonzero derivative at ς^*, among other technical conditions. Thus, our results show that the conclusion in Bühlmann and Yu [2002] and Banerjee and McKeague [2007] are not uniformly valid over the class of conditional expectation functions: the exclusion of the constant regression function from the allowed class of data generating processes is necessary for their results to hold for all values of the scalar covariate.

6.6 Invalidity of Inference Methods

Theorem 1 establishes lower bounds on the uniform convergence rate of causal decision tree estimators. The main technical observation is that these estimation procedures will generate a partition of $\mathscr{X}$ with highly unbalanced cells, where potentially many cells will have a very small number of samples. These results are established under Assumption 1, which does not assume a parametric family of distributions on the data, but rather only independence and moment conditions.

From an inference perspective, our results also show that a valid (Gaussian or otherwise) distributional approximation for the causal decision tree estimators, after perhaps properly centering and scaling, does not hold in general. The main obstacle is that the effective sample size may not even increase for the approximation to apply in many regions of $\mathscr{X}$. In particular, standard inference methods, such as the usual confidence intervals of the form $\hat{\tau}^q_l(\mathbf{x}) \pm z_\alpha \cdot \text{Sd.Err.}(\hat{\tau}^q_l(\mathbf{x}))$ with z_α denoting the usual quantile of the standad Gaussian distribution, Sd.Err.() a standard error estimator, and $q \in \{\mathtt{NSS}, \mathtt{HON}, \mathtt{X}\}$, will not deliver asymptotically valid inference for the parameter of interest $\tau(\mathbf{x})$.

7 Simulations

We illustrate the implications of Theorem 1 in the univariate case $p = 1$. Figure 1 reports the pointwise root mean squared error $\text{RMSE}(x) = \big\{\mathbb{E}\big[(\hat{\tau}^q_\ell(x) - \tau)^2\big]\big\}^{1/2}$, for $\ell \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$ and $q \in \{\mathtt{NSS}, \mathtt{HON}, \mathtt{X}\}$, estimated from 2,000 Monte Carlo replications under $\tau = \mu_0 = \mu_1 = 0$, $\varepsilon_i(0), \varepsilon_i(1) \overset{\text{i.i.d.}}{\sim} \mathsf{N}(0,1)$, $X_i \sim \text{Uniform}[0,1]$, and $n = 1{,}000$. For each of the nine causal-tree estimators, we consider depths $K \in \{1, \dots, 5\}$, where curves are color-coded by K.

