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

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

March 17, 2026

Abstract

Recursive decision trees are widely used to estimate heterogeneous causal treatment effects in experimental and observational studies. These methods are typically implemented using CART-type recursive partitioning and are often viewed as adaptive procedures capable of discovering treatment effect heterogeneity in high-dimensional settings. We study causal tree estimators based on adaptive recursive partitioning and establish lower bounds on their estimation accuracy. Under basic conditions, we show that causal trees constructed via standard CART-type splitting rules cannot achieve polynomial-in-n convergence rates in the uniform norm (where n denotes the sample size). The underlying mechanism is that greedy recursive partitioning selects highly imbalanced splits with non-vanishing probability, producing terminal nodes containing very few observations and leading to large estimation variance. We further show that sample splitting ("honesty") yields at most negligible improvements in convergence rates. As a consequence, causal tree estimators may converge arbitrarily slowly and can even be inconsistent in some settings. Our results also clarify the role of balanced partition assumptions in existing theoretical guarantees for causal forests and related ensemble methods. The analysis develops new probabilistic tools for studying adaptive recursive partitioning procedures, including non-asymptotic approximations for suprema of partial sums and Gaussian processes. As a technical by-product, we also identify and correct an error in Eicker [1979].

Keywords: machine learning, recursive partitioning, adaptive decision trees, minimax convergence rates

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

Contents

1 Introduction

Recursive decision trees have become a popular tool for estimating heterogeneous causal effects in both experimental and observational settings. These methods adapt the classical CART (Classification and Regression Tree) algorithm [Breiman et al., 1984] to causal inference by modifying the splitting criterion and (sometimes) using sample splitting to separate tree construction from treatment effect estimation. Due to their simplicity and interpretability, causal trees have been widely adopted in both academic research and industry applications over the last decade. A leading example is the *honest causal tree* framework of Athey and Imbens [2016], which is often understood as allowing the data to "discover relevant subgroups while preserving the validity of confidence intervals constructed on treatment effects within subgroups" [Athey and Imbens, 2016, page 7353].

Despite their widespread use, the fundamental statistical properties of causal tree estimators and their associated inference procedures remain poorly understood. Existing theoretical analyses typically rely on strong assumptions about the tree-growing process that are incompatible with canonical implementations of CART-type recursive partitioning under basic conditions. This paper studies heterogeneous treatment effect estimators based on adaptive recursive partitioning and establishes theoretical lower bounds on their estimation accuracy. More broadly, our results reveal a fundamental limitation of CART-type recursive partitioning for uniform estimation of heterogeneous treatment effects.

Our main finding is that causal tree estimators constructed via standard CART-type greedy splitting cannot achieve polynomial-in-n convergence rates in the uniform norm under basic conditions, where n denotes the sample size. More precisely, we show that these estimators converge more slowly than any polynomial-in-n rate uniformly over the covariate space. This failure arises even in the simplest setting in which the true treatment effect is constant over the feature space.

The mechanism behind this phenomenon is intrinsic to greedy recursive partitioning. CART-type splitting rules select highly imbalanced splits with non-vanishing probability, thereby generating terminal nodes with very few observations. Because treatment effects within each node are estimated by local averaging, these small cells induce large estimation variance and prevent the estimator from concentrating uniformly around the true conditional average treatment effect.

A natural response to this small-cell phenomenon is to impose additional regularization on the tree construction. In practice, this often takes the form of minimum node-size requirements. Although such modifications may reduce finite-sample variance, they also alter the class of feasible partitions of the procedure. In particular, ruling out small cells necessarily introduces additional approximation bias whenever the conditional average treatment effect (CATE) is not locally constant, and this problem can be even more pronounced under conditional heteroskedasticity. More broadly, the more aggressively one regularizes the tree through minimum node-size requirements, the more one weakens the adaptivity that makes CART attractive in the first place. For this reason, in practice, causal tree methods are typically regularized through fixed or lightly tuned minimum node-size hyperparameters, often on the order of only a few tens of observations, rather than through carefully tuned sample-size-dependent hyperparameters that grow with n, as would be required to address the kind of convergence issues raised here.

At present, however, the statistical behavior of CART-type estimators under such algorithmic regularization remains only partially understood in general settings where the partition is selected adaptively from the data. For this reason, regularization alone does not provide a theoretically satisfactory resolution of the phenomenon we document unless one can formally characterize the resulting bias–variance trade-off and establish valid inference for the resulting estimator. Moreover, even apart from these statistical concerns, it has been argued that imbalanced splits can be algorithmically beneficial for deeper trees, because they preserve sample size after a bad split and allow the tree to recover downstream [Ishwaran, 2015].

Another frequently discussed modification of adaptive recursive partitioning methods is sample splitting, also known as *honesty*. By constructing the tree on one subsample and estimating treatment effects on an independent subsample, honesty aims to reduce overfitting and improve the validity of inference. While honesty does yield modest improvements in the achievable convergence rate, our results demonstrate that it does not resolve the fundamental limitations of adaptive recursive partitioning. Even under honest sample splitting, causal tree estimators cannot achieve polynomial-in-n uniform convergence rates under basic conditions.

Our findings also have implications for theoretical guarantees established for causal forests and related ensemble methods. Existing polynomial-rate analyses of forest estimators often rely on conditions ensuring that each constituent tree generates approximately balanced partitions. These conditions are commonly formalized through the so-called α-regularity assumption [Wager and Athey, 2018], requiring each split to allocate a non-negligible fraction of the observations in the parent node to each child node. While such assumptions facilitate theoretical analysis, they are incompatible with canonical CART-type greedy splitting rules: under basic conditions, CART constructions can select highly imbalanced splits with non-vanishing probability. Consequently, theoretical guarantees for causal forests that rely on quasi-uniform partitions or α-regularity do not directly extend to implementations based on standard CART-type splitting procedures. Moreover, even within the α-regular framework itself, these guarantees lead to a counterintuitive conclusion. In the class analyzed by Wager and Athey [2018, see Theorem 3.2 and condition (14) in Theorems 3.1 and 4.1], both the bias exponent and the subsampling-rate exponent improve monotonically with α. The available upper bounds therefore favor increasingly balanced splits. If one extrapolates these expressions over $\alpha \in (0, 1/2]$, the most favorable bound is achieved at $\alpha = 1/2$, that is, at the maximally balanced case. See Section 6 for further discussion.

