# Universal Consistency of Decision Trees in High Dimensions

**Jason M. Klusowski***

*Department of Operations Research and Financial Engineering, Princeton University*
*e-mail:* jason.klusowski@princeton.edu

**Abstract:** This paper shows that decision trees constructed with Classification and Regression Trees (CART) methodology are universally consistent in an additive model context, even when the number of predictor variables grows exponentially with the sample size, under certain 1-norm and 0-norm sparsity constraints. The theory applies to a wide range of additive models, including those with component functions that are of bounded variation and, more generally, measurable. Consistency is universal in the sense that there are no a priori assumptions on the distribution of the predictor variables. Amazingly, this adaptivity to (approximate or exact) model sparsity is achieved with a single tree, as opposed to what might be expected for an ensemble. Finally, we show that these qualitative properties of individual trees are inherited by Breiman's random forests. Another surprise is that consistency holds even when the "mtry" tuning parameter vanishes as a fraction of the number of predictor variables, thus enabling more rapid computation of the forest. A key step in the analysis is the establishment of an oracle inequality, which precisely characterizes the goodness-of-fit and complexity tradeoff for a misspecified model.

**MSC2020 subject classifications:** Primary 62G08, 68T05; secondary 68Q32, 68Q25, 68W40.

**Keywords and phrases:** Decision trees; classification and regression trees; random forests; ensemble learning; interpretable machine learning; greedy algorithms; additive models; high dimensional; consistency; oracle inequality.

## 1. Introduction

Decision trees are one of the most elemental methods for predictive modeling. Accordingly, they are the cornerstone of many celebrated algorithms in statistical learning. For example, decision trees are often employed in ensemble learning, i.e., bagging [4], random forests [5], and gradient tree boosting [11]. From an applied perspective, decision trees are intuitive and have an appealing interpretability that makes them easy to explain to statistical nonexperts. They are also supplemented by a rich set of analytic and visual diagnostic tools for exploratory data analysis. These qualities have led to the prominence of decision tree learning in disciplines—such as medicine and

*This research was supported in part by the National Science Foundation through grants DMS-2054808 and HDR TRIPODS DATA-INSPIRE DCCF-1934924.

business—which place high importance on the ability to understand and interpret the output from the training algorithm, even at the expense of predictive accuracy.

Though our primary focus is theoretical, to make this paper likewise relevant to the applied user of decision trees, we focus exclusively on Classification and Regression Trees (CART) [6] methodology—undoubtedly the most popular one for regression and classification problems. On the theoretical side, this methodology raises a number of technical challenges which stem from the top down greedy recursive splitting and line search needed to find the best split points, thereby making CART notoriously difficult to study. These subtle mechanisms are of course desirable from a statistical standpoint, as they endow the decision tree with the ability to adapt to structural and qualitative properties of the underlying statistical model (such as sparsity and smoothness). Notwithstanding these major challenges, we take a significant step forward in advancing the theory of decision trees and prove the following (informal) statement in this paper:

> *Decision trees constructed with CART methodology are universally consistent for high dimensional additive models, where the number of predictor variables is allowed to grow exponentially fast with the sample size.*

The consistency (in mean squared error) is universal in the sense that there are no a priori assumptions on the input distribution, thereby improving upon most past work which requires the predictor variables to be continuous and either independent or near-independent (e.g., joint densities which are bounded above and below by fixed positive constants). Here we allow the input distribution to be discrete (so as to handle count data) and the predictor variables to have arbitrary dependence between each other.

Expectedly, our results for individual trees also carry over to ensembles, namely, Breiman's random forests [5], which among other things, use CART methodology for the constituent trees. The surprising finding here is that the "mtry" tuning parameter (which equals the number of randomly selected candidate variables for splitting at each node) can be significantly smaller than the number of predictor variables. Specifically, "mtry" can vanish as a fraction of the number of predictor variables and random forests can still be consistent. The conclusion of this result clearly has implications for the computational cost of growing each tree, since the optimal split points only need to be computed for a (possibly) very small subset of the variables at each node. Informally stated:

> *Breiman's random forests are universally consistent for high dimensional additive models, where "mtry" is allowed to vanish as a fraction of the number of predictor variables.*

Let us emphasize that, in our theoretical treatment of decision trees, we are not content with merely exhibiting the standard certificates for good predictors, such as asymptotic consistency. Rather, we aim to identify and explore the *unique* advantages of decision tree learning over other nonparametric tools, such as nearest neighbors or kernel based methods. In doing so, we show that decision trees automatically adapt to the unknown sparsity of the statistical model as well as satisfy a type of adaptive prediction error

bound known as an *oracle inequality*, which cleanly reveals the goodness-of-fit and complexity tradeoff. We believe that these results reveal distinct properties of decision trees that have not been uncovered in past work.

### *1.1. Prior art*

We now review some of the past theoretical work on CART. The first consistency result for CART was provided in the original book that proposed the methodology [6], albeit under very strong assumptions on the tree construction. Thirty years later, [23] showed asymptotic consistency of CART for (fixed dimensional) additive models with continuous component functions, en route to establishing asymptotic consistency of Breiman's random forests. This paper was an important technical achievement because it did not require any of the strong assumptions on the tree made in [6]. Subsequent work by [27], [8], and [15] provide finite sample consistency rates in a high dimensional setting with exact sparsity, though under strong assumptions on the regression model and tree construction. Another notable paper [12] provides oracle-type inequalities for pruned CART, though the theory does not extend to out-out-sample prediction.

Motivated by Stone's conditions for consistency in nonparametric regression [24], most existing convergence results for decision trees follow an approach in which the approximation error is bounded by the mesh of the induced partition. Conditions are then imposed to ensure that the mesh approaches zero as the depth of the tree increases. This is then combined with a standard empirical process argument to show vanishing estimation error, which in turn, implies that the prediction error vanishes also [9, 6, 27, 26]. In contrast, the aforementioned paper [23] controls the variation of the regression function inside the cells of the partition, without explicitly controlling the mesh, though the theoretical consequences are similar. While these techniques can be useful to prove consistency statements, they are not generally delicate enough to capture the adaptive properties of the tree or handle high dimensional situations. To address this shortcoming, [8] and [15] developed techniques to bypass having to use the partition mesh as a proxy for the approximation error and instead directly analyze the prediction error by exploiting the greedy optimization construction of CART. Both works provide consistency rates for models with exact sparsity in a moderately high dimensional regime (i.e., when the dimensionality grows polynomially with the sample size), however, they make very strong assumptions on the tree construction and data generating process. For example, the results of [15] apply only to the noise free setting. Furthermore, [8] require an "edge" condition to ensure the prediction error decreases by a constant factor after each split. Except in a narrow set of cases for which the condition can be verified (namely, additive models with isotonic or piecewise linear component functions and independent predictor variables), its generality remains unclear. Our theory reveals that these assumptions are not strictly necessary in a general additive setting, most notably when the predictor variables are dependent and the dimensionality grows exponentially with the sample size.

### 1.2. Organization

This paper is organized according the following schema. In Section 2, we describe the statistical model and introduce various important quantities, including those that control the sparsity of the model. We review basic terminology associated with CART methodology and describe how to construct the decision tree in Section 3. Our main results for CART are contained in Section 4; specifically, a training error bound, oracle inequality, and high dimensional asymptotic consistency statement. Analogous theory for Breiman's random forests is stated in Section 5. Finally, all proofs and technical lemmas are deferred to Section 7.

## 2. Preliminaries

### 2.1. Learning setting

Throughout this paper, we operate under a standard regression framework. The statistical model is $Y = \mu(\mathbf{X}) + \varepsilon$, where $\mu(\mathbf{X}) \coloneqq \mathbb{E}(Y|\mathbf{X})$ is a regression function, $\mathbf{X} \coloneqq (X_1, X_2, \dots, X_p)^T \in \mathbb{R}^p$ is a $p$-dimensional vector of predictor variables, and $\varepsilon \coloneqq Y - \mu(\mathbf{X})$ is statistical noise. We observe data $\mathcal{D}_n \coloneqq \{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \dots, (\mathbf{X}_n, Y_n)\}$ drawn i.i.d. from the regression model $\mathbb{P}_{(\mathbf{X},Y)} = \mathbb{P}_{\mathbf{X}}\mathbb{P}_{Y|\mathbf{X}}$. For simplicity and ease of exposition, we assume throughout this paper that the response variable is almost surely bounded, i.e., there exists $B > 0$ such that

$$|Y| \le B \tag{1}$$

almost surely; however, our proofs can be readily modified to accommodate unbounded sub-Gaussian errors with only minor adjustments to the forthcoming theoretical statements.

The object of interest is the regression function $\mu(\cdot)$; that is, from the data, we would like to predict the conditional mean of $Y$ given a new observation $\mathbf{X}' \in \mathbb{R}^p$. Here we measure the efficacy of a predictor via the *expected* mean squared prediction error; however, high probability bounds (over the data) on the mean squared prediction error can be developed with relative ease. To understand the predictive properties of decision trees in an (ultra) high dimensional setting, under suitable conditions on $\mu(\cdot)$, the dimensionality $p = p_n$ is permitted to grow exponentially with the sample size.

A natural playground to illustrate the high dimensional properties of CART is in the context of additive modeling, i.e., models for which the regression function can be expressed as a sum of univariate functions of the predictor variables. Indeed, these models are often used in (ultra) high dimensional settings where notions of approximate sparsity make sense. They are, however, typically trained with regularized procedures based on splines, trend filtering, or reproducing kernel Hilbert spaces [18, 21, 22, 28] that, unlike the agnostic nature of CART, take explicit advantage of the additive structure.

We implicitly work with the class of functions $\mathcal{G}$ that admit an additive form

$$g(\mathbf{X}) = g_1(X_1) + g_2(X_2) + \cdots + g_p(X_p), \tag{2}$$

where $g_1(X_1), g_2(X_2), \ldots, g_p(X_p)$ is a collection of $p$ univariate functions of *bounded variation* over the respective supports of $X_1, X_2, \ldots, X_p$ (i.e., each $g_j(\cdot)$ has finite total variation, denoted by $v(g_j) \coloneqq \mathrm{TV}(g_j)$). That is, if $\mathbf{x} = (x_1, x_2, \ldots, x_p)^T \in \mathbb{R}^p$, then

$$\mathcal{G} \coloneqq \big\{ g(\mathbf{x}) = g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) : g_j(\cdot) \text{ has bounded variation} \big\}.$$

Of course, additive models may be too rigid for many applications because of their inability to capture interaction effects among the predictor variables. To address this shortcoming, we importantly allow for model misspecification and therefore *do not* require the regression function $\mu(\cdot)$ to be additive or to belong to $\mathcal{G}$.

For $g \in \mathcal{G}$, we define the norm $\|g\|_{\mathrm{TV}}$ as the infimum of

$$v(g_1) + v(g_2) + \cdots + v(g_p)$$

over all representations of $g(\cdot)$ as (2), i.e., $\|g\|_{\mathrm{TV}}$ is the $\ell_1$ aggregated total variation of the individual component functions (see [25] and the references therein). Throughout, however, we assume $g(\cdot)$ has a canonical representation such that $\|g\|_{\mathrm{TV}}$ achieves this infimum. One can think of $\|g\|_{\mathrm{TV}}$ as a measure of the "capacity" of $g(\cdot)$ and, as we shall see, it will play a central role in the paper. Later on, we will relax the bounded variation assumption and develop results for additive models with component functions that are merely bounded and measurable.

In the case that all the component functions $g_j(\cdot)$ are smooth over a compact domain $\mathcal{X} \subset \mathbb{R}$, the total variation $\ell_1$ norm can be expressed as the multiple Riemann integral

$$\|g\|_{\mathrm{TV}} = \int_{\mathcal{X}^p} \|\nabla g(\mathbf{x})\|_{\ell_1} d\mathbf{x},$$

where $\nabla(\cdot)$ is the gradient operator and $\|\cdot\|_{\ell_1}$ is the usual $\ell_1$ norm of a vector in $\mathbb{R}^p$. We note that if $g(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}$ is linear over the domain $[0,1]^p$, then $\|g\|_{\mathrm{TV}}$ equals $\|\boldsymbol{\beta}\|_{\ell_1}$, the $\ell_1$ norm of the coefficient vector. Furthermore, if $g(\cdot)$ is piecewise constant on $V$ regions, then $\|g\|_{\mathrm{TV}} \le 2V\|g\|_\infty$, where $\|\cdot\|_\infty$ is the supremum norm.

