

Universal Consistency of Decision Trees in High Dimensions

Jason M. Klusowski*

Department of Operations Research and Financial Engineering,
Princeton University

August 4, 2021

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 natural 1-norm and 0-norm sparsity constraints. The theory applies to a wide range of additive models, including those with component functions that are continuous, of bounded variation, or, more generally, Borel measurable. Consistency is universal in the sense that there are no a priori assumptions on the distribution of the predictor variables, thereby accommodating continuous, discrete, and/or dependent data. Finally, we show that these qualitative properties of individual trees are inherited by Breiman's random forests. A key step in the analysis is the establishment of an oracle inequality, which allows for a precise characterization of the goodness-of-fit and complexity tradeoff for a misspecified model.

Keywords: Classification and regression trees; random forests; interpretable machine learning; greedy algorithms; oracle inequality

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

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 (Breiman, 1996), random forests (Breiman, 2001), and gradient tree boosting (Friedman, 2001). 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 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) (Breiman et al., 1984) 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 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 accommodate 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 (Breiman, 2001), which among other things, use CART methodology for the constituent trees.

1.1 Prior art

We now review some of the past theoretical work on CART. The first consistency result was provided in the original book that proposed the methodology (Breiman et al., 1984), albeit under very strong assumptions on the tree construction, such as a minimum node size condition and shrinking cell condition. Thirty years later, Scornet et al. (2015) 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 development because it did not require any of the strong assumptions on the tree made in [Breiman et al. \(1984\)](#). Subsequent work by [Chi et al. \(2020\)](#); [Klusowski \(2020\)](#); [Syrkkanis and Zampetakis \(2020\)](#); [Wager and Walther \(2015\)](#) provide finite sample consistency rates in a high dimensional setting with exact sparsity, though again, like [Breiman et al. \(1984\)](#), they operate under a set of conditions that may or may not hold in practice. Another notable paper by [Gey and Nedelec \(2005\)](#) provides oracle-type inequalities for pruned CART, but the theory does not extend to out-of-sample prediction.

Motivated by Stone’s conditions for consistency in nonparametric regression ([Stone, 1977](#)), most existing convergence results for decision trees follow an approach in which the approximation error is bounded by the mesh of the induced partition of the input space. Conditions are then imposed, either explicitly or implicitly, 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 ([Denil et al., 2014](#); [Breiman et al., 1984](#); [Wager and Walther, 2015](#); [Wager and Athey, 2018](#)). In contrast, the aforementioned paper ([Scornet et al., 2015](#)) 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. More recently, [Chi et al. \(2020\)](#); [Klusowski \(2020\)](#); [Syrkkanis and Zampetakis \(2020\)](#) developed techniques to directly analyze the approximation error (instead of using the granularity of the partition as a proxy) by exploiting the greedy optimization inherent in CART methodology. These papers provide consistency rates for models with exact sparsity in a high dimensional regime (i.e., when the ambient dimensionality grows with the sample size), however, they make a number of assumptions that lead to an unsatisfactory theory. For example, the results of [Klusowski \(2020\)](#) apply only to the noise free setting, [Chi et al. \(2020\)](#) require the ambient dimensionality to grow at most polynomially with the sample size, and [Syrkkanis and Zampetakis \(2020\)](#) work with binary 0-1 valued predictor variables. In addition, a local accuracy gain condition (akin to an edge condition in boosting literature) is required in these works ([Chi et al., 2020](#); [Klusowski, 2020](#); [Syrkkanis and Zampetakis, 2020](#)) to ensure the approximation error decreases by a constant factor after splitting at each level. In [Chi et al. \(2020\)](#), this local accuracy gain condition is verified to hold for some simple classes of additive models, such as those with isotonic or piecewise linear component functions and independent predictor variables. [Syrkkanis and Zampetakis \(2020\)](#) do not provide any concrete examples of models that satisfy what they call a “submodularity” property, and the reader is required to accept its validity. In summary, it is difficult to verify which models satisfy these technical conditions and, therefore, the general applicability of the theory remains unclear.

In order to fairly compare CART with other general nonparametric methods like nearest neighbors or kernel regression, our prime directive will be to develop theory that applies to a broad class of models, yet requires minimal operating assumptions. As we shall see from our novel method of analysis, the aforementioned technical conditions in extant literature are not strictly necessary in a general additive setting. Our statistical guarantees for CART remain valid even when the predictor variables are continuous, discrete, and/or dependent,

the model is only approximately sparse with a large number of predictor variables that make small contributions to the model output, 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 contained in the supplementary material.

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

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

almost surely; however, our proofs can be readily modified to accommodate unbounded sub-Gaussian errors (via a truncation argument) 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.

While CART methodology is a general nonparametric tool for prediction on unstructured data, a natural playground to illustrate its high dimensional properties 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 or exact) sparsity are easy to define and intuitive. They are, however, typically trained with regularized procedures

based on splines, trend filtering, or reproducing kernel Hilbert spaces (Meier et al., 2009; Ravikumar et al., 2009; Sadhanala and Tibshirani, 2019; Yuan and Zhou, 2016) that, unlike the agnostic nature of CART, take explicit advantage of the additive structure.

Before we proceed, let us emphasize that we do not advocate using CART specifically for additive modeling. Indeed, other methods (such as those aforementioned) that are tailored to this setting will almost always be a superior alternative, unless one values interpretability over accuracy. Instead, for our purposes, additive models will serve to illustrate how CART is able to do things that would not be possible with vanilla nearest neighbors and kernel methods, like handle sparsity when the predictor variables are dependent.

2.2 Sparse regression models

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

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

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

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

Later on, we will relax the bounded variation assumption and develop results for additive models with component functions that are merely bounded and Borel measurable.

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} .

As we have already mentioned, we would like to consider models that are *approximately* sparse. To this end, for $g \in \mathcal{G}$, we define the norm $\|g\|_{\text{TV}}$ as the infimum of

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

over all representations of $g(\cdot)$ as (2), i.e., $\|g\|_{\text{TV}}$ is the ℓ_1 aggregated total variation of the individual component functions (see Tan and Zhang (2019) and the references therein). Throughout, however, we assume $g(\cdot)$ has a canonical representation such that $\|g\|_{\text{TV}}$ achieves this infimum. One can think of $\|g\|_{\text{TV}}$ as a measure of the capacity of $g(\cdot)$ and, as we shall see, it will play a central role in the paper. The ℓ_1 total variation norm is desirable because it allows for some predictor variables to make very small yet meaningful contributions to the regression function.

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 ℓ_1 norm can be expressed as the multiple Riemann integral

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

where $\nabla(\cdot)$ is the gradient operator and $\|\cdot\|_{\ell_1}$ is the usual ℓ_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 $\|g\|_{\text{TV}}$ equals $\|\boldsymbol{\beta}\|_{\ell_1}$, the ℓ_1 norm of the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$. Furthermore, if $g(\cdot)$ is piecewise constant on V regions, then $\|g\|_{\text{TV}} \leq 2V\|g\|_{\infty}$, where $\|\cdot\|_{\infty}$ is the supremum norm.

