On the Pointwise Behavior of Recursive Partitioning and Its Implications for Heterogeneous Causal Effect Estimation

Matias D. Cattaneo*[a], Jason M. Klusowski†[a], and Peter M. Tian‡[b]

[a]Department of Operations Research and Financial Engineering , Princeton University
[b]Two Sigma Investments, LP

January 28, 2024

Abstract

Decision tree learning is increasingly being used for pointwise inference. Important applications include causal heterogenous treatment effects and dynamic policy decisions, as well as conditional quantile regression and design of experiments, where tree estimation and inference is conducted at specific values of the covariates. In this paper, we call into question the use of decision trees (trained by adaptive recursive partitioning) for such purposes by demonstrating that they can fail to achieve polynomial rates of convergence in uniform norm with non-vanishing probability, even with pruning. Instead, the convergence may be arbitrarily slow or, in some important special cases, such as *honest* regression trees, fail completely. We show that random forests can remedy the situation, turning poor performing trees into nearly optimal procedures, at the cost of losing interpretability and introducing two additional tuning parameters. The two hallmarks of random forests, subsampling and the random feature selection mechanism, are seen to each distinctively contribute to achieving nearly optimal performance for the model class considered.

Keywords: recursive partitioning, decision trees, random forests, pointwise estimation, causal inference, heterogeneous treatment effects

*cattaneo@princeton.edu
†jason.klusowski@princeton.edu
‡ptian@twosigma.com

1 Introduction

As data-driven technologies continue to be adopted and deployed in high-stakes decision-making environments, the need for fast, interpretable algorithms has never been more important. As one such candidate, it has become increasingly common to use decision trees, constructed by adaptive recursive partitioning, for inferential tasks on a predictive or causal model. These applications are spurred by the appealing connection between decision trees and rule-based decision-making, particularly in clinical, legal, or business contexts, as the determination of the output mimics the way a human user may think and reason [Berk, 2020]. Decision trees are ubiquitous in empirical work not only because they offer an interpretable decision-making methodology [Murdoch et al., 2019, Rudin, 2019], but also because their construction relies on data-adaptive implementations that take into account the specific features of the underlying data generating process. See Hastie et al. [2009] for a textbook introduction.

While data-adaptive, rule-based tree learning is powerful, it is not without its pitfalls. In this paper, we provide theoretical evidence of these shortcomings in commonly encountered data situations. Focusing on the simplest possible data generating process (i.e., a homoskedastic constant regression/treatment effect model), we show that decision trees cannot converge faster any polynomial function of the sample size n, uniformly over the entire support of the covariates, with non-vanishing probability. Furthermore, when adding honesty to the tree construction, which is often regarded as an improvement over canonical tree fitting [Athey and Imbens, 2016], we show that the resulting decision trees can be inconsistent, uniformly over the covariate support, as soon as the depth of the tree is at least a constant multiple of $\log\log(n)$ (e.g., $\log\log(n) \approx 3$ for $n = 1$ billion observations).

Our results paint a rather bleak picture of decision trees, if the goal is to use them for statistical learning *pointwise* (or *uniformly*) over the entire support of the covariates; they can produce unreliable estimates even in large samples for the simplest possible statistical model underlying the data generation. Thankfully, in such settings, we are able to show that random forests are provably superior and exhibit optimal performance when the constituent trees do not. This improvement comes at the cost of losing interpretability and introducing two additional tuning parameters (subsample size and number of candidate variables to consider at each node).

To formalize our results, we consider the canonical regression model where the observed data $\{(y_i, \mathbf{x}_i^T) : i = 1, 2, \dots n\}$ is a random sample satisfying

$$y_i = \mu(\mathbf{x}_i) + \varepsilon_i, \qquad \mathbb{E}[\varepsilon_i \mid \mathbf{x}_i] = 0, \qquad \mathbb{E}[\varepsilon_i^2 \mid \mathbf{x}_i] = \sigma^2(\mathbf{x}_i), \tag{1}$$

with $\mathbf{x}_i = (x_{i1}, x_{i2}, \dots, x_{ip})^T$ a vector of p covariates taking values on some support set $\mathcal{X}$. The parameter of interest is the conditional mean response function $\mu(\mathbf{x}_i) = \mathbb{E}[y_i \mid \mathbf{x}_i]$, which may be assumed to belong to some smooth, or otherwise appropriately restricted, set of functions. The goal is to use the observed data together with an algorithmic procedure

to learn $\mu(\mathbf{x})$ for all values of $\mathbf{x} \in \mathcal{X}$. While there are many ways to grow a decision tree (i.e., a partition of $\mathcal{X}$), our focus throughout this paper will be on the CART algorithm [Breiman et al., 1984], by far the most popular in practice.

A decision tree is a hierarchically organized data structure constructed in a top down, greedy manner through recursive binary splitting. According to conventional CART methodology, a parent node t (i.e., a region in $\mathcal{X}$) in the tree is divided into two child nodes, t_L and t_R, by minimizing the sum-of-squares error (SSE)

$$\sum_{\mathbf{x}_i \in \mathrm{t}} (y_i - \beta_1 \mathbb{1}(x_{ij} \leq \tau) - \beta_2 \mathbb{1}(x_{ij} > \tau))^2, \tag{2}$$

with respect to the child node outputs, split point, and split direction, $(\beta_1, \beta_2, \tau, j)$, with $\mathbb{1}(\cdot)$ denoting the indicator function.

Because the splits occur along values of a single covariate, the induced partition of the input space $\mathcal{X}$ is a collection of hyper-rectangles. The solution of (2) yields estimates $(\hat{\beta}_1, \hat{\beta}_2, \hat{\tau}, \hat{\jmath})$, and the resulting refinement of t produces child nodes $\mathrm{t}_L = \{\mathbf{x} \in \mathrm{t} : x_{\hat{\jmath}} \leq \hat{\tau}\}$ and $\mathrm{t}_R = \{\mathbf{x} \in \mathrm{t} : x_{\hat{\jmath}} > \hat{\tau}\}$. The normal equations imply that $\hat{\beta}_1 = \bar{y}_{\mathrm{t}_L} = \frac{1}{\#\{\mathbf{x}_i \in \mathrm{t}_L\}} \sum_{\mathbf{x}_i \in \mathrm{t}_L} y_i$ and $\hat{\beta}_2 = \bar{y}_{\mathrm{t}_R} = \frac{1}{\#\{\mathbf{x}_i \in \mathrm{t}_R\}} \sum_{\mathbf{x}_i \in \mathrm{t}_R} y_i$, the respective sample means after splitting the parent node at $x_{\hat{\jmath}} = \hat{\tau}$, where $\#A$ denotes the cardinality of the set A. These child nodes become new parent nodes at the next level of the tree and can be further refined in the same manner, and so on and so forth, until a desired depth is reached. To obtain a maximal decision tree T_K of depth K, the procedure is iterated K times until (i) the node contains a single data point $(y_i, \mathbf{x}_i^T)$ or (ii) all input values $\mathbf{x}_i$ and/or all response values y_i within the node are the same.

In a conventional regression problem, where the goal is to estimate the conditional mean response $\mu(\mathbf{x})$, the tree output for $\mathbf{x} \in \mathrm{t}$ is the within-node sample mean $\bar{y}_{\mathrm{t}}$, i.e., if T is a decision tree, then $\hat{\mu}(T)(\mathbf{x}) = \bar{y}_{\mathrm{t}} = \frac{1}{\#\{\mathbf{x}_i \in \mathrm{t}\}} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i$. However, one can aggregate the data in the node in a number of ways, depending on the target estimand. For example, CART methodology is also commonly used for classification tasks (e.g., propensity score estimation in causal inference settings), in particular, where the outcome variable $y_i \in \{0, 1\}$ takes on binary values. In this case, the classification tree output is the majority vote of the class instances in the node. Because the canonical splitting criterion for binary classification, the *Gini index*, is equivalent to (2), the results presented in this paper are directly applicable. In addition, decision tree methodology can also be employed for conditional quantile regression and its various downstream tasks, such as estimating quantiles, constructing confidence intervals, or performing outlier detection [Meinshausen, 2006, and references therein]. These methods also require high pointwise accuracy of decision trees, and thus our results will have methodological implications in those settings as well.

Furthermore, in multi-step semiparametric settings, it is often the case that preliminary unknown functions (e.g., propensity scores in causal inference settings) are estimated

using modern machine learning methods such as CART [see, for example, Chernozhukov et al., 2022, and references therein]. Our results reveal that reliance on fast uniform convergence rates for decision tree methodology may not be guaranteed, as we show below that decision trees will have a convergence rate slower than any polynomial-in-n, over the entire support $\mathcal{X}$. This finding implies that other machine learning procedures such as neural networks [Farrell et al., 2021, and references therein] may be preferable in those multi-step semiparametric settings, if such methods could be shown to be uniformly consistent with sufficiently fast rates of convergence.

From a big picture perspective, our main methodological message is to warn against mechanical application of flexible, adaptive machine learning methodologies for tasks that require good quality estimates at specific covariate values of interest. Machine learning procedures that are currently deployed in practice (for canonical regression problems) are trained to approximately minimize the empirical mean squared error. As such, they enjoy good out-of-sample accuracy for an average-case value of the covariates, i.e., if accuracy is measured via the integrated mean squared error (IMSE). However, if the task requires a more stringent form of convergence, such as uniform convergence, it is unknown if those procedures meet such additional demands. Our results are the first to formally show that this is not the case for decision trees, despite them having small IMSE.

2 Causal Inference and Policy Decisions

As mentioned earlier, recursive partitioning is now a common tool of choice in the analysis of heterogeneous causal treatment effects and the design of heterogeneous policy interventions [Athey and Imbens, 2019, Yao et al., 2021, and references therein]. Here the observed data is a random sample $\{(y_i, \mathbf{x}_i^T, d_i) : i = 1, 2, \ldots, n\}$, where y_i is the outcome of interest, $\mathbf{x}_i$ is a set of pre-treatment covariates, and d_i is a binary treatment indicator variable. Employing standard potential outcomes notation,

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

where $y_i(1)$ is the potential outcome under treatment ($d_i = 1$) and $y_i(0)$ is the potential outcome under control ($d_i = 0$). This paradigm is fundamental to most applied sciences; for example, it can be used to model the effectiveness of a drug therapy, behavioral intervention, marketing campaign, or government program.

In cases where the individual treatment effect $y_i(1) - y_i(0)$ varies across different subgroups, a natural goal is to estimate the *heterogeneous* average treatment effect (ATE) for each covariate value $\mathbf{x} \in \mathcal{X}$, namely, $\theta(\mathbf{x}) = \mathbb{E}[y_i(1) - y_i(0) \mid \mathbf{x}_i = \mathbf{x}]$. In recent years, there has been an explosion of machine learning technologies adapted for heterogeneous causal effect estimation, thanks to the abundance of data produced from large-scale experiments and observational studies. Among these machine learning algorithms, recursive partitioning estimators (specifically, *causal decision trees*) stand out as natural contenders, as they

are well-suited for grouping data according to the treatment effect size, conditional on observable characteristics [e.g., Su et al., 2009, Athey and Imbens, 2016].

We now discuss CART methodology in the context of heterogeneous causal effect estimation, one popular application of decision trees where accurate pointwise estimates over the entire support $\mathcal{X}$ are essential. In experimental settings, where $(y_i(0), y_i(1), \mathbf{x}_i^T) \perp\!\!\!\perp d_i$, the *conditional* ATE is identifiable because

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

where the probability of treatment assignment $\xi = \mathbb{P}(d_i = 1)$ is known by virtue of the known randomization mechanism. It follows that $\theta(\mathbf{x})$, $\mathbf{x} \in \mathcal{X}$, can be estimated using decision tree methodology in at least two ways, namely, for a decision tree T,

$$\hat{\theta}_{\text{reg}}(T)(\mathbf{x}) = \frac{1}{\#\{\mathbf{x}_i \in \text{t} : d_i = 1\}} \sum_{\mathbf{x}_i \in \text{t}: d_i = 1} y_i - \frac{1}{\#\{\mathbf{x}_i \in \text{t} : d_i = 0\}} \sum_{\mathbf{x}_i \in \text{t}: d_i = 0} y_i, \tag{3}$$

or

$$\hat{\theta}_{\text{ipw}}(T)(\mathbf{x}) = \frac{1}{\#\{\mathbf{x}_i \in \text{t}\}} \sum_{\mathbf{x}_i \in \text{t}} y_i \frac{d_i - \xi}{\xi(1-\xi)}, \tag{4}$$

where recall t denotes the unique (terminal) node containing $\mathbf{x} \in \mathcal{X}$.

In this spirit, we consider a tree-based approach for analyzing treatment effect heterogeneity in randomized control trials, which may also be used to design personalized treatment assignments based on pre-intervention observable characteristics. While our forthcoming results are stated for the regression problem (1), they are also directly applicable to the causal decision tree estimators above that involve minimizing the SSE criterion. This is precisely because $\hat{\theta}_{\text{reg}}(T)(\mathbf{x})$ and $\hat{\theta}_{\text{ipw}}(T)(\mathbf{x})$ can be implemented using conventional CART methodology. That is, we implement $\hat{\theta}_{\text{reg}}(T)(\mathbf{x})$ following a plug-in approach that estimates $\mathbb{E}[y_i \mid \mathbf{x}_i,\ d_i = 1]$ and $\mathbb{E}[y_i \mid \mathbf{x}_i,\ d_i = 0]$ separately with regression trees and conventional CART methodology. Alternatively, we fit a regression tree with CART methodology to the transformed outcome $y_i(d_i - \xi)/(\xi(1-\xi))$ to implement $\hat{\theta}_{\text{ipw}}(T)(\mathbf{x})$. Yet another (more principled) approach [Athey and Imbens, 2016] implements $\hat{\theta}_{\text{reg}}(T)(\mathbf{x})$ by growing a decision tree using a slightly modified version of the SSE criterion (2) (referred to as *adjusted expected MSE*) that more directly targets the conditional ATE, together with an *honest* property, where different samples are used for constructing the partition and estimating the effects of each subpopulation.

