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

Matias D. Cattaneo[†] Jason M. Klusowski[†] Peter M. Tian[†]

July 9, 2023

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, even with pruning. Instead, the convergence may be poly-logarithmic 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.

[*] We thank Jianqing Fan, Max Farrell, Boris Hanin, Joowon Klusowski, and Rocio Titiunik for comments. Cattaneo gratefully acknowledges financial support from the National Science Foundation through grants SES-1947805, SES-2019432, and SES-2241575. Klusowski gratefully acknowledges financial support from the National Science Foundation through grants DMS-2054808 and CCF-1934924.

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

Contents

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, Singh, Kumbier, Abbasi-Asl, and Yu, 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, Tibshirani, and Friedman (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 will exhibit, at best, poly-logarithmic rates of convergence, as a function of the sample size n, uniformly over the entire support of the covariates. Furthermore, when adding sample-splitting 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, Friedman, Olshen, and Stone, 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)

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

with respect to the child node outputs, split point, and split direction, $(\beta_1, \beta_2, \tau, j)$, with $\mathbf{1}(\cdot)$ 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 $\mathrm{t}_L = \{\mathbf{x} \in \mathrm{t} : x_{\hat{\jmath}} \le \hat{\tau}\}$ and $\mathrm{t}_R = \{\mathbf{x} \in \mathrm{t} : x_{\hat{\jmath}} > \hat{\tau}\}$. The normal equations imply that $\hat{\beta}_1 = \overline{y}_{\mathrm{t}_L} = \frac{1}{\#\{\mathbf{x}_i \in \mathrm{t}_L\}} \sum_{\mathbf{x}_i \in \mathrm{t}_L} y_i$ and $\hat{\beta}_2 = \overline{y}_{\mathrm{t}_R} = \frac{1}{\#\{\mathbf{x}_i \in \mathrm{t}_R\}} \sum_{\mathbf{x}_i \in \mathrm{t}_R} y_i$, 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 $\overline{y}_{\mathrm{t}}$, i.e., if T is a decision tree, then $\hat{\mu}(T)(\mathbf{x}) = \overline{y}_{\mathrm{t}} = \frac{1}{\#\{\mathbf{x}_i \in \mathrm{t}\}} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i$. 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.

Finally, 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, Escanciano, Ichimura, Newey, and Robins, 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, Liang, and Misra, 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, Chu, Li, Li, Gao, and Zhang, 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, owing 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, Tsai, Wang, Nickerson, and Li, 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

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

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,

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

or

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

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 a SSE criterion. This is precisely because $\hat{\theta}_{\mathrm{reg}}(T)(\mathbf{x})$ and $\hat{\theta}_{\mathrm{ipw}}(T)(\mathbf{x})$ can be implemented using conventional CART methodology. That is, we implement $\hat{\theta}_{\mathrm{reg}}(T)(\mathbf{x})$ 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 $\hat{\theta}_{\mathrm{ipw}}(T)(\mathbf{x})$. Yet another (more principled) approach (Athey and Imbens, 2016) implements $\hat{\theta}_{\mathrm{reg}}(T)(\mathbf{x})$ by growing a decision tree using a slightly modified version of the SSE criterion (2) (referred to as *adjusted expected MSE*) that directly targets the conditional ATE, together with sample-splitting, where different samples are used for constructing the partition and estimating the effects of each subpopulation (a procedure referred to as *honest*).

Our theory implies that, for a constant treatment effect model, the aforementioned causal decision tree estimators will exhibit, at best, poly-logarithmic rates of convergence. 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 $\mathbf{x} \in \mathcal{X}$, even in settings with very large sample sizes. 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. $\mathbb{E}[|\varepsilon_i|^{2+\nu}] < \infty$ *for some* $\nu > 0$.

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:

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

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, Kohler, Krzyzak, and Walk (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 functions 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. 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 $j = 1, 2, \dots, p$, the data $\{x_{ij} : \mathbf{x}_i \in \mathrm{t}\}$ is relabeled so that x_{ij} is increasing in the index $i = 1, 2, \dots, n(\mathrm{t})$, where $n(\mathrm{t}) = \#\{\mathbf{x}_i \in \mathrm{t}\}$. Then, minimization of the objective (2) can be equivalently recast as maximizing the so-called *impurity gain*:

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

with respect to the index i and variable j. 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

$$\hat\mu(T_1)(\mathbf{x}) = \hat\beta_1 \mathbf{1}(x_{\hat\jmath} \le \hat\tau) + \hat\beta_2 \mathbf{1}(x_{\hat\jmath} > \hat\tau) = \begin{cases} \frac{1}{\#\{\mathbf{x}_i : x_{i\hat\jmath} \le x_{\hat\imath\hat\jmath}\}} \sum_{\mathbf{x}_i : x_{i\hat\jmath} \le x_{\hat\imath\hat\jmath}} y_i, & x_{\hat\jmath} \le x_{\hat\imath\hat\jmath} \\ \frac{1}{\#\{\mathbf{x}_i : x_{i\hat\jmath} > x_{\hat\imath\hat\jmath}\}} \sum_{\mathbf{x}_i : x_{i\hat\jmath} > x_{\hat\imath\hat\jmath}} y_i, & x_{\hat\jmath} > x_{\hat\imath\hat\jmath} \end{cases}, \tag{6}$$

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

The following theorem characterizes the location of the CART split index $\hat\imath$ at the root node.

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 $\delta \in (0, 1)$, there exist $\gamma = \gamma(\delta) \in (0, 1)$ satisfying (17), and a positive integer $N = N(\delta)$ such that, for all $n \ge N$,*

$$\mathbb{P}\Big(n^{\gamma^2} \le \hat\imath \le n^{\gamma} \ \ or \ \ n - n^{\gamma} \le \hat\imath \le n - n^{\gamma^2}\Big) \ge 1 - \delta. \tag{7}$$

Furthermore, there exist positive constants a, b, A, and B, and a positive integer $N = N(a, b, A, B)$ such that, for all $n \ge N$,

$$\mathbb{P}(\hat\imath \le A \log^a(n) \ \ or \ \ \hat\imath \ge n - A \log^a(n)) \ge B(\log(n))^{-b}. \tag{8}$$

This theorem formally characterizes the regions of the support $\mathcal{X}$ where the first CART split index $\hat\imath$, at the root node, is most likely to realize. 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 $\hat\imath = \#\{\mathbf{x}_i : x_{ij} \le \hat\tau\}$).

First, Theorem 4.1 shows that with arbitrary high probability, $\hat\imath$ will concentrate near its extremes, from the beginning of any tree construction. The slow 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, Wang, Li, and Yu, 2022, and references therein). The slow convergence rates for the decision stump occur because the optimal split point concentrates near the boundary of the support (Ishwaran, 2015), i.e., $\hat{\tau} \approx 0$ or $\hat{\tau} \approx 1$, 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 high $(1 - \delta)$ probability; for some $\gamma \in (0, 1)$, either $n^{\gamma^2} \leq n(\mathrm{t}_L) \leq n^{\gamma}$ or $n^{\gamma^2} \leq n(\mathrm{t}_R) \leq n^{\gamma}$ for large n.

Second, and more importantly for our purposes, Theorem 4.1 shows that just the first split of a decision tree construction can generate a cell containing, at most, $\log^a(n)$ observations, with probability at least $(\log(n))^{-b}$, up to a constant factor. It will follow from this result that, on the event considered in (8), 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 slow poly-logarithmic rates, for some $\mathbf{x} \in \mathcal{X}$.

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. We employ a generalization (Berkes and Weber, 2006, Equation (3)) of the classic Darling-Erdös limit law for the maximum of normalized sums of i.i.d. mean zero random variables (Darling and Erdös, 1956), the proof of which relies on the scaling property of Brownian motion. That is, under Assumption 1, for any non-decreasing function $1 \leq h(m) \leq m$ for which $\lim_{m\to\infty} h(m) = \infty$ and any $w \in \mathbb{R}$,

$$\mathbb{P}\left(\max_{m/h(m) \leq i \leq m} \left| \frac{1}{\sqrt{i}} \sum_{l=1}^{i} (y_l - \mu) \right| < \lambda(h(m), w) \right) \to \exp(-\exp(-w)), \tag{9}$$

as $m \to \infty$, where

$$\lambda(h(m), w) = \sqrt{2\sigma^2 \log\log h(m)} + \frac{\sigma \log\log\log h(m)}{2\sqrt{2\log\log h(m)}} + \frac{\sigma(w - \log(\sqrt{\pi}))}{\sqrt{2\log\log h(m)}}.$$

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 each $\delta \in (0, 1)$, there exist $\gamma = \gamma(\delta) \in (0, 1)$ satisfying (17), positive constants $C = C(\delta)$ and $D = D(\delta)$, and a positive integer $N = N(\delta)$ such that, for all $n \geq N$,

$$\mathbb{P}\left(\sup_{x \in \mathcal{X}} |\hat{\mu}(T_1)(x) - \mu| \geq C\sigma n^{-\gamma/2} \sqrt{\log\log(n)} \right) \geq 1 - \delta, \tag{10}$$

and

$$\mathbb{P}\Big(|\hat{\mu}(T_1)(x) - \mu| \geq C\sigma n^{-\gamma/2}\sqrt{\log\log(n)}\Big) \geq 1/2 - \delta, \tag{11}$$

for all $x \in [0, Dn^{\gamma^2-1}) \cup (1 - Dn^{\gamma^2-1}, 1]$.

