# Accuracy Limits of Causal Trees for Individualized Treatment Effects

**Matias D. Cattaneo**[a], **Jason M. Klusowski**[b], **and Ruiqi (Rae) Yu**[b]

[a]Department of Economics, Princeton University, Princeton, NJ 08544.; [b]Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ 08544.

**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, with splitting criteria designed to identify variation in treatment effects across subgroups defined by covariates. We establish lower bounds on the uniform estimation error of causal trees that choose splits using treatment effect contrasts or squared error minimization, both with and without sample splitting. Even in a benchmark with randomized treatment assignment and a constant treatment effect, causal trees constructed via these splitting rules need not achieve any fixed polynomial rate of uniform convergence in the sample size. The underlying mechanism is that greedy recursive partitioning selects highly imbalanced splits with nonvanishing probability, producing terminal nodes containing very few observations and leading to large estimation variance. Sample splitting, often called "honesty," does not, however, remove this limitation. The same estimators can be accurate on average despite making large errors in regions of the covariate space with small probability mass. With a single covariate, maximally grown trees based on inverse probability weighting whose depth grows at a log–log rate in the sample size can be uniformly inconsistent under sample splitting yet have vanishing average error. Our results also clarify the role of balanced partition assumptions in existing theoretical guarantees for causal forests and related ensemble methods.**

recursive partitioning | decision trees | causal inference | heterogeneous treatment effects

### Significance Statement

Causal decision trees are widely used to estimate how treatment effects vary across individuals, with applications from precision medicine to recommendation systems. We show, however, that common versions of these methods based on greedy recursive partitioning, including causal trees based on sample splitting, can have intrinsic accuracy limits. Their worst-case errors, in particular, need not converge to zero at any fixed polynomial rate in the sample size. The problem arises when terminal nodes (the subgroups used for estimation) contain very few observations. Preventing these small nodes requires user-chosen restrictions, which change the estimator and introduce additional bias-variance tradeoffs. Thus, good average accuracy across a population need not imply reliable effect estimates or inference at individual covariate values.

M.D.C., J.M.K., and R.Y. designed research, performed research, contributed new analytic tools, analyzed results, and wrote the paper.

The authors declare no competing interests.

Corresponding authors are Matias D. Cattaneo (cattaneo@princeton.edu), Jason M. Klusowski (jason.klusowski@princeton.edu), and Ruiqi (Rae) Yu (rae.yu@princeton.edu).

## 1. Introduction

Recursive decision trees have become a popular tool for estimating heterogeneous causal effects in both experimental and observational settings. These methods, more specifically, adapt the classical CART (Classification and Regression Trees) algorithm [1] to causal inference by modifying the splitting criterion and, in some implementations, using sample splitting to separate tree construction from treatment effect estimation. Their simplicity and interpretability have made tree and recursive partitioning methods common in academic research and applied practice.

Tree methods approach treatment effect heterogeneity through subgroup identification, treatment effect contrasts, and local model fitting. Methods for subgroup identification include, for example, interaction trees [2, 3], differential effect search [4], Virtual Twins [5], and qualitative interaction trees [6]. Regression-based approaches include regression trunks [7], marginal trees [8], GUIDE regression trees [9], and model-based recursive partitioning [10]. Honest causal trees separate partition selection from treatment effect estimation through sample splitting [11]. Tree methods also accommodate instrumental variable settings [12, 13], censored survival outcomes [14], and observational studies [15, 16]. In randomized trials, conditional inference trees have also been combined with prognostic score matching to assess treatment effect heterogeneity [17].

Despite their widespread use across these different causal settings, the fundamental statistical properties of causal tree estimators and their associated inference procedures remain poorly understood for canonical CART-type implementations. Many existing analyses obtain positive guarantees, in particular, by imposing balance or regularity conditions on tree construction. We instead examine what can happen without these restrictions by establishing lower bounds on the uniform estimation error of CART-type causal trees.

Under basic conditions, causal trees constructed using CART-type greedy splitting rules need not achieve any fixed polynomial rate of uniform convergence in the sample size $n$, even though their integrated error can remain small. Thus, good average performance can coexist with large estimation errors for individualized or subgroup-level decisions in regions of the covariate space with small probability mass. This limitation, in fact, arises even in the simplest setting in which the true treatment effect is constant over the covariate space.

These greedy splitting rules, more specifically, select highly imbalanced splits with nonvanishing probability, 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 lead to the lower bounds established below.

One way to address this small cell phenomenon is to regular-

ize tree construction through minimum node size requirements. In common implementations such as `causalTree` and `htetree`, the default minimum node sizes are fixed and do not change with the sample size $n$ (Supplemental Appendix, Section SA-1.4). Larger minimum node sizes may reduce finite-sample variance, but they also restrict the partitions the procedure can select, changing the adaptivity that makes CART attractive, and they can introduce approximation bias when the conditional average treatment effect (CATE) is not locally constant. Conditional heteroskedasticity can further complicate this choice by making it harder to distinguish local treatment effect variation from sampling noise. Imbalanced splits may also have algorithmic benefits for deeper trees because they preserve sample size after a bad split and allow the tree to recover downstream [18].

The statistical effects of these restrictions, however, remain only partially understood for CART-type estimators with adaptively selected partitions. A satisfactory theoretical resolution of the small cell phenomenon therefore requires characterizing the resulting bias-variance tradeoff and establishing valid inference for the regularized estimator.

Another frequently discussed modification of adaptive recursive partitioning methods is sample splitting, also known as *honesty*. Sample splitting constructs the tree on one subsample and estimates treatment effects on an independent subsample. This separation of partition selection from within-node estimation has in turn been used to reduce overfitting and facilitate inference [11]. However, a region selected using the construction sample may still have small probability mass and contain few observations in the independent estimation sample. Our results show that honesty does not remove the small cell mechanism behind the lower bounds. Even under honest sample splitting, causal tree estimators need not achieve any fixed polynomial rate of uniform convergence in the sample size $n$ under basic conditions.

Our findings also have implications for theoretical guarantees established for quantile regression forests [19], honest causal forests [20], generalized random forests [21], and random forest ensembles for Boolean interaction recovery [22]. These analyses impose balance conditions on the constituent trees, although their targets and guarantees differ. In a common formulation, $\alpha$-regularity requires each split to place at least an $\alpha$ fraction of the observations in the parent node into each child node. These assumptions facilitate theoretical analysis by restricting split balance, whereas canonical CART-type greedy rules can select highly imbalanced splits with nonvanishing probability. Guarantees based on approximately uniform partitions or $\alpha$-regularity therefore concern a more constrained class of tree construction procedures. See Section 5 for further discussion.

The small cells behind our lower bounds also create a challenge for subgroup-level inference based on causal trees. 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. Distributional approximations centered at the true conditional average treatment effect in turn require controlling small-cell effects and approximation bias. For CART-type causal trees and forests, this is challenging without approximately uniform partition conditions. For forests, inference also depends on how trees are aggregated across subsamples (Section 5).

**Contributions and Related Literature.** This paper makes three main contributions. First, we establish lower bounds on uniform estimation error and upper bounds on integrated error for causal tree estimators constructed via CART-type recursive partitioning. The lower bounds show, in particular, that such estimators need not achieve any fixed polynomial rate of uniform convergence in the sample size $n$ under basic conditions, even in settings where the underlying treatment effect is constant. Second, we show that the lower bounds on uniform error persist under honest sample splitting, while the bounds without sample splitting contain an additional slowly varying $\sqrt{\log\log n}$ factor. For a scalar covariate, maximally grown honest trees based on inverse probability weighting (IPW) can, moreover, be uniformly inconsistent at depth of order $\log\log n$ and remain consistent in integrated $L^q$ loss, as shown in Remark 1. Third, we clarify the implications of these findings for causal inference using trees more broadly, including commonly used regularization strategies and theoretical guarantees for causal forests that rely on approximately uniform partitions or $\alpha$-regularity conditions.