Our theory implies that, for a constant treatment effect model, the aforementioned causal decision tree estimators cannot converge faster than any polynomial-in-n. Furthermore, in more interesting cases, shallow (honest) causal decision tree estimators will be shown to be inconsistent, as a function of the sample size n, for some $\mathbf{x} \in \mathcal{X}$. Finally, we will also

show that random forest methodology, while hurting the interpretability and introducing additional tuning parameters, can overcome the limitations of decision trees by restoring nearly optimal pointwise (for all $\mathbf{x} \in \mathcal{X}$) convergence rates.

3 Homoskedastic Constant Regression Model

To formalize the pitfalls of pointwise regression estimation using decision trees, we consider the simplest possible data generating process.

Assumption 1 (Location Regression Model)**.** *The observed data* $\{(y_i, \mathbf{x}_i^T) : i = 1, 2, \ldots, n\}$ *is a random sample satisfying* (1) *and the following:*

1. $\mu(\mathbf{x}) \equiv \mu$ *is constant for all* $\mathbf{x} \in \mathcal{X} \subseteq \mathbb{R}^p$.
2. $\mathbf{x}_i$ *has a continuous distribution.*
3. $\mathbf{x}_i \perp\!\!\!\perp \varepsilon_i$ *for all* $i = 1, 2, \ldots, n$.
4. $\mathbb{E}[\varepsilon_i^2 \log\log(|\varepsilon_i| + 1)] < \infty$ *and* $\mathbb{E}[\varepsilon_i^2] > 0$.

Because trees are invariant with respect to monotone transformations of the coordinates of $\mathbf{x}$, without loss of generality, we assume henceforth that the marginal distributions of the covariates are uniformly distributed on $\mathcal{X} = [0, 1]^p$, i.e., $x_j \sim U([0, 1])$ for $j = 1, 2, \ldots, p$.

Under Assumption 1, the regression model (1) becomes the standard location (or *intercept-only regression*) model with homoskedastic errors:

$$y_i = \mu + \varepsilon_i, \qquad \sigma^2 = \mathbb{E}[\varepsilon_i^2].$$

In the causal setting, the assumption corresponds to the constant treatment effect model, in which $\theta(\mathbf{x}) \equiv \theta$ is constant for all pre-treatment covariates $\mathbf{x}$.

This statistical model is perhaps the most canonical member of any interesting set of data generating processes. In particular, the regression function belongs to all classical smoothness function classes, as well as to the set of functions with bounded total variation. See, for example, Györfi et al. [2002] for review and further references. As a consequence, our results will also shed light in settings where uniformity over any of the aforementioned classes of functions is of interest, since our lower bounds can be applied directly in those cases. To be more precise, if $\hat{\mu}(T)(\mathbf{x})$ is the output from a decision tree T, then for any class of data generating processes $\mathcal{P}$ containing the model defined by Assumption 1, $\sup_{\mathbb{P} \in \mathcal{P}} \mathbb{P}(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}(T)(\mathbf{x}) - \mu(\mathbf{x})| > \epsilon) \geq \mathbb{P}(\sup_{\mathbf{x} \in \mathcal{X}} |\hat{\mu}(T)(\mathbf{x}) - \mu| > \epsilon)$, for any $\epsilon > 0$. Because $\mathcal{P}$ will include the model defined by Assumption 1 in all relevant (both theoretically and practically) cases, our results also highlight fundamental limitations of CART regression methods from a uniform (over $\mathcal{P}$) perspective, whenever interest lies on estimation of $\mu(\mathbf{x})$ for all $\mathbf{x} \in \mathcal{X}$.

Since the main purpose of this paper is to explore the limits of decision tree methodology, we do not aim for generality, but rather consider the simplest possible data generating process (Assumption 1). In the context of causal inference and treatment effects (e.g., Section 2), the assumptions correspond to a constant treatment effect model, the most basic case of practical interest. Importantly, Assumption 1 removes issues related to smoothing (or misspecification) bias because the regression function $\mu(\mathbf{x})$ is constant for all $\mathbf{x} \in \mathcal{X}$, which shows that our results will not be driven by standard (boundary or other smoothing) bias in nonparametrics [Fan and Gijbels, 1996]. Indeed, if the distribution of ε_i is symmetric, then we have $\mathbb{E}[\hat{\mu}(T)(\mathbf{x}) - \mu] = -\mathbb{E}[\hat{\mu}(T)(\mathbf{x}) - \mu] \implies \mathbb{E}[\hat{\mu}(T)(\mathbf{x})] = \mu$, owing to the fact that the split points $\hat{\tau}$ are symmetric statistics of the ε_i. Our results will be driven instead by the fact that decision tree methodology can generate small cells containing only a handful of observations, thereby making the estimator imprecise in certain regions of $\mathcal{X}$. In other words, inconsistency is due to a *large variance* problem, not a large bias problem.

The location (or constant treatment effect) model is the simplest instantiation of a regression model of practical interest because the regression function is supersmooth and the curse of dimensionality is absent. Furthermore, all smooth regression functions can be seen as locally constant. Thus, we should expect any competitive nonparametric estimator to separate a constant signal from noise or, in the language of causal inference, to estimate accurately (constant) treatment effects when they happen to be homogeneous. Assumption 1 also approximately captures another common modeling situation in machine learning and data science, in which the marginal distribution $y_i \mid x_{ij}$ is noisy (i.e., the marginal projections $\mathbb{E}[y_i \mid x_{ij}]$ are constant and contain no signal). Because splits in trees are determined using only marginal information, here the split at the root node would be essentially fitting the location model.

4 Decision Stumps

For each variable $j = 1, 2, \ldots, p$, the data $\{x_{ij} : \mathbf{x}_i \in \mathrm{t}\}$ is relabeled so that x_{ij} is increasing in the index $i = 1, 2, \ldots, n(\mathrm{t})$, where $n(\mathrm{t}) = \#\{\mathbf{x}_i \in \mathrm{t}\}$. Then, minimizing the objective (2) can be equivalently recast as maximizing the so-called *impurity gain*:

$$\begin{aligned} &\sum_{\mathbf{x}_l \in \mathrm{t}} (y_l - \bar{y}_{\mathrm{t}})^2 - \sum_{\mathbf{x}_l \in \mathrm{t}} (y_l - \bar{y}_{\mathrm{t}_L} \mathbb{1}(x_{lj} \leq \tau) - \bar{y}_{\mathrm{t}_R} \mathbb{1}(x_{lj} > \tau))^2 \\ &\quad = \frac{\left(\frac{1}{\sqrt{n(\mathrm{t})}} \sum_{l=1}^{i} (y_l - \mu) - \frac{i}{n(\mathrm{t})} \frac{1}{\sqrt{n(\mathrm{t})}} \sum_{l=1}^{n(\mathrm{t})} (y_l - \mu) \right)^2}{i(n(\mathrm{t}) - i)}, \end{aligned} \tag{5}$$

with respect to the index i and variable j; see [Breiman et al., 1984]. The maximizers are denoted by $(\hat{\imath}, \hat{\jmath})$, and the optimal split point $\hat{\tau}$ that minimizes (2) can be expressed as $x_{\hat{\imath}\hat{\jmath}}$.

We start by considering the case when the tree is depth one ($K = 1$), i.e., a decision stump.

The tree output can then be written as

$$\hat{\mu}(T_1)(\mathbf{x}) = \hat{\beta}_1 \mathbb{1}(x_{\hat{\jmath}} \leq \hat{\tau}) + \hat{\beta}_2 \mathbb{1}(x_{\hat{\jmath}} > \hat{\tau}) = \begin{cases} \frac{1}{\#\{\mathbf{x}_i : x_{ij} \leq x_{i\hat{\jmath}}\}} \sum_{\mathbf{x}_i : x_{ij} \leq x_{i\hat{\jmath}}} y_i, & x_{\hat{\jmath}} \leq x_{\hat{\imath}\hat{\jmath}} \\ \frac{1}{\#\{\mathbf{x}_i : x_{ij} > x_{i\hat{\jmath}}\}} \sum_{\mathbf{x}_i : x_{ij} > x_{i\hat{\jmath}}} y_i, & x_{\hat{\jmath}} > x_{\hat{\imath}\hat{\jmath}} \end{cases}, \tag{6}$$

where $x_{\hat{\jmath}}$ denotes the value of the $\hat{\jmath}$-th component of $\mathbf{x}$.

The following theorem formally (and very precisely) characterizes the regions of the support $\mathcal{X}$ where the first CART split index $\hat{\imath}$, at the root node, has non-vanishing probability of realizing. As a consequence, the theorem also characterizes the effective sample size of the resulting cells (recall the data is ordered so that $\hat{\tau} = x_{\hat{\imath}\hat{\jmath}}$ and hence $\hat{\imath} = \#\{\mathbf{x}_i : x_{ij} \leq \hat{\tau}\}$).

Theorem 4.1. *Suppose Assumption 1 holds and $p = 1$, and let $\hat{\imath}$ be the CART split index at the root node. For each $a, b \in (0, 1)$ with $a < b$, we have*

$$\liminf_{n\to\infty} \mathbb{P}(n^a \leq \hat{\imath} \leq n^b) = \liminf_{n\to\infty} \mathbb{P}(n - n^b \leq \hat{\imath} \leq n - n^a) \geq \frac{b-a}{e}. \tag{7}$$

Remark 1. *We conjecture that* $\lim_{n\to\infty} \mathbb{P}(n^a \leq \hat{\imath} \leq n^b) = \lim_{n\to\infty} \mathbb{P}(n - n^b \leq \hat{\imath} \leq n - n^a) = (b - a)/2$. *Another way of stating this conjecture is that the asymptotic conditional distribution of* $\log(\hat{\imath})/\log(n)$ *given that* $\hat{\imath} \leq n/2$ *is* $U([0, 1])$. *Evidence for this limit will be given in the proof of Theorem 4.1.*

First, Theorem 4.1 shows that with non-vanishing probability, $\hat{\imath}$ will realize near its extremes, from the beginning of any tree construction. The arbitrarily slow polynomial-in-n rates do not contradict, but are rather precluded by, existing polynomial convergence guarantees [e.g., Wager and Athey, 2018], which a priori require that each split generates two child nodes that contain a constant fraction of the number of observations in the parent node, i.e., $n(\mathrm{t}_L) \gtrsim n(\mathrm{t})$ and $n(\mathrm{t}_R) \gtrsim n(\mathrm{t})$. By implication, Theorem 4.1 shows that such assumptions requiring *balanced* cells almost surely, which are typically imposed in the literature, are in general incompatible with standard decision tree constructions employing conventional CART methodology [e.g., Behr et al., 2022, and references therein]. The slow convergence rates for the decision stump occur because the optimal split point is realized near the boundary of the support [Ishwaran, 2015] with non-vanishing probability, i.e., $\hat{\tau} \approx 0$ or $\hat{\tau} \approx 1$ with non-vanishing probability, causing the two nodes in the stump to be imbalanced, with one containing a much smaller number of samples, and therefore rendering a situation where local averaging is less accurate. To be more precise, after the first split when $n(\mathrm{t}) = n$, CART will generate two unbalanced cells with non-vanishing probability; for any $a, b \in (0, 1)$ with $a < b$, either $n^a \leq n(\mathrm{t}_L) \leq n^b$ or $n^a \leq n(\mathrm{t}_R) \leq n^b$ for large n, where $n(\mathrm{t}_L) + n(\mathrm{t}_R) = n$. It will follow from this result that, on the events considered in (7), too few observations will be available on one of the cells after the first split for CART to deliver a polynomial-in-n consistent estimator of μ, thereby making the decision tree procedure exhibit arbitrarily slow rates, for some $x \in \mathcal{X}$.

4.1 Convergence Rates

Theorem 4.1 appears to be new in the literature. It arises from a careful study of the maximum of (5) over different ranges of the split index, which turns out to be asymptotically similar to the suprema of a standardized Brownian bridge over different time intervals.

Once the location of the first CART split point is well-understood, we can study the resulting CART estimator $\hat{\mu}(T_1)(\mathbf{x})$ of the unknown regression function. The following statements hold for the pointwise prediction error of the decision stump.

Theorem 4.2. *Suppose Assumption 1 holds and $p = 1$, and let $\hat{\mu}(T_1)(x)$ be the CART estimator of the regression function at the root node. For any $a, b \in (0, 1)$ with $a < b$, we have*

$$\liminf_{n\to\infty} \mathbb{P}\left(\sup_{x\in\mathcal{X}} |\hat{\mu}(T_1)(x) - \mu| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log\log(n)} \right) \geq \frac{2b}{e}, \tag{8}$$

and

$$\liminf_{n\to\infty} \inf_{x\in\mathcal{X}_n} \mathbb{P}\left(|\hat{\mu}(T_1)(x) - \mu| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log\log(n)} \right) \geq \frac{b - a}{e}, \tag{9}$$

where $\mathcal{X}_n = [0,\ (1 + o(1))n^{a-1}) \cup (1 - (1 + o(1))n^{a-1},\ 1]$.

The theorem above shows that decision stumps can have, at most, $n^{b/2}$ (suboptimal) convergence for evaluation points that are within n^{a-1} distance from the boundary of $\mathcal{X}$ (see (9)), for *any* $a, b \in (0, 1)$ with $a < b$. This happens because the two nodes in the stump are highly imbalanced with non-trivial probability under Assumption 1, with one containing a much smaller number of samples—thereby making local estimation difficult. An immediate implication of Theorem 4.2 in the context of heterogeneous (in $\mathbf{x} \in \mathcal{X}$) causal effect estimation is that the CART estimators discussed in Section 2 can have poor performance in some regions of the covariate support, particularly near the boundaries of $\mathcal{X}$.