Furthermore, there exist positive constants c, d, C, *and* D, *and a positive integer* $N = N(c, d, C, D)$ *such that, for all* $n \geq N$,

$$\mathbb{P}\Big(|\hat{\mu}(T_1)(x) - \mu| \geq C\sigma(\log(n))^{-c}\Big) \geq D(\log(n))^{-d}, \tag{12}$$

for all $x \in \{0, 1\}$.

The theorem above shows that decision stumps can have, at most, $n^{\gamma/2}$ (suboptimal) convergence for evaluation points that are within n^{γ^2-1} distance from the boundary of $\mathcal{X}$ (see (11)), for some $\gamma \in (0, 1)$, and, at most, poly-logarithmic convergence at the boundaries of the covariate space (see (12)). 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, Friedman, Olshen, and Stone, 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, Friedman, Olshen, and Stone (1984, Theorem 11.1) and Ishwaran (2015, Theorem 4) showed that for each $\epsilon \in (0, 1)$, $\mathbb{P}(\hat{\imath} \leq \epsilon n \text{ or } \hat{\imath} \geq (1-\epsilon)n) \to 1$ as $n \to \infty$, but, unlike our Theorem 4.1, this result is too weak to imply rates of estimation slower than the optimal $\sqrt{n}$ rate. 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, Garreau, and von Luxburg (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) show 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 $\hat{\mu}(T_1)(x) = \hat{\beta}_1\mathbf{1}(x \leq \hat{\tau}) + \hat{\beta}_2\mathbf{1}(x > \hat{\tau})$, 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, Markov's inequality applied to (12) from Theorem 4.2 yields

$$\sup_{\mathbb{P}\in\mathcal{P}} \mathbb{E}\left[\sup_{x\in\mathcal{X}} (\hat{\mu}(T)(x) - \mu(x))^2\right] \gtrsim \sigma^2 (\log n)^{-(2c+d)},$$

where T is any tree constructed using conventional CART methodology with at least one split. Therefore, decision trees grown with CART methodology exhibit, at best, poly-logarithmic uniform convergence rates, when uniformity over the full support of the data $\mathcal{X}$, and over possible data generating processes, is of interest.

4.4 Higher Dimensions

The high probability bound (10) in Theorem 4.2 continues to hold in higher dimensions ($p > 1$). For example, suppose that $x_{i1}, x_{i2}, \ldots, x_{ip}$ are independent, which, by symmetry, implies the optimal splitting direction $\hat{\jmath}$ satisfies $\mathbb{P}(\hat{\jmath} = j) = 1/p$ for all j. Then, granting Assumption 1, for each $\delta \in (0, 1)$, there exists $\gamma \in (0, 1)$, a constant $C = C(\delta)$, and a positive integer $N = N(\delta)$ such that, for all $n \geq N$,

$$\mathbb{P}\left(\sup_{\mathbf{x}\in\mathcal{X}} |\hat{\mu}(T_1)(\mathbf{x}) - \mu| \geq C\sigma n^{-\gamma/2}\sqrt{\log\log(n)}\right) \geq 1 - \delta. \tag{13}$$

Thus, the decision stump in multiple dimensions *cannot* converge at the optimal $\sqrt{n}$ rate.

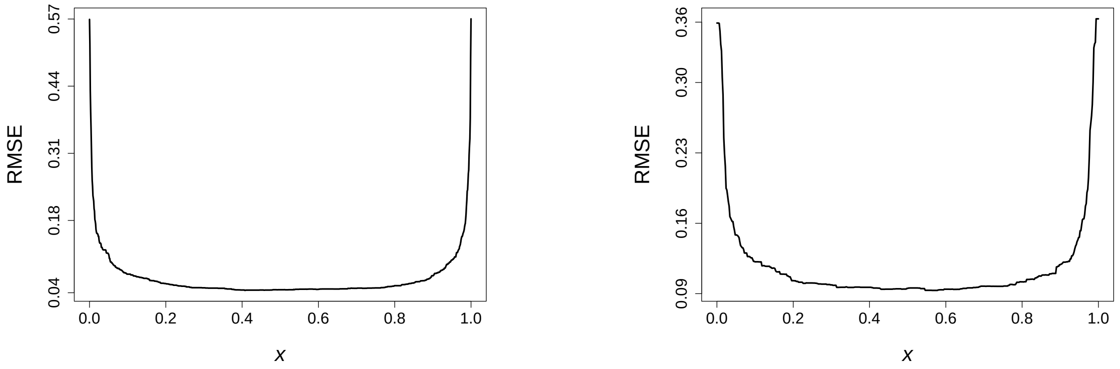
4.5 Honest Trees

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

$$\tilde{\mu}(T)(\mathbf{x}) = \frac{1}{\#\{\tilde{\mathbf{x}}_i \in \mathrm{t}\}} \sum_{\tilde{\mathbf{x}}_i\in\mathrm{t}} \tilde{y}_i, \quad \mathbf{x} \in \mathrm{t}, \tag{14}$$

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



(a) *Pointwise RMSE of decision stump.*

(b) *Pointwise RMSE of causal decision stump.*

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

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 high probability. More rigorously, using a conditioning argument and (8), it follows that $\tilde{\mu}(T_1)(x)$ converges uniformly no faster than

$$\mathbb{E}\Big[\sup_{x\in\mathcal{X}}(\tilde{\mu}(T_1)(x)-\mu)^2\Big] \geq \sigma^2\mathbb{E}\Big[\frac{(1-2^{-\hat{\imath}})^2}{\hat{\imath}}\Big] \gtrsim \sigma^2(\log(n))^{-(a+b)}. \tag{15}$$

4.6 Simulation Evidence

We illustrate the implications of Theorems 4.1 and 4.2 numerically with $p = 1$. In Fig. 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 Fig. 1b, we consider the context of the causal model discussed in Section 2, with a constant treatment effect $\theta(x) = 1$ and $\mathbb{E}[y_i(0)] = 0$, $d_i \sim \text{Bern}(0.5)$, and $\varepsilon_i \sim N(0,1)$. We plot the pointwise RMSE for an honest causal decision stump with output based on the regression estimator $\hat{\theta}_{\text{reg}}(T_1)(x)$ constructed using the adjusted expected MSE splitting criterion proposed by Athey and Imbens (2016). The transformed outcome tree, $\hat{\theta}_{\text{ipw}}(T_1)(x)$, 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 sample-splitting (i.e., 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 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 already demonstrated this fact for multi-dimensional covariates in (13), and 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 (at best, poly-logarithmic) 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 it is customary in the literature, we will focus on trees constructed using sample-splitting (honesty), which are believed to offer better empirical performance (Athey and Imbens, 2016).

Definition 5.1 (CART with sample-splitting (CART+)). At each level of the tree, generate new data $\{(\tilde{y}_i, \tilde{\mathbf{x}}_i^T) : i = 1, 2, \ldots, n\}$. 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 $\{(\tilde{y}_i, \tilde{\mathbf{x}}_i^T) : \tilde{\mathbf{x}}_i \in \mathrm{t}\}$. The output of the tree T at a point $\mathbf{x}$ belonging to a terminal node t is $\tilde{\mu}(T)(\mathbf{x}) = \frac{1}{\#\{\tilde{\mathbf{x}}_i \in \mathrm{t}\}} \sum_{\tilde{\mathbf{x}}_i \in \mathrm{t}} \tilde{y}_i$ if $n(\mathrm{t}) = \#\{\tilde{\mathbf{x}}_i \in \mathrm{t}\} > 0$ and $\tilde{\mu}(T)(\mathbf{x}) = \mu(\mathbf{x})$ if $n(\mathrm{t}) = 0$.

The only difference between conventional CART and CART with sample-splitting (CART+) is that the split points at each level are chosen using a fresh (statistically independent) random sample. In fact, all the results in this section remain valid if sample-splitting is not done at the output level, but only for symmetry do we consider CART+ herein. The adaptive properties of the tree are retained either way, as the nodes are still refined by minimizing the empirical squared error (2). One can construct a depth K tree with this methodology by splitting the original dataset $\{(y_i, \mathbf{x}_i^T) : i = 1, 2, \ldots, n\}$ into K disjoint subsets of size $\lfloor n/K \rfloor$. For our purposes, problems will arise as soon as the depth K is approximately $\log\log(n)$ and so there is 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+ ensures that the split points have a desirable Markovian property: a split point $\tilde{\tau} \in [a, b]$ conditioned on its immediate ancestor split points $\tilde{a} = a$ and $\tilde{b} = b$ is independent of all ancestor split points, including $\tilde{a}$ and $\tilde{b}$.

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 Q and a positive integer N such that, for all $n \geq N$,*

$$\mathbb{P}\Big(\sup_{x \in \mathcal{X}} |\tilde{\mu}(T_{K_n})(x) - \mu| > Q\Big) > Q^2.$$

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 used in practice.

The pointwise error in Theorem 5.2 should be contrasted with the IMSE. Under Assumption 1,

$$\mathbb{E}\Big[\int_{\mathcal{X}} (\tilde{\mu}(T_K)(x) - \mu)^2 \mathbb{P}_x(dx)\Big] \leq \frac{2^{K+1}\sigma^2}{n+1}. \tag{16}$$

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.

