JOURNAL OF THE AMERICAN STATISTICAL ASSOCIATION
2022, VOL. 00, NO. 0, 1–13: Theory and Methods
https://doi.org/10.1080/01621459.2022.2126782

Taylor & Francis
Taylor & Francis Group

Large Scale Prediction with Decision Trees

Jason M. Klusowski and Peter M. Tian

Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ

ABSTRACT

This article shows that decision trees constructed with Classification and Regression Trees (CART) and C4.5 methodology are consistent for regression and classification tasks, even when the number of predictor variables grows sub-exponentially with the sample size, under natural 0-norm and 1-norm sparsity constraints. The theory applies to a wide range of models, including (ordinary or logistic) additive regression models with component functions that are continuous, of bounded variation, or, more generally, Borel measurable. Consistency holds for arbitrary joint distributions 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 mis-specified model. Supplementary materials for this article are available online.

ARTICLE HISTORY
Received July 2021
Accepted August 2022

KEYWORDS
C4.5; CART; Interpretable machine learning; Oracle inequality; Random forests;

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, that is, bagging (Breiman 1996), random forests (Breiman 2001), and gradient tree boosting (Friedman 2001). From an applied perspective, decision trees scale well to large datasets, and are intuitive and interpretable, the latter of which makes them easy to explain to statistical nonexperts, particularly in the context of rule-based decision-making. 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 trees 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.[1]

Though our primary focus is theoretical, to make this article likewise relevant to the applied user of decision trees, we focus on Classification And Regression Trees (CART) (Breiman et al. 1984) and C4.5 (Quinlan 1993) methodology—undoubtedly the most popular varieties for regression and classification problems. On the theoretical side, these approaches raise 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 and C4.5 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 article:

> Decision trees constructed with CART and C4.5 methodology are consistent for large scale predictive models, where the number of predictor variables is allowed to grow sub-exponentially with the sample size, under ℓ_0 or ℓ_1 sparsity constraints.

The consistency (with respect to mean squared error risk for regression and excess mis-classification risk for classification) is shown under essentially no assumptions on the predictor variables, thereby improving upon most past work which requires the them to be continuous and either independent or near-independent (e.g., uniformly distributed or with joint densities which are bounded above and below by fixed positive constants).

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 decision trees, starting with CART in the regression setting, and then for C4.5 in the classification setting.

Regression trees. The first consistency result for CART was provided in the original book that proposed the methodology (Breiman et al. 1984), albeit under very strong assumptions on

[1] The oft-touted interpretability of CART and C4.5 is sometimes compromised by the instability of the splits, particularly among deeper nodes where less data is available. Therefore, one should be cautious when attaching any meaning or interpretation to such splits.

CONTACT Jason M. Klusowski jason.klusowski@princeton.edu Department of Operations Research and Financial Engineering, Princeton University, Princeton, NJ.
Supplementary materials for this article are available online. Please go to www.tandfonline.com/r/JASA.

the tree construction, such as a minimum node size condition and shrinking cell condition. Thirty years later, Scornet, Biau, and Vert (2015) showed asymptotic consistency of CART for (fixed dimensional) additive regression models with continuous component functions, en route to establishing asymptotic consistency of Breiman's random forests. This article was an important technical breakthrough 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), Syrgkanis and Zampetakis (2020), and 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 risk vanishes also (Breiman et al. 1984; Denil, Matheson, and De Freitas 2014; Wager and Walther 2015; Wager and Athey 2018). In contrast, the aforementioned paper (Scornet, Biau, and Vert 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), and Syrgkanis 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 Syrgkanis and Zampetakis (2020) work with binary valued predictor variables. In addition, a local accuracy gain condition (akin to an edge or progress condition in boosting literature) is required in these works (Chi et al. 2020; Klusowski 2020; Syrgkanis 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 regression models, such as those with isotonic or piece-wise linear component functions and independent predictor variables. Syrgkanis and Zampetakis (2020) do not provide any concrete examples of models that satisfy what they call a *sub-modularity* 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.

Classification trees. The story for the classification setting is far less complete. Theory for regression trees is far easier to assemble—and has therefore been overwhelmingly the focus of past literature—since one does not have to deal with the discrete nature of the model. This lack of attention is somewhat unfortunate as decision trees are more often successfully deployed in problems with discrete outputs, that is, clinical decision support systems. As an exception, one stand-alone paper to tackle the classification problem is Kearns and Mansour (1999), where the authors show that classification trees constructed with CART and C4.5 methodology have small mis-classification risk under a weak hypothesis assumption, that is, the decision tree output in each node performs slightly better than random guessing as approximations to the target function. Their results, however, do not account for the effect of dimensionality, nor do they accommodate standard statistical models for classification, such as logistic regression.

1.2. Related Work

In closing, we mention a few papers that study other tree based procedures, but with different aims and from different perspectives. Mondrian random forests (Mourtada, Gaïffas, and Scornet 2020, 2021), unlike some of the aforementioned variants of Breiman's random forests, provably attain near-optimal minimax rates for various levels of smoothness regularity. The dyadic CART procedure of Donoho (1997), obtained by optimal (non-greedy), dyadic recursive partitioning, was also shown to achieve near-optimal rates (and adaptation to unknown smoothness regularity) for the case of two predictor variables. (Dyadic CART unfortunately scales poorly with the dimensionality p, since finding the optimal dyadic partition may require up to $\mathcal{O}(2^p Np)$ operations (Chatterjee and Goswami 2021)). An interesting line of work from a Bayesian perspective explores Bayesian trees and forests. For example, Ročková and van der Pas (2020) and Jeong and Ročková (2020) obtain near-optimal posterior concentration rates for nonparametric regression, similar to the optimal minimax rates discussed in Section 4.5, that adapt to both exact sparsity and smoothness regularity (for Bayesian CART and BART) and additive structures (for BART).

1.3. Organization

This article is organized according the following schema. In Section 2, we describe a unified statistical framework for regression and classification problems, and introduce various important quantities for performance assessment. We review basic terminology associated with decision trees and describe CART and C4.5 methodology in Section 3. Our main results for CART and C4.5 are contained in Section 4; specifically, an empirical risk bound, oracle inequality, and asymptotic consistency statement for (ordinary or logistic) additive regression models. We then show that C4.5 achieves a faster consistency rate for separable, large margin binary data in Section 5. Our main consistency results in Section 4 are extended to models with interactions in Section 6 and Breiman's random forests in Section 7. We conclude with a discussion in Section 8. Finally, all proofs and technical lemmas are contained in the supplementary materials.

2. Statistical Framework

Throughout this article, we operate under a standard predictive framework for regression and classification; that is, from training data, we desire to predict a response value y for a new set of p predictor variables $\mathbf{x}$. More formally, we observe training data $\mathcal{D} := \{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$ drawn iid from the statistical 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 $\mathbb{R}^p$. To understand the predictive properties of decision trees in an (ultra) high dimensional setting, the dimensionality $p = p_N$ is permitted to grow sub-exponentially with the sample size N.

In order to unify both the regression and classification settings, we consider a discriminative statistical model in which the conditional mean of y given $\mathbf{x}$ is modeled indirectly via a (possibly nonlinear) link function. To this end, we assume that there is a fixed and known link function $h : \mathbb{R} \to \mathbb{R}$ such that

$$g^*(\mathbf{x}) := h(\mathbb{E}(y|\mathbf{x})) \in \mathcal{G}, \tag{1}$$

where $\mathcal{G}$ is some class of $\mathbb{R}^p \to \mathbb{R}$ functions to be chosen ahead. A loss function $\mathcal{L}(\cdot,\cdot)$ is chosen so that

$$g^*(\cdot) \in \operatorname{argmin}_{g(\cdot)\in\mathcal{G}} \mathbb{E}_{(\mathbf{x},y)}\mathcal{L}(y, g(\mathbf{x})). \tag{2}$$

The *empirical and population risk* corresponding to the loss function $\mathcal{L}(\cdot,\cdot)$ are denoted by

$$\widehat{\mathcal{R}}(g) := \frac{1}{N}\sum_{i=1}^{N}\mathcal{L}(y_i, g(\mathbf{x}_i)), \qquad \mathcal{R}(g) := \mathbb{E}_{(\mathbf{x},y)}\mathcal{L}(y, g(\mathbf{x})),$$

respectively. In light of (2), we aim to estimate the true model $g^*(\cdot)$ by finding a $\mathcal{D}$-dependent fit $g(\cdot)$, not necessarily in $\mathcal{G}$, which comes close to minimizing the empirical risk $\widehat{\mathcal{R}}(g)$.

2.1. Regression Setting

We assume a real-valued response variable $y \in \mathbb{R}$ and consider the identity link function $h(\mu) = \mu$ so that $g^*(\mathbf{x}) = \mathbb{E}(y|\mathbf{x})$ is the conditional mean response. We choose the loss function to be the squared error

$$\mathcal{L}(y, g(\mathbf{x})) = (y - g(\mathbf{x}))^2, \tag{3}$$

which satisfies (2). The out-of-sample performance is measured with the squared $\mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$ norm

$$\|g - g^*\|^2 := \mathbb{E}_{\mathbf{x}}((g(\mathbf{x}) - g^*(\mathbf{x}))^2), \tag{4}$$

which also equals the excess squared error risk $\mathcal{R}(g) - \mathcal{R}(g^*)$ under these model specifications.