Our analysis thereby complements theoretical work on random forests [23, 24], recursive partitioning estimators [25, 26], and oblique regression trees [27], and contributes to a growing body of formal limitations for methods based on trees. For example, risk lower bounds show that honest, axis-aligned trees that estimate leaf values by averaging can fail to attain the minimax rate in additive regression models, even when the partition is optimized over an admissible partition class [28]. On the computational side, recent work [29] on a modification of the Bayesian Additive Regression Trees (BART) sampler [30] shows that with discrete covariates and initialization at trivial trees, the expected time to reach certain regions of high posterior density can increase with the training sample size. Moreover, greedily trained regression trees with Boolean covariates can require exponentially many samples in the dimension for certain sparse regression functions and can therefore be inconsistent in high-dimensional settings [31].

Our Supplemental Appendix also develops companion lower bounds for standard CART regression trees. These results thereby clarify the relationship between our lower bound framework and earlier large-sample analyses of CART decision stumps (trees with a single split) [32, 33].

The proofs rely on new probabilistic arguments concerning the behavior of adaptive recursive partitioning procedures. The Supplemental Appendix develops, to this end, approximations for suprema of partial sums and related Gaussian processes using strong approximations [34, 35], Brownian motion and stochastic calculus [36], Gaussian correlation inequalities [37], Ornstein–Uhlenbeck extreme-value limits [38], excursion estimates over fixed intervals [39], Markov chain comparisons and Doob transforms [40], elliptic regularity [41], and Dirichlet spectral characterizations [42]. As a technical byproduct of our analysis, we also identify and correct a normalization error in Eicker [43, Theorem 5].

**Organization.** The remainder of the paper is organized as follows. Section 2 first introduces the causal tree estimators and the data generating framework studied in the paper. Section 3 then presents the assumptions underlying our theoretical analysis. Section 4 establishes lower bounds on the uniform

estimation error of causal tree estimators and provides complementary results on their integrated $L^q$ error. Section 5 discusses the implications of our results for recursive partitioning methods, causal forests, and inference procedures. Section 6 concludes. Additional theoretical results, an overview of the proof strategy, and all technical proofs are reported separately in the Supplemental Appendix. The Supplemental Appendix also presents simulations and an empirical resampling illustration showing large local estimation errors for trees with and without sample splitting.

**Notation.** We write $\mathbb{P}$, $\mathbb{E}$, and $\mathbb{V}$ for probability, expectation, and variance, and $\rightsquigarrow$ for convergence in distribution. For positive sequences, $a_n = o(b_n)$ means $a_n/b_n \to 0$, while $a_n = O(b_n)$ means that $a_n/b_n$ remains bounded as $n \to \infty$. Unless indicated otherwise, log is the natural logarithm. For random sequences, $a_n = O_{\mathbb{P}}(b_n)$ means that $|a_n/b_n|$ is bounded in probability.

## 2. Setup

The causal tree estimators studied here combine (i) an estimator of the conditional average treatment effect (CATE) computed within each node, (ii) a recursive partitioning rule used to construct the tree, and (iii) a data usage scheme determining whether sample splitting is employed.

We observe a random sample $\mathcal{D} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n\}$, where $y_i$ is an outcome variable, $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,p})^\top$ is a vector of pretreatment covariates, and $d_i$ is a binary treatment indicator. Employing standard potential outcomes notation [see, e.g., 44], 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. We consider, in particular, a simple randomized experiment in which 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, in this setting, is the 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 pretreatment covariates. Under this design, the CATE function is identified by

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

where the constant treatment probability $\xi = \mathbb{P}(d_i = 1) \in (0, 1)$ is known from the randomization mechanism. The first equality represents the CATE as a single conditional expectation, $\mathbb{E}[\tilde{y}_i \mid \mathbf{x}_i = \mathbf{x}]$, where

$$\tilde{y}_i = y_i \frac{d_i - \xi}{\xi(1-\xi)}$$

is the transformed outcome. The second equality, in turn, expresses the CATE as the difference of two conditional expectation functions based on observed data.

The conditional expectations above can be estimated using nonparametric regression methods. 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. Recursive partitioning, in this context, is a nonparametric approach that adapts the partition to the data, constructing subgroups defined by covariates and estimating treatment effects within the resulting terminal nodes.

We specifically study six causal tree estimators obtained by combining three tree construction rules with two data usage schemes. The three rules are variance maximization using IPW estimates, variance maximization using difference in means (DIM) estimates, and sum of squared errors (SSE) minimization. The two data usage schemes, in turn, are no sample splitting and honest sample splitting. These choices together represent the transformed outcome, treatment contrast, local model fit, and sample splitting designs used in causal tree implementations and related recursive partitioning procedures.

**CATE Estimator.** Although we study three tree construction rules, only two estimators are needed to estimate effects within their terminal nodes: IPW trees use the IPW estimator, whereas both DIM and SSE trees use the DIM estimator. We accordingly define these estimators for a tree $\mathsf{T}$ and an estimation sample $\mathcal{D}_\tau$, with the next two subsections specifying how the tree and sample are chosen. Let $\mathbb{1}(\cdot)$ be the indicator function.

**Definition 1** (CATE Estimators)**.** *Suppose $\mathsf{T}$ is a tree, and $\mathcal{D}_\tau = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \dots, n_\tau\}$, with $n_\tau \le n$, is the dataset used to estimate treatment effects within the terminal nodes. All sums and counts in this definition accordingly use $\mathcal{D}_\tau$. Let $\mathtt{t}$ be the unique terminal node (leaf) in $\mathsf{T}$ containing $\mathbf{x}$ in the covariate support $\mathcal{X}$.*

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

$$\hat{\tau}_{\mathtt{IPW}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau) = \frac{1}{n(\mathtt{t})} \sum_{i:\mathbf{x}_i \in \mathtt{t}} \tilde{y}_i,$$

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

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

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

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

The terminal nodes of $\mathsf{T}$ partition the covariate support $\mathcal{X}$, and both estimators use only observations in the node containing $\mathbf{x}$. For a given tree, both are therefore nonparametric partitioning-based estimators of $\tau(\mathbf{x})$. See [45–47] and references therein.

Since the estimators $\hat{\tau}_{\mathtt{IPW}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau)$ and $\hat{\tau}_{\mathtt{DIM}}(\mathbf{x}; \mathsf{T}, \mathcal{D}_\tau)$ are constant within each terminal node of $\mathsf{T}$, we define

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

for all terminal nodes $\mathtt{t}$ of $\mathsf{T}$.

**Tree Construction.** An axis-aligned recursive decision tree partitions the covariate space through a sequence of splits on individual covariates. At each node, the algorithm uses the observed outcomes and covariates to choose a covariate and threshold according to a criterion such as squared error, Gini impurity, or entropy. It then repeats this process until a stopping condition is met, such as maximum depth or pure terminal nodes. See [48, 49] and references therein for further background.