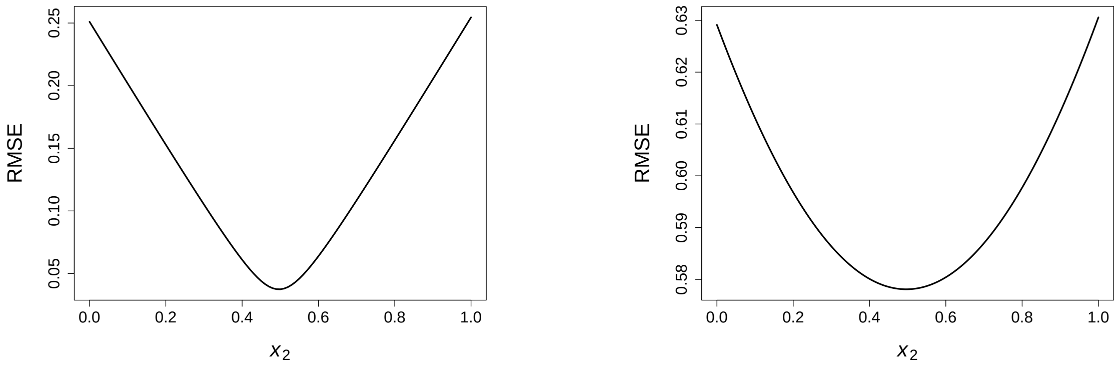
The intuition for Theorem 5.2 is based similarly on Theorem 4.1, but for depth K trees constructed with CART+ methodology. That is, honest trees of depth only $K \approx \log\log(n)$ will generate cells near the boundaries of the support $\mathcal{X}$ containing a finite number of observations with probability 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 $n(\mathrm{t}') \leq (n(\mathrm{t}))^{\gamma}$. 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, $n(\mathrm{t}) \leq n^{\gamma^K}$ observations in terminal nodes near the boundary of $\mathcal{X}$. Since $\gamma < 1$, the bound n^{γ^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 depth $K \gtrsim \log\log(n)$. In other words, after the first (finite) $J \geq 1$ splits, inconsistency will occur at any of the (approximately) $2^J + 1$ endpoints associated with the 2^J cells, whenever $J + K \gtrsim \log\log(n)$ and $n \to \infty$. Therefore, Theorem 5.2 establishes that there are at least countably many locations on $\mathcal{X}$ where a maximal depth K decision tree will be inconsistent, whenever $K \gtrsim \log\log(n)$ and $n \to \infty$ (since J is finite, it can be absorbed in the constant underlying the condition $K \gtrsim \log\log(n)$).

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, Delalleau, and Simard, 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. Fig. 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, Fig. 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



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

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

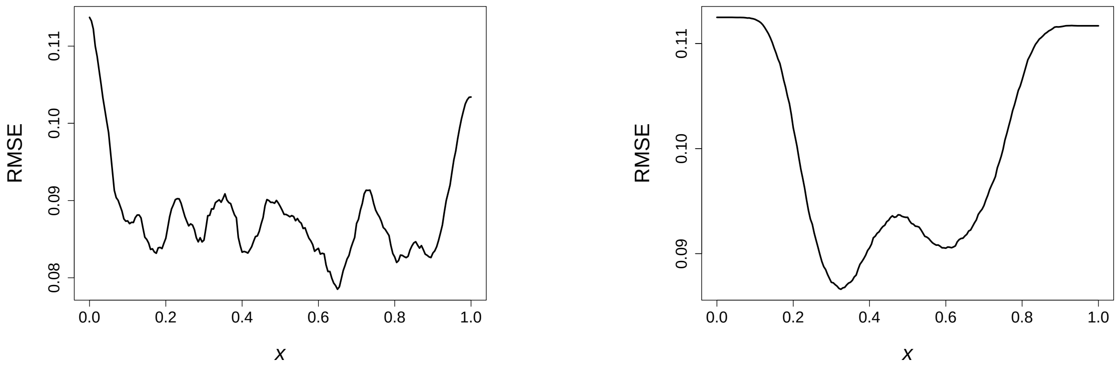
Figure 2: *Pointwise RMSE of pruned trees for models where* $\mathbf{x}$ *and* y *are dependent.*

from the model $y_i = d_i(x_{i1} - 0.5)(x_{i2} - 0.5) + \varepsilon_i$, where $d_i \sim \text{Bin}(0.5)$, $\mathbf{x}_i \sim U([0,1]^2)$, and $\varepsilon_i \sim N(0,1)$ are independent, and $n = 1000$. We do not include the transformed outcome tree $\hat{\theta}_{\text{ipw}}(T)(\mathbf{x})$ as it also produces a similar plot. In both cases, the numerical evidence indicates that pruning does not mitigate the lack of 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 $S \subset \{1, 2, \dots, n\}$ 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 $S \subset \{1, 2, \dots, n\}$ 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 $\mathcal{S} = \{S = S_1 \cup S_0 \subset \{1, 2, \dots, n\} : S_0 \cap S_1 = \emptyset, \ |S_0| = |S_1| = s/2\}$. In addition, at each node, a particular variable is split if it yields the smallest SSE (2) among a random selection $M \subset \{1, 2, \dots, p\}$ of $m = mtry$ candidate directions. The set of all candidate variable selections is denoted by $\mathcal{M} = \{M \subset \{1, 2, \dots p\} : |M| = m\}$. 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 (14) 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 $(M_1, S_1), (M_2, S_2), \dots, (M_B, S_B)$ are independent



(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.*

copies of (M, S). When the number of trees B is large, the honest random forest can 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 $s = o(n^{1/3})$ and $m = o(p/s)$, then for all $\mathbf{x} \in \mathcal{X}$,*

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

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 slow poly-logarithmic or polynomial 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^{γ} rate for individual trees. (This specification corroborates with recent empirical work by Mentch and Zhou (2020), which suggests that m should be small when the signal-to-noise ratio is low.) 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 Fig. 3, we plot the pointwise RMSE of a regression forest and causal forest for the models in Section 4.6. Compared to Fig. 1a and Fig. 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 formally 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, at best, poly-logarithmic uniform convergence rates, with pointwise convergence rates slower than any polynomial in boundary regions of the support of the covariates. Even more dramatic, when using sample-splitting (honesty), 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.

A Proofs

In this appendix, we include proofs of the formal statements in the main text. Throughout the proofs below, by working with the standardized response variable $(y_i - \mu)/\sigma$, we can assume without loss of generality that $\mu = 0$ and $\sigma^2 = 1$. As our results are asymptotic in nature, to avoid redundancy in the proofs, unless necessary to state so explicitly, we assume throughout that n is sufficiently large so that any event in question occurs with the stated probability.

A.1 Decision Stumps

In this section, we prove (7) and (8) in Theorem 4.1; (10), (11), and (12) in Theorem 4.2; and (13), (15), and (16). Throughout this section, we denote the partial sum by $V_i = y_1 + \cdots + y_i$, for $i \geq 1$.

Proof of (7) *in Theorem 4.1.* We first show that for each $\gamma \in (0,1)$, positive constants a, b, and real numbers w_1, w_2, we have the following limit:

$$\lim_{n\to\infty}(\lambda^2(an^\gamma, w_1) - \lambda^2(bn^{1-\gamma}, w_2)) = 2\sigma^2 \log\Big(\frac{\gamma}{1-\gamma}\Big) + 2\sigma^2(w_1 - w_2).$$

For simplicity, we only provide the proof when $a = b = 1$. The more general case follows a similar argument. The limit follows from noting that

$$\begin{aligned}\lambda^2(m, w) &= 2\sigma^2 \log\log(m)\bigg(1 + \frac{\log\log\log(m) + 2(w - \log(\sqrt{\pi}))}{4\log\log(m)}\bigg)^2 \\ &= 2\sigma^2\log\log(m) + \sigma^2\log\log\log(m) + 2\sigma^2(w - \log(\sqrt{\pi})) + O\bigg(\frac{\log\log\log(m)}{\log\log(m)}\bigg),\end{aligned}$$

and hence

$$\begin{aligned}\lim_{n\to\infty}(\lambda^2(n^\gamma, w_1) - \lambda^2(n^{1-\gamma}, w_2)) &= \lim_{n\to\infty}\Big(2\sigma^2\log\log(n^\gamma) + \sigma^2\log\log\log(n^\gamma) + 2\sigma^2(w_1 - \log(\sqrt{\pi})) \\ &\qquad - (2\sigma^2\log\log(n^{1-\gamma}) + \sigma^2\log\log\log(n^{1-\gamma}) + 2\sigma^2(w_2 - \log(\sqrt{\pi})))\Big) \\ &= 2\sigma^2\log\Big(\frac{\gamma}{1-\gamma}\Big) + 2\sigma^2(w_1 - w_2).\end{aligned}$$

In the above, we used the fact that

$$\lim_{n\to\infty}\Big(\log\log\log(n^\gamma) - \log\log\log(n^{1-\gamma})\Big) = 0.$$

In what follows, let $w_z := -\log(-\log(z))$. Fix $\delta \in (0,1)$ and choose $\gamma = \gamma(\delta) \in (0,1)$ such that

$$2\sigma^2\log\Big(\frac{\gamma}{1-\gamma}\Big) + 2\sigma^2(w_{\delta/16} - w_{1-\delta/32}) > 4\log(64/\delta), \tag{17}$$

which can always be done since $\lim_{\gamma\uparrow 1}\log\left(\frac{\gamma}{1-\gamma}\right)=\infty$. In this case, for n large enough,

$$\lambda^2(n^\gamma, w_{\delta/16}) > 4\log(64/\delta) + \lambda^2(n^{1-\gamma}/2, w_{1-\delta/32}).$$

To show (7) in Theorem 4.1, it suffices to show that, with probability at least $1-\delta$, we have

$$\hat{\imath} \in \operatorname*{argmax}_{1\le i\le n-1}\left(\frac{V_i^2}{i}+\frac{(V_n-V_i)^2}{n-i}\right) \subseteq \left[n^{\gamma^2},\, n^\gamma\right]\cup\left[n-n^\gamma,\, n-n^{\gamma^2}\right]. \tag{18}$$

We will show that, with probability at least $1-\delta/2$,

$$\operatorname*{argmax}_{1\le i\le n/2}\left(\frac{V_i^2}{i}+\frac{(V_n-V_i)^2}{n-i}\right) \subseteq \left[n^{\gamma^2},\, n^\gamma\right]. \tag{19}$$

If (19) is true, then by symmetry, with probability at least $1-\delta/2$, we have

$$\operatorname*{argmax}_{n/2<i\le n-1}\left(\frac{V_i^2}{i}+\frac{(V_n-V_i)^2}{n-i}\right) \subseteq \left[n-n^\gamma,\, n-n^{\gamma^2}\right], \tag{20}$$

so that a union bound of (19) and (20) proves (18).

