Analyzing CART

Jason M. Klusowski Department of Statistics
Rutgers University—New Brunswick
Piscataway, New Jersey 8019
jason.klusowski@rutgers.edu

Abstract

Decision trees with binary splits are popularly constructed using Classification and Regression Trees (CART) methodology. For regression models, this approach recursively divides the data into two near-homogenous daughter nodes according to a split point that maximizes the reduction in sum of squares error (the impurity) along a particular variable. This paper aims to study the statistical properties of regression trees constructed with CART. In doing so, we find that the training error is governed by the Pearson correlation between the optimal decision stump and response data in each node, which we bound by solving a quadratic program. We leverage this connection to show that CART with cost-complexity pruning achieves a good bias-variance tradeoff when the depth scales with the logarithm of the sample size. Data dependent quantities, which adapt to the local dimensionality and structure of the regression surface, are seen to govern the rates of convergence of the prediction error.

1 Introduction

Decision trees are the building blocks of some of the most important and powerful algorithms in statistical learning. For example, ensembles of decision trees are used for some bootstrap aggregated prediction rules (e.g., bagging [2] and random forests [3]). In addition, each iteration of gradient tree boosting (e.g., TreeBoost [8]) fits the pseudo-residuals with decision trees as base learners. From an applied perspective, decision trees have an appealing interpretability and are accompanied by a rich set of analytic and visual diagnostic tools. These attributes make tree-based learning particularly well-suited for applied sciences and related disciplines—which may rely heavily on understanding and interpreting output from a statistical model and the system that generated the data. Although, as with many aspects of statistical learning, good empirical performance often comes at the expense of rigor. Tree-structured learning with decision trees is no exception—statistical guarantees for popular variants, i.e., those that are actually used in practice, are hard to find. Indeed, the recursive manner in which decision trees are constructed makes them unamenable to analysis, especially

when the split protocol involves *both* the input and output data. Despite these challenges, we take a step forward in advancing the theory of decision trees and aim to tackle the following fundamental question.

When do decision trees generalize well to out-of-sample data?

To make our work informative to the applied user of decision trees, we strive to make the least departure from practice and therefore focus specifically on Classification and Regression Tree (CART) [4] methodology—by far the most popular for regression and classification problems. With this methodology, the tree construction importantly depends on both the input and output data and is therefore *data-dependent*. This aspect lends itself favorably to the empirical performance of CART, but poses unique mathematical challenges. It is perhaps not surprising then that, despite the widespread use of CART, there have been only a small handful of papers that study its theoretical properties. For example, [14] study the asymptotic properties of CART in a fixed dimensional regime, en route to establishing consistency of Breiman's random forests for additive regression models. Another notable paper [9] provides oracle-type inequalities for the CART pruning algorithm proposed by [4], though the theory does not imply guarantees for out-out-sample prediction. What the existing literature currently lacks, however, is a more fine-grained analysis that reveals the advantages of tree learning with CART over other unstructured regression procedures, like vanilla nearest neighbors or other kernel based estimators.

Arguably the most difficult aspect of decision trees (and for that matter, any adaptive partitioning-based predictor) is understanding their bias or approximation properties and pinning down structural conditions on the data that enable this. Indeed, most existing convergence results [14] for decision trees or ensembles thereof begin with a study of the size (i.e., the diameter) of the terminal nodes and show that they vanish with the depth of the tree, ensuring that the bias does so also. While this can be useful to prove consistency statements, it is not generally delicate enough to capture the adaptive properties of the tree on the data. To address this shortcoming, one of our crucial insights is that we can avoid using the node diameters as a proxy for the bias and, instead, directly bound the training error in terms of data-dependent quantities that are more transparent and interpretable.

1.1 Learning setting

Let us now describe the learning setting and framework that we will operate under for the rest of the paper. For clarity and ease of exposition, we focus specifically on *regression trees*, where the target outcome is a continuous real value.

We assume the training data is $\mathcal{D}_n = \{(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)\}$, where $(\mathbf{X}_i, Y_i)$, $1 \leq i \leq n$ are i.i.d. with common joint distribution $\mathbb{P}_{\mathbf{X},Y}$. Here, $\mathbf{X}_i \in [0,1]^d$ is the input and $Y_i \in \mathbb{R}$ is a continuous outcome response (or output) variable. A generic pair of variables will be denoted as $(\mathbf{X}, Y)$. A generic coordinate of

$\mathbf{X}$ will be denoted by X, unless there is a need to highlight the dependence on a coordinate index j, denoted X_j, or additionally on a data point i, denoted X_{ij}.Using squared error loss $L(Y, Y') = (Y - Y')^2$ as the performance metric, our goal is to predict Y at a new point $\mathbf{X} = \mathbf{x}$ via a tree structured prediction rule $\widehat{Y}(\mathbf{x}) = \widehat{Y}(\mathbf{x}; \mathcal{D}_n)$. The training error on the training data is $\overline{\text{err}}_n(\widehat{Y}) = \frac{1}{n}\sum_{i=1}^n (Y_i - \widehat{Y}(\mathbf{X}_i))^2$ and the test error on a test sample $\mathcal{D}'_m = \{(\mathbf{X}'_i, Y'_i)\}_{i=1}^m$ of size m is $\overline{\text{Err}}_m(\widehat{Y}) = \frac{1}{m}\sum_{i=1}^m (Y'_i - \widehat{Y}(\mathbf{X}'_i))^2$. For brevity, we let $\widehat{\sigma}^2_Y$ denote the sample variance of $Y_1, \ldots, Y_n$. We state our performance guarantees in terms of the population level prediction error $\text{Err}(\widehat{Y}) = \mathbb{E}_{(\mathbf{X}', Y')}[(Y' - \widehat{Y}(\mathbf{X}'))^2]$. Proofs of all forthcoming results are given in the supplement.

2 Preliminaries

As mentioned earlier, regression trees are commonly constructed with Classification and Regression Tree (CART) [4] 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, CART trees are greedily grown with a procedure in which binary splits recursively partition the tree into near-homogeneous terminal nodes. That is, an effective binary split partitions the data from the parent tree node into two daughter nodes so that the resultant homogeneity of the daughter nodes, as measured through their *impurity*, is improved from the homogeneity of the parent node.

The CART algorithm is comprised of two elements—a growing procedure and a pruning procedure. The growing procedure constructs from the data a maximal binary tree T_{max} by the recursive partitioning scheme; the pruning procedure selects, among all the subtrees of T_{max}, a sequence of subtrees that greedily optimize a cost function.