The most popular implementation of recursive decision trees is the CART algorithm, which proceeds in a top-down, greedy manner through recursive binary splitting: at each step, it chooses the best split at the current node without looking ahead to future splits. Given a construction sample $\mathcal{D}_{\mathsf{T}} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n_{\mathsf{T}}\}$ with $n_{\mathsf{T}} \leq n$ (used for all sums, counts, and fitted child means throughout this subsection), regression CART divides a parent node $\mathtt{t}$ in the tree (i.e., a region in $\mathcal{X}$) into two child nodes, $\mathtt{t}_{\mathtt{L}}$ and $\mathtt{t}_{\mathtt{R}}$, by minimizing the sum of squared errors (SSE),

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

over candidate splits leaving at least one observation in each child node (using the threshold and tie-breaking conventions in Definition 2; a node with no valid split is left terminal). The solution $(\hat{\beta}_{\mathtt{L}}, \hat{\beta}_{\mathtt{R}}, \hat{\varsigma}, \hat{\jmath})$ thus gives the two fitted child values, the split point, and the split coordinate, respectively. Because each split uses a single covariate, the resulting cells are axis-aligned hyperrectangles intersected with $\mathcal{X}$. The selected split, more explicitly, divides $\mathtt{t}$ into the child nodes $\mathtt{t}_{\mathtt{L}} = \{\mathbf{x} \in \mathtt{t} : \mathbf{e}_{\hat{\jmath}}^\top \mathbf{x} \leq \hat{\varsigma}\}$ and $\mathtt{t}_{\mathtt{R}} = \{\mathbf{x} \in \mathtt{t} : \mathbf{e}_{\hat{\jmath}}^\top \mathbf{x} > \hat{\varsigma}\}$. Here $\mathbf{e}_j$ is the $j$th standard basis vector.

The normal equations imply that $\hat{\beta}_{\mathtt{L}} = \frac{1}{n(\mathtt{t}_{\mathtt{L}})} \sum_{\mathbf{x}_i\in\mathtt{t}_{\mathtt{L}}} y_i$ and $\hat{\beta}_{\mathtt{R}} = \frac{1}{n(\mathtt{t}_{\mathtt{R}})} \sum_{\mathbf{x}_i\in\mathtt{t}_{\mathtt{R}}} y_i$, the respective sample means after splitting the parent node at $\mathbf{e}_{\hat{\jmath}}^\top \mathbf{x} = \hat{\varsigma}$. These child nodes may, in turn, become parent nodes at the next level of the tree construction, and the procedure continues recursively up to a maximum depth $K$, with the root at depth zero and each child one level deeper than its parent.

Treatment effect trees adapt the CART algorithm by replacing outcome prediction error with a splitting criterion that targets treatment effect heterogeneity or local model fit. Existing procedures, for example, split on treatment-by-covariate interactions or subgroup contrasts [2–6], use criteria based on estimated child-node treatment effects across experimental, survival, instrumental variable, or observational settings [11–16], or use regression models or hypothesis tests to select splits [7–10, 17]. We focus here on three representative CART-type criteria: IPW and DIM use variance maximization with different estimators of the child-node effects, whereas SSE minimizes the residual sum of squares from local treatment models. Definition 2 formalizes these two types of splitting rule.

**Definition 2** (Tree Construction)**.** *Suppose* $\mathcal{D}_{\mathsf{T}} = \{(y_i, d_i, \mathbf{x}_i^\top) : i = 1, 2, \ldots, n_{\mathsf{T}}\}$*, with* $n_{\mathsf{T}} \leq n$*, is the dataset used to construct the tree* $\mathsf{T}$*. There is a unique node* $\mathtt{t}_0 = \mathcal{X}$ *at initialization, and child nodes are generated by iterative axis-aligned splitting of the parent node based on either of the following two rules. Each split divides a terminal node* $\mathtt{t}$ *of the current tree* $\mathsf{T}'$ *into two child nodes,* $\mathtt{t}_{\mathtt{L}}$ *and* $\mathtt{t}_{\mathtt{R}}$*, to form the new tree* $\mathsf{T}$*.*

*For IPW, each child must contain at least one construction observation. For DIM and SSE, each child must contain at least one treated and one control construction observation. A candidate split is valid precisely when these requirements hold. The optimization therefore excludes invalid splits, and a node with no valid split is terminal.*

*For each candidate split, the threshold is the largest value of the split coordinate among construction observations in the left child. Among tied optimal splits, choose the smallest coordinate index and then the smallest cut rank within that coordinate.*

- *Variance Maximization. Choose the valid split that maximizes*

$$\frac{n(\mathtt{t}_{\mathtt{L}})n(\mathtt{t}_{\mathtt{R}})}{n(\mathtt{t})}\left(\hat{\tau}_l(\mathtt{t}_{\mathtt{L}}; \mathsf{T}, \mathcal{D}_{\mathsf{T}}) - \hat{\tau}_l(\mathtt{t}_{\mathtt{R}}; \mathsf{T}, \mathcal{D}_{\mathsf{T}})\right)^2,$$

  *where* $l \in \{\mathtt{IPW}, \mathtt{DIM}\}$*. The resulting causal trees are denoted by* $\mathsf{T}_{\mathtt{IPW}}(\mathcal{D}_{\mathsf{T}})$ *and* $\mathsf{T}_{\mathtt{DIM}}(\mathcal{D}_{\mathsf{T}})$*, respectively.*

- *SSE Minimization. For each valid split, compute*

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

  *using only the data* $\mathcal{D}_{\mathsf{T}}$*. Choose the coordinate and threshold that minimize this profiled objective. The resulting causal tree is denoted by* $\mathsf{T}_{\mathtt{SSE}}(\mathcal{D}_{\mathsf{T}})$*.*

The variance maximization criterion thus uses treatment effect contrasts in place of the outcome prediction errors in the original CART criterion (1). It selects, more specifically, the split with the largest weighted squared difference between estimated treatment effects in the two child nodes; related criteria in the literature incorporate additional weights or variance adjustments [11, 14–16].

For the `IPW` estimator, this rule is exactly equivalent to applying the CART criterion in (1) to the transformed outcome $\tilde{y}_i$, whose conditional mean is the CATE, $\mathbb{E}[\tilde{y}_i \mid \mathbf{x}_i = \mathbf{x}] = \tau(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$. The `DIM` estimator likewise uses the same splitting criterion, estimating the treatment effect in each child as the difference between the treated and control means.

SSE minimization instead selects splits by fitting local treatment models. For each candidate split, it fits separate regressions of the outcome on an intercept and a treatment indicator in the two child nodes, then selects the split with the smallest combined residual sum of squares. Supplemental Appendix, Section SA-3.3, derives the representation of the SSE objective as a sum of treated and control variance gains.

The three construction rules thus generate data-driven partitions that need not be approximately uniform (often called quasi-uniform in the partitioning literature). Existing guarantees that require quasi-uniformity [e.g., 46, Assumption 2] or related regularity conditions [e.g., 47, Assumption A] therefore do not directly apply. The Supplemental Appendix also considers other recursive partitioning constructions, including the standard CART algorithm and variants thereof.

**Sample Splitting.** Having specified the splitting criteria, we now distinguish the observations used to construct the partition from those used to estimate treatment effects within leaves. Sample splitting, often called "honesty," uses separate

observations for tree construction and effect estimation. This separation has been used to reduce overfitting and facilitate inference [11]. We then compare it with no sample splitting, which uses the same observations for both tasks, as formalized in Definition 3.

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

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

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

- *Honesty (HON). The dataset* $\mathcal{D}$*, by contrast, is divided into two independent datasets* $\mathcal{D}_{\mathsf{T}}$ *and* $\mathcal{D}_\tau$ *with positive sample sizes* $n_{\mathsf{T}}$ *and* $n_\tau$*, respectively, satisfying* $n_{\mathsf{T}} + n_\tau = n$*. The causal tree estimators are correspondingly*

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

Separate samples are also used for tree construction and estimation of leaf effects in conditional inference treatment effect trees [17]. Both data usage schemes are common in the literature, so our results inform theoretical, methodological, and empirical work that uses them.