To begin, by (9), with $h(m)=m=n^\gamma$, we have

$$\mathbb{P}\left(\max_{1\le i\le n^\gamma}\left(\frac{V_i^2}{i}+\frac{(V_n-V_i)^2}{n-i}\right) > \lambda^2(n^\gamma, w_{\delta/16})\right) \ge \mathbb{P}\left(\max_{1\le i\le n^\gamma}\frac{|V_i|}{\sqrt{i}} > \lambda(n^\gamma, w_{\delta/16})\right) \to 1-\frac{\delta}{16},$$

as $n\to\infty$. Therefore, we have

$$\mathbb{P}\left(\max_{1\le i\le n^\gamma}\left(\frac{V_i^2}{i}+\frac{(V_n-V_i)^2}{n-i}\right) > \lambda^2(n^\gamma, w_{\delta/16})\right) \ge 1-\frac{\delta}{8}, \tag{21}$$

for large enough n. Similarly, by (9) with $m=n/2$ and $h(m)=n^{1-\gamma}/2$, we have

$$\mathbb{P}\left(\max_{n^\gamma<i\le n/2}\frac{|V_i|}{\sqrt{i}} < \lambda(n^{1-\gamma}/2, w_{1-\delta/32})\right) \ge 1-\frac{\delta}{16}, \tag{22}$$

for large enough n. By Berkes and Weber (2006, Equation (5)), we have

$$\max_{1\le i\le n/2}\frac{|V_n-V_i|}{\sqrt{n-i}} \to \sup_{1\le t\le 2}\frac{|W_t|}{\sqrt{t}}, \quad n\to\infty,$$

where W_t is standard Brownian motion. Therefore, for n large enough, we have

$$\begin{aligned}\mathbb{P}\left(\max_{1\le i\le n/2}\frac{|V_n-V_i|}{\sqrt{n-i}} \ge 2\sqrt{\log(64/\delta)}\right) &\le 2\mathbb{P}\left(\sup_{1\le t\le 2}\frac{|W_t|}{\sqrt{t}} \ge 2\sqrt{\log(64/\delta)}\right)\\ &\le 2\mathbb{P}\left(\sup_{1\le t\le 2}|W_t| \ge 2\sqrt{\log(64/\delta)}\right).\end{aligned}$$

By Doob's maximal inequality for the margingale W_t (Revuz and Yor, 1999, Proposition 1.8) and a union bound, we have

$$\mathbb{P}\left(\max_{1\le i\le n/2} \frac{|V_n - V_i|}{\sqrt{n-i}} \ge 2\sqrt{\log(64/\delta)}\right) \le \frac{\delta}{16}. \tag{23}$$

Combining (22) and (23) and a union bound, we have

$$\mathbb{P}\left(\max_{n^\gamma < i\le n/2} \left(\frac{V_i^2}{i} + \frac{(V_n - V_i)^2}{n-i}\right) < 4\log(64/\delta) + \lambda^2(n^{1-\gamma}/2, w_{1-\delta/32})\right) \ge 1 - \frac{\delta}{8}. \tag{24}$$

Then a union bound of (21) and (24) implies that, with probability at least $1-\delta/4$, we have

$$\max_{1\le i\le n^\gamma} \left(\frac{V_i^2}{i} + \frac{(V_n - V_i)^2}{n-i}\right) > \max_{n^\gamma < i\le n/2} \left(\frac{V_i^2}{i} + \frac{(V_n - V_i)^2}{n-i}\right). \tag{25}$$

By a similar argument as before, with probability at least $1-\delta/4$, we have

$$\begin{aligned}\max_{1\le i< n^\gamma} \left(\frac{V_i^2}{i} + \frac{(V_n - V_i)^2}{n-i}\right) &< 4\log(64/\delta) + \lambda^2(n^\gamma, w_{1-\delta/32}) < \lambda^2(n^{\gamma^2}, w_{\delta/16}) \\ &< \max_{n^{\gamma^2}\le i\le n^\gamma} \left(\frac{V_i^2}{i} + \frac{(V_n - V_i)^2}{n-i}\right).\end{aligned} \tag{26}$$

Combining (25) and (26) with a union bound, the desired result (19) holds true. □

Now we prove the second half of Theorem 4.1.

Proof of (8) *in Theorem 4.1.* By symmetry, it suffices to show that $\mathbb{P}(\hat{\imath} \le 2h) \ge B(\log n)^{-b}$, where $2h \le A\log^a(n)$. Consider the event

$$\mathcal{E}_h = \left\{\frac{V_h}{\sqrt{h}} \in \left[\sqrt{5\log\log n}, \sqrt{7\log\log n}\right] \text{ and } \frac{V_{2h} - V_h}{\sqrt{h}} \in \left[-\sqrt{7\log\log n}, -\sqrt{5\log\log n}\right]\right\},$$

where h is a positive integer to be chosen later. By (9), under $\mathcal{E}_h$, with probability at least 0.99, we have

$$\max_{2h< i\le n-1} \frac{(V_n - V_i)^2}{n-i} \le 2.01\log\log n. \tag{27}$$

Under $\mathcal{E}_h$, we have $|V_{2h}| \le (\sqrt{7} - \sqrt{5})\sqrt{h\log\log n} \le 0.3\sqrt{2h\log\log n}$. Note that $V_n - V_i$ for $i \ge 2h$ is independent of $\mathcal{E}_h$. Therefore, under $\mathcal{E}_h$, for $h \ge 1$, we have

$$\max_{2h<i\le n-1} \frac{|V_i|}{\sqrt{i}} \le \max_{2h<i\le n-1} \frac{|V_i - V_{2h}|}{\sqrt{i-2h}} + \frac{|V_{2h}|}{\sqrt{2h}} \le \max_{1\le i\le n-2h-1} \frac{|V_{i+2h} - V_{2h}|}{\sqrt{i}} + 0.3\sqrt{\log\log n}.$$

Since $V_{i+2h} - V_{2h}$ has the same distribution as V_i, applying (9) to $\max_{1\le i\le n-2h-1} \frac{|V_{i+2h} - V_{2h}|}{\sqrt{i}}$, under $\mathcal{E}_h$, with probability at least 0.99, we have

$$\max_{2h<i\le n-1} \frac{|V_i|}{\sqrt{i}} \le (\sqrt{2.01} + 0.3)\sqrt{\log\log n}. \tag{28}$$

Therefore, by a union bound of (27) and (28), under $\mathcal{E}_h$, with probability at least 0.98, we have

$$\begin{aligned}\max_{2h<i\le n-1}\left(\frac{V_i^2}{i}+\frac{(V_n-V_i)^2}{n-i}\right) &\le ((\sqrt{2.01}+0.3)^2+2.01)\log\log n \le 5\log\log n \le \frac{V_h^2}{h} \\ &\le \max_{1\le i\le 2h}\left(\frac{V_i^2}{i}+\frac{(V_n-V_i)^2}{n-i}\right),\end{aligned} \tag{29}$$

which implies that $\hat{\imath} \le 2h$. Now it remains to choose h so that $\mathbb{P}(\mathcal{E}_h)$ is sufficiently large. By a generalization of the Berry-Esseen theorem (Petrov, 1975, Theorem 5, page 112), if $\rho := \mathbb{E}[|\varepsilon|^{2+\nu}] < \infty$, there exists a positive constant κ for which

$$\left|\mathbb{P}\left(\frac{V_h}{\sqrt{h}} < \sqrt{5\log\log n}\right) - \Phi\left(\sqrt{5\log\log n}\right)\right| \le \frac{\kappa\rho}{h^{\nu/2}},$$

and

$$\left|\mathbb{P}\left(\frac{V_h}{\sqrt{h}} < \sqrt{7\log\log n}\right) - \Phi\left(\sqrt{7\log\log n}\right)\right| \le \frac{\kappa\rho}{h^{\nu/2}},$$

where $\Phi(\cdot)$ is the standard normal cumulative distribution function. Since $\frac{\exp(-z^2/2)}{\sqrt{2\pi}}(1/z-1/z^3) < 1-\Phi(z) < \frac{\exp(-z^2/2)}{z\sqrt{2\pi}}$, for $z>0$, we have

$$\begin{aligned}\mathbb{P}\left(\frac{V_h}{\sqrt{h}} \in \left[\sqrt{5\log\log n},\ \sqrt{7\log\log n}\right]\right) &\ge \Phi\left(\sqrt{7\log\log n}\right) - \Phi\left(\sqrt{5\log\log n}\right) - \frac{2\kappa\rho}{h^{\nu/2}} \\ &\ge \frac{1}{(\log n)^{5/2}\sqrt{2\pi}}\left(\frac{1}{\sqrt{5\log\log n}} - \frac{1}{\sqrt{(5\log\log n)^3}}\right) \\ &\quad - \frac{1}{\sqrt{14\pi\log\log n}(\log n)^{7/2}} - \frac{2\kappa\rho}{h^{\nu/2}} \\ &\ge \frac{1}{5(\log n)^{5/2}\sqrt{\pi\log\log n}} - \frac{2\kappa\rho}{h^{\nu/2}}.\end{aligned}$$