2.2. Classification Setting

We assume a binary response variable $y \in \{-1, 1\}$ and consider the logit link function $h(\mu) = \log(\frac{1+\mu}{1-\mu})$, so that $g^*(\mathbf{x}) = h(2\eta^*(\mathbf{x}) - 1) = \log(\frac{\eta^*(\mathbf{x})}{1-\eta^*(\mathbf{x})})$ is the log-odds of the conditional class probability $\eta^*(\mathbf{x}) := \mathbb{P}(y = 1|\mathbf{x}) = 1/(1 + \exp(-g^*(\mathbf{x})))$. Note that if $g^*(\cdot)$ is a linear function, then this model is linear logistic regression. We choose the loss function to be the logistic loss

$$\mathcal{L}(y, g(\mathbf{x})) = \log(1 + \exp(-yg(\mathbf{x}))), \tag{5}$$

which satisfies (2) and, appealingly, corresponds to maximum likelihood estimation of conditionally Bernoulli distributed data with probability model $\mathbb{P}(y|\mathbf{x}) = 1/(1 + \exp(-yg(\mathbf{x})))$. Instead of assessing performance with the logistic risk, we more familiarly consider the mis-classification risk

$$\operatorname{Err}(g) := \mathbb{P}_{(\mathbf{x},y)}(c(\mathbf{x}) \neq y) \tag{6}$$

of the plug-in classifier

$$c(\mathbf{x}) := \begin{cases} +1 & \text{if } g(\mathbf{x}) \geq 0 \\ -1 & \text{if } g(\mathbf{x}) < 0 \end{cases}.$$

The out-of-sample performance is then measured by the excess mis-classification risk

$$\operatorname{Err}(g) - \operatorname{Err}(g^*), \tag{7}$$

where $\operatorname{Err}(g^*)$ is the Bayes error rate.

Definition 2.1. When no context is provided, *prediction risk* will refer to excess squared error risk (4) in the regression setting (Section 2.1) or excess mis-classification risk (7) in the classification setting (Section 2.2).

3. CART and C4.5 Methodology

As mentioned earlier, regression trees are commonly constructed with Classification and Regression Trees (CART) methodology. In the context of classification, one can use either CART or its contemporary counterpart, C4.5. For technical reasons that will be explained at the end of this section, we will focus on the latter methodology. While CART and C4.5 are algorithmically similar, they differ in important ways. The notable distinction lies in the criterion used to determine the split points, which turns out to be the key to their success for the predictive models we consider in Sections 2.1 and 2.2. In a nutshell, the objective of decision tree learning is to find partitions of the predictor variables that produce minimal empirical risk of the constant (average response) values over the partition. Because of the computational infeasibility of choosing the best overall partition, CART and C4.5 operate in a greedy, top down fashion (with a mere $\mathcal{O}(pN\log^2(N))$ average-case complexity (Louppe 2014, sec. 5)) using a procedure in which a sequence of locally optimal splits recursively partition the input space.

In Section 3.1, we first discuss the splitting rule, stopping criterion, and tree output for a generic (top down, greedy) tree construction algorithm. We then specialize our treatment of decision trees to CART and C4.5 in Sections 3.2 and 3.3, respectively.

3.1. Greedy Tree Construction

Consider splitting a decision tree T at a node t (a hyperrectangular region in $\mathbb{R}^p$). Let s be a candidate split point for a variable $x_j \in \mathbb{R}$ that divides the parent node t into left and right daughter nodes t_L and t_R according to whether $x_j \leq s$ or $x_j > s$, respectively. These two nodes will be denoted by $\mathrm{t}_L := \{\mathbf{x} \in \mathrm{t} : x_j \leq s\}$ and $\mathrm{t}_R := \{\mathbf{x} \in \mathrm{t} : x_j > s\}$.

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 maximally reduced from that of the parent node. The impurity is determined by the within-node empirical risk

$$\widehat{\mathcal{R}}_{\mathrm{t}}(g) := \frac{1}{N_{\mathrm{t}}} \sum_{\mathbf{x}_i \in \mathrm{t}} \mathcal{L}(y_i, g(\mathbf{x}_i)), \quad N_{\mathrm{t}} := \#\{\mathbf{x}_i \in \mathrm{t}\}. \tag{8}$$

In accordance with the directive of minimizing the empirical risk, it is equal to the smallest within-node empirical risk over all constant predictors in the node, that is,

$$\mathcal{I}(\mathrm{t}) = \min_{\beta \in \mathbb{R}} \widehat{\mathcal{R}}_{\mathrm{t}}(\beta) = \widehat{\mathcal{R}}_{\mathrm{t}}(h(\overline{y}_{\mathrm{t}})), \quad \overline{y}_{\mathrm{t}} := \frac{1}{N_{\mathrm{t}}} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i. \tag{9}$$

The parent node t is split into two daughter nodes using a variable $x_{j_{\mathrm{t}}}$ and split point s_{t} which produce the largest impurity gain (Breiman et al. 1984, Definition 8.13), (Quinlan 1993, p. 22)

$$\begin{aligned} \mathcal{IG}(j, s, \mathrm{t}) &:= \mathcal{I}(\mathrm{t}) - P_{\mathrm{t}_L}\mathcal{I}(\mathrm{t}_L) - P_{\mathrm{t}_R}\mathcal{I}(\mathrm{t}_R), \\ \mathcal{IG}(\mathrm{t}) &:= \max_{(j,s)} \mathcal{IG}(j, s, \mathrm{t}), \end{aligned} \tag{10}$$

breaking ties arbitrarily, where $P_{\mathrm{t}_L} := N_{\mathrm{t}_L}/N_{\mathrm{t}}$ and $P_{\mathrm{t}_R} := N_{\mathrm{t}_R}/N_{\mathrm{t}}$ are the proportions of data points within t that are contained in t_L and t_R, respectively. Equivalently, the variable and split point, chosen to maximize (10), also minimize the within-node empirical risk (8) over all *decision stumps*, since

$$P_{\mathrm{t}_L}\mathcal{I}(\mathrm{t}_L) + P_{\mathrm{t}_R}\mathcal{I}(\mathrm{t}_R) = \min_{\beta_0, \beta_1 \in \mathbb{R}} \widehat{\mathcal{R}}_{\mathrm{t}}(\beta_0 + \beta_1 \mathbf{1}(x_j > s)). \tag{11}$$

We can thus view the maximum impurity gain $\mathcal{IG}(\mathrm{t})$ as the amount by which the optimal decision stump decreases the empirical risk in the node.

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 K 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 article, we use the following stopping rule.

Definition 3.1 (Stopping rule). We stop splitting a node if (i) the node contains a single data point, (ii) all input values and/or all response values within the node are the same, or (iii) a depth of K is reached, whichever occurs sooner.

Finally, the output of the tree T at a terminal node t is the best constant predictor in the node:

$$\widehat{g}(T)(\mathbf{x}) := h(\overline{y}_{\mathrm{t}}) \approx g^*(\mathbf{x}), \quad \mathbf{x} \in \mathrm{t}. \tag{12}$$

When we wish to emphasize the dependence of the tree T on the depth K, we will write T_K.

3.2. CART Algorithm

Recall the regression setting in Section 2.1, where $h(\cdot)$ is the identity link and $\mathcal{L}(\cdot, \cdot)$ is the squared error loss (3). In this case, the impurity (9) becomes the within-node sample variance of the response variable, that is,

$$\mathcal{I}(\mathrm{t}) = \frac{1}{N_{\mathrm{t}}} \sum_{\mathbf{x}_i \in \mathrm{t}} (y_i - \overline{y}_{\mathrm{t}})^2. \tag{13}$$

According to these model specifications, the tree output (12) makes a prediction by returning the within-node sample mean of the response variable, that is,

$$\widehat{g}(T^{\mathrm{CART}})(\mathbf{x}) = h(\overline{y}_{\mathrm{t}}) = \overline{y}_{\mathrm{t}}, \quad \mathbf{x} \in \mathrm{t}.$$

3.3. C4.5 Algorithm

Recall the classification setting in Section 2.2, where $h(\cdot)$ is the logit link and $\mathcal{L}(\cdot, \cdot)$ is the logistic loss (5). In this case, the impurity (9) becomes the binary entropy of the within-node empirical class probability, that is,

$$\begin{aligned} \mathcal{I}(\mathrm{t}) &= \eta_{\mathrm{t}} \log(1/\eta_{\mathrm{t}}) + (1 - \eta_{\mathrm{t}}) \log(1/(1 - \eta_{\mathrm{t}})), \\ \eta_{\mathrm{t}} &:= \frac{1}{N_{\mathrm{t}}} \sum_{\mathbf{x}_i \in \mathrm{t}} \mathbf{1}(y_i = 1). \end{aligned}$$

As entropy quantifies the information content of a random variable, the impurity gain (10) is sometimes called the *information gain* or the *mutual information*. According to these model specifications, the tree output (12) is the log-odds of the within-node empirical class probability

$$\widehat{g}(T^{\mathrm{C4.5}})(\mathbf{x}) = h(\overline{y}_{\mathrm{t}}) = \log\left(\frac{\eta_{\mathrm{t}}}{1 - \eta_{\mathrm{t}}}\right), \quad \mathbf{x} \in \mathrm{t}.$$

The tree makes a class prediction by returning the majority vote of the classes in the node, that is,

$$\widehat{c}(T^{\mathrm{C4.5}})(\mathbf{x}) = \begin{cases} +1 & \text{if } \widehat{g}(T^{\mathrm{C4.5}})(\mathbf{x}) \geq 0 \\ -1 & \text{if } \widehat{g}(T^{\mathrm{C4.5}})(\mathbf{x}) < 0 \end{cases}.$$

Definition 3.2. A decision tree T constructed with CART methodology (Section 3.2) in the regression setting (Section 2.1) is denoted by T^{CART}. Similarly, a decision tree T constructed with C4.5 methodology (Section 3.3) in the classification setting (Section 2.2) is denoted by $T^{\mathrm{C4.5}}$. An arbitrary unnamed decision tree T refers to either T^{CART} or $T^{\mathrm{C4.5}}$.