To account for *exact* sparsity, we also define the ℓ_0 norm of an additive function $g(\mathbf{x}) = \sum_{j=1}^p g_j(x_j)$ as

$$\|g\|_{\ell_0} := \#\{j : g_j(\cdot) \text{ is nonconstant}\}.$$

In other words, the ℓ_0 norm counts the number of relevant variables that affect $g(\cdot)$. Note that if $g(\mathbf{x}) = \sum_{j=1}^p g_j(x_j)$ belongs to \mathcal{G} , then we have the relation

$$\|g\|_{\text{TV}} \leq \|g\|_{\ell_0} \cdot \max_j v(g_j).$$

Thus, the ℓ_1 total variation norm captures *both* exact and approximate sparsity, if present.

2.3 Notation

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

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

denote the squared norm and inner product, respectively, with respect to the empirical measure on the data. We view the response data vector $(Y_1, Y_2, \dots, Y_n)^T \in \mathbb{R}^n$ as a function defined on the design matrix $(\mathbf{X}_1, \mathbf{X}_2, \dots, \mathbf{X}_n)^T \in \mathbb{R}^{n \times p}$ 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 $\llbracket z \rrbracket$ to 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) (Breiman et al., 1984) 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 t_L and t_R according to whether $X \leq s$ or $X > s$, respectively. These two nodes will be denoted by $t_L := \{\mathbf{X} \in t : X \leq s\}$ and $t_R := \{\mathbf{X} \in t : X > s\}$. 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

$$\|W\|_t^2 := \frac{1}{N(t)} \sum_{\mathbf{X}_i \in t} W_i^2 \quad \text{and} \quad \langle W, W' \rangle_t := \frac{1}{N(t)} \sum_{\mathbf{X}_i \in t} W_i W'_i$$

for the squared norm and inner product, respectively, with respect to the conditional empirical measure given $\mathbf{X} \in t$. Here $N(t) := \#\{\mathbf{X}_i \in t\}$ 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

$$\|Y - \bar{Y}_t\|_t^2 := \frac{1}{N(t)} \sum_{\mathbf{X}_i \in t} (Y_i - \bar{Y}_t)^2, \quad (4)$$

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

$$\|Y - \bar{Y}_{t_L}\|_{t_L}^2 = \frac{1}{N(t_L)} \sum_{\mathbf{X}_i \in t_L} (Y_i - \bar{Y}_{t_L})^2, \quad \|Y - \bar{Y}_{t_R}\|_{t_R}^2 = \frac{1}{N(t_R)} \sum_{\mathbf{X}_i \in t_R} (Y_i - \bar{Y}_{t_R})^2,$$

where \bar{Y}_{t_L} is the sample mean for t_L and $N(t_L)$ is the sample size of t_L (similar definitions apply to \bar{Y}_{t_R} and $N(t_R)$). For a candidate variable X_j and corresponding split point $s = s_j$ of X_j , the impurity gain is defined as (Breiman et al., 1984, Definition 8.13)

$$\hat{\Delta}(s, j, t) := \|Y - \bar{Y}_t\|_t^2 - P(t_L)\|Y - \bar{Y}_{t_L}\|_{t_L}^2 - P(t_R)\|Y - \bar{Y}_{t_R}\|_{t_R}^2, \quad (5)$$

where $P(t_L) := N(t_L)/N(t)$ and $P(t_R) := N(t_R)/N(t)$ are the proportions of data points within t that are contained in t_L and t_R , respectively. We also define $w(t) := N(t)/n$ 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 $X_{\hat{j}}$, where $\hat{j} = \hat{j}(t)$, and corresponding split point $\hat{s} = \hat{s}(\hat{j}, t)$ of $X_{\hat{j}}$ that minimize the sum of squares error

$$\sum_{\mathbf{X}_i \in t_L} (Y_i - \bar{Y}_{t_L})^2 + \sum_{\mathbf{X}_i \in t_R} (Y_i - \bar{Y}_{t_R})^2,$$

or, equivalently, maximize the impurity gain $\hat{\Delta}(s, j, t)$, 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 t_L and 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. There are many criteria that can be used to determine when to stop splitting, each one giving rise to a different tree structure. In this paper, we use the following stopping rule.

Definition 3.1. Let T_K denote a fully grown binary tree of depth $K \in \mathbb{N}$, 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, $\bar{Y} := n^{-1} \sum_{i=1}^n Y_i$.

The output $\hat{\mu}(T) = \hat{\mu}(T, \mathcal{D}_n)$ of the tree T at a terminal (leaf) node t is the least squares (constant) predictor for data within t , namely, $(\hat{\mu}(T))(\mathbf{x}) \equiv \bar{Y}_t$ for all $\mathbf{x} \in t$.

3.2 Orthogonal decision tree expansions

Our first lemma, Lemma 3.2, shows that maximizing $\hat{\Delta}(s, j, t)$ is equivalent to maximizing the inner product $\langle Y - \bar{Y}_t, \Psi_t \rangle_t = \frac{1}{N(t)} \sum_{\mathbf{x}_i \in t} (Y_i - \bar{Y}_t) \Psi_t(\mathbf{X}_i)$ between the residuals $Y - \bar{Y}_t$ and the family of standardized decision stumps

$$\Psi_t(\mathbf{x}) := \frac{\mathbf{1}(\mathbf{x} \in t_L)P(t_R) - \mathbf{1}(\mathbf{x} \in t_R)P(t_L)}{\sqrt{P(t_L)P(t_R)}}, \quad (6)$$

which split X_j ($j = 1, 2, \dots, p$) at s within the parent node t .

Furthermore, Lemma 3.2 also shows that the tree output $\hat{\mu}(T)$ is equal to the (empirical) orthogonal projection of Y onto the linear span of orthonormal decision stumps

$$\psi_t(\mathbf{x}) := \Psi_t(\mathbf{x})\sqrt{1/w(t)}, \quad (7)$$

for internal (nonterminal) parent nodes t in T . (By slightly expanding the notion of an internal node to include the null node (i.e., the empty set), we define $\Psi_t(\mathbf{x}) \equiv \psi_t(\mathbf{x}) \equiv 1$ if t is the null node.)