From an inferential perspective, the same mechanism underlying our lower bounds also creates a fundamental obstacle for subgroup-level inference based on causal trees and their ensembles. In regions where terminal nodes contain very few observations, the effective sample size may fail to increase with the overall sample size. In such settings, classical Gaussian approximations and standard error formulas need not be valid. Establishing distributional approximations for CART-based causal tree or forest estimators without imposing quasi-uniform partition conditions therefore remains an open theoretical problem.

Before introducing the formal setup, it is useful to briefly summarize the intuition behind our results. Recursive decision trees select splits by optimizing a data-dependent criterion over many candidate partitions of the covariate space. When the underlying conditional expectation function is locally flat, as in the constant treatment effect model considered in this paper, the splitting criterion is driven primarily by stochastic fluctuations in the data. As a consequence, with non-vanishing probability the optimal split occurs near the boundary of a parent node, producing highly imbalanced child nodes. Because treatment effects within each node are estimated by local averaging, such splits generate terminal nodes containing very few observations, leading to large estimation variance in those regions of the covariate space. The recursive nature of the tree construction then propagates this phenomenon across deeper levels of the tree, creating multiple regions with small effective sample sizes and ultimately preventing uniform convergence of the estimator.

1.1 Contributions and Related Literature

This paper makes three main contributions. First, we establish lower bounds on the uniform convergence rate of causal tree estimators constructed via CART-type recursive partitioning. These results show that such estimators cannot achieve polynomial-in-n convergence rates under basic conditions, even in settings where the underlying treatment effect is constant. Second, we analyze the role of sample splitting (honesty) in causal tree estimation. While honesty improves the achievable convergence rate by removing a slowly varying $\sqrt{\log \log n}$ factor, our results demonstrate that it does not eliminate the fundamental limitations of adaptive recursive partitioning. Third, we clarify the implications of these findings for tree-based causal inference more broadly, including commonly used regularization strategies and theoretical guarantees for causal forests that rely on quasi-uniform partitions or α-regularity conditions. Our analysis shows that these assumptions are incompatible with canonical CART-type implementations and highlights the resulting challenges for valid inference based on tree-grown partitions.

Our work complements the rich theoretical literature on 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 a growing body of negative results for tree-based methods. For example, Tan et al. [2022] demonstrated that regression trees are inefficient at estimating additive structure, regardless of the optimization strategy employed. Tan et al. [2024] established that the mixing time of Bayesian Additive Regression Trees (BART) [Chipman et al., 2010] can increase with the training sample size. Finally, Tan et al. [2026] showed that adaptive regression trees with Boolean covariates may require exponentially many samples in the dimension and can be inconsistent in high-dimensional settings.

The present paper supersedes the unpublished manuscript of Cattaneo, Klusowski, and Tian [2022], which studied the behavior of a one-dimensional regression stump constructed via CART and showed that its estimation error can converge arbitrarily slowly. That work conjectured that similar phenomena would arise for causal trees and that honesty would not eliminate the problem. The present paper confirms both conjectures and substantially generalizes those findings. In particular, we establish lower bounds for causal tree estimators in arbitrary covariate dimension and for any causal tree structure containing at least one split, allowing for arbitrary tree depth. Furthermore, the supplemental appendix reports analogous results for standard adaptive regression trees of arbitrary depth, and with and without sample splitting. These theoretical results clarify the role of necessary and sufficient conditions in the work of Bühlmann and Yu [2002] and Banerjee and McKeague [2007], among others. See Section 6 for more discussion.

The proofs rely on a collection of new probabilistic insights concerning the behavior of adaptive recursive partitioning procedures. Our analysis develops non-asymptotic approximations for suprema of partial sums and related Gaussian processes, leveraging high-dimensional central limit theorems, Gaussian comparison inequalities, and strong approximation techniques. Key technical tools include results from Chernozhukov et al. [2017], Chernozhuokov et al. [2022], Csörgö and Révész [1981], Csörgö and Horváth [1997], El-Yaniv and Pechyony [2009], Horváth [1993], Göing-Jaeschke and Yor [2003], Latała and Matlak [2017], Petrov [2007], Shorack and Smythe [1976], and Skorski [2023]. As a technical by-product of our analysis, we also identify and correct an error in Eicker [1979].

1.2 Organization

The remainder of the paper is organized as follows. Section 2 introduces the causal tree estimators and the data generating framework studied in the paper. Section 3 presents the assumptions underlying our theoretical analysis. Section 4 establishes lower bounds on the uniform convergence rate of causal tree estimators and provides complementary results on their integrated mean squared error. Section 5 studies **X**-adaptive variants of causal tree estimators and demonstrates their uniform inconsistency under increasing tree depth. Section 6 discusses the implications of our results for recursive partitioning methods, causal forests, and inference procedures. Section 7 presents simulation evidence illustrating the practical implications of the theoretical results. Appendix A identifies and corrects an error in Eicker [1979]. Additional theoretical results and all technical proofs are reported in the supplemental appendix.

2 Setup

We introduce the class of causal tree estimators studied in the paper. These estimators combine three components: (i) a within-node estimator of the conditional average treatment effect (CATE), (ii) a recursive partitioning rule used to construct the tree, and (iii) a data usage scheme determining whether sample splitting (honesty) is employed. Different choices of these components lead to the family of estimators analyzed in this paper.

The available data $\mathscr{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n\}$ is a random sample, where y_i is an outcome variable, $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,p})^\top$ 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], we assume that

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

where $y_i(1)$ and $y_i(0)$ denote the potential outcomes under treatment and control, respectively. In experimental settings the treatment assignment 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

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

which captures how treatment effects vary with observable pre-treatment covariates. In experimental settings, the CATE function is identifiable because

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

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) expresses the CATE as the difference of two conditional expectation functions based on observed data, while the second equality (2) represents the CATE as a single conditional expectation of the transformed outcome $y_i \frac{d_i - \xi}{\xi(1-\xi)}$.

Traditional semiparametric approaches estimate heterogeneous treatment effects by replacing the conditional expectations in (1) or (2) with nonparametric estimators. However, such methods may perform poorly in high-dimensional settings or when the structure of the regression functions (e.g., sparsity or additive separability) is unknown. Motivated by the recent success of adaptive machine learning methods, Athey and Imbens [2016] proposed estimating $\tau(\mathbf{x})$ using recursive decision trees. Their approach retains the greedy recursive construction of standard CART while modifying the splitting criterion to target treatment effect heterogeneity. They also proposed using sample splitting (the so-called "honesty" property) to decouple tree construction from the estimation of treatment effects within terminal nodes.

