Universal Consistency of Decision Trees for High Dimensional Additive Models

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 for additive models, even when the dimensionality scales exponentially with the sample size, under certain ℓ_1 sparsity constraints. The consistency is universal in the sense that there are no a priori assumptions on the distribution of the input variables. Surprisingly, this adaptivity to (approximate or exact) 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. A key step in the analysis is the establishment of an oracle inequality, which precisely characterizes the goodness-of-fit and complexity tradeoff.

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

Keywords and phrases: Decision trees; classification and regression trees; 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 [3], random forests [4], and gradient tree boosting [9]. From an applied perspective, decision trees are intuitive and have an appealing interpretability that makes them easy to explain to statistical non-experts. They are also supplemented by a rich set of analytic and visual diagnostic tools for exploratory data analysis. These qualities make decision tree learning particularly appropriate for 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 specifically on Classification and Regression Trees (CART) [5] methodology—undoubtedly the most popular one for regression

*This research was supported in part by NSF DMS-1915932 and NSF HDR TRIPODS DATA-INSPIRE DCCF-1934924.

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.*

The consistency 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 either independent or near-independent (e.g., joint densities which are bounded above and below by universal positive constants).

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 tree-structured learning. In doing so, we show that decision trees 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 this result reveals a distinct property of decision trees that has 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 [5], albeit under very strong assumptions on the tree construction. Thirty years later, [20] 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 [5]. Subsequent work by [23], [7], and [13] 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 [10] 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 [21], 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 [8, 5, 23, 22]. In contrast, the aforementioned paper [20] 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, [7] and [13] developed techniques to bypass using the mesh as a proxy for the approximation error and instead directly analyze the prediction error using the construction of CART based on greedy optimization. However, they make very strong (essentially unverifiable) assumptions on the tree construction and data generating process, as well as operate in only a moderately high dimensional regime. Our theory reveals that these assumptions are not strictly necessary in many settings.

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 6.

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 $\mu(\mathbf{X}) \coloneqq \mathbb{E}(Y \mid \mathbf{X})$ 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 ε 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, 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 adapted to accommodate unbounded sub-Gaussian errors with only minor modifications 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 $\mathbf{X} = \mathbf{x}$. To understand the properties of prediction 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, namely, as large as $[\![\exp(cn^{1-\xi})]\!]$ for some positive constants c and $\xi \in (1/2, 1]$.

It turns out that the high dimensional properties of CART are most naturally and cleanly illustrated with additive models, i.e., models for which the regression function may be expressed as a sum of univariate functions of the input variables. Indeed, these models are often used in high dimensional contexts where sparsity is present, and are typically trained with regularized procedures based on splines or reproducing kernel Hilbert spaces [16, 19, 24].

We therefore 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., finite total variation, denoted by $\mathrm{TV}(g_j)$). However, importantly, we allow for model mis-specification and therefore *do not* require that $\mu(\cdot)$ belongs to $\mathcal{G}$.

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

$$\mathrm{TV}(g_1) + \mathrm{TV}(g_2) + \cdots + \mathrm{TV}(g_p)$$

over all representations of $g(\cdot)$ as (2). Throughout, however, we assume $g(\cdot)$ has a canonical representation such that $\|g\|_{\mathrm{TV}}$ achieves this infimum. One can think of this quantity as a measure of the "capacity" of $g(\cdot)$ and, as we shall see, it will play a central role in the paper.

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

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

where $\nabla(\cdot)$ is the gradient operator and $\|\cdot\|_{\ell_1}$ is the usual ℓ_1 norm of a vector in $\mathbb{R}^p$. We note that when $g(\cdot)$ is linear over the domain $[0,1]^p$, $\|g\|_{\mathrm{TV}}$ is the ℓ_1 norm of the parameter vector. Furthermore, if $g(\cdot)$ is piecewise constant on V regions, then $\|g\|_{\mathrm{TV}} \le 2V\|g\|_\infty$.

We now introduce some notation that will be used throughout the paper. For functions $f, g \in L_2(P_{\mathbf{X}})$, let

$$\|f\|^2 \coloneqq \int_{\mathbb{R}^p} |f(\mathbf{x})|^2 dP_{\mathbf{X}}(\mathbf{x})$$

be the $L_2(P_{\mathbf{X}})$ squared norm and

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

denote the empirical squared norm and inner product, respectively. 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 $(\mathbf{X}_1, \ldots, \mathbf{X}_n)^T \in \mathbb{R}^{n\times p}$ 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 smallest integer greater than or equal to z.

3. CART methodology

As mentioned earlier, regression trees are commonly constructed with Classification and Regression Trees (CART) [5] methodology. The primary objective of CART is to find partitions of the input 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 $\mathcal{O}(pn\log(n))$ complexity) using a procedure in which a sequence of locally optimal splits recursively partitions the input space. We now describe the procedure 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 that divides the parent node t into left and right daughter nodes t_L and t_R according to whether $X \le s$ or $X > s$, respectively. These two nodes will be denoted by $\mathrm{t}_L = \{\mathbf{X} \in \mathrm{t} : X \le s\}$ and $\mathrm{t}_R = \{\mathbf{X} \in \mathrm{t} : X > s\}$. For a node t in T and data vectors $(W_1, W_2, \dots, W_n)^T$ and $(W_1', W_2', \dots, W_n')^T$, 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 node-specific empirical squared norm and inner product, respectively. Here $N(\mathrm{t}) = \#\{\mathbf{X}_i \in \mathrm{t}\}$ is the number of sample points $\mathbf{X}_i$ in t.