2.1 Growing the tree

Let us now describe the tree construction algorithm with additional detail. Consider splitting a regression tree T at a node t. Let s be a candidate split for a variable X that splits t into left and right daughter nodes t_L and t_R according to whether $X \le s$ or $X > s$. These two nodes will be denoted by $\text{t}_L = \{\mathbf{X} \in \text{t} : X \le s\}$ and $\text{t}_R = \{\mathbf{X} \in \text{t} : X > s\}$. As mentioned previously, a tree is grown by recursively reducing node impurity. Impurity for regression trees is determined by the within node sample variance

$$\widehat{\Delta}(\text{t}) \coloneqq \widehat{\text{VAR}}(Y \mid \mathbf{X} \in \text{t}) = \frac{1}{N(\text{t})} \sum_{\mathbf{X}_i \in \text{t}} (Y_i - \overline{Y}_\text{t})^2, \tag{1}$$

where $\overline{Y}_\text{t} = \frac{1}{N(\text{t})}\sum_{\mathbf{X}_i \in \text{t}} Y_i$ is the sample mean for t and $N(\text{t}) = \#\{i : \mathbf{X}_i \in \text{t}\}$ is the number of data points in t. Similarly, the within sample variance for a

daughter node is

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

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). The parent node t is split into two daughter nodes using the variable and split point producing the largest decrease in impurity. For a candidate split s for X, this decrease in impurity equals [4, Definition 8.13]

$$\widehat{\Delta}(s, \mathrm{t}) = \widehat{\Delta}(\mathrm{t}) - [\widehat{P}(\mathrm{t}_L)\widehat{\Delta}(\mathrm{t}_L) + \widehat{P}(\mathrm{t}_R)\widehat{\Delta}(\mathrm{t}_R)], \tag{2}$$

where $\widehat{P}(\mathrm{t}_L) = N(\mathrm{t}_L)/N(\mathrm{t})$ and $\widehat{P}(\mathrm{t}_R) = N(\mathrm{t}_R)/N(\mathrm{t})$ are the proportions of data points in t that are contained in t_L and t_R, respectively.

To summarize, the tree T is grown recursively by finding the variable $\hat{\jmath}$ and split point $\hat{s}$ that maximizes $\widehat{\Delta}(s, \mathrm{t})$. Note that for notational brevity, we suppress the dependence on directions j. The output of the tree at a terminal node t is the least squares predictor, namely, $\widehat{Y}(T, \mathbf{x}) = \overline{Y}_{\mathrm{t}}$ for all $\mathbf{x} \in \mathrm{t}$.

2.2 Pruning the tree

The CART growing procedure stops once a maximal binary tree T_{max} is grown (i.e., when the terminal nodes contain at least a single data point). However, $\widehat{Y}(T_{max})$ is generally not a good predictor, since it will tend to overfit the data and therefore generalize poorly to unseen data. This effect can be mitigated by complexity regularization. Removing portions of the overly complex tree (i.e., via pruning) is one way of reducing its complexity and improving performance. We will now describe such a procedure.

We say that T is a pruned subtree of T', written as $T \preceq T'$, if T can be obtained from T' by collapsing any number of its internal nodes. A pruned subtree of T_{max} is defined as any binary subtree of T_{max} having the same root node as T_{max}. The number of terminal nodes in a tree T is denoted $|T|$. Given a subtree T and temperature $\alpha > 0$, we define the penalized cost function

$$R_\alpha(\widehat{Y}(T)) = \overline{\mathrm{err}}_n(\widehat{Y}(T)) + \alpha|T|. \tag{3}$$

As shown in [4, Section 10.2], the smallest minimizing subtree for the temperature α,

$$\widehat{T} \in \operatorname*{arg\,min}_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T)),$$

exists and is unique (smallest in the sense that if $T' \in \arg\min_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T))$, then $\widehat{T} \preceq T'$). The optimal subtree $\widehat{T}$ can be found efficiently by weakest link pruning [4], i.e., by successively collapsing the internal node that produce the smallest per-node increase in $\overline{\mathrm{err}}_n(\widehat{Y}(T))$ until we obtain the tree consisting of the root node. Good values of α can be selected using cross-validation, for

example, though analyzing the effect of such a procedure is outside the scope of the present paper.

Our first result shows that, with high probability, the test error of the pruned tree $\widehat{T}$ on new data is bounded by a multiple of $\min_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T))$.

Theorem 1. *Let $\widehat{T}$ be the smallest minimizer of (3). Suppose $Y = f(\mathbf{X})$, $B = \|f\|_\infty < \infty$, $n > (d+1)/2$, and $\alpha > \frac{18B^2(d+1)\log(2en/(d+1))}{n}$. Then, with probability at least $1-\delta$ over $\mathcal{D}_n$,*

$$Err(\widehat{Y}(\widehat{T})) \leq 2 \min_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T)) + \frac{18B^2 \log(1/\delta^2)}{n} + 4B^2\delta.$$

Remark 1. *Similar bounds hold for the binary classification context, i.e., $Y \in \{0,1\}$, since the squared error impurity (1) equals one-half of the so-called* Gini *impurity used for classification trees.*

In what follows, we let $T_K \preceq T_{max}$ denote a fully grown binary tree of depth K, i.e., we grow the tree until each node contains a single data point or a depth of K is reached, whichever occurs sooner. We also let $\widehat{T}$ be the smallest minimizer of the cost function (3) with temperature $\alpha = \Theta((d/n)\log(n/d))$.

3 Bounded the training error

In the previous section, Theorem 1 showed that, with high probability, the test error is bounded by the cost function (3) at its minimum. Since the cost function is defined as the training error plus penalty term, the next step in our course of study is to understand how the training error of CART behaves.

3.1 Splitting criterion and Pearson correlation

Before we begin our analysis of the training error, we first digress back to the tree construction algorithm and give an alternative characterization of the objective. Now, the use of the sum of squares impurity criterion $\widehat{\Delta}(s, \mathrm{t})$ with averages in the terminal nodes permits further simplifications of the formula (2) above. For example, using the sum of squares decomposition, $\widehat{\Delta}(s, \mathrm{t})$ can equivalently be expressed as [4, Section 9.3]

$$\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)\big(\overline{Y}_{\mathrm{t}_L} - \overline{Y}_{\mathrm{t}_R}\big)^2, \tag{4}$$

which is commonly used for its computational appeal—that is, one can find the best split for a continuous variable with just a single pass over the data, without the need to calculate multiple averages and sums of squared differences for these averages, as required with (2). Yet another way to view $\widehat{\Delta}(s, \mathrm{t})$, which appears to not have been considered in past literature and will prove to be useful for our

purposes, is via its equivalent representation as $\widehat{\Delta}(\mathrm{t}) \times \widehat{\rho}^2(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t})$, where

$$\widehat{\rho}\,(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \coloneqq \frac{\frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} (\widetilde{Y}_i - \overline{Y}_{\mathrm{t}})(Y_i - \overline{Y}_{\mathrm{t}})}{\sqrt{\frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} (\widetilde{Y}_i - \overline{Y}_{\mathrm{t}})^2 \times \frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} (Y_i - \overline{Y}_{\mathrm{t}})^2}} \tag{5}$$

is the Pearson product-moment correlation coefficient between the decision stump

$$\widetilde{Y} \coloneqq \overline{Y}_{\mathrm{t}_L} \mathbf{1}_{\{X \leq s\}} + \overline{Y}_{\mathrm{t}_R} \mathbf{1}_{\{X > s\}} \tag{6}$$

and response variable Y within t (see Lemma 4 in the supplement). Hence, at each node, CART seeks the decision stump most correlated in magnitude with the response variable along a particular variable, i.e.,

$$\hat{s} \in \arg\max_s \widehat{\Delta}(s, \mathrm{t}) = \arg\max_s |\widehat{\rho}\,(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t})|. \tag{7}$$

We let $\widehat{Y}$ denote the decision stump $\widetilde{Y}$ with the optimal direction $\hat{\jmath} \in \arg\max_{j=1,2,\ldots,d} \widehat{\Delta}(\hat{s}, \mathrm{t})$ with corresponding optimal split $\hat{s}$. It should be stressed that this alternative characterization of the splitting criterion (2) is unique to the squared error impurity with (constant) averages in the terminal nodes of the tree.