4.2 Past Work

Theorem 4.2 contributes to the literature in several ways. Our results indicate that when the goal is to approximate the unknown conditional expectation pointwise for all $\mathbf{x} \in \mathcal{X}$, as it is the case in the analysis of heterogeneity in causal inference settings, decision trees will exhibit extremely slow convergence rates in some regions of the support, making those methods suboptimal from an approximation perspective. The phenomenon revealed in Theorems 4.1 and 4.2 has been observed in various forms since the inception of CART [Breiman et al., 1984, Section 11]. Historically, the phenomenon characterized in Theorem 4.1 has been called the *end-cut preference*, where splits along noisy directions tend to concentrate along the end points of the parent node. More specifically, Breiman et al. [1984, Theorem 11.1] and Ishwaran [2015, Theorem 4] showed that for each $\delta \in (0, 1)$, $\mathbb{P}(\hat{\imath} \leq \delta n \text{ or } \hat{\imath} \geq (1 - \delta)n) \to 1$ as $n \to \infty$. However, unlike (9) in Theorem 4.1 which

characterizes regions of the support where the pointwise rates of estimation are slower than any *polynomial-in-n*, their result only implies rates in uniform norm slower than any *constant multiple* of the already nearly optimal rate $\sqrt{n/\log\log(n)}$, i.e., for any $C > 0$,

$$\lim_{n\to\infty} \mathbb{P}\Big(\sup_{x\in\mathcal{X}} |\hat{\mu}(T_1)(x) - \mu| \geq C\sigma n^{-1/2} \sqrt{\log\log(n)} \Big) = 1.$$

Thus, past theoretical work is not strong enough to illustrate the weaknesses of decision trees for pointwise estimation (i.e., prior lower bounds in the literature were too loose to be informative).

In accordance with Theorem 4.2, simulation results from Wager and Athey [2018, Supplement, Section B], and many others, also suggested that adaptive causal trees can have slow convergence at the boundaries of the support $\mathcal{X}$, but no formal theory supporting that numerical evidence was available in the literature until now. Tang et al. [2018] give sufficient theoretical conditions under which non-adaptive random forests (i.e., where the decision nodes are independent of the data) will be inconsistent, but those conditions do not apply to commonly used forest implementations nor are they shown to be realized by the data generating mechanism.

Bühlmann and Yu [2002] and Banerjee and McKeague [2007] showed that the minimizers $(\hat{\beta}_1, \hat{\beta}_2, \hat{\tau})$ of (2) at the root node converge to the population minimizers $(\beta_1^*, \beta_2^*, \tau^*)$ at a cube-root $n^{1/3}$ rate when the regression model (1) satisfies specific regularity assumptions. Because the decision stump (6) can be expressed as $\hat{\mu}(T_1)(x) = \hat{\beta}_1 \mathbb{1}(x \leq \hat{\tau}) + \hat{\beta}_2 \mathbb{1}(x > \hat{\tau})$, their results can be used to study the asymptotic properties of $\hat{\mu}(T_1)(x)$. Among other things, they posit that the population minimizers $(\beta_1^*, \beta_2^*, \tau^*)$ are unique and that the regression function $\mu(x)$ is continuously differentiable and has nonzero derivative at τ^*. Theorem 4.2 shows that the results in Bühlmann and Yu [2002] and Banerjee and McKeague [2007] are not uniformly valid in the sense that excluding the constant regression function from the allowed class of data generating processes is necessary for their results to hold for $x \in \mathcal{X}$.

4.3 Uniform Minimax Rates

Letting $\mathcal{P}$ be any set of data generating processes of interest that includes the location model in Assumption 1, for any $b \in (0, 1)$, we immediately obtain from (8) that

$$\liminf_{n\to\infty} \sup_{\mathbb{P}\in\mathcal{P}} \mathbb{P}\left(\sup_{x\in\mathcal{X}} |\hat{\mu}(T)(x) - \mu| \geq \sigma n^{-b/2} \sqrt{(2 + o(1)) \log\log(n)} \right) \geq (2/e)b,$$

where T is any tree constructed using conventional CART methodology with at least one split. Therefore, decision trees grown with CART methodology cannot converge faster than any polynomial-in-n, when uniformity over the full support of the data $\mathcal{X}$, and over possible data generating processes, is of interest.

4.4 Honest Trees

While Theorem 4.2 deals with depth $K = 1$ adaptive trees (i.e., the same data is used for determining the split points and terminal node output), analogous results hold for honest trees. The honest tree output is

$$\tilde{\mu}(T)(\mathbf{x}) = \frac{1}{\#\{\tilde{\mathbf{x}}_i \in \mathrm{t}\}} \sum_{\tilde{\mathbf{x}}_i \in \mathrm{t}} \tilde{y}_i, \quad \mathbf{x} \in \mathrm{t}, \tag{10}$$

where $(\tilde{y}_i, \tilde{\mathbf{x}}_i^T)$, $i = 1, 2, \ldots, n$, are independent samples from those which were used to construct the decision nodes (i.e., the partition of $\mathcal{X}$), and $n(\mathrm{t}) = \#\{\tilde{\mathbf{x}}_i \in \mathrm{t}\} > 0$. To simplify calculations, we define $\tilde{\mu}(T)(\mathbf{x}) = \mu(\mathbf{x})$ if $n(\mathrm{t}) = 0$, an event that occurs with vanishingly small probability.

Conditional on the data used to construct the partition, the honest decision stump $\tilde{\mu}(T_1)(x)$ at $x = 0$ is an average of (approximately) $\hat{\imath}$ response values, and so we expect its variance (equal to mean squared error) to be approximately $\sigma^2/\hat{\imath}$. The problem is that, according to Theorem 4.1, the split index $\hat{\imath}$ is much smaller than n, with non-vanishing probability. More rigorously, using a conditioning argument and (7), it follows that $\tilde{\mu}(T_1)(x)$ converges uniformly no faster than

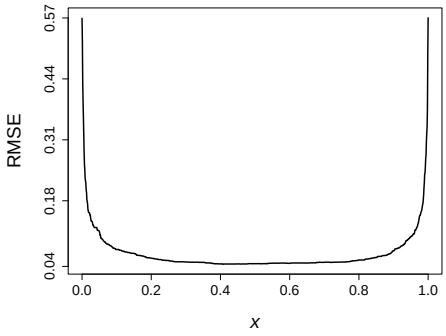
$$\mathbb{E}\left[\sup_{x \in \mathcal{X}} \left(\tilde{\mu}(T_1)(x) - \mu\right)^2\right] \geq \sigma^2 \mathbb{E}\left[\frac{(1 - 2^{-\hat{\imath}})^2}{\hat{\imath}}\right] \gtrsim \frac{\sigma^2}{n^b}, \tag{11}$$

for any $b \in (0, 1)$, and n large enough.

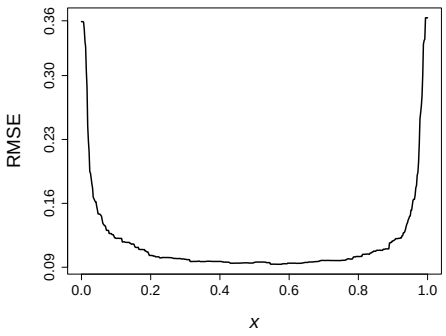
4.5 Simulation Evidence

We illustrate the implications of Theorems 4.1 and 4.2 numerically with $p = 1$. In Figure 1a, we plot the pointwise root mean squared error (RMSE) $\sqrt{\mathbb{E}[(\hat{\mu}(T_1)(x) - \mu)^2]}$, approximated by 500 replications, when $\mu = 0$, $\varepsilon_i \sim N(0, 1)$, and $n = 1000$. In Figure 1b, we consider the context of the causal model discussed in Section 2, with a constant treatment effect $\theta(x) = 1$ and $\mathbb{E}[y_i(0)] = 0$, $d_i \sim \mathrm{Bern}(0.5)$, and $\varepsilon_i \sim N(0, 1)$, again with $n = 1000$ and 500 replications. We plot the pointwise RMSE for an honest causal decision stump with output based on the regression estimator $\hat{\theta}_{\mathrm{reg}}(T_1)(x)$ constructed using the adjusted expected MSE splitting criterion proposed by Athey and Imbens [2016]. The transformed outcome tree, $\hat{\theta}_{\mathrm{ipw}}(T_1)(x)$, exhibits similar empirical behavior. Both plots corroborate with Theorem 4.2: the decision stump has smallest pointwise RMSE near the center of the covariate space, but the performance degrades as the evaluation points move closer to the boundary.

The following section investigates further the role of honesty in the construction of deeper trees, and shows an even stronger result: honest trees will be inconsistent on some (at least countably many) regions of $\mathcal{X}$ whenever the trees are grown up to depth $K \approx \log\log(n)$. In other words, shallow (honest) regression trees can be uniformly inconsistent, a result



(a) Pointwise RMSE of decision stump.



(b) Pointwise RMSE of causal decision stump.

Figure 1: Pointwise RMSE of decision stumps for location model.

that is intuitively anticipated from Theorems 4.1 and 4.2 because even after one single split there is non-trivial probability of having small cells with only a few observations, and repeating this process further down the tree can only exacerbate the issue.

The main results in this section were derived in the simplest possible case (constant regression model, $p = 1$, $K = 1$, etc.), but the main conclusions are applicable more generally. The key phenomenon captured by Theorems 4.1 and 4.2 are only exacerbated in multi-dimensional settings ($p > 1$) or for multi-level decision trees ($K > 1$). We will formalize the shortcomings associated with deeper honest trees in the next section.

5 Inconsistency with Deeper Trees

The previous section provides a pessimistic view on depth one ($K = 1$) decision trees: decision stumps can have slow convergence for the simplest regression models in some regions of $\mathcal{X}$. We now discuss formally situations where decision trees can be *inconsistent* (i.e., fail to converge) altogether, if grown only to depth $K \approx \log\log(n)$. As is customary in the literature, we will focus on trees that are honest, which are believed to offer better empirical performance [Athey and Imbens, 2016].

Definition 5.1 (Honest CART (CART+)). At each level of the tree, including the output in the terminal nodes, generate new response values $\{\tilde{y}_i : i = 1, 2, \ldots, n\}$. Each node t from the parent level is further refined by selecting a split direction and split point that minimizes the CART squared error criterion (2) with data $\{(\tilde{y}_i, \mathbf{x}_i^T) : \mathbf{x}_i \in \mathrm{t}\}$. The output of the tree T at a point $\mathbf{x}$ belonging to a terminal node t is $\tilde{\mu}(T)(\mathbf{x}) = \frac{1}{\#\{\mathbf{x}_i \in \mathrm{t}\}} \sum_{\mathbf{x}_i \in \mathrm{t}} \tilde{y}_i$.

The only difference between conventional CART and CART+ is that the split points at each level are determined using a new, statistically independent set of response values, although the input values remain the same. Importantly, the adaptive properties of the tree

are retained, as the nodes are still refined by minimizing the empirical squared error (2). For our purposes, problems will arise as soon as the depth K is approximately $\log\log(n)$ and so we expect there to be little practical difference between CART+ and the original CART algorithm when the sample size is large.

CART+ serves as a phenomenological model of conventional CART and allows us to analyze its pointwise (and uniform in $\mathbf{x} \in \mathcal{X}$) behavior. Importantly, the formulation of CART+ and Assumption 1 together ensure that the split points have a desirable Markovian property: a split point $\tilde{\tau} \in [\tau_1, \tau_2]$ conditioned on its immediate ancestor split points $\tilde{\tau}_1 = \tau_1$ and $\tilde{\tau}_2 = \tau_2$ is independent of all ancestor split points, including $\tilde{\tau}_1$ and $\tilde{\tau}_2$.

Theorem 5.2. *Suppose Assumption 1 holds and $p = 1$. Consider a maximal depth $K_n \gtrsim \log\log(n)$ tree T_{K_n} constructed with CART+ methodology. Then, there exists a positive constant C such that*

$$\liminf_{n\to\infty} \mathbb{P}\left(\sup_{x\in\mathcal{X}} |\tilde{\mu}(T_{K_n})(x) - \mu| > C \right) > 0.$$

This theorem shows that very shallow trees grown with the conventional squared error criterion can be pointwise (and hence uniform in $\mathbf{x} \in \mathcal{X}$) inconsistent. To put the iterated logarithm scaling of the depth K into perspective, if $n = 1$ billion, then $\log\log(n) \approx 3$, a typical depth seen in practice.

The pointwise error in Theorem 5.2 should be contrasted with the IMSE. Under Assumption 1, if $K \asymp \log\log(n)$, then

$$\mathbb{E}\left[\int_{\mathcal{X}} (\tilde{\mu}(T_K)(x) - \mu)^2 \mathbb{P}_x(dx) \right] \leq \frac{2^{K+1}\sigma^2}{n+1} = O\left(\frac{\sigma^2 \text{poly-log}(n)}{n} \right), \tag{12}$$

and hence by Markov's inequality,

$$\lim_{n\to\infty} \mathbb{P}\left(\int_{\mathcal{X}} (\tilde{\mu}(T_K)(x) - \mu)^2 \mathbb{P}_x(dx) > \frac{\sigma^2/n}{\text{poly-log}(n)} \right) = 0.$$

Therefore, the IMSE of the pointwise inconsistent depth $K \asymp \log\log(n)$ decision tree decays at the optimal $\sqrt{n}$ rate, up to poly-logarithmic factors. This shows that the performance of the tree varies widely depending on whether the input x is average or worst case.