By the same calculations, we also have

$$\mathbb{P}\left(\frac{V_{2h}-V_h}{\sqrt{h}} \in \left[-\sqrt{5\log\log n}, -\sqrt{7\log\log n}\right]\right) \ge \frac{1}{5(\log n)^{5/2}\sqrt{\pi\log\log n}} - \frac{2\kappa\rho}{h^{\nu/2}},$$

Recall that, under $\mathcal{E}_h$, the event $\hat{\imath} \le 2h$ occurs with probability at least 0.98. Therefore, choosing $h := \lceil(400\pi\kappa^2\rho^2(\log n)^5\log\log n)^{1/\nu}\rceil$ yields

$$\begin{aligned}&\mathbb{P}(\hat{\imath} \le 2h) \ge 0.98\cdot\mathbb{P}(\mathcal{E}_h) \\ &\quad = 0.98\cdot\mathbb{P}\left(\frac{V_h}{\sqrt{h}} \in \left[\sqrt{5\log\log n},\ \sqrt{7\log\log n}\right]\right)\mathbb{P}\left(\frac{V_{2h}-V_h}{\sqrt{h}} \in \left[-\sqrt{5\log\log n}, -\sqrt{7\log\log n}\right]\right) \\ &\quad \ge \frac{1}{100\pi(\log n)^5\log\log n},\end{aligned} \tag{30}$$

which completes the proof. □

Now we are ready to use Theorem 5.2 to prove Theorem 4.2.

Proof of (10) *in Theorem 4.2.* From (25) and (26), we know that

$$\mathbb{P}\left(\min\{\hat\imath,\, n-\hat\imath\} \in [n^{\gamma^2}, n^{\gamma}] \text{ and } \frac{V_{\hat\imath}^2}{\hat\imath} + \frac{(V_n - V_{\hat\imath})^2}{n-\hat\imath} \geq \lambda^2(n^{\gamma^2}, w_{\delta/16})\right) \geq 1-\delta/2. \tag{31}$$

By (23), we have that, with probability at least $1-\delta/4$,

$$\frac{(V_n - V_{\hat\imath})^2}{n-\hat\imath}\mathbf{1}(\hat\imath \leq n/2) \leq 4\log(16/\delta). \tag{32}$$

By symmetry, with probability at least $1-\delta/4$, we also have

$$\frac{V_{\hat\imath}^2}{\hat\imath}\mathbf{1}(\hat\imath > n/2) \leq 4\log(16/\delta). \tag{33}$$

Therefore, by (32) and (33) and a union bound, we have that, with probability at least $1-\delta/2$,

$$\frac{V_{\hat\imath}^2}{\hat\imath}\mathbf{1}(\hat\imath > n/2) + \frac{(V_n - V_{\hat\imath})^2}{n-\hat\imath}\mathbf{1}(\hat\imath \leq n/2) \leq 4\log(16/\delta). \tag{34}$$

Then, combining (31) and (34) and using a union bound implies that, with probability at least $1-\delta$, we have

$$\begin{aligned}
\sup_x |\hat\mu(T_1)(x)|^2 &\geq \frac{V_{\hat\imath}^2}{\hat\imath^2}\mathbf{1}(\hat\imath \leq n/2) + \frac{(V_n - V_{\hat\imath})^2}{(n-\hat\imath)^2}\mathbf{1}(\hat\imath > n/2) \\
&\geq \frac{1}{\min\{\hat\imath,\, n-\hat\imath\}}\left(\frac{V_{\hat\imath}^2}{\hat\imath} + \frac{(V_n - V_{\hat\imath})^2}{n-\hat\imath} - \left(\frac{V_{\hat\imath}^2}{\hat\imath}\mathbf{1}(\hat\imath > n/2) + \frac{(V_n - V_{\hat\imath})^2}{n-\hat\imath}\mathbf{1}(\hat\imath \leq n/2)\right)\right) \\
&\geq \frac{1}{\min\{\hat\imath,\, n-\hat\imath\}}\left(\lambda^2(n^{\gamma^2}, w_{\delta/16}) - 4\log(16/\delta)\right) \\
&\geq \frac{2\log\log(n^{\gamma^2}) + o(\log\log(n))}{n^{\gamma}}.
\end{aligned}$$

□

Proof of (11) *in Theorem 4.2.* From (31) and symmetry of the optimal split index, we have that

$$\mathbb{P}\left(\hat\imath \in \left[n^{\gamma^2}, n^{\gamma}\right] \text{ and } \frac{V_{\hat\imath}^2}{\hat\imath} + \frac{(V_n - V_{\hat\imath})^2}{n-\hat\imath} \geq \lambda^2(n^{\gamma^2}, w_{\delta/16})\right) \geq \frac{1-\delta/2}{2}. \tag{35}$$

Notice that a union bound of (32) and (35) implies that, with probability at least $(1-\delta)/2$, we have $\hat\imath \leq n^{\gamma}$ and

$$|\hat\mu(T_1)(x)|^2 = \frac{V_{\hat\imath}^2}{\hat\imath^2} \geq \frac{1}{\hat\imath}\left(\lambda^2(n^{\gamma^2}, w_{\delta/16}) - 4\log(16/\delta)\right) \geq \frac{2\log\log(n^{\gamma^2}) + o(\log\log(n))}{n^{\gamma}}, \tag{36}$$

for any $x < \hat\tau$. Notice that $\mathbb{E}[x_{(\lceil n^{\gamma}\rceil)}] = \frac{\lceil n^{\gamma}\rceil}{n+1}$ and $\mathrm{Var}(x_{(\lceil n^{\gamma}\rceil)}) = \frac{\lceil n^{\gamma}\rceil(n-\lceil n^{\gamma}\rceil+1)}{(n+1)^2(n+2)}$, owing to the fact that $x_{(i)} \sim$

Beta(i, $n+1-i$). Therefore, by Chebyshev's inequality, we have

$$\mathbb{P}\left(x_{(\lceil n^\gamma \rceil)} > \frac{\lceil n^\gamma \rceil}{2(n+1)}\right) \geq 1 - \frac{\frac{\lceil n^\gamma \rceil (n - \lceil n^\gamma \rceil + 1)}{(n+1)^2(n+2)}}{\left(\frac{\lceil n^\gamma \rceil}{2(n+1)}\right)^2} \geq 1 - \frac{4}{\lceil n^\gamma \rceil} \geq 1 - \delta/2. \tag{37}$$

A union bound of (36) and (37) implies that, with probability at least $1/2 - \delta$, we have

$$|\hat{\mu}(T_1)(x)|^2 \geq \frac{2 \log\log(n^{\gamma^2}) + o(\log\log(n))}{n^\gamma},$$

for all $x \in [0, \hat{\tau})$, where $\hat{\tau} = x_{(\hat{\imath})} \geq x_{(\lceil n^\gamma \rceil)} > \frac{\lceil n^\gamma \rceil}{2(n+1)} > D n^{\gamma - 1}$ and D is sufficiently small. By symmetry, the same bound is true for $x \in (\hat{\tau}, 1]$ as well, where $\hat{\tau} < 1 - D n^{\gamma-1}$. □

Proof of (12) *in Theorem 4.2*. By symmetry, we just need to show (12) for $x = 0$. Define $\delta := \frac{1}{200\pi (\log n)^5 \log\log n}$. From (29) and (30) in the proof of (8), we have that

$$\mathbb{P}\left(\hat{\imath} \leq 2h \text{ and } \frac{V_{\hat{\imath}}^2}{\hat{\imath}} + \frac{(V_n - V_{\hat{\imath}})^2}{n - \hat{\imath}} \geq 5 \log\log n\right) \geq 2\delta. \tag{38}$$

Similar to (32), we have that, with probability at least $1 - \delta$,

$$\frac{(V_n - V_{\hat{\imath}})^2}{n - \hat{\imath}} \mathbf{1}(\hat{\imath} \leq n/2) \leq 4 \log(4/\delta). \tag{39}$$

Now, by a union bound of (38) and (39), we see that, with probability at least δ, we have

$$|\hat{\mu}(T_1)(0)|^2 = \frac{1}{\hat{\imath}} \frac{V_{\hat{\imath}}^2}{\hat{\imath}} \geq \frac{5 \log\log n - 4 \log(4/\delta)}{2h} \geq \frac{5 \log\log n - 4 \log(4/\delta)}{2\lceil (400 \pi \kappa^2 \rho^2 (\log n)^5 \log\log n)^{1/\nu} \rceil}. \qquad \square$$

Proof of (13). Recall that, by symmetry, the optimal direction $\hat{\jmath}$ satisfies $\mathbb{P}(\hat{\jmath} = j) = 1/p$ for all j. Hence, by (10) in Theorem 4.2, we have,

$$\begin{aligned}
&\mathbb{P}\left(\sup_{\mathbf{x} \in \{0,1\}^p} |\hat{\mu}(T_1)(\mathbf{x}) - \mu| \geq C n^{-\gamma} \sqrt{\log\log(n)}\right) \\
&\quad = \sum_{j=1}^p \mathbb{P}\left(\sup_{x_j = 0,1} |\hat{\mu}(T_1)(x_j) - \mu| \geq C n^{-\gamma} \sqrt{\log\log(n)},\ \hat{\jmath} = j\right) \\
&\quad \geq \sum_{j=1}^p \left(\mathbb{P}\left(\sup_{x_j = 0,1} |\hat{\mu}(T_1)(x_j) - \mu| \geq C n^{-\gamma} \sqrt{\log\log(n)}\right) - \mathbb{P}(\hat{\jmath} \neq j)\right) \\
&\quad \geq \sum_{j=1}^p ((1 - \delta') - (1 - 1/p)) = 1 - \delta' p,
\end{aligned}$$

where $\hat{\mu}(T_1)(\mathbf{x})$ and $\hat{\mu}(T_1)(x_j)$ are the multi-dimensional and one-dimensional versions of the decision stump (6), respectively. Choosing $\delta' = \delta/p$, we obtain the desired result. □