3.2 Location of splits and Pearson correlation

Having already revealed the intimate role the correlation between the decision stump and response values (5) plays in the tree construction, it is instructive to explore this relationship with the location of the splits. In order to study this cleanly, let us for the moment work in an asymptotic data setting to determine the directions to split and their split points, i.e.,

$$\widehat{\Delta}(s, \mathrm{t}) \underset{n \to \infty}{\to} \Delta(s, \mathrm{t}) \coloneqq \Delta(\mathrm{t}) - [P(\mathrm{t}_L)\Delta(\mathrm{t}_L) + P(\mathrm{t}_R)\Delta(\mathrm{t}_R)], \tag{8}$$

where quantities without hats are the population level counterparts of the empirical quantities discussed previously. The decision stump (6) with the optimal theoretical direction j^* with corresponding optimal theoretical split s^* is denoted by $\widehat{Y}^*$. Now, if the number of data points within t is large and $\Delta(s, \mathrm{t})$ has a unique global maximum, then we can expect $\hat{s} \approx s^*$ (via an empirical process argument) and hence the infinite sample setting is a good approximation to CART with empirical splits, giving us some insights into its dynamics. Indeed, if s^* is unique, [11, Theorem 2] shows that $\hat{s}$ converges in probability to s^*. With additional assumptions, one can go even further and characterize the rate of convergence. For example, [5, Section 3.4.2] provide cube root asymptotics for $\hat{s}$, i.e., $n^{1/3}(\hat{s} - s^*)$ converges in distribution.

Each node t is a Cartesian product of intervals. As such, the interval along variable X in t is denoted by $[a, b]$, where $a < b$. The next theorem characterizes the relationship between the optimal theoretical split s^* and infinite sample

correlation $\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t}) \stackrel{a.s.}{=} \lim_n \widehat{\rho}(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})$, which, in turn, can be used to bound the bias of the tree. The proof is based on the first-order optimality condition, namely, $\frac{\partial}{\partial s}\Delta(s, \mathrm{t}) \mid_{s=s^*} = 0$.

Theorem 2. *Suppose* $\mathbf{X}$ *is uniformly distributed and* $\Delta(s^*, \mathrm{t}) > 0$. *Then the optimal theoretical split* $s^* \in [a, b]$ *along variable* X *has the form*

$$\frac{a+b}{2} \pm \frac{b-a}{2}\sqrt{\frac{v}{v + \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})}}, \tag{9}$$

where $v = \frac{(\mathbb{E}[Y|\mathbf{X}\in \mathrm{t},\, X=s^*] - \mathbb{E}[Y|\mathbf{X}\in \mathrm{t}])^2}{VAR(Y|\mathbf{X}\in \mathrm{t})}$.

Expression (9) in Theorem 2 reveals that the optimal theoretical split s^* is a perturbation of the median $(a+b)/2$ of the conditional distribution $X \mid \mathbf{X} \in \mathrm{t}$, where the gap is governed by the correlation $\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})$. In particular, splits along directions that contain a strong signal, i.e., $|\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})| \gg 0$, tend to be further away from the parent node edges, thereby producing terminal node lengths that are on average narrower. At the other extreme, the correlation is weakest when there is no signal in the splitting direction or when the response values in the node are not fit well by a decision stump—yielding either $s^* \approx a$ or $s^* \approx b$—and hence the predicted output in one of the daughter nodes does not change by much. For example, if $Y = g(X)$ is a sinusoidal waveform with large frequency w (not fit well by a single decision stump) and t is the root node $[0,1]^d$, then $v = \Theta(1)$ and $|\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})| = \Theta(1/\sqrt{w})$, and hence by (9), either $s^* = \Theta(1/w)$ or $s^* = 1 - \Theta(1/w)$ (see Lemma 5 in the supplement). This phenomenon has been dubbed 'end-cut preference' in the literature [11], [4, Section 11.8] and has been known empirically since the inception of CART [4, Section 11.8]. The theory above is also consistent with empirical studies on the adaptive properties of Breiman's random forests which use CART [12, Section 4].

3.3 Training error, bias, and Pearson correlation

In addition to determining the location of the splits, the correlation is also intimately connected to the training error. Intuitively, the training error should small when CART finds decision stumps that have strong correlation with the response values in each node. More precisely, the following lemma reveals the importance of the correlation (5) in controlling the training error. It shows that each time a node t is split, the training error in t is reduced by a constant factor, namely, $\exp(-\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}))$. Recursing this inequality over nodes at each level of the tree leads to the conclusion that the training error should be exponentially small in the depth, provided the correlation at each node is large.

Lemma 1. *With probability one,*

$$\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}} (Y_i - \widehat{Y}_i)^2 \leq \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}} (Y_i - \overline{Y}_{\mathrm{t}})^2 \times \exp(-\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})), \tag{10}$$

and hence

$$\overline{err}_n(\widehat{Y}(T_K)) \leq \widehat{\sigma}_Y^2 \exp(-K\,\widehat{r}^{\,2}), \tag{11}$$

where $\widehat{r} \coloneqq \min_{\mathrm{t}} \widehat{\rho}^{\,2}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})$ *and the minimum extends over all internal nodes* t *of* T_K*.*

Lemma 1 also has an analog in the infinite sample setting—with a similar conclusion for the asymptotic bias

$$\mathrm{Bias}(\widehat{Y}(T^*, \mathbf{x})) \coloneqq f(\mathbf{x}) - \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t}], \quad \mathbf{x} \in \mathrm{t},$$

where t is a terminal node from a tree T^* that is grown to depth K with the optimal theoretical directions j^* and splits s^*. If $\mathbf{X}$ is uniformly distributed and $f(\cdot)$ is an additive function with d_0 components that are not too 'flat', then we show (see the proof of Theorem 3 in the supplement) that for each fixed node t, the squared correlation between the response values Y and decision stump $\widehat{Y}$ at the optimal empirical direction $\hat{\jmath}$ and split $\hat{s}$ is asymptotically $\Omega(1/d_0)$, i.e., $\liminf_n \widehat{\rho}^{\,2}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \overset{a.s.}{=} \Omega(1/d_0)$. As suggested by (11), this can be used to bound the bias and, by doing so, illustrate the approximation quality of CART.