An effective split divides the data from the parent node into two daughter nodes so that the homogeneity 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. (The sample mean in the root node is denoted by $\overline{Y}$.) 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 t_L and $N(\mathrm{t}_L)$ is the sample size of t_L (similar definitions apply to t_R). For a candidate variable X_j and split $s = s_j$, the decrease in impurity is defined as [5, 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 in t that are contained in t_L and t_R, respectively. We also define $w(\mathrm{t}) \coloneqq N(\mathrm{t})/n$ to be the portion of samples $\mathbf{X}_i$ that belong to node t.

At each node t of the tree, we find the variable $\hat{\jmath} = \hat{\jmath}(\mathrm{t})$ and split point $\hat{s} = \hat{s}(\hat{\jmath}, \mathrm{t})$ that maximizes $\widehat{\Delta}(s, j, \mathrm{t})$, 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 decrease in impurity. 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. The output $\widehat{\mu}(\mathrm{t})$ of the tree at a terminal (leaf) node t is the least squares (constant) predictor for data in 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 (1) the node contains a single data point, (2) all response values in the node are the same, or (3) a depth of K is reached, whichever occurs sooner.

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

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

which split X_j at s in t.

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

$$\widetilde{Y}_{\mathrm{t}}(\mathbf{x}) \coloneqq \widehat{Y}_{\mathrm{t}}(\mathbf{x}) \sqrt{1/w(\mathrm{t})}, \tag{7}$$

for internal (non-terminal) nodes t in T. 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. 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" [2] in which one projects the data onto the linear span of a finite collection of dictionary elements). For brevity, all proofs are deferred to Section 6.

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

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

Furthermore, the output of CART admits the orthogonal expansion

$$\widehat{\mu}(T) = \overline{Y} + \sum_{\mathrm{t}} \langle Y, \widetilde{Y}_{\mathrm{t}} \rangle_n \widetilde{Y}_{\mathrm{t}}, \tag{9}$$

where the sum runs over all internal nodes in T, $\|\widetilde{Y}_{\mathrm{t}}\|_n = 1$, and $\langle \widetilde{Y}_{\mathrm{t}}, \widetilde{Y}_{\mathrm{t}'} \rangle_n = 0$ 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 $\{1\} \cup \{\widetilde{Y}_{\mathrm{t}}\}_{\mathrm{t}}$.

4. Main results

Our first theorem 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 reader should recognize and appreciate that we make no assumptions on the decision tree itself (such as a minimum node side condition or shrinking cell condition that typifies extant literature). We also do not deviate from the original CART procedure out of theoretical convenience.

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

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

Remark 1. *Though we have assumed at the outset that Y is almost surely bounded by B, we emphasize that Theorem 4.1 holds for an arbitrary dataset.*

Our second result in Theorem 4.2 establishes an adaptive risk bound (also known as an "oracle inequality") for CART under model mis-specification. Essentially, it says that the CART algorithm 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) of the approximation. That is, 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 Theorem 4.1 and the complexity comes from the fact that the $(1/n)$-covering number for decision trees of depth K constructed from n sample points and p variables is of order $(np)^{2^K}$, with logarithm $2^K \log(np)$.

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

$$\mathbb{E}(\|\mu - \widehat{\mu}(T_K)\|^2) \leq 2 \inf_{g \in \mathcal{G}} \Big\{ \|\mu - g\|^2 + \frac{\|g\|_{TV}^2}{K+3} + C \frac{2^K \log(np)}{n} \Big\}, \tag{10}$$

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

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

$$\frac{4\|\mu\|_{\mathrm{TV}}}{\log_2(n)+6} + 2C\frac{\log(np)}{n^{1/2}}. \tag{11}$$

This bound serves as the basis for our next theorem. Simply put, if the total variation ℓ_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.

Theorem 4.3 (High dimensional consistency of CART for well-specified models)**.** *Suppose $\{\mu_n\}$ is a sequence of p_n-dimensional additive regression functions that belong to $\mathcal{G}$ and $\|\mu_n\|_{TV} \le \Lambda$, where $\Lambda > 0$ is independent of p_n. 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 Theorem 4.3 are satisfied if, for example, $K_n = \llbracket (1/2)\log_2(n) \rrbracket$ and $p_n = \llbracket \exp(cn^{1-\xi}) \rrbracket$ for some universal positive constants c and $\xi \in (1/2, 1]$. In this case, from Theorem 4.2, the consistency rate of the CART algorithm is

$$\frac{4\Lambda}{\log_2(n)+6} + \frac{4C\log(2n)}{n^{1/2}} + \frac{4Cc}{n^{\xi-1/2}} = \mathcal{O}\Big(\frac{1}{\log(n)}\Big).$$

Similar statements as Theorem 4.2 and Theorem 4.3 hold for the excess risk for binary classification, i.e., $Y \in \{0,1\}$, since, in this case, 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) [15, Section 3].

The reader might be somewhat surprised at Theorem 4.2 and Theorem 4.3, especially since they are qualitatively similar to existing results for ensemble learners (such as boosting) or neural networks [14, 1], yet are based on very different principles. For example, [6, Theorem 1] states that boosting with linear learners is also consistent for linear models $\mu(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x}$ on $[0,1]^p$ in the high dimensional regime, i.e., when $p = \llbracket \exp(cn^{1-\xi}) \rrbracket$ and $\|\mu\|_{\mathrm{TV}} = \|\boldsymbol{\beta}\|_{\ell_1} \le \Lambda$, where c is a positive constant and $\Lambda > 0$ is independent of p.

Remark 2. *The proof of Theorem 4.3 uses an explicit choice of depth, namely, $K_n = \llbracket (1/2)\log_2(n) \rrbracket$. However, in practice, it is best to let the data decide and therefore cost complexity pruning (i.e., weakest link pruning [5]) 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 α is a positive constant and $|T|$ is the number of terminal nodes of T. This enables one to obtain oracle inequalities of the form (10), but with the advantage of having the infimum over both the depth $K \geq 1$ and functions $g \in \mathcal{G}$.