We now introduce some notation that will be used throughout the paper. For a function $f \in L_2(\mathbb{P}_{\mathbf{X}})$, let $\|f\|^2 \coloneqq \int_{\mathbb{R}^p} (f(\mathbf{x}))^2 d\mathbb{P}_{\mathbf{X}}(\mathbf{x})$ be the squared $L_2(\mathbb{P}_{\mathbf{X}})$ norm and for multivariate functions $f(\cdot)$ and $g(\cdot)$, let

$$\|f\|_n^2 \coloneqq \frac{1}{n}\sum_{i=1}^n (f(\mathbf{X}_i))^2 \quad \text{and} \quad \langle f, g\rangle_n \coloneqq \frac{1}{n}\sum_{i=1}^n f(\mathbf{X}_i) g(\mathbf{X}_i) \tag{3}$$

denote the squared norm and inner product, respectively, with respect to the empirical measure on the data. We view the response data vector $(Y_1, Y_2, \ldots, Y_n)^T \in \mathbb{R}^n$ as a function defined on the design matrix $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_n)^T \in \mathbb{R}^{n\times p}$ such that $Y(\mathbf{X}_i) = Y_i$. Thus, to be consistent with (3), we write, for example, $\|Y - f\|_n^2 = \frac{1}{n}\sum_{i=1}^n (Y_i - f(\mathbf{X}_i))^2$ and $\langle Y, f\rangle_n = \frac{1}{n}\sum_{i=1}^n Y_i f(\mathbf{X}_i)$. For a positive real number $z$, we use the notation $[\![z]\!]$ to the denote the largest integer less than or equal to $z$, i.e., the floor function.

## 3. CART methodology

As mentioned earlier, regression trees are commonly constructed with Classification and Regression Trees (CART) [6] methodology. The primary objective of CART is to find partitions of the predictor variables that produce minimal variance of the response values (i.e., minimal sum of squares error with respect to the average response values). Because of the computational infeasibility of choosing the best overall partition, decision trees with CART methodology are constructed in a greedy top down fashion (with a mere $\mathcal{O}(pn\log(n))$ average case complexity) using a procedure in which a sequence of locally optimal splits recursively partitions the input space. We now describe the algorithm in further detail.

### *3.1. Growing the tree*

Consider splitting a regression tree $T$ at a node t (a hyperrectangular region in $\mathbb{R}^p$). Let $s$ be a candidate split point for a generic variable $X \in \mathbb{R}$ that divides the parent node t into left and right daughter nodes $\mathrm{t}_L$ and $\mathrm{t}_R$ according to whether $X \le s$ or $X > s$, respectively. These two nodes will be denoted by $\mathrm{t}_L \coloneqq \{\mathbf{X} \in \mathrm{t} : X \le s\}$ and $\mathrm{t}_R \coloneqq \{\mathbf{X} \in \mathrm{t} : X > s\}$. For a node t in $T$ and data vectors $(W_1, W_2, \dots, W_n)^T$ and $(W_1', W_2', \dots, W_n')^T$ in $\mathbb{R}^n$ (again, viewed as functions defined on the design matrix), we introduce the notation

$$\|W\|_{\mathrm{t}}^2 \coloneqq \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} |W_i|^2 \quad \text{and} \quad \langle W, W'\rangle_{\mathrm{t}} \coloneqq \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} W_i W_i'$$

for the squared norm and inner product, respectively, with respect to the conditional empirical measure given $\mathbf{X} \in \mathrm{t}$. Here $N(\mathrm{t}) \coloneqq \#\{\mathbf{X}_i \in \mathrm{t}\}$ is the number of sample points $\mathbf{X}_i$ within t.

An effective split divides the data from the parent node into two daughter nodes so that the heterogeneity in each of the daughter nodes, as measured through the *impurity*, is reduced from that of the parent node. Impurity for regression trees is determined by the within-node sample variance

$$\|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 \coloneqq \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} (Y_i - \overline{Y}_{\mathrm{t}})^2, \tag{4}$$

where $\overline{Y}_{\mathrm{t}} \coloneqq \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} Y_i$ is the sample mean for t. Similarly, the within-node sample variances for the daughter nodes are

$$\|Y - \overline{Y}_{\mathrm{t}_L}\|_{\mathrm{t}_L}^2 = \frac{1}{N(\mathrm{t}_L)}\sum_{\mathbf{X}_i\in\mathrm{t}_L} (Y_i - \overline{Y}_{\mathrm{t}_L})^2, \quad \|Y - \overline{Y}_{\mathrm{t}_R}\|_{\mathrm{t}_R}^2 = \frac{1}{N(\mathrm{t}_R)}\sum_{\mathbf{X}_i\in\mathrm{t}_R} (Y_i - \overline{Y}_{\mathrm{t}_R})^2,$$

where $\overline{Y}_{\mathrm{t}_L}$ is the sample mean for $\mathrm{t}_L$ and $N(\mathrm{t}_L)$ is the sample size of $\mathrm{t}_L$ (similar definitions apply to $\overline{Y}_{\mathrm{t}_R}$ and $N(\mathrm{t}_R)$). For a candidate variable $X_j$ and split point $s = s_j$, the impurity gain is defined as [6, Definition 8.13]

$$\widehat{\Delta}(s, j, \mathrm{t}) \coloneqq \|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - P(\mathrm{t}_L)\|Y - \overline{Y}_{\mathrm{t}_L}\|_{\mathrm{t}_L}^2 - P(\mathrm{t}_R)\|Y - \overline{Y}_{\mathrm{t}_R}\|_{\mathrm{t}_R}^2, \tag{5}$$

where $P(\mathrm{t}_L) \coloneqq N(\mathrm{t}_L)/N(\mathrm{t})$ and $P(\mathrm{t}_R) \coloneqq N(\mathrm{t}_R)/N(\mathrm{t})$ are the proportions of data points within t that are contained in $\mathrm{t}_L$ and $\mathrm{t}_R$, respectively. We also define $w(\mathrm{t}) \coloneqq N(\mathrm{t})/n$ to be the proportion of samples $\mathbf{X}_i$ that belong to node t.

At each node t of the tree, we find the direction $X_{\hat{\jmath}}$, where $\hat{\jmath} = \hat{\jmath}(\mathrm{t})$, and split point $\hat{s} = \hat{s}(\hat{\jmath}, \mathrm{t})$ that minimize the sum of squares error

$$\sum_{\mathbf{X}_i \in \mathrm{t}_L} (Y_i - \overline{Y}_{\mathrm{t}_L})^2 + \sum_{\mathbf{X}_i \in \mathrm{t}_R} (Y_i - \overline{Y}_{\mathrm{t}_R})^2,$$

or, equivalently, maximize the impurity gain $\widehat{\Delta}(s, j, \mathrm{t})$, breaking ties arbitrarily. In other words, the parent node t is split into two daughter nodes using the variable and split point producing the largest impurity gain. The daughter nodes $\mathrm{t}_L$ and $\mathrm{t}_R$ of t become new parent nodes at the next level of the tree and are themselves further divided according to the previous scheme, and so on and so forth, until a desired depth is reached. The output $\widehat{\mu}(T) = \widehat{\mu}(T, \mathcal{D}_n)$ of the tree $T$ at a terminal (leaf) node t is the least squares (constant) predictor for data within t, namely, $(\widehat{\mu}(T))(\mathbf{x}) \equiv \overline{Y}_{\mathrm{t}}$ for all $\mathbf{x} \in \mathrm{t}$.

In what follows, we let $T_K$ denote a fully grown binary tree of depth $K$, constructed with the CART methodology described above. More specifically, we stop splitting a node if (i) the node contains a single data point, (ii) all response values within the node are the same, or (iii) a depth of $K$ is reached, whichever occurs sooner. The tree $T_0$ consisting of the root node outputs the mean of the entire dataset, $\overline{Y} \coloneqq n^{-1} \sum_{i=1}^n Y_i$.

Our first lemma, Lemma 3.1, shows that maximizing $\widehat{\Delta}(s, j, \mathrm{t})$ is equivalent to maximizing the inner product $\langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}} \rangle_{\mathrm{t}} = \frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} (Y_i - \overline{Y}_{\mathrm{t}}) \Psi_{\mathrm{t}}(\mathbf{X}_i)$ between the residuals $Y - \overline{Y}_{\mathrm{t}}$ and the family of standardized decision stumps

$$\Psi_{\mathrm{t}}(\mathbf{x}) \coloneqq \frac{\mathbf{1}(\mathbf{x} \in \mathrm{t}_L) P(\mathrm{t}_R) - \mathbf{1}(\mathbf{x} \in \mathrm{t}_R) P(\mathrm{t}_L)}{\sqrt{P(\mathrm{t}_L) P(\mathrm{t}_R)}}, \tag{6}$$

which split $X_j$ $(j = 1, 2, \ldots, p)$ at $s$ within t.

Furthermore, Lemma 3.1 shows that the tree output $\widehat{\mu}(T)$ is equal to the (empirical) orthogonal projection of $Y$ onto the linear span of orthonormal decision stumps

$$\psi_{\mathrm{t}}(\mathbf{x}) \coloneqq \Psi_{\mathrm{t}}(\mathbf{x}) \sqrt{1/w(\mathrm{t})}, \tag{7}$$

for internal (nonterminal) nodes t in $T$ (We define $\Psi_{\mathrm{t}}(\mathbf{x}) \equiv \psi_{\mathrm{t}}(\mathbf{x}) \equiv 1$ if t is the root node.). The factor $\sqrt{1/w(\mathrm{t})}$ ensures that $\psi_{\mathrm{t}}(\cdot)$ has unit empirical norm with respect to the empirical measure on $\mathcal{D}_n$. This lemma suggests that there may be some connections between CART and greedy optimization in Hilbert spaces. Indeed, as we shall see, the CART algorithm can be viewed as a "local" orthogonal greedy procedure, in which one projects the data onto the space of all constant predictors within the node. The proofs show that this local greedy optimization is in fact very similar to standard greedy optimization in Hilbert spaces. The reader familiar with greedy algorithms in Hilbert spaces for over-complete dictionaries will recognize some

similarities in the analysis (in particular, the "orthogonal greedy algorithm" [3] in which one projects the data onto the linear span of a finite collection of dictionary elements). A similar parallel with orthogonal basis algorithms was explored for "dyadic" CART in [10]. However, in contrast to the irregularly scattered data assumed here, recursive partitions formed by dyadic CART arise from midpoint splits on equispaced data on a rectangular lattice.

**Lemma 3.1.** *The impurity gain is equal to the squared within-node inner product between the residuals and a standardized decision stump that splits a variable $X_j$ at $s$, i.e.,*

$$\widehat{\Delta}(s, j, \mathrm{t}) = |\langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}} \rangle_{\mathrm{t}}|^2. \tag{8}$$

*Furthermore, if $T$ is a decision tree constructed with CART methodology, then its output admits the orthogonal expansion*

$$\widehat{\mu}(T) = \sum_{\mathrm{t}} \langle Y, \psi_{\mathrm{t}} \rangle_n \psi_{\mathrm{t}}, \tag{9}$$

*where the sum runs over all internal nodes $\mathrm{t}$ in $T$, $\|\psi_{\mathrm{t}}\|_n = 1$, and $\langle \psi_{\mathrm{t}}, \psi_{\mathrm{t}'} \rangle_n = 0$ for distinct internal nodes $\mathrm{t}$ and $\mathrm{t}'$ in $T$. In other words, $\widehat{\mu}(T)$ is the (empirical) orthogonal projection of $Y$ onto the linear span of $\{\psi_{\mathrm{t}}\}_{\mathrm{t}}$.*

The orthogonal decomposition (9) of $\widehat{\mu}(T_K)$ in Lemma 3.1 is useful because it allows us to write down a recursive expression for the training error of CART. That is, from $\widehat{\mu}(T_K) = \widehat{\mu}(T_{K-1}) + \sum_{\mathrm{t} \in T_{K-1}} \langle Y, \psi_{\mathrm{t}} \rangle_n \psi_{\mathrm{t}}$, we have

$$\|Y - \widehat{\mu}(T_K)\|_n^2 = \|Y - \widehat{\mu}(T_{K-1})\|_n^2 - \sum_{\mathrm{t} \in T_{K-1}} |\langle Y, \psi_{\mathrm{t}} \rangle_n|^2, \tag{10}$$

where the sums run over all terminal nodes $\mathrm{t}$ of $T_{K-1}$. This identity is the starting point of our analysis.

## 4. Main results

Our first lemma in this section is the key to all forthcoming results for CART and ensembles thereof. It provides a purely algorithmic guarantee, namely, that the (excess) training error of a depth $K$ regression tree constructed with CART methodology decays like $1/K$. To the best of our knowledge, this result is the first of its kind for CART, or, for that matter, any decision tree algorithm. The math behind it is surprisingly clean; in particular, unlike most past work, we do not need to directly analyze the partition of the input space that is induced by recursively splitting the variables. Nor do we need to rely on concentration of measure to show that certain "local" (i.e., node-specific) empirical quantities concentrate around their population level versions. Because we are able to circumvent these technical aspects with a new method of analysis, the astute reader will notice and appreciate that we make no assumptions on the decision tree itself (such as a minimum node size condition or shrinking cell condition that typifies extant literature). Unlike the recent work [8], we

also do not have an "edge" condition that ensures the prediction error decreases by a constant factor after each split. In sum, we do not deviate from the original CART procedure out of theoretical convenience.