Theorem 3. *Suppose* $\mathbf{X}$ *is uniformly distributed,* $f(\mathbf{X}) = \sum_{j \in \mathcal{S}} g_j(X_j)$*,* $\#\mathcal{S} = d_0 \leq d$*, and each* $g_j(\cdot)$ *admits a power series representation. Then there exists a constant* $\Lambda > 0$ *that depends only on each* $g_j(\cdot)$ *such that*

$$\mathbb{E}_{\mathbf{X}}[Bias^2(\widehat{Y}(T^*, \mathbf{X}))] \leq VAR(Y) \exp(-\Lambda K/d_0).$$

3.4 Size of Pearson correlation

Due to the importance of the correlation in controlling the training error, it is natural to ask when it will be large. We accomplish this by studying its size relative to the correlation between the data and another more flexible basis function. That is, we fit an arbitrary function $g(X)$ to the data in the node and ask how large $|\widehat{\rho}\,(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})|$ is relative to $|\widehat{\rho}\,(g(X), Y \mid \mathbf{X} \in \mathrm{t})|$. Such a relationship will enable us to conclude that if Y is locally correlated with $g(X)$ in the node, then so will Y with the optimal decision stump $\widehat{Y}$. Before we continue, let us mention that studying $\widehat{\rho}\,(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})$ directly is hopeless since it almost never admits a closed form expression. Furthermore, it is difficult to rely on concentration of measure when t contains insufficient data; a likely situation among deep nodes. Nevertheless, by definition of $\widehat{Y}$ via (7), we can construct a prior $\Pi(j, s)$ on coordinates j and splits s, and lower bound $|\widehat{\rho}\,(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})|$ by

$$\int |\widehat{\rho}\,(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t})| d\Pi(j, s), \tag{12}$$

which is much less burdensome to analyze. Importantly, the prior can involve unknown quantities from the distribution of $(\mathbf{X}, Y)$. The next set of results are based on studying (12) for a special choice of prior Π, which, curiously, involves solving a quadratic program. Our first result shows that, relative to a general

univariate function, the optimal decision stump produces a correlation with the data that is always above the $1/\sqrt{N(\mathrm{t})}$ noise level in the node.

Lemma 2. *With probability one, uniformly over all functions $g(\cdot)$ of X that change from (strictly) increasing to decreasing and vice versa at most V times in the node, we have*

$$|\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})| \geq \frac{1}{\sqrt{N(\mathrm{t}) \times (V+1)}} \times |\widehat{\rho}(g(X), Y \mid \mathbf{X} \in \mathrm{t})|. \tag{13}$$

Additional assumptions on $g(\cdot)$ are required to obtain a useful lower bound. The next result shows that despite fitting the data with a decision stump with *one* degree of freedom, CART behaves almost as if it fit the data with an arbitrary monotone function with $N(\mathrm{t}) - 1$ degrees of freedom, at the expense of a sublogarithmic factor in $N(\mathrm{t})$. For example, the correlation between the response variable and the decision stump is at least as strong (up to a sublogarithmic factor) as the correlation between the response variable and a linear or isotonic fit.

Lemma 3. *With probability one, uniformly over all monotone functions $g(\cdot)$ of X in the node, we have*

$$|\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})| \geq \frac{1}{\sqrt{1 + \log(2N(\mathrm{t}))}} \times |\widehat{\rho}(g(X), Y \mid \mathbf{X} \in \mathrm{t})|. \tag{14}$$

The previous lemma also suggests that CART is quite good at fitting response surfaces that have a local, low-dimensional, monotone relationship with the input variables. We will exploit this fact when we obtain a bound on the test error. Note that because correlation is merely a measure of linear association, $|\widehat{\rho}(g(X), Y \mid \mathbf{X} \in \mathrm{t})|$ can still be large for some monotone $g(\cdot)$, even if Y is not approximately monotone in one coordinate. That is, Y need only be locally correlated with such a function.

4 Main results

In this section, we use the correlation inequalities from Section 3.4 to give bounds on prediction error of CART. We first give the high-level strategy to obtain our next set of results. By Theorem 1, with high probability, the leading behavior of the prediction error $\mathrm{Err}(\widehat{Y}(\widehat{T}))$ is governed by $\inf_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T))$, which is smaller than the minimum of $R_\alpha(\widehat{Y}(T_K)) = \overline{\mathrm{err}}_n(\widehat{Y}(T_K)) + \alpha|T_K|$ over all fully grown trees T_K of depth K with $|T_K| \leq 2^K$, i.e.,

$$\inf_{K \geq 1} \{\overline{\mathrm{err}}_n(\widehat{Y}(T_K)) + \alpha 2^K\}. \tag{15}$$

Coupled with an informative bound on $\overline{\mathrm{err}}_n(\widehat{Y}(T_K))$, (15) can then be further bounded and solved. The proofs reveal that a good balance between the model fit and complexity typically occurs when $K = \Theta(\log_2(n))$.

4.1 Consistency rates in high-dimensional settings

In this section, we provide rates of convergence for the CART algorithm under a mild assumption on the size of N_k—the largest number of data points in a node at level k in T_K. As we shall see, this assumptions yields finite sample rates of convergence for the prediction error, which, in turn, reveal how the local nature of trees enable adaptation to the response surface. Importantly, our theory only requires that $N(\mathrm{t})$ is *upper* bounded at each level of the tree. This allows for nodes that have very few data points, which is typical for trees trained in practice. Contrast this assumption with past work on tree learning algorithms that requires each $N(\mathrm{t})$ to be *lower* bounded so that local estimation is valid. For example, [4, Section 12.2], en route to establishing asymptotic consistency of CART, assumes that each terminal node contains at least $\omega(\log n)$ data points. The assumption of large $N(\mathrm{t})$ is often accompanied by further assuming the diameters of the terminal nodes converge to zero. We make no such assumption. Note that for methods using ensembles of trees, existing theoretical results require comparable regularity conditions—see the extant literature on random forest models, e.g., [13, 15].