Remark 3. *While the prediction error bound in Theorem 4.2 depends only explicitly on p through $\log(p)$, for fixed p, $\mu \in \mathcal{G}$, and $K_n = [\![(1/2)\log_2(n)]\!]$, the prediction error unfortunately decays at a slow logarithmic rate, i.e., $(\log(n))^{-1}$. This is in line with recent work that also shows logarithmic rates for CART, albeit with some additional technical conditions on the regression model (see [13, Theorem 5] and [7, Theorem 2], where p is allowed to scale only polynomially with n). 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" [12, Section 9.2.4]. The issue with trees seems to be due to their high complexity relative to their goodness-of-fit, which is characterized by Theorem 4.2. At the moment, we do not know if the rate $(\log(n))^{-1}$ is optimal for CART within the class of bounded variation additive models. One direction for future work would be determining converse results for CART to supplement our upper bounds.*

4.1. 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 $\varepsilon \equiv 0$) that if each component in the additive model is a step function and the total number of constant pieces of $\mu(\cdot)$ is V, then the squared prediction error achieves the parametric rate $V(p/n)\log(n/p)$ [13, Theorem 4]. Furthermore, if the input 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 input variables, we have the polynomial rate $((p/n)\log(n/p))^{\Omega(1/q)}$ [13, Theorem 3]. Similar polynomial rates in which the exponent of the rate depends only on the sparsity level (and where p scales polynomially with n) were provided in [7] for special classes of additive models (e.g., isotonic or piecewise linear component functions).

The author has also attempted to understand the convergence rates of CART for general classes of univariate regression functions with bounded higher order derivatives. Unfortunately, at the moment, the setting is rather restrictive. For example, using some of the ideas presented in this paper (in particular those which yield Theorem 4.1), one can show convergence rates for noiseless univariate regression that adapt to the smoothness level. To be precise, if $Y = \mu(X)$, where X is uniformly distributed on $[0,1]$, and $\mu(\cdot)$ has an r^{th} order derivative uniformly bounded by B, then if $K_n = [\![(1/2)\log_2(n)]\!]$, we have $\mathbb{E}(\|\mu - \widehat{\mu}(T_{K_n})\|^2) \leq C(\log(n))^{-r}$, where C is a positive constant that depends only on r and B. Interestingly, this bound generalizes (11) in one dimension ($p = 1$) if we think of bounded variation functions as roughly corresponding to differentiable functions ($r = 1$).

4.2. Beyond additive models

In this paper, we considered additive models in part because they are 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. To state succinctly the issues, we note that a crucial step in the proof of Theorem 4.1 involves lower bounding the "information gain" $\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t})$ by the squared excess empirical risk $\|g\|_{\mathrm{TV}}^{-1}(\|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y - g\|_{\mathrm{t}}^2)^2$, 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 6.1). This type of inequality is no longer possible, in general, for models with interaction terms, as one can have $\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \approx 0$ for some node t, even though $\mu(\cdot)$ is nonconstant on t.

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 [8, Proposition 3 and Proposition 4], [4, Section 11], or [3, Section 4.1]). Indeed, our results on the consistency of CART for high dimensional additive models also carry over to Breiman's random forests [4] with relative ease, as we now explain.

To this end, consider a subsample $\mathcal{D}_n'$ of size a_n ($\leq n$) from the original dataset $\mathcal{D}_n$ (selected uniformly at random, without replacement). From this training subsample, 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 $\mathcal{S} = \mathcal{S}(\mathrm{t}) \subset \{1, 2, \ldots, p\}$ of size q and set $\hat{\jmath} = \max_{j \in \mathcal{S}} \widehat{\Delta}(\hat{s}, j, \mathrm{t})$.

We grow M of these depth K regression trees separately using, respectively, M independent realizations $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \ldots, \Theta_M)^T$ of a random variable Θ. Here Θ is distributed according to the law that generates the subsampled training data and candidate variables for splitting at each of the nodes. The regression tree outputs are denoted by $\{\widehat{\mu}(\Theta_m)\}_{m=1}^M$. The random forest output is simply the empirical average of the M regression tree outputs, namely,

$$\widehat{\mu}(\boldsymbol{\Theta}) \coloneqq M^{-1} \sum_{m=1}^{M} \widehat{\mu}(\Theta_m). \tag{12}$$

By a modification of the proofs of Theorem 4.1 and Theorem 4.2, it is possible to show the following oracle inequality for random forests. The proofs are furnished in Section 6.[1]

[1] We deviate slightly from Breiman's original random forests [4] by sampling the data without replacement and not considering fully grown trees (i.e., trees where the number of terminal nodes are of the same order as the sample size, or when $K = [\![\log_2(a_n)]\!]$). For consistency of random forests with fully grown trees, see [20, Theorem 2].

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

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

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

To the best of our knowledge, Theorem 5.1 is the first result in the literature that shows explicitly the impact of randomly choosing subsets of variables as candidates for splitting at the nodes. Even though it is not captured by Theorem 5.1, empirically, the random feature selection mechanism of forests has the effect of de-correlating the constituent trees, which can greatly improve the performance. What Theorem 5.1 does reveal, however, is that this mechanism cannot hurt the prediction error beyond a benign factor of $(p/q)^2$. In fact, standard implementations of regression forests use default values of q and a_n equal to $[\![p/3]\!]$ and $[\![0.632n]\!]$, respectively, the latter of which mimics a bootstrap sample from $\mathcal{D}_n$. With these choices, we see by comparing Theorem 5.1 and Theorem 4.2 that there is essentially no loss in performance (at most, a factor of $(p/q)^2 = 9$) 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 bounds for individual trees when $q = p$ and $a_n = n$.

As with Theorem 4.3, we also have a consistency result for forests.