Remark 1. The curious reader may wonder why we do not analyze classification trees constructed with CART methodology—which use the so-called *Gini* splitting criterion—and instead focus on C4.5 methodology. Gini impurity for classification trees is equivalent to squared error impurity (13) for regression trees (Louppe 2014, sec. 3), and thus both types of trees produce identical estimates of the conditional mean response, namely, $\overline{y}_{\mathrm{t}} \approx \mathbb{E}(y|\mathbf{x})$ for $\mathbf{x} \in \mathrm{t}$. However, our forthcoming results for regression trees are relegated to additive $\mathbb{E}(y|\mathbf{x})$, which are appropriate for regression, but awkward for classification. For this reason, C4.5 methodology allows us to work with more common models for discrete responses $y \in \{-1, 1\}$, such as the logistic regression model in Section 2.2.

4. Main Results

In this section, we first describe a class of large scale predictive models. For this class of models, we establish an adaptive prediction risk bound, which in turn, leads to consistency of CART and C4.5.

4.1. Large Scale Predictive Models

We illustrate the high dimensional properties of CART and C4.5 in the context of *generalized additive models*, whereby an *additive function* is related to the conditional mean of the response variable by a link function (1). While there are many link functions that could be used, the canonical choices for regression and classification tasks are the aforementioned identity and logit functions from Sections 2.1 and 2.2, respectively.

In particular, note that additive logistic regression models are, importantly, different from ordinary additive regression in the sense that $\mathbb{E}(y|\mathbf{x}) = 2\mathbb{P}(y = 1|\mathbf{x}) - 1$ is not equal to an additive function of the predictor variables. This means that one cannot deduce consistency from existing results for regression trees (Scornet, Biau, and Vert 2015), even in the fixed dimensional setting, since they are limited to additive $\mathbb{E}(y|\mathbf{x})$.

In practice, additive logistic regression models are typically fit with backfitting or boosting algorithms (Hastie and Tibshirani 1990; Tutz and Binder 2006). Despite a rich literature on theoretical guarantees in the fixed dimensional setting (Horowitz and Mammen 2004), consistency results in the (ultra) high dimensional setting (i.e., $\log(p) \asymp N^{1-\xi}$, $\xi \in (0, 1)$) with either ℓ_0 or ℓ_1 sparsity constraints do not appear to be available, unless the logistic regression model is linear and the sparsity pattern is the number of relevant predictor variables (Abramovich and Grinshtein 2019). Therefore, sparse logistic regression models render a situation in which decision trees are unrivaled as a scalable, theoretically grounded method.

Generalized additive models. We now describe the generalized additive modeling framework with additional precision. Consider the additive function class

$$\mathcal{G}^1 := \left\{g(\mathbf{x}) := g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p)\right\},$$

where $g_1(x_1), g_2(x_2), \ldots, g_p(x_p)$ is a collection of p univariate (Borel measurable) functions. The generalized additive modeling framework involves finding a $g(\cdot) \in \mathcal{G}^1$ for which

$$g(\mathbf{x}) = g_1(x_1) + g_2(x_2) + \cdots + g_p(x_p) \qquad (14)$$

approximates the true model (1).

Generalized additive models are often used in high dimensional settings, in part because notions of exact ℓ_0 or approximate ℓ_1 sparsity are easy to define and interpret. We now describe these two types of sparsity patterns in detail.

Approximate ℓ_1 sparsity. As we have already mentioned, we would like to consider models with *approximate* sparsity. To this end, for $g(\cdot) \in \mathcal{G}^1$, we define the *total variation ℓ_1 norm* $\|g\|_{\mathrm{TV}}$ as the infimum of

$$\mathrm{TV}(g_1) + \mathrm{TV}(g_2) + \cdots + \mathrm{TV}(g_p) \qquad (15)$$

over all representations of $g(\cdot)$ as (14), that is, $\|g\|_{\mathrm{TV}}$ is the aggregated total variation of the individual component functions (see Tan and Zhang (2019) and the references therein). To simplify the arguments, we henceforth assume $g(\cdot)$ has a canonical representation (14) such that (15) achieves this infimum. One can think of $\|g\|_{\mathrm{TV}}$ as a measure of the capacity of $g(\cdot)$ and, as we shall see, it will play a central role in the paper. The total variation ℓ_1 norm is a desirable quantification of sparsity because it allows for some predictor variables to make very small yet meaningful contributions to the model.

In the case that all the component functions $g_j(\cdot)$ are smooth over a domain $\mathcal{X}$ with Lebesgue measure one, 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 $\mathbb{R}^p$. In particular, if $g(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x}$ is linear over $\mathcal{X}^p$, then $\|g\|_{\mathrm{TV}} = \|\boldsymbol{\beta}\|_{\ell_1}$, the ℓ_1 norm of the coefficient vector $\boldsymbol{\beta} \in \mathbb{R}^p$.

Exact ℓ_0 sparsity. To account for *exact* sparsity, we also define the ℓ_0 norm $\|g\|_{\ell_0}$ of $g(\cdot) \in \mathcal{G}^1$ as the infimum of

$$\#\{j : g_j(\cdot) \text{ is nonconstant}\}$$

over all representations of $g(\cdot)$ as (14). In other words, the ℓ_0 norm counts the number of relevant variables that affect $g(\cdot)$. For $g(\cdot) \in \mathcal{G}^1$, we have the relation

$$\|g\|_{\mathrm{TV}} \le \|g\|_{\ell_0} \cdot \max_j \mathrm{TV}(g_j),$$

provided $\max_j \mathrm{TV}(g_j) < \infty$. Thus, a small total variation ℓ_1 norm captures *either* exact or approximate sparsity, whichever is present.

Finally, our results will require us to have uniform control on the magnitude of a function $g : \mathbb{R}^p \to \mathbb{R}$, which we do through the supremum norm $\|g\|_\infty := \sup_{\mathbf{x}} |g(\mathbf{x})|$.

4.2. Empirical Risk Bound

The empirical risk bound in this section is the key to all forthcoming results. We prove a purely algorithmic guarantee, namely, that the (excess) empirical risk of a depth K regression tree constructed with CART or C4.5 methodology is of order $1/K$. To the best of our knowledge, this result is the first of its kind for any decision tree algorithm. The math behind it is surprisingly simple; 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., within-node) 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). In contrast with the recent work of Chi et al. (2020) and 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.

We now describe the empirical risk bound in detail. We aim to upper bound the excess empirical risk of the decision tree T_K. In view of (11), as we grow deeper trees, the empirical risk reduces by the impurity gain, as can be seen from the recursion (which holds generically)

$$\widehat{\mathcal{R}}(\widehat{g}(T_K)) = \widehat{\mathcal{R}}(\widehat{g}(T_{K-1})) - \sum_{\mathrm{t}\in T_{K-1}} \frac{N_{\mathrm{t}}}{N}\mathcal{IG}(\mathrm{t}), \tag{16}$$

where the notation "$\mathrm{t} \in T$" means that t is a terminal node of a tree T. In view of (16), to obtain an inductive upper bound on $\widehat{\mathcal{R}}(\widehat{g}(T_K))$, we further aim to the lower bound the impurity gain $\mathcal{IG}(\mathrm{t})$ for $\mathrm{t} \in T_{K-1}$ in terms of the within-node excess risk $\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-1})) - \widehat{\mathcal{R}}_{\mathrm{t}}(g)$ for a candidate model $g(\cdot)$. Lemma 4.1 accomplishes this goal.

Lemma 4.1 (Impurity gain for CART and C4.5). Let $g(\cdot) \in \mathcal{G}^1$ be any additive function and $K \geq 1$ be any depth. Then for any terminal node t of the tree T_{K-1} such that $\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-1})) > \widehat{\mathcal{R}}_{\mathrm{t}}(g)$, we have

$$\mathcal{IG}(\mathrm{t}) \geq \frac{(\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-1})) - \widehat{\mathcal{R}}_{\mathrm{t}}(g))^2}{V^2(g)},$$

where $V(g) = \|g\|_{\mathrm{TV}}$ for CART and $V(g) = \|g\|_{\mathrm{TV}} + \|g\|_\infty + 3$ for C4.5.

Plugging Lemma 4.1 into (16) and subtracting $\widehat{\mathcal{R}}(g)$ from both sides, we see that

$$\mathcal{E}_K \leq \mathcal{E}_{K-1} - \frac{1}{V^2(g)} \sum_{\mathrm{t}\in T_{K-1}:\mathcal{E}_{K-1}(\mathrm{t})>0} \frac{N_{\mathrm{t}}}{N}\mathcal{E}^2_{K-1}(\mathrm{t}), \tag{17}$$

where

$$\mathcal{E}_K := \widehat{\mathcal{R}}(\widehat{g}(T_K)) - \widehat{\mathcal{R}}(g), \quad \mathcal{E}_K(\mathrm{t}) := \widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_K)) - \widehat{\mathcal{R}}_{\mathrm{t}}(g)$$

are the global and within-node excess empirical risks, respectively. Next, using Jensen's inequality on (17) and the fact that $\mathcal{E}_{K-1} = \sum_{\mathrm{t}\in T_{K-1}} \frac{N_{\mathrm{t}}}{N}\mathcal{E}_{K-1}(\mathrm{t})$, it can be shown that (see Lemma D.1 in supplementary materials D)

$$\mathcal{E}_K \leq \mathcal{E}_{K-1}\Big(1 - \frac{\mathcal{E}_{K-1}}{V^2(g)}\Big), \quad \mathcal{E}_{K-1} \geq 0, \quad K \geq 1. \tag{18}$$

Iterating the recursion (18), we establish the following upper bound on the excess empirical risk $\mathcal{E}_K$.