Assumption 1. *For some positive constant $A > 0$, the largest number of data points in a node at level k in T_K satisfies $N_k \leq An/2^k$, $k = 1, 2, \ldots, K = \mathcal{O}(\log_2(n))$.*

Theorem 4. *Let $Y = f(\mathbf{X})$ with $\|f\|_\infty < \infty$. Under Assumption 1, with probability one*

$$\overline{err}_n(\widehat{Y}(T_K)) \leq \widehat{\sigma}_Y^2 \Big(1 - \frac{K}{\log_2(4An)}\Big)^{\widehat{\rho}^{\,2}}, \tag{16}$$

where

$$\widehat{\rho} \coloneqq \min_{\mathrm{t}} \sup_{g \in \mathcal{G},\ j=1,2,\ldots,d} |\widehat{\rho}\,(g(X_j), Y \mid \mathbf{X} \in \mathrm{t})|;$$

the minumum runs over all internal nodes t in T_K and $\mathcal{G}$ is the set of all monotone functions $g : \mathbb{R} \to \mathbb{R}$. Furthermore, with probability at least $1 - \delta$,

$$Err(\widehat{Y}(\widehat{T})) = \mathcal{O}\Big(\widehat{\sigma}_Y^2 \Big(\frac{\log((d/\widehat{\sigma}_Y^2)(\log(n))^2)}{\log(n)}\Big)^{\widehat{\rho}^{\,2}} + \frac{\log(1/\delta)}{n} + \delta\Big). \tag{17}$$

According to Lemma 3, the empirical correlation (5) is large when Y is locally correlated with a monotone function in one of the input coordinates. Thus, the quantity $\widehat{\rho}$ from Theorem 4 is a measure of the local monotone dependence between $\mathbf{X}$ and Y for a worst-case node. By (16), the training error decreases with the depth if CART partitions the response surface into pieces that are locally monotone in a few of the input coordinates. We will now argue that $\widehat{\rho}$ is an empirical measure of the local dimensionality of Y; in particular, we argue that if CART effectively partitions the response surface so that beyond a certain level, in each node, Y is locally correlated with an additive function with d_0 monotone components (see [6, 1] for applications of this model in data analysis), then $\widehat{\rho}^{\,2} = \Omega(1/d_0)$.

Theorem 5. *Suppose* $g(\mathbf{X}) = \sum_{j \in \mathcal{S}} g_j(X_j)$ *with* $\#\mathcal{S} = d_0 \leq d$*, where each* $g_j(\cdot)$ *has nonnegative correlation with the others. Then, for each node* t*, with probability one,*

$$\max_{j=1,2,\ldots,d} \widehat{\rho}^2(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}) \geq \frac{\widehat{\rho}^2(g(\mathbf{X}), Y \mid \mathbf{X} \in \mathrm{t})}{d_0}. \tag{18}$$

Note that (18) holds regardless of the correlation structure between the d_0 input coordinates in $\mathcal{S}$ with strong output signals and the $d - d_0$ input coordinates in $\mathcal{S}^c$ with noisy output signals.

Remark 2. *If* $Y = g(X_j)$ *for some variable* X_j *(not known a priori) and monotone function* $g(\cdot)$*, then* $\widehat{\rho} = 1$ *and according to Theorem 4, with high probability,* $Err(\widehat{Y}(\widehat{T})) = \mathcal{O}(\log(d)/\log(n)) \to 0$*, provided* $d = e^{o(\log(n))}$*. More generally, (18) suggests that it is possible to achieve rates of the form* $(\log(d)/\log(n))^{\Omega(1/d_0)}$ *—which vanish even when* $d = e^{o(\log(n))}$ *—where* d_0 *is the intrinsic dimension of the regression surface.*

4.2 Polynomial consistency rates

Despite its ubiquity in statistical learning methodology, CART suffers from two major pitfalls: instability (i.e., small perturbations in the training samples may significantly change the structure of the optimal tree) and inability to accurately approximate even simple response surfaces, such as linear or, more generally, additive [7]. In many applied settings, though, the primary focus is on interpretability and transparency, and thus CART may still be the tool of choice. In such cases, the user may be interested to know *when* CART will work well—in terms fast *polynomial* rates of convergence. To this end, applying (11) to Theorem 1 with $K = \frac{1}{\widehat{r}^2 + \log 2} \log(\widehat{\sigma}_Y^2/\alpha)$, we have that with probability at least $1 - \delta$,

$$\mathrm{Err}(\widehat{Y}(\widehat{T})) = \mathcal{O}\Big(\widehat{\sigma}_Y^2 \Big(\frac{d \log(n/d)}{n \widehat{\sigma}_Y^2}\Big)^{\frac{\widehat{r}^2}{\widehat{r}^2 + \log 2}} + \frac{\log(1/\delta)}{n} + \delta\Big). \tag{19}$$

What remains then is to determine the size of $\widehat{r}$. Before we continue, the two aforementioned issues with CART must first be addressed. Firstly, a decision stump fit to the data with squared error loss tends to be highly sensitive to outliers and other influential data points—yielding small values of $|\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})|$, depending on how the data in arranged in the node. Therefore, to reduce the instability of CART, we will require the data to be spread out (as opposed to it being sparse in certain places and dense in others). Secondly, to address the issue of the approximation accuracy of CART, we will assume that the response surface Y is piecewise constant and lies on a one-dimensional subspace determined by one of the input coordinates. These two assumptions are reflected in the following.

Assumption 2. *There exists a fixed real number* $0 < q < 1$ *such that, in each internal node* t *in* T_K*,* $Y_i = g(X_{ij})$ *for some variable* X_j *and piecewise constant*

function $g(\cdot)$, where each constant piece with at least one data point contains at least $q \times N(\mathrm{t})$ data points.

Our next theorem shows that polynomial rates of convergence can be achieved if the data satisfies Assumption 2. Importantly, the bound is precise enough in its dependence on the ambient dimension d to conclude that the prediction error $\mathrm{Err}(\widehat{Y}(\widehat{T}))$ goes to zero even when $d = o(n)$. The proof is based on the device (12) for studying the correlations in each node, which we, in turn, use to bound the training error (see Lemma 1) and finally the prediction error via (15).

Theorem 6. *Let $Y = f(\mathbf{X})$ with $\|f\|_\infty < \infty$. Under Assumption 2, with probability one,*

$$\overline{err}_n(\widehat{Y}(T_K)) \le \widehat{\sigma}_Y^2 \exp(-Kq^2). \tag{20}$$

Furthermore, with probability at least $1 - \delta$,

$$Err(\widehat{Y}(\widehat{T})) = \mathcal{O}\Big(\widehat{\sigma}_Y^2 \Big(\frac{d \log(n/d)}{n \widehat{\sigma}_Y^2}\Big)^{\frac{q^2}{q^2 + \log 2}} + \frac{\log(1/\delta)}{n} + \delta\Big). \tag{21}$$

5 Discussion

5.1 Automatic dimension reduction

One key strength of decision trees is that they can exploit, if present, low local dimensionality of the response surface. This is particularly useful since many real-world input/output systems are locally approximated by simple model forms with only a few variables. That is, even though the input/output relationship is determined by a large number of variables overall, the dependence may be locally characterized by only a small subset of variables. Such local adaptivity is made possible by the recursive partitioning of the input space, in which optimal splits are increasingly affected by local qualities of the data as the tree is grown. The data dependent quantity $\widehat{\rho}$, which measures the local monotone dependence between the response variable and an input coordinate, in essence, captures this ability.