Theorem 5.2 (High dimensional consistency of random forests for well-specified models)**.** *Suppose $\{\mu_n\}$ is a sequence of p_n-dimensional additive regression functions that belong to $\mathcal{G}$ and $\|\mu_n\|_{TV} \leq \Lambda$, where $\Lambda > 0$ is independent of p_n. If $a_n \to \infty$, $(q_n/p_n)^2 K_n \to \infty$, and $2^{K_n}(\log(a_n p_n))/a_n \to 0$ 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 Theorem 4.3, the hypotheses of Theorem 5.2 are satisfied if, for example, $K_n = [\![(1/2)\log_2(a_n)]\!]$, $p_n = [\![\exp(c a_n^{1-\xi})]\!]$, $q_n = [\![p_n/3]\!]$, and $a_n \to \infty$ as $n \to \infty$, for some universal positive constants c and $\xi \in (1/2, 1]$. 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., $q_n = o(p_n)$ so long as $(q_n/p_n)K_n \to \infty$ as $n \to \infty$.

Theorem 5.2 is a strengthening of [20, Theorem 1], which states that random forests are consistent for additive models when p is fixed, the component functions are continuous, and (in our notation) $a_n \to \infty$, $K_n \to \infty$, and $2^{K_n}(\log(a_n))/a_n \to 0$ as $n \to \infty$.[2] In contrast, here we allow the dimensionality to grow exponentially with the sample size and also for the component functions to be possibly discontinuous. The latter point has practical implications as certain processes encountered in natural and social sciences exhibit, for example, thresholding behavior.

[2]The condition $2^{K_n}(\log(a_n))/a_n \to 0$ is actually stated as $2^{K_n}(\log(a_n))^9/a_n \to 0$ in [20, Theorem 1], where Guassian errors are assumed, but the former condition is sufficient when the noise is bounded.

The author chose not to pursue this avenue any further because, as can be seen from Theorem 5.1 and Theorem 5.2, there is no theoretical advantage of forests over individual trees. Indeed, 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. Fascinating recent work by [17] 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 design and randomly selected subsets of variables, then asymptotically as the number of models goes to infinity, the parameter vector of the ensemble is "shrunk" by a factor of q/p (in our notation). It is possible that a similar phenomenon is occurring in our setting and it would be interesting to explore this further. Certainly more work needs to be done to answer these questions.

6. Proofs

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

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

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

where the final equality comes from [5, Section 9.3].

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

$$\begin{aligned}&\langle Y-\overline{Y}_{\mathrm{t}},\widetilde{Y}_{\mathrm{t}}\rangle_n\widetilde{Y}_{\mathrm{t}}(\mathbf{x})\\ &=\Big\langle Y-\overline{Y}_{\mathrm{t}},\frac{\mathbf{1}(\mathbf{X}\in \mathrm{t}_L)P(\mathrm{t}_R)-\mathbf{1}(\mathbf{X}\in \mathrm{t}_R)P(\mathrm{t}_L)}{\sqrt{w(\mathrm{t})P(\mathrm{t}_L)P(\mathrm{t}_R)}}\Big\rangle_n\frac{\mathbf{1}(\mathbf{x}\in \mathrm{t}_L)P(\mathrm{t}_R)-\mathbf{1}(\mathbf{x}\in \mathrm{t}_R)P(\mathrm{t}_L)}{\sqrt{w(\mathrm{t})P(\mathrm{t}_L)P(\mathrm{t}_R)}}\\ &=(\overline{Y}_{\mathrm{t}_L}-\overline{Y}_{\mathrm{t}})\mathbf{1}(\mathbf{x}\in \mathrm{t}_L)+(\overline{Y}_{\mathrm{t}_R}-\overline{Y}_{\mathrm{t}})\mathbf{1}(\mathbf{x}\in \mathrm{t}_R).\end{aligned}\tag{13}$$

For a terminal node t of T, let $\mathrm{t}_0,\mathrm{t}_1,\dots,\mathrm{t}_K$ be the unique downward path from the root node t_0 to the terminal node $\mathrm{t}_K=\mathrm{t}$. Next, sum up (13) 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:

$$\sum_{k=0}^{K-1}(\overline{Y}_{\mathrm{t}_{k+1}}-\overline{Y}_{\mathrm{t}_k})=\overline{Y}_{\mathrm{t}_K}-\overline{Y}_{\mathrm{t}_0}=\overline{Y}_{\mathrm{t}}-\overline{Y}. \tag{14}$$

Combining (13) and (14), we have that

$$\sum_{\mathrm{t}}\overline{Y}_{\mathrm{t}}\mathbf{1}(\mathbf{x}\in\mathrm{t})=\overline{Y}+\sum_{\mathrm{t}}\langle Y,\widetilde{Y}_{\mathrm{t}}\rangle_n\widetilde{Y}_{\mathrm{t}}(\mathbf{x}),$$

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 standardized decision stumps $\{1\}\cup\{\widetilde{Y}_{\mathrm{t}}\}_{\mathrm{t}}$ are orthonormal. First, notice that each $\widetilde{Y}_{\mathrm{t}}$ has unit empirical norm since

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

where we used $N(\mathrm{t}_L)+N(\mathrm{t}_R)=N(\mathrm{t})$. Next, each $\widetilde{Y}_{\mathrm{t}}$ is orthogonal to the constant function 1 since

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

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

$$\langle\widetilde{Y}_{\mathrm{t}},\widetilde{Y}_{\mathrm{t}'}\rangle_n=0. \tag{15}$$

To see this, first note that (15) 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 $\widetilde{Y}_{\mathrm{t}}$ and $\widetilde{Y}_{\mathrm{t}'}$ will zero each other out when multiplied, i.e., $\widetilde{Y}_{\mathrm{t}}(\mathbf{x})\widetilde{Y}_{\mathrm{t}'}(\mathbf{x})=0$. On the other hand, assume t is an ancestor of t′ and (without loss of generality) that t′ is a descendant of t_L or is equal to t_L. Then we have