The introduction of honesty has been widely 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 does not overcome the fundamental limitations of recursive partitioning for heterogeneous causal effect estimation (or even for adaptive regression trees), yielding at best negligible logarithmic improvements in the effective sample size or dimension.

We study causal tree estimators obtained by combining three components: two within-node treatment effect estimators (difference-in-means and inverse probability weighting), two tree construction criteria (variance maximization and SSE minimization), and two sample splitting schemes (with and without honesty). These combinations yield nine causal tree estimators corresponding to commonly used implementations in practice.

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 $\mathbf{x} \in \mathcal{X}$.*

- *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 $\mathcal{X}$, and estimation of $\tau(\mathbf{x})$ uses only observations with covariates $\mathbf{x}_i$ belonging to the cell in the partition covering $\mathbf{x} \in \mathcal{X}$. 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. [2026], and references therein.

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

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

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 $\mathtt{t}$ in the tree (i.e., a region in $\mathscr{X}$) is divided into two child nodes, $\mathtt{t_L}$ and $\mathtt{t_R}$, by minimizing the sum-of-squares error (SSE),

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

where the solution yields estimates $(\hat{\beta}_{\mathtt{L}}, \hat{\beta}_{\mathtt{R}}, \hat{\varsigma}, \hat{\jmath})$, 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 $\mathtt{t}$ produces child nodes $\mathtt{t_L} = \{\mathbf{x} \in \mathtt{t} : \mathbf{e}_{\hat{\jmath}}^\top \mathbf{x} \leq \hat{\varsigma}\}$ and $\mathtt{t_R} = \{\mathbf{x} \in \mathtt{t} : \mathbf{e}_{\hat{\jmath}}^\top \mathbf{x} > \hat{\varsigma}\}$. More precisely, the normal equations imply that $\hat{\beta}_{\mathtt{L}} = \frac{1}{n(\mathtt{t_L})}\sum_{\mathbf{x}_i\in\mathtt{t_L}} y_i$ and $\hat{\beta}_{\mathtt{R}} = \frac{1}{n(\mathtt{t_R})}\sum_{\mathbf{x}_i\in\mathtt{t_R}} y_i$, 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 the procedure continues recursively 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 $\mathtt{t}_0 = \mathscr{X}$ 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 $\mathtt{t}$ (i.e., a terminal node partitioning $\mathscr{X}$) in a previous tree T' is divided into two child nodes, $\mathtt{t_L}$ and $\mathtt{t_R}$, forming the new tree T, by maximizing*

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

  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 $\mathtt{t}$ (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 differs from the original CART criterion (3) in that it explicitly selects splits that maximize the squared difference between the estimated treatment effects in the two child nodes. In this sense, the rule directly targets treatment effect heterogeneity across the partition. 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 underlying our analysis is that these recursive procedures do not generate quasi-uniform partitions of the covariate space, and thus known results in the nonparametric partitioning-based estimation literature [Györfi et al., 2002, Cattaneo et al., 2020, 2026] 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, procedures without sample splitting are not "dishonest" in any formal sense; they are simply harder to analyze theoretically.

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* $\mathscr{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n\}$ *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*

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

- *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*

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

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

The following assumption describes the data generating process used throughout the analysis.

Assumption 1 (Data Generating Process)**.** $\mathscr{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : 1 \le i \le n\}$ *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* $\mathbf{x} \in \mathcal{X}$*, where* c_d *is some constant and* $\mathcal{X}$ *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 a simple randomized experiment with treatment probability $\xi \in (0,1)$. Assumption 1(ii) specifies a canonical regression representation for the potential outcomes, where $\mu_d(\mathbf{x})$ denotes the conditional mean function and $\varepsilon_i(d)$ is a mean-zero error term independent of $\mathbf{x}_i$. Assumption 1(iii) imposes a constant treatment effect, $\tau = c_1 - c_0$, across the covariate space. Assumption 1(iv) requires the covariates to be independent and continuously distributed. Because recursive decision trees are invariant to monotone transformations of the covariates, this condition can be replaced without loss of generality by assuming that $\mathbf{x}_i$ is uniformly distributed on $\mathcal{X} = [0,1]^p$. Finally, Assumption 1(v) requires the potential outcome errors to be sub-exponential, or equivalently to satisfy a Bernstein moment condition.

Because our goal is to establish lower bounds on the estimation accuracy of the causal tree estimators defined in Definition 3, it suffices to consider the constant treatment effect model in Assumption 1. The constant function belongs to all commonly studied smoothness classes, including Hölder classes and functions with bounded total variation. Consequently, lower bounds established under this model immediately extend to larger classes of data generating processes. Formally, for any estimator $\hat{\tau}(\mathbf{x})$ and any class of distributions $\mathscr{P}$ containing the distribution $\mathbb{P}_1$ satisfying Assumption 1,

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

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.

Assumption 1 also removes issues related to smoothing (or misspecification) bias, heteroskedasticity, and heavy-tailed errors. In particular, because the CATE function $\tau(\mathbf{x})$ is constant over $\mathcal{X}$, the results below are not driven by the usual (boundary or smoothing) bias arising in nonparametric estimation. For example, if the distributions of $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are symmetric about zero, then

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

for $l \in \{\texttt{DIM}, \texttt{IPW}, \texttt{SSE}\}$ and $\mathbf{x} \in \mathtt{t}$ where $\mathtt{t}$ is a terminal node in the tree. Unbiasedness of $\hat{\tau}_l^{\texttt{NSS}}(\mathbf{x})$ follows from the fact that the split points are symmetric functions of the residuals. In the case of $\hat{\tau}_l^{\texttt{HON}}(\mathbf{x})$, 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. In particular, $\hat{\tau}_l^{\texttt{HON}}(\mathbf{x})$ is unbiased when $\tau = 0$ (or for any other known treatment effect value), as well as in tree constructions ensuring that $\mathbb{P}(n(\mathtt{t}) = 0) = 0$. More generally, $\hat{\tau}_l^{\texttt{HON}}(\mathbf{x})$ is asymptotically unbiased whenever $\mathbb{P}(n(\mathtt{t}) = 0) \to 0$ as $n \to \infty$.

Consequently, the negative results established below are not driven by bias. Instead, they arise because adaptive recursive tree constructions can generate highly imbalanced partitions with non-vanishing probability, producing terminal nodes that contain very few observations. These small cells lead to large estimation variance in some regions of $\mathcal{X}$, which ultimately prevents the estimator from achieving polynomial convergence rates.