**Lemma 4.1** (Training error bound for CART)**.** *Let $\widehat{\mu}(T_K)$ be the output of a depth $K$ regression tree $T_K$ constructed with CART methodology. Then, for any $K \geq 1$ and any additive function $g \in \mathcal{G}$,*

$$\|Y - \widehat{\mu}(T_K)\|_n^2 \leq \|Y - g\|_n^2 + \frac{\|g\|_{TV}^2}{K+3}.$$

**Remark 1.** *Though we have assumed at the outset that $Y$ is almost surely bounded by $B$, we emphasize that Lemma 4.1 holds for an arbitrary dataset $\mathcal{D}_n$.*

### *4.1. Oracle inequality for CART*

Our second result, in Theorem 4.2, establishes an adaptive risk bound (also known as an "oracle inequality") for CART under model misspecification. Essentially, it says that the CART algorithm adapts to the class $\mathcal{G}$ of bounded variation additive models and performs as if it were finding the best additive approximation (even though it is agnostic to such structure) to the regression function, while accounting for the "capacity" (the total variation $\ell_1$ norm $\|\cdot\|_{\text{TV}}$) of the approximation. In particular, Theorem 4.2 reveals the tradeoff between the goodness-of-fit and complexity relative to sample size. The goodness-of-fit stems from the training error bound in Lemma 4.1 and the descriptive complexity (relative to sample size) comes from the fact that the $\epsilon$-metric entropy for depth $K$ decision trees constructed from $n$ sample points and $p$ variables is of order $2^K \log(np/\epsilon)$.

**Theorem 4.2** (Oracle inequality for CART)**.** *Let $\widehat{\mu}(T_K)$ be the output of a depth $K$ regression tree $T_K$ constructed with CART methodology. Then, for any $K \geq 1$,*

$$\mathbb{E}(\|\mu - \widehat{\mu}(T_K)\|^2) \leq 2 \inf_{g \in \mathcal{G}} \Big\{ \|\mu - g\|^2 + \frac{\|g\|_{TV}^2}{K+3} + C \frac{2^K \log(np)}{n} \Big\}, \tag{11}$$

*where $C$ is a positive constant that depends only on $B$.*

**Remark 2.** *While the prediction error bound in Theorem 4.2 depends only explicitly on $p$ through $\log(p)$, for $\mu \in \mathcal{G}$, the prediction error at the optimal depth unfortunately decays at a slow logarithmic rate, i.e., $(\log(n/\log(p)))^{-1}$. This is a far cry from the standard rate of $\sqrt{(\log(np))/n}$ for bounded variation additive models in high dimensions [25]. The slow logarithmic rate for CART is in line with recent work (see [15, Theorem 5] and [8, Theorem 2]), where $p$ is allowed to grow only polynomially with $n$, i.e., $p = \lfloor\!\lfloor n^c \rfloor\!\rfloor$ for some positive constant $c$. The rate also corroborates with empirical evidence suggesting that the CART algorithm may require very large samples in order to be accurate and may have "difficulty in modeling additive structure" [14, Section 9.2.4]. At the moment, we do not know if the rate $(\log(n/\log(p)))^{-1}$ is optimal for CART within the class of bounded variation additive models. Going forward, one direction for future work would be to derive complementary lower bounds on the mean squared prediction error for the CART algorithm.*

**Remark 3.** *While tree models are highly flexible, their recursive construction can be highly suboptimal, thereby producing inefficiencies in how the data is utilized. For example, consider a single dimension and suppose the regression function is a sinusoid with frequency $K$ on the unit interval, normalized by $4K$ to have unit total variation. Then the CART algorithm tends to produce many unnecessary and wasteful splits near the endpoints of the unit interval (that consequently inflate the tree complexity), instead of splitting closer to the midpoint. This means that even though the number of terminal nodes in a depth $K$ tree is large, i.e., $\Omega(2^K)$, the training error is only of the same order as the squared norm of the regression function, or, $\Omega(1/K^2)$.*

### *4.2. Consistency of CART*

Next, we consider the case when the model is well-specified, i.e., $\mu \in \mathcal{G}$. In this case, choosing a depth of $K = \llbracket (\xi/2)\log_2(n) \rrbracket$ for some $\xi \in (0,1]$, Theorem 4.2 states that the prediction error is bounded by

$$\frac{4\|\mu\|_{\mathrm{TV}}^2}{\xi \log_2(n) + 6} + 2C\frac{\log(np)}{n^{1-\xi/2}}. \tag{12}$$

This bound serves as the basis for our next corollary. Simply put, if the total variation $\ell_1$ norm of the regression function, i.e., $\|\mu\|_{\mathrm{TV}}$, is controlled, then the CART algorithm is consistent even when the dimensionality grows exponentially with the sample size. We should point out that this type of result is impossible with nonadaptive predictors, such vanilla $k$-nearest neighbors or kernel regression, unless the data is preprocessed via some sort of dimensionality reduction technique.

**Corollary 4.3** (High dimensional consistency of CART for well-specified models)**.** *Suppose $\mu_n(\mathbf{x}) = \sum_{j=1}^{p_n} g_{jn}(x_j)$ is a sequence of $p_n$-dimensional additive regression functions that belong to $\mathcal{G}$ and $\sup_n \|\mu_n\|_{TV} < \infty$. If $K_n \to \infty$ and $2^{K_n}(\log(np_n))/n \to 0$ as $n \to \infty$, then the CART algorithm is consistent, that is,*

$$\lim_{n\to\infty} \mathbb{E}(\|\mu_n - \widehat{\mu}(T_{K_n})\|^2) = 0.$$

The hypotheses of Corollary 4.3 are satisfied if, for example, $K_n = \llbracket (\xi/2)\log_2(n) \rrbracket$ and $p_n = \llbracket \exp(cn^{1-\xi}) \rrbracket$ for some constants $c > 0$ and $\xi \in (0,1]$. In this case, from Theorem 4.2, the consistency rate of the CART algorithm is

$$\frac{4\sup_n \|\mu_n\|_{\mathrm{TV}}^2}{\xi \log_2(n) + 6} + \frac{2C\log(n)}{n^{1-\xi/2}} + \frac{2Cc}{n^{\xi/2}} = \mathcal{O}\Big(\frac{1}{\log(n)}\Big).$$

The dependence on the total variation $\ell_1$ norm $\|\mu_n\|_{\mathrm{TV}}$ in the consistency rate above also shows that CART can tolerate an approximate sparsity level that grows as fast as $o(\sqrt{\log(n)})$.

A similar statement to Corollary 4.3 is true for the excess misclassification error (difference between the expected 0-1 loss error for classifying a new observation and

the Bayes risk) for binary classification, i.e., $Y \in \{0,1\}$ and $\mathbb{P}(Y=1|\mathbf{X}) = g(\mathbf{X})$ for some additive function $g \in \mathcal{G}$. This is because the squared error impurity (4) equals one-half of the so-called *Gini* impurity used for classification trees (e.g., trees which output the majority vote in each terminal node) [17, Section 3]. In fact, by a well known inequality for plug-in classifiers [13, Theorem 2.2], the excess misclassification error for the output of classification tree is bounded by twice the root mean squared prediction error for the output of a regression tree (where both trees operate on the same dataset with binary labels).

The reader might be somewhat surprised at Theorem 4.2 and Corollary 4.3, especially since they are qualitatively similar to existing performance guarantees for predictors based on very different principles, like boosting [7] or neural networks [16, 2]. For example, [7, Theorem 1] states that boosting with linear learners is also consistent for a sequence of linear models $\mu_n(\mathbf{x}) = \boldsymbol{\beta}_n^T \mathbf{x}$ on $[0,1]^p$ in the high dimensional regime, i.e., when $p_n = \llbracket \exp(cn^{1-\xi}) \rrbracket$ and $\sup_n \|\mu_n\|_{\mathrm{TV}} = \sup_n \|\boldsymbol{\beta}_n\|_{\ell_1} < \infty$, where $c$ is a positive constant and $\xi \in (0,1]$.

**Remark 4.** *Corollary 4.3 does not offer guidance on how to choose the depth $K_n$. In practice, it is best to let the data decide and therefore cost complexity pruning (i.e., weakest link pruning [6]) is recommended. This would have one first grow a full tree $T_{max}$ (to maximum depth) and then minimize*

$$\|Y - \widehat{\mu}(T)\|_n^2 + \frac{\alpha \log(np)}{n} \#T$$

*over all trees $T$ that can be obtained from $T_{max}$ by iteratively merging its internal nodes, where $\alpha$ is a positive constant and $\#T$ is the number of terminal nodes of $T$. Working with the resulting pruned tree enables one to obtain oracle inequalities of the form (11), but with the advantage of having the infimum over both the depth $K \geq 1$ and additive functions $g \in \mathcal{G}$.*

### *4.3. Rate improvements*

When $p$ is moderately sized and the model class is further restricted, it is likely that the consistency rates of CART can be improved. For example, it was shown in the well-specified, noiseless regression setting (i.e., $\mu \in \mathcal{G}$ and $\varepsilon = Y - \mu(\mathbf{X}) \equiv 0$) that if $\mu(\cdot)$ is a univariate step function and the total number of constant pieces of $\mu(\cdot)$ is $V$, then the squared prediction error achieves the parametric rate $V(\log(n))/n$ [15, Theorem 4]. Furthermore, if the predictor variables are independent, each component function admits a power series representation, and $\mu(\cdot)$ only depends on a subset of $q \ll p$ of the predictor variables, we have the polynomial rate $((p/n)\log(n/p))^{\Omega(1/q)}$ [15, Theorem 3]. Similar polynomial rates in which the exponent of the rate depends only on the sparsity level (and where $p$ grows polynomially with $n$) were provided in [8] for special classes of additive models (e.g., isotonic or piecewise linear component functions with independent predictor variables).

### *4.4. Beyond additive models*

In this paper, we considered additive models primarily because notions of approximate sparsity are easier to define and more intuitive in (ultra) high dimensional settings. Separately, additive models are also amendable to the mathematical study of CART. Interactions terms are more difficult to incorporate into the greedy analysis and, consequently, it is unclear whether a general consistency theory can be developed. We note that, in a departure with the conventional heuristic explanation for the inconsistency of CART, the problem is not always caused by a lack of "identifiability" of the regression function from the marginal projections; that is, a situation where the marginal projection $\mathbb{E}(Y|X_j)$ is constant in some variable $X_j$ (which leads to a zero impurity gain along that direction at the population level) and yet the full projection $\mu(\mathbf{X}) = \mathbb{E}(Y|\mathbf{X})$ is nonconstant in $X_j$. This scenario can occur with additive models and correlated predictor variables, for which our consistency guarantees still hold in the affirmative.[1] Rather, all of our theory for CART rests solely on being able to lower bound the "information gain" $\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t})$ by the squared within-node excess training error $\|g\|_{\mathrm{TV}}^{-2}(\|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y - g\|_{\mathrm{t}}^2)^2$, provided $\|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 \geq \|Y - g\|_{\mathrm{t}}^2$, for *all* nodes t and *all* functions $g \in \mathcal{G}$ (see Lemma 7.1).

For general regression models, an inspection of the proof of Lemma 7.1 reveals that $\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t})$ is lower bounded by the more unwieldy expression $(\omega_g(\mathrm{t})N(\mathrm{t}))^{-2}(\|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y - g\|_{\mathrm{t}}^2)^2$, where $\omega_g(\mathrm{t}) \coloneqq \sup_{\mathbf{x},\mathbf{x}' \in \mathrm{t}} |g(\mathbf{x}) - g(\mathbf{x}')|$ is the oscillation of $g(\cdot)$ within t. This gives rise to the sufficient conditions

$$K_n \to \infty, \quad 2^{K_n}(\log(np_n))/n \to 0, \quad \mathbb{E}\big(\max_{\mathrm{t} \in T_{K_n}} \{\omega_{\mu_n}(\mathrm{t})N(\mathrm{t})\}\big) \to 0, \quad n \to \infty,$$

for mean squared consistency. Here "$\mathrm{t} \in T_{K_n}$" means that t is a terminal node of $T_{K_n}$. This is of course only a partial result as $\mathbb{E}(\max_{\mathrm{t} \in T_{K_n}} \{\omega_{\mu_n}(\mathrm{t})N(\mathrm{t})\}) \to 0$ is impossible to verify without knowledge of the distribution of $(\mathbf{X}, Y)$. To see why this condition may be reasonable, consider the case when $\mu_n(\cdot)$ is Lipschitz smooth and the input data is uniformly distributed on the unit hypercube $[0,1]^p$. Then $\omega_{\mu_n}(\mathrm{t})$ is at most a constant multiple of the diameter of t, which is approximately the fraction of samples that the node contains, or, $N(\mathrm{t})/n$. Thus, it suffices to have $\mathbb{E}(\max_{\mathrm{t} \in T_{K_n}} N^2(\mathrm{t})/n) \to 0$. If the data is roughly balanced across the ($\leq 2^{K_N}$) terminal nodes of $T_{K_n}$, we expect $N(\mathrm{t})$ to be approximately $n/2^{K_n}$. Consequently, it further suffices to have $n/4^{K_n} \to 0$, which is true if $K_n = \llbracket c \log_2(n) \rrbracket$ for any $c > 1/2$.