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

thus completing the proof. □

Proof of Theorem 4.1. To begin, let $g \in \mathcal{G}$ and define $R_K \coloneqq \|Y - \widehat{\mu}(T_K)\|_n^2 - \|Y - g\|_n^2$. If $R_{K-1} < 0$, then there is nothing to prove because $R_K \le R_{K-1} < 0$ and hence $\|Y - \widehat{\mu}(T_K)\|_n^2 < \|Y - g\|_n^2$. Therefore, we assume throughout that $R_{K-1} \ge 0$.[3] For a terminal node t of T_K, we define $R_K(\mathrm{t}) \coloneqq \|Y - \overline{Y}_\mathrm{t}\|_\mathrm{t}^2 - \|Y - g\|_\mathrm{t}^2$ so that $R_K = \sum_{\mathrm{t} \in T_K} w(\mathrm{t}) R_K(\mathrm{t})$, where 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

$$\|Y - \widehat{\mu}(T_K)\|_n^2 = \|Y - \widehat{\mu}(T_{K-1})\|_n^2 - \sum_{\mathrm{t} \in T_{K-1}} |\langle Y - \overline{Y}_\mathrm{t}, \widetilde{Y}_\mathrm{t} \rangle_n|^2, \tag{16}$$

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 (16) and using the definition of R_K, we have

$$R_K = R_{K-1} - \sum_{\mathrm{t} \in T_{K-1}} |\langle Y - \overline{Y}_\mathrm{t}, \widetilde{Y}_\mathrm{t} \rangle_n|^2. \tag{17}$$

Furthermore, according to the definitions of $\widehat{Y}_\mathrm{t}$ in (6) and $\widetilde{Y}_\mathrm{t}$ in (7), we have $\langle Y - \overline{Y}_\mathrm{t}, \widetilde{Y}_\mathrm{t} \rangle_n = \sqrt{w(\mathrm{t})} \langle Y - \overline{Y}_\mathrm{t}, \widehat{Y}_\mathrm{t} \rangle_\mathrm{t}$, and thus (17) can be rewritten as

$$R_K = R_{K-1} - \sum_{\mathrm{t} \in T_{K-1}} w(\mathrm{t}) |\langle Y - \overline{Y}_\mathrm{t}, \widehat{Y}_\mathrm{t} \rangle_\mathrm{t}|^2. \tag{18}$$

Throwing away terms in the sum for $\mathrm{t} \in T_{K-1}$ such that $R_{K-1}(\mathrm{t}) < 0$, note that (18) satisfies the inequality

$$R_K \le R_{K-1} - \sum_{\mathrm{t} \in T_{K-1} : R_{K-1}(\mathrm{t}) \ge 0} w(\mathrm{t}) |\langle Y - \overline{Y}_\mathrm{t}, \widehat{Y}_\mathrm{t} \rangle_\mathrm{t}|^2. \tag{19}$$

By Lemma 6.1, if t is a terminal node of T_{K-1} and $R_{K-1}(\mathrm{t}) \ge 0$, we have

$$\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \ge \frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathrm{TV}}^2}. \tag{20}$$

The identity (8) from Lemma 3.1 and (20) together imply that

$$|\langle Y - \overline{Y}_\mathrm{t}, \widehat{Y}_\mathrm{t} \rangle_\mathrm{t}|^2 \ge \frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathrm{TV}}^2}. \tag{21}$$

Next, we apply (21) to each term in the sum from (19) to obtain

$$R_K \le R_{K-1} - \frac{1}{\|g\|_{\mathrm{TV}}^2} \sum_{\mathrm{t} \in T_{K-1} : R_{K-1}(\mathrm{t}) \ge 0} w(\mathrm{t}) R_{K-1}^2(\mathrm{t}). \tag{22}$$

[3] It can be seen from (16) that the training error of CART decreases with the depth so that $R_1 \ge R_2 \ge \cdots \ge R_K$.

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

$$\sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t})R_{K-1}^2(\mathrm{t}) \geq \Big(\sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t})R_{K-1}(\mathrm{t})\Big)^2. \tag{23}$$

Also,

$$\begin{aligned}\sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t})R_{K-1}(\mathrm{t}) &\geq \sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})\geq 0} w(\mathrm{t})R_{K-1}(\mathrm{t}) + \\ &\quad \sum_{\mathrm{t}\in T_{K-1}:R_{K-1}(\mathrm{t})<0} w(\mathrm{t})R_{K-1}(\mathrm{t}) \\ &= R_{K-1},\end{aligned} \tag{24}$$

which is nonnegative by assumption. Applying (24) to (23), we obtain

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

Thus, combining (22) and (25), we have established the recursion

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

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 6.1. *Let $R_K(\mathrm{t}) \coloneqq \|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y - g\|_{\mathrm{t}}^2$ for $g \in \mathcal{G}$ and $w(\mathrm{t}) \coloneqq N(\mathrm{t})/n$ for all terminal nodes t of T_K. Let t be a terminal node of T_{K-1} and suppose $R_{K-1}(\mathrm{t}) \geq 0$. Then,*

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

Proof of Lemma 6.1. We use recent tools for the analysis of decision trees from [13], namely, Lemma A.4 in the supplement. Define an empirical distribution $\Pi(s,j)$ on split points s and variables X_j in the daughter nodes t_L and t_R via