Finally, the constant treatment effect model can also be interpreted as a local approximation to smooth heterogeneous treatment effect functions. Recursive partitioning estimators approximate $\tau(\mathbf{x})$ using piecewise constant functions over the tree partition, which corresponds to a Haar basis representation. Our results therefore extend to shrinking neighborhoods of smooth functions around the constant function when the signal-to-noise ratio is sufficiently small.

4 Main Results

The following theorem establishes our main lower bound on the uniform accuracy of causal tree estimators. Let e denote the base of the natural logarithm.

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 $b \in (0,1)$,*

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

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

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

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 provides an overview of the proof strategy for Theorem 1, with all technical details deferred to the supplemental appendix (see Section SA-1.2). The proof relies on several non-asymptotic approximations for suprema of partial sums and Gaussian processes, building on 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, Appendix A corrects an error in Eicker [1979].

Theorem 1 establishes lower bounds on the uniform convergence rate of the six causal tree estimators introduced in Section 2. For procedures without sample splitting, the estimators $\hat{\tau}_{\mathtt{DIM}}^{\mathtt{NSS}}(\mathbf{x})$, $\hat{\tau}_{\mathtt{IPW}}^{\mathtt{NSS}}(\mathbf{x})$, and $\hat{\tau}_{\mathtt{SSE}}^{\mathtt{NSS}}(\mathbf{x})$ cannot achieve a uniform convergence rate of order $n^{-b/2}\sqrt{\log\log n}$ for any $b > 0$. In particular, these estimators converge more slowly than any polynomial rate in n, implying that their accuracy must deteriorate in some regions of the covariate space $\mathcal{X}$.

Athey and Imbens [2016] and subsequent work argue that sample splitting (the so-called "honesty" property) can improve the performance of machine learning estimators by decoupling model selection from estimation. The second result in Theorem 1 analyzes the corresponding honest causal tree estimators, $\hat{\tau}^{\mathtt{HON}}_{\mathtt{DIM}}(\mathbf{x})$, $\hat{\tau}^{\mathtt{HON}}_{\mathtt{IPW}}(\mathbf{x})$, and $\hat{\tau}^{\mathtt{HON}}_{\mathtt{SSE}}(\mathbf{x})$. The theorem shows that these estimators also fail to achieve polynomial convergence rates. While honesty improves the attainable rate slightly by removing the $\sqrt{\log\log n}$ factor, the resulting improvement is not even logarithmic in sample size.

Theorem 1 therefore paints a pessimistic picture for adaptive decision tree methods when the goal is to estimate heterogeneous treatment effects uniformly over the covariate space. In contrast, the same estimators can achieve near-optimal estimation accuracy when their performance is measured on average over $\mathscr{X}$, as shown by the following result. Let $F_{\mathbf{X}}(\mathbf{x}) = \mathbb{P}(\mathbf{x}_i \leq \mathbf{x})$.

Theorem 2 (Mean Square Accuracy)**.** *Suppose Assumption 1 holds and the underlying causal tree has depth at most $K \geq 1$. 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 Theorem 2 is given in the supplemental appendix (Section SA-1.2) and builds on ideas from Györfi et al. [2002] and Klusowski and Tian [2024]. Importantly, the result applies only under Assumption 1, that is, when the CATE function is constant. The purpose of this theorem is to highlight that, even in the same setting where uniform convergence fails, causal decision trees can still achieve favorable performance in an integrated mean-squared sense. It remains an open question whether near-optimal mean square convergence rates can be achieved over larger classes of functions by adaptive decision trees constructed using CART-type procedures.

A natural interpretation of the contrast between Theorem 1 and Theorem 2 relates to the often-discussed tension between causal inference and prediction in machine learning. Adaptive causal trees may perform poorly pointwise (and hence uniformly), yet perform well on average over the covariate space. This contrast highlights a fundamental limitation of adaptive recursive partitioning for learning heterogeneous treatment effects: while tree-based methods can deliver strong predictive performance on average, they may fail to provide reliable estimation guarantees uniformly over the covariate space.

From a technical perspective, the results in Theorem 2 are new in the context of causal tree estimation, particularly for the formal comparison between no-sample-splitting and honest implementations. Moreover, our theoretical analysis in the supplemental appendix establishes high-probability bounds for the integrated mean-squared error, allowing for a sharper comparison with

Theorem 1 (see Corollaries SA-12, SA-14, SA-32, SA-34 and Theorems SA-22 and SA-24). For example, in the case of no sample splitting we show that

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

for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$, where C_1 is the constant appearing in Theorem 2.

4.1 Proof Strategy of Theorem 1

The key idea behind the proof is that greedy recursive partitioning tends to select highly imbalanced splits with non-vanishing probability. In particular, even in the simplest case of a decision stump (a tree of depth one), the optimal split often occurs near the boundary of the covariate space, producing child nodes with very small sample sizes. These small cells lead to large estimation variance in some regions of $\mathscr{X}$, which ultimately prevents causal tree estimators from achieving polynomial uniform convergence rates.

Underlying our theoretical results are several technical properties of decision stumps, and hence trees of depth one. For each tree splitting criterion and sample-splitting design, we first analyze the probabilistic behavior of the split location at the root node. This analysis characterizes the regions of the covariate space $\mathscr{X}$ where the first split is most likely to occur and determines the effective sample sizes of the resulting child nodes.

Our results show that, with non-vanishing probability, the optimal split concentrates near the boundary of the parent node (a cell in the partition of $\mathscr{X}$). As a consequence, one of the child nodes may contain only a very small number of observations. This phenomenon arises at the very first step of the recursive tree construction and ultimately drives the slow uniform convergence rate. 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, for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$. Note that the first split coincide for no-sample-splitting and honest constructions. For each $a, b \in (0,1)$ with $a < b$ and $j \in \{1, 2, \dots, p\}$, we establish that

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

The slow uniform convergence rate of the decision stump estimator arises because the optimal split tends to concentrate near the boundary of the support, producing highly imbalanced partitions. In such cases, one child node contains only a small number of observations, making the corresponding local average estimator highly variable. Relation (6) quantifies this phenomenon: for each coordinate $j = 1, \dots, p$ and each $b \in (0,1)$, since $\mathbb{P}(\hat{\jmath} = j) = 1/p$ by symmetry, there is non-vanishing probability, $b/(pe)$, that one of the child cells $\{\mathbf{x} \in \mathscr{X} : x_j \leq \hat{\varsigma}\}$ or $\{\mathbf{x} \in \mathscr{X} : x_j > \hat{\varsigma}\}$ are highly anisotropic and will contain at most n^b samples. Consequently, with non-vanishing probability the estimator exhibits arbitrarily slow convergence in some region of $\mathscr{X}$. These insights extend to deeper trees. Because the partition is constructed recursively and $p > 1$, the problematic regions propagate throughout the tree and appear as many hyper-rectangles scattered across $\mathscr{X}$

with non-vanishing probability.