Another way to ensure consistency (without additional assumptions) for more general models is to, for example, augment the set of $p$ input variables with $p(p-1)/2$ interaction variables $X_j X_{j'}$ for $j < j'$. Therefore, at each node of the tree, we consider splitting along $p + p(p-1)/2$ different variables. Under this setup, statements analogous to Theorem 4.2 and Corollary 4.3 hold for two-way ANOVA models of the form $g(\mathbf{X}) = \sum_j g_j(X_j) + \sum_{j<j'} g_{jj'}(X_j X_{j'})$, for some bounded variation functions $g_j(\cdot)$ and $g_{jj'}(\cdot)$, where now $\|g\|_{\mathrm{TV}} \coloneqq \sum_j v(g_j) + \sum_{j<j'} v(g_{jj'})$. The above reasoning

[1] For example, take $\mu(X_1, X_2) = g_1(X_1) + g_2(X_2)$, where $g_2(X_2) = -\mathbb{E}(g_1(X_1)|X_2)$ and $X_1$ and $X_2$ are correlated. Then $\mathbb{E}(Y|X_2) \equiv 0$, even though $\mu(\cdot)$ depends explicitly on $X_2$.

can be generalized to arbitrary order interactions in the model, but at the cost of greater computational complexity in constructing the tree. [2]

## 5. Random forests

The predictive abilities of individual decision trees should intuitively be inherited by random forests due to the ensemble principle and convexity of squared error (see [9, Proposition 3 and Proposition 4], [5, Section 11], or [4, Section 4.1]). Indeed, our results on the consistency of CART for high dimensional additive models also carry over to Breiman's random forests [5] with relative ease, as we now explain.

### 5.1. Growing the forest

Consider a bootstrap sample $\mathcal{D}_n'$ from the original dataset $\mathcal{D}_n$. From this bootstrapped training sample, we construct a depth $K$ regression tree $T_K$ with CART methodology in the usual way, except that, at each internal node, we select $q$ (also known as "mtry") of the $p$ variables uniformly at random, without replacement, as candidates for splitting. That is, for each internal node t of $T_K$, we generate a random subset $\mathcal{S} \subset \{1, 2, \dots, p\}$ of size $q$ and split along the variable $X_{\hat{j}}$, where $\hat{j} \in \arg\max_{j \in \mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t})$.[3]

We grow $M$ of these depth $K$ regression trees separately using, respectively, $M$ independent realizations $\mathbf{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_M)^T$ of a random variable $\Theta$. Here $\Theta$ is distributed according to the law that generates the bootstrapped training data and candidate variables for splitting at each of the nodes. The output of the $m^{\text{th}}$ regression tree is denoted by $\widehat{\mu}(\Theta_m) = \widehat{\mu}(\Theta_m, \mathcal{D}_n)$. With this notation in place, the random forest output is then simply the empirical average of the $M$ regression tree outputs, namely,

$$\widehat{\mu}(\mathbf{\Theta}) = \widehat{\mu}(\mathbf{\Theta}, \mathcal{D}_n) \coloneqq M^{-1} \sum_{m=1}^{M} \widehat{\mu}(\Theta_m). \tag{13}$$

### 5.2. Oracle inequality for random forests

By a modification of the proofs of Lemma 4.1 and Theorem 4.2, it is possible to show the following oracle inequality for random forests. The proof is furnished in Section 7.

**Theorem 5.1** (Oracle inequality for random forests)**.** *For all $K \geq 1$,*

$$\mathbb{E}_{\mathbf{\Theta}, \mathcal{D}_n}(\|\mu - \widehat{\mu}(\mathbf{\Theta})\|^2) \leq 2 \inf_{g \in \mathcal{G}} \Big\{ \|\mu - g\|^2 + \frac{p}{q} \frac{\|g\|_{TV}^2}{K+3} + C \frac{2^K \log(np)}{n} \Big\},$$

*where $C$ is a positive constant that depends only on $B$.*

[2] The main caveat of this approach is that the computational complexity of growing the tree increases from $\mathcal{O}(pn \log(n))$ to $\mathcal{O}(p^2 n \log(n))$, or more generally for $d$-way interactions, $\mathcal{O}(p^d n \log(n))$.

[3] We deviate slightly from Breiman's original random forests [5] by not growing the constituent trees to maximum depth (i.e., trees whose terminal nodes contain only a single observation). For consistency of random forests with fully grown trees, see [23, Theorem 2].

To the best of our knowledge, Theorem 5.1 is one of the first results in the literature that shows explicitly the impact on the prediction error from randomly choosing subsets of variables as candidates for splitting at the nodes, without any restrictive assumptions on the tree or data generating process. Even though it is not captured by Theorem 5.1, empirically, the random variable selection mechanism of forests has the effect of de-correlating and encouraging diversity among the constituent trees, which can greatly improve the performance. It also reduces the computational time of constructing each tree, since the optimal split points do not need to be calculated for every variable at each node. What Theorem 5.1 does reveal, however, is that this mechanism cannot hurt the prediction error beyond a benign factor of $p/q$. In fact, standard implementations of regression forests use a default value of $q$ equal to $[\![p/3]\!]$. With this choice, we see by comparing Theorem 5.1 and Theorem 4.2 that there is essentially no loss in performance (at most, a factor of $p/q = 3$) over individual trees, despite not optimizing over the full set of variables at the internal nodes or not using the full dataset $\mathcal{D}_n$ in the constituent trees. It is also interesting to note that we recover the bound (11) for individual trees when $q = p$.

**Remark 5.** *As the number of trees $M$ approaches infinity, we note that $\widehat{\mu}(\boldsymbol{\Theta}) \to \mathbb{E}_{\Theta|\mathcal{D}_n}(\widehat{\mu}(\Theta))$ almost surely, by the strong law of large numbers. Thus, as $M \to \infty$,*

$$\mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}_n}(\|\mu - \widehat{\mu}(\boldsymbol{\Theta})\|^2) \to \mathbb{E}_{\mathcal{D}_n}(\|\mu - \mathbb{E}_{\Theta|\mathcal{D}_n}(\widehat{\mu}(\Theta))\|^2).$$

### *5.3. Consistency of random forests*

We also have a consistency result for forests, analogous to Corollary 4.3 for individual trees.

**Corollary 5.2** (High dimensional consistency of random forests for well-specified models)**.** *Suppose $\mu_n(\mathbf{x}) = \sum_{j=1}^{p_n} g_{jn}(x_j)$ is a sequence of $p_n$-dimensional additive regression functions that belong to $\mathcal{G}$ and $\sup_n \|\mu_n\|_{TV} < \infty$. If $(q_n/p_n)K_n \to \infty$ and $2^{K_n}(\log(np_n))/n \to 0$ as $n \to \infty$, then random forests are consistent, that is,*

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}_n}(\|\mu_n - \widehat{\mu}(\boldsymbol{\Theta})\|^2) = 0.$$

As with the consistency statement for CART in Corollary 4.3, the hypotheses of Corollary 5.2 are satisfied if, for example, $K_n = [\![(\xi/2)\log_2(n)]\!]$, $p_n = [\![\exp(cn^{1-\xi})]\!]$, and $q_n = [\![p_n/3]\!]$, for some constants $c > 0$ and $\xi \in (0, 1]$. It is also interesting to note that consistency is still possible even if only a vanishing fraction of variables are randomly selected at each node, i.e.,

$$q_n = o(p_n) \quad \text{as long as} \quad (q_n/p_n)K_n \to \infty \quad \text{and} \quad 2^{K_n}(\log(np_n))/n \to 0.$$

This paves the way for some further comments. For classification problems, the default value of $q_n$ is $[\![\sqrt{p_n}]\!]$, in which case consistency is possible when $p_n = o(\log^2(n))$ and $K_n = [\![(1/2)\log_2(n)]\!]$. At the extreme end, consistency holds even when $q_n \equiv 1$; that is, only a single coordinate is selected at random at each node, provided $p_n = o(\log(n))$

and $K_n = [\![(1/2)\log_2(n)]\!]$. Again, these specifications could produce major savings in the computational cost of growing the forest.

Corollary 5.2 provides a partial answer to a problem posed by [23]: "It remains that a substantial research effort is still needed to understand the properties of forests in a high-dimensional setting, when $p = p_n$ may be substantially larger than the sample size." More specifically, Corollary 5.2 strengthens [23, Theorem 1] from the same authors, who show that random forests are consistent for additive models when $p$ is fixed, the component functions are continuous, and (in our notation) $K_n \to \infty$ and $2^{K_n}(\log(n))/n \to 0$ as $n \to \infty$.[4] In contrast, here we allow the dimensionality to grow exponentially with the sample size and also for the component functions to be possibly discontinuous (see the next subsection for even weaker requirements on the component functions). The latter point has practical implications as certain processes encountered in natural and social sciences exhibit, for example, thresholding behavior. Another (minor) difference from [23] is that here, congruent with Breiman's original algorithm, we grow the constituent trees with bootstrap samples from $\mathcal{D}_n$ instead of subsamples of size $a_n$ ($\leq n$), i.e., random sampling without replacement. However, the two sampling schemes are roughly the same when $a_n = [\![0.632n]\!]$.

### *5.4. Beyond bounded variation component functions*

Up to this point, we have only considered additive regression functions whose components have bounded variation. While this collection is broad, it still may be restrictive, especially if the data is highly oscillatory. The allowance for model misspecification means that we can go far beyond bounded variation component functions, as our final result reveals. As a byproduct of our new analysis, we generalize the aforementioned theory in [23] and show that random forests are consistent for sparse additive models with bounded (and measurable) component functions—a much weaker requirement than continuity or bounded variation. Before we state this result, we define the $\ell_0$ norm of an additive function $g(\mathbf{x}) = \sum_{j=1}^{p} g_j(x_j)$ as $\|g\|_{\ell_0} := \{j : g_j(\cdot) \text{ is nonconstant}\}$. Roughly speaking, this norm counts the number of relevant variables that affect $g(\cdot)$.

**Corollary 5.3** (Consistency for bounded and measurable component functions)**.** *Suppose $\mu_n(\mathbf{x}) = \sum_{j=1}^{p_n} g_j(x_j)$ is a sequence of $p_n$-dimensional additive regression functions, where each component function $g_j(\cdot)$ is bounded and measurable. Furthermore, assume that $\sup_n \|\mu_n\|_{\ell_0} < \infty$. If $(q_n/p_n)K_n \to \infty$ and $2^{K_n}(\log(np_n))/n \to 0$ as $n \to \infty$, then random forests are consistent, that is,*

$$\lim_{n\to\infty} \mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}_n}(\|\mu_n - \widehat{\mu}(\boldsymbol{\Theta})\|^2) = 0.$$

**Remark 6.** *A similar consistency statement as Corollary 5.3 is also valid for single decision trees. More precisely, suppose $\mu_n(\mathbf{x}) = \sum_{j=1}^{p_n} g_j(x_j)$ is a sequence of $p_n$-dimensional additive regression functions, where each component function $g_j(\cdot)$ is*

[4]The condition $2^{K_n}(\log(n))/n \to 0$ is actually stated as $2^{K_n}(\log(n))^9/n \to 0$ in [23, Theorem 1], where Gaussian errors are assumed, but the former condition is sufficient when the noise is bounded.

*bounded and measurable. Furthermore, assume that* $\sup_n \|\mu_n\|_{\ell_0} < \infty$. *If* $K_n \to \infty$ *and* $2^{K_n}(\log(np_n))/n \to 0$ *as* $n \to \infty$, *then the CART algorithm is consistent, that is,*

$$\lim_{n\to\infty} \mathbb{E}(\|\mu_n - \widehat{\mu}(T_{K_n})\|^2) = 0.$$

## 6. Conclusion

We end the paper with a few remarks. It is still largely a mystery (at least theoretically) why bagging and the random variable selection mechanism of random forests are so effective at reducing the prediction error. Our results only show that these apparatuses do not degrade the performance beyond small factors. Fascinating recent work by [19] shows that $q$ ("mtry") plays a similar role as the shrinkage penalty in explicitly regularized procedures. More specifically, when $p > n$, they show that if an ensemble predictor is formed by averaging over many linear regression models with orthogonal designs and randomly selected subsets of variables, then asymptotically as the number of models goes to infinity, the coefficient vector of the ensemble is "shrunk" by a factor of $q/p$. It is possible that a similar implicit regularization phenomenon is occurring in our high dimensional additive setting, which may lead to even better performance over individual trees. Certainly more work needs to be done to answer these questions.

## 7. Proofs

In this section, we provide complete proofs of Lemma 3.1, Lemma 4.1, Theorem 4.2, Corollary 4.3, Theorem 5.1, Corollary 5.2, and Corollary 5.3.