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 iteratively projects the data onto the space of all constant predictors within a greedily obtained node. The proofs show that this local greedy approach has a very similar structure to standard global greedy algorithms in Hilbert spaces. The reader familiar with greedy algorithms in Hilbert spaces for over-complete dictionaries will recognize some similarities in the analysis (see the ‘‘orthogonal greedy algorithm’’ (Barron et al., 2008) in which one iteratively projects the data onto the linear span of a finite collection of greedily obtained dictionary elements). A similar parallel with orthogonal basis algorithms was explored for ‘‘dyadic’’ CART in Donoho (1997). 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 3.2. *The impurity gain is equal to the squared within-node inner product between the residuals $Y - \bar{Y}_t$ and a standardized decision stump $\Psi_t(\cdot)$ that splits a variable X_j at s within a parent node t , i.e.,*

$$\hat{\Delta}(s, j, t) = |\langle Y - \bar{Y}_t, \Psi_t \rangle_t|^2. \quad (8)$$

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

$$\widehat{\mu}(T) = \sum_t \langle Y, \psi_t \rangle_n \psi_t, \quad (9)$$

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

The orthogonal decomposition (9) of $\widehat{\mu}(T_K)$ in Lemma 3.2 is useful because it allows us to write down a recursive expression for the training error of CART. That is, from $\widehat{\mu}(T_K) = \widehat{\mu}(T_{K-1}) + \sum_{t \in T_{K-1}} \langle Y, \psi_t \rangle_n \psi_t$, we have

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

where the sums run over all terminal nodes 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 of Chi et al. (2020); Syrgkanis and Zampetakis (2020), we also do not need to assume a local accuracy gain condition so that the approximation error decreases by a constant factor after splitting at each level.

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 ℓ_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 $\mathcal{O}(1/K)$ training error bound in Lemma 4.1 and the descriptive complexity (relative to sample size) comes from the fact that the ϵ -metric entropy for depth K decision trees constructed from n sample points and p variables is of order $2^K \log(np/\epsilon)$. We note that both of these bounds for the goodness-of-fit and descriptive complexity are nearly optimal. For example, consider a single dimension ($p = 1$) and suppose the regression function is a sinusoid with frequency K on the unit interval, normalized by $4K$ to have unit total variation, i.e., $\mu(x) = (4K)^{-1} \sin(2\pi Kx)$ for $x \in [0, 1]$. Then the CART algorithm tends to produce many unnecessary splits near the endpoints of the unit interval (that consequently inflate the tree complexity), instead of reducing the training error by splitting closer to the midpoint. In fact, even though the descriptive complexity (i.e., the number of terminal nodes) of a tree of depth K is large, namely, $\Omega(2^K)$, the goodness-of-fit (i.e., the training error) is only of the same order as the squared norm of the regression function, namely, $\Omega(1/K^2)$.

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

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

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

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 $K = \lceil (\xi/2) \log_2(n) \rceil$ for some $\xi \in (0, 1]$, Theorem 4.2 states that the prediction error is bounded by

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

This bound serves as the basis for our next corollary. Simply put, if the total variation ℓ_1 norm of the regression function, i.e., $\|\mu\|_{\text{TV}}$, 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 as 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, where*

each component function $g_{j_n}(\cdot)$ has bounded variation and $\|\mu_n\|_{TV} = o(\sqrt{K_n})$. If $K_n \rightarrow \infty$ and $2^{K_n}(\log(np_n))/n \rightarrow 0$ as $n \rightarrow \infty$, then the CART algorithm is consistent, that is,

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

Remark 2. Note that because $\|g\|_{TV} \leq \|g\|_{\ell_0} \cdot \max_j v(g_j)$ for $g \in \mathcal{G}$, the consistency statement in Corollary 4.3 also applies to models with sparsity patterns defined by the number of relevant variables. Thus, the condition $\|\mu_n\|_{TV} = o(\sqrt{K_n})$ can be replaced with $\|\mu_n\|_{\ell_0} = o(\sqrt{K_n})$.

The hypotheses of Corollary 4.3 are satisfied if, for example, $K_n = \lceil (\xi/2) \log_2(n) \rceil$ and $p_n = \lceil \exp(cn^{1-\xi}) \rceil$ 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

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

The dependence on the total variation ℓ_1 norm $\|\mu_n\|_{TV}$ 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) (Louppe, 2014, Section 3). In fact, by a well known inequality for plug-in classifiers (Györfi et al., 2002, 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 (Bühlmann, 2006) or neural networks (Lee et al., 1996; Barron, 1994). For example, (Bühlmann, 2006, Theorem 1) states that boosting with linear learners is also consistent for a sequence of ℓ_1 constrained linear models $\mu_n(\mathbf{x}) = \beta_n^T \mathbf{x}$ on $[0, 1]^p$ in the high dimensional regime, i.e., when $p_n = \lceil \exp(cn^{1-\xi}) \rceil$ and $\sup_n \|\mu_n\|_{TV} = \sup_n \|\beta_n\|_{\ell_1} < \infty$, where c is a positive constant and $\xi \in (0, 1]$.

Remark 3. 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 (Breiman et al., 1984)) is recommended. This would have one first grow a full tree T_{max} (to maximum depth) and then minimize

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

over all trees T that can be obtained from T_{max} by iteratively merging its internal nodes, where α 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 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 amenable 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. 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 . Rather, all of our theory for CART rests solely on being able to lower bound the information gain $\widehat{\Delta}(\hat{s}, \hat{j}, t)$ by the squared within-node excess training error $\|g\|_{\text{TV}}^{-2}(\|Y - \bar{Y}_t\|_t^2 - \|Y - g\|_t^2)^2$, provided $\|Y - \bar{Y}_t\|_t^2 \geq \|Y - g\|_t^2$, for *all* nodes t and *all* functions $g \in \mathcal{G}$ (see Lemma 7.1). If we had we made this inequality part of our hypotheses, with the factor $\|g\|_{\text{TV}}^{-2}$ replaced by some quantity that encodes the complexity of $g(\cdot)$ (with $g(\cdot)$ not necessarily additive), then every statement made in this paper would remain valid for such regression models. However, this would place the present work in the same category as Chi et al. (2020); Syrgkanis and Zampetakis (2020) in which the theory is predicated on the sort of unverifiable assumptions we would like to avoid making.

One 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 $\|g\|_{\text{TV}} := \sum_j v(g_j) + \sum_{j < j'} v(g_{jj'})$. The above reasoning can be generalized to arbitrary order interactions in the model, but at the cost of greater computational complexity in constructing the tree. Indeed, 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, to $\mathcal{O}(p^d n \log(n))$.

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 (Denil et al., 2014, Proposition 3 and Proposition 4), (Breiman, 2001, Section 11), or (Breiman, 1996, Section

4.1)). Indeed, our results on the consistency of CART for high dimensional additive models also carry over to Breiman’s random forests (Breiman, 2001) 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 $\hat{j} \in \arg \max_{j \in \mathcal{S}} \hat{\Delta}(\hat{s}, j, t)$.¹

We grow M of these depth K regression trees separately using, respectively, M independent realizations $\Theta = (\Theta_1, \Theta_2, \dots, \Theta_M)^T$ of a random variable Θ . Here Θ 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^{th} regression tree is denoted by $\hat{\mu}(\Theta_m) = \hat{\mu}(\Theta_m, \mathcal{D}_n)$. With this notation in place, the random forest output is then simply the empirical average of the M regression tree outputs, namely,

$$\hat{\mu}(\Theta) = \hat{\mu}(\Theta, \mathcal{D}_n) := M^{-1} \sum_{m=1}^M \hat{\mu}(\Theta_m). \quad (13)$$

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.