For either scheme, the estimators $\hat{\tau}_l(\mathbf{x})$ and $\check{\tau}_l(\mathbf{x})$, $l \in \{\texttt{IPW}, \texttt{DIM}, \texttt{SSE}\}$, depend on the maximum tree depth $K$ (which may depend on $n$). We suppress $K$ in the estimator notation for simplicity, while keeping the dependence on $K$ explicit in the theoretical results below.

## 3. Assumptions

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

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

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

(ii) $y_i(d) = \mu_d(\mathbf{x}_i) + \varepsilon_i(d)$*, with* $\mathbb{E}[\varepsilon_i(d)] = 0$ *and* $\mathbf{x}_i \perp\!\!\!\perp (\varepsilon_i(0), \varepsilon_i(1))$*.*

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

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

(v) *There exists* $\chi > 0$ *such that* $\mathbb{E}[\exp(\lambda \varepsilon_i(d))] < \infty$ *for all* $|\lambda| < 1/\chi$ *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), in turn, specifies a canonical regression representation for the potential outcomes, where $\mu_d(\mathbf{x})$ denotes the conditional mean function and the potential outcome error vector is jointly independent of $\mathbf{x}_i$. Arbitrary dependence between $\varepsilon_i(0)$ and $\varepsilon_i(1)$, however, is allowed unless stated otherwise. Assumption 1(iii) requires the conditional mean of each potential outcome to be constant, implying a constant CATE, $\tau(\mathbf{x}) \equiv \tau^\star = \mu_1^\star - \mu_0^\star$, across the covariate space. Below, $\tau$ also denotes this constant.

Assumption 1(iv) requires the covariates to be independent and continuously distributed. Applying the marginal cumulative distribution functions then yields independent uniform coordinates on $[0, 1]^p$ and preserves sample ranks almost surely. Because the candidate splits and tie-breaking rule in Definition 2 depend on the covariates only through their coordinatewise orderings, we may work without loss of generality on $\mathcal{X} = [0, 1]^p$, using the same split inequalities on the boundary of $[0, 1]^p$ without changing the uniform error. Throughout the asymptotic results, $p$ is fixed.

Assumption 1(v), finally, requires the potential outcome errors to be subexponential, or equivalently, to satisfy a Bernstein moment condition.

To establish lower bounds for the causal tree estimators in Definition 3, it suffices to consider the 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 $\mathcal{P}$ containing the distribution $\mathbb{P}_1$ satisfying Assumption 1,

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

for all $\epsilon > 0$, where $\tau(\mathbf{x})$ and $\mathcal{X}$ denote the CATE function and covariate support under the corresponding distribution.

Assumption 1 also removes issues related to smoothing (or misspecification) bias, heteroskedasticity across covariate values, 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 honest estimators, any bias under this model comes from the zero prediction used when a terminal node lacks the observations needed to estimate an effect. For a fixed $\mathbf{x} \in \mathcal{X}$, let $\mathsf{t}_l = \mathsf{t}_l(\mathbf{x})$ denote the terminal node containing $\mathbf{x}$ in the corresponding honest tree, with counts computed from $\mathcal{D}_\tau$. Lemma SA-34 (Supplemental Appendix, Section SA-3.5) explicitly gives

$$\begin{aligned}\mathbb{E}[\check{\tau}_{\texttt{IPW}}(\mathbf{x})] &= \tau(\mathbf{x})\big\{1 - \mathbb{P}\big(n(\mathsf{t}_{\texttt{IPW}}) = 0\big)\big\},\\ \mathbb{E}[\check{\tau}_l(\mathbf{x})] &= \tau(\mathbf{x})\big\{1 - \mathbb{P}\big(n_0(\mathsf{t}_l) = 0 \text{ or } n_1(\mathsf{t}_l) = 0\big)\big\},\end{aligned}$$

where the second equality holds for $l \in \{\texttt{DIM}, \texttt{SSE}\}$ and includes nodes missing either treatment arm.

For NSS trees grown using Definition 2, the same lemma gives the following identities for any $\tau$ under its conditions on partition invariance and symmetry:

$$\begin{aligned}&\mathbb{E}[\hat{\tau}_{\texttt{IPW}}(\mathbf{x})] = \tau(\mathbf{x}),\\ &\mathbb{E}[\hat{\tau}_l(\mathbf{x})] = \tau(\mathbf{x})\big\{1 - \xi^n - (1 - \xi)^n\big\}, \qquad l \in \{\texttt{DIM}, \texttt{SSE}\}.\end{aligned}$$

The symmetry conditions, more specifically, concern the marginal potential outcome errors for DIM and SSE, and the transformed residuals $\tilde{y}_i - \tau$ for IPW. With covariates and treatment assignments held fixed, the final partition, including stopping, pruning, and tie breaking, must be unchanged when all residual signs are reversed. For $n \geq 2$, DIM and SSE are exactly unbiased conditional on both treatment arms appearing in the sample.

Under the growth conditions of the theorems below, the lower bounds hold even when $\mu_0 = \mu_1 = 0$ and the preceding conditions on partition invariance and symmetry hold, in which case all six estimators are exactly unbiased.

Finally, beyond its role in the lower bounds, 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 tree-adapted Haar basis representation. This perspective suggests that analogous lower bounds may hold for smooth treatment effect functions in shrinking neighborhoods of a constant function when the signal is sufficiently weak relative to the noise.

## 4. Main Results

The following theorem establishes our main lower bound on the uniform estimation error of causal tree estimators (Supplemental Appendix, Sections SA-1.2 and SA-3.1–SA-3.3).

**Theorem 1** (Uniform Accuracy)**.** *Suppose Assumption 1 holds. Assume that the root is split according to the prescribed criterion whenever a valid split exists and that the resulting tree has depth at most a deterministic integer $K \geq 1$, possibly depending on $n$. Fix $b \in (0,1)$. Then the following conclusions hold for each $l \in \{\texttt{IPW}, \texttt{DIM}, \texttt{SSE}\}$.*

- *No Sample Splitting (NSS). Also assume that $2^K \log n = o\big(n^{b/4}\sqrt{\log\log n}\big)$ if $l \in \{\texttt{DIM}, \texttt{SSE}\}$. Then, there exist positive constants $C_1$ and $C_2$ such that*

$$\liminf_{n\to\infty} \mathbb{P}\left( \sup_{\mathbf{x}\in\mathcal{X}} \left|\hat{\tau}_l(\mathbf{x}) - \tau(\mathbf{x})\right| \geq C_1 \frac{\sqrt{\log\log n}}{n^{b/2}} \right) \geq C_2 b,$$

  *where $C_1$ and $C_2$ only depend on $p$ and the distribution of $(y_i(0), y_i(1), d_i)$.*

- *Honesty (HON). Also assume that $\rho \leq n_{\mathsf{T}}/n_\tau \leq \rho^{-1}$ for some $\rho \in (0,1)$. Then, there exist positive constants $C_3$ and $C_4$ such that*

$$\liminf_{n\to\infty} \mathbb{P}\left( \sup_{\mathbf{x}\in\mathcal{X}} \left|\check{\tau}_l(\mathbf{x}) - \tau(\mathbf{x})\right| \geq C_3 \frac{1}{n^{b/2}} \right) \geq C_4 b,$$

  *where $C_3$ and $C_4$ only depend on $\rho$, $p$, and the distribution of $(y_i(0), y_i(1), d_i)$.*

The constants in Theorem 1 do not depend on $K$ or $n$. For NSS-DIM and NSS-SSE, any depth sequence satisfying $K = o(\log n)$, including fixed depth, meets the depth condition for every fixed $b \in (0,1)$. NSS-IPW and HON procedures, by contrast, do not require this depth restriction.