*Proof of Lemma 3.1.* To prove the first assertion (8), note that

$$\begin{aligned}
|\langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}}\rangle_{\mathrm{t}}|^2 &= \Big|\Big\langle Y - \overline{Y}_{\mathrm{t}}, \frac{\mathbf{1}(\mathbf{X} \in \mathrm{t}_L)N(\mathrm{t}_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}_R)N(\mathrm{t}_L)}{\sqrt{N(\mathrm{t}_L)N(\mathrm{t}_R)}}\Big\rangle_{\mathrm{t}}\Big|^2 \\
&= \Big|\frac{1}{N(\mathrm{t})}\frac{\overline{Y}_{\mathrm{t}_L}N(\mathrm{t}_L)N(\mathrm{t}_R) - \overline{Y}_{\mathrm{t}_R}N(\mathrm{t}_R)N(\mathrm{t}_L)}{\sqrt{N(\mathrm{t}_L)N(\mathrm{t}_R)}}\Big|^2 \\
&= \frac{N(\mathrm{t}_L)N(\mathrm{t}_R)}{N^2(\mathrm{t})}(\overline{Y}_{\mathrm{t}_L} - \overline{Y}_{\mathrm{t}_R})^2 = \widehat{\Delta}(s, j, \mathrm{t}),
\end{aligned}$$

where the final equality comes from [6, Section 9.3].

We now turn our attention to (9). That is, we decompose the tree output $\widehat{\mu}(T)$ into an additive expansion of the orthonormal decision stumps $\psi_{\mathrm{t}}$. To this end, we first associate each internal node of $T$ with a decision stump $\psi_{\mathrm{t}}$. Then for each internal node t, notice that

$$\begin{aligned}
&\langle Y - \overline{Y}_{\mathrm{t}}, \psi_{\mathrm{t}}\rangle_n \psi_{\mathrm{t}}(\mathbf{x}) \\
&= \Big\langle Y - \overline{Y}_{\mathrm{t}}, \frac{\mathbf{1}(\mathbf{X} \in \mathrm{t}_L)P(\mathrm{t}_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}_R)P(\mathrm{t}_L)}{\sqrt{w(\mathrm{t})P(\mathrm{t}_L)P(\mathrm{t}_R)}}\Big\rangle_n \frac{\mathbf{1}(\mathbf{x} \in \mathrm{t}_L)P(\mathrm{t}_R) - \mathbf{1}(\mathbf{x} \in \mathrm{t}_R)P(\mathrm{t}_L)}{\sqrt{w(\mathrm{t})P(\mathrm{t}_L)P(\mathrm{t}_R)}} \\
&= (\overline{Y}_{\mathrm{t}_L} - \overline{Y}_{\mathrm{t}})\mathbf{1}(\mathbf{x} \in \mathrm{t}_L) + (\overline{Y}_{\mathrm{t}_R} - \overline{Y}_{\mathrm{t}})\mathbf{1}(\mathbf{x} \in \mathrm{t}_R).
\end{aligned} \tag{14}$$

For a terminal node t of $T$, let $\mathrm{t}_0, \mathrm{t}_1, \ldots, \mathrm{t}_K$ be the unique downward path from the root node $\mathrm{t}_0$ to the terminal node $\mathrm{t}_K = \mathrm{t}$. Next, we sum up (14) over all internal nodes in $T$. The tree output $\overline{Y}_{\mathrm{t}}$ corresponding to a terminal region $\mathbf{x} \in \mathrm{t}$ minus the output $\overline{Y}$ in the root node is a telescoping sum of the successive (internal) node outputs:

$$\sum_{k=0}^{K-1} (\overline{Y}_{\mathrm{t}_{k+1}} - \overline{Y}_{\mathrm{t}_k}) = \overline{Y}_{\mathrm{t}_K} - \overline{Y}_{\mathrm{t}_0} = \overline{Y}_{\mathrm{t}} - \overline{Y} = \overline{Y}_{\mathrm{t}} - \langle Y, \psi_{\mathrm{t}_0} \rangle_n \psi_{\mathrm{t}_0}. \tag{15}$$

Combining (14) and (15), we have the expansion

$$(\widehat{\mu}(T))(\mathbf{x}) = \sum_{\mathrm{t} \in T} \overline{Y}_{\mathrm{t}} \mathbf{1}(\mathbf{x} \in \mathrm{t}) = \sum_{\mathrm{t}} \langle Y, \psi_{\mathrm{t}} \rangle_n \psi_{\mathrm{t}}(\mathbf{x}), \tag{16}$$

where the sum on the left side extends over all terminal nodes of $T$ and the sum on the right side extends over all internal nodes in $T$. Next, we show that the decision stumps $\{\psi_{\mathrm{t}}\}_{\mathrm{t}}$ in the expansion (16) are orthonormal. First, notice that each $\psi_{\mathrm{t}}$ has unit empirical norm since

$$\begin{aligned} \|\psi_{\mathrm{t}}\|_n^2 &= \frac{\|\mathbf{1}(\mathbf{X} \in \mathrm{t}_L) P(\mathrm{t}_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}_R) P(\mathrm{t}_L)\|_n^2}{w(\mathrm{t}) P(\mathrm{t}_L) P(\mathrm{t}_R)} \\ &= \frac{1}{n} \frac{N(\mathrm{t}_L) N^2(\mathrm{t}_R) + N(\mathrm{t}_R) N^2(\mathrm{t}_L)}{(N(\mathrm{t})/n) N(\mathrm{t}_L) N(\mathrm{t}_R)} = 1, \end{aligned}$$

where we used $N(\mathrm{t}_L) + N(\mathrm{t}_R) = N(\mathrm{t})$. Next, each $\psi_{\mathrm{t}}$ (where t is not the root node) is orthogonal to the constant function since

$$\begin{aligned} \langle 1, \psi_{\mathrm{t}} \rangle_n &= \Big\langle 1, \frac{\mathbf{1}(\mathbf{X} \in \mathrm{t}_L) P(\mathrm{t}_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}_R) P(\mathrm{t}_L)}{\sqrt{w(\mathrm{t}) P(\mathrm{t}_L) P(\mathrm{t}_R)}} \Big\rangle_n \\ &= \frac{1}{n} \frac{N(\mathrm{t}_L) N(\mathrm{t}_R) - N(\mathrm{t}_R) N(\mathrm{t}_L)}{\sqrt{w(\mathrm{t}) N(\mathrm{t}_L) N(\mathrm{t}_R)}} = 0. \end{aligned}$$

Furthermore, for distinct internal nodes t and t′ in $T$, we have

$$\langle \psi_{\mathrm{t}}, \psi_{\mathrm{t}'} \rangle_n = 0. \tag{17}$$

To see this, first note that (17) is clear if t and t′ are not connected via a downward path in the tree, since the corresponding indicator variables in the definitions of $\psi_{\mathrm{t}}$ and $\psi_{\mathrm{t}'}$ will zero each other out when multiplied, i.e., $\psi_{\mathrm{t}}(\mathbf{x}) \psi_{\mathrm{t}'}(\mathbf{x}) = 0$. On the other hand, assume t is an ancestor of t′ and (without loss of generality) that t′ is a descendant of $\mathrm{t}_L$ or is equal to $\mathrm{t}_L$. Then we have

$$\begin{aligned} &\langle \psi_{\mathrm{t}}, \psi_{\mathrm{t}'} \rangle_n \\ &= \Big\langle \frac{\mathbf{1}(\mathbf{X} \in \mathrm{t}_L) P(\mathrm{t}_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}_R) P(\mathrm{t}_L)}{\sqrt{w(\mathrm{t}) P(\mathrm{t}_L) P(\mathrm{t}_R)}}, \frac{\mathbf{1}(\mathbf{X} \in \mathrm{t}'_L) P(\mathrm{t}'_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}'_R) P(\mathrm{t}'_L)}{\sqrt{w(\mathrm{t}') P(\mathrm{t}'_L) P(\mathrm{t}'_R)}} \Big\rangle_n \\ &= \frac{\langle \mathbf{1}(\mathbf{X} \in \mathrm{t}_L) N(\mathrm{t}_R), \mathbf{1}(\mathbf{X} \in \mathrm{t}'_L) N(\mathrm{t}'_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}'_R) N(\mathrm{t}'_L) \rangle_n}{\sqrt{w(\mathrm{t}) N(\mathrm{t}_L) N(\mathrm{t}_R) w(\mathrm{t}') N(\mathrm{t}'_L) N(\mathrm{t}'_R)}} \\ &= \frac{N(\mathrm{t}'_L) N(\mathrm{t}_R) N(\mathrm{t}'_R) - N(\mathrm{t}'_R) N(\mathrm{t}_R) N(\mathrm{t}'_L)}{n \sqrt{w(\mathrm{t}) N(\mathrm{t}_L) N(\mathrm{t}_R) w(\mathrm{t}') N(\mathrm{t}'_L) N(\mathrm{t}'_R)}} = 0, \end{aligned}$$

thus completing the proof. □

*Proof of Lemma 4.1.* To begin, let $g \in \mathcal{G}$ and define $R_K \coloneqq \|Y - \widehat{\mu}(T_K)\|_n^2 - \|Y - g\|_n^2$ as the excess training error. If $R_{K-1} < 0$, then there is nothing to prove because $R_K \le R_{K-1} < 0$ and hence $\|Y - \widehat{\mu}(T_K)\|_n^2 < \|Y - g\|_n^2$.[5] Therefore, we assume throughout that $R_{K-1} \ge 0$. For a terminal node t of $T_K$, we define the node-specific excess training error as $R_K(\mathrm{t}) \coloneqq \|Y - \overline{Y}_\mathrm{t}\|_\mathrm{t}^2 - \|Y - g\|_\mathrm{t}^2$ so that $R_K = \sum_{\mathrm{t} \in T_K} w(\mathrm{t}) R_K(\mathrm{t})$. We recall that $w(\mathrm{t}) = N(\mathrm{t})/n$ so that $\sum_{\mathrm{t} \in T_K} w(\mathrm{t}) = 1$. Here we use the notation "$\mathrm{t} \in T_K$" in the sum to mean that t is a terminal node of $T_K$. By the orthogonal decomposition of the tree output $\widehat{\mu}(T_K)$ from (9) of Lemma 3.1, we have

$$\|Y - \widehat{\mu}(T_K)\|_n^2 = \|Y - \widehat{\mu}(T_{K-1})\|_n^2 - \sum_{\mathrm{t} \in T_{K-1}} |\langle Y, \psi_\mathrm{t} \rangle_n|^2, \tag{18}$$

where again we emphasize that the sum is taken over all terminal nodes t of $T_{K-1}$. Subtracting $\|Y - g\|_n^2$ from both sides of (18) and using the definition of $R_K$, we have

$$R_K = R_{K-1} - \sum_{\mathrm{t} \in T_{K-1}} |\langle Y, \psi_\mathrm{t} \rangle_n|^2. \tag{19}$$

Furthermore, according to the definitions of $\Psi_\mathrm{t}$ in (6) and $\psi_\mathrm{t}$ in (7), we have $\langle Y, \psi_\mathrm{t} \rangle_n = \sqrt{w(\mathrm{t})} \langle Y - \overline{Y}_\mathrm{t}, \Psi_\mathrm{t} \rangle_\mathrm{t}$, and thus (19) can be rewritten as

$$R_K = R_{K-1} - \sum_{\mathrm{t} \in T_{K-1}} w(\mathrm{t}) |\langle Y - \overline{Y}_\mathrm{t}, \Psi_\mathrm{t} \rangle_\mathrm{t}|^2. \tag{20}$$

Throwing away terms in the sum for $\mathrm{t} \in T_{K-1}$ such that $R_{K-1}(\mathrm{t}) < 0$, note that (20) satisfies the inequality

$$R_K \le R_{K-1} - \sum_{\mathrm{t} \in T_{K-1} : R_{K-1}(\mathrm{t}) \ge 0} w(\mathrm{t}) |\langle Y - \overline{Y}_\mathrm{t}, \Psi_\mathrm{t} \rangle_\mathrm{t}|^2. \tag{21}$$

By Lemma 7.1, if t is a terminal node of $T_{K-1}$ and $R_{K-1}(\mathrm{t}) \ge 0$, we have

$$\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \ge \frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathrm{TV}}^2}. \tag{22}$$

The identity (8) from Lemma 3.1 and (22) together imply that

$$|\langle Y - \overline{Y}_\mathrm{t}, \Psi_\mathrm{t} \rangle_\mathrm{t}|^2 \ge \frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathrm{TV}}^2}. \tag{23}$$

Next, we apply (23) to each term in the sum from (21) to obtain

$$R_K \le R_{K-1} - \frac{1}{\|g\|_{\mathrm{TV}}^2} \sum_{\mathrm{t} \in T_{K-1} : R_{K-1}(\mathrm{t}) \ge 0} w(\mathrm{t}) R_{K-1}^2(\mathrm{t}). \tag{24}$$

[5] It can be seen from (10) that the training error of CART decreases with the depth so that $R_1 \ge R_2 \ge \cdots \ge R_K$.

Let $R_{K-1}^+ \coloneqq \sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t})R_{K-1}(\mathrm{t})$ and $R_{K-1}^- \coloneqq \sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})< 0} w(\mathrm{t})R_{K-1}(\mathrm{t})$ and note that $R_{K-1} = R_{K-1}^+ + R_{K-1}^-$.