Theorem 4.2 (Empirical risk bound for CART and C4.5). Let T_K be a depth $K \geq 1$ decision tree. Then we have

$$\widehat{\mathcal{R}}(\widehat{g}(T_K)) \leq \inf_{g(\cdot)\in\mathcal{G}^1} \Big\{\widehat{\mathcal{R}}(g) + \frac{V^2(g)}{K+3}\Big\},$$

where $V(g)$ is the constant specified in Lemma 4.1.

The above theorem says that a decision tree of depth K minimizes the empirical risk (for squared error loss and logistic loss) over all additive functions, up to a slackness term of order $1/K$.

4.3. Oracle Inequality

Our main theorem establishes an adaptive prediction risk bound (also known as an *oracle inequality*) for decision trees under model mis-specification; that is, when the true model (1) may not belong to $\mathcal{G}^1$. Essentially, the result says that CART and C4.5 adapt to the class of (ordinary or logistic) additive regression models, performing as if they were finding the best additive approximation of the true model (1), while accounting for the capacity (the total variation ℓ_1 norm $\|\cdot\|_{\mathrm{TV}}$) of the approximation.

In the regression setting, for simplicity and ease of exposition, we assume that the error $\varepsilon = y - g^*(\mathbf{x}) = y - \mathbb{E}(y|\mathbf{x})$ is sub-Gaussian, that is, there exists $\sigma^2 > 0$ such that for all $u \geq 0$,

$$\mathbb{P}(|\varepsilon| \geq u) \leq 2\exp(-u^2/(2\sigma^2)). \tag{19}$$

Before we state our main theorem, we remind the reader that $\|\cdot\|$ denotes the $\mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$ norm (4) and $\mathrm{Err}(\cdot)$ denotes the misclassification risk (6). Using these risk measures, we evaluate the performance of CART for the regression model in Section 2.1 and the performance of C4.5 for the logistic regression model in Section 2.2.

Theorem 4.3 (Oracle inequalities for CART and C4.5). Let $K \geq 1$ be any depth. Granting the noise condition (19), we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}(\|\widehat{g}(T_K^{\mathrm{CART}}) - g^*\|^2) \\ &\quad \leq 2 \inf_{g(\cdot)\in\mathcal{G}^1} \Big\{\|g - g^*\|^2 + \frac{\|g\|^2_{\mathrm{TV}}}{K+3} + C_1\frac{2^K\log^2(N)\log(Np)}{N}\Big\},\end{aligned} \tag{20}$$

where C_1 is a positive constant that depends only on $\|g^*\|_\infty$ and σ^2. Furthermore, we have

$$\begin{aligned}&\mathbb{E}_{\mathcal{D}}(\mathrm{Err}(\widehat{g}(T_K^{\mathrm{C4.5}}))) - \mathrm{Err}(g^*) \\ &\quad \leq \inf\nolimits_{g(\cdot)\in\mathcal{G}^1} \Big\{\|g - g^*\| + 2\tfrac{\|g\|_{\mathrm{TV}}+\|g\|_\infty+3}{\sqrt{K+3}} + \\ &\qquad C_2\Big(\tfrac{2^K\log^2(N)\log(Np)}{N}\Big)^{1/4}\Big\},\end{aligned} \tag{21}$$

where C_2 is a positive universal constant.

Remark 2. Throughout this article, we measure the accuracy of a predictor via the *expected* prediction risk over the data $\mathcal{D}$, as in Theorem 4.3. However, at the expense of more complicated expressions, one can also obtain statements that hold with high probability.

Theorem 4.3 reveals the tradeoff between the goodness of fit and complexity relative to sample size. The goodness of fit terms involving $\|g-g^*\|$, $\|g\|_{\mathrm{TV}}$, and $\|g\|_\infty$ stem from the excess empirical risk bound in Theorem 4.2, and the descriptive complexity term $2^K\log(Np)$ comes from the fact that the empirical ϵ-metric entropy for depth K decision trees with p predictor variables is of order $2^K\log(Np/\epsilon)$.

4.4. Consistency

We now explore the case where the true model $g^*(\cdot)$ has the freedom to change with the sample size, which is common in other literature on high dimensional consistency (Bühlmann

2006). To this end, Theorem 4.3 immediately implies consistency when the model is well-specified (i.e., $g^*(\cdot) \in \mathcal{G}^1$) and has a controlled sparsity pattern. More specifically, choosing $g(\cdot) = g^*(\cdot)$ in Theorem 4.3 and stipulating that the total variation ℓ_1 norm of $g^*(\cdot)$ does not grow too fast, we find that CART and C4.5 are consistent, even when the dimensionality grows sub-exponentially with the sample size. We note that this type of result is impossible with nonadaptive procedures that do not automatically adjust the amount of smoothing along a particular dimension according to how much the predictor variable affects the response variable. Such procedures perform local estimation at a query point using data that are close in every single dimension, making them prone to the curse of dimensionality even if the true model is sparse. This is the case with conventional multivariate (Nadaraya-Watson or local polynomial) kernel regression in which the bandwidth is the same for all directions, or k-nearest neighbors with Euclidean distance. Indeed, one can compute asymptotic expansions of their bias and variance (Mack 1981; Ruppert and Wand 1994), which evidently do not exploit low dimensional structure in the regression model.

Corollary 4.4 (Consistency of CART and C4.5). Consider a sequence of prediction problems (1) with true models $\{g_N^*(\cdot)\}_{N=1}^{\infty}$. Assume that $g_N^*(\mathbf{x}) = \sum_{j=1}^{p_N} g_j(x_j) \in \mathcal{G}^1$ and $\sup_N \|g_N^*\|_\infty < \infty$. Suppose that $K_N \to \infty$, $\|g_N^*\|_{\mathrm{TV}} = o(\sqrt{K_N})$, and $\frac{2^{K_N}\log^2(N)\log(Np_N)}{N} \to 0$ as $N \to \infty$. Granting the noise condition (19), regression trees are consistent, that is,

$$\lim_{N\to\infty} \mathbb{E}_{\mathcal{D}}(\|\widehat{g}(T_{K_N}^{\mathrm{CART}}) - g_N^*\|^2) = 0.$$

Furthermore, classification trees are consistent, that is,

$$\lim_{N\to\infty} (\mathbb{E}_{\mathcal{D}}(\mathrm{Err}(\widehat{g}(T_{K_N}^{\mathrm{C4.5}}))) - \mathrm{Err}(g_N^*)) = 0.$$

Remark 3. Note that because $\|g_N^*\|_{\mathrm{TV}} \leq \|g_N^*\|_{\ell_0} \cdot \max_{j\leq p_N} \mathrm{TV}(g_j^*)$, the consistency statement in Corollary 4.4 also applies to models with sparsity patterns defined by the number of relevant variables. Thus, the condition $\|g_N^*\|_{\mathrm{TV}} = o(\sqrt{K_N})$ can be replaced with $\|g_N^*\|_{\ell_0} = o(\sqrt{K_N})$, provided $\max_{j\leq p_N} \mathrm{TV}(g_j^*)$ is independent of N. The same reasoning applies to all forthcoming results that use $\|\cdot\|_{\mathrm{TV}}$ to measure sparsity.

4.4.1. Consistency Rates

We now describe the effect of specific choices of the depth and regimes of the ambient dimension on the consistency rate for CART and C4.5. The hypotheses of Corollary 4.4 are satisfied if, for example, $K_N = \lfloor(\xi/2)\log_2(N)\rfloor$, $\log(p_N) \asymp N^{1-\xi}$ for $\xi \in (0,1)$, $\sup_N \|g_N^*\|_\infty < \infty$, and $\sup_N \|g_N^*\|_{\mathrm{TV}} < \infty$. In this case, from (20) in Theorem 4.3, the consistency rate of the CART algorithm is

$$\frac{4\sup_N \|g_N^*\|_{\mathrm{TV}}^2}{\xi\log_2(N)+6} + 2C_1\frac{\log^3(N)}{N^{1-\xi/2}} + 2C_1\frac{\log^2(N)}{N^{\xi/2}} \\ = \mathcal{O}(1/\log(N)) \tag{22}$$

and from (21) in Theorem 4.3, the consistency rate of the C4.5 algorithm is

$$4\frac{\sup_N \|g_N^*\|_{\mathrm{TV}} + \sup_N \|g_N^*\|_\infty + 3}{\sqrt{2\xi\log_2(N)+12}} \tag{23} \\ + C_2\left(\frac{\log^3(N)}{N^{1-\xi/2}} + \frac{\log^2(N)}{N^{\xi/2}}\right)^{1/4} = \mathcal{O}(1/\sqrt{\log(N)}).$$

The dependence on the total variation ℓ_1 norm $\|g_N^*\|_{\mathrm{TV}}$ in the consistency rates (22) and (23) shows that CART and C4.5 can tolerate an approximate sparsity level that grows as fast as $o(\sqrt{\log(N)})$. As per Remark 3, a similar growth is tolerated for the ℓ_0 norm $\|g_N^*\|_{\ell_0}$.

4.4.2. Consistency for Unbounded Variation Component Functions

Corollary 4.4 implicitly considers (ordinary or logistic) additive regression models whose component functions have bounded variation, per the finiteness of $\|g_N^*\|_{\mathrm{TV}}$. In fact, consistency holds when the component functions $g_j(\cdot)$ are merely Borel measurable, as Corollary 4.5 reveals.