The theorem uses a deliberately favorable statistical benchmark in which treatment assignment is randomized with known probability, both potential outcomes have constant conditional means, and their errors are jointly independent of the covariates and have light tails. The trees use the canonical CART-type greedy splitting rules in Definition 2, without additional minimum terminal node size or balance restrictions. Thus, the lower bounds are not driven by treatment effect complexity, confounding, or smoothing bias.

For each fixed $b$ satisfying the conditions of Theorem 1, the uniform estimation error exceeds a threshold of order $n^{-b/2}\sqrt{\log\log n}$ without sample splitting and of order $n^{-b/2}$ under honesty, with probability bounded away from zero. Taking $b$ arbitrarily small along a single sequence of trees therefore rules out any polynomial rate of uniform convergence in $n$ for all three honest estimators, for NSS-IPW, and, when $K = o(\log n)$, for NSS-DIM and NSS-SSE. These lower bounds, however, do not by themselves determine the attainable uniform convergence rates or imply that the uniform error fails to converge to zero.

Although Theorem 1 rules out polynomial uniform convergence rates under its stated conditions, the same estimators can achieve favorable accuracy when error is averaged over $\mathcal{X}$. The next theorem establishes bounds for this integrated error (Supplemental Appendix, Sections SA-3.1–SA-3.3). Let $F_X(\mathbf{x}) = \mathbb{P}(\mathbf{x}_i \leq \mathbf{x})$. For $1 \leq q < \infty$, write $\|g\|_q = \left(\int_{\mathcal{X}} |g(\mathbf{x})|^q \, \mathrm{d}F_X(\mathbf{x})\right)^{1/q}$.

**Theorem 2** (Integrated $L^q$ Accuracy)**.** *Suppose Assumption 1 holds, $n \geq 2$, and the underlying causal tree has depth at most a deterministic integer $K \geq 1$, possibly depending on $n$. Fix $1 \leq q < \infty$. Then the following conclusions hold for each $l \in \{\texttt{IPW}, \texttt{DIM}, \texttt{SSE}\}$.*

- *No Sample Splitting (NSS). There exists a positive constant $C_1$ such that*

$$\left(\mathbb{E}\left[\|\hat{\tau}_l - \tau\|_q^q\right]\right)^{\frac{1}{q}} \leq C_1 \left(\frac{2^K \log^{\max\{3,q+1\}}(n)}{n}\right)^{\frac{1}{\max\{2,q\}}},$$

  *where $C_1$ only depends on $q$, $p$, and the distribution of $(y_i(0), y_i(1), d_i)$.*

- *Honesty (HON). There exists a positive constant $C_2$ such that, for any positive sample sizes $n_{\mathsf{T}}$ and $n_\tau$,*

$$\left(\mathbb{E}\left[\|\check{\tau}_l - \tau\|_q^q\right]\right)^{\frac{1}{q}} \leq C_2 \left(\frac{2^K}{n_\tau + 1}\right)^{\frac{1}{\max\{2,q\}}},$$

  *where $C_2$ only depends on $q$ and the distribution of $(y_i(0), y_i(1), d_i)$.*

The constants in Theorem 2 likewise do not depend on $K$ or the sample sizes. For each fixed $q$, the bounds imply integrated $L^q$ consistency along any sequence $K = K_n$ satisfying $2^{K_n}\log^{\max\{3,q+1\}}(n)/n \to 0$ for no sample splitting and $2^{K_n}/(n_\tau + 1) \to 0$ for honest sample splitting. By Markov's inequality, these moment bounds also yield the corresponding convergence rates in probability for $\|\hat{\tau}_l - \tau\|_q$ and $\|\check{\tau}_l - \tau\|_q$ (Supplemental Appendix, Section SA-1.3).

As tree depth grows, the contrast between uniform and integrated accuracy can become more pronounced. The next remark, in particular, shows that integrated consistency can coexist with uniform inconsistency.

*Remark* 1 (Uniform Inconsistency at Log–Log Depth). Suppose Assumption 1 holds with $p = 1$. Consider a maximally

grown honest IPW tree with $\rho \le n_{\mathsf{T}}/n_\tau \le \rho^{-1}$ for some $\rho \in (0,1)$, and depth cap

$$K_{n_{\mathsf{T}}} = \lceil C \log\log n_{\mathsf{T}} \rceil.$$

Here maximal growth means splitting every node with at least two construction observations up to the depth cap, even when the gain is zero, without pruning. For sufficiently large fixed $C$, Corollary SA-15 in the Supplemental Appendix gives

$$\liminf_{n_{\mathsf{T}}\to\infty} \mathbb{P}\left( \sup_{\mathbf{x}\in\mathcal{X}} \left| \check{\tau}_{\mathtt{IPW}}(\mathbf{x}) - \tau(\mathbf{x}) \right| > \epsilon \right) \ge c,$$

for some $\epsilon, c > 0$. Since

$$\frac{2^{K_{n_{\mathsf{T}}}}}{n_\tau + 1} = O\left( \frac{(\log n_{\mathsf{T}})^{C \log 2}}{n_{\mathsf{T}}} \right) \to 0,$$

Theorem 2 implies

$$\mathbb{E}\left[ \|\check{\tau}_{\mathtt{IPW}} - \tau\|_q^q \right] \to 0 \qquad \text{for every fixed } 1 \le q < \infty.$$

Trees with depth growing at a log–log rate can therefore be uniformly inconsistent despite vanishing integrated errors.

## 5. Discussion

Theorems 1 and 2 together show that polynomial uniform convergence rates can fail even when integrated accuracy is favorable. This is possible because large errors may be confined to cells containing only a small fraction of the population. We now discuss where such cells arise, how regularization and averaging change their effects, and the implications for inference, treatment rules, and practice.

**Decision Stumps.** Studying decision stumps isolates the role of the first split. Adaptive recursive partitioning, for example, can favor highly unbalanced splits, a phenomenon recognized since the early development of CART and known as the *end-cut preference*. When the signal is weak relative to sampling noise, the empirical splitting criterion may favor a threshold near the boundary of the parent cell.

For standard CART regression without sample splitting in one dimension ($p = 1$), Breiman et al. [1, Theorem 11.1] established the following end-cut preference [see also 18, Theorem 4]. If the response is independent of a continuously distributed covariate and has finite positive variance, then, for each $\delta \in (0,1)$, the first split produces a child containing at most a $\delta$ fraction of the sample with probability approaching one. For scalar regression trees and IPW trees without sample splitting, this earlier work on end-cut preference can, in turn, be used to obtain a lower bound on uniform error close to the parametric rate. That bound still allows relatively fast convergence, whereas our results rule out every polynomial rate of uniform convergence in $n$.