By convexity of the square function and the fact that $\sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t}) \leq 1$, we have from Jensen's inequality that

$$\sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t})R_{K-1}^2(\mathrm{t}) \geq \Big(\sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t})R_{K-1}(\mathrm{t})\Big)^2 = (R_{K-1}^+)^2. \tag{25}$$

Applying (25) to (24), we obtain

$$R_K \leq R_{K-1} - \frac{1}{\|g\|_{\mathrm{TV}}^2}(R_{K-1}^+)^2. \tag{26}$$

Since $R_{K-1}^+ \geq R_{K-1}^+ + R_{K-1}^- = R_{K-1}$ and $R_{K-1}$ is nonnegative by assumption, from (26), we have established the recursion

$$R_K \leq R_{K-1}(1 - R_{K-1}/\|g\|_{\mathrm{TV}}^2), \tag{27}$$

provided $R_{K-1} \geq 0$. It is now easy to prove that $R_K \leq \|g\|_{\mathrm{TV}}^2/(K+3)$. The base case $K = 1$ is established by noticing that $R_1 \leq R_0(1 - R_0/\|g\|_{\mathrm{TV}}^2) \leq \|g\|_{\mathrm{TV}}^2/4$. For $K > 1$, assume that $R_{K-1} \leq \|g\|_{\mathrm{TV}}^2/(K+2)$. Then either $R_{K-1} \leq \|g\|_{\mathrm{TV}}^2/(K+3)$, in which case we are done since $R_K \leq R_{K-1}$, or $R_{K-1} > \|g\|_{\mathrm{TV}}^2/(K+3)$, in which case

$$R_K \leq R_{K-1}\Big(1 - \frac{R_{K-1}}{\|g\|_{\mathrm{TV}}^2}\Big) \leq \frac{\|g\|_{\mathrm{TV}}^2}{K+2}\Big(1 - \frac{1}{K+3}\Big) = \frac{\|g\|_{\mathrm{TV}}^2}{K+3}. \qquad \square$$

**Lemma 7.1.** *Let* $\mathrm{t}$ *be a terminal node of* $T_{K-1}$ *and* $g \in \mathcal{G}$ *and define* $R_{K-1}(\mathrm{t}) \coloneqq \|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y - g\|_{\mathrm{t}}^2$. *Then,*

$$\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \geq \frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{TV}^2},$$

*provided* $R_{K-1}(\mathrm{t}) \geq 0$.

*Proof of Lemma 7.1.* We use recent tools for the analysis of decision trees from [15], namely, Lemma A.4 in the supplement therein. For daughter nodes $\mathrm{t}_L$ and $\mathrm{t}_R$, define an empirical measure $\Pi(s, j)$ on split points $s$ and variables $X_j$, having Radon-Nikodym derivative (with respect to Lebesgue measure and counting measure)

$$\frac{d\Pi(s,j)}{d(s,j)} \coloneqq \frac{|g_j'(s)|\sqrt{P(\mathrm{t}_L)P(\mathrm{t}_R)}}{\sum_{j'=1}^p \int |g_{j'}'(s')|\sqrt{P(\mathrm{t}_L')P(\mathrm{t}_R')}ds'},$$

where $g_j'(\cdot)$ is shorthand for the divided difference of $g_j(\cdot)$ for the successive ordered data points along the $j^{\text{th}}$ direction within $\mathrm{t}$. That is, if $X_1' \leq X_2' \leq \cdots \leq X_{N(\mathrm{t})}'$ denotes the ordered data along the $j^{\text{th}}$ direction within $\mathrm{t}$, then $g_j'(s) \coloneqq (g_j(X_{i+1}') -$

$g_j(X'_i))/(X'_{i+1} - X'_i)$ for $X'_i \le s < X'_{i+1}$ and $i = 1, 2, \dots, N(\mathrm{t}) - 1$.[6] For notational brevity, we omit the explicit dependence of $P(\mathrm{t}_L)$ and $P(\mathrm{t}_R)$ on $s$ and $j$ and $P(\mathrm{t}'_L)$ and $P(\mathrm{t}'_R)$ on $s'$ and $j'$.

Since $\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t})$ is by definition the maximum of $\widehat{\Delta}(s, j, \mathrm{t})$ over $s$ and $j$, we have from the fact that a maximum is larger than average

$$\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \ge \int \widehat{\Delta}(s, j, \mathrm{t}) d\Pi(s, j) = \int |\langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}} \rangle_{\mathrm{t}}|^2 d\Pi(s, j), \tag{28}$$

where the last identity follows from (8) in Lemma 3.1. Next, by Jensen's inequality for the square function, (28) is further lower bounded by

$$\int |\langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}} \rangle_{\mathrm{t}}|^2 d\Pi(s, j) \ge \Big( \int |\langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}} \rangle_{\mathrm{t}}| d\Pi(s, j) \Big)^2. \tag{29}$$

We next evaluate the expectation in (29) with respect to the measure $\Pi$, giving

$$\int |\langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}} \rangle_{\mathrm{t}}| d\Pi(s, j) = \frac{\sum_{j=1}^p \int |g'_j(s)| |\langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}}| ds}{\sum_{j'=1}^p \int |g'_{j'}(s')| \sqrt{P(\mathrm{t}'_L) P(\mathrm{t}'_R)} ds'}, \tag{30}$$

where we used the identity $\sqrt{P(\mathrm{t}_L) P(\mathrm{t}_R)} \langle Y - \overline{Y}_{\mathrm{t}}, \Psi_{\mathrm{t}} \rangle_{\mathrm{t}} = -\langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}}$, which follows from $\mathbf{1}(\mathbf{X} \in \mathrm{t}_L) P(\mathrm{t}_R) - \mathbf{1}(\mathbf{X} \in \mathrm{t}_R) P(\mathrm{t}_L) = -(\mathbf{1}(X_j > s) - P(\mathrm{t}_R)) \mathbf{1}(\mathbf{X} \in \mathrm{t})$.

Continuing on the numerator in (30), we use the fact that the integral of the absolute value is at least the absolute value of the integral, yielding

$$\sum_{j=1}^p \int |g'_j(s)| |\langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}}| ds \ge \Big| \sum_{j=1}^p \int g'_j(s) \langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}} ds \Big|. \tag{31}$$

Using linearity of the inner product and integration and the fundamental theorem of calculus, the expression in the right hand side of (31) can be simplified as follows

$$\begin{aligned} \sum_{j=1}^p \int g'_j(s) \langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}} ds &= \Big\langle Y - \overline{Y}_{\mathrm{t}}, \sum_{j=1}^p \int g'_j(s) \mathbf{1}(X_j > s) ds \Big\rangle_{\mathrm{t}} \\ &= \Big\langle Y - \overline{Y}_{\mathrm{t}}, \sum_{j=1}^p g_j \Big\rangle_{\mathrm{t}} \\ &= \langle Y - \overline{Y}_{\mathrm{t}}, g \rangle_{\mathrm{t}}. \end{aligned} \tag{32}$$

Combining all of these inequalities, namely, (28)-(32), proves that

$$\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \ge \frac{|\langle Y - \overline{Y}_{\mathrm{t}}, g \rangle_{\mathrm{t}}|^2}{\big( \sum_{j'=1}^p \int |g'_{j'}(s')| \sqrt{P(\mathrm{t}'_L) P(\mathrm{t}'_R)} ds' \big)^2}. \tag{33}$$

[6] Observe that this definition of $g'_j(\cdot)$ coincides with the derivative of the function that linearly interpolates the points $\{(X'_1, g_j(X'_1)), (X'_2, g_j(X'_2)), \dots, (X'_{N(\mathrm{t})}, g_j(X'_{N(\mathrm{t})}))\}$.

Our next goal is to provide respective lower and upper bounds on the numerator and denominator of (33). For the denominator of (33), note that for each $j'$,

$$\begin{aligned}\int |g'_{j'}(s')|\sqrt{P(\mathrm{t}'_L)P(\mathrm{t}'_R)}ds' &= \sum_{i=0}^{N(\mathrm{t})}\int_{N(\mathrm{t}'_L)=i}|g'_{j'}(s')|\sqrt{(i/N(\mathrm{t}))(1-i/N(\mathrm{t}))}ds'\\ &= \sum_{i=1}^{N(\mathrm{t})-1}\int_{X'_i}^{X'_{i+1}}|g'_{j'}(s')|ds'\sqrt{(i/N(\mathrm{t}))(1-i/N(\mathrm{t}))}\\ &= \sum_{i=1}^{N(\mathrm{t})-1}|g_{j'}(X'_{i+1})-g_{j'}(X'_i)|\sqrt{(i/N(\mathrm{t}))(1-i/N(\mathrm{t}))}\\ &\le 2^{-1}\sum_{i=1}^{N(\mathrm{t})-1}|g_{j'}(X'_{i+1})-g_{j'}(X'_i)|\\ &\le 2^{-1}v(g_{j'}),\end{aligned}$$

and hence, summing over $j' = 1, 2, \dots, p$, we obtain

$$\sum_{j'=1}^{p}\int |g'_{j'}(s')|\sqrt{P(\mathrm{t}'_L)P(\mathrm{t}'_R)}ds' \le 2^{-1}\|g\|_{\mathrm{TV}}. \tag{34}$$

For the numerator of (33), we use the Cauchy-Schwarz inequality to lower bound $\langle Y-\overline{Y}_{\mathrm{t}}, g\rangle_{\mathrm{t}}$ by

$$\langle Y-\overline{Y}_{\mathrm{t}}, g\rangle_{\mathrm{t}} = \langle Y-\overline{Y}_{\mathrm{t}}, Y\rangle_{\mathrm{t}} + \langle Y-\overline{Y}_{\mathrm{t}}, g-Y\rangle_{\mathrm{t}} \ge \|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}\|Y-g\|_{\mathrm{t}},$$

where we also used $\langle Y-\overline{Y}_{\mathrm{t}}, Y\rangle_{\mathrm{t}} = \|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2$. Now, by the AM–GM inequality, $\|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}\|Y-g\|_{\mathrm{t}} \le \frac{\|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2+\|Y-g\|_{\mathrm{t}}^2}{2}$, and hence $\langle Y-\overline{Y}_{\mathrm{t}}, g\rangle_{\mathrm{t}} \ge 2^{-1}(\|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y-g\|_{\mathrm{t}}^2) = 2^{-1}R_{K-1}(\mathrm{t})$. Squaring both sides and using the assumption $R_{K-1}(\mathrm{t}) \ge 0$, we have

$$|\langle Y-\overline{Y}_{\mathrm{t}}, g\rangle_{\mathrm{t}}|^2 \ge 4^{-1}(\|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y-g\|_{\mathrm{t}}^2)^2 = 4^{-1}R_{K-1}^2(\mathrm{t}). \tag{35}$$

Applying inequalities (34) and (35) to (33), we therefore have shown that

$$\widehat{\Delta}(\hat{s},\hat{j},\mathrm{t}) \ge \frac{4^{-1}R_{K-1}^2(\mathrm{t})}{(2^{-1}\|g\|_{\mathrm{TV}})^2} = \frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathrm{TV}}^2},$$

which completes the proof. □

*Proof of Theorem 4.2.* We first write $\|\mu-\widehat{\mu}(T_K)\|^2 = E_1 + E_2$, where

$$E_1 := \|\mu-\widehat{\mu}(T_K)\|^2 - 2(\|Y-\widehat{\mu}(T_K)\|_n^2 - \|Y-\mu\|_n^2) - \alpha - \beta \tag{36}$$

and

$$E_2 := 2(\|Y-\widehat{\mu}(T_K)\|_n^2 - \|Y-\mu\|_n^2) + \alpha + \beta,$$

and $\alpha$ and $\beta$ are positive constants to be chosen later. Notice that by Lemma 4.1, we have

$$E_2 \le 2(\|Y-g\|_n^2 - \|Y-\mu\|_n^2) + \frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + \alpha + \beta, \tag{37}$$

for any $g \in \mathcal{G}$. Taking expectations on both sides of (37) and using $\mathbb{E}(\|Y-g\|_n^2 - \|Y-\mu\|_n^2) = \|\mu - g\|^2$ yields

$$\begin{aligned}\mathbb{E}(E_2) &\le 2\mathbb{E}(\|Y-g\|_n^2 - \|Y-\mu\|_n^2) + \frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + \alpha + \beta \\ &= 2\|\mu - g\|^2 + \frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + \alpha + \beta.\end{aligned} \tag{38}$$

To bound $E_1$, we first introduce a few useful concepts and definitions due to [20] for studying data-dependent partitions. Let

$$\Lambda_n := \{\mathcal{P}(\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \dots, (\mathbf{x}_n, y_n)\}) : (\mathbf{x}_i, y_i) \in \mathbb{R}^p \times \mathbb{R}\}$$

be the family of all achievable partitions $\mathcal{P}$ by growing a depth $K$ binary tree on $n$ points (in particular, note that $\Lambda_n$ contains all data-dependent partitions). We also define

$$M(\Lambda_n) := \max\{\#\mathcal{P} : \mathcal{P} \in \Lambda_n\}$$

to be the maximum number of terminal nodes among all partitions in $\Lambda_n$. Note that $M(\Lambda_n) \le 2^K$. Given a set $\mathbf{z}^n = \{\mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n\} \subset \mathbb{R}^p$, define $\Gamma(\mathbf{z}^n, \Lambda_n)$ to be the number of distinct partitions of $\mathbf{z}^n$ induced by elements of $\Lambda_n$, that is, the number of different partitions $\{\mathbf{z}^n \cap A : A \in \mathcal{P}\}$, for $\mathcal{P} \in \Lambda_n$. The partitioning number $\Gamma_n(\Lambda_n)$ is defined by

$$\Gamma_n(\Lambda_n) := \max\{\Gamma(\mathbf{z}^n, \Lambda_n) : \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_n \in \mathbb{R}^p\},$$

i.e., the maximum number of different partitions of any $n$ point set that can be induced by members of $\Lambda_n$. Finally, let $\mathcal{F}_n$ denote the collection of all piecewise constant functions (bounded by $B$) on partitions $\mathcal{P} \in \Lambda_n$.