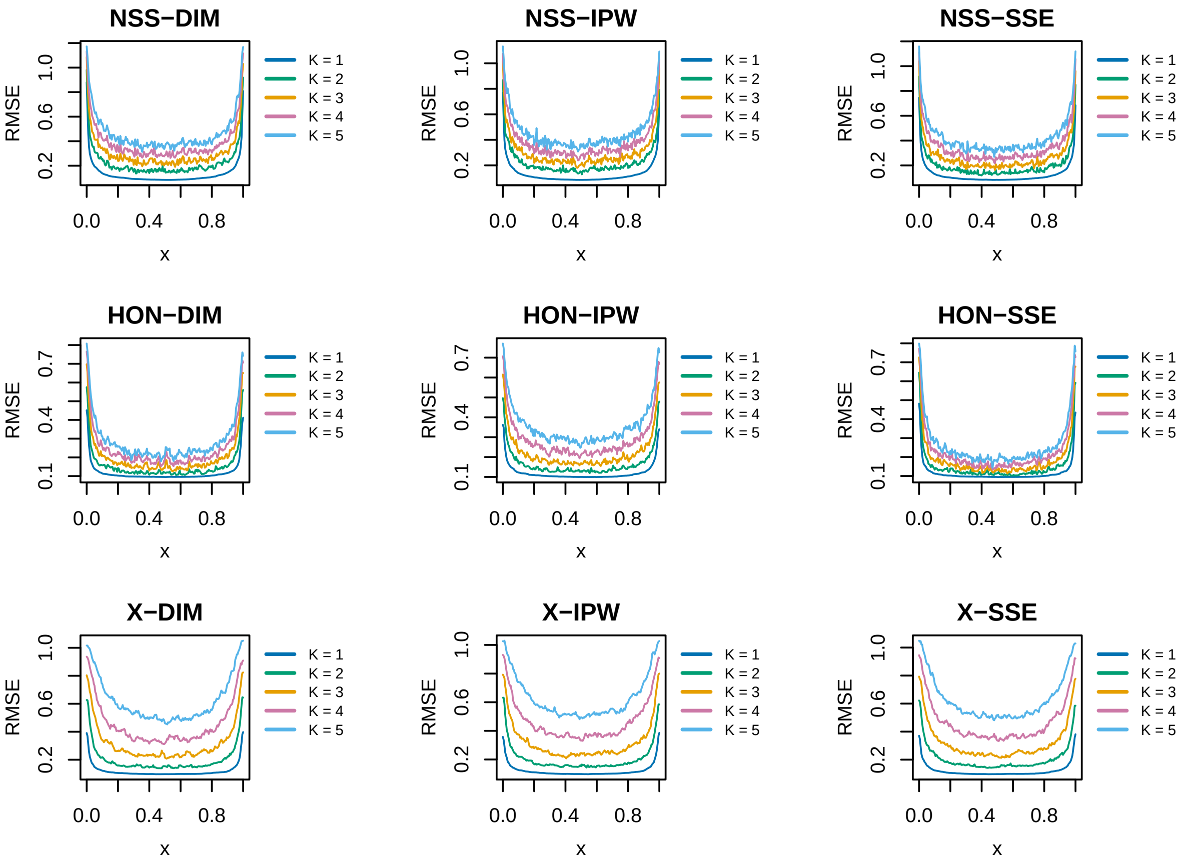

Figure 1: Plots of root mean-squared error (RMSE) of heterogeneous treatment effect estimation using nine distinct causal tree methods with depth $K = 1, 2, \cdots, 5$. We chose $p = 1$, and the univariate covariate X is supported on $[0, 1]$. For all methods and depths, the causal tree has smallest pointwise RMSE near the center of the covariate space, but the performance degrades as the evaluation points move closer to the boundary. The experiment is conducted with 2,000 Monte-Carlo simulations.

Two patterns emerge across all nine methods: (i) For any fixed K, the pointwise RMSE is smallest near the center of the covariate space and increases as x approaches the boundary; (ii) For any fixed $x \in [0, 1]$, the RMSE increases with tree depth K. The first pattern is due to the small cells near boundary predicted by (6), rendering a situation where local averaging is less accurate. The second is consistent with the `X`-results of Theorem 1 and, heuristically, extends to `NSS` and `HON`: at higher depths, a larger fraction of evaluation points lie near terminal node boundaries, where the same boundary effects that govern decision stumps degrade performance, leading to increased RMSE even for interior points.

Acknowledgments

The authors thank Benjamin Budway, Max Farrell, Boris Hanin, Felix Hoefer, Michael Jansson, Joowon Klusowski, Boris Shigida, Jantje Sönksen, Jennifer Sun, Rocio Titiunik, and Kevin Zhang for comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through SES-1947805, SES-2019432, and SES-2241575. Klusowski gratefully acknowledges financial support from the National Science Foundation through CAREER DMS-2239448.

References

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

Moulinath Banerjee and Ian W. McKeague. Confidence sets for split points in decision trees. *Annals of Statistics*, 35(2):543 – 574, 2007.

Merle Behr, Yu Wang, Xiao Li, and Bin Yu. Provable boolean interaction recovery from tree ensemble obtained via random forests. *Proceedings of the National Academy of Sciences*, 119 (22):e2118636119, 2022.

Richard A Berk. *Statistical learning from a regression perspective.* Springer Series in Statistics. Springer Nature, 2020.

Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone. *Classification and Regression Trees.* Chapman and Hall/CRC, 1984.

Peter Bühlmann and Bin Yu. Analyzing bagging. *Annals of Statistics*, 30(4):927 – 961, 2002.

Matias D. Cattaneo, Max H. Farrell, and Yingjie Feng. Large sample properties of partitioning-based series estimators. *Annals of Statistics*, 48(3):1718–1741, 2020.

Matias D Cattaneo, Jason M Klusowski, and Peter M Tian. On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *Technical report, arXiv preprint arXiv:2211.10805*, 2022.