Corollary 4.5 (Consistency of CART and C4.5 for unbounded variation components). Consider a sequence of prediction problems with true models $\{g_N^*(\cdot)\}_{N=1}^{\infty}$. Assume $g_N^*(\mathbf{x}) = \sum_{j=1}^{p_N} g_j(x_j) \in \mathcal{G}^1$, $\sup_N \|g_N^*\|_\infty < \infty$, and $\sup_N \|g_N^*\|_{\ell_0} < \infty$. Suppose that $K_N \to \infty$ and $\frac{2^{K_N}\log^2(N)\log(Np_N)}{N} \to 0$ as $N \to \infty$. Granting the noise condition (19), regression trees are consistent, that is,

$$\lim_{N\to\infty} \mathbb{E}_{\mathcal{D}}(\|\widehat{g}(T_{K_N}^{\mathrm{CART}}) - g_N^*\|^2) = 0.$$

Furthermore, classification trees are consistent, that is,

$$\lim_{N\to\infty} (\mathbb{E}_{\mathcal{D}}(\mathrm{Err}(\widehat{g}(T_{K_N}^{\mathrm{C4.5}}))) - \mathrm{Err}(g_N^*)) = 0.$$

While Corollary 4.5 allows the component functions of the model to have unbounded variation, it requires the number of relevant variables $\|g_N^*\|_{\ell_0}$ to be uniformly bounded in N, in contrast to the $o(\sqrt{K_N})$ growth of $\|g_N^*\|_{\mathrm{TV}}$ tolerated in Corollary 4.4.

Remark 4. Corollaries 4.4 and 4.5 do 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

$$\widehat{\mathcal{R}}(\widehat{g}(T)) + \lambda|T|$$

over all trees T that can be obtained from $T_{\max}$ by iteratively merging its internal nodes, where λ 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 (20), but with the advantage of having the infimum over both the depth $K \geq 1$ and additive functions $g(\cdot) \in \mathcal{G}^1$.

4.5. Related Consistency Results and Optimality

Here we compare our consistency results for CART and C4.5 to those of other prediction methods and the optimal minimax rates.

The reader might be somewhat surprised by the consistency statements in Corollaries 4.4 and 4.5, especially since they are qualitatively similar to existing performance guarantees for predictors based on very different principles, like boosting or neural networks. For example, (Bühlmann 2006, Theorem 1) states that boosting with linear learners is also consistent for a sequence of ℓ_1 constrained linear models $g^*_N(\mathbf{x}) = \boldsymbol{\beta}^T_N\mathbf{x}$ on $[0,1]^p$ in the high dimensional regime, that is, when $\log(p_N) \asymp N^{1-\xi}$ for $\xi \in (0,1)$ and $\sup_N \|g^*_N\|_{\mathrm{TV}} = \sup_N \|\boldsymbol{\beta}_N\|_{\ell_1} < \infty$.

To provide another frame of reference, we also compare the consistency rates of CART (22) and C4.5 (23) to the corresponding minimax rates for the model classes we consider. For regression, the optimal minimax rate (with respect to excess squared error risk) for s-sparse additive regression with continuously differentiable component functions is $\max\{(s/N)\log(p/s), sN^{-2/3}\}$ (Raskutti, Wainwright, and Yu 2012), while the optimal rate for additive regression models with bounded total variation ℓ_1 norm (when $p \gg N$) is $\sqrt{(\log(p))/N}$ (Tan and Zhang 2019). Both of these settings, corresponding to ℓ_0 and ℓ_1 sparsity, respectively, are covered by our theory for regression trees, and so our $1/\log(N)$ rate (22) is evidently sub-optimal. For classification, the literature is less developed, though there are results for linear logistic regression. For example, Abramovich and Grinshtein (2019) show that the optimal minimax rate (with respect to excess mis-classification risk) for s-sparse linear logistic regression is $\sqrt{(s/N)\log(p/s)}$, again much faster than our $1/\sqrt{\log(N)}$ rate (23).

The sub-optimality of our rates is due to a combination of two factors. First, the form of the decision tree predictions—averaging the response data in the terminal nodes—introduces an inductive bias, which, in the aforementioned s-sparse additive regression setting, leads to best-case (yet still sub-optimal) rates of order $N^{-2/(2+s)}$ (Tan, Agarwal, and Yu 2021). The second source of sub-optimality stems from our method of analysis. Recall from Lemma 4.1 that a key step in our proofs is to lower bound the impurity gain at each node by a constant multiple of the *squared* excess risk, viz., $\mathcal{IG}(\mathrm{t}) \gtrsim (\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-1})) - \widehat{\mathcal{R}}_{\mathrm{t}}(g^*))^2$ whenever $\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-1})) > \widehat{\mathcal{R}}_{\mathrm{t}}(g^*)$. Had we been able to show (or rather presumed in the form of an assumption) a lower bound $\mathcal{IG}(\mathrm{t}) \gtrsim \widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-1})) - \widehat{\mathcal{R}}_{\mathrm{t}}(g^*)$ whenever $\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-1})) > \widehat{\mathcal{R}}_{\mathrm{t}}(g^*)$, then it would have been possible for us to obtain polynomial convergence rates, instead of logarithmic. In the next section, we obtain a similar lower bound on the impurity gain (Lemma 5.1) for separable, large margin data, which does indeed lead to a faster consistency rate.

5. Beyond Discriminative Models

In Section 4, we operated under a discriminative statistical model of the data; that is, we study decision trees under an explicit form of the conditional distribution $\mathbb{P}_{y|\mathbf{x}}$. In lieu of the logistic regression model from Section 2.2, here we consider another ubiquitous classification setting in which the data can be perfectly separated into two classes by an additive decision boundary, with margin $\gamma > 0$. As we shall see, this setting will allow us to obtain consistency rates which are exponentially faster than those from Section 4.4. Separable data assumptions, such as the one formalized in Assumption 1, are prevalent in statistical learning literature, especially in the context of (kernel) support vector machines (Boser, Guyon, and Vapnik 1992) and boosting (Bartlett et al. 1998). Note that such an assumption is particularly appropriate for our high dimensional setting—when p is large, there is more freedom for the data to be separated by an additive decision boundary.

Assumption 1 (Additively separable, large margin). There exists $\gamma \in (0,1]$ and $f^*(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j) \in \mathcal{G}^1$ with $\max\{\|f^*\|_{\mathrm{TV}}, \|f^*\|_\infty\} \le 1$ such that for almost all pairs $(\mathbf{x}, y)$ drawn from the joint distribution $\mathbb{P}_{(\mathbf{x},y)}$,

$$yf^*(\mathbf{x}) \ge \gamma.$$

Remark 5. If the additive function $f^*(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x}$ is linear over $[0,1]^p$, then γ corresponds to the maximum ℓ_1-margin subject to $\|\boldsymbol{\beta}\|_{\ell_1} \le 1$, matching the standard margin framework for linearly separable data.

Note that any separable, large margin distributional assumption implies that the Bayes risk is zero; therefore, we aim to show that a tree constructed with C4.5 methodology will have mis-classification risk converging to zero. While there are some similarities with the proof of Corollary 4.4, the key difference is an improved lower bound on the impurity gain that establishes nearly linear, rather than quadratic (see Lemma 4.1), dependence on the within-node empirical risk.

Lemma 5.1 (Information gain for C4.5 with separable, large margin data). Grant Assumption 1. Let $K \ge 1$ be any depth and let t be any terminal node of $T^{\mathrm{C4.5}}_{K-1}$. We have that

$$\mathcal{IG}(\mathrm{t}) \ge \frac{\gamma^2}{30}\cdot\frac{\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T^{\mathrm{C4.5}}_{K-1}))}{\log(1/\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T^{\mathrm{C4.5}}_{K-1})))}.$$

Using Lemma 5.1 in conjunction with (16) shows that the empirical risk decays sub-exponentially fast in the depth K, much faster than the polynomial rate of decay for a logistic regression model (Theorem 4.2).

Theorem 5.2 (Empirical risk for C4.5 with separable, large margin data). Granting Assumption 1, for all depths $K \ge 1$,

$$\widehat{\mathcal{R}}(\widehat{g}(T^{\mathrm{C4.5}}_{K})) \le \exp\Big(-\Big(\frac{\gamma^2 K}{30}\Big)^{1/2}\Big).$$

Employing the above empirical risk bound, we can establish the following mis-classification risk bound.

Theorem 5.3 (Mis-classification risk for C4.5 with separable, large margin data). Granting Assumption 1, for all depths $K \ge 1$,

$$\mathbb{E}_{\mathcal{D}}(\mathrm{Err}(\widehat{g}(T^{\mathrm{C4.5}}_{K}))) \le 2\exp\Big(-\Big(\frac{\gamma^2 K}{120}\Big)^{1/2}\Big) + C_2\Big(\frac{2^K\log^2(N)\log(Np)}{N}\Big)^{1/4},$$

where C_2 is the same constant in the statement of Theorem 4.3.

Theorem 5.3 immediately implies the following consistency result.

Corollary 5.4 (Consistency of C4.5 with separable, large margin data). Consider a sequence of decision boundaries $\{f_N^*(\cdot)\}_{N=1}^{\infty}$ with respective margins $\{\gamma_N(\cdot)\}_{N=1}^{\infty}$ such that Assumption 1 holds. Suppose $f_N^*(\mathbf{x}) = \sum_{j=1}^{p_N} f_j(x_j) \in \mathcal{G}^1$ and that $K_N \to \infty$, $\gamma_N = \omega(1/\sqrt{K_N})$, and $\frac{2^{K_N}\log^2(N)\log(Np_N)}{N} \to 0$ as $N \to \infty$. Then, classification trees are consistent, that is,

$$\lim_{N\to\infty} \mathbb{E}_{\mathcal{D}}(\mathrm{Err}(\widehat{g}(T_{K_N}^{\mathrm{C4.5}}))) = 0.$$

The hypotheses of Corollary 5.4 are satisfied if, for example, $K_N = \lfloor(\xi/2)\log_2(N)\rfloor$, $\log(p_N) \asymp N^{1-\xi}$ for $\xi \in (0,1)$, and $\gamma_N = \gamma$ for some $\gamma \in (0,1]$. In this case, from Theorem 5.3, the consistency rate of C4.5 is

$$\mathcal{O}\big(\exp\big(-\sqrt{\gamma^2\xi\log(N)/240}\big)\big),$$

which is sub-polynomial in N, and exponentially faster than the $1/\sqrt{\log(N)}$ rate (23) for a logistic regression model.