Proof of (15). Note that $\mathbb{E}[\,\sup_{x\in\mathcal{X}}(\tilde{\mu}(T_1)(x)-\mu)^2] \geq \mathrm{Var}(\tilde{\mu}(T_1)(0)) = \mathbb{E}\Big[\frac{\mathbf{1}(\#\{\tilde{x}_i\leq x_{(\hat{\imath})}\}>0)}{\#\{\tilde{x}_i\leq x_{(\hat{\imath})}\}}\Big]$. Using the Cauchy-Schwarz inequality together with the independence between $\tilde{x}_i$ and $x_{(\hat{\imath})}$ and the fact that $x_{(\hat{\imath})} \sim \mathrm{Beta}(\hat{\imath},\, n+1-\hat{\imath})$, conditional on $\hat{\imath}$, we have

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

where we used $\mathbb{P}(\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} > 0 \mid \hat{\imath}) = 1-\binom{2n-\hat{\imath}}{n}/\binom{2n}{n} = 1-\prod_{i=1}^{\hat{\imath}} \frac{n-i+1}{2n-i+1} \geq 1-2^{-\hat{\imath}}$ and $\mathbb{E}[\#\{\tilde{x}_i \leq x_{(\hat{\imath})}\} \mid \hat{\imath}] = \frac{n}{n+1}\hat{\imath}$. The fact that $\mathbb{E}\Big[\frac{(1-2^{-\hat{\imath}})^2}{\hat{\imath}}\Big] \gtrsim (\log(n))^{-(a+b)}$ follows directly from (8). □

Proof of (16). Let $|T_K|$ denote the number of terminal nodes in T_K and $0 = \tilde{\tau}_0 \leq \tilde{\tau}_1 \leq \cdots \leq \tilde{\tau}_{|T_K|-1} \leq \tilde{\tau}_{|T_K|} = 1$ denote the successive splits at the terminal level of the tree, which are independent of the data $\{(\tilde{y}_i, \tilde{x}_i) : i = 1, 2, \ldots, n\}$ at the output level. Then, using Lemma A.2 below, the IMSE can be bounded as follows:

$$\begin{aligned}
\mathbb{E}\left[\int_{\mathcal{X}} (\tilde{\mu}(T_K)(x)-\mu)^2 \mathbb{P}_x(dx)\right] &= \sum_{k=1}^{|T_K|} \mathbb{E}\left[(\tilde{\tau}_k - \tilde{\tau}_{k-1})\frac{\mathbf{1}(\#\{\tilde{\tau}_{k-1} \leq \tilde{x}_i \leq \tilde{\tau}_k\} > 0)}{\#\{\tilde{\tau}_{k-1} \leq \tilde{x}_i \leq \tilde{\tau}_k\}}\right] \\
&\leq \sum_{k=1}^{|T_K|} \mathbb{E}\left[\frac{2(\tilde{\tau}_k - \tilde{\tau}_{k-1})}{1 + \#\{\tilde{\tau}_{k-1} \leq \tilde{x}_i \leq \tilde{\tau}_k\}}\right] \\
&\leq \frac{2|T_K|}{n+1} \leq \frac{2^{K+1}}{n+1}.
\end{aligned}$$

□

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 $\{(y_i(k), x_i(k)) : i = 1, 2, \ldots, n\}$ denote the samples at the k-th level. The i-th largest sample covariate at the k-th level is denoted as $x_{(i)}(k)$. We also 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 $[0, x_{(\tilde{\imath}_{k-1})}(k-1)]$. Because each $x_i(k)$ is uniformly distributed on $[0, 1]$, it follows that $x_{(i)}(k) \sim \mathrm{Beta}(i,\, n+1-i)$.

Lemma A.1. *There exist $r > 1$, a positive constant R, and a positive integer M such that for any depth $k \geq 1$ and $m \geq M$, we have $\mathbb{P}(1 \leq \tilde{n}_k \leq m) \geq (7/8) \cdot \mathbb{P}(1 \leq \tilde{n}_{k-1} \leq m) + (1/8) \cdot \mathbb{P}(1 \leq \tilde{n}_{k-1} \leq Rm^r)$.*

Proof. Observe that if v is a positive integer, then $\tilde{\imath}_{k-1} \mid \tilde{n}_{k-1} = v$ has the same distribution as $\tilde{\imath}_0 \mid \tilde{n}_0 = v$, because of the honest tree construction. Therefore, we can apply (7) with $\delta = 1/2$ to obtain

$$\min_{v\geq N} \mathbb{P}\Big(v^{\gamma^2} \leq \tilde{\imath}_{k-1} \leq v^{\gamma} \mid \tilde{n}_{k-1} = v\Big) \geq \frac{1-\delta}{2} = \frac{1}{4}, \tag{40}$$

for some positive integer N. To relate $\tilde{\imath}_{k-1}$ to $\tilde{n}_k$, notice that there are $\tilde{n}_k = \sum_{i=1}^n \mathbf{1}(x_i(k) \leq x_{(\tilde{\imath}_{k-1})}(k-1))$ samples in the left child node after splitting the parent node at $x_{(\tilde{\imath}_{k-1})}(k-1)$. Observe further that, by the law of iterated expectations,

$$\mathbb{E}[\tilde{n}_k \mid \tilde{\imath}_{k-1}] = \mathbb{E}[\mathbb{E}[\tilde{n}_k \mid \tilde{\imath}_{k-1}, \, x_{(\tilde{\imath}_{k-1})}(k-1)] \mid \tilde{\imath}_{k-1}] = n \cdot \mathbb{E}[x_{(\tilde{\imath}_{k-1})}(k-1) \mid \tilde{\imath}_{k-1}] = \frac{n}{n+1}\tilde{\imath}_{k-1}, \tag{41}$$

and, by the law of total variance,

$$\begin{aligned} \operatorname{Var}(\tilde{n}_k \mid \tilde{\imath}_{k-1}) &= n^2 \operatorname{Var}(x_{(\tilde{\imath}_{k-1})}(k-1) \mid \tilde{\imath}_{k-1}) + n\mathbb{E}[x_{(\tilde{\imath}_{k-1})}(k-1)(1 - x_{(\tilde{\imath}_{k-1})}(k-1)) \mid \tilde{\imath}_{k-1}] \\ &= \frac{n^2 \tilde{\imath}_{k-1}(n - \tilde{\imath}_{k-1} + 1)}{(n+1)^2(n+2)} + \frac{n\tilde{\imath}_{k-1}(n - \tilde{\imath}_{k-1} + 1)}{(n+1)(n+2)} \leq 2\tilde{\imath}_{k-1}. \end{aligned} \tag{42}$$

In both calculations (41) and (42), we used the fact that $x_{(\tilde{\imath}_{k-1})}(k-1) \sim \operatorname{Beta}(\tilde{\imath}_{k-1}, \, n+1-\tilde{\imath}_{k-1})$, conditional on $\tilde{\imath}_{k-1}$. Hence, by Chebyshev's inequality,

$$\mathbb{P}(1 \leq \tilde{n}_k \leq 2\tilde{\imath}_{k-1} \mid \tilde{\imath}_{k-1}) \geq 1 - \mathbb{P}\left(\left|\tilde{n}_k - \frac{n}{n+1}\tilde{\imath}_{k-1}\right| > \frac{\tilde{\imath}_{k-1}}{4} \mid \tilde{\imath}_{k-1}\right) \geq 1 - \frac{2\tilde{\imath}_{k-1}}{(\tilde{\imath}_{k-1}/4)^2} \geq \frac{1}{2}, \tag{43}$$

provided $\tilde{\imath}_{k-1} \geq 64$. By (40), if $m \geq N$, we have

$$\begin{aligned} \mathbb{P}(1 \leq \tilde{n}_k \leq m \mid m \leq \tilde{n}_{k-1} \leq (m/2)^{1/\gamma}) &\geq \min_{m \leq v \leq (m/2)^{1/\gamma}} \mathbb{P}\Big(v^{\gamma^2} \leq \tilde{\imath}_{k-1} \leq v^{\gamma} \mid \tilde{n}_{k-1} = v\Big)\mathbb{P}\Big(1 \leq \tilde{n}_k \leq m \mid v^{\gamma^2} \leq \tilde{\imath}_{k-1} \leq v^{\gamma}\Big) \\ &\geq \frac{1}{4} \min_{m \leq v \leq (m/2)^{1/\gamma}} \mathbb{P}\Big(1 \leq \tilde{n}_k \leq m \mid v^{\gamma^2} \leq \tilde{\imath}_{k-1} \leq v^{\gamma}\Big). \end{aligned} \tag{44}$$

Assume $m \geq (64)^{1/\gamma^2}$. If $m \geq 2v^{\gamma} \geq 2\tilde{\imath}_{k-1}$ and $\tilde{\imath}_{k-1} \geq v^{\gamma^2} \geq m^{\gamma^2} \geq 64$, we have $\min_{m \leq v \leq (m/2)^{1/\gamma}} \mathbb{P}\Big(1 \leq \tilde{n}_k \leq m \mid v^{\gamma^2} \leq \tilde{\imath}_{k-1} \leq v^{\gamma}\Big) \geq \min_{u \geq 64} \mathbb{P}(1 \leq \tilde{n}_k \leq 2u \mid \tilde{\imath}_{k-1} = u)$ and hence, by (43) and (44),

$$\mathbb{P}(1 \leq \tilde{n}_k \leq m \mid m \leq \tilde{n}_{k-1} \leq (m/2)^{1/\gamma}) \geq \frac{1}{4} \min_{u \geq 64} \mathbb{P}(1 \leq \tilde{n}_{k-1} \leq 2u \mid \tilde{\imath}_{k-1} = u) \geq \frac{1}{8}. \tag{45}$$