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

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

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

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

¹We deviate slightly from Breiman’s original random forests (Breiman, 2001) 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 (Scornet et al., 2015, Theorem 2).

default value of q equal to $\lfloor p/3 \rfloor$. 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 4. *As the number of trees M approaches infinity, we note that $\hat{\mu}(\Theta) \rightarrow \mathbb{E}_{\Theta|\mathcal{D}_n}(\hat{\mu}(\Theta))$ almost surely, by the strong law of large numbers. Thus, as $M \rightarrow \infty$,*

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

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, where each component function $g_{jn}(\cdot)$ has bounded variation and $\|\mu_n\|_{TV} = o(\sqrt{(q_n/p_n)K_n})$. If $(q_n/p_n)K_n \rightarrow \infty$ and $2^{K_n}(\log(np_n))/n \rightarrow 0$ as $n \rightarrow \infty$, then random forests are consistent, that is,*

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

As with the consistency statement for CART in Corollary 4.3, the hypotheses of Corollary 5.2 are satisfied if, for example, $K_n = \lfloor (\xi/2) \log_2(n) \rfloor$, $p_n = \lfloor \exp(cn^{1-\xi}) \rfloor$, and $q_n = \lfloor p_n/3 \rfloor$, 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.,

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

This paves the way for some further comments. For classification problems, the default value of q_n is $\lfloor \sqrt{p_n} \rfloor$, in which case consistency is possible when $p_n = o(\log^2(n))$ and $K_n = \lfloor (1/2) \log_2(n) \rfloor$. 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 $K_n = \lfloor (1/2) \log_2(n) \rfloor$. 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 Scornet et al. (2015): “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 (Scornet et al., 2015, Theorem 1), who show that random forests are consistent for additive models when p is fixed, the component functions are continuous, and (in our notation) $K_n \rightarrow \infty$ and $2^{K_n}(\log(n))/n \rightarrow 0$ as $n \rightarrow \infty$. (The condition $2^{K_n}(\log(n))/n \rightarrow 0$ is actually stated as $2^{K_n}(\log(n))^9/n \rightarrow 0$ in (Scornet et al., 2015, Theorem 1), where Gaussian errors are assumed, but the condition stated here is sufficient when the noise is bounded.) 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 [Scornet et al. \(2015\)](#) 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 $a_n = \lceil 0.632n \rceil$, and, practically speaking, produce similar results when $a_n = \lceil n/2 \rceil$ ([Friedman and Hall, 2007](#)).

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 [Scornet et al. \(2015\)](#) and show that random forests are consistent for additive models with growing dimensionality and Borel measurable component functions—a much weaker requirement than continuity or bounded variation. Before we state this result, we recall that the ℓ_0 norm of an additive function $g(\mathbf{x}) = \sum_{j=1}^p g_j(x_j)$ is the number of variables that are relevant to $g(\cdot)$, namely, $\|g\|_{\ell_0} = \#\{j : g_j(\cdot) \text{ is nonconstant}\}$.

Corollary 5.3 (Consistency for additive models with Borel 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 Borel measurable. Furthermore, assume that $\sup_n \|\mu_n\|_{\ell_0} < \infty$. If $(q_n/p_n)K_n \rightarrow \infty$ and $2^{K_n}(\log(np_n))/n \rightarrow 0$ as $n \rightarrow \infty$, then random forests are consistent, that is,*

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

Remark 5. *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 bounded and Borel measurable. Furthermore, assume that $\sup_n \|\mu_n\|_{\ell_0} < \infty$. If $K_n \rightarrow \infty$ and $2^{K_n}(\log(np_n))/n \rightarrow 0$ as $n \rightarrow \infty$, then the CART algorithm is consistent, that is,*

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

6 Conclusion

In this paper, we showed that CART and random forests perform well under sparsity and can handle continuous, discrete, and/or dependent predictor variables. In contrast with past work, our theory is not predicated on assumptions that one must accept to be true or verify for models on a case by case basis.

Regarding random forests, it is still largely a mystery (at least theoretically) why bagging and the random variable selection mechanism 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 [Mentch and Zhou \(2020\)](#) 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 form of implicit regularization 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.

References

- Andrew R Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14(1):115–133, 1994.
- Andrew R Barron, Albert Cohen, Wolfgang Dahmen, and Ronald A. DeVore. Approximation and learning by greedy algorithms. *Annals of Statistics*, 36(1):64–94, 2008. ISSN 0090-5364. doi: 10.1214/009053607000000631. URL <http://dx.doi.org/10.1214/009053607000000631>.
- Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.
- Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.
- Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 1984.
- Peter Bühlmann. Boosting for high-dimensional linear models. *The Annals of Statistics*, 34(2):559–583, 2006.
- Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *arXiv preprint arXiv:2004.13953*, 2020.
- Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning (ICML)*, 2014.
- David L. Donoho. CART and best-ortho-basis: a connection. *The Annals of Statistics*, 25(5):1870 – 1911, 1997. doi: 10.1214/aos/1069362377. URL <https://doi.org/10.1214/aos/1069362377>.
- Jerome H Friedman. Greedy function approximation: A gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.
- Jerome H Friedman and Peter Hall. On bagging and nonlinear estimation. *Journal of statistical planning and inference*, 137(3):669–683, 2007.
- Servane Gey and Elodie Nedelec. Model selection for CART regression trees. *IEEE Transactions on Information Theory*, 51(2):658–670, 2005.

- László Györfi, Adam Krzyżak, Michael Kohler, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.
- Jason M. Klusowski. Sparse learning with CART. In *Advances in Neural Information Processing Systems*, 2020.
- Wee Sun Lee, Peter L. Bartlett, and Robert C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Trans. Inform. Theory*, 42(6, part 2): 2118–2132, 1996. ISSN 0018-9448. doi: 10.1109/18.556601. URL <http://dx.doi.org/10.1109/18.556601>.
- Gilles Louppe. Understanding random forests: From theory to practice. *arXiv preprint: arXiv:1407.7502*, 2014.
- L. Meier, V. Geer, , and P. Bühlmann. High-dimensional additive modeling. *The Annals of Statistics*, 37:3779–3821, 2009.
- Lucas Mentch and Siyu Zhou. Randomization as regularization: A degrees of freedom explanation for random forest success. *Journal of Machine Learning Research*, 21(171): 1–36, 2020.
- Pradeep Ravikumar, Han Liu, John Lafferty, and Larry Wasserman. Spam: Sparse additive models. *Journal of the Royal Statistical Society. Series B*, 71(5):1009–1030, 2009.
- Veeranjaneyulu Sadhanala and Ryan J. Tibshirani. Additive models with trend filtering. *The Annals of Statistics*, 47(6):3032 – 3068, 2019. doi: 10.1214/19-AOS1833. URL <https://doi.org/10.1214/19-AOS1833>.
- Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741, 2015.
- Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, 5(4): 595–620, 1977.
- Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Conference on Learning Theory*, pages 3453–3454. PMLR, 2020.
- Zhiqiang Tan and Cun-Hui Zhang. Doubly penalized estimation in additive regression with high-dimensional data. *The Annals of Statistics*, 47(5):2567 – 2600, 2019. doi: 10.1214/18-AOS1757. URL <https://doi.org/10.1214/18-AOS1757>.
- Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, pages 1–15, 2018.
- Stefan Wager and Guenther Walther. Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*, 2015.
- Ming Yuan and Ding-Xuan Zhou. Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics*, 44(6):2564–2593, 2016.