The intuition for Theorem 5.2 is based similarly on Theorem 4.1, but for depth K trees constructed with CART+ methodology. That is, honest trees of depth only $K \approx \log\log(n)$ will generate cells near the boundaries of the support $\mathcal{X}$ containing a finite number of observations with probability bounded away from zero. The inequality (7) implies that, with probability bounded away from zero, the number of observations in a child node t' of a parent node t (near the boundary of $\mathcal{X}$) satisfies $n(\mathrm{t}') \leq (n(\mathrm{t}))^b$. It turns out that

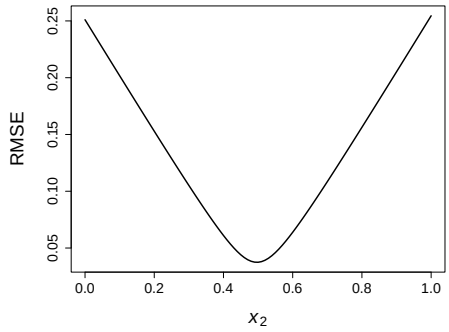
the number of times this occurs after K splits is stochastically dominated by a negative binomial random variable, providing a lower bound on the probability that a maximal depth K tree will have, at most, $n(\mathrm{t}) \leq n^{b^K}$ observations in terminal nodes near the boundary of $\mathcal{X}$. Since $b \in (0, 1)$, the bound n^{b^K} is a constant whenever K exceeds a constant multiple of $\log\log(n)$.

It is important to note that the aforementioned inconsistency of honest regression trees need not occur at the boundary of the support $\mathcal{X}$. By a symmetry argument, if $\tilde{\tau}$ is any split point that occurs at a fixed depth in the tree, then $\tilde{\mu}(T_K)(\tilde{\tau})$ will also fail to converge to μ if the tree has maximal depth $K \gtrsim \log\log(n)$. In other words, after reaching depth $J \geq 1$, inconsistency will occur at any of the (at most) $2^J + 1$ endpoints associated with the 2^J cells, whenever we grow the tree to a total depth of $J + K$ such that $K \gtrsim \log\log(n)$ as $n \to \infty$.

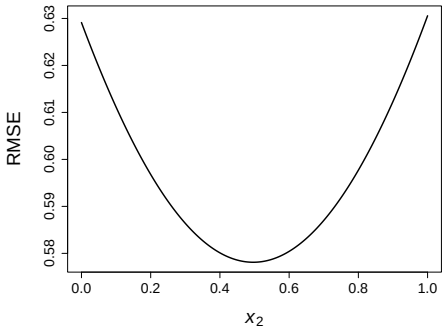
6 Pruning

Pruning is a well-established strategy for mitigating some of the ill consequences of working with trees, such as overfitting. In some cases, however, pruning will not help. Indeed, as the previous section has revealed, depth one trees can have extremely slow convergence near the boundary of the covariate space. While this phenomenon holds for location models, it can also manifest with models that have a strong dependence on the covariates. For example, if the first split at the root node is along a variable x_j such that the marginal projection $\mathbb{E}[y \mid x_j]$ is constant—resembling the location model in Assumption 1 marginally—then, according to the previous discussion, the tree will almost always produce one cell with very few observations, but no amount of pruning at lower depths will help. The *checkerboard* model [Bengio et al., 2010] in $p = 2$ dimensions is an example where y is marginally independent of both covariates. That is, if $y_i = \mathrm{sgn}(x_{i1} - 0.5)\mathrm{sgn}(x_{i2} - 0.5) + \varepsilon_i$, where $\mathbf{x}_i \sim U([0, 1]^2)$ and $\varepsilon_i \sim N(0, 1)$ are independent, then y_i given $x_{ij} = x_j$ is distributed as a symmetric two-component Gaussian mixture, free from x_j.

To illustrate the point above numerically on a model with a smooth regression function, suppose $y_i = (x_{i1} - 0.5)(x_{i2} - 0.5) + \varepsilon_i$, where $\mathbf{x}_i \sim U([0, 1]^2)$ and $\varepsilon_i \sim N(0, 1)$ are independent. As $\mathbb{E}[y_i \mid x_{ij} = x_j] = 0$ for $j = 1, 2$, the response variable has no marginal dependence on either covariate. Figure 2a displays the results of a computer experiment with $n = 1000$ and 500 replications. The plot shows the pointwise RMSE of a pruned tree T with output $\hat{\mu}(T)(\mathbf{x})$ at $\mathbf{x} = (0, x_2)$ as x_2 ranges from 0 to 1. Similarly, Figure 2b shows the result of fitting a pruned causal tree T with output $\hat{\theta}_{\mathrm{reg}}(T)(\mathbf{x})$, constructed using honesty and the adjusted expected MSE splitting criterion proposed by Athey and Imbens [2016]. The experiment consists of 500 replications from the model $y_i = d_i(x_{i1} - 0.5)(x_{i2} - 0.5) + \varepsilon_i$, where $d_i \sim \mathrm{Bin}(0.5)$, $\mathbf{x}_i \sim U([0, 1]^2)$, and $\varepsilon_i \sim N(0, 1)$ are independent, and $n = 1000$. We do not include the transformed outcome tree $\hat{\theta}_{\mathrm{ipw}}(T)(\mathbf{x})$ as it also produces a similar plot. In both cases, the numerical evidence indicates that pruning does not mitigate the lack of



(a) Pointwise RMSE for pruned tree at $\mathbf{x} = (0, x_2)^T$.



(b) Pointwise RMSE for pruned causal tree at $\mathbf{x} = (0, x_2)^T$.

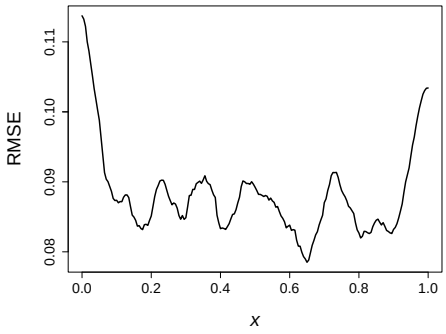
Figure 2: Pointwise RMSE of pruned trees for models where $\mathbf{x}$ and y are dependent.

uniform consistency over $\mathcal{X}$ and the poor performance near the boundary persists.

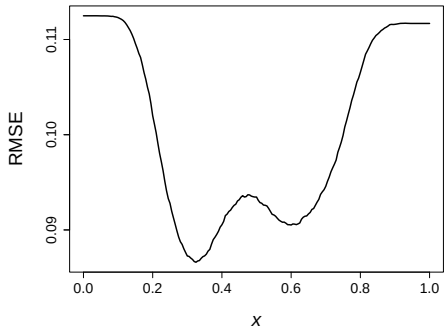
7 Random Forests

At this point, the curious reader may wonder whether ensemble learning can address some of the convergence issues with decision trees. Here we consider *honest random forests*, developed by Wager and Athey [2018]. Specifically, for each tree in the ensemble, we randomly sample a subset $S \subset \{1, 2, \ldots, n\}$ of size s and, among the data $\{(\mathbf{x}_i, y_i)\}_{i \in S}$, use half for determining the splits and the other half for estimating the conditional mean in the terminal nodes (the division of S into two equally sized subsets occurs randomly). More specifically, for each $S \subset \{1, 2, \ldots, n\}$ with $|S| = s$, let S_0 denote the portion used for determining the splits and S_1 be the portion used for estimating the conditional mean in the terminal nodes. The set of all such subsamples is denoted by $\mathcal{S} = \{S = S_1 \cup S_0 \subset \{1, 2, \ldots, n\} : S_0 \cap S_1 = \emptyset,\ |S_0| = |S_1| = s/2\}$. In addition, at each node, a particular variable is split if it yields the smallest SSE (2) among a random selection $M \subset \{1, 2, \ldots, p\}$ of $m = mtry$ candidate directions. The set of all candidate variable selections is denoted by $\mathcal{M} = \{M \subset \{1, 2, \ldots p\} : |M| = m\}$. This idea can be applied to regression trees to obtain a regression forest, or causal decision tree estimators (3) or (4) to obtain a causal forest, though, for simplicity, here we only consider the regression setting.

To get a sense of the improvement that forests offer over trees, we specialize to the case where the constituent trees in the forest are honest decision stumps (i.e., honest trees (10) with depth $K = 1$). The decision stump output $\hat{\mu}(T_1)(\mathbf{x})$ constructed in this way is denoted by $\hat{\mu}(T(M, S))(\mathbf{x})$ and the (regression) random forest output is $\hat{\mu}_B(\mathbf{x}) = B^{-1} \sum_{b=1}^{B} \hat{\mu}(T(M_b, S_b))(\mathbf{x})$, where $(M_1, S_1), (M_2, S_2), \ldots, (M_B, S_B)$ are independent copies of (M, S). When the number of trees B is large, the honest random forest can



(a) Pointwise RMSE of random forest with $s = 100$ and $m = 1$.



(b) Pointwise RMSE of causal forest with $s = 100$ and $m = 1$.

Figure 3: Pointwise RMSE of random forests for location model.

be approximated by

$$\hat{\mu}(\mathbf{x}) = \frac{1}{\binom{n}{s}\binom{s}{s/2}\binom{p}{m}} \sum_{S \in \mathcal{S}} \sum_{M \in \mathcal{M}} \hat{\mu}(T(M, S))(\mathbf{x}).$$

The next theorem provides an upper bound on its pointwise error.

Theorem 7.1. *Suppose Assumption 1 holds, and, additionally, that $x_{i1}, x_{i2}, \ldots, x_{ip}$ are independent. If $n \to \infty$, $p \to \infty$, $s = o(n^{1/3})$, and $m = o(p/s)$, then for all $\mathbf{x} \in \mathcal{X}$,*

$$\mathbb{E}[(\hat{\mu}(\mathbf{x}) - \mu)^2] \leq (\sigma^2/n)(1 + (s/2)(m/p) + o(1)).$$

This theorem showcases explicitly the effect of both subsampling and the random variable selection mechanism—each is important for reducing variance. According to past work that utilizes the Hoeffding-Serfling variance inequality for U-statistics [Wager and Athey, 2018, Bühlmann and Yu, 2002], subsampling allows us to achieve a pointwise error

$$\mathbb{E}[(\hat{\mu}(\mathbf{x}) - \mu)^2] \lesssim \sigma^2 s/n,$$

which is significantly better than the arbitrarily slow polynomial-in-n rates for individual trees (see Theorem 4.2), but still suboptimal since s is typically chosen to grow with the sample size to reduce bias when it exists. The result becomes more interesting when we account for the random variable selection mechanism, because it further reduces the error by decorrelating the constituent trees. Therefore, if the dimensionality p is large relative to s and $m = o(p/s)$, then it is possible to achieve the *exact* optimal $\sqrt{n}$ rate—a vast improvement over the $n^{b/2}$ rate for individual trees. The price paid for such improvement is the inclusion of two additional tuning parameters for implementation (s and m), and the loss of interpretability for the resulting estimates.

In Figure 3, we plot the pointwise RMSE of a regression forest and causal forest for the respective models in Section 4.5, each time using $B = 2000$ trees and the same sample sizes and number of replications as before. Compared to Figure 1a and Figure 1b, we see that random forests have considerably better performance than a single tree near the boundary.

When $p = 1$, Banerjee and McKeague [2007] and Bühlmann and Yu [2002] investigated the properties of decision trees under assumptions that rule out the location model in Assumption 1. They also showed that subsampling can reduce variance, similar to our result in Theorem 7.1. However, because the decision stump exhibits large bias in their setting, one cannot deduce from their results how random forests would improve the pointwise mean square error, which accounts for both bias and variance. Additionally, unlike Theorem 7.1, the random variable selection mechanism was not explored by Banerjee and McKeague [2007] and Bühlmann and Yu [2002] because their results are limited to the one-dimensional setting $p = 1$. As a consequence, Theorem 7.1 complements prior literature by studying the pointwise mean squared error performance of random forest under the the location model with $p \geq 1$, and thus formalizes a beneficial aspect of random feature selection for decision tree ensembles.

Finally, while Theorem 7.1 concerns a depth one ($K = 1$) random forest construction, it is possible to explore multi-level honest tree ensembles. Theorem 5.2 showed that shallow honest trees constructed with the CART+ procedure can produce pointwise inconsistent estimates of the regression function μ. In contrast, using the Hoeffding-Serfling variance inequality for U-statistics, it can be shown that an ensemble of depth $K \asymp \log\log(n)$ trees constructed with CART+ methodology on subsampled data will have pointwise error $\sqrt{\mathbb{E}[(\hat{\mu}(\mathbf{x}) - \mu)^2]} = O(\sigma\sqrt{s/n})$, for all $\mathbf{x} \in \mathcal{X}$. This result provides a concrete example where an ensemble of shallow inconsistent decision trees can be consistent with nearly optimal convergence rates, and is, to the best of our knowledge, the first time that such a result has been shown in the literature for practical trees based on CART methodology.

8 Conclusion

This article studied the delicate pointwise properties of axis-aligned recursive partitioning, focusing on heterogeneous causal effect estimation, where accurate pointwise estimates over the entire support of the covariates are essential for valid statistical learning (e.g., point estimation, testing hypotheses, confidence interval construction). Specifically, we called into question the use of causal decision trees for such purposes by demonstrating that, for a standard location model, depth one decision trees (e.g., decision stumps) constructed using CART methodology exhibit pointwise convergence rates slower than any polynomial-in-n in boundary regions of the support of the covariates. Even more dramatic, honest shallow decision trees were shown to be inconsistent even in large samples. Pruning was unable to overcome these limitations, but ensemble learning with both subsampling

and random feature selection was successful at restoring near-optimal convergence rates for pointwise estimation for the specific simple class of data generating processes that we considered. While our emphasis was on direct use of decision trees for causal effect estimation, the methodological implications are similar for multi-step semi-parametric settings, where preliminary unknown functions (e.g., propensity scores) are estimated with machine learning tools, as well as conditional quantile regression, both of which require estimators with high pointwise accuracy.

In conclusion, our results have important implications for heterogeneous prediction and causal inference learning tasks employing decision trees. Whenever the goal is to produce accurate pointwise regression estimates over the entire support of the conditioning variables, even shallow decision trees trained with a large number of samples can exhibit poor performance. Consequently, adaptive recursive partitioning should be used with caution for heterogeneous prediction or causal inference purposes, especially in high-stakes environments where high pointwise accuracy is crucial.

Acknowledgements

The authors thank Jianqing Fan, Max Farrell, Boris Hanin, Joowon Klusowski, Jantje Sönksen, and Rocio Titiunik for comments.