6. Models with Interactions

Our main results in Section 4 focused on (ordinary or logistic) additive regression models primarily because notions of approximate and exact sparsity are easier to define and more interpretable in high dimensional settings. However, the interpretation of CART and C4.5 in an (ordinary or logistic) additive regression setting will be muddled by the fact that all interactions it finds will be spurious. It is therefore desirable to have a more comprehensive theory, particularly for data settings where decision trees could be useful.

As we now explain, it is possible to motivate reasonable assumptions so that our main results for (ordinary or logistic) additive regression models can be extended to models with interactions. To this end, recall the proof outline in Section 4. We started with the recursion (16) and then used a lower bound on the impurity gain (Lemma 4.1), tailored for (ordinary or logistic) additive regression models, to bound the empirical risk (Theorem 4.2) and ultimately show consistency (Corollary 4.4). In fact, under suitable assumptions, many of these same ideas work when model class $\mathcal{G}$ has interaction terms.

To illustrate the broad strokes in obtaining these extensions, we will consider the class of d-way interaction models

$$\mathcal{G}^d := \Big\{g(\mathbf{x}) := \sum_{j_1} g_{j_1}(x_{j_1}) + \sum_{j_1<j_2} g_{j_1,j_2}(x_{j_1},x_{j_2}) + \cdots + \sum_{j_1<j_2<\cdots<j_d} g_{j_1,j_2,\ldots,j_d}(x_1,x_2,\ldots,x_d)\Big\},$$

which encompasses the additive function class $\mathcal{G}^1$. Thus, models belonging to $\mathcal{G}^d$ have interactions involving up to d predictor variables. Conversely, any square-integrable function with interactions involving at most d variables admits a functional ANOVA decomposition in the form above, where the functional components in the expansion have zero mean and are orthogonal to each other (Hooker 2007). We note in passing that, while decision trees can discover interaction effects through the way they are constructed, they do not directly leverage additive structure in the model and so are unlikely to achieve the optimal rates of convergence on $\mathcal{G}^d$ (this is certainly the case for $\mathcal{G}^1$; see Section 4.5).

By recursing (16), we can write the empirical risk of the tree as

$$\widehat{\mathcal{R}}(\widehat{g}(T_K)) = \widehat{\mathcal{R}}(\widehat{g}(T_{K-d})) - \sum_{\mathrm{t}\in T_{K-d}} \tfrac{N_{\mathrm{t}}}{N}\mathcal{IG}_d(\mathrm{t}), \qquad (24)$$

where $\mathcal{IG}_d(\mathrm{t}) := \mathcal{IG}(\mathrm{t}) + \sum_{\mathrm{t}'}\frac{N_{\mathrm{t}'}}{N_{\mathrm{t}}}\mathcal{IG}(\mathrm{t}')$ is the *d-level impurity gain* of a node $\mathrm{t} \in T_{K-d}$. Here in the definition of $\mathcal{IG}_d(\mathrm{t})$, the sum extends over all descendent nodes t' of t up to depth $K-1$. Another way of thinking about $\mathcal{IG}_d(\mathrm{t})$ is that it measures the decrease in risk from greedily splitting d times in t and thus captures interactions involving up to d predictor variables. For example, according to the representations given by (10) and (11),

$$\mathcal{IG}(\mathrm{t}) = \widehat{\mathcal{R}}_{\mathrm{t}}(h(\overline{y}_{\mathrm{t}})) - \min_{\beta_0,\beta_1}\widehat{\mathcal{R}}_{\mathrm{t}}(\beta_0 + \beta_1\mathbf{1}(x_{j_{\mathrm{t}}} > s_{\mathrm{t}})),$$

which captures only main effects from splitting *once* in t. This explains why Lemma 4.1 relates the impurity gain to the empirical risk of an (ordinary or logistic) additive regression model. On the other hand, $\mathcal{IG}_2(\mathrm{t}) = \mathcal{IG}(\mathrm{t}) + P_{\mathrm{t}_L}\mathcal{IG}(\mathrm{t}_L) + P_{\mathrm{t}_R}\mathcal{IG}(\mathrm{t}_R)$ equals

$$\widehat{\mathcal{R}}_{\mathrm{t}}(h(\overline{y}_{\mathrm{t}})) - \min_{\beta_0,\beta_1,\beta_2,\beta_3,j_1,s_1,j_2,s_2}\widehat{\mathcal{R}}_{\mathrm{t}}(\beta_0 + \beta_1\mathbf{1}(x_{j_{\mathrm{t}}} > s_{\mathrm{t}}) + \beta_2\mathbf{1}(x_{j_1} > s_1,\ x_{j_{\mathrm{t}}} \le s_{\mathrm{t}}) + \beta_3\mathbf{1}(x_{j_2} > s_2,\ x_{j_{\mathrm{t}}} > s_{\mathrm{t}})),$$

and thus captures both main effects and second order effects from greedily splitting *twice* in t. It is then reasonable to postulate that one could relate $\mathcal{IG}_2(\mathrm{t})$ to the empirical risk of a function in $\mathcal{G}^2$. Consequently, a natural generalization of the impurity gain inequality in Lemma 4.1 to functions in $\mathcal{G}^d$ would be the following condition.

Assumption 2. Let $g^*(\cdot) \in \mathcal{G}^d$ and $K \ge d$. For any terminal node t of the tree T_{K-d} such that $\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-d})) > \widehat{\mathcal{R}}_{\mathrm{t}}(g^*)$, we have

$$\mathcal{IG}_d(\mathrm{t}) \ge \frac{(\widehat{\mathcal{R}}_{\mathrm{t}}(\widehat{g}(T_{K-d})) - \widehat{\mathcal{R}}_{\mathrm{t}}(g^*))^2}{V^2(g^*)}, \qquad (25)$$

for some complexity constant $V(g^*)$ that depends only on $g^*(\cdot)$.

Following a similar argument to the one outlined in Section 4, we can substitute the purported lower bound (25) into (24) and use Jensen's inequality to produce the recursion

$$\mathcal{E}_K \le \mathcal{E}_{K-d}\Big(1 - \frac{\mathcal{E}_{K-d}}{V^2(g^*)}\Big), \quad \mathcal{E}_{K-d} \ge 0, \quad K \ge d.$$

Thus, granting the impurity gain condition (25), by Lemma D.1 in supplementary materials D, we have that

$$\widehat{\mathcal{R}}(\widehat{g}(T_K)) \le \widehat{\mathcal{R}}(g^*) + \frac{V^2(g^*)d}{K + 2d + 1}, \quad K \ge d, \qquad (26)$$

a direct analogue to Theorem 4.2. Thus, the excess empirical risk for a d-way interaction model is of order d/K, which means that the depth K should exceed d for it to be small. This is to be expected since the depth K controls the interaction order of the tree. Using (26) and the same steps as the proof of Theorem 4.3

and Corollary 4.4, we can easily deduce that if $\{g^*_N(\cdot)\}_{N=1}^{\infty}$ is a sequence of true models in $\mathcal{G}^d$ with $\sup_N \|g^*_N\|_\infty < \infty$ and $K_N \to \infty$, $V(g^*_N) = o(\sqrt{K_N})$, and $\frac{2^{K_N}\log^2(N)\log(Np_N)}{N} \to 0$ as $N \to \infty$, then both regression trees and classification trees are consistent, that is,

$$\mathbb{E}_{\mathcal{D}}(\|\widehat{g}(T^{\mathrm{CART}}_{K_N}) - g^*_N\|^2) \to 0 \quad \text{and}$$
$$\mathbb{E}_{\mathcal{D}}(\mathrm{Err}(\widehat{g}(T^{\mathrm{C4.5}}_{K_N}))) - \mathrm{Err}(g^*_N) \to 0, \quad \text{as} \quad N \to \infty.$$

7. 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, Matheson, and De Freitas 2014, Propositions 3 and 4; Breiman 2001, sec. 11, or Breiman 1996, sec. 4.1). Indeed, our main results for individual trees in Section 4 also carry over to Breiman's random forests (Breiman 2001) with relative ease, as we now explain. To keep redundancy to a minimum, we will restrict ourselves to the regression setting of Section 2.1.

7.1. Growing the Forest

Consider a sub-sample $\mathcal{D}'$ of size a_N from the original dataset $\mathcal{D}$, whereby each sample point is drawn uniformly at random without replacement.[2] From this sub-sample, we train a depth K tree T^{CART}_K with CART methodology in the usual way, except that, at each internal node, we select m (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^{CART}_K, we generate a random subset $\mathcal{S} \subset \{1, 2, \dots, p\}$ of size m and split along a variable $x_{\widehat{j}_{\mathrm{t}}}$ with split point $\widehat{s}_{\mathrm{t}}$, where $(\widehat{j}_{\mathrm{t}}, \widehat{s}_{\mathrm{t}}) \in \arg\max_{(j \in \mathcal{S},\, s \in \mathbb{R})} \mathcal{IG}(j, s, \mathrm{t})$.

We grow B of these depth K regression trees separately using, respectively, B independent realizations $\boldsymbol{\Theta} = (\Theta_1, \Theta_2, \dots, \Theta_B)^T$ of a random variable Θ. Here Θ is distributed according to the law that generates the sub-sampled training data $\mathcal{D}'$ and candidate variables $\mathcal{S}$ for splitting at each of the nodes. The output of the bth regression tree is denoted by $\widehat{g}(T^{\mathrm{CART}}_K(\Theta_b))$.