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

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

$X_i \le s < X_{i+1}$ and $i = 1, 2, \dots, N(\mathrm{t}) - 1$. 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,

$$\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \ge \int \widehat{\Delta}(s, j, \mathrm{t}) d\Pi(s, j) = \int |\langle Y - \overline{Y}_{\mathrm{t}}, \widehat{Y}_{\mathrm{t}} \rangle_{\mathrm{t}}|^2 d\Pi(s, j), \tag{26}$$

where the last identity follows from (8) in Lemma 3.1. Next, by Jensen's inequality for the square function, (26) is further lower bounded by

$$\int |\langle Y - \overline{Y}_{\mathrm{t}}, \widehat{Y}_{\mathrm{t}} \rangle_{\mathrm{t}}|^2 d\Pi(s, j) \ge \Big(\int |\langle Y - \overline{Y}_{\mathrm{t}}, \widehat{Y}_{\mathrm{t}} \rangle_{\mathrm{t}}| d\Pi(s, j) \Big)^2. \tag{27}$$

We next evaluate the expectation in (27) with respect to the measure Π, giving

$$\int |\langle Y - \overline{Y}_{\mathrm{t}}, \widehat{Y}_{\mathrm{t}} \rangle_{\mathrm{t}}| d\Pi(s, j) = \frac{\sum_{j=1}^p \int |g'_j(s)| |\langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}}| ds}{\sum_{j'=1}^p \int |g'_{j'}(s')| \sqrt{P(\mathrm{t}'_L) P(\mathrm{t}'_R)} ds'}, \tag{28}$$

where we used the identity $\sqrt{P(\mathrm{t}_L) P(\mathrm{t}_R)} \langle Y - \overline{Y}_{\mathrm{t}}, \widehat{Y}_{\mathrm{t}} \rangle_{\mathrm{t}} = -\langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}}$, 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, we use Jensen's inequality for the absolute value function on the numerator in (28), yielding

$$\sum_{j=1}^p \int |g'_j(s)| |\langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}}| ds \ge \Big| \sum_{j=1}^p \int g'_j(s) \langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}} ds \Big|. \tag{29}$$

Using linearity of the inner product and integration and the fundamental theorem of calculus, the expression in the right hand side of (29) can be simplified as follows

$$\begin{aligned} \sum_{j=1}^p \int g'_j(s) \langle Y - \overline{Y}_{\mathrm{t}}, \mathbf{1}(X_j > s) \rangle_{\mathrm{t}} ds &= \Big\langle Y - \overline{Y}_{\mathrm{t}}, \sum_{j=1}^p \int g'_j(s) \mathbf{1}(X_j > s) ds \Big\rangle_{\mathrm{t}} \\ &= \Big\langle Y - \overline{Y}_{\mathrm{t}}, \sum_{j=1}^p g_j(X_j) \Big\rangle_{\mathrm{t}} \\ &= \langle Y - \overline{Y}_{\mathrm{t}}, g \rangle_{\mathrm{t}}. \end{aligned} \tag{30}$$

Combining all of these inequalities, namely, (26)-(30), proves that

$$\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t}) \ge \frac{|\langle Y - \overline{Y}_{\mathrm{t}}, g \rangle_{\mathrm{t}}|^2}{\big(\sum_{j'=1}^p \int |g'_j(s')| \sqrt{P(\mathrm{t}'_L) P(\mathrm{t}'_R)} ds' \big)^2}. \tag{31}$$

Our next goal is to provide respective lower and upper bounds on the numerator and

denominator of (31). For the denominator of (31), note that for each j',

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

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

$$\sum_{j'=1}^{p}\int |g'_{j'}(s')|\sqrt{P(\mathrm{t}'_L)P(\mathrm{t}'_R)}ds' \le 2^{-1}\|g\|_{\mathrm{TV}}. \tag{32}$$

For the numerator of (31), 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

$$|\langle Y-\overline{Y}_{\mathrm{t}}, g\rangle_{\mathrm{t}}|^2 \ge 4^{-1}(\|Y-\overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y-g\|_{\mathrm{t}}^2)^2 = 4^{-1}R_{K-1}^2(\mathrm{t}). \tag{33}$$

Applying inequalities (32) and (33) to (31), we therefore have shown that

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

which completes the proof. □

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

$$E_1 \coloneqq \|\mu-\widehat{\mu}(T_K)\|^2 - 2(\|Y-\widehat{\mu}(T_K)\|_n^2 - \|Y-\mu\|_n^2) - \alpha - \beta$$

and

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

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

$$E_2 \le 2(\|Y-g\|_n^2 - \|Y-\mu\|_n^2) + \frac{2\|g\|_{\mathrm{TV}}}{K+3} + \alpha + \beta, \tag{34}$$

for any $g \in \mathcal{G}$. Taking expectations on both sides of (34) and using $\mathbb{E}(\|Y-g\|_n^2 - \|Y-\mu\|_n^2) = \|\mu - g\|^2$ yields

$$\mathbb{E}(E_2) \le 2\mathbb{E}(\|Y-g\|_n^2 - \|Y-\mu\|_n^2) + \frac{2\|g\|_{\mathrm{TV}}}{K+3} + \alpha = 2\|\mu-g\|^2 + \frac{2\|g\|_{\mathrm{TV}}}{K+3} + \alpha + \beta. \tag{35}$$

To bound E_1, we first introduce some useful definitions due to [18]. 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 tree on n points (in particular, note that Λ_n contains all data-dependent partitions). 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. Let $\mathcal{F}_n$ denote the collection of all piecewise constant functions (bounded by B) on partitions $\mathcal{P} \in \Lambda_n$.