To quantify this end-cut behavior, we characterize the size of the smaller root child under Assumption 1. The root is split according to the prescribed criterion whenever a valid split exists. For each $l \in \{\mathtt{IPW}, \mathtt{DIM}, \mathtt{SSE}\}$, let $\hat{\imath}$ be the number of construction observations in the left root child and $\hat{\jmath}$ the selected coordinate. For each $1 \le j \le p$, Section SA-1.2 of the Supplemental Appendix then gives $\mathbb{P}(\hat{\jmath} = j) \to 1/p$ and

$$\frac{\log \min\{\hat{\imath}, n_{\mathsf{T}} - \hat{\imath}\}}{\log n_{\mathsf{T}}} \,\Big|\, \hat{\jmath} = j \rightsquigarrow \mathsf{Uniform}(0,1).$$

Here $(\hat{\imath}, \hat{\jmath}) = (1,1)$ on the event that no valid split exists, whose probability tends to zero. The displayed limit also holds unconditionally, so the first split produces a child containing at most $\sqrt{n_{\mathsf{T}}}$ construction observations with probability converging to $1/2$.

Going beyond split locations to their consequences for estimation accuracy, our decision stump results rule out any fixed polynomial rate of uniform convergence for every fixed $p \ge 1$, both with and without sample splitting. For NSS, we also give pointwise lower bounds along deterministic sequences approaching the support boundary. These conclusions thus quantify the estimation consequences of end-cut preference and cover causal tree constructions that differ from standard CART regression trees. See the CART, IPW, DIM, and SSE results in Sections SA-2 and SA-3 of the Supplemental Appendix for details.

**Deeper Trees, Multivariate Covariates, and the Location of Small Cells.** Under Assumption 1, the first split can isolate a small child cell with nonvanishing probability. Because recursive partitioning is hierarchical, further splits within this region only subdivide the observations it already contains. With multivariate covariates, subsequent splits can occur along any coordinate, producing cells with very different side lengths and with boundaries inside the original support while remaining within the small region created by the first split. Thus, the small cell phenomenon is not confined to one-dimensional boundary intervals.

**Regularization and Bias.** A natural response to the small cell phenomenon is therefore to regularize the tree construction. For instance, the algorithm may impose minimum node size constraints or incorporate penalties designed to discourage overfitting. These modifications can reduce estimation variance, but they can also limit the ability of the tree to capture treatment effect heterogeneity.

Adaptive tree procedures can generate small cells because local refinement is useful for reducing approximation bias, or because sampling noise leads to an imbalanced split. Sampling noise can, in particular, drive imbalanced splits even when all six estimators are exactly unbiased. These two sources are, however, difficult to separate in empirical applications. When treatment effects vary across the covariate space, regularization rules designed to eliminate small cells may introduce approximation bias and change overall convergence behavior. Supplemental Appendix, Section SA-1.4, gives additional discussion and a threshold example.

**$\alpha$-Regularity and Causal Random Forests.** Subsampling and averaging can also mitigate the local estimation errors associated with small terminal cells. Different trees, for example, may use different observations to estimate the treatment effect at the same covariate value, reducing the influence of individual noisy outcomes. Our lower bounds concern individual causal trees and therefore do not automatically extend to a forest average. How much averaging reduces variance depends on how trees share observations and whether the forest spreads weight across many observations or concentrates it on a few. Observation weights are also central to generalized random forests, which aggregate neighborhood weights before solving a local estimating equation [21]. For these and other

treatment effect estimators, the analysis in [50] examines how implementation choices affect the outcome weights.

To isolate the effect of subsampling at a fixed covariate value $\mathbf{x}$, let $T_s(\mathbf{x})$ be a symmetric, square-integrable tree rule, where symmetry means that reordering the training observations leaves its output unchanged. Fit the rule entirely on each size-$s$ subsample drawn without replacement from $n$ i.i.d. observations, averaging over any randomization independent of the data, and let $F_{n,s}(\mathbf{x})$ denote the equally weighted average of these fits over all such subsamples. The variance bound for averaging over subsamples [32, Proposition 3.1] then gives

$$\mathbb{V}[F_{n,s}(\mathbf{x})] \leq \frac{s}{n}\mathbb{V}[T_s(\mathbf{x})].$$

If $\mathbb{V}[T_s(\mathbf{x})]$ stays bounded, forest variance vanishes when $s/n \to 0$, but the bound gives no such guarantee for a fixed positive subsampling fraction. Under Assumption 1, each of the six tree estimators is bounded by a constant multiple of the largest absolute outcome in its subsample, so $\mathbb{V}[T_s(\mathbf{x})] = O(\log^2 s)$ for $s \geq 2$ and $s\log^2 s/n \to 0$ suffices for the forest variance to vanish. These bounds thus control variation around the mean prediction at one fixed covariate value. They do not, however, control bias or errors across the full covariate space, so uniform convergence rates for forests built from unrestricted CART-type trees require a separate analysis.

Subsample averaging, more specifically, leaves the mean prediction unchanged, $\mathbb{E}[F_{n,s}(\mathbf{x})] = \mathbb{E}[T_s(\mathbf{x})]$, so it cannot correct the bias of a tree fitted to $s$ observations, even if forest variance tends to zero. For a smooth, nonconstant CATE, such bias can arise from assigning weight to observations far from $\mathbf{x}$, even if the prediction averages over many observations. To control bias from distant observations in honest causal forests, the analysis in [20, Sections 3.2 and 4] uses Lipschitz conditional means in both treatment arms and shrinking terminal node diameters under the identification and tree regularity conditions imposed there.

These regularity conditions include restrictions on split balance, and Theorem 1 complements existing positive results by showing what can happen when those restrictions are removed. In particular, $\alpha$-regularity for honest causal trees and forests [20, Definitions 4 and 4b] requires each child node to contain at least a fixed fraction $\alpha \in (0, 1/2]$ of the observations in its parent node, so $n(\mathtt{t}_\mathtt{L}) \geq \alpha\, n(\mathtt{t})$ and $n(\mathtt{t}_\mathtt{R}) \geq \alpha\, n(\mathtt{t})$. Here $n(\cdot)$ counts observations in the sample on which balance is imposed, which is the estimation sample for double-sample trees. Related restrictions on child node observation counts also appear in quantile regression forests [19] and generalized random forests [21]. A balance condition on child node population volumes is also used in work on Boolean interaction recovery from random forest ensembles [22].

As with other forms of regularization discussed above, this balance requirement can substantially alter the adaptive behavior of recursive partitioning. For uniformly distributed covariates, a single unrestricted split can, for example, isolate a subgroup defined by a threshold on one coordinate and occupying less than an $\alpha$ fraction of the support. If the same balance requirement is imposed on conditional population masses, it rules out that split at the root, so isolating the subgroup requires further refinements. Supplemental Appendix, Section SA-1.5, gives the full threshold example and a lower bound on the required number of splits.

**Inference for Treatment Effects.** Small terminal nodes can also make inference on treatment effects at specific covariate values unreliable. When a terminal node contains too few estimation observations, Gaussian approximations can be inaccurate and estimated standard errors can be unstable. For DIM and SSE, the relevant sample sizes are the treated and control counts within the terminal node. A large overall sample therefore does not ensure reliable inference throughout the covariate support.

Honesty uses separate samples to define subgroups and estimate their treatment effects, but some subgroups may still contain too few observations for reliable inference. To justify conventional confidence intervals, we need to analyze estimation bias, how the estimates vary across repeated samples, and whether the estimated standard errors accurately measure that variation.

**Treatment Rules and Policy Value.** The local estimation errors documented in this paper nevertheless need not lead to poor treatment outcomes on average. To see why, suppose there are no treatment costs or constraints and treatment is assigned whenever the estimated treatment effect is positive. A tree that overestimates a positive effect, for example, still makes the correct decision to treat. Even when a decision is wrong, errors affecting only a small share of the population need not greatly reduce the population mean outcome.