The core of the proof studies the tree construction as the maximizer of the split criteria in (4) and (5), indexed by the split location and the covariate coordinate. The analysis relies on 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]. The argument proceeds in four main steps.

Step 1: Split Criterion Approximation. Using empirical process techniques, we establish an asymptotic equivalence between the split criterion underlying each causal tree estimator and the split criterion of a standard regression tree employing CART. For $l = \texttt{DIM}$ and $l = \texttt{IPW}$, this corresponds to a regression tree applied to the transformed outcomes $y_i \frac{d_i - \xi}{\xi(1-\xi)}$. For $l = \texttt{SSE}$, the approximating process is the sum of two independent split criterion processes, one based on the transformed outcome $\frac{d_i}{\xi} y_i$ for treated units and one based on $\frac{1-d_i}{1-\xi} y_i$ for control units. A truncation argument removes extremely small or large split indices where empirical process approximations are less reliable [Csörgö and Horváth, 1997, Theorem A.4.1].

Step 2: Conditional Gaussian Approximation. Conditional on the ordering of the covariates, the square root of the split criterion process can be approximated by a Gaussian process with the same conditional covariance structure. For $l = \texttt{DIM}$ and $l = \texttt{IPW}$, the split criterion can be written as a sum of i.i.d. high-dimensional random vectors indexed by split location and coordinate. Applying the high-dimensional central limit theorem of [Chernozhukov et al., 2017, Theorem 2.1], we obtain a Gaussian approximation conditional on the ordering. Because of the structure of the split criterion, a high-dimensional central limit theory over hyper-rectangles suffices. For $l = \texttt{SSE}$, the treated and control components are stacked into a higher-dimensional vector and a central limit theory for convex sets [Chernozhukov et al., 2017, Proposition 3.1] is employed.

Step 3: Unconditional Gaussian Approximation. When $p > 1$, different covariate coordinates induce different orderings of the observations. We therefore show that the conditional Gaussian process from Step 2 is close to an unconditional Gaussian process in which splits across different coordinates are asymptotically uncorrelated. This implies asymptotic independence of the corresponding subprocesses and reduces the problem to studying the maximization of the split criterion along a single coordinate. The approximation is established using a Gaussian comparison inequality [Chernozhuokov et al., 2022, Proposition 2.1] together with bounds on the difference between the conditional and unconditional covariance matrices. For $l = \texttt{DIM}$ and $l = \texttt{IPW}$, the argument follows directly from a high-dimensional central limit theory for hyper-rectangles. For $l = \texttt{SSE}$, additional approximation error is controlled using Nazarov's inequality [Nazarov, 2003].

Step 4: Lower bound on imbalanced split probability. The unconditional Gaussian processes obtained in Step 3 correspond to the squared norm of a univariate ($l \in \{\texttt{DIM}, \texttt{IPW}\}$) or bivariate ($l = \texttt{SSE}$) Ornstein–Uhlenbeck process, with a one-to-one transformation between split index for the tree and time for the O-U process [Csörgö and Révész, 1981, Göing-Jaeschke and Yor, 2003]. The Darling–Erdös theorem [Eicker, 1979, Horváth, 1993] then characterizes the distribution of the maximum of this process over an interval. Combining this result with the Gaussian correlation inequality [Latała and Matlak, 2017, Remark 3(i)] yields the lower bound in (6), which in turn determines the effective sample sizes of the child nodes.

The remaining arguments apply these insights recursively to deeper trees, allowing us to characterize the concentration properties of the resulting CATE estimators.

5 X-Adaptivity and Inconsistency

The estimators considered in Theorem 1 either employ the full sample throughout their construction or rely on a two-sample independent split (honesty), where one subsample is used to construct the tree and the other is used to estimate the conditional average treatment effects. As discussed in Devroye et al. [2013] and references therein, **X**-adaptivity offers a middle ground between these two data-usage designs: the tree construction and the final estimation step share the same covariates but use different outcome variables, so that the two subsamples are independent conditional on the covariates.

We leverage the idea of **X**-adaptivity and study causal tree estimators in which the outcome variable and treatment indicator are independent across all stages of the tree construction and the final CATE estimation step, while the same covariates are used throughout. This **X**-adaptive design is of theoretical interest because it provides a bridge between no-sample-splitting and honesty. The following definition formalizes the construction of the corresponding **X**-adaptive causal tree estimators.

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

1. *The dataset* $\mathscr{D}$ *is divided into* $K+1$ *datasets* $(\mathscr{D}_{\mathsf{T}_1}, \ldots, \mathscr{D}_{\mathsf{T}_K}, \mathscr{D}_\tau)$*, with sample sizes given by* $(n_{\mathsf{T}_1}, \ldots, n_{\mathsf{T}_K}, n_\tau)$*, respectively, and satisfying* $n_{\mathsf{T}_1} = \cdots = n_{\mathsf{T}_K} = n_\tau$ *(possibly after randomly discarding at most* $n \mod K$ *observations). 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*, denoted* $\mathsf{T}_K^l(\mathscr{D}_{\mathsf{T}_1}, \cdots, \mathscr{D}_{\mathsf{T}_K})$*, is constructed by iterating* K *times the splitting procedure* $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$ *in Definition 2, splitting* all *terminal nodes until either (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*

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

As in the previous cases, if the distributions of $\varepsilon_i(0)$ and $\varepsilon_i(1)$ are symmetric about zero, then the $\mathbf{X}$-adaptive estimators are unbiased: $\mathbb{E}[\hat{\tau}_l^{\mathtt{X}}(\mathbf{x}; K)] = \tau$, for $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$. The next theorem shows that even under this intermediate sampling design, causal tree estimators remain uniformly inconsistent.

Theorem 3 (Accuracy of $\mathbf{X}$-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* $l \in \{\mathtt{DIM}, \mathtt{IPW}, \mathtt{SSE}\}$,

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

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* $K \geq 1$,

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

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 $\mathbf{X}$-adaptive causal tree estimator whenever $K_n \gtrsim \log\log n$. To put this rate condition in perspective, note that even for extremely large samples, the quantity $\log(\log(n))$ grows very slowly. For example, if $n/K_n \approx 1$ billion then $\log\log(10^9) \approx 3$. Thus, the inconsistency phenomenon appears even for very shallow trees, with only a few recursive splits.