SUPPLEMENTARY MATERIAL

In this supplement, we provide complete proofs of Lemma 3.2, Lemma 4.1, Theorem 4.2, Corollary 4.3, Theorem 5.1, Corollary 5.2, and Corollary 5.3.

7 Proofs

Proof of Lemma 3.2. To prove the first assertion (8), note that

$$\begin{aligned} |\langle Y - \bar{Y}_t, \Psi_t \rangle_t|^2 &= \left| \left\langle Y - \bar{Y}_t, \frac{\mathbf{1}(\mathbf{X} \in t_L)N(t_R) - \mathbf{1}(\mathbf{X} \in t_R)N(t_L)}{\sqrt{N(t_L)N(t_R)}} \right\rangle_t \right|^2 \\ &= \left| \frac{1}{N(t)} \frac{\bar{Y}_{t_L}N(t_L)N(t_R) - \bar{Y}_{t_R}N(t_R)N(t_L)}{\sqrt{N(t_L)N(t_R)}} \right|^2 \\ &= \frac{N(t_L)N(t_R)}{N^2(t)} (\bar{Y}_{t_L} - \bar{Y}_{t_R})^2 = \hat{\Delta}(s, j, t), \end{aligned}$$

where the final equality comes from (Breiman et al., 1984, Section 9.3).

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

$$\begin{aligned} &\langle Y - \bar{Y}_t, \psi_t \rangle_n \psi_t(\mathbf{x}) \\ &= \left\langle Y - \bar{Y}_t, \frac{\mathbf{1}(\mathbf{X} \in t_L)P(t_R) - \mathbf{1}(\mathbf{X} \in t_R)P(t_L)}{\sqrt{w(t)P(t_L)P(t_R)}} \right\rangle_n \frac{\mathbf{1}(\mathbf{x} \in t_L)P(t_R) - \mathbf{1}(\mathbf{x} \in t_R)P(t_L)}{\sqrt{w(t)P(t_L)P(t_R)}} \\ &= (\bar{Y}_{t_L} - \bar{Y}_t)\mathbf{1}(\mathbf{x} \in t_L) + (\bar{Y}_{t_R} - \bar{Y}_t)\mathbf{1}(\mathbf{x} \in t_R). \end{aligned} \tag{14}$$

For a terminal node t of T , let t_0, t_1, \dots, t_K be the unique downward path from the root node t_0 to the terminal node $t_K = t$. Next, we sum up (14) over all internal nodes in T . The tree output \bar{Y}_t corresponding to a terminal region $\mathbf{x} \in t$ minus the output \bar{Y} in the root node is a telescoping sum of the successive (internal) node outputs:

$$\sum_{k=0}^{K-1} (\bar{Y}_{t_{k+1}} - \bar{Y}_{t_k}) = \bar{Y}_{t_K} - \bar{Y}_{t_0} = \bar{Y}_t - \bar{Y} = \bar{Y}_t - \langle Y, \psi_{t_0} \rangle_n \psi_{t_0}. \tag{15}$$

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

$$(\hat{\mu}(T))(\mathbf{x}) = \sum_{t \in T} \bar{Y}_t \mathbf{1}(\mathbf{x} \in t) = \sum_t \langle Y, \psi_t \rangle_n \psi_t(\mathbf{x}), \tag{16}$$

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 $\{\psi_t\}_t$ in the expansion (16) are orthonormal. First, notice that each ψ_t has unit empirical norm since

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

where we used $N(t_L) + N(t_R) = N(t)$. Next, each ψ_t (where t is not the null node) is orthogonal to the constant function since

$$\begin{aligned} \langle 1, \psi_t \rangle_n &= \left\langle 1, \frac{\mathbf{1}(\mathbf{X} \in t_L)P(t_R) - \mathbf{1}(\mathbf{X} \in t_R)P(t_L)}{\sqrt{w(t)P(t_L)P(t_R)}} \right\rangle_n \\ &= \frac{1}{n} \frac{N(t_L)N(t_R) - N(t_R)N(t_L)}{\sqrt{w(t)N(t_L)N(t_R)}} = 0. \end{aligned}$$

Furthermore, for distinct internal nodes t and t' in T , we have

$$\langle \psi_t, \psi_{t'} \rangle_n = 0. \quad (17)$$

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 ψ_t and $\psi_{t'}$ will zero each other out when multiplied, i.e., $\psi_t(\mathbf{x})\psi_{t'}(\mathbf{x}) = 0$. On the other hand, assume t is an ancestor of t' and (without loss of generality) that t' is a descendant of t_L or is equal to t_L . Then we have

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

thus completing the proof. \square

Proof of Lemma 4.1. To begin, let $g \in \mathcal{G}$ and define $R_K := \|Y - \hat{\mu}(T_K)\|_n^2 - \|Y - g\|_n^2$ as the excess training error. If $R_{K-1} < 0$, then there is nothing to prove because $R_K \leq R_{K-1} < 0$ and hence $\|Y - \hat{\mu}(T_K)\|_n^2 < \|Y - g\|_n^2$. (It can be seen from (10) that the training error of CART decreases with the depth so that $R_1 \geq R_2 \geq \dots \geq R_K$.) Therefore, we assume throughout that $R_{K-1} \geq 0$. For a terminal node t of T_K , we define the node-specific excess training error as $R_K(t) := \|Y - \bar{Y}_t\|_t^2 - \|Y - g\|_t^2$ so that $R_K = \sum_{t \in T_K} w(t)R_K(t)$. We recall that $w(t) = N(t)/n$ so that $\sum_{t \in T_K} w(t) = 1$. Here we use the notation “ $t \in T_K$ ” in the sum to mean that t is a terminal node of T_K . By the orthogonal decomposition of the tree output $\hat{\mu}(T_K)$ from (9) of Lemma 3.2, we have

$$\|Y - \hat{\mu}(T_K)\|_n^2 = \|Y - \hat{\mu}(T_{K-1})\|_n^2 - \sum_{t \in T_{K-1}} |\langle Y, \psi_t \rangle_n|^2, \quad (18)$$

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

$$R_K = R_{K-1} - \sum_{t \in T_{K-1}} |\langle Y, \psi_t \rangle_n|^2. \quad (19)$$