We measure this average performance more formally by policy value, the population mean outcome under the rule. Regret, in turn, is the shortfall from the best policy value achievable using the true CATE. The uniform error lower bounds in Theorem 1 therefore do not by themselves imply lower bounds on the population share receiving an incorrect treatment assignment or on regret.

To relate regret to estimation error, let $\hat{\tau}$ be any measurable CATE estimator and $R(\hat{\tau})$ the regret of the resulting treatment rule. A wrong assignment reduces the mean outcome by the magnitude of the true CATE at that covariate value. Conditional on the fitting data, averaging these losses across the population gives

$$\begin{aligned} R(\hat{\tau}) &= \int_{\mathcal{X}} |\tau(\mathbf{x})| \left| \mathbb{1}(\hat{\tau}(\mathbf{x}) > 0) - \mathbb{1}(\tau(\mathbf{x}) > 0) \right| \mathrm{d}F_X(\mathbf{x}) \\ &\leq \|\hat{\tau} - \tau\|_1. \end{aligned}$$

Whether the CATE is constant or varies across covariate values, this inequality bounds regret by the population average absolute error of the effect estimates. To apply it to the six tree estimators studied here, we take expectations over the fitting data. Theorem 2 with $q = 1$ then gives expected regret bounds under Assumption 1, with and without sample splitting. A related connection appears in binary classification, where $L^1$ regression estimation error controls the excess risk of plug-in rules even without a margin condition [51, eq. (1.3)].

In this constant-effect benchmark, treating everyone is optimal when $\tau > 0$, and treating no one is optimal when $\tau < 0$. When $\tau = 0$, all treatment rules have the same value. Because these regret bounds are derived under a constant treatment effect, they do not establish that a tree can identify subgroups that benefit on average when treatment effects vary across the population.

The bounds also do not provide confidence regions for the set of covariate values at which treatment is beneficial or justify

inference on the unknown optimal policy value. Inference for that value has been studied separately [52].

**Practical Implications.** Theorems 1 and 2 together show why good average accuracy does not by itself ensure that a subgroup estimate is reliable. Before interpreting apparent differences in treatment effects, we therefore suggest checking how many treated and control observations support each estimate, even in a shallow or honest tree. For honest trees, these counts include only the observations reserved for effect estimation. If a conclusion rests on few observations, refitting with larger minimum node sizes or stricter treated and control count requirements can show how much it depends on these choices. Because refitting can change subgroup boundaries, compare estimated effects and estimation counts at the same covariate values, with the data and sample split held fixed.

Across these refits, changes in the sign or size of the estimates indicate sensitivity to the chosen restrictions. They do not, however, determine which fit is more accurate, since larger nodes can average over real differences in treatment effects and introduce approximation bias. Agreement across fits likewise does not establish accurate estimates or valid statistical inference.

## 6. Conclusion

Our results thus show that adaptive CART-type causal trees can have sharply different integrated and uniform accuracy, including under sample splitting. Even in a benchmark with a constant treatment effect, greedy splitting can, with nonvanishing probability, create terminal nodes containing few observations. As a result, causal trees may have favorable integrated $L^q$ accuracy while failing to achieve any fixed polynomial rate of uniform convergence in the sample size, a distinction that has implications for individualized decisions, subgroup conclusions, and inference.

Regularization is therefore important when causal trees are used for individualized treatment effect estimation or inference. Minimum node size rules, balance restrictions, pruning, and related modifications can limit small cells, but they also change the estimator and introduce tuning choices and bias-variance tradeoffs. To obtain guarantees for such estimators, existing positive theory often imposes balance or regularity conditions, rather than deriving them from canonical greedy CART splitting. Our results still leave open how practical regularization rules affect the accuracy of adaptively selected causal tree partitions. Beyond the constant-effect benchmark, another open question is whether adaptive CART-type causal trees can achieve nearly optimal integrated mean square convergence rates over larger function classes.

Trees, forests, and related ensembles can still be useful for learning and summarizing how treatment effects vary. Their suitability relative to other nonparametric methods depends, however, on the scientific objective, the dimension and structure of the problem, and whether accuracy is needed on average or at individual covariate values.

### Materials and Methods

We study i.i.d. observations from randomized experiments with binary treatment, a fixed number of independent and continuously distributed pretreatment covariates, and potential outcomes. Both potential outcomes, in particular, have constant conditional means, and the errors have light tails and are jointly independent of the covariates. Under this model, we analyze three CART-type splitting rules based on inverse probability weighting, difference in means, and squared error minimization. Applying each rule with either no sample splitting or honest sample splitting thus gives six estimators. The proofs use approximations for partial sums and Gaussian processes to characterize boundary splits and analyze estimation errors in the resulting small terminal nodes, accounting for tree depth and the observations available for effect estimation. See the Supplemental Appendix for details.

We used OpenAI Codex (GPT-6 Astra), Anthropic Claude Code (Fable 5.1 and Opus 5.5), and Google Antigravity (Gemini 4 Argon) to assist with preparing the manuscript, checking mathematical arguments, and developing and testing the replication code. The authors reviewed the resulting material and take responsibility for the final manuscript and code.

## Data, Materials, and Software Availability

Replication code and data are available at https://github.com/mdcattaneo/replication-CKY_2027_PNAS.

**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. We also thank the anonymous editor and two reviewers for their helpful comments and suggestions, which improved the paper. This paper supersedes the earlier unpublished work [53]. A previous version of this paper circulated under the title "The Honest Truth About Causal Trees: Accuracy Limits for Heterogeneous Treatment Effect Estimation" (arXiv:2509.11381) and contained additional technical results concerning **X**-adaptive recursive partitioning [see 54, Section 20.2, for related methods].

## Funding

Cattaneo gratefully acknowledges support from the National Science Foundation through grants SES-2019432, DMS-2210561, SES-2241575, and SES-2342226, from the National Institute of Food and Agriculture (NIFA) through grant 2024-67023-42704, and from the John Simon Guggenheim Memorial Foundation through a 2026 Guggenheim Fellowship. 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.

1. L Breiman, JH Friedman, RA Olshen, CJ Stone, *Classification and Regression Trees*. (Chapman and Hall/CRC), (1984).
2. X Su, T Zhou, X Yan, J Fan, S Yang, Interaction trees with censored survival data. *The International Journal of Biostatistics* **4**, Article 2 (2008).
3. X Su, CL Tsai, H Wang, DM Nickerson, B Li, Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research* **10**, 141–158 (2009).
4. I Lipkovich, A Dmitrienko, J Denne, G Enas, Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Statistics in Medicine* **30**, 2601–2621 (2011).
5. JC Foster, JMG Taylor, SJ Ruberg, Subgroup identification from randomized clinical trial data. *Statistics in Medicine* **30**, 2867–2880 (2011).
6. E Dusseldorp, I Van Mechelen, Qualitative interaction trees: A tool to identify qualitative treatment–subgroup interactions. *Statistics in Medicine* **33**, 219–237 (2014).
7. E Dusseldorp, JJ Meulman, The regression trunk approach to discover treatment covariate interaction. *Psychometrika* **69**, 355–374 (2004).
8. J Kang, X Su, B Hitsman, K Liu, D Lloyd-Jones, Tree-structured analysis of treatment effects with large observational data. *Journal of Applied Statistics* **39**, 513–529 (2012).
9. WY Loh, X He, M Man, A regression tree approach to identifying subgroups with differential treatment effects. *Statistics in Medicine* **34**, 1818–1833 (2015).
10. H Seibold, A Zeileis, T Hothorn, Model-based recursive partitioning for subgroup analyses. *The International Journal of Biostatistics* **12**, 45–63 (2016).
11. S Athey, G Imbens, Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences* **113**, 7353–7360 (2016).
12. FJ Bargagli Stoffi, G Gnecco, Causal tree with instrumental variable: An extension of the causal tree framework to irregular assignment mechanisms. *International Journal of Data Science and Analytics* **9**, 315–337 (2020).
13. G Wang, J Li, WJ Hopp, An instrumental variable forest approach for detecting heterogeneous treatment effects in observational studies. *Management Science* **68**, 3399–3418 (2022).
14. W Zhang, TD Le, L Liu, ZH Zhou, J Li, Mining heterogeneous causal effects for personalized cancer treatment. *Bioinformatics* **33**, 2372–2378 (2017).