Now, taking $R = (1/2)^{1/\gamma}$ and $r = 1/\gamma$, note that (45) implies Lemma A.1 since

$$\begin{aligned} \mathbb{P}(1 \leq \tilde{n}_k \leq m) &\geq (1/8 + 7/8) \cdot \mathbb{P}(1 \leq \tilde{n}_{k-1} \leq m) + (1/8) \cdot \mathbb{P}(m \leq \tilde{n}_{k-1} \leq Rm^r) \\ &= (7/8) \cdot \mathbb{P}(1 \leq \tilde{n}_{k-1} \leq m) + (1/8) \cdot \mathbb{P}(1 \leq \tilde{n}_{k-1} \leq Rm^r). \end{aligned}$$

□

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 $n_\ell = R^{-1}(nR)^{(1/r)^\ell}$. We will show by induction that for any $k \geq 0$ and $\ell \geq 1$ such that $n_\ell \geq M$,

$$\mathbb{P}(1 \leq \tilde{n}_k \leq n_\ell) \geq \sum_{k'=\ell}^{k} \binom{k'-1}{\ell-1} (7/8)^{k'-\ell}(1/8)^\ell. \tag{46}$$

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

$$\mathbb{P}(1 \leq \tilde{n}_{k-1} \leq n_{\ell'}) \geq \sum_{k'=\ell'}^{k-1} \binom{k'-1}{\ell'-1} (7/8)^{k'-\ell'} (1/8)^{\ell'}. \tag{47}$$

If $\ell \geq 2$, then substituting our induction hypothesis (47) with $\ell' = \ell$ and $\ell' = \ell - 1$ into Lemma A.1, we get that

$$\begin{aligned}\mathbb{P}(1 \leq \tilde{n}_k \leq n_\ell) &\geq (7/8) \sum_{k'=\ell}^{k-1} \binom{k'-1}{\ell-1} (7/8)^{k'-\ell} (1/8)^{\ell} + (1/8) \sum_{k'=\ell-1}^{k-1} \binom{k'-1}{\ell-2} (7/8)^{k'-\ell+1} (1/8)^{\ell-1} \\ &= \sum_{k'=\ell}^{k} \binom{k'-1}{\ell-1} (7/8)^{k'-\ell} (1/8)^{\ell},\end{aligned}$$

where we used Pascal's identity. This completes the inductive proof of (46).

Let $X \sim \mathrm{NB}(L, 1/8)$, i.e., the number of independent trials, each occurring with probability $1/8$, until L successes. Choose

$$L = \lceil \log_r \log_r(nR) - \log_r \log_r(MR) - 1 \rceil \asymp \log\log(n), \quad n_L = R^{-1}(nR)^{(1/r)^L} \in [M, M^r R^{r-1}].$$

By (46) and Markov's inequality applied to the tail probability of X, we have that

$$\begin{aligned}\mathbb{P}(1 \leq \tilde{n}_K \leq M^r R^{r-1}) &\geq \mathbb{P}(1 \leq \tilde{n}_K \leq n_L) \\ &\geq \sum_{k'=L}^{K} \binom{k'-1}{L-1} (7/8)^{k'-L} (1/8)^{L} = 1 - \mathbb{P}(X \geq K+1) \\ &\geq 1 - \frac{L}{(1/8)(K+1)} \\ &\geq \frac{1}{2},\end{aligned}$$

as long as $K = K_n \geq \frac{2L}{1/8} \gtrsim \log\log(n)$. We therefore have $\mathrm{Var}(\tilde{\mu}(T_{K_n})(0)) = \sum_{m \geq 1} \frac{\mathbb{P}(\tilde{n}_{K_n} = m)}{m} \geq \frac{\mathbb{P}(1 \leq \tilde{n}_{K_n} \leq M^r R^{r-1})}{M^r R^{r-1}} \geq \frac{1}{2M^r R^{r-1}}$. This implies that there exists a set $\mathcal{A}$ and a positive constant U for which $\mathbb{P}(\mathcal{A}) > 0$ and $|\tilde{\mu}(T_{K_n})(0)| \geq U$ on $\mathcal{A}$; hence, we have that $\mathbb{E}[|\tilde{\mu}(T_{K_n})(0)|] \geq \mathbb{P}(\mathcal{A})U > 0$. By the Paley-Zygmund inequality (Petrov, 2007), we have

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

Thus, $\mathbb{P}(|\tilde{\mu}(T_{K_n})(0)| > Q) > Q^2$, where $Q = \frac{1}{2}\mathbb{P}(\mathcal{A})U$. □

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

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

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

Proof. We have

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

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*

$$\frac{1}{\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2}\sum_{S_0,S_1}\sum_{S'_0,S'_1} |S'_1 \cap S_0||S_1 \cap S'_0| \leq \frac{s^4}{16n(n-s/2)}.$$

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

$$\begin{aligned} \mathbb{E}[|S_1 \cap S'_0| \mid S'_1, S_0] &= \sum_{i \notin S'_1 \cup S_0} \mathbb{P}(i \in S_1 \cap S'_0 \mid S'_1, S_0) = \sum_{i \notin S'_1 \cup S_0} \mathbb{P}(i \in S_1 \mid S'_1, S_0)\mathbb{P}(i \in S'_0 \mid S'_1, S_0), \\ &= (n - |S'_1 \cup S_0|)\left(\frac{s/2}{n-s/2}\right)^2 \leq \frac{s^2}{4(n-s/2)}. \end{aligned} \tag{48}$$

Combining (48) and Lemma A.3, we have

$$\frac{1}{\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2}\sum_{S_0,S_1}\sum_{S'_0,S'_1} |S'_1 \cap S_0||S_1 \cap S'_0| = \mathbb{E}[|S'_1 \cap S_0| \cdot \mathbb{E}[|S_1 \cap S'_0| \mid S'_1, S_0]] \leq \frac{s^4}{16n(n-s/2)}. \qquad \square$$

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 $|S_1 \cap S'_1| \geq 1$, we have*

$$\frac{1}{\binom{n-s/2}{s/2}}\sum_{S_0}\frac{1}{|S'_1 \backslash S_0|} - \frac{2}{s} = \frac{1}{\binom{n-s/2}{s/2}}\sum_{S'_0}\frac{1}{|S_1 \backslash S'_0|} - \frac{2}{s} \leq \frac{2n}{s(n-s+2)}. \tag{49}$$

Furthermore,

$$\frac{1}{\binom{n-s/2}{s/2}^2}\left(\sum_{S_0}\frac{1}{|S'_1 \backslash S_0|} - \frac{2}{s}\right)\left(\sum_{S'_0}\frac{1}{|S_1 \backslash S'_0|} - \frac{2}{s}\right) \leq \frac{4n^2}{s^2(n-s+2)^2}. \tag{50}$$

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 (49) holds regardless of (S_1, S_1'). This implies (50), 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

$$\begin{aligned}
\mathbb{E}[\hat{\mu}(\mathbf{x})^2] &= \frac{1}{\binom{p}{m}^2\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2} \sum_{M,M'} \sum_{S,S'} \mathbb{E}[\hat{\mu}(T(M,S))(\mathbf{x})\hat{\mu}(T(M',S'))(\mathbf{x})] \\
&= \frac{1}{\binom{p}{m}^2\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2} \sum_{M,M'} \sum_{S,S'} \sum_{\substack{j\in M\\ j'\in M'}} \sum_{\substack{i\in S_1\\ i'\in S_1'}} \mathbb{E}[LL' + LR' + RL' + RR'],
\end{aligned} \tag{51}$$

where

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

We evaluate (51) by considering five cases on the indices (i, i', j, j').

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 $\mathbb{E}[LL'] = \mathbb{E}[LR'] = \mathbb{E}[RL'] = \mathbb{E}[RR'] = 0$.

Case 2: $i' \in S_1' \backslash S_0$ and $i \neq i'$. As with Case 1, we have that $\mathbb{E}[LL'] = \mathbb{E}[LR'] = \mathbb{E}[RL'] = \mathbb{E}[RR'] = 0$.