Funding

Cattaneo gratefully acknowledges financial support from the National Science Foundation through SES-1947805, SES-2019432, and SES-2241575. Klusowski gratefully acknowledges financial support from the National Science Foundation through CAREER DMS-2239448, DMS-2054808, and HDR TRIPODS CCF-1934924. Part of this research was conducted by Tian during his doctoral studies at Princeton University, and it constitutes a chapter in his dissertation.

Disclaimer

This document is being distributed for informational and educational purposes only and is not an offer to sell or the solicitation of an offer to buy any securities or other instruments. The information contained herein is not intended to provide, and should not be relied upon for, investment advice. The views expressed herein are not necessarily the views of Two Sigma Investments, LP or any of its affiliates (collectively, "Two Sigma"). Such views reflect the assumptions of the author(s) of the document and are subject to change without notice. The document may employ data derived from third-party sources. No representation is made by Two Sigma as to the accuracy of such information and the use

of such information in no way implies an endorsement of the source of such information or its validity.

The copyrights and/or trademarks in some of the images, logos or other material used herein may be owned by entities other than Two Sigma. If so, such copyrights and/or trademarks are most likely owned by the entity that created the material and are used purely for identification and comment as fair use under international copyright and/or trademark laws. Use of such image, copyright or trademark does not imply any association with such organization (or endorsement of such organization) by Two Sigma, nor vice versa.

A Proofs

In this appendix, we include proofs of the formal statements in the main text. Throughout the proofs below, because the quantities of interest are location invariant and homogeneous with respect to scale, by working with the standardized response variable $(y_i - \mu)/\sigma$, we can assume without loss of generality that $\mu = 0$ and $\sigma^2 = 1$.

A.1 Decision Stumps

In this section, we prove (7) in Theorem 4.1; (8) and (9) in Theorem 4.2; and (11) and (12). Throughout this section, we denote the partial sum by $S_k = y_1 + \cdots + y_k$, for $k \geq 1$.

Proof of (7) *in Theorem 4.1.* Fix $a, b \in (0, 1)$ with $a < b$. According to (5), the desired probability is

$$\begin{aligned} &\mathbb{P}(n^a \leq \hat{\imath} \leq n^b) \\ &\quad = \mathbb{P}\left(\max_{1 \leq k < n} \frac{\left(\frac{1}{\sqrt{n}} S_k - \frac{k}{n} \frac{1}{\sqrt{n}} S_n\right)^2}{(k/n)(1 - k/n)} > \max_{1 \leq k < n^a,\, n^b < k < n} \frac{\left(\frac{1}{\sqrt{n}} S_k - \frac{k}{n} \frac{1}{\sqrt{n}} S_n\right)^2}{(k/n)(1 - k/n)} \right). \end{aligned} \tag{13}$$

By Csörgö and Horváth [1997, Equation A.4.37], we can define a sequence of Brownian bridges $\{B_n(t) : 0 \leq t \leq 1\}$ on a suitable probability space such that

$$\left| \max_{1 \leq k < n} \frac{\left|\frac{1}{\sqrt{n}} S_k - \frac{k}{n} \frac{1}{\sqrt{n}} S_n\right|}{\sqrt{(k/n)(1 - k/n)}} - \sup_{1/n \leq t \leq 1 - 1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} \right| = \epsilon_n, \tag{14}$$

where $\epsilon_n = o_P((\log\log(n))^{-1/2})$. We note that while Csörgö and Horváth [1997, Equation A.4.37] bounds the approximation error of the maximum over the full range $1 \leq k < n$ as in (14), its proof, which relies on invariance principles for partial sums of i.i.d. random variables, can be generalized to bound the approximation error over $1 \leq k < n^a, n^b < k < n$. Thus,

$$\left| \max_{1 \leq k < n^a,\, n^b < k < n} \frac{\left|\frac{1}{\sqrt{n}} S_k - \frac{k}{n} \frac{1}{\sqrt{n}} S_n\right|}{\sqrt{(k/n)(1 - k/n)}} - \sup_{1/n \leq t < n^{a-1},\, n^{b-1} < t \leq 1 - 1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} \right| = \epsilon_n. \tag{15}$$

Combining the approximations (14) and (15), the probability (13) can thus be lower bounded by

$$\mathbb{P}\left(\sup_{1/n \le t \le 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} > \sup_{1/n \le t < n^{a-1},\, n^{b-1} < t \le 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} + 2\epsilon_n \right).$$

Next, we note that the standardized Brownian bridge $\{B_n(t)/\sqrt{t(1-t)} : 0 < t < 1\}$ is distributionally equivalent to a time-transformed Ornstein-Uhlenbeck (O-U) process $\{U(\log(t/(1-t))) : 0 < t < 1\}$, where $\{U(t) : t \in \mathbb{R}\}$ is a zero-mean O-U process [Csörgö and Révész, 1981, Section 1.9], and thus

$$\begin{aligned}
&\mathbb{P}\left(\sup_{1/n \le t \le 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} > \sup_{1/n \le t < n^{a-1},\, n^{b-1} < t \le 1-1/n} \frac{|B_n(t)|}{\sqrt{t(1-t)}} + 2\epsilon_n \right) \\
&= \mathbb{P}\left(\sup_{-\log(n-1) \le t \le \log(n-1)} |U(t)| > \sup_{-\log(n-1) \le t < \log(n^{a-1}/(1-n^{a-1})),\, \log(n^{b-1}/(1-n^{b-1})) < t \le \log(n-1)} |U(t)| + 2\epsilon_n \right) \\
&= \mathbb{P}\left(\sup_{0 \le t \le 2\log(n-1)} |U(t)| > \sup_{0 \le t < \log(n^{a-1}(n-1)/(1-n^{a-1})),\, \log(n^{b-1}(n-1)/(1-n^{b-1})) < t \le 2\log(n-1)} |U(t)| + 2\epsilon_n \right) \\
&= \mathbb{P}\left(\sup_{0 \le t \le 2\log(n-1)} |U(t)| > \sup_{0 \le t \le \log((n-1)^2 n^{a-b}(1-n^{b-1})/(1-n^{a-1}))} |U(t)| + 2\epsilon_n \right),
\end{aligned} \tag{16}$$

where the last two equalities result from, respectively, stationarity and the Markov property of the process $|U(t)|$, the square of which is a Cox-Ingersoll-Ross (CIR) process [Göing-Jaeschke and Yor, 2003].

By the Darling-Erdős Limit Theorem for the O-U process [Csörgö and Horváth, 1997, Theorem A.3.1] and [Eicker, 1979, Theorem 2.2], for all $c > 0$ and $z \in \mathbb{R}$, we have

$$\begin{aligned}
\lim_{n\to\infty} \mathbb{P}\left(\sup_{0 \le t \le (c+o(1))\log(n)} |U(t)| < \frac{2\log\log(n) + (1/2)\log\log\log(n) + z - (1/2)\log(\pi)}{\sqrt{2\log\log(n)}} \right) \\
= e^{-e^{-(z-2\log(c))}}.
\end{aligned} \tag{17}$$

Let z^* maximize $z \mapsto e^{-e^{-(z-2\log(2-(b-a)))}} - e^{-e^{-(z-2\log(2))}}$. Simple calculus yields

$$z^* = \log\left(\frac{v(4-v)}{\log(4/(2-v)^2)} \right), \qquad v = b - a.$$

Define

$$u_n = \frac{2\log\log(n) + (1/2)\log\log\log(n) + z^* - (1/2)\log(\pi)}{\sqrt{2\log\log(n)}}.$$

Continuing from (16), using (17) twice with $c = 2$ and $c = 2 - (b-a)$ and the fact that

$\epsilon_n = o_P((\log\log(n))^{-1/2})$, we obtain

$$
\begin{aligned}
&\liminf_{n\to\infty} \mathbb{P}\left(\sup_{0\le t\le 2\log(n-1)} |U(t)| > \sup_{0\le t\le \log((n-1)^2 n^{a-b}(1-n^{b-1})/(1-n^{a-1}))} |U(t)| + 2\epsilon_n \right) \\
&\quad \ge \liminf_{n\to\infty} \mathbb{P}\left(\sup_{0\le t\le 2\log(n-1)} |U(t)| \ge u_n, \sup_{0\le t\le \log((n-1)^2 n^{a-b}(1-n^{b-1})/(1-n^{a-1}))} |U(t)| < u_n - 2\epsilon_n \right) \\
&\quad \ge \lim_{n\to\infty} \mathbb{P}\left(\sup_{0\le t\le \log((n-1)^2 n^{a-b}(1-n^{b-1})/(1-n^{a-1}))} |U(t)| < u_n - 2\epsilon_n \right) - \lim_{n\to\infty} \mathbb{P}\left(\sup_{0\le t\le 2\log(n-1)} |U(t)| < u_n \right) \\
&\quad = \lim_{n\to\infty} \left(e^{-e^{-(z^* + o(1) - 2\log(2-(b-a)))}} + o(1) \right) - e^{-e^{-(z^* - 2\log(2))}} \\
&\quad = e^{-e^{-(z^* - 2\log(2-(b-a)))}} - e^{-e^{-(z^* - 2\log(2))}} \\
&\quad = v \frac{(4-v)(1-v/2)^{8/(v(4-v))}}{(2-v)^2} \\
&\quad \ge v \cdot \lim_{u\downarrow 0} \frac{(4-u)(1-u/2)^{8/(u(4-u))}}{(2-u)^2} \\
&\quad = (b-a)/e.
\end{aligned}
$$

We have thus shown that $\liminf_{n\to\infty} \mathbb{P}(n^a \le \hat{\imath} \le n^b) \ge (b-a)/e$. By symmetry, we obtain $\liminf_{n\to\infty} \mathbb{P}(n - n^b \le \hat{\imath} \le n - n^a) \ge (b-a)/e$, and by disjointness of the events $n^a \le \hat{\imath} \le n^b$ and $n - n^b \le \hat{\imath} \le n - n^a$ when $n > 2^{1/(1-b)}$, we also have

$$
\liminf_{n\to\infty} \mathbb{P}(n^a \le \hat{\imath} \le n^b \ \text{ or } \ n - n^b \le \hat{\imath} \le n - n^a) \ge (2/e)(b-a). \qquad \square
$$

Remark 2. *Alternatively, for any $0 \le A < B$, we have*

$$
\mathbb{P}\left(\sup_{0\le t\le B} |U(t)| > \sup_{0\le t\le A} |U(t)| \right) = \frac{B-A}{B}. \tag{18}
$$

This can readily be shown using the fact that the absolute value of a zero-mean O-U process is stationary, Markov, and has continuous paths. Consequently, ignoring the stochastic error ϵ_n from approximating the impurity gain (5) by the square of a standardized Brownian bridge (not yet justified), using (18), we can approximate the probability (13) by

$$
\begin{aligned}
&\mathbb{P}\left(\sup_{0\le t\le 2\log(n-1)} |U(t)| > \sup_{0\le t\le \log((n-1)^2 n^{a-b}(1-n^{b-1})/(1-n^{a-1}))} |U(t)| \right) \\
&\qquad = \frac{2\log(n-1) - \log((n-1)^2 n^{a-b}(1-n^{b-1})/(1-n^{a-1}))}{2\log(n-1)} \to \frac{b-a}{2}, \quad n \to \infty.
\end{aligned}
$$

We are therefore led to conjecture that

$$
\lim_{n\to\infty} \mathbb{P}(n^a \le \hat{\imath} \le n^b) = \frac{b-a}{2}.
$$

Proof of (8) *and* (9) *in Theorem 4.2.* By Csörgö and Horváth [1997, Theorem A.4.1] and Donsker's Theorem which says that $\max_{n/2<k<n} |S_k|/\sqrt{k}$ and $\max_{1\le k\le n/2} |S_n - S_k|/\sqrt{n-k}$ converge in distribution to $\sup_{0\le t\le \log(2)} |U(t)|$, we have

$$\frac{\max_{1\le k<n} \frac{|S_k|}{\sqrt{k}}}{\sqrt{2\log\log(n)}} = 1 + o_P(1), \qquad \frac{\max_{n/2<k<n} \frac{|S_k|}{\sqrt{k}} + \max_{1\le k\le n/2} \frac{|S_n - S_k|}{\sqrt{n-k}}}{\sqrt{2\log\log(n)}} = o_P(1). \tag{19}$$

To prove (8), we note that, on the event $\hat{\imath} \le n^b$ or $\hat{\imath} \ge n - n^b$ which occurs with asymptotic probability at least $2b/e$, we have

$$\begin{aligned}
\sup_{x\in\mathcal{X}} |\hat{\mu}(T_1)(x)|^2 &\ge \frac{S_{\hat{\imath}}^2}{\hat{\imath}^2}\mathbb{1}(\hat{\imath} \le n/2) + \frac{(S_n - S_{\hat{\imath}})^2}{(n-\hat{\imath})^2}\mathbb{1}(\hat{\imath} > n/2) \\
&\ge \frac{1}{\min\{\hat{\imath},\, n-\hat{\imath}\}}\left(\frac{S_{\hat{\imath}}^2}{\hat{\imath}} + \frac{(S_n - S_{\hat{\imath}})^2}{n-\hat{\imath}} - \left(\frac{S_{\hat{\imath}}^2}{\hat{\imath}}\mathbb{1}(\hat{\imath} > n/2) + \frac{(S_n - S_{\hat{\imath}})^2}{n-\hat{\imath}}\mathbb{1}(\hat{\imath} \le n/2)\right)\right) \\
&\ge \frac{1}{\min\{\hat{\imath},\, n-\hat{\imath}\}}\left(\max_{1\le k<n}\left(\frac{S_k^2}{k} + \frac{(S_n - S_k)^2}{n-k}\right) - \max_{n/2<k<n}\frac{S_k^2}{k} - \max_{1\le k\le n/2}\frac{(S_n - S_k)^2}{n-k}\right) \\
&\ge \frac{(2 + o_P(1))\log\log(n)}{\min\{\hat{\imath},\, n-\hat{\imath}\}} \\
&\ge \frac{(2 + o_P(1))\log\log(n)}{n^b}.
\end{aligned}$$

Here we have used the fact that $\hat{\imath}$ equivalently maximizes $k \mapsto S_k^2/k + (S_n - S_k)^2/(n-k) = \sum_{i=1}^n (y_i - \mu)^2 - \sum_{i=1}^n (y_i - (S_k/k)\mathbb{1}(i \le k) - ((S_n - S_k)/(n-k))\mathbb{1}(i > k))^2$ over $1 \le k < n$.