Let us conclude by making a few comments on ensembles of decision trees. One of the key insights for our analysis of CART was the ability to connect the training error to the objective function of the growing procedure, as in Lemma 1. This connection enabled us to reveal how certain data-dependent quantities adapt to the intrinsic dimensionality of the response surface and control the rates of convergence. Establishing similar relationships is not as easy with tree ensembles like bagging or random forests, due to the subsampling step and form of predicted output as an average of trees. Nevertheless, by convexity [3, Section 11] or [2, Section 4.1] show that the prediction error of a random forest or bagged regression trees is at most the weighted correlation between the residuals of the trees times the average prediction error of the individual trees. In that sense, tree ensembles should, on average, perform at least as well as individual trees, and so our theory for CART can be taken as a benchmark for such refinements.

References

[1] Peter Bacchetti. Additive isotonic models. *Journal of the American Statistical Association*, 84(405):289–294, 1989.

[2] Leo Breiman. Bagging predictors. *Machine learning*, 24(2):123–140, 1996.

[3] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001.

[4] Leo Breiman, Jerome Friedman, RA Olshen, and Charles J Stone. *Classification and regression trees*. Chapman and Hall/CRC, 1984.

[5] Peter Bühlmann and Bin Yu. Analyzing bagging. *The Annals of Statistics*, 30(4): 927–961, 2002.

[6] Yining Chen and Richard J Samworth. Generalized additive and index models with shape constraints. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(4):729–754, 2016.

[7] Jerome H Friedman. Multivariate adaptive regression splines. *The Annals of Statistics*, 19(1):1–67, 1991.

[8] Jerome H Friedman. Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, pages 1189–1232, 2001.

[9] Servane Gey and Elodie Nedelec. Model selection for CART regression trees. *IEEE Transactions on Information Theory*, 51(2):658–670, 2005.

[10] Cong Huang, Gerald HL Cheang, and Andrew R Barron. *Risk of penalized least squares, greedy selection and L1-penalization for flexible function libraries*. PhD thesis.

[11] Hemant Ishwaran. The effect of splitting on random forests. *Machine Learning*, 99(1):75–118, 2015.

[12] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

[13] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7(Jun):983–999, 2006.

[14] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741, 2015.

[15] Stefan Wager and Susan Athey. Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, pages 1–15, 2018.

Appendix

In Appendix A, we provide proofs of Theorem 1, Theorem 2, Lemma 1, Theorem 3, Lemma 3, Lemma 2, Theorem 5, Theorem 4, and Theorem 6 from the main body of the paper. We also state and prove any supporting lemmas in Appendix B.

Henceforth, we denote the daughter nodes with the optimal split $\hat{s}$ by $\hat{\mathrm{t}}_L$ and $\hat{\mathrm{t}}_R$, i.e., $\hat{\mathrm{t}}_L = \{\mathbf{X} \in \mathrm{t} : X \leq \hat{s}\}$ and $\hat{\mathrm{t}}_R = \{\mathbf{X} \in \mathrm{t} : X > \hat{s}\}$, and those with the optimal theoretical split s^* by t_L^* and t_R^*, respectively. As a general rule, if the coordinate index j is omitted on any quantity that should otherwise depend on j, it should be understood that we are considering a generic variable X.

A Proofs of main theorems and lemmas

Lemma 4 (Equivalence between the decrease in impurity and Pearson correlation)**.**

$$\widehat{\rho}\,(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) = \sqrt{\widehat{\Delta}(s, \mathrm{t})/\widehat{\Delta}(\mathrm{t})}.$$

Proof. By expanding the sum of squares in (2), it can easily be shown that $\widehat{\Delta}(s, \mathrm{t})$ equals

$$\widehat{P}(\mathrm{t}_L)(\overline{Y}_{\mathrm{t}_L})^2 + \widehat{P}(\mathrm{t}_R)(\overline{Y}_{\mathrm{t}_R})^2 - (\overline{Y}_{\mathrm{t}})^2,$$

which is further equal to both $\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}(\widetilde{Y}_i - \overline{Y}_{\mathrm{t}})^2$ and $\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}(\widetilde{Y}_i - \overline{Y}_{\mathrm{t}})(Y_i - \overline{Y}_{\mathrm{t}})$. Thus,

$$\begin{aligned}
\widehat{\rho}\,(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) &= \frac{\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}(\widetilde{Y}_i - \overline{Y}_{\mathrm{t}})(Y_i - \overline{Y}_{\mathrm{t}})}{\sqrt{\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}(\widetilde{Y}_i - \overline{Y}_{\mathrm{t}})^2 \times \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}(Y_i - \overline{Y}_{\mathrm{t}})^2}} \qquad (22)\\
&= \frac{\widehat{P}(\mathrm{t}_L)(\overline{Y}_{\mathrm{t}_L})^2 + \widehat{P}(\mathrm{t}_R)(\overline{Y}_{\mathrm{t}_R})^2 - (\overline{Y}_{\mathrm{t}})^2}{\sqrt{(\widehat{P}(\mathrm{t}_L)(\overline{Y}_{\mathrm{t}_L})^2 + \widehat{P}(\mathrm{t}_R)(\overline{Y}_{\mathrm{t}_R})^2 - (\overline{Y}_{\mathrm{t}})) \times \widehat{\Delta}(\mathrm{t})}}\\
&= \sqrt{\frac{\widehat{P}(\mathrm{t}_L)(\overline{Y}_{\mathrm{t}_L})^2 + \widehat{P}(\mathrm{t}_R)(\overline{Y}_{\mathrm{t}_R})^2 - (\overline{Y}_{\mathrm{t}})^2}{\widehat{\Delta}(\mathrm{t})}}\\
&= \sqrt{\widehat{\Delta}(s, \mathrm{t})/\widehat{\Delta}(\mathrm{t})}.
\end{aligned}$$

Note that the mean of the decision stump $\widetilde{Y}$ in t is in fact $\overline{Y}_{\mathrm{t}}$, which is why it appears in the formula (22) for the Pearson correlation. □

Lemma 5 (Example from Section 3.2)**.** *Let* $Y = \sin(2\pi w X)$ *for some positive integer* w *and* $\mathrm{t} = [0, 1]^d$*. Then,*

$$|\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})| = \Theta(1/\sqrt{w}),\ s^* = \Theta(1/w),\ \textit{and}\ s^* = 1 - \Theta(1/w).$$

Proof. Elementary calculations reveal that $\Delta(s, \mathrm{t}) = \frac{(1-\cos(2\pi w s))^2}{4\pi^2 w^2 s(1-s)} = \frac{(1-\cos(2\pi w(1-s)))^2}{4\pi^2 w^2 s(1-s)}$. It can be seen from this expression that the maximizers satisfy $s^* = \Theta(1/w)$ and $s^* = 1 - \Theta(1/w)$ and $\Delta(s^*, \mathrm{t}) = \Theta(1/w)$. Since $\Delta(\mathrm{t}) = 1/2$, we have from the infinite sample analog of Lemma 4 that $|\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})| = \sqrt{\Delta(s^*, \mathrm{t})/\Delta(\mathrm{t})} = \Theta(1/\sqrt{w})$. □