Now, by [11, Theorem 11.4] (with, in their notation, $\epsilon = 1/2$), we have

$$\begin{aligned}\mathbb{P}\big(\exists\, f \in \mathcal{F}_n : \|\mu - f\|^2 \ge 2(\|Y-f\|_n^2 - \|Y-\mu\|_n^2) + \alpha + \beta\big) \le \\ 14 \sup_{\mathbf{x}^n} \mathcal{N}\Big(\frac{\beta}{40B}, \mathcal{F}_n, L_1(\nu_{\mathbf{x}^n})\Big) \exp\Big(-\frac{\alpha n}{2568 B^4}\Big),\end{aligned} \tag{36}$$

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

Next, we use [11, Lemema 13.1 and Theorem 9.4] to bound the empirical covering number by

$$\mathcal{N}\Big(\frac{\beta}{40B}, \mathcal{F}_n, L_1(\nu_{\mathbf{x}^n})\Big) \le \Gamma_n(\Lambda_n)\Big(\frac{40}{32}\frac{333eB^2}{\beta}\Big)^{2^{K+1}} \le \Gamma_n(\Lambda_n)\Big(\frac{417eB^2}{\beta}\Big)^{2^{K+1}}. \tag{37}$$

We further bound (37) by noting that $\Gamma_n(\Lambda_n) \le ((n-1)p)^{2^K - 1} \le (np)^{2^K}$ (letting $\gamma_K = \Gamma_n(\Lambda_n)$ for depth K trees, we note that $\gamma_K \le ((n-1)p)\gamma_{K-1}^2$ with $\gamma_1 \le (n-1)p$, since there are at most $(n-1)p$ possible splits at the root node).[4]

Returning to (36), since $\widehat{\mu}(T_K) \in \mathcal{F}_n$, we thus have

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

[4]The author was inspired by [20] for this estimate of $\Gamma_n(\Lambda_n)$.

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 $P(E_1 \geq 0) \leq 1/n$. Since $E_1 \leq \|\mu - \widehat{\mu}(T_K)\|^2 + 2\|Y-\mu\|_n^2 \leq 12B^2$, 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 (35) and plugging in the choices of α and β, we have

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

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

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

Proof of Theorem 5.1. The proof follows similar lines as Theorem 4.1 and Theorem 4.2, but with some important differences.

We first remind the reader of how the base regression trees in the forest are constructed. Consider a subsample $\mathcal{D}_n'$ of size a_n ($\leq n$) from the original dataset $\mathcal{D}_n$ (drawn uniformly at random, without replacement). From this training subsample, we grow a depth K regression tree T_K with CART methodology in the usual way, except that, at each internal node t, the objective $\widehat{\Delta}(\hat{s}, j, \mathrm{t})$ is optimized over $j \in \mathcal{S}$, where $\mathcal{S} = \mathcal{S}(\mathrm{t}) \subset \{1, 2, \ldots, p\}$ is a random subset formed by selecting q of the p variables uniformly at random, without replacement. For $k = 1, 2, \ldots, K$, let Ψ_k denote the random variable whose law generates the subsets $\mathcal{S}$ for all internal nodes in T_k, conditional on the subsampled training data $\mathcal{D}_n'$.

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

We proceed by first bounding the training error averaged with respect to Ψ_K; that is, we aim to bound $\mathbb{E}_{\Psi_K}(\|Y - \widehat{\mu}(T_K)\|_{a_n}^2)$. The same proof as Theorem 4.1 goes through, except now in Lemma 6.1 one shows that, for each terminal node t of T_{K-1},

$$\mathbb{E}_{\Psi_K|\Psi_{K-1}}(\widehat{\Delta}(\hat{s}, \hat{\jmath}, \mathrm{t})) \geq \frac{q^2}{p^2}\frac{(\|Y - \overline{Y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|Y-g\|_{\mathrm{t}}^2)^2}{\|g\|_{\mathrm{TV}}^2} = \frac{q^2}{p^2}\frac{R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathrm{TV}}^2}, \tag{38}$$

provided $R_{K-1}(\mathrm{t}) \geq 0$. Let us first establish the validity of the above inequality (38).

Following a similar argument as the proof of Lemma 6.1, we have

$$\begin{aligned}\mathbb{E}_{\Psi_K|\Psi_{K-1}}(\widehat{\Delta}(\hat{s},\hat{j},\mathrm{t})) &\geq \mathbb{E}_{\Psi_K|\Psi_{K-1}}\Big(\int \widehat{\Delta}(s,j,\mathrm{t})d\Pi(s,j)\Big)\\ &= \mathbb{E}_{\Psi_K|\Psi_{K-1}}\Big(\int |\langle Y-\overline{Y}_{\mathrm{t}},\widehat{Y}_{\mathrm{t}}\rangle_{\mathrm{t}}|^2 d\Pi(s,j)\Big)\\ &\geq \Big(\mathbb{E}_{\Psi_K|\Psi_{K-1}}\int \langle Y-\overline{Y}_{\mathrm{t}},\widehat{Y}_{\mathrm{t}}\rangle_{\mathrm{t}} d\Pi(s,j)\Big)^2,\end{aligned}$$

where the last line follows from Jensen's inequality for the square function (note that, given a set $\mathcal{S}$ of candidate variables, the measure Π is now defined for splits along variables X_j with $j\in\mathcal{S}$). Thus, (32) is still bounded by

$$\sum_{j'\in\mathcal{S}}\int |g'_{j'}(s')|\sqrt{P(\mathrm{t}'_L)P(\mathrm{t}'_R)}ds' \leq \sum_{j'=1}^{p}\int |g'_{j'}(s')|\sqrt{P(\mathrm{t}'_L)P(\mathrm{t}'_R)}ds' \leq 2^{-1}\|g\|_{\mathrm{TV}},$$