The other result (9) follows again from (19) and from the fact that $x_{(n^a)}/n^{a-1} = 1 + o_P(1)$ and $x_{(n-n^a)}/(1-n^{a-1}) = 1 + o_P(1)$. Thus, on the event $n^a \le \hat{\imath} \le n^b$ which occurs with asymptotic probability at least $(b-a)/e$, if $x_n \le (1 + o_P(1))n^{a-1} = x_{(n^a)} \le x_{(\hat{\imath})} = \hat{\tau}$, we have

$$\begin{aligned}
|\hat{\mu}(T_1)(x_n)|^2 = \frac{S_{\hat{\imath}}^2}{\hat{\imath}^2} &= \frac{1}{\hat{\imath}}\left(\frac{S_{\hat{\imath}}^2}{\hat{\imath}} + \frac{(S_n - S_{\hat{\imath}})^2}{n-\hat{\imath}} - \frac{(S_n - S_{\hat{\imath}})^2}{n-\hat{\imath}}\right) \\
&\ge \frac{1}{\hat{\imath}}\left(\max_{1\le k<n}\left(\frac{S_k^2}{k} + \frac{(S_n - S_k)^2}{n-k}\right) - \max_{1\le k\le n^b}\frac{(S_n - S_k)^2}{n-k}\right) \\
&= \frac{(2 + o_P(1))\log\log(n)}{\hat{\imath}} \\
&\ge \frac{(2 + o_P(1))\log\log(n)}{n^b}.
\end{aligned}$$

By symmetry, on the event $n - n^b \le \hat{\imath} \le n - n^a$, the same lower bound holds for $x_n > 1 - (1 + o_P(1))n^{a-1}$. □

Proof of (11). We first observe that

$$\begin{aligned}\mathbb{E}\Big[\sup_{x\in\mathcal{X}}(\tilde{\mu}(T_1)(x))^2\Big] &\geq \mathrm{Var}(\tilde{\mu}(T_1)(0)) \\ &= \mathbb{E}\left[\left(\frac{\mathbb{1}(\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} > 0)}{\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\}} \sum_{i=1}^{n} \tilde{y}_i \mathbb{1}(\tilde{x}_i \leq x_{(\hat{\imath})})\right)^2\right] \\ &= \mathbb{E}\left[\frac{\mathbb{1}(\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} > 0)}{\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\}}\right],\end{aligned}$$

where we used the independence between $\tilde{y}_i$ and $\tilde{x}_i$ and x_i, per the honest construction and Assumption 1. By the Cauchy-Schwarz inequality, we have

$$\mathbb{E}\left[\frac{\mathbb{1}(\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} > 0)}{\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\}}\right] \geq \mathbb{E}\left[\frac{(\mathbb{P}(\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} > 0 \mid \hat{\imath}))^2}{\mathbb{E}[\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} \mid \hat{\imath}]}\right]. \tag{20}$$

Again, by the honest construction and Assumption 1, we note that $\tilde{x}_i$, x_i, and $\hat{\imath}$ are mutually independent. In particular, $x_{(\hat{\imath})}$ given $\hat{\imath} = i$ is distributed Beta$(i, n - i + 1)$, allowing us to compute

$$\begin{aligned}\mathbb{P}(\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} > 0 \mid \hat{\imath}) &= 1 - \mathbb{E}[(1 - x_{(\hat{\imath})})^n \mid \hat{\imath}] \\ &= 1 - \binom{2n-\hat{\imath}}{n} \Big/ \binom{2n}{n} \\ &= 1 - \prod_{i=1}^{\hat{\imath}} \frac{n-i+1}{2n-i+1} \\ &\geq 1 - 2^{-\hat{\imath}},\end{aligned}$$

and

$$\mathbb{E}[\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} \mid \hat{\imath}] = \frac{n}{n+1}\hat{\imath}.$$

We may thus lower bound (20) via

$$\mathbb{E}\left[\frac{\left(1 - \binom{2n-\hat{\imath}}{n} \Big/ \binom{2n}{n}\right)^2}{n\hat{\imath}/(n+1)}\right] \geq \mathbb{E}\left[\frac{(1-2^{-\hat{\imath}})^2}{\hat{\imath}}\right].$$

The fact that $\mathbb{E}\left[\frac{(1-2^{-\hat{\imath}})^2}{\hat{\imath}}\right] \gtrsim n^{-b}$ follows directly from (7). □

A.2 Inconsistency with Deeper Trees

In this section, we prove Theorem 5.2. First, we define some notation related to the tree construction which will be used in the proofs. Let $\tilde{n}_k$ be the number of observations in the left-most cell (i.e., the node containing $x = 0$) at depth k and $\tilde{\imath}_k$ be the CART split index of this node, with $\tilde{n}_0 = n$ and $\tilde{\imath}_0 = \hat{\imath}$ (recall that $\hat{\imath}$ is the split index for the decision stump (6)). Then, the left-most cell at the k-th level can be expressed as $[0, x_{(\tilde{\imath}_{k-1})}]$ and $\tilde{n}_k = \tilde{\imath}_{k-1} = \#\{x_i \leq x_{(\tilde{\imath}_{k-1})}\}$.

Lemma A.1. *There exist $\delta \in (0,1)$, $c > 1$, and a positive integer M such that for any depth $k \geq 1$ and $m \geq M$, we have $\mathbb{P}(\tilde{n}_k \leq m) \geq (1-\delta) \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m) + \delta \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m^c)$.*

Proof. Observe that if m is a positive integer, then $\tilde{\imath}_{k-1} \mid \tilde{n}_{k-1} = m$ has the same distribution as $\tilde{\imath}_0 \mid \tilde{n}_0 = m$, because of the honest tree construction and Assumption 1. Therefore, we can apply (7) to obtain

$$\mathbb{P}(m^a \leq \tilde{\imath}_{k-1} \leq m^b \mid \tilde{n}_{k-1} = m) \geq \delta > 0, \tag{21}$$

for some $\delta > 0$ and sufficiently large m. Hence, by (21), we have for m sufficiently large,

$$\begin{aligned}
&\mathbb{P}(\tilde{n}_k \leq m \mid m < \tilde{n}_{k-1} \leq m^{1/b}) \\
&\quad \geq \min_{m < i \leq m^{1/b}} \mathbb{P}(i^a \leq \tilde{\imath}_{k-1} \leq i^b \mid \tilde{n}_{k-1} = i) \mathbb{P}(\tilde{n}_k \leq m \mid i^a \leq \tilde{\imath}_{k-1} \leq i^b) \\
&\quad \geq \delta \min_{m < i \leq m^{1/b}} \mathbb{P}(\tilde{n}_k \leq m \mid i^a \leq \tilde{\imath}_{k-1} \leq i^b) \\
&\quad \geq \delta \min_{m^a < i \leq m} \mathbb{P}(\tilde{n}_k \leq \tilde{\imath}_{k-1} \mid \tilde{\imath}_{k-1} = i) \\
&\quad = \delta.
\end{aligned} \tag{22}$$

Now, taking $c = 1/b$, note that (22) implies Lemma A.1 since, for m sufficiently large, we have

$$\begin{aligned}
\mathbb{P}(\tilde{n}_k \leq m) &= \mathbb{P}(\tilde{n}_k \leq m,\ \tilde{n}_{k-1} > m^c) + \mathbb{P}(\tilde{n}_k \leq m,\ \tilde{n}_{k-1} \leq m^c) \\
&\geq \mathbb{P}(\tilde{n}_k \leq m,\ \tilde{n}_{k-1} \leq m^c) \\
&= \mathbb{P}(\tilde{n}_k \leq m,\ \tilde{n}_{k-1} \leq m) + \mathbb{P}(\tilde{n}_k \leq m,\ m < \tilde{n}_{k-1} \leq m^c) \\
&\geq \mathbb{P}(\tilde{n}_{k-1} \leq m) + \delta \cdot \mathbb{P}(m < \tilde{n}_{k-1} \leq m^c) \\
&= (1-\delta) \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m) + \delta \cdot \mathbb{P}(\tilde{n}_{k-1} \leq m^c).
\end{aligned}$$

□

Next, we use Lemma A.1 to finish the proof of Theorem 5.2. The main idea is to establish that the terminal nodes in a shallow tree will be small with constant probability.

Proof of Theorem 5.2. Define $n_\ell = n^{(1/c)^\ell}$. We will show by induction that for any $k \geq 0$ and $\ell \geq 1$ such that $n_\ell \geq M$,

$$\mathbb{P}(\tilde{n}_k \leq n_\ell) \geq \sum_{k'=\ell}^{k} \binom{k'-1}{\ell-1} (1-\delta)^{k'-\ell} \delta^\ell. \tag{23}$$

The base case of $k = 0$ is trivial since $\tilde{n}_0 = n$. Now, assume that for some fixed $k \geq 1$ and any $\ell' \geq 1$ such that $n_{\ell'} \geq M$, we have

$$\mathbb{P}(\tilde{n}_{k-1} \leq n_{\ell'}) \geq \sum_{k'=\ell'}^{k-1} \binom{k'-1}{\ell'-1} (1-\delta)^{k'-\ell'} \delta^{\ell'}. \tag{24}$$

If $\ell \geq 2$, then substituting our induction hypothesis (24) with $\ell' = \ell$ and $\ell' = \ell - 1$ into Lemma A.1, we get that

$$\begin{aligned}\mathbb{P}(\tilde{n}_k \leq n_\ell) &\geq (1-\delta)\sum_{k'=\ell}^{k-1}\binom{k'-1}{\ell-1}(1-\delta)^{k'-\ell}\delta^{\ell} + \delta\sum_{k'=\ell-1}^{k-1}\binom{k'-1}{\ell-2}(1-\delta)^{k'-\ell+1}\delta^{\ell-1} \\ &= \sum_{k'=\ell}^{k}\binom{k'-1}{\ell-1}(1-\delta)^{k'-\ell}\delta^{\ell},\end{aligned}$$

where we used Pascal's identity. This completes the inductive proof of (23).

Let $X \sim \mathrm{NB}(L, \delta)$, i.e., the number of independent trials, each occurring with probability δ, until L successes. Choose

$$L = \lceil \log_c \log_c(n) - \log_c \log_c(M) - 1 \rceil \asymp \log\log(n), \quad n_L = n^{(1/c)^L} \in [M, M^c].$$

By (23) and Markov's inequality applied to the tail probability of X, we have that

$$\begin{aligned}\mathbb{P}(\tilde{n}_K \leq n_L) &\geq \sum_{k'=L}^{K}\binom{k'-1}{L-1}(1-\delta)^{k'-L}\delta^{L} \\ &= 1 - \mathbb{P}(X \geq K+1) \\ &\geq 1 - \frac{\mathbb{E}[X]}{K+1} \\ &= 1 - \frac{L}{\delta(K+1)} \\ &\geq \frac{1}{2},\end{aligned} \tag{25}$$

as long as $K \geq 2L/\delta \gtrsim \log\log(n)$. By the Paley-Zygmund inequality [Petrov, 2007] and the fact that $\mathrm{Var}(\tilde{\mu}(T_K)(0)) = \mathbb{E}[1/\tilde{n}_K] \leq 1$, we have

$$\mathbb{P}\left(|\tilde{\mu}(T_K)(0)| > \frac{\mathbb{E}[|\tilde{\mu}(T_K)(0)|]}{2}\right) \geq \frac{(\mathbb{E}[|\tilde{\mu}(T_K)(0)|])^2}{4\mathrm{Var}(\tilde{\mu}(T_K)(0))} \geq \frac{(\mathbb{E}[|\tilde{\mu}(T_K)(0)|])^2}{4}. \tag{26}$$

By the honest construction of the tree and (25), we have the lower bound

$$\begin{aligned}\mathbb{E}[|\tilde{\mu}(T_K)(0)|] &= \sum_{k=1}^{n}\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}\tilde{y}_i\right|\right]\mathbb{P}(\tilde{n}_K = k) \\ &\geq \min_{k=1,2,\ldots,\lceil n_L\rceil}\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}\tilde{y}_i\right|\right]\mathbb{P}(\tilde{n}_K \leq \lceil n_L\rceil) \\ &\geq \frac{1}{2}\min_{k=1,2,\ldots,\lceil n_L\rceil}\mathbb{E}\left[\left|\frac{1}{k}\sum_{i=1}^{k}\tilde{y}_i\right|\right].\end{aligned} \tag{27}$$

Since a sum of independent random variables is almost surely constant if and only if the individual random variables are almost surely constant, it follows that the last expression in (27) is bounded away from zero. Returning to (26) completes the proof. □

Proof of (12). Let $0 = \tilde{\imath}_0 < \tilde{\imath}_1 \leq \cdots \leq \tilde{\imath}_{2^K-1} < \tilde{\imath}_{2^K} = n$ and $0 = \tilde{\tau}_0 < \tilde{\tau}_1 \leq \cdots \leq \tilde{\tau}_{2^K-1} < \tilde{\tau}_{2^K} = 1$ denote the successive splits indices and values, respectively, at the terminal level of the tree (if a node cannot be further refined, we duplicate the split indices and values at the next level). Note that the split indices are independent of the $\tilde{y}_i$ data by the honest condition and the x_i data per Assumption 1. In particular, note that $\tilde{\tau}_k = x_{(\tilde{\imath}_k)}$ given $\tilde{\imath}_k = i$ is distributed Beta$(i, n - i + 1)$. Thus, the IMSE can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathcal{X}} (\tilde{\mu}(T_K)(x))^2 \mathbb{P}_x(dx)\right] &= \sum_{k=1}^{2^K} \mathbb{E}\left[(\tilde{\tau}_k - \tilde{\tau}_{k-1})\left(\frac{\mathbb{1}(\tilde{\imath}_k > \tilde{\imath}_{k-1})}{\tilde{\imath}_k - \tilde{\imath}_{k-1}} \sum_{i=1}^{n} \tilde{y}_i \mathbb{1}(\tilde{\tau}_{k-1} \leq x_i < \tilde{\tau}_k)\right)^2\right] \\
&= \mathbb{E}\left[\frac{\tilde{\tau}_1}{\tilde{\imath}_1}\right] + \sum_{k=2}^{2^K-1} \mathbb{E}\left[\frac{\tilde{\tau}_k - \tilde{\tau}_{k-1}}{\tilde{\imath}_k - \tilde{\imath}_{k-1}} \mathbb{1}(\tilde{\imath}_k > \tilde{\imath}_{k-1})\right] + \mathbb{E}\left[\frac{1 - \tilde{\tau}_{2^K-1}}{n - \tilde{\imath}_{2^K-1}}\right] \\
&\leq \mathbb{E}\left[\frac{1}{n+1}\right] + \sum_{k=2}^{2^K-1} \mathbb{E}\left[\frac{1}{n+1}\right] + \mathbb{E}\left[\frac{1}{n+1} \frac{n - \tilde{\imath}_{2^K-1} + 1}{n - \tilde{\imath}_{2^K-1}}\right] \\
&\leq \frac{2^{K+1}}{n+1}.
\end{aligned}$$

□

A.3 Random Forests

In this section, we prove Theorem 7.1. The following lemmas will be helpful.