Now, by [13, Theorem 11.4] (choosing, in their notation, $\epsilon = 1/2$), we have

$$\begin{aligned}&\mathbb{P}\big(\exists\, f \in \mathcal{F}_n : \|\mu - f\|^2 \ge 2(\|Y-f\|_n^2 - \|Y-\mu\|_n^2) + \alpha + \beta\big) \le \\ &\qquad 14 \sup_{\mathbf{x}^n} \mathcal{N}\Big(\frac{\beta}{40B}, \mathcal{F}_n, L_1(\mathbb{P}_{\mathbf{x}^n})\Big) \exp\Big(-\frac{\alpha n}{2568 B^4}\Big),\end{aligned} \tag{39}$$

where $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n\} \subset \mathbb{R}^p$ and $\mathcal{N}(r, \mathcal{F}_n, L_1(\mathbb{P}_{\mathbf{x}^n}))$ is the covering number for $\mathcal{F}_n$ by balls of radius $r > 0$ in $L_1(\mathbb{P}_{\mathbf{x}^n})$ with respect to the empirical discrete measure $\mathbb{P}_{\mathbf{x}^n}$ on $\mathbf{x}^n$.

Next, we use [13, Lemma 13.1 and Theorem 9.4] as in the proof of [13, Theorem 13.1] to bound the empirical covering number by

$$\mathcal{N}\Big(\frac{\beta}{40B}, \mathcal{F}_n, L_1(\mathbb{P}_{\mathbf{x}^n})\Big) \le \Gamma_n(\Lambda_n)\Big(\frac{40}{32} \cdot \frac{333eB^2}{\beta}\Big)^{2M(\Lambda_n)} \le \Gamma_n(\Lambda_n)\Big(\frac{417eB^2}{\beta}\Big)^{2^{K+1}}. \tag{40}$$

We further bound (40) by noting that $\Gamma_n(\Lambda_n) \le ((n-1)p)^{2^K-1} \le (np)^{2^K}$. To see this, let $\gamma_K := \Gamma_n(\Lambda_n)$ for depth $K$ trees. Then $\gamma_K$ satisfies the recursion $\gamma_K \le ((n-1)p)\gamma_{K-1}^2$ with $\gamma_1 \le (n-1)p$, since there are at most $(n-1)p$ possible split points among any $n$ point set in $\mathbb{R}^p$ at the root node.[7] We therefore can bound the covering number by

$$\mathcal{N}\Big(\frac{\beta}{40B}, \mathcal{F}_n, L_1(\mathbb{P}_{\mathbf{x}^n})\Big) \le (np)^{2^K}\Big(\frac{417eB^2}{\beta}\Big)^{2^{K+1}}. \tag{41}$$

Returning to (39) and applying (41) to bound the covering number, since $\widehat{\mu}(T_K) \in \mathcal{F}_n$, we thus have

$$\mathbb{P}(E_1 \ge 0) \le 14(np)^{2^K}\Big(\frac{417eB^2}{\beta}\Big)^{2^{K+1}} \exp\Big(-\frac{\alpha n}{2568B^4}\Big).$$

We choose $\alpha = \frac{2568B^4(2^K\log(np)+2^{K+1}\log(417eB^2/\beta)+\log(14n))}{n}$ and $\beta = \frac{417eB^2}{n}$ so that $P(E_1 \ge 0) \le 1/n$. Furthermore, since $E_1 \le \|\mu - \widehat{\mu}(T_K)\|^2 + 2\|Y-\mu\|_n^2 \le 12B^2$, we have $\mathbb{E}(E_1) \le 12B^2\mathbb{P}(E_1 \ge 0) \le 12B^2/n$. Adding this bound on $\mathbb{E}(E_1)$ to the bound on $\mathbb{E}(E_2)$ from (38) and plugging in the choices of $\alpha$ and $\beta$, we have

$$\begin{aligned}
\mathbb{E}(\|\mu - \widehat{\mu}(T_K)\|^2) &= \mathbb{E}(E_1) + \mathbb{E}(E_2) \\
&\le \frac{12B^2}{n} + 2\|\mu - g\|^2 + \frac{2\|g\|_{\mathrm{TV}}^2}{K+3} \\
&\quad + \frac{2568B^4(2^K\log(np) + 2^{K+1}\log(n) + \log(14n))}{n} + \frac{417eB^2}{n} \\
&\le 2\|\mu - g\|^2 + \frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + C\frac{2^{K+1}\log(np)}{n},
\end{aligned}$$

where $C$ is a positive constant that depends only on $B$. □

*Proof of Corollary 4.3.* The proof follows immediately from Theorem 4.2. □

*Proof of Theorem 5.1.* The proof follows similar lines as Lemma 4.1 and Theorem 4.2, but with some interesting twists.

We first remind the reader of how the base regression trees in the forest are constructed. We begin by drawing a bootstrap sample $\mathcal{D}_n'$ from the original dataset $\mathcal{D}_n$. From this bootstrapped training sample, we grow a depth $K$ regression tree $T_K$ with CART methodology in the usual way, except that, at each internal node t, we split along a variable $X_{\hat{\jmath}}$ with $\hat{\jmath} \in \arg\max_{j\in\mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t})$, where $\mathcal{S} \subset \{1, 2, \ldots, p\}$ is a random subset formed by selecting $q$ of the $p$ variables uniformly at random, without replacement. Let $\Xi_K$ denote the random variable whose law generates the subsets $\mathcal{S}$ of candidate splitting variables at all internal nodes in $T_K$, conditional on the bootstrapped training data $\mathcal{D}_n'$.

[7] The author was inspired by [23] for this estimate of $\Gamma_n(\Lambda_n)$.

In order to prove Theorem 5.1, it is necessary to first establish a training error bound akin to Lemma 4.1. Following the notation in the proof of Lemma 4.1, for a terminal node t of $T_K$ and additive function $g \in \mathcal{G}$, we define $R_K(\mathrm{t})$ to be $\|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y - g\|_{\mathrm{t}}^2$, but based on the bootstrap sample $\mathcal{D}_n'$. We also analogously let $R_K$ denote the excess training error $\|Y - \widehat{\mu}(T_K)\|_n^2 - \|Y - g\|_n^2$, but again based on the bootstrap sample $\mathcal{D}_n'$.

Because of the additional randomness injected into the trees, we proceed by bounding the training error averaged with respect to $\Xi_K$; that is, we aim to bound $\mathbb{E}_{\Xi_K}(\|Y - \widehat{\mu}(T_K)\|_n^2)$. Note that due to independence of the subsets $\mathcal{S}$ across nodes, any terminal node t of $T_{K-1}$ is conditionally independent of $\Xi_K$ given $\Xi_{K-1}$. Thus, we study the $\Xi_K$ randomness of the tree at depth $K$ conditional on the randomness in the previous $K-1$ levels of the tree. As with the analysis for Lemma 4.1, we begin with the identity

$$\mathbb{E}_{\Xi_K|\Xi_{K-1}}(\|Y - \widehat{\mu}(T_K)\|_n^2) = \|Y - \widehat{\mu}(T_{K-1})\|_n^2 - \sum_{\mathrm{t}\in T_{K-1}} \mathbb{E}_{\Xi_K|\Xi_{K-1}}(|\langle Y, \psi_{\mathrm{t}}\rangle_n|^2), \quad (42)$$

which results from taking the expected value of (10) with respect to the conditional distribution of $\Xi_K$ given $\Xi_{K-1}$ (again, recognizing that the terminal nodes of $T_{K-1}$ are conditionally independent of $\Xi_K$ given $\Xi_{K-1}$).

Following a similar argument as the proof of Lemma 4.1, we have $\mathbb{E}_{\Xi_K|\Xi_{K-1}}(|\langle Y, \psi_{\mathrm{t}}\rangle_n|^2) = w(\mathrm{t})\mathbb{E}_{\Xi_K|\Xi_{K-1}}\big(\max_{j\in\mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t})\big)$, except now in lieu of Lemma 7.1, one shows that, for each terminal node t of $T_{K-1}$,

$$\mathbb{E}_{\Xi_K|\Xi_{K-1}}\big(\max_{j\in\mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t})\big) \geq \frac{q}{p}\frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathrm{TV}}^2}, \quad (43)$$

provided $R_{K-1}(\mathrm{t}) \geq 0$. Towards the goal of establishing (43), we first gain some intuition. Let $(\hat{\jmath}_1, \hat{\jmath}_2, \ldots, \hat{\jmath}_p)^T$ be a ranking of the variables such that $\widehat{\Delta}(\hat{s}, \hat{\jmath}_1, \mathrm{t}) < \widehat{\Delta}(\hat{s}, \hat{\jmath}_2, \mathrm{t}) < \cdots < \widehat{\Delta}(\hat{s}, \hat{\jmath}_p, \mathrm{t})$. For simplicity, at the moment, we have assumed there are no tie-breakers in the ranking. Then the expectation in (43) can be evaluated exactly as

$$\mathbb{E}_{\Xi_K|\Xi_{K-1}}\big(\max_{j\in\mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t})\big) = \sum_{k=q}^{p} \frac{\binom{k-1}{q-1}}{\binom{p}{q}} \widehat{\Delta}(\hat{s}, \hat{\jmath}_k, \mathrm{t}), \quad (44)$$

where we used $\mathbb{P}_{\Xi_K|\Xi_{K-1}}(\max_{j\in\mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t}) = \hat{\jmath}_k) = \binom{k-1}{q-1}/\binom{p}{q}$. Each summand in (44) is positive and so the final term $(\binom{p-1}{q-1}/\binom{p}{q})\widehat{\Delta}(\hat{s}, \hat{\jmath}_k, \mathrm{t}) = (q/p)\widehat{\Delta}(\hat{s}, \hat{\jmath}_p, \mathrm{t})$ provides a simple lower bound. On the other hand, even if there are tie-breakers in the ranking of the variables, we can still lower bound the expectation by

$$\mathbb{E}_{\Xi_K|\Xi_{K-1}}\big(\max_{j\in\mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t})\big) \geq \mathbb{E}_{\Xi_K|\Xi_{K-1}}\big(\mathbf{1}(\hat{\jmath}_p \in \mathcal{S})\widehat{\Delta}(\hat{s}, \hat{\jmath}_p, \mathrm{t})\big) \geq \frac{q}{p}\widehat{\Delta}(\hat{s}, \hat{\jmath}_p, \mathrm{t}), \quad (45)$$

where $\hat{\jmath}_p \in \arg\max_{j=1,2,\ldots,p} \widehat{\Delta}(\hat{s}, j, \mathrm{t})$ and the last inequality follows from the fact that the probability that a specific coordinate index belongs to $\mathcal{S}$ is $\binom{p-1}{q-1}/\binom{p}{q} = q/p$. We

then use Lemma 7.1 directly to conclude that $\widehat{\Delta}(\hat{s}, \hat{j}_p, \mathrm{t}) \geq \|g\|_{\mathrm{TV}}^{-2} R_{K-1}^2(\mathrm{t})$, which, when combined with (45), yields (43) as purported.