At the same time, the theorem shows that the integrated mean squared error of a uniformly inconsistent $\mathbf{X}$-adaptive causal tree estimator can still decay at the optimal $\sqrt{n}$ rate, up to polylogarithmic factors. As in the previous results, the performance of causal tree estimators can therefore vary dramatically depending on whether the covariate value $\mathbf{x}$ corresponds to an average region of the covariate space or to a worst-case region.

6 Discussion

This section interprets the theoretical implications of Theorems 1–3 and related results for adaptive regression trees established in the supplemental appendix.

6.1 Decision Stumps

The generation of highly unbalanced cells in adaptive recursive partitioning has been recognized since the early development of CART and is often referred to as the *end-cut preference*. Informally, when the signal is weak relative to sampling noise, the empirical splitting criterion may be optimized by thresholds located near the boundary of the parent node.

For example, in the context of standard CART regression without sample splitting, Breiman et al. [1984, Theorem 11.1] and Ishwaran [2015, Theorem 4] showed that in one-dimension ($p = 1$), for each $\delta \in (0, 1)$, $\mathbb{P}(n(\mathtt{t_L}) \leq \delta n \text{ or } n(\mathtt{t_R}) \geq (1 - \delta)n) \to 1$ as $n \to \infty$. If applicable in our setting, their result would only imply uniform convergence rates 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 precisely characterize the regions of the support $\mathcal{X}$ where the pointwise rates of estimation are slower than any polynomial-in-n (see Theorems SA-2 and SA-20, as well as Corollaries SA-10 and SA-30 in the supplemental appendix). Thus, existing theoretical results are not strong enough to reveal the limitations of causal trees for pointwise estimation. Moreover, our analysis covers settings with sample splitting (honesty) and shows that this modification does not mitigate the slow convergence of adaptive causal trees under Assumption 1. Finally, our results apply to causal tree constructions that differ from, and are more complex than, standard CART regression trees.

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

Our theoretical results show that, under Assumption 1, the first split of an adaptive decision tree generates a small child cell with non-vanishing probability. Because recursive partitioning propagates this imbalance across subsequent levels of the tree, deeper trees necessarily produce multiple regions with very small effective sample sizes, again with non-vanishing probability. This phenomenon becomes more pronounced as the covariate dimension increases ($p > 1$), precisely the regime in which tree-based methods are often employed to detect treatment effect heterogeneity.

Importantly, these problematic regions need not occur near the boundary of $\mathcal{X}$. At each stage of the recursive construction, a parent node $\mathtt{t}$ produces one large and one small child node, but the split can occur anywhere within $\mathtt{t}$ and along any coordinate direction. As a consequence, adaptive tree constructions can generate collections of hyper-rectangular cells with extremely small sample sizes scattered throughout the interior of the covariate space $\mathcal{X} \subseteq \mathbb{R}^p$.

6.3 Regularization and Bias

A natural response to the small-cell phenomenon is to regularize the tree construction in order to prevent highly imbalanced splits. For instance, the algorithm may impose minimum node size constraints or incorporate penalties designed to discourage overfitting. Such modifications can reduce estimation variance but necessarily alter the class of admissible partitions.

Adaptive tree procedures generate small cells for two distinct reasons. When the conditional expectation function exhibits strong local curvature, finer partitions reduce approximation bias and constitute a key advantage of adaptive methods. At the same time, as shown in this paper, highly imbalanced cells also arise with non-vanishing probability even when the regression function is locally flat. In practice, these two sources of small cells cannot be distinguished reliably.

Our theoretical analysis isolates the variance mechanism by focusing on data generating processes with constant conditional expectation functions. In empirical applications, however, treatment effects may vary across the covariate space, so regularization rules designed to eliminate small cells may introduce substantial approximation bias and thereby affect overall convergence rates.

Even when small or imbalanced cells appear undesirable from a purely statistical point of view, they may still play a useful role in recursive partitioning. In particular, Ishwaran [2015] argues that imbalanced splits (e.g., the aforementioned end-cut preference) can be beneficial because they preserve a relatively large sample in one descendant node after a poor earlier split. To illustrate, consider the threshold CATE function $\tau(\mathbf{x}) = \mathbb{1}(x_1 \leq 1/2)$ and suppose the tree first makes a poor split on an irrelevant coordinate, say $x_2 = 0.05$. Although this split carries no information about the treatment effect, it is highly imbalanced. That is, one child contains only a small fraction of the sample, and the other contains the majority of the observations, including the entire signal region $\{x_1 \leq 1/2\}$. The algorithm may therefore continue splitting within this large node and still recover the relevant threshold in x_1. In contrast, if one imposes a minimum node-size requirement, then an early poor split on an irrelevant coordinate must divide the sample much more evenly, thereby shrinking the sample size on every downstream branch and leaving less room for later correction. In this sense, some imbalanced splits are not merely an incidental by-product of adaptation, but rather a key part of the mechanism by which the algorithm corrects itself locally. Accordingly, a blanket regularization that suppresses small cells may remove not only splits that are needed to reduce misspecification bias, but also splits that help deep trees recover from earlier mistakes.

6.4 α-Regularity and Causal Random Forests

Theorem 1 does not contradict existing (suboptimal) polynomial-in-n convergence rate results for honest causal trees and forests [Wager and Athey, 2018]. Rather, it shows that those results rely on strong algorithmic balance conditions that are incompatible with standard CART-type greedy splitting. In particular, prior analyses assume that each split allocates a fixed proportion of observations to both child nodes, that is, $n(\mathtt{t_L}) \geq \alpha\, n(\mathtt{t})$ and $n(\mathtt{t_R}) \geq \alpha\, n(\mathtt{t})$, where $\alpha \in (0, 1/2]$. This assumption is often called α-regularity, because it assumes that the tree construction generates an α-proportion of the data in each child node (cell). Although α-regularity is technically convenient and underlies positive consistency results for sufficiently deep trees and forest ensembles, it can force the resulting partition to be substantially less parsimonious than the underlying data generating mechanism would warrant, with corresponding losses in accuracy, adaptivity, and interpretability.