Proof of Theorem 1. Recall that $\overline{\mathrm{Err}}_n(\widehat{Y}(\widehat{T}))$ is the test error on a new sample $\mathcal{D}'_n = \{(\mathbf{X}'_i, Y'_i)\}_{i=1}^n$ of size n and $\mathrm{Err}(\widehat{Y}(\widehat{T})) = \mathbb{E}_{(\mathbf{X}',Y')}[(Y' - \widehat{Y}(\widehat{T}, \mathbf{X}'))^2]$ is the mean-squared prediction error. Let $\mathcal{T}$ denote the collection of tree-structured partitions constructed on the grid $\{\mathbf{X}_i\}_{i=1}^n \cup \{\mathbf{X}'_i\}_{i=1}^n$ with $2n$ points. Note that the VC-dimension of the collection of axis-parallel splits is at most $d+1$. Lemma B.2 in [9] shows that the number of trees in $\mathcal{T}$ with exactly $|T|$ nodes is at most $(2ne/(d+1))^{|T|(d+1)}$. Using this, we have

$$\sum_{T\in\mathcal{T}} e^{-L(T)} \leq \sum_{k:|T|=k\geq 1} \exp\Big(-L(T) + |T|(d+1)\log(2ne/(d+1))\Big) \leq 1,$$

if $L(T) \geq 2|T|(d+1)\log(2en/(d+1)) \geq |T|(\log(2) + (d+1)\log(2ne/(d+1)))$. Thus, a penalty equal to $L(T) \coloneqq 2|T|(d+1)\log(2en/(d+1)) \geq |T|(\log(2)+(d+1)\log(2ne/(d+1)))$ satisfies Kraft's inequality, i.e., $\sum_{T\in\mathcal{T}} e^{-L(T)} \leq 1$. Observe also that $\mathcal{T}$ is invariant with respect to permutations of the data points on the grid $\{\mathbf{X}_i\}_{i=1}^n \cup \{\mathbf{X}'_i\}_{i=1}^n$. By Lemma 2.1 in [10], for all $u \geq 0$,

$$\mathbb{P}\left(\max_{T\in\mathcal{T}} \frac{\overline{\mathrm{Err}}_n(\widehat{Y}(T)) - \overline{\mathrm{err}}_n(\widehat{Y}(T))}{u + \frac{\gamma L(T)}{n} + \frac{1}{2\gamma}V^2(\widehat{Y}(T))} < 1\right) \geq 1 - \exp\Big(-\frac{nu}{\gamma}\Big), \tag{23}$$

where $V^2(\widehat{Y}(T)) = \frac{1}{n}\sum_{i=1}^n ((Y'_i - \widehat{Y}(\mathbf{X}'_i))^2 - (Y_i - \widehat{Y}(\mathbf{X}_i))^2)^2$. Using the fact that $V^2(\widehat{Y}(T)) \leq 8B^2(\overline{\mathrm{Err}}_n(\widehat{Y}(T)) + \overline{\mathrm{err}}_n(\widehat{Y}(T)))$ and $\widehat{T} \in \mathcal{T}$, and choosing $u = \frac{\gamma\log(1/\delta^2)}{n}$, we find that the event whose probability is lower bounded in (23) is contained in the event

$$\overline{\mathrm{Err}}_n(\widehat{Y}(\widehat{T})) < \frac{\gamma+4B^2}{\gamma-4B^2} R_\alpha(\widehat{Y}(\widehat{T})) + \frac{\gamma^2}{\gamma-4B^2}\frac{\log(1/\delta^2)}{n},$$

which occurs with probability at least $1-\delta^2$. Next, we use a truncation argument to obtain high probability bounds in terms of $\mathrm{Err}(\widehat{Y}(\widehat{T}))$. To this end, we have

$$\begin{aligned}
&\mathrm{Err}(\widehat{Y}(\widehat{T})) = \mathbb{E}_{\mathcal{D}'_n}[\overline{\mathrm{Err}}_n(\widehat{Y}(\widehat{T}))] \\
&\leq \frac{\gamma+4B^2}{\gamma-4B^2} R_\alpha(\widehat{Y}(\widehat{T})) + \frac{\gamma^2}{\gamma-4B^2}\frac{\log(1/\delta^2)}{n} + \\
&\quad 4B^2\mathbb{P}_{\mathcal{D}'_n}\Big(\overline{\mathrm{Err}}_n(\widehat{Y}(\widehat{T})) > \frac{\gamma+4B^2}{\gamma-4B^2} R_\alpha(\widehat{Y}(\widehat{T})) + \frac{\gamma^2}{\gamma-4B^2}\frac{\log(1/\delta^2)}{n}\Big),
\end{aligned}$$

where we used the fact that $\overline{\mathrm{Err}}_n(\widehat{Y}(\widehat{T})) \leq 4B^2$. By Markov's inequality and (23), the random variable $\mathbb{P}_{\mathcal{D}'_n}\Big(\overline{\mathrm{Err}}_n(\widehat{Y}(\widehat{T})) > \frac{\gamma+4B^2}{\gamma-4B^2} R_\alpha(\widehat{Y}(\widehat{T})) + \frac{\gamma^2}{\gamma-4B^2}\frac{\log(1/\delta^2)}{n}\Big)$ is at least δ with probability less than δ.

Hence, with probability at least $1-\delta$,

$$\mathrm{Err}(\widehat{Y}(\widehat{T})) \leq \frac{\gamma+4B^2}{\gamma-4B^2} R_\alpha(\widehat{Y}(\widehat{T})) + \frac{\gamma^2}{\gamma-4B^2}\frac{\log(1/\delta^2)}{n} + 4B^2\delta.$$

The conclusion of the theorem follows from the definition of $\widehat{T}$ as a minimizer of $R_\alpha(\widehat{Y}(T))$ and choosing $\gamma = 12B^2$. □

Proof of Theorem 2. The identity (9) is shown by first noting that, in the special case of uniform $\mathbf{X}$, the probability $\mathbb{P}(X \leq s^* \mid \mathbf{X} \in \mathrm{t})$ from Lemma 6 is equal to $(s^*-a)/(b-a)$. Finally, rearranging the resulting expression yields the desired identity. □