Matias D. Cattaneo, Rajita Chandak, and Jason M. Klusowski. Convergence rates of oblique regression trees for flexible function libraries. *Annals of Statistics*, 52(2):466 – 490, 2024.

Matias D Cattaneo, Yingjie Feng, and Boris Shigida. Uniform estimation and inference for nonparametric partitioning-based m-estimators. *arXiv preprint arXiv:2409.05715*, 2025.

Victor Chernozhukov, Denis Chetverikov, and Kengo Kato. Central limit theorems and bootstrap in high dimensions. *Annals of Probability*, 45(4):2309 – 2352, 2017.

Victor Chernozhuokov, Denis Chetverikov, Kengo Kato, and Yuta Koike. Improved central limit theorem and bootstrap approximations in high dimensions. *Annals of Statistics*, 50(5):2562–2586, 2022.

Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, December 2022.

Hugh A. Chipman, Edward I. George, and Robert E. McCulloch. BART: Bayesian additive regression trees. *Annals of Applied Statistics*, 4(1):266 – 298, 2010.

M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis.* Wiley, 1997.

M. Csörgö and P. Révész. *Strong Approximations in Probability and Statistics.* Probability and Mathematical Statistics : a series of monographs and textbooks. Academic Press, 1981.

Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*, volume 31. Springer Science & Business Media, 2013.

F. Eicker. The asymptotic distribution of the suprema of the standardized empirical processes. *Annals of Statistics*, 7(1):116 – 138, 1979.

Ran El-Yaniv and Dmitry Pechyony. Transductive rademacher complexity and its applications. *Journal of Artificial Intelligence Research*, 35:193–234, 2009.

Anja Göing-Jaeschke and Marc Yor. A survey and some generalizations of bessel processes. *Bernoulli*, 9(2):313 – 349, 2003.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression.* Springer-Verlag, 2002.

Miguel A. Hernán and James M. Robins. *Causal Inference: What If.* Boca Raton: Chapman & Hall/CRC, 2020.

Lajos Horváth. The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of statistics*, 21(2):671–680, 1993.

Hemant Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99(1):75–118, 2015.

Jason M Klusowski and Peter M Tian. Large scale prediction with decision trees. *Journal of the American Statistical Association*, 119(545):525–537, 2024.

Rafał Latała and Dariusz Matlak. *Royen's Proof of the Gaussian Correlation Inequality*, pages 265–275. Springer International Publishing, 2017.

Rahul Mazumder and Haoyue Wang. On the convergence of CART under sufficient impurity decrease condition. *Advances in Neural Information Processing Systems*, 36, 2024.

Fedor Nazarov. On the maximal perimeter of a convex set in $\mathbb{R}^n$ with respect to a Gaussian measure. In *Geometric Aspects of Functional Analysis: Israel Seminar, 2001–2002*, pages 169–187. Springer, 2003.

Valentin V. Petrov. On lower bounds for tail probabilities. *Journal of Statistical Planning and Inference*, 137(8):2703–2705, 2007.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716 – 1741, 2015.

Galen R Shorack and RT Smythe. Inequalities for max— sk—/bk where k ∈ nr. *Proceedings of the American Mathematical Society*, pages 331–336, 1976.

Maciej Skorski. Bernstein-type bounds for beta distribution. *Modern Stochastics: Theory and Applications*, 10(2):211–228, 2023.

Yan Shuo Tan, Abhineet Agarwal, and Bin Yu. A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds. In *International Conference on Artificial Intelligence and Statistics*, pages 9663–9685. PMLR, 2022.

Yan Shuo Tan, Jason M Klusowski, and Krishnakumar Balasubramanian. Statistical-computational trade-offs for recursive adaptive partitioning estimators. *arXiv preprint arXiv:2411.04394*, 2024a.

Yan Shuo Tan, Omer Ronen, Theo Saarinen, and Bin Yu. The computational curse of big data for bayesian additive regression trees: A hitting time analysis. *arXiv preprint arXiv:2406.19958*, 2024b.

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

Heping Zhang and Burton H Singer. *Recursive Partitioning and Applications*. Springer, 2010.