In particular, α-regularity may substantially alter the adaptive behavior of recursive partitioning. To illustrate, consider again the threshold CATE function

$$\tau(\mathbf{x}) = \mathbb{1}(x_1 \leq \eta), \qquad \eta \in (0, 1),$$

with $\mathbf{x} = (x_1, \ldots, x_p)^\top$ uniformly distributed on $[0, 1]^p$. At the population level, an unconstrained CART procedure can recover this function with a single split on the relevant variable x_1 at $x_1 = \eta$. In contrast, if $\eta < \alpha$, an α-regular tree cannot make that split at the root. It must instead approximate the threshold through a sequence of progressively finer splits, requiring at least

$$\left\lceil \log_{1/\alpha}(1/\eta) \right\rceil$$

successive splits along x_1. In addition to reducing parsimony, it also weakens variable-selection adaptivity. Rather than isolating the relevant subgroup immediately by one decisive split on x_1, the procedure is forced to spend multiple splits refining the same coordinate. In finite samples, especially when p is large relative to n, this repeated refinement creates more opportunities for spurious splits on irrelevant covariates.

More generally, α-regularity limits the ability of the tree to isolate small but substantively meaningful regions of the covariate space. While sufficiently deep balanced trees may still approximate such structures in an integrated sense, the resulting partitions need not recover the underlying subgroup itself. In applications where subgroup identification is a primary objective, ensemble averaging (e.g., random forests) does not resolve this limitation.

It is also important to note that α-regularity is primarily a theoretical device rather than a feature of standard applied implementations. In practice, regularization is typically governed by software defaults or light tuning, with minimum node-size choices fixed at small values (e.g. 2, 5, 20, or 30) rather than scaled with the sample size. Consequently, convergence guarantees derived under balance conditions apply to a more constrained algorithm than the CART-type procedures commonly used in empirical work. Extending such guarantees to canonical implementations would require additional regularization that modifies the estimator and introduces further bias-variance trade-offs.

6.5 Invalidity of Inference Methods

Theorem 1 also has direct implications for statistical inference based on adaptive causal trees. Because recursive partitioning generates highly imbalanced cells with non-vanishing probability, the effective sample size within some regions of the covariate space need not increase with the overall sample size. As a consequence, standard distributional approximations for ("honest") causal tree estimators may fail to hold even after appropriate centering and scaling. In particular, Gaussian asymptotic approximations can break down in regions where terminal nodes contain only a small number of observations.

This phenomenon undermines commonly used inference procedures based on asymptotic normality. For example, confidence intervals of the form

$$\hat{\tau}_l^q(\mathbf{x}) \pm z_\alpha \cdot \text{Sd.Err.}(\hat{\tau}_l^q(\mathbf{x})),$$

where z_α denoting the usual quantile of the standard Gaussian distribution, Sd.Err.() a standard error estimator, and $q \in \{\texttt{NSS}, \texttt{HON}, \texttt{X}\}$, need not provide asymptotically valid coverage for $\tau(\mathbf{x})$ over many regions of $\mathscr{X} \subseteq \mathbb{R}^p$.

6.6 Decision Tree Regression

The supplemental appendix establishes analogous lower bounds for standard CART regression trees used for nonparametric estimation of conditional expectation functions. In particular, Section SA-2 shows that adaptive regression trees can exhibit similarly slow uniform convergence, and may even be inconsistent depending on the sample-splitting design.

These findings complement earlier large-sample analyses of the decision stump without sample splitting by Bühlmann and Yu [2002] and Banerjee and McKeague [2007]. In the univariate setting ($p = 1$, $K = 1$), those studies showed that the empirical minimizers at the root node converge to well-defined population minimizers at cube-root rate under smoothness and identification conditions on the regression function.

Our analysis demonstrates that such conclusions are not uniformly valid over broad classes of regression functions. In particular, constant regression functions must be excluded from the admissible class for cube-root convergence results to hold uniformly over the support of the covariate. This distinction highlights the importance of uniform-in-function performance guarantees when evaluating adaptive recursive partitioning procedures.

7 Simulations

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

Two patterns emerge across all nine methods. First, for any fixed depth K, the pointwise RMSE is the smallest near the center of the covariate space and increases as $\mathbf{x}$ approaches the boundary. This pattern reflects the small-cell phenomenon predicted by (6): splits occurring near the boundary produce highly imbalanced nodes, reducing the effective sample size available for local averaging.

Second, for any fixed $\mathbf{x} \in [0, 1]^2$, the RMSE increases as the tree depth K increases. This behavior is consistent with the X-adaptive results in Theorem 1 and appears heuristically in the NSS and HON cases as well. As the depth increases, a larger fraction of evaluation points lie close to terminal node boundaries, where the same boundary effects that govern decision stumps lead to higher estimation variance. Consequently, the RMSE increases even for points located in the interior of the covariate space.