Lemma A.2. *If $W \sim Bin(w, r)$, where $w \in \mathbb{N}$ and $r \in (0, 1]$, then $\mathbb{E}[\frac{1}{W+1}] \leq \frac{1}{(w+1)r}$.*

Proof. We have

$$\mathbb{E}\left[\frac{1}{W+1}\right] = \sum_{i=0}^{w} \frac{1}{i+1}\binom{w}{i} r^i (1-r)^{w-i} = \frac{1}{(w+1)r} \sum_{i=1}^{w+1} \binom{w+1}{i} r^i (1-r)^{w+1-i} \leq \frac{1}{(w+1)r}. \quad \square$$

Lemma A.3. *Let m and a be positive integers and A and A' be two independent random subsets of $\{1, 2, \ldots, m\}$ of size a. Then, $\frac{1}{\binom{m}{a}^2} \sum_{A,A'} |A \cap A'| = \frac{a^2}{m}$.*

Proof. We have

$$\mathbb{E}_{A,A'}[|A \cap A'|] = \sum_{i \in \{1,2,\ldots,m\}} \mathbb{E}[\mathbb{1}(i \in A \cap A')] = \sum_{i \in \{1,2,\ldots,m\}} \mathbb{P}(i \in A)\mathbb{P}(i \in A') = m \cdot \frac{a}{m} \cdot \frac{a}{m} = \frac{a^2}{m}. \quad \square$$

Lemma A.4. *Let (S_0, S_1) and (S_0', S_1') be two independent subsamples from the honest forest construction. Then, we have*

$$\frac{1}{\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2}\sum_{S_0,S_1}\sum_{S_0',S_1'}|S_1'\cap S_0||S_1\cap S_0'|\le\frac{s^4}{16n(n-s/2)}.$$

Proof. First, assume that S_1' and S_0 are fixed. Notice that $S_1\cap S_0'$ is disjoint from $S_1'\cup S_0$. Thus, we have

$$\begin{aligned}&\mathbb{E}[|S_1\cap S_0'|\mid S_1', S_0]\\&=\sum_{i\notin S_1'\cup S_0}\mathbb{P}(i\in S_1\cap S_0'\mid S_1', S_0)=\sum_{i\notin S_1'\cup S_0}\mathbb{P}(i\in S_1\mid S_1', S_0)\mathbb{P}(i\in S_0'\mid S_1', S_0),\\&=(n-|S_1'\cup S_0|)\left(\frac{s/2}{n-s/2}\right)^2\le\frac{s^2}{4(n-s/2)}.\end{aligned}\tag{28}$$

Combining (28) and Lemma A.3, we have

$$\begin{aligned}\frac{1}{\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2}\sum_{S_0,S_1}\sum_{S_0',S_1'}|S_1'\cap S_0||S_1\cap S_0'|&=\mathbb{E}[|S_1'\cap S_0|\cdot\mathbb{E}[|S_1\cap S_0'|\mid S_1', S_0]]\\&\le\frac{s^4}{16n(n-s/2)}.\end{aligned}$$

□

Lemma A.5. *Let (S_0, S_1) and (S_0', S_1') be two independent subsamples from the honest forest construction. Given a fixed S_1 and S_1' such that $|S_1\cap S_1'|\ge 1$, we have*

$$\frac{1}{\binom{n-s/2}{s/2}}\sum_{S_0}\frac{1}{|S_1'\backslash S_0|}-\frac{2}{s}=\frac{1}{\binom{n-s/2}{s/2}}\sum_{S_0'}\frac{1}{|S_1\backslash S_0'|}-\frac{2}{s}\le\frac{2n}{s(n-s+2)}.\tag{29}$$

Furthermore,

$$\frac{1}{\binom{n-s/2}{s/2}^2}\left(\sum_{S_0}\frac{1}{|S_1'\backslash S_0|}-\frac{2}{s}\right)\left(\sum_{S_0'}\frac{1}{|S_1\backslash S_0'|}-\frac{2}{s}\right)\le\frac{4n^2}{s^2(n-s+2)^2}.\tag{30}$$

Proof. Fix S_1 and S_1' and note that $\mathbb{P}(|S_1 \cap S_0'| = k \mid S_1, S_1') = \frac{\binom{s/2-|S_1\cap S_1'|}{k}\binom{n-s+|S_1\cap S_1'|}{s/2-k}}{\binom{n-s/2}{s/2}}$. Then,

$$\begin{aligned}
\frac{1}{\binom{n-s/2}{s/2}} \sum_{S_0'} \frac{1}{|S_1 \backslash S_0'|} &= \sum_{k=0}^{s/2-|S_1\cap S_1'|} \frac{1}{s/2-k} \mathbb{P}(|S_1 \cap S_0'| = k \mid S_1, S_1') \\
&\leq \sum_{k=0}^{s/2-|S_1\cap S_1'|} \frac{2}{s/2-k+1} \frac{\binom{s/2-|S_1\cap S_1'|}{k}\binom{n-s+|S_1\cap S_1'|}{s/2-k}}{\binom{n-s/2}{s/2}} \\
&\leq \frac{2(n-s/2+1)}{(n-s+|S_1\cap S_1'|+1)(s/2+1)} \sum_{k=0}^{s/2-|S_1\cap S_1'|} \frac{\binom{s/2-|S_1\cap S_1'|}{k}\binom{n-s+|S_1\cap S_1'|+1}{s/2-k+1}}{\binom{n-s/2+1}{s/2+1}} \\
&\leq \frac{4(n-s/2+1)}{s(n-s+2)},
\end{aligned}$$

which implies that (29) holds regardless of (S_1, S_1'). This implies (30), since $S_1 \backslash S_0'$ is conditionally independent of $S_1' \backslash S_0$ given (S_1, S_1'). □

Proof of Theorem 7.1. We use the notation $(\hat{s}(M, S_0), \hat{j}(M, S_0))$ to denote the split point and direction, respectively, for a given pair (M, S_0). First, notice that

$$\begin{aligned}
\mathbb{E}[(\hat{\mu}(\mathbf{x}))^2] &= \frac{1}{\binom{p}{m}^2\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2} \sum_{M,M'} \sum_{S,S'} \mathbb{E}[\hat{\mu}(T(M,S))(\mathbf{x})\hat{\mu}(T(M',S'))(\mathbf{x})] \\
&= \frac{1}{\binom{p}{m}^2\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2} \sum_{M,M'} \sum_{S,S'} \sum_{\substack{j\in M \\ j'\in M'}} \sum_{\substack{i\in S_1 \\ i'\in S_1'}} \mathbb{E}[LL' + LR' + RL' + RR'],
\end{aligned} \tag{31}$$

where

$$\begin{aligned}
L &= \frac{y_i \mathbb{1}(\hat{j}(M,S_0) = j)\mathbb{1}(x_{ij} \leq \hat{s}(M,S_0))\mathbb{1}(x_j \leq \hat{s}(M,S_0))}{1 + \#\{k \in S_1 \backslash \{i\} : x_{kj} \leq \hat{s}(M,S_0)\}}, \\
L' &= \frac{y_{i'} \mathbb{1}(\hat{j}(M',S_0') = j')\mathbb{1}(x_{i'j'} \leq \hat{s}(M',S_0'))\mathbb{1}(x_{j'} \leq \hat{s}(M',S_0'))}{1 + \#\{k' \in S_1' \backslash \{i'\} : x_{k'j'} \leq \hat{s}(M',S_0')\}}, \\
R &= \frac{y_i \mathbb{1}(\hat{j}(M,S_0) = j)\mathbb{1}(x_{ij} \geq \hat{s}(M,S_0))\mathbb{1}(x_j > \hat{s}(M,S_0))}{1 + \#\{k \in S_1 \backslash \{i\} : x_{kj} > \hat{s}(M,S_0)\}}, \text{ and} \\
R' &= \frac{y_{i'} \mathbb{1}(\hat{j}(M',S_0') = j')\mathbb{1}(x_{i'j'} > \hat{s}(M',S_0'))\mathbb{1}(x_{j'} \geq \hat{s}(M',S_0'))}{1 + \#\{k' \in S_1' \backslash \{i'\} : x_{k'j'} > \hat{s}(M',S_0')\}}.
\end{aligned}$$

We evaluate (31) by considering five cases on the indices $(i,\ i',\ j,\ j')$.

A.3.1 Case 1: $i \in S_1 \backslash S_0'$ and $i \neq i'$

In this case, y_i is independent of $(\{(\mathbf{x}_k, y_k) : k \in S_0 \cup S_0'\}, \{\mathbf{x}_k : k \in S_1 \cup S_1'\}, y_{i'})$ and $\mathbb{E}[y_i] = 0$, so we have that $\mathbb{E}[LL'] = \mathbb{E}[LR'] = \mathbb{E}[RL'] = \mathbb{E}[RR'] = 0$.

A.3.2 Case 2: $i' \in S_1' \backslash S_0$ and $i \neq i'$

As with Case 1, we have that $\mathbb{E}[LL'] = \mathbb{E}[LR'] = \mathbb{E}[RL'] = \mathbb{E}[RR'] = 0$.

A.3.3 Case 3: $i \in S_1 \cap S_0'$ and $i' \in S_1' \cap S_0$

By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}(\mathbb{E}[LL'])^2 &\leq \mathbb{E}\left[\frac{y_i^2 \mathbb{1}(\hat{\jmath}(M, S_0) = j)\mathbb{1}(x_{ij} \leq \hat{s}(M, S_0))\mathbb{1}(x_j \leq \hat{s}(M, S_0))}{(1 + \#\{k \in S_1 \backslash \{i\} : x_{kj} \leq \hat{s}(M, S_0)\})^2}\right] \\ &\quad \cdot \mathbb{E}\left[\frac{y_{i'}^2 \mathbb{1}(\hat{\jmath}(M', S_0') = j')\mathbb{1}(x_{i'j'} \leq \hat{s}(M', S_0'))\mathbb{1}(x_{j'} \leq \hat{s}(M', S_0'))}{(1 + \#\{k' \in S_1' \backslash \{i'\} : x_{k'j} \leq \hat{s}(M', S_0')\})^2}\right] \\ &\leq \mathbb{E}\left[\frac{\mathbb{1}(x_{ij} \leq \hat{s}(M, S_0))\mathbb{1}(\hat{\jmath}(M, S_0) = j)}{1 + \#\{k \in S_1 \backslash \{i\} : x_{kj} \leq \hat{s}(M, S_0)\}}\right] \cdot \mathbb{E}\left[\frac{\mathbb{1}(x_{i'j'} \leq \hat{s}(M', S_0'))\mathbb{1}(\hat{\jmath}(M', S_0') = j')}{1 + \#\{k' \in S_1' \backslash \{i'\} : x_{k'j} \leq \hat{s}(M', S_0')\}}\right], \end{aligned} \tag{32}$$

where we used the fact that y_i is independent of $(\{\mathbf{x}_{k'} : k' \in S_1\}, \hat{s}(M, S_0), \hat{\jmath}(M, S_0))$ and $y_{i'}$ is independent of $(\{\mathbf{x}_{k'} : k' \in S_1'\}, \hat{s}(M', S_0'), \hat{\jmath}(M', S_0'))$. Now, since $(\{x_{kj} : k \in S_1\}, \hat{\jmath}(M, S_0))$ is independent of $\hat{s}(M, S_0)$ and $(\{x_{k'j'} : k' \in S_1'\}, \hat{\jmath}(M', S_0'))$ is independent of $\hat{s}(M', S_0')$, by applying Lemma A.2 to (32), we have

$$\mathbb{E}[LL'] \leq \frac{2}{s}\sqrt{\mathbb{P}(\hat{\jmath}(M, S_0) = j)\mathbb{P}(\hat{\jmath}(M', S_0') = j')}.$$

By symmetry, we have that

$$\mathbb{E}[LL' + LR' + RL' + RR'] \leq \frac{8}{s}\sqrt{\mathbb{P}(\hat{\jmath}(M, S_0) = j)\mathbb{P}(\hat{\jmath}(M', S_0') = j')}.$$