Furthermore, according to the definitions of Ψ_t in (6) and ψ_t in (7), we have $\langle Y, \psi_t \rangle_n = \sqrt{w(t)} \langle Y - \bar{Y}_t, \Psi_t \rangle_t$, and thus (19) can be rewritten as

$$R_K = R_{K-1} - \sum_{t \in T_{K-1}} w(t) |\langle Y - \bar{Y}_t, \Psi_t \rangle_t|^2. \quad (20)$$

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

$$R_K \leq R_{K-1} - \sum_{t \in T_{K-1}: R_{K-1}(t) \geq 0} w(t) |\langle Y - \bar{Y}_t, \Psi_t \rangle_t|^2. \quad (21)$$

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

$$\widehat{\Delta}(\hat{s}, \hat{j}, t) \geq \frac{R_{K-1}^2(t)}{\|g\|_{\text{TV}}^2}. \quad (22)$$

The identity (8) from Lemma 3.2 and (22) together imply that

$$|\langle Y - \bar{Y}_t, \Psi_t \rangle_t|^2 \geq \frac{R_{K-1}^2(t)}{\|g\|_{\text{TV}}^2}. \quad (23)$$

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

$$R_K \leq R_{K-1} - \frac{1}{\|g\|_{\text{TV}}^2} \sum_{t \in T_{K-1}: R_{K-1}(t) \geq 0} w(t) R_{K-1}^2(t). \quad (24)$$

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

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

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

Applying (25) to (24), we obtain

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

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

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

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

$$R_K \leq R_{K-1} \left(1 - \frac{R_{K-1}}{\|g\|_{\text{TV}}^2} \right) \leq \frac{\|g\|_{\text{TV}}^2}{K + 2} \left(1 - \frac{1}{K + 3} \right) = \frac{\|g\|_{\text{TV}}^2}{K + 3}. \quad \square$$

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

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

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

Proof of Lemma 7.1. We use recent tools for the analysis of decision trees from [Klusowski \(2020\)](#), namely, Lemma A.4 in the supplement therein. For daughter nodes t_L and 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)

$$\frac{d\Pi(s, j)}{d(s, j)} := \frac{|g'_j(s)|\sqrt{P(t_L)P(t_R)}}{\sum_{j'=1}^p \int |g'_{j'}(s')|\sqrt{P(t'_L)P(t'_R)}ds'},$$

where $g'_j(\cdot)$ is shorthand for the divided difference of $g_j(\cdot)$ for the successive ordered data points along the j^{th} direction within t . That is, if $X'_1 \leq X'_2 \leq \dots \leq X'_{N(t)}$ denotes the ordered data along the j^{th} direction within t , then $g'_j(s) := (g_j(X'_{i+1}) - g_j(X'_i))/(X'_{i+1} - X'_i)$ for $X'_i \leq s < X'_{i+1}$ and $i = 1, 2, \dots, N(t) - 1$. Observe that this definition of $g'_j(\cdot)$ coincides with the derivative of the function that linearly interpolates the points $\{(X'_1, g_j(X'_1)), (X'_2, g_j(X'_2)), \dots, (X'_{N(t)}, g_j(X'_{N(t)}))\}$. For notational brevity, we omit the explicit dependence of $P(t_L)$ and $P(t_R)$ on s and j and $P(t'_L)$ and $P(t'_R)$ on s' and j' .

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

$$\widehat{\Delta}(\hat{s}, \hat{j}, t) \geq \int \widehat{\Delta}(s, j, t) d\Pi(s, j) = \int |\langle Y - \bar{Y}_t, \Psi_t \rangle_t|^2 d\Pi(s, j), \quad (28)$$

where the last identity follows from (8) in Lemma 3.2. Next, by Jensen's inequality for the square function, (28) is further lower bounded by

$$\int |\langle Y - \bar{Y}_t, \Psi_t \rangle_t|^2 d\Pi(s, j) \geq \left(\int |\langle Y - \bar{Y}_t, \Psi_t \rangle_t| d\Pi(s, j) \right)^2. \quad (29)$$

We next evaluate the expectation in (29) with respect to the measure Π , giving

$$\int |\langle Y - \bar{Y}_t, \Psi_t \rangle_t| d\Pi(s, j) = \frac{\sum_{j=1}^p \int |g'_j(s)| |\langle Y - \bar{Y}_t, \mathbf{1}(X_j > s) \rangle_t| ds}{\sum_{j'=1}^p \int |g'_{j'}(s')|\sqrt{P(t'_L)P(t'_R)}ds'}, \quad (30)$$

where we used the identity $\sqrt{P(t_L)P(t_R)}\langle Y - \bar{Y}_t, \Psi_t \rangle_t = -\langle Y - \bar{Y}_t, \mathbf{1}(X_j > s) \rangle_t$, which follows from $\mathbf{1}(\mathbf{X} \in t_L)P(t_R) - \mathbf{1}(\mathbf{X} \in t_R)P(t_L) = -(\mathbf{1}(X_j > s) - P(t_R))\mathbf{1}(\mathbf{X} \in t)$.

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

$$\sum_{j=1}^p \int |g'_j(s)| |\langle Y - \bar{Y}_t, \mathbf{1}(X_j > s) \rangle_t| ds \geq \left| \sum_{j=1}^p \int g'_j(s) \langle Y - \bar{Y}_t, \mathbf{1}(X_j > s) \rangle_t ds \right|. \quad (31)$$

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

$$\begin{aligned}
\sum_{j=1}^p \int g'_j(s) \langle Y - \bar{Y}_t, \mathbf{1}(X_j > s) \rangle_t ds &= \left\langle Y - \bar{Y}_t, \sum_{j=1}^p \int g'_j(s) \mathbf{1}(X_j > s) ds \right\rangle_t \\
&= \left\langle Y - \bar{Y}_t, \sum_{j=1}^p g_j \right\rangle_t \\
&= \langle Y - \bar{Y}_t, g \rangle_t.
\end{aligned} \tag{32}$$

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

$$\hat{\Delta}(\hat{s}, \hat{j}, t) \geq \frac{|\langle Y - \bar{Y}_t, g \rangle_t|^2}{\left(\sum_{j'=1}^p \int |g'_{j'}(s')| \sqrt{P(t'_L)P(t'_R)} ds' \right)^2}. \tag{33}$$

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' ,

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

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

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

For the numerator of (33), we use the Cauchy-Schwarz inequality to lower bound $\langle Y - \bar{Y}_t, g \rangle_t$ by

$$\langle Y - \bar{Y}_t, g \rangle_t = \langle Y - \bar{Y}_t, Y \rangle_t + \langle Y - \bar{Y}_t, g - Y \rangle_t \geq \|Y - \bar{Y}_t\|_t^2 - \|Y - \bar{Y}_t\|_t \|Y - g\|_t,$$

where we also used $\langle Y - \bar{Y}_t, Y \rangle_t = \|Y - \bar{Y}_t\|_t^2$. Now, by the AM-GM inequality, $\|Y - \bar{Y}_t\|_t \|Y - g\|_t \leq \frac{\|Y - \bar{Y}_t\|_t^2 + \|Y - g\|_t^2}{2}$, and hence $\langle Y - \bar{Y}_t, g \rangle_t \geq 2^{-1} (\|Y - \bar{Y}_t\|_t^2 - \|Y - g\|_t^2) = 2^{-1} R_{K-1}(t)$. Squaring both sides and using the assumption $R_{K-1}(t) \geq 0$, we have

$$|\langle Y - \bar{Y}_t, g \rangle_t|^2 \geq 4^{-1} (\|Y - \bar{Y}_t\|_t^2 - \|Y - g\|_t^2)^2 = 4^{-1} R_{K-1}^2(t). \tag{35}$$