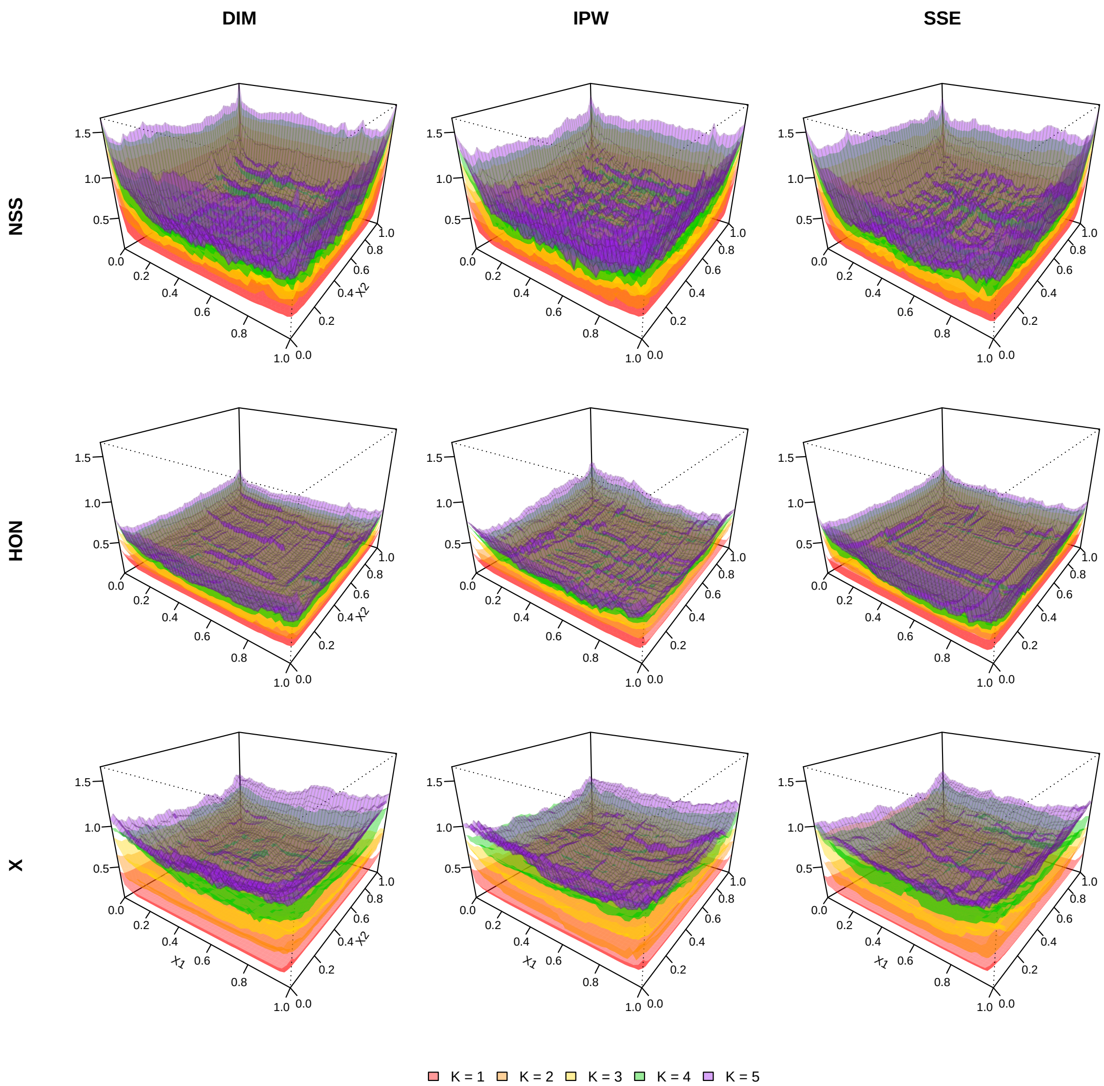

Figure 1: Root mean-squared error (RMSE) of heterogeneous treatment effect estimation using nine causal tree estimators with depth $K = 1, 2, \dots, 5$. The covariate is bivariate ($p = 2$) with support $[0, 1]^2$. Across all methods and depths, the smallest RMSE occurs near the center of the covariate space, while performance deteriorates as the evaluation point approaches the boundary. Results are based on $2,000$ Monte Carlo replications.

Appendix A Correction of Eicker [1979]

As part of the technical arguments used in this paper, we correct a statement concerning the limiting distribution of the maximum of an Ornstein-Uhlenbeck (O-U) process in Eicker [1979, Theorem 5]. Specifically, the term $\log(c)$ should appear in place of $2\log(c)$, where $c > 0$. This matters because replacing $\log(c)$ by $2\log(c)$ changes the factor $e^{-(z-\log(c))} = ce^{-z}$ to $e^{-(z-2\log(c))} = c^2e^{-z}$, so the limiting exponent depends quadratically rather than linearly on the window length. The problem is that this is incompatible with the Gaussian correlation inequality argument we used for adjacent windows. That is, for two adjacent windows of lengths $c_1 \log n$ and $c_2 \log n$, the event over the combined window is the intersection of the corresponding events on the two pieces, and the Gaussian correlation inequality implies that the probability for the combined window must be at least the product of the two individual window probabilities. With the correct term $\log(c)$, this is perfectly consistent, because the limiting probability for a window of length $(c_1 + c_2) \log n$ is

$$\exp(-(c_1 + c_2)e^{-z}) = \exp(-c_1 e^{-z}) \exp(-c_2 e^{-z}),$$

so the dependence on window length adds in exactly the way required by the inequality. If one instead used $2\log(c)$, then the same reasoning would force

$$\exp(-(c_1 + c_2)^2 e^{-z}) \geq \exp(-(c_1^2 + c_2^2)e^{-z}),$$

because the left-hand side would be the limiting probability for the combined window, while the right-hand side would come from the product lower bound for the two adjacent pieces. But this inequality is false, since $(c_1 + c_2)^2 > c_1^2 + c_2^2$ whenever $c_1, c_2 > 0$. Thus the extra factor 2 leads to a direct contradiction with the probability inequality implied by the Gaussian correlation argument. For completeness, we state below a corrected version of the result in a slightly more general form, allowing for the maximum of the norm of a possibly multivariate O–U process. Let $\Gamma(\cdot)$ denote Euler Gamma function.

Lemma 4 (Vector-Valued Markov-type Darling-Erdős)**.** *Let $\{(V_1(t), \ldots, V_d(t)) : 0 \leq t < \infty\}$ be d independent identically distributed Ornstein-Uhlenbeck processes with $\mathbb{E}[V_i(t)] = 0$ and $\mathbb{E}[V_i(t)V_i(s)] = \exp(-|t - s|/2)$, $1 \leq i \leq d$. Define*

$$N(t) = \Big(\sum_{1 \leq i \leq d} V_i^2(t) \Big)^{1/2}.$$

For any $c > 0$, $z \in \mathbb{R}$,

$$\lim_{n \to \infty} \mathbb{P}\Big(a(\log(n)) \sup_{0 \leq t \leq c \log(n)} N(t) - b_d(\log(n)) \leq z \Big) = \exp\Big(- e^{-(z - \log(c))} \Big),$$

where $a(t) = (2\log(t))^{1/2}$ and $b_d(t) = 2\log(t) + \frac{d}{2} \log\log(t) - \log \Gamma(d/2)$.

The proof of this result is given in the supplemental appendix (Section 4.1).

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.

Funding

Cattaneo gratefully acknowledges support from the National Science Foundation through grants SES-2019432, DMS-2210561, SES-2241575, and SES-2342226 and from the National Institute for Food and Agriculture (NIFA) through grant 2024-67023-42704. Klusowski gratefully acknowledges support from the National Science Foundation through NSF CAREER grant DMS-2239448 and from the Alfred P. Sloan Foundation through a Sloan Research Fellowship.

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.

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. *Annals of Statistics*, 2026.

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. *Annals of Statistics*, 50(6):3415–3438, 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, 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*, 2024.

Yan Shuo Tan, Jason M Klusowski, and Krishnakumar Balasubramanian. Statistical-computational trade-offs for recursive adaptive partitioning estimators. *Annals of Statistics*, 2026.

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.