as before, but (30) becomes

$$\begin{aligned}&\mathbb{E}_{\Psi_K|\Psi_{K-1}}\Big(\sum_{j\in\mathcal{S}}\int g'_j(s)\langle Y-\overline{Y}_{\mathrm{t}},\mathbf{1}(X_j>s)\rangle_{\mathrm{t}}ds\Big)\\ &=\Big\langle Y-\overline{Y}_{\mathrm{t}},\mathbb{E}_{\Psi_K|\Psi_{K-1}}\Big(\sum_{j\in\mathcal{S}}\int g'_j(s)\mathbf{1}(X_j>s)ds\Big)\Big\rangle_{\mathrm{t}}\\ &=\Big\langle Y-\overline{Y}_{\mathrm{t}},\mathbb{E}_{\mathcal{S}}\Big(\sum_{j\in\mathcal{S}}g_j(X_j)\Big)\Big\rangle_{\mathrm{t}}\\ &=\Big\langle Y-\overline{Y}_{\mathrm{t}},\frac{q}{p}\sum_{j=1}^{p}g_j(X_j)\Big\rangle_{\mathrm{t}}=\frac{q}{p}\langle Y-\overline{Y}_{\mathrm{t}},g\rangle_{\mathrm{t}}.\end{aligned}$$

Note that the above identity holds when $\mathcal{S}$ is formed by selecting a subset of size q from $\{1,2,\dots,p\}$ uniformly at random *without* replacement. If sampling is done with replacement, the factor q/p is replaced by $1-(1-1/p)^q$.

Following the same steps as the proof of Lemma 6.1 yields (38), as claimed. Proceeding in the same way as before with the proof of Theorem 4.1, using (38), one can easily establish the inequality

$$\mathbb{E}_{\Psi_K|\Psi_{K-1}}(R_K)\leq R_{K-1}\big(1-\frac{q^2}{p^2}\frac{R_{K-1}}{\|g\|^2_{\mathrm{TV}}}\big). \tag{39}$$

Taking expected values of (39) with respect to Ψ_{K-1} and using, in turn, the law of iterated expectations and Jensen's inequality for the square function, i.e., $\mathbb{E}_{\Psi_{K-1}}(\mathbb{E}_{\Psi_K|\Psi_{K-1}}(R_K))=\mathbb{E}_{\Psi_K}(R_K)$ and $\mathbb{E}_{\Psi_{K-1}}(R^2_{K-1})\geq(\mathbb{E}_{\Psi_{K-1}}(R_{K-1}))^2$, we have

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

Iterating this recursion as in the proof of Theorem 4.1 yields $\mathbb{E}_{\Psi_K}(R_K) \le (p/q)^2\|g\|_{\mathrm{TV}}^2/(K+3)$, or equivalently,

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

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

$$\begin{aligned}\mathbb{P}\big(\mathbb{E}_{\Psi_K}(\|\mu-\widehat{\mu}(T_K)\|^2) \ge 2(\mathbb{E}_{\Psi_K}(\|Y-\widehat{\mu}(T_K)\|_{a_n}^2) - \|Y-\mu\|_{a_n}^2) + \alpha + \beta\big) \le \\ \mathbb{P}\big(\exists\, f \in \mathcal{F}_n : \|\mu-f\|^2 \ge 2(\|Y-f\|_{a_n}^2 - \|Y-\mu\|_{a_n}^2) + \alpha + \beta\big),\end{aligned}$$

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

$$\mathbb{E}_{\Psi_K,\mathcal{D}_n'|\mathcal{I}}(\|\mu-\widehat{\mu}(T_K)\|^2) \le 2\|\mu-g\|^2 + \frac{p^2}{q^2}\frac{2\|g\|_{\mathrm{TV}}^2}{K+3} + C\frac{2^{K+1}\log(a_n p)}{a_n}, \tag{40}$$

where C is a positive constant that depends only on B. Thus, taking expectations of (40) with respect to $\mathcal{I}$, we have

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

The output of a random forest (12) 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 squared error yields $\mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}_n}(\|\mu-\widehat{\mu}(\boldsymbol{\Theta})\|^2) \le \mathbb{E}_{\Theta,\mathcal{D}_n}(\|\mu-\widehat{\mu}(\Theta)\|^2)$. This proves Theorem 5.1, from which Theorem 5.2 follows immediately. □

References

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

[2] 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

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

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

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

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

[7] Chi, C.-M., Vossler, P., Fan, Y. and Lv, J. (2020). Asymptotic Properties of High-Dimensional Random Forests. *arXiv preprint arXiv:2004.13953*.
[8] 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)*.
[9] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 1189–1232.
[10] Gey, S. and Nedelec, E. (2005). Model selection for CART regression trees. *IEEE Transactions on Information Theory* **51** 658–670.
[11] Györfi, L., Krzyżak, A., Kohler, M. and Walk, H. (2002). *A distribution-free theory of nonparametric regression*. Springer.
[12] 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.
[13] Klusowski, J. M. (2020). Sparse Learning with CART. In *Advances in Neural Information Processing Systems*.
[14] 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
[15] Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint: arXiv:1407.7502*.
[16] Meier, L., Geer, V., and Bühlmann, P. (2009). High-Dimensional Additive Modeling. *The Annals of Statistics* **37** 3779–3821.
[17] 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.
[18] Nobel, A. et al. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics* **24** 1084–1105.
[19] 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.
[20] Scornet, E., Biau, G. and Vert, J.-P. (2015). Consistency of random forests. *Annals of Statistics* **43** 1716–1741.
[21] Stone, C. J. (1977). Consistent nonparametric regression. *Annals of Statistics* 595–620.
[22] Wager, S. and Athey, S. (2018). Estimation and Inference of Heterogeneous Treatment Effects using Random Forests. *Journal of the American Statistical Association* 1–15.
[23] Wager, S. and Walther, G. (2015). Adaptive concentration of regression trees, with application to random forests. *arXiv preprint arXiv:1503.06388*.
[24] Yuan, M. and Zhou, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *The Annals of Statistics* **44** 2564–2593.