15. W Zhang, TD Le, L Liu, J Li, Estimating heterogeneous treatment effect by balancing heterogeneity and fitness. *BMC Bioinformatics* **19**, 518 (2018).
16. J Yang, IJ Dahabreh, JA Steingrimsson, Causal interaction trees: Finding subgroups with heterogeneous treatment effects in observational data. *Biometrics* **78**, 624–635 (2022).
17. A Venkatasubramaniam, et al., Assessing effect heterogeneity of a randomized treatment using conditional inference trees. *Statistical Methods in Medical Research* **31**, 549–562 (2022).
18. H Ishwaran, The effect of splitting on random forests. *Machine Learning* **99**, 75–118 (2015).
19. N Meinshausen, Quantile regression forests. *Journal of Machine Learning Research* **7**, 983–999 (2006).
20. S Wager, S Athey, Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association* **113**, 1228–1242 (2018).
21. S Athey, J Tibshirani, S Wager, Generalized random forests. *Annals of Statistics* **47**, 1148–1178 (2019).
22. M Behr, Y Wang, X Li, B Yu, Provable Boolean interaction recovery from tree ensemble obtained via random forests. *Proceedings of the National Academy of Sciences* **119**, e2118636119 (2022).
23. E Scornet, G Biau, JP Vert, Consistency of random forests. *Annals of Statistics* **43**, 1716–1741 (2015).
24. CM Chi, P Vossler, Y Fan, J Lv, Asymptotic properties of high-dimensional random forests. *Annals of Statistics* **50**, 3415–3438 (2022).
25. JM Klusowski, PM Tian, Large scale prediction with decision trees. *Journal of the American Statistical Association* **119**, 525–537 (2024).
26. R Mazumder, H Wang, On the convergence of CART under sufficient impurity decrease condition in *Advances in Neural Information Processing Systems*. (Curran Associates, Inc.), Vol. 36, pp. 57754–57782 (2023).
27. MD Cattaneo, R Chandak, JM Klusowski, Convergence rates of oblique regression trees for flexible function libraries. *Annals of Statistics* **52**, 466–490 (2024).
28. YS Tan, A Agarwal, B Yu, A cautionary tale on fitting decision trees to data from additive models: Generalization lower bounds in *Proceedings of the 25th International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research. (PMLR), Vol. 151, pp. 9663–9685 (2022).
29. YS Tan, O Ronen, T Saarinen, B Yu, On the computational efficiency of Bayesian additive regression trees: An asymptotic analysis. *arXiv preprint arXiv:2406.19958* (2026).
30. HA Chipman, EI George, RE McCulloch, BART: Bayesian additive regression trees. *Annals of Applied Statistics* **4**, 266–298 (2010).
31. YS Tan, JM Klusowski, K Balasubramanian, Statistical-computational trade-offs for recursive adaptive partitioning estimators. *Annals of Statistics* **54**, 1262–1288 (2026).
32. P Bühlmann, B Yu, Analyzing bagging. *Annals of Statistics* **30**, 927–961 (2002).
33. M Banerjee, IW McKeague, Confidence sets for split points in decision trees. *Annals of Statistics* **35**, 543–574 (2007).
34. M Csörgő, P Révész, *Strong Approximations in Probability and Statistics*, Probability and Mathematical Statistics : a series of monographs and textbooks. (Academic Press), (1981).
35. I Waudby-Smith, M Larsson, A Ramdas, Nonasymptotic and distribution-uniform Komlós–Major–Tusnády approximation (2025) arXiv:2502.06188.
36. J Pitman, M Yor, A guide to Brownian motion and related stochastic processes (2018) arXiv:1802.09679.
37. R Latała, D Matlak, Royen's proof of the Gaussian correlation inequality in *Geometric Aspects of Functional Analysis: Israel Seminar (GAFA) 2014–2016*, Lecture Notes in Mathematics, eds. B Klartag, E Milman. (Springer International Publishing) Vol. 2169, pp. 265–275 (2017).
38. L Horváth, The maximum likelihood method for testing changes in the parameters of normal observations. *Annals of Statistics* **21**, 671–680 (1993).
39. AI Zhdanov, High excursions of a quadratic form for a Gaussian stationary vector process. *Fundamentalnaya i Prikladnaya Matematika* **23**, 123–144 (2020) In Russian; English translation in Journal of Mathematical Sciences 262(4) (2022), 476–492.
40. DA Levin, Y Peres, *Markov Chains and Mixing Times.* (American Mathematical Society, Providence, RI), 2nd edition, (2017) With contributions by Elizabeth L. Wilmer.
41. JK Hunter, Notes on partial differential equations (2014) Lecture notes, University of California, Davis. Revised June 18, 2014. https://www.math.ucdavis.edu/~hunter/pdes/pde_notes.pdf.
42. A Carbotti, S Cito, DA La Manna, D Pallara, Stability of the Gaussian Faber–Krahn inequality. *Annali di Matematica Pura ed Applicata* **203**, 2185–2198 (2024).
43. F Eicker, The asymptotic distribution of the suprema of the standardized empirical processes. *Annals of Statistics* **7**, 116–138 (1979).
44. MA Hernán, JM Robins, *Causal Inference: What If.* (Boca Raton: Chapman & Hall/CRC), (2020).
45. L Györfi, M Kohler, A Krzyżak, H Walk, *A Distribution-Free Theory of Nonparametric Regression.* (Springer-Verlag), (2002).
46. MD Cattaneo, MH Farrell, Y Feng, Large sample properties of partitioning-based series estimators. *Annals of Statistics* **48**, 1718–1741 (2020).
47. MD Cattaneo, Y Feng, B Shigida, Uniform estimation and inference for nonparametric partitioning-based M-estimators. *Annals of Statistics* **54**, 1740–1767 (2026).
48. RA Berk, *Statistical learning from a regression perspective*, Springer Texts in Statistics. (Springer Nature), Third edition, (2020).
49. H Zhang, BH Singer, *Recursive Partitioning and Applications*, Springer Series in Statistics. (Springer), Second edition, (2010).
50. MC Knaus, Treatment effect estimators as weighted outcomes (arXiv preprint arXiv:2411.11559) (2024).
51. JY Audibert, AB Tsybakov, Fast learning rates for plug-in classifiers. *Annals of Statistics* **35**, 608–633 (2007).
52. AR Luedtke, MJ van der Laan, Statistical inference for the mean outcome under a possibly non-unique optimal treatment strategy. *Annals of Statistics* **44**, 713–742 (2016).
53. MD Cattaneo, JM Klusowski, PM Tian, On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *Technical report, arXiv preprint arXiv:2211.10805v1* (2022).
54. L Devroye, L Györfi, G Lugosi, *A Probabilistic Theory of Pattern Recognition*, Stochastic Modelling and Applied Probability. (Springer) Vol. 31, (1996).