With this notation in place, the random forest output is then simply the empirical average of the B regression tree outputs, namely,

$$\widehat{g}(\boldsymbol{\Theta})(\mathbf{x}) := B^{-1}\sum_{b=1}^{B}\widehat{g}(T^{\mathrm{CART}}_K(\Theta_b))(\mathbf{x}).$$

7.2. Oracle Inequality for Random Forests

By a modification of the proofs of Theorems 4.2 and 4.3, it is possible to show the following oracle inequality for random forests.

Theorem 7.1 (Oracle inequality for CART random forests). Granting the noise condition (19), for all depths $K \geq 1$, we have

$$\mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}}(\|\widehat{g}(\boldsymbol{\Theta}) - g^*\|^2)$$
$$\leq 2 \inf_{g(\cdot)\in\mathcal{G}^1}\left\{\|g - g^*\|^2 + \frac{p}{m}\frac{\|g\|^2_{\mathrm{TV}}}{K+3} + C_1\frac{2^K\log^2(a_N)\log(a_N p)}{a_N}\right\},$$

where C_1 is the same constant in the statement of Theorem 4.3.

Remark 6. As the number of trees B in the forest approaches infinity, by (Breiman 2001, Theorem 11.1),

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

almost surely.

Thus, when B is large, Theorem 7.1 resembles results such as (20) in Theorem 4.3 which measure the accuracy of a predictor via the *expected* prediction risk over the data $\mathcal{D}$.

To the best of our knowledge, Theorem 7.1 is one of the first results in the literature that shows explicitly the impact on the prediction risk 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 7.1, empirically, the random variable selection mechanism of forests has the effect of decorrelating 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 7.1 does reveal, however, is that this mechanism cannot hurt the prediction risk beyond a benign factor of p/m. In fact, standard implementations of regression forests use a default value of m equal to $\lfloor p/3 \rfloor$. With this choice, we see by comparing Theorem 7.1 and Corollary 4.4 that there is essentially no loss in performance (at most a factor of $p/m = 3$) over individual trees, despite not optimizing over the full set of variables at the internal nodes. It is also interesting to note that we recover the bound (20) for individual regression trees when $m = p$.

7.3. Consistency of Random Forests

Theorem 7.1 immediately implies a consistency result for regression forests, analogous to Corollary 4.4 for individual regression trees.

Corollary 7.2 (Consistency of random forests). Consider a sequence of prediction problems (1) with true models $\{g^*_N(\cdot)\}_{N=1}^{\infty}$. Assume that $g^*_N(\mathbf{x}) = \sum_{j=1}^{p_N} g_j(x_j) \in \mathcal{G}^1$ and $\sup_N \|g^*_N\|_\infty < \infty$. Suppose that $K_N \to \infty$, $\|g^*_N\|_{\mathrm{TV}} = o(\sqrt{(m_N/p_N)K_N})$, and $\frac{2^{K_N}\log^2(a_N)\log(a_N p_N)}{a_N} \to 0$ as $N \to \infty$. Then, granting the noise condition (19), regression forests are consistent, that is,

$$\lim_{N\to\infty} \mathbb{E}_{\boldsymbol{\Theta},\mathcal{D}}(\|\widehat{g}(\boldsymbol{\Theta}) - g^*\|^2) = 0.$$

Like the consistency statement for CART in Corollary 4.4, the hypotheses of Corollary 7.2 are satisfied if, for example, $K_N =$

[2] We deviate slightly from Breiman's original random forests (Breiman 2001), as we do not grow the constituent trees to maximum depth with bootstrapped data. Note, however, that sampling with and without replacement produce similar results when $a_N = \lfloor N/2 \rfloor$ (Friedman and Hall 2007).

$\lfloor(\xi/2)\log_2(a_N)\rfloor$, $\log(p_N) \asymp a_N^{1-\xi}$, and $m_N = \lfloor p_N/3\rfloor$, for some constant $\xi \in (0,1)$, yielding the same $\mathcal{O}(1/\log(N))$ rate as a single tree (see (22)). 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, that is, $m_N = o(p_N)$. At the extreme end, consistency holds even when $m_N \equiv 1$; that is, only a single coordinate is selected at random at each node, provided appropriate restrictions are placed on p_N and $\|g_N^*\|_{\mathrm{TV}}$.

Corollary 7.2 provides a partial answer to a problem posed by Scornet, Biau, and Vert (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 7.2 strengthens (Scornet, Biau, and Vert 2015, Theorem 1), which shows that random forests are consistent for additive regression when p is fixed, the component functions are continuous, and $K_N \to \infty$ and $\frac{2^{K_N}(\log^9(a_N))}{a_N} \to 0$ as $N \to \infty$. In contrast, here we allow the dimensionality p_N to grow sub-exponentially with the sample size N (under ℓ_1 sparsity constraints) and also for the component functions to be possibly discontinuous. In fact, the proof of Corollary 4.5 can be modified to establish consistency of random forests for additive regression with growing dimensionality p_N (under ℓ_0 sparsity constraints) and Borel measurable component functions—without assuming continuity or finite total variation. More specifically, suppose that $\sup_N \|g_N^*\|_\infty < \infty$, $\sup_N \|g_N^*\|_{\ell_0} < \infty$, and $m_N \asymp p_N$. If $K_N \to \infty$ and $\frac{2^{K_N}\log^2(a_N)\log(a_N p_N)}{a_N} \to 0$ as $N \to \infty$, then granting the noise condition (19), we have

$$\lim_{N\to\infty} \mathbb{E}_{\mathbf{\Theta},\mathcal{D}}(\|\widehat{g}(\mathbf{\Theta}) - g_N^*\|^2) = 0.$$

We close this section by saying that it is still largely a mystery (at least theoretically) why bagging and the random feature selection mechanism are so effective at reducing the prediction risk. Our bounds in Theorem 7.1 only show that these apparatuses do not degrade the performance beyond small factors. Certainly more work needs to be done to answer these questions.

7.4. Empirical Studies

Because random forests are now a classic topic in machine learning and data science, they have been subject to thorough empirical scrutiny and investigation under a variety of model specifications. This includes the high dimensional regime when $p \gg N$. As the literature is vast, we only mention a few experimental papers below.

Regarding the m parameter, Genuer, Poggi, and Tuleau (2008) provide a detailed empirical analysis of random forests in the high dimensional setting. They consider both synthetic and real-world data for regression and classification tasks. Among the simulated examples for regression are Friedman's benchmark models (Breiman 1996), which, in our notation, belong to the classes $\mathcal{G}^2$ and $\mathcal{G}^4$ and involve either 4 or 5 relevant predictor variables. For a sample size of $N = 100$, the simulation results (see Figures 3–5 in Genuer, Poggi, and Tuleau 2008) reveal that forests are quite robust to the inclusion of noisy predictor variables (with ambient dimensionality ranging from $p = 100$ to $p = 1000$). The author finds that forests achieve best performance when m is set to be equal to the ambient dimension p, corresponding to bagging. Similar observations are made in (Segal 2004, Table I) for one of Friedman's models with $N = 200$ and $p = 510$.

An investigation of the tree depth in random forests can be found in Zhou and Mentch (2021, Figures 10 and 11), where they sample $N = 50$ and $N = 100$ data points from a sparse linear model with 5 and 10 relevant variables, and ambient dimensionality $p = 1000$. Shallow trees are found to be advantageous when the model has a low signal-to-noise ratio. In Duroux and Scornet (2018, Figures A.1 and A.3) reveal that forests with smaller trees achieve similar performance to forests with fully grown trees if the number of terminal nodes is properly tuned. One of their examples (Model 8) consists of a sparse additive model with 4 relevant variables, $p = 1000$, and $N = 500$.

Other references on the impact of hyperparameter tuning in random forests can be found in the review article (Probst, Wright, and Boulesteix 2019).

It is still largely a mystery (at least theoretically) why bagging and the random feature selection mechanism are so effective at reducing the prediction risk. Our bounds in Theorem 7.1 only show that these apparatuses do not degrade the performance beyond small factors. Fascinating recent work by Mentch and Zhou (2020) shows that m 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 m/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.

8. Conclusion

In this article, we showed that decision trees and random forests adapt to ℓ_0 or ℓ_1 forms of sparsity and can accommodate (essentially) arbitrary types of predictor variables, such as continuous, discrete, and/or dependent. Our work is primarily of theoretical value, since we study existing decision tree methodology, namely, CART and C4.5. Nevertheless, given their widespread popularity in many applied disciplines, we believe our results can be used as a theoretical justification for practical application on high dimensional regression and classification problems. Specifically, Theorems 4.3 and 7.1 give an explicit characterization of how the various quantities (e.g., tree depth, mtry, sample size, ambient dimension, sparsity level) interact with each other and determine the performance, thereby motivating specific choices of the parameters (e.g., tree depth, mtry). Furthermore, consistency in Corollaries 4.4 and 7.2 serves as a *stress-test* and shows how practical implementations of decision trees and random forests can be accurate even with a very high predictor variable count.

Supplementary Materials

The supplementary materials contain detailed proofs for all the theoretical results presented in the paper.

Funding

Jason M. Klusowski was supported in part by the National Science Foundation through grants DMS-2054808 and HDR TRIPODS DATA-INSPIRE DCCF-1934924. Peter M. Tian was supported in part by the Gordon Wu Fellowship of Princeton University.