Case 3: $i \in S_1 \cap S'_0$ and $i' \in S'_1 \cap S_0$. By the Cauchy-Schwartz inequality, we have

$$\begin{aligned}(\mathbb{E}[LL'])^2 \le \mathbb{E}&\left[\frac{y_i^2 \mathbf{1}(\hat{\jmath}(M,S_0)=j)\mathbf{1}(x_{ij} \le \hat{s}(M,S_0))\mathbf{1}(x_j \le \hat{s}(M,S_0))}{(1+\#\{k \in S_1 \backslash \{i\} : x_{kj} \le \hat{s}(M,S_0)\})^2}\right] \\ &\cdot \mathbb{E}\left[\frac{y_{i'}^2 \mathbf{1}(\hat{\jmath}(M',S'_0)=j')\mathbf{1}(x_{i'j'} \le \hat{s}(M',S'_0))\mathbf{1}(x_{j'} \le \hat{s}(M',S'_0))}{(1+\#\{k' \in S'_1 \backslash \{i'\} : x_{k'j} \le \hat{s}(M',S'_0)\})^2}\right] \\ &\le \mathbb{E}\left[\frac{\mathbf{1}(x_{ij} \le \hat{s}(M,S_0))\mathbf{1}(\hat{\jmath}(M,S_0)=j)}{1+\#\{k \in S_1 \backslash \{i\} : x_{kj} \le \hat{s}(M,S_0)\}}\right] \cdot \mathbb{E}\left[\frac{\mathbf{1}(x_{i'j'} \le \hat{s}(M',S'_0))\mathbf{1}(\hat{\jmath}(M',S'_0)=j')}{1+\#\{k' \in S'_1 \backslash \{i'\} : x_{k'j} \le \hat{s}(M',S'_0)\}}\right],\end{aligned} \tag{52}$$

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 (52), we have

$$\mathbb{E}[LL'] \le \frac{2}{s}\sqrt{\mathbb{P}(\hat{\jmath}(M,S_0)=j)\mathbb{P}(\hat{\jmath}(M',S'_0)=j')}.$$

By symmetry, we have that

$$\mathbb{E}[LL' + LR' + RL' + RR'] \le \frac{8}{s}\sqrt{\mathbb{P}(\hat{\jmath}(M,S_0)=j)\mathbb{P}(\hat{\jmath}(M',S'_0)=j')}.$$

Therefore, by the Cauchy-Schwarz inequality,

$$\begin{aligned}\sum_{\substack{j\in M\\ j'\in M'}} \sum_{\substack{i\in S_1\cap S'_0\\ i'\in S'_1\cap S_0}} \mathbb{E}[LL' + LR' + RL' + RR'] &\le \frac{8|S_1 \cap S'_0||S'_1 \cap S_0|}{s} \sum_{\substack{j\in M\\ j'\in M'}} \sqrt{\mathbb{P}(\hat{\jmath}(M,S_0)=j)\mathbb{P}(\hat{\jmath}(M',S'_0)=j')} \\ &\le \frac{8|S_1 \cap S'_0||S'_1 \cap S_0|}{s}\sqrt{\sum_{j\in M}\mathbb{P}(\hat{\jmath}(M,S_0)=j)\sum_{j'\in M'}\mathbb{P}(\hat{\jmath}(M',S'_0)=j')} \\ &= \frac{8|S_1 \cap S'_0||S'_1 \cap S_0|}{s},\end{aligned}$$

so that, by Lemma A.4, we have

$$\frac{1}{\binom{p}{m}^2\binom{n}{s/2}^2\binom{n-s/2}{s/2}^2}\sum_{M,M'}\sum_{S,S'}\sum_{\substack{j\in M\\ j'\in M'}}\sum_{\substack{i\in S_1\cap S'_0\\ i'\in S'_1\cap S_0}} \mathbb{E}[LL' + LR' + RL' + RR'] \le \frac{s^3}{2n(n-s/2)}.$$

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 $(\{\mathbf{x}_k : k \in S_1\}, \hat{s}(M,S_0), \hat{\jmath}(M,S_0), \hat{s}(M',S'_0))$ and $\mathbb{E}[y_i^2] = 1$. Therefore,

$$\begin{aligned}\mathbb{E}[LL'] &\le \mathbb{E}\left[\frac{\mathbf{1}(\hat{\jmath}(M,S_0)=j)\mathbf{1}(x_{ij} \le \hat{s}(M,S_0))\mathbf{1}(x_j \le \hat{s}(M,S_0))\mathbf{1}(x_j \le \hat{s}(M',S'_0))}{1+\#\{k \in S_1 \backslash \{i\} : x_{kj} \le \hat{s}(M,S_0)\}}\right] \\ &\le \frac{\mathbb{P}(\hat{\jmath}(M,S_0)=j,\ x_j \le \hat{s}(M,S_0),\ \text{and } x_j \le \hat{s}(M',S'_0))}{s/2},\end{aligned}$$

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

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

Applying Lemma A.3 twice, we see that

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

Case 5: $j \ne j'$ 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

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

Similarly, we also have

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

Therefore, we have

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

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

$$\begin{aligned}&\sum_{\substack{j\in M\\ j\in M'}}\ \sum_{i\in S_1\cap S_1'} \mathbb{E}[LL' + LR' + RL' + RR']\\
&\le \sum_{\substack{j\in M\\ j\in M'}}\ \sum_{i\in S_1\cap S_1'} \mathbb{P}(\hat{j}(M,S_0) = j,\ \hat{j}(M',S_0') = j')\left(\frac{\mathbf{1}(j\notin M')}{s/2} + \frac{\mathbf{1}(j\in M')}{|S_1\backslash S_0'|}\right)\left(\frac{\mathbf{1}(j'\notin M)}{s/2} + \frac{\mathbf{1}(j'\in M)}{|S_1\backslash S_0'|}\right)\\
&\le \frac{|S_1\cap S_1'|}{m^2}\left(\frac{m-|M\cap M'|}{s/2} + \frac{|M\cap M'|}{|S_1\backslash S_0'|}\right)\left(\frac{m-|M\cap M'|}{s/2} + \frac{|M\cap M'|}{|S_1'\backslash S_0|}\right)\\
&\le |S_1\cap S_1'|\left(\frac{4}{s^2} + \frac{2|M\cap M'|}{sm}\left(\frac{1}{|S_1\backslash S_0'|} + \frac{1}{|S_1'\backslash S_0|} - \frac{4}{s}\right) + \frac{|M\cap M'|}{m}\left(\frac{1}{|S_1\backslash S_0'|} - \frac{2}{s}\right)\left(\frac{1}{|S_1'\backslash S_0|} - \frac{2}{s}\right)\right).\end{aligned}\tag{53}$$

Since $i \in S_1 \cap S'_1$, we have $|S_1 \cap S'_1| \geq 1$, so by (53) 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

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

Athey, S., and G. W. Imbens (2019): "Machine learning methods that economists should know about," *Annual Review of Economics*, 11(1), 685–725.

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

Behr, M., Y. Wang, X. Li, and B. Yu (2022): "Provable Boolean interaction recovery from tree ensemble obtained via random forests," *Proceedings of the National Academy of Sciences*, 119(22), e2118636119.

Bengio, Y., O. Delalleau, and C. Simard (2010): "Decision Trees Do Not Generalize To New Variations," *Computational Intelligence*, 26(4), 449–467.

Berk, R. A. (2020): *Statistical learning from a regression perspective*, Springer Series in Statistics. Springer Nature.

Berkes, I., and M. Weber (2006): "Almost sure versions of the Darling–Erdős theorem," *Statistics and Probability Letters*, 76(3), 280–290.

Breiman, L., J. Friedman, R. Olshen, and C. J. Stone (1984): *Classification and Regression Trees*. Chapman and Hall/CRC.

Bühlmann, P., and B. Yu (2002): "Analyzing bagging," *The Annals of Statistics*, 30(4), 927 – 961.

Chernozhukov, V., J. C. Escanciano, H. Ichimura, W. K. Newey, and J. M. Robins (2022): "Locally Robust Semiparametric Estimation," *Econometrica*, 90(4), 1501–1535.

Darling, D. A., and P. Erdös (1956): "A limit theorem for the maximum of normalized sums of independent random variables," *Duke Mathematical Journal*, 23(1), 143 – 155.

Fan, J., and I. Gijbels (1996): *Local polynomial modelling and its applications*, vol. 66 of *Monographs on Statistics and Applied Probability*. Chapman and Hall.

Farrell, M. H., T. Liang, and S. Misra (2021): "Deep neural networks for estimation and inference," *Econometrica*, 89(1), 181–213.

Györfi, L., M. Kohler, A. Krzyzak, and H. Walk (2002): *A Distribution-Free Theory of Nonparametric Regression*. Springer Science & Business Media.

Hastie, T., R. Tibshirani, and J. Friedman (2009): *The elements of statistical learning*, Springer Series in Statistics. Springer-Verlag, New York.

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

Meinshausen, N. (2006): "Quantile Regression Forests," *Journal of Machine Learning Research*, 7(35), 983–999.

Mentch, L., and S. Zhou (2020): "Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success," *Journal of Machine Learning Research*, 21(171), 1–36.

Murdoch, W. J., C. Singh, K. Kumbier, R. Abbasi-Asl, and B. Yu (2019): "Definitions, methods, and applications in interpretable machine learning," *Proceedings of the National Academy of Sciences*, 116(44), 22071–22080.

Petrov, V. V. (1975): *Sums of Independent Random Variables*, vol. 82 of *Ergebnisse der Mathematik und ihrer Grenzgebiete*. Springer-Verlag, New York, Heidelberg, Berlin.

Petrov, V. V. (2007): "On lower bounds for tail probabilities," *Journal of Statistical Planning and Inference*, 137(8), 2703–2705, 5th St. Petersburg Workshop on Simulation.

Revuz, D., and M. Yor (1999): *Continuous martingales and Brownian motion*, no. 293 in Grundlehren der mathematischen Wissenschaften. Springer, Berlin.

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

Su, X., C.-L. Tsai, H. Wang, D. M. Nickerson, and B. Li (2009): "Subgroup Analysis via Recursive Partitioning," *Journal of Machine Learning Research*, 10(5), 141–158.

Tang, C., D. Garreau, and U. von Luxburg (2018): "When do random forests fail?," in *Advances in Neural Information Processing Systems*, ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, vol. 31. Curran Associates, Inc.

WAGER, S., AND S. ATHEY (2018): "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113(523), 1228–1242.

YAO, L., Z. CHU, S. LI, Y. LI, J. GAO, AND A. ZHANG (2021): "A Survey on Causal Inference," *ACM Trans. Knowl. Discov. Data*, 15(5).