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

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

which completes the proof. \square

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

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

and

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

and α and β are positive constants to be chosen later. Notice that by Lemma 4.1, we have

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

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

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

To bound E_1 , we first introduce a few useful concepts and definitions due to Nobel et al. (1996) 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 Λ_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 Λ_n . Note that $M(\Lambda_n) \leq 2^K$. 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 Λ_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 Λ_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 (Györfi et al., 2002, Theorem 11.4) (choosing, in their notation, $\epsilon = 1/2$), we have

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

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 (Györfi et al., 2002, Lemma 13.1 and Theorem 9.4) as in the proof of (Györfi et al., 2002, Theorem 13.1) with $\epsilon = (32/40)\beta$ to bound the empirical covering number by

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

We further bound (40) by noting that $\Gamma_n(\Lambda_n) \leq ((n-1)p)^{2^K-1} \leq (np)^{2^K}$. To see this, let $\gamma_K := \Gamma_n(\Lambda_n)$ to emphasize the dependence on the depth K . Now, for each depth K tree constructed from a set of n points in \mathbb{R}^p , there are at most $(n-1)p$ possible split points at the root node. The left and right (sub)trees resulting from this split have depth $K-1$, and hence they each independently contribute to the count in γ_{K-1} . Thus, γ_K satisfies the recursion $\gamma_K \leq ((n-1)p) \cdot \gamma_{K-1} \cdot \gamma_{K-1} = ((n-1)p)\gamma_{K-1}^2$ with $\gamma_1 \leq (n-1)p$. It can easily be shown that the solution is $\gamma_K \leq ((n-1)p)^{2^K-1}$ for $K \geq 1$. (The author was inspired by Scornet et al. (2015) for this estimate of $\Gamma_n(\Lambda_n)$.) We therefore can bound the covering number by

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

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

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

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 $\mathbb{P}(E_1 \geq 0) \leq 1/n$. Furthermore, since $E_1 \leq \|\mu - \hat{\mu}(T_K)\|^2 + 2\|Y - \mu\|_n^2 \leq 12B^2$, we have $\mathbb{E}(E_1) \leq 12B^2 \mathbb{P}(E_1 \geq 0) \leq 12B^2/n$. Adding this bound on $\mathbb{E}(E_1)$ to the bound on $\mathbb{E}(E_2)$ from (38) and plugging in the choices of α and β , we have

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

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_j with $\hat{j} \in \arg \max_{j \in \mathcal{S}} \widehat{\Delta}(\hat{s}, j, t)$, where $\mathcal{S} \subset \{1, 2, \dots, p\}$ is a random subset formed by selecting q of the p variables uniformly at random, without replacement. Let Ξ_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 .

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 $R_K(t)$ to be $\|Y - \bar{Y}_t\|_t^2 - \|Y - g\|_t^2$, but based on the bootstrap sample \mathcal{D}'_n . We also analogously let R_K denote the excess training error $\|Y - \hat{\mu}(T_K)\|_n^2 - \|Y - g\|_n^2$, 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 Ξ_K ; that is, we aim to bound $\mathbb{E}_{\Xi_K}(\|Y - \hat{\mu}(T_K)\|_n^2)$. Note that due to independence of the subsets \mathcal{S} across nodes, any terminal node t of T_{K-1} is conditionally independent of Ξ_K given Ξ_{K-1} . Thus, we study the Ξ_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

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

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

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

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

provided $R_{K-1}(t) \geq 0$. Towards the goal of establishing (43), we first gain some intuition. Let $(\hat{j}_1, \hat{j}_2, \dots, \hat{j}_p)^T$ be a ranking of the variables such that $\widehat{\Delta}(\hat{s}, \hat{j}_1, t) < \widehat{\Delta}(\hat{s}, \hat{j}_2, t) < \dots < \widehat{\Delta}(\hat{s}, \hat{j}_p, t)$. 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

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

where we used $\mathbb{P}_{\Xi_K|\Xi_{K-1}}(\max_{j \in \mathcal{S}} \widehat{\Delta}(\hat{s}, j, t) = \hat{j}_k) = \binom{k-1}{q-1} / \binom{p}{q}$. Each summand in (44) is positive and so the final term $(\binom{p-1}{q-1} / \binom{p}{q}) \widehat{\Delta}(\hat{s}, \hat{j}_p, t) = (q/p) \widehat{\Delta}(\hat{s}, \hat{j}_p, t)$ 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

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

where $\hat{j}_p \in \arg \max_{j=1,2,\dots,p} \widehat{\Delta}(\hat{s}, j, t)$ 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 $\widehat{\Delta}(\hat{s}, \hat{j}_p, t) \geq \|g\|_{\text{TV}}^{-2} R_{K-1}^2(t)$, 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

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

where we remind the reader that $R_{K-1}^+ = \sum_{t \in T_{K-1}: R_{K-1}(t) \geq 0} w(t) R_{K-1}(t) \geq R_{K-1}$. Taking expected values of (46) with respect to Ξ_{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

$$\begin{aligned} \mathbb{E}_{\Xi_K}(R_K) &\leq \mathbb{E}_{\Xi_{K-1}}(R_{K-1}) - \frac{q/p}{\|g\|_{\text{TV}}^2} \mathbb{E}_{\Xi_{K-1}}((R_{K-1}^+)^2) \\ &\leq \mathbb{E}_{\Xi_{K-1}}(R_{K-1}) - \frac{q/p}{\|g\|_{\text{TV}}^2} (\mathbb{E}_{\Xi_{K-1}}(R_{K-1}^+))^2. \end{aligned} \quad (47)$$

Since $\mathbb{E}_{\Xi_{K-1}}(R_{K-1}^+) \geq \mathbb{E}_{\Xi_{K-1}}(R_{K-1})$ (which follows from $R_{K-1}^+ \geq R_{K-1}$), from (47) we finally obtain

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

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 $\mathbb{E}_{\Xi_K}(R_K) \leq (p/q)\|g\|_{\text{TV}}^2/(K+3)$, or equivalently,