ORCID

Jason M. Klusowski http://orcid.org/0000-0001-6484-8682

References

Abramovich, F., and Grinshtein, V. (2019), "High-Dimensional Classification by Sparse Logistic Regression," *IEEE Transactions on Information Theory*, 65, 3068–3079. *https://doi.org/10.1109/TIT.2018.2884963*. [5,8]

Bartlett, P., Freund, Y., Lee, W. S., and Schapire, R. E. (1998), "Boosting the Margin: A New Explanation for the Effectiveness of Voting Methods," *The Annals of Statistics*, 26, 1651–1686. *https://doi.org/10.1214/aos/1024691352*. [8]

Boser, B. E., Guyon, I. M., and Vapnik, V. N. (1992), "A Training Algorithm for Optimal Margin Classifiers," in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, pp. 144–152, New York, NY, USA, 1992. Association for Computing Machinery. ISBN 089791497X. *https://doi.org/10.1145/130385.130401*. [8]

Breiman, L. (1996), "Bagging Predictors," *Machine Learning*, 24, 123–140. *https://doi.org/10.1007/BF00058655*. [1,10,11]

——— (2001), "Random Forests," *Machine Learning*, 45, 5–32. *https://doi.org/10.1023/A:1010933404324*. [1,10]

Breiman, L., Friedman, J., Olshen, R. A., and Stone, C. J. (1984), *Classification and Regression Trees*, Boca Raton, FL: Chapman and Hall/CRC. [1,2,4,7]

Bühlmann, P. (2006), "Boosting for High-Dimensional Linear Models," *The Annals of Statistics*, 34, 559–583. *https://doi.org/10.1214/009053606000000092*. [7,8]

Chatterjee, S., and Goswami, S. (2021), "Adaptive Estimation of Multivariate Piecewise Polynomials and Bounded Variation Functions by Optimal Decision Trees," *The Annals of Statistics*, 49, 2531–2551. *https://doi.org/10.1214/20-AOS2045*. [2]

Chi, C.-M., Vossler, P., Fan, Y., and Lv, J. (2020), "Asymptotic Properties of High-Dimensional Random Forests," arXiv preprint arXiv:2004.13953. [2,5]

Denil, M., Matheson, D., and De Freitas, N. (2014), "Narrowing the Gap: Random Forests in Theory and in Practice," in *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of Proceedings of Machine Learning Research, eds. E. P. Xing and T. Jebara, pp. 665–673, Bejing, China, 22–24 Jun 2014. PMLR. Available at *https://proceedings.mlr.press/v32/denil14.html*. [2,10]

Donoho, D. L. (1997), "CART and Best-Ortho-Basis: A Connection," *The Annals of Statistics*, 25, 1870–1911. *https://doi.org/10.1214/aos/1069362377*. [2]

Duroux, R., and Scornet, E. (2018), "Impact of Subsampling and Tree Depth on Random Forests," *ESAIM: PS*, 22, 96–128. *https://doi.org/10.1051/ps/2018008*. [11]

Friedman, J. H. (2001), "Greedy Function Approximation: A Gradient Boosting Machine, *The Annals of Statistics*, 29, 1189–1232. *https://doi.org/10.1214/aos/1013203451*. [1]

Friedman, J. H., and Hall, P. (2007), "On Bagging and Nonlinear Estimation," *Journal of Statistical Planning and Inference*, 137, 669–683. *https://doi.org/10.1016/j.jspi.2006.06.002*; *https://www.sciencedirect.com/science/article/pii/S0378375806001339*. Special Issue on Nonparametric Statistics and Related Topics: In honor of M.L. Puri. [10]

Genuer, R., Poggi, J.-M., and Tuleau, C. (2008), "Random Forests: Some Methodological Insights," arXiv preprint arXiv:0811.3619. [11]

Gey, S., and Nedelec, E. (2005), "Model Selection for Cart Regression Trees," *IEEE Transactions on Information Theory*, 51, 658–670. *https://doi.org/10.1109/TIT.2004.840903*. [2]

Hastie, T., and Tibshirani, R. (1990), *Generalized Additive Models*, London and New York: Chapman and Hall. [5]

Hooker, G. (2007), "Generalized Functional Anova Diagnostics for High-Dimensional Functions of Dependent Variables," *Journal of Computational and Graphical Statistics*, 16, 709–732. *https://doi.org/10.1198/106186007X237892*. [9]

Horowitz, J. L., and Mammen, E. (2004), "Nonparametric Estimation of an Additive Model with a Link Function," *The Annals of Statistics*, 32, 2412–2443. *https://doi.org/10.1214/009053604000000814*. [5]

Jeong, S., and Ročková, V. (2020), "The Art of BART: On Flexibility of Bayesian Forests," arXiv preprint arXiv:2008.06620. [2]

Kearns, M., and Mansour, Y. (1999), "On the Boosting Ability of Top–Down Decision Tree Learning Algorithms," *Journal of Computer and System Sciences*, 58, 109–128. *https://doi.org/10.1006/jcss.1997.1543*; *https://www.sciencedirect.com/science/article/pii/S0022000097915439*. [2]

Klusowski, J. (2020), "Sparse Learning with CART," in *Advances in Neural Information Processing Systems* (Vol. 33), eds. H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, pp. 11612–11622, Curran Associates, Inc. Available at *https://proceedings.neurips.cc/paper/2020/file/85fc37b18c57097425b52fc7afbb6969-Paper.pdf*. [2]

Louppe, G. (2014), "Understanding Random Forests: From Theory to Practice," arXiv preprint: arXiv:1407.7502. [3,4]

Mack, Y. P. (1981), "Local Properties of k-nn Regression Estimates," *SIAM Journal on Algebraic Discrete Methods*, 2, 311–323. *https://doi.org/10.1137/0602035*. [7]

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. Available at *http://jmlr.org/papers/v21/19-905.html*. [11]

Mourtada, J., Gaïffas, S., and Scornet, E. (2020), "Minimax Optimal Rates for Mondrian Trees and Forests," *The Annals of Statistics*, 48, 2253–2276. *https://doi.org/10.1214/19-AOS1886*. [2]

Mourtada, J., Gaïffas, S., and Scornet, E. (2021), "Amf: Aggregated Mondrian Forests for Online Learning," *Journal of the Royal Statistical Society*, Series B, 83, 505–533. *https://doi.org/10.1111/rssb.12425*; *https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12425*. [2]

Probst, P., Wright, M. N., and Boulesteix, A.-L. (2019), "Hyperparameters and Tuning Strategies for Random Forest," *WIREs Data Mining and Knowledge Discovery*, 9, e1301. *https://doi.org/10.1002/widm.1301*; *https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1301*. [11]

Quinlan, J. R. (1993), *C4.5 : Programs for Machine Learning*, San Mateo, CA: Morgan Kaufmann Publishers. [1,4]

Raskutti, G., Wainwright, M. J., and Yu, B. (2012), "Minimax-Optimal Rates for Sparse Additive Models over Kernel Classes via Convex Programming," *Journal of Machine Learning Research*, 13, 389–427. Available at *http://jmlr.org/papers/v13/raskutti12a.html*. [8]

Ročková, V., and van der Pas, S. (2020), "Posterior Concentration for Bayesian Regression Trees and Forests," *The Annals of Statistics*, 48, 2108–2131. *https://doi.org/10.1214/19-AOS1879*. [2]

Ruppert, D., and Wand, M. P. (1994), "Multivariate Locally Weighted Least Squares Regression, "*The Annals of Statistics*, 22, 1346–1370. *https://doi.org/10.1214/aos/1176325632*. [7]

Scornet, E., Biau, G., and Vert, J.-P. (2015), "Consistency of Random Forests," *The Annals of Statistics*, 43, 1716–1741. *https://doi.org/10.1214/15-AOS1321*. [2,5,11]

Segal, M. R. (2004), "Machine Learning Benchmarks and Random Forest Regression," *UCSF: Center for Bioinformatics and Molecular Biostatistics*. Available at *https://escholarship.org/uc/item/35x3v9t4*. [11]

Stone, C. J. (1977), "Consistent Nonparametric Regression," *The Annals of Statistics*, 5, 595–620. *https://doi.org/10.1214/aos/1176343886*. [2]

Syrgkanis, V., and Zampetakis, M. (2020), "Estimation and Inference with Trees and Forests in High Dimensions," in *Proceedings of Thirty Third*

Conference on Learning Theory, volume 125 of Proceedings of Machine Learning Research, eds. J. Abernethy and S. Agarwal, pp. 3453–3454. PMLR, 09–12 Jul 2020. Available at *https://proceedings.mlr.press/v125/syrgkanis20a.html*. [2,5]

Tan, Y. S., Agarwal, A., and Yu, B. (2021), "A Cautionary Tale on Fitting Decision Trees to Data from Additive Models: Generalization Lower Bounds," arXiv preprint arXiv:2110.09626. [8]

Tan, Z., and Zhang, C.-H. (2019), "Doubly Penalized Estimation in Additive Regression with High-Dimensional Data," *The Annals of Statistics*, 47, 2567–2600. *https://doi.org/10.1214/18-AOS1757*. [5,8]

Tutz, G., and Binder, H. (2006), "Generalized Additive Modeling with Implicit Variable Selection by Likelihood-based Boosting," *Biometrics*, 62, 961–971. *https://doi.org/10.1111/j.1541-0420.2006.00578.x*; *https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1541-0420.2006.00578.x*. [5]

Wager, S., and Athey, S. (2018), "Estimation and Inference of Heterogeneous Treatment Effects using Random Forests," *Journal of the American Statistical Association*, 113, 1228–1242. *https://doi.org/10.1080/01621459.2017.1319839*. [2]

Wager, S., and Walther, G. (2015), "Adaptive Concentration of Regression Trees, with Application to Random Forests," arXiv preprint arXiv:1503.06388. [2]

Zhou, S., and Mentch, L. (2021), "Trees, Forests, Chickens, and Eggs: When and Why to Prune Trees in a Random Forest," arXiv preprint arXiv:2103.16700. [11]