Therefore, by the Cauchy-Schwarz inequality,

$$\begin{aligned}&\sum_{\substack{j \in M \\ j' \in M'}} \sum_{\substack{i \in S_1 \cap S_0' \\ i' \in S_1' \cap S_0}} \mathbb{E}[LL' + LR' + RL' + RR'] \\ &\leq \frac{8|S_1 \cap S_0'||S_1' \cap S_0|}{s} \sum_{\substack{j \in M \\ j' \in M'}} \sqrt{\mathbb{P}(\hat{\jmath}(M, S_0) = j)\mathbb{P}(\hat{\jmath}(M', S_0') = j')} \\ &\leq \frac{8|S_1 \cap S_0'||S_1' \cap S_0|}{s} \sqrt{\sum_{j \in M} \mathbb{P}(\hat{\jmath}(M, S_0) = j) \sum_{j' \in M'} \mathbb{P}(\hat{\jmath}(M', S_0') = j')} \\ &= \frac{8|S_1 \cap S_0'||S_1' \cap S_0|}{s},\end{aligned}$$

so that, by Lemma A.4, we have

$$\frac{1}{\binom{p}{m}^2\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2} \sum_{M, M'} \sum_{S, S'} \sum_{\substack{j \in M \\ j' \in M'}} \sum_{\substack{i \in S_1 \cap S_0' \\ i' \in S_1' \cap S_0}} \mathbb{E}[LL' + LR' + RL' + RR'] \leq \frac{s^3}{2n(n - s/2)}.$$

A.3.4 Case 4: $j = j' \in M \cap M'$ and $i = i'$

In this case, $i \in S_1 \cap S_1'$ is not in S_0 or S_0' so $y_i y_{i'} = y_i^2$ is independent of $(\{\mathbf{x}_k : k \in S_1\}, \hat{s}(M, S_0), \hat{j}(M, S_0), \hat{s}(M', S_0'))$ and $\mathbb{E}[y_i^2] = 1$. Therefore,

$$\begin{aligned}\mathbb{E}[LL'] &\le \mathbb{E}\left[\frac{\mathbb{1}(\hat{j}(M, S_0) = j)\mathbb{1}(x_{ij} \le \hat{s}(M, S_0))\mathbb{1}(x_j \le \hat{s}(M, S_0))\mathbb{1}(x_j \le \hat{s}(M', S_0'))}{1 + \#\{k \in S_1 \backslash \{i\} : x_{kj} \le \hat{s}(M, S_0)\}}\right] \\ &\le \frac{\mathbb{P}(\hat{j}(M, S_0) = j,\ x_j \le \hat{s}(M, S_0),\ \text{and } x_j \le \hat{s}(M', S_0'))}{s/2},\end{aligned}$$

where we similarly applied Lemma A.2. By symmetry, we have

$$\begin{aligned}&\sum_{j=j' \in M \cap M'} \sum_{i=i' \in S_1 \cap S_1'} \mathbb{E}[LL' + LR' + RL' + RR'] \\ &\le \sum_{j=j' \in M \cap M'} \sum_{i=i' \in S_1 \cap S_1'} \frac{\mathbb{P}(\hat{j}(M, S_0) = j)}{s/2} \le \frac{2|S_1 \cap S_1'||M \cap M'|}{sm}.\end{aligned}$$

Applying Lemma A.3 twice, we see that

$$\frac{1}{\binom{p}{m}^2 \binom{n}{s/2}^2 \binom{n-s/2}{s/2}^2} \sum_{M, M'} \sum_{S, S'} \sum_{j=j' \in M \cap M'} \sum_{i=i' \in S_1 \cap S_1'} \mathbb{E}[LL' + LR' + RL' + RR'] = \frac{sm}{2np}.$$

A.3.5 Case 5: $j \neq j'$ and $i = i'$

If $j \notin M'$, then $\#\{x_{kj} : k \in S_1 \backslash \{i\}\}$ is independent of $(y_i, \hat{s}(M, S_0), \{\hat{j}(M, S_0) = j\}, L', y_i)$. Otherwise $\#\{x_{kj} : k \in S_1 \backslash \{S_0' \cup i\}\}$ (which is less than $\#\{x_{kj} : k \in S_1 \backslash \{i\}\}$) is independent of $(\hat{s}(M, S_0), \{\hat{j}(M, S_0) = j\}, L')$. Therefore, by applying Lemma A.2, we have

$$\begin{aligned}&\mathbb{E}[L \mid y_i, \hat{s}(M, S_0), \{\hat{j}(M, S_0) = j\}, L'] \\ &\le y_i \mathbb{1}(\hat{j}(M, S_0) = j)\left(\frac{\mathbb{1}(j \notin M')}{s/2} + \frac{\mathbb{1}(j \in M')}{|S_1 \backslash S_0'|}\right)\mathbb{1}(x_j \le \hat{s}(M, S_0)).\end{aligned}$$

Similarly, we also have

$$\begin{aligned}&\mathbb{E}[L' \mid y_i, \{\hat{j}(M, S_0) = j\}, \hat{s}(M', S_0'), \{\hat{j}(M', S_0') = j'\}] \\ &\le y_i \mathbb{1}(\hat{j}(M', S_0') = j')\left(\frac{\mathbb{1}(j' \notin M)}{s/2} + \frac{\mathbb{1}(j' \in M)}{|S_1' \backslash S_0|}\right)\mathbb{1}(x_{j'} \le \hat{s}(M', S_0')).\end{aligned}$$

Therefore, we have

$$\begin{aligned}\mathbb{E}[LL'] \le \mathbb{P}(\hat{j}(M, S_0) = j,\ \hat{j}(M', S_0') = j',\ x_j \le \hat{s}(M, S_0),\ \text{and } x_{j'} \le \hat{s}(M', S_0')) \\ \cdot \left(\frac{\mathbb{1}(j \notin M')}{s/2} + \frac{\mathbb{1}(j \in M')}{|S_1 \backslash S_0'|}\right)\left(\frac{\mathbb{1}(j' \notin M)}{s/2} + \frac{\mathbb{1}(j' \in M)}{|S_1 \backslash S_0'|}\right),\end{aligned}$$

where we used the fact that y_i^2 is independent of the data indices in $S_0 \cup S_0'$, for $i = i' \in S_1 \cap S_1'$, and $\mathbb{E}[y_i^2] = 1$. By symmetry, we have

$$
\begin{aligned}
&\sum_{\substack{j\in M\\ j\in M'}} \sum_{i\in S_1\cap S_1'} \mathbb{E}[LL' + LR' + RL' + RR'] \\
&\leq \sum_{\substack{j\in M\\ j\in M'}} \sum_{i\in S_1\cap S_1'} \mathbb{P}(\hat{j}(M, S_0) = j,\ \hat{j}(M', S_0') = j')\left(\frac{\mathbb{1}(j \notin M')}{s/2} + \frac{\mathbb{1}(j \in M')}{|S_1 \backslash S_0'|}\right)\left(\frac{\mathbb{1}(j' \notin M)}{s/2} + \frac{\mathbb{1}(j' \in M)}{|S_1 \backslash S_0'|}\right) \\
&\leq \frac{|S_1 \cap S_1'|}{m^2}\left(\frac{m - |M \cap M'|}{s/2} + \frac{|M \cap M'|}{|S_1 \backslash S_0'|}\right)\left(\frac{m - |M \cap M'|}{s/2} + \frac{|M \cap M'|}{|S_1' \backslash S_0|}\right) \\
&\leq |S_1 \cap S_1'|\left(\frac{4}{s^2} + \frac{2|M \cap M'|}{sm}\left(\frac{1}{|S_1 \backslash S_0'|} + \frac{1}{|S_1' \backslash S_0|} - \frac{4}{s}\right)\right. \\
&\qquad \left. + \frac{|M \cap M'|}{m}\left(\frac{1}{|S_1 \backslash S_0'|} - \frac{2}{s}\right)\left(\frac{1}{|S_1' \backslash S_0|} - \frac{2}{s}\right)\right).
\end{aligned} \tag{33}
$$

Since $i \in S_1 \cap S_1'$, we have $|S_1 \cap S_1'| \geq 1$, so by (33) and Lemma A.5, we have

$$
\begin{aligned}
&\frac{\sum_{M,M'} \sum_{S,S'} \sum_{j\neq j'} \sum_{i=i'} \mathbb{E}[LL' + LR' + RL' + RR']}{\binom{p}{m}^2 \binom{n}{s/2}^2 \binom{n-s/2}{s/2}^2} \\
&\leq \frac{\sum_{M,M'} \sum_{S_1,S_1'} |S_1 \cap S_1'|\left(\frac{4}{s^2} + \frac{8n|M\cap M'|}{s^2(n-s+2)m} + \frac{4n^2|M\cap M'|}{s^2(n-s+2)^2 m}\right)}{\binom{p}{m}^2 \binom{n}{s/2}^2} \\
&\leq \frac{1}{n} + \frac{2m}{(n-s+2)p} + \frac{nm}{(n-s+2)^2 p} \\
&\leq \frac{1}{n}\left(1 + \frac{3m}{p}\left(\frac{n}{n-s+2}\right)^2\right),
\end{aligned}
$$

where we applied Lemma A.3 in the second inequality. Combining Cases 1-5, we have thus shown that

$$
\mathbb{E}[(\hat{\mu}(\mathbf{x}))^2] \leq \frac{1}{n}\left(1 + \frac{sm}{2p} + \frac{3m}{p}\left(\frac{n}{n-s+2}\right)^2 + \frac{s^3}{2(n-s/2)}\right). \qquad \square
$$

References

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016. doi: 10.1073/pnas.1510489113. URL https://www.pnas.org/doi/abs/10.1073/pnas.1510489113.

Susan Athey and Guido W Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11(1):685–725, 2019. doi: 10.1146/annurev-economics-080217-053433. URL https://doi.org/10.1146/annurev-economics-080217-053433.

Moulinath Banerjee and Ian W. McKeague. Confidence sets for split points in decision trees. *The Annals of Statistics*, 35(2):543 – 574, 2007. doi: 10.1214/009053606000001415. URL https://doi.org/10.1214/009053606000001415.

Merle Behr, Yu Wang, Xiao Li, and Bin Yu. Provable boolean interaction recovery from tree ensemble obtained via random forests. *Proceedings of the National Academy of Sciences*, 119(22):e2118636119, 2022. doi: 10.1073/pnas.2118636119. URL https://www.pnas.org/doi/abs/10.1073/pnas.2118636119.

Yoshua Bengio, Olivier Delalleau, and Clarence Simard. Decision trees do not generalize to new variations. *Computational Intelligence*, 26(4):449–467, 2010. doi: https://doi.org/10.1111/j.1467-8640.2010.00366.x. URL https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-8640.2010.00366.x.

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. *The Annals of Statistics*, 30(4):927 – 961, 2002. doi: 10.1214/aos/1031689014. URL https://doi.org/10.1214/aos/1031689014.

Victor Chernozhukov, Juan Carlos Escanciano, Hidehiko Ichimura, Whitney K. Newey, and James M. Robins. Locally robust semiparametric estimation. *Econometrica*, 90(4):1501–1535, 2022. doi: https://doi.org/10.3982/ECTA16294. URL https://onlinelibrary.wiley.com/doi/abs/10.3982/ECTA16294.

M. Csörgö and L. Horváth. *Limit Theorems in Change-Point Analysis*. Wiley Series in Probability and Statistics. Wiley, 1997. ISBN 9780471955221. URL https://books.google.co.kr/books?id=iyXvAAAAMAAJ.

M. Csörgö and P. Révész. *Strong Approximations in Probability and Statistics*. Probability and Mathematical Statistics : a series of monographs and textbooks. Academic Press, 1981. ISBN 9780121985400. URL https://books.google.co.kr/books?id=sw2oAAAAIAAJ.

F. Eicker. The Asymptotic Distribution of the Suprema of the Standardized Empirical Processes. *The Annals of Statistics*, 7(1):116 – 138, 1979. doi: 10.1214/aos/1176344559. URL https://doi.org/10.1214/aos/1176344559.

Jianqing Fan and Irene Gijbels. *Local polynomial modelling and its applications*, volume 66 of *Monographs on Statistics and Applied Probability*. Chapman and Hall, 1996.

Max H Farrell, Tengyuan Liang, and Sanjog Misra. Deep neural networks for estimation and inference. *Econometrica*, 89(1):181–213, 2021. doi: https://doi.org/10.3982/ECTA16901.

Anja Göing-Jaeschke and Marc Yor. A survey and some generalizations of Bessel processes. *Bernoulli*, 9(2):313 – 349, 2003. doi: 10.3150/bj/1068128980. URL https://doi.org/10.3150/bj/1068128980.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media, 2002.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The elements of statistical learning*. Springer Series in Statistics. Springer-Verlag, New York, 2009.

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

Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(35):983–999, 2006. URL http://jmlr.org/papers/v7/meinshausen06a.html.

W. James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019. doi: 10.1073/pnas.1900654116. URL https://www.pnas.org/doi/abs/10.1073/pnas.1900654116.

Valentin V. Petrov. On lower bounds for tail probabilities. *Journal of Statistical Planning and Inference*, 137(8):2703–2705, 2007. ISSN 0378-3758. doi: https://doi.org/10.1016/j.jspi.2006.02.015. URL https://www.sciencedirect.com/science/article/pii/S0378375807000213. 5th St. Petersburg Workshop on Simulation.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5): 206–215, 2019.

Xiaogang Su, Chih-Ling Tsai, Hansheng Wang, David M. Nickerson, and Bogong Li. Subgroup analysis via recursive partitioning. *Journal of Machine Learning Research*, 10(5):141–158, 2009. URL http://jmlr.org/papers/v10/su09a.html.

Cheng Tang, Damien Garreau, and Ulrike von Luxburg. When do random forests fail? In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper/2018/file/204da255aea2cd4a75ace6018fad6b4d-Paper.pdf.

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. doi: 10.1080/01621459.2017.1319839. URL https://doi.org/10.1080/01621459.2017.1319839.

Liuyi Yao, Zhixuan Chu, Sheng Li, Yaliang Li, Jing Gao, and Aidong Zhang. A survey on causal inference. *ACM Trans. Knowl. Discov. Data*, 15(5), may 2021. ISSN 1556-4681. doi: 10.1145/3444944. URL https://doi.org/10.1145/3444944.