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

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

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

since if $\mathbb{E}_{\Xi_K}(\|\mu - \hat{\mu}(T_K)\|^2) - 2\|Y - \hat{\mu}(T_K)\|_n^2 + 2\|Y - \mu\|_n^2 - \alpha - \beta \geq 0$, then there exists a realization from Ξ_K (i.e., a piecewise constant function $\hat{\mu}(T'_K)$ in \mathcal{F}_n) for which the inequality $\|\mu - \hat{\mu}(T'_K)\|^2 - 2\|Y - \hat{\mu}(T'_K)\|_n^2 + 2\|Y - \mu\|_n^2 - \alpha - \beta \geq 0$ 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 ,

$$\mathbb{E}_{\Xi_K, \mathcal{D}'_n|\mathcal{I}}(\|\mu - \hat{\mu}(T_K)\|^2) \leq 2\|\mu - g\|^2 + \frac{p}{q} \frac{2\|g\|_{\text{TV}}^2}{K+3} + C' \frac{2^{K+1} \log(\#\mathcal{I}p)}{\#\mathcal{I}}, \quad (48)$$

where C' is a positive constant that depends only on B . Taking expectations of (48) with respect to \mathcal{I} and using $1 \leq \#\mathcal{I} \leq n$, we have

$$\mathbb{E}_{\Theta, \mathcal{D}_n}(\|\mu - \hat{\mu}(\Theta)\|^2) \leq 2\|\mu - g\|^2 + \frac{p}{q} \frac{2\|g\|_{\text{TV}}^2}{K+3} + C' \mathbb{E}_{\mathcal{I}}\left(\frac{2^{K+1} \log(np)}{\#\mathcal{I}}\right). \quad (49)$$

Next, we use the fact that there exist universal positive constants c_1 and c_2 such that $\mathbb{P}(\#\mathcal{I} < c_1 n) \leq c_2/n$. This can be shown from known expressions for the mean and variance of $\#\mathcal{I}$ (Abadie and Imbens, 2008, Lemma A.4 (vi) and (vii)), which are both approximately linear in n , and an application of Chebyshev's inequality. Thus,

$$\mathbb{E}_{\mathcal{I}}\left(\frac{1}{\#\mathcal{I}}\right) = \mathbb{E}_{\mathcal{I}}\left(\frac{1}{\#\mathcal{I}} \mathbf{1}(\#\mathcal{I} \geq c_1 n)\right) + \mathbb{E}_{\mathcal{I}}\left(\frac{1}{\#\mathcal{I}} \mathbf{1}(\#\mathcal{I} < c_1 n)\right) \leq \frac{1}{c_1 n} + \frac{c_2}{n} = \frac{C''}{n}, \quad (50)$$

where $C'' := 1/c_1 + c_2$. Combining this bound (50) with (49) shows that

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

where $C := C' C''$. 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 Θ . Therefore, Jensen's inequality applied to the (convex) training error loss yields $\mathbb{E}_{\Theta, \mathcal{D}_n}(\|\mu - \hat{\mu}(\Theta)\|^2) \leq \mathbb{E}_{\Theta, \mathcal{D}_n}(\|\mu - \hat{\mu}(\Theta)\|^2)$. \square

Proof of Corollary 5.2. The proof follows immediately from Theorem 5.1. \square

Remark 6. *Our analysis does not exploit the de-correlation effect that occurs from the additional randomness in the ensemble. If we had first conditioned on Ξ_{K-1} and then considered the training error of the partially randomized ensemble $\mathbb{E}_{\Xi_K | \Xi_{K-1}}(\hat{\mu}(T_K))$ directly, we would have been led to*

$$\begin{aligned} \|Y - \mathbb{E}_{\Xi_K | \Xi_{K-1}}(\hat{\mu}(T_K))\|_n^2 &= \|Y - \hat{\mu}(T_{K-1})\|_n^2 - \sum_{t \in T_{K-1}} \mathbb{E}_{\Xi_K | \Xi_{K-1}}(|\langle Y, \psi_t \rangle_n|^2) \\ &\quad - \sum_{t \in T_{K-1}} \mathbb{E}_{\Xi_K | \Xi_{K-1}}(\|\langle Y, \psi_t \rangle_n \psi_t - \mathbb{E}_{\Xi_K | \Xi_{K-1}}(\langle Y, \psi_t \rangle_n \psi_t)\|_n^2). \end{aligned}$$

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

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

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

Proof of Corollary 5.3. By (Stein and Shakarchi, 2005, Theorem 4.3), there exists a sequence of step functions $(s_{j_n}(\cdot))_n$ that converges pointwise almost surely to $g_j(\cdot)$. (The proof of (Stein and Shakarchi, 2005, Theorem 4.3) extends to any Borel probability measure on \mathbb{R}^p .) By expanding the sequence to include successive duplicate elements, we can assume without loss of generality that $\|s_{j_n}\|_{\text{TV}} = o(((q_n/p_n)K_n)^{1/2})$. Define $h_{j_n} := \text{sgn}(s_{j_n}) \min\{\|g_j\|_{\infty}, |s_{j_n}|\}$ if $g_j(\cdot)$ is nonconstant; otherwise, take $h_{j_n}(\cdot)$ to equal $g_j(\cdot)$ if $g_j(\cdot)$ is constant. Next, define

$\tilde{\mu}_n(\mathbf{x}) := \sum_{j=1}^{p_n} h_{jn}(x_j)$ so that by the Lebesgue dominated convergence theorem and the assumption that $\sup_n \|\mu_n\|_{\ell_0} < \infty$, we have $\lim_{n \rightarrow \infty} \|\mu_n - \tilde{\mu}_n\| = 0$. Furthermore, since each component $h_{jn}(\cdot)$ is a step function with total variation $o(((q_n/p_n)K_n)^{1/2})$, we likewise have $\|\tilde{\mu}_n\|_{\text{TV}} = o(((q_n/p_n)K_n)^{1/2})$, 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

$$\begin{aligned} \mathbb{E}_{\Theta, \mathcal{D}_n} (\|\mu_n - \hat{\mu}(\Theta)\|^2) &\leq \\ 2\|\mu_n - \tilde{\mu}_n\|^2 + \frac{2p_n}{q_n} \frac{\|\tilde{\mu}_n\|_{\text{TV}}^2}{K_n + 3} + 2C \frac{2^{K_n} \log(np_n)}{n} &\rightarrow 0, \quad \text{as } n \rightarrow \infty. \quad \square \end{aligned}$$

References

- Alberto Abadie and Guido W Imbens. On the failure of the bootstrap for matching estimators. *Econometrica*, 76(6):1537–1557, 2008.
- Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 1984.
- László Györfi, Adam Krzyżak, Michael Kohler, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer, 2002.
- Jason M. Klusowski. Sparse learning with CART. In *Advances in Neural Information Processing Systems*, 2020.
- Andrew Nobel et al. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084–1105, 1996.
- Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741, 2015.
- Elias M Stein and Rami Shakarchi. *Real analysis: measure theory, integration, and Hilbert spaces*. Princeton lectures in analysis. Princeton Univ. Press, Princeton, NJ, 2005. URL <https://cds.cern.ch/record/1385521>.