Proceeding in the same way as before with the proof of Lemma 4.1 that produced (26), and applying (43) to (42), one can easily establish the inequality

$$\mathbb{E}_{\Xi_K|\Xi_{K-1}}(R_K) \leq R_{K-1} - \frac{q/p}{\|g\|_{\mathrm{TV}}^2}(R_{K-1}^+)^2, \tag{46}$$

where we remind the reader that $R_{K-1}^+ = \sum_{\mathrm{t}\in T_{K-1}: R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t}) R_{K-1}(\mathrm{t}) \geq R_{K-1}$. Taking expected values of (46) with respect to $\Xi_{K-1}$ and using, in turn, the law of iterated expectations, i.e., $\mathbb{E}_{\Xi_{K-1}}(\mathbb{E}_{\Xi_K|\Xi_{K-1}}(R_K)) = \mathbb{E}_{\Xi_K}(R_K)$, and Jensen's inequality for the square function, i.e., $\mathbb{E}_{\Xi_{K-1}}((R_{K-1}^+)^2) \geq (\mathbb{E}_{\Xi_{K-1}}(R_{K-1}^+))^2$, we have

$$\begin{aligned}\mathbb{E}_{\Xi_K}(R_K) &\leq \mathbb{E}_{\Xi_{K-1}}(R_{K-1}) - \frac{q/p}{\|g\|_{\mathrm{TV}}^2}\mathbb{E}_{\Xi_{K-1}}((R_{K-1}^+)^2) \\ &\leq \mathbb{E}_{\Xi_{K-1}}(R_{K-1}) - \frac{q/p}{\|g\|_{\mathrm{TV}}^2}(\mathbb{E}_{\Xi_{K-1}}(R_{K-1}^+))^2.\end{aligned}$$

Since $\mathbb{E}_{\Xi_{K-1}}(R_{K-1}^+) \geq \mathbb{E}_{\Xi_{K-1}}(R_{K-1})$ (which follows from $R_{K-1}^+ \geq R_{K-1}$), we finally obtain

$$\mathbb{E}_{\Xi_K}(R_K) \leq \mathbb{E}_{\Xi_{K-1}}(R_{K-1})\Big(1 - \frac{q}{p}\frac{\mathbb{E}_{\Xi_{K-1}}(R_{K-1})}{\|g\|_{\mathrm{TV}}^2}\Big),$$

provided $\mathbb{E}_{\Xi_{K-1}}(R_{K-1}) \geq 0$. Iterating this recursion as with (27) in the proof of Lemma 4.1 yields $\mathbb{E}_{\Xi_K}(R_K) \leq (p/q)\|g\|_{\mathrm{TV}}^2/(K+3)$, or equivalently,

$$\mathbb{E}_{\Xi_K}(\|Y - \widehat{\mu}(T_K)\|_n^2) \leq \|Y - g\|_n^2 + \frac{p}{q}\frac{\|g\|_{\mathrm{TV}}^2}{K+3},$$

for $K \geq 1$. We next turn our attention to modifying the proof of Theorem 4.2 to accommodate the current setting. The most notable difference is in bounding the probability that $E_1 \geq 0$, where $E_1$ is defined in (36). However, this can easily be done in the same manner as before by noting that

$$\begin{aligned}&\mathbb{P}\big(\mathbb{E}_{\Xi_K}(\|\mu - \widehat{\mu}(T_K)\|^2) \geq 2(\mathbb{E}_{\Xi_K}(\|Y - \widehat{\mu}(T_K)\|_n^2) - \|Y - \mu\|_n^2) + \alpha + \beta\big) \leq \\ &\qquad \mathbb{P}\big(\exists\, f \in \mathcal{F}_n : \|\mu - f\|^2 \geq 2(\|Y - f\|_n^2 - \|Y - \mu\|_n^2) + \alpha + \beta\big),\end{aligned}$$

since if $\mathbb{E}_{\Xi_K}(\|\mu - \widehat{\mu}(T_K)\|^2 - 2\|Y - \widehat{\mu}(T_K)\|_n^2 + 2\|Y - \mu\|_n^2 - \alpha - \beta) \geq 0$, then there exists a realization from $\Xi_K$ (i.e., a piecewise constant function $\widehat{\mu}(T_K')$ in $\mathcal{F}_n$) for which the inequality $\|\mu - \widehat{\mu}(T_K')\|^2 - 2\|Y - \widehat{\mu}(T_K')\|_n^2 + 2\|Y - \mu\|_n^2 - \alpha - \beta \geq 0$ also holds. Following the same lines as the rest of the proof of Theorem 4.2, we have that for all $K \geq 1$ and all $g \in \mathcal{G}$, conditional on the indices $\mathcal{I}$ (corresponding to distinct observations) of the bootstrap sample $\mathcal{D}_n'$,

$$\mathbb{E}_{\Xi_K, \mathcal{D}_n'|\mathcal{I}}(\|\mu - \widehat{\mu}(T_K)\|^2) \leq 2\|\mu - g\|^2 + \frac{p}{q}\frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + C'\frac{2^{K+1}\log(\#\mathcal{I}p)}{\#\mathcal{I}}, \tag{47}$$

where $C'$ is a positive constant that depends only on $B$. Taking expectations of (47) with respect to $\mathcal{I}$ and using $1 \le \#\mathcal{I} \le n$, we have

$$\mathbb{E}_{\Theta,\mathcal{D}_n}(\|\mu - \widehat{\mu}(\Theta)\|^2) \le 2\|\mu - g\|^2 + \frac{p}{q}\frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + C'\mathbb{E}_{\mathcal{I}}\Big(\frac{2^{K+1}\log(np)}{\#\mathcal{I}}\Big). \tag{48}$$

Next, we use the fact that there exist universal positive constants $c_1$ and $c_2$ such that $\mathbb{P}(\#\mathcal{I} < c_1 n) \le c_2/n$. This can be shown from known expressions for the mean and variance of $\#\mathcal{I}$ [1, Lemma A.4 (vi) and (vii)], which are both approximately linear in $n$, and an application of Chebyshev's inequality. Thus,

$$\mathbb{E}_{\mathcal{I}}\Big(\frac{1}{\#\mathcal{I}}\Big) = \mathbb{E}_{\mathcal{I}}\Big(\frac{1}{\#\mathcal{I}}\mathbf{1}(\#\mathcal{I} \ge c_1 n)\Big) + \mathbb{E}_{\mathcal{I}}\Big(\frac{1}{\#\mathcal{I}}\mathbf{1}(\#\mathcal{I} < c_1 n)\Big) \le \frac{1}{c_1 n} + \frac{c_2}{n} = \frac{C''}{n}, \tag{49}$$

where $C'' \coloneqq 1/c_1 + c_2$. Combining this bound (49) with (48) shows that

$$\mathbb{E}_{\Theta,\mathcal{D}_n}(\|\mu - \widehat{\mu}(\Theta)\|^2) \le 2\|\mu - g\|^2 + \frac{p}{q}\frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + C\frac{2^{K+1}\log(np)}{n},$$

where $C \coloneqq C'C''$. The output of a random forest (13) is simply an empirical average (e.g., an equally weighted convex combination) of the individual tree outputs, each of which is generated according to the law of $\Theta$. Therefore, Jensen's inequality applied to the (convex) training error loss yields $\mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}_n}(\|\mu - \widehat{\mu}(\boldsymbol{\Theta})\|^2) \le \mathbb{E}_{\Theta,\mathcal{D}_n}(\|\mu - \widehat{\mu}(\Theta)\|^2)$. □

*Proof of Corollary 5.2.* The proof follows immediately from Theorem 5.1. □

**Remark 7.** *Our analysis does not exploit the de-correlation effect that occurs from the additional randomness in the ensemble. If we had first conditioned on $\Xi_{K-1}$ and then considered the training error of the partially randomized ensemble $\mathbb{E}_{\Xi_K|\Xi_{K-1}}(\widehat{\mu}(T_K))$ directly, we would have been led to*

$$\|Y - \mathbb{E}_{\Xi_K|\Xi_{K-1}}(\widehat{\mu}(T_K))\|_n^2 = \|Y - \widehat{\mu}(T_{K-1})\|_n^2 - \sum_{\mathrm{t}\in T_{K-1}} \mathbb{E}_{\Xi_K|\Xi_{K-1}}(|\langle Y, \psi_{\mathrm{t}}\rangle_n|^2)$$
$$- \sum_{\mathrm{t}\in T_{K-1}} \mathbb{E}_{\Xi_K|\Xi_{K-1}}(\|\langle Y, \psi_{\mathrm{t}}\rangle_n \psi_{\mathrm{t}} - \mathbb{E}_{\Xi_K|\Xi_{K-1}}(\langle Y, \psi_{\mathrm{t}}\rangle_n \psi_{\mathrm{t}})\|_n^2).$$

*Note the similarity with the identity (42), except for the presence of the additional term*

$$\sum_{\mathrm{t}\in T_{K-1}} \mathbb{E}_{\Xi_K|\Xi_{K-1}}(\|\langle Y, \psi_{\mathrm{t}}\rangle_n \psi_{\mathrm{t}} - \mathbb{E}_{\Xi_K|\Xi_{K-1}}(\langle Y, \psi_{\mathrm{t}}\rangle_n \psi_{\mathrm{t}})\|_n^2) > 0,$$

*which is attributable to the variance reduction effect of the ensemble.*

*Proof of Corollary 5.3.* We approximate each nonconstant $g_j(\cdot)$ by a sequence of step functions $h_{jn}(x_j) \coloneqq g_j(\llbracket x_j m_n \rrbracket/m_n)\mathbf{1}(|x_j| \le m_n)$, where $m_n = o(((q_n/p_n)K_n)^{1/4})$ and $m_n \to \infty$ for large $n$. Note that $g_j(\cdot)$ is the almost sure pointwise limit of the sequence $\{h_{jn}(\cdot)\}$ as $n \to \infty$. We take $h_{jn}(\cdot)$ to equal $g_j(\cdot)$ if $g_j(\cdot)$ is constant. Next, define $\widetilde{\mu}_n(\mathbf{x}) \coloneqq \sum_{j=1}^{p_n} h_{jn}(x_j)$ so that by Lebesgue's dominated convergence

theorem and the assumption that $\sup_n \|\mu_n\|_{\ell_0} < \infty$, we have $\lim_{n\to\infty} \|\mu_n - \widetilde{\mu}_n\| = 0$. Furthermore, since each component $h_{jn}(\cdot)$ (approximating a nonconstant $g_j(\cdot)$) is a step function with $\mathcal{O}(m_n^2)$ constant pieces, we have $\|\widetilde{\mu}_n\|_{\mathrm{TV}} = \mathcal{O}(m_n^2)$, where we again used the assumption that $\sup_n \|\mu_n\|_{\ell_0} < \infty$. Thus, by Theorem 5.1 and the hypotheses of Corollary 5.3, it follows that

$$\mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}_n}(\|\mu_n - \widehat{\mu}(\boldsymbol{\Theta})\|^2) \le$$
$$2\|\mu_n - \widetilde{\mu}_n\|^2 + \frac{2p_n}{q_n}\frac{\|\widetilde{\mu}_n\|_{\mathrm{TV}}^2}{K_n+3} + 2C\frac{2^{K_n}\log(np_n)}{n} \to 0, \quad \text{as} \quad n\to\infty. \qquad \square$$

## References


[1] Abadie, A. and Imbens, G. W. (2008). On the failure of the bootstrap for matching estimators. *Econometrica* **76** 1537–1557.

[2] Barron, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning* **14** 115-133.

[3] Barron, A. R., Cohen, A., Dahmen, W. and DeVore, R. A. (2008). Approximation and learning by greedy algorithms. *Annals of Statistics* **36** 64–94. MR2387964

[4] Breiman, L. (1996). Bagging predictors. *Machine learning* **24** 123–140.

[5] Breiman, L. (2001). Random forests. *Machine Learning* **45** 5–32.

[6] Breiman, L., Friedman, J., Olshen, R. and Stone, C. J. (1984). *Classification and regression trees*. Chapman and Hall/CRC.

[7] Bühlmann, P. (2006). Boosting for high-dimensional linear models. *The Annals of Statistics* **34** 559–583.

[8] Chi, C.-M., Vossler, P., Fan, Y. and Lv, J. (2020). Asymptotic Properties of High-Dimensional Random Forests. *arXiv preprint arXiv:2004.13953*.

[9] Denil, M., Matheson, D. and De Freitas, N. (2014). Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning (ICML)*.

[10] Donoho, D. L. (1997). CART and best-ortho-basis: a connection. *The Annals of Statistics* **25** 1870 – 1911.

[11] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 1189–1232.

[12] Gey, S. and Nedelec, E. (2005). Model selection for CART regression trees. *IEEE Transactions on Information Theory* **51** 658–670.

[13] Györfi, L., Krzyżak, A., Kohler, M. and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer.

[14] Hastie, T., Tibshirani, R. and Friedman, J. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition. Springer Series in Statistics*. Springer New York.

[15] Klusowski, J. M. (2020). Sparse Learning with CART. In *Advances in Neural Information Processing Systems*.

[16] Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1996). Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory* **42** 2118–2132. MR1447518

[17] Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint: arXiv:1407.7502*.
[18] Meier, L., Geer, V., and Bühlmann, P. (2009). High-Dimensional Additive Modeling. *The Annals of Statistics* **37** 3779–3821.
[19] Mentch, L. and Zhou, S. (2020). Randomization as Regularization: A Degrees of Freedom Explanation for Random Forest Success. *Journal of Machine Learning Research* **21** 1–36.
[20] Nobel, A. et al. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics* **24** 1084–1105.
[21] Ravikumar, P., Liu, H., Lafferty, J. and Wasserman, L. (2009). SpAM: Sparse Additive Models. *Journal of the Royal Statistical Society. Series B* **71** 1009-1030.
[22] Sadhanala, V. and Tibshirani, R. J. (2019). Additive models with trend filtering. *The Annals of Statistics* **47** 3032 – 3068.
[23] Scornet, E., Biau, G. and Vert, J.-P. (2015). Consistency of random forests. *Annals of Statistics* **43** 1716–1741.
[24] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics* 595–620.
[25] Tan, Z. and Zhang, C.-H. (2019). Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics* **47** 2567 – 2600.
[26] Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* 1–15.
[27] Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.
[28] Yuan, M. and Zhou, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics* **44** 2564–2593.