Proof of Lemma 1. We first prove (10). The training error in t after splitting at a point is

$$\begin{aligned}\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}}(Y_i-\widetilde{Y}_i)^2 &= \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}_L}(Y_i-\overline{Y}_{\mathrm{t}_L})^2+\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}_R}(Y_i-\overline{Y}_{\mathrm{t}_R})^2\\ &= \widehat{\Delta}(\mathrm{t})\bigg(1-\frac{\widehat{\Delta}(\hat{s},\mathrm{t})}{\widehat{\Delta}(\mathrm{t})}\bigg)\\ &= \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}}(Y_i-\overline{Y}_{\mathrm{t}})^2\times(1-\widehat{\rho}^2(\widehat{Y},Y\mid\mathbf{X}\in\mathrm{t})),\end{aligned}$$

where the last equality follows from Lemma 4. Finally, $1-\widehat{\rho}^2(\widehat{Y},Y\mid\mathbf{X}\in\mathrm{t})\le\exp(-\widehat{\rho}^2(\widehat{Y},Y\mid\mathbf{X}\in\mathrm{t}))$ follows from $1-z\le e^{-z}$ for $z\ge 0$. To show (11), we use (10) recursively together with the identity

$$\overline{\mathrm{err}}_n(\widehat{Y}(T_K))=\sum_{\mathrm{t}}\widehat{P}(\mathrm{t})\widehat{\Delta}(\mathrm{t}),$$

where the sum extends over all terminal nodes t of T_K. □

Proof of Theorem 3. The proof is based on recursing the inequality

$$\mathrm{VAR}(\widehat{Y}^*\mid\mathbf{X}\in\mathrm{t})\le\mathrm{VAR}(Y\mid\mathbf{X}\in\mathrm{t})\exp(-\rho^2(\widehat{Y}^*,Y\mid\mathbf{X}\in\mathrm{t})),$$

which is the infinite sample analog of (10) in Lemma 1. Using the relation $\mathbb{E}_{\mathbf{X}}[\mathrm{Bias}^2(\widehat{Y}(T_k^*,\mathbf{X}))]=\sum_{\mathrm{t}}P(\mathrm{t})\mathrm{VAR}(Y\mid\mathbf{X}\in\mathrm{t})$, the sum extending over all terminal nodes t of T_K, yields

$$\mathbb{E}_{\mathbf{X}}[\mathrm{Bias}^2(\widehat{Y}(T_K^*,\mathbf{X}))]\le\mathrm{VAR}(Y)\exp(-K\min_{\mathrm{t}}\rho^2(\widehat{Y}^*,Y\mid\mathbf{X}\in\mathrm{t})),$$

where the minimum extends over all internal nodes of t of T_K. Next, we employ a technique similar to establishing (29) in the proofs of Lemma 3 and Lemma 2 to lower bound each $\rho^2(\widehat{Y}^*,Y\mid\mathbf{X}\in\mathrm{t})$—that is, for each function $g(\cdot)$ of X and node t,

$$\rho^2(\widehat{Y}^*,Y\mid\mathbf{X}\in\mathrm{t})\ge\Lambda\times\rho^2(g(X),Y\mid\mathbf{X}\in\mathrm{t}),\tag{24}$$

where

$$\Lambda:=\frac{\mathrm{VAR}(g(X)\mid X\in[a,b])}{\big(\int_a^b|g'(s)|\sqrt{\frac{s-a}{b-a}\frac{b-s}{b-a}}ds\big)^2}.$$

Now, (24) is valid for any function $g_j(\cdot)$ of any coordinate X_j and so we can consider the maximum over all such functions. By the infinite sample analog of Theorem 5, we have $\max_{j\in\mathcal{S}}\rho^2(g_j(X_j),Y\mid\mathbf{X}\in\mathrm{t})\ge\frac{\rho^2(f(\mathbf{X}),Y\mid\mathbf{X}\in\mathrm{t})}{d_0}$, where $Y=f(\mathbf{X})=\sum_{j\in\mathcal{S}}g_j(X_j)$ and $\#\mathcal{S}=d_0$, and hence

$$\rho^2(\widehat{Y}^*,Y\mid\mathbf{X}\in\mathrm{t})\ge\frac{\Lambda\rho^2(f(\mathbf{X}),Y\mid\mathbf{X}\in\mathrm{t})}{d_0}=\frac{\Lambda}{d_0}.\tag{25}$$

Note next that Λ is continuous and strictly positive for all $a<b$ and, furthermore by Lemma 7,

$$\inf_c\liminf_{(a,b)\to(c,c)}\Lambda=\Omega(1/R),$$

where $R = \sup_{c\in[0,1]} \inf_{r\geq 1}\{r : g^{(r)}(\cdot) \text{ is nonzero and continuous at } c\}$—which means that $\inf_{(a,b)} \Lambda > 0$. Note that R is finite if $g(\cdot)$ admits a power series representation. The result follows from taking the minimum of $\inf_{(a,b)} \Lambda$ over all $g_j(\cdot)$—each of which has finite R—which results in a positive quantity that depends only on each $g_j(\cdot)$.

Remark 3. *The same inequality (25) also holds for* $\liminf_n \widehat{\rho}^{\,2}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})$*, since by definition of* $\widehat{Y}$ *and the law of large numbers,*

$$\liminf_n \widehat{\rho}^{\,2}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \geq \liminf_n \widehat{\rho}^{\,2}(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t}) \overset{a.s.}{=} \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t}).$$

□

Proof of Lemma 3 and Lemma 2. Let $g(\cdot)$ be any function of a generic coordinate X and assume that the datapoints in the node are labeled for simplicity as $\{X_i : \mathbf{X} \in \mathrm{t}\} = \{X_1, X_2, \ldots, X_{N(\mathrm{t})}\}$. Without loss of generality, we can assume that $g(\cdot)$ linearly interpolates between the values $g(X_1), g(X_2), \ldots, g(X_{N(\mathrm{t})})$. We look at the (empirical Bayesian) prior Π on splits $s \in [0,1]$ with density

$$\frac{d\Pi(s)}{ds} = \frac{|g'(s)|\sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)}}{\int_0^1 |g'(s')|\sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)}ds'},$$

where we remind the reader that $\widehat{P}(\mathrm{t}_L) = 1 - \widehat{P}(\mathrm{t}_R) = \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} \mathbf{1}_{\{X_i \leq s\}}$. From the piecewise linear form of $g(\cdot)$, observe that Π has a piecewise constant density with knots at the data points. Since, by definition, $\widehat{Y}$ maximizes $s \mapsto \widehat{\rho}^{\,2}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t})$ and a maximum is larger than an average, we have

$$\begin{aligned}\widehat{\rho}^{\,2}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) &= \max_s \widehat{\rho}^{\,2}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \\ &\geq \int_0^1 \widehat{\rho}^{\,2}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) d\Pi(s) = \int_0^1 \frac{\widehat{\Delta}(s,\mathrm{t})}{\Delta(\mathrm{t})} d\Pi(s), \end{aligned} \tag{26}$$

where the last equality follows from Lemma 4. Next, note that the reduction in impurity admits the form

$$\widehat{\Delta}(s,\mathrm{t}) = \left(\frac{1}{\sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)}}\left(\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} (\mathbf{1}_{\{s<X_i\}} - \widehat{P}(\mathrm{t}_R))(Y_i - \overline{Y}_{\mathrm{t}})\right)\right)^2, \tag{27}$$

and, furthermore, integrating inside the square in (27) against $g'(s)\sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)}$, we have

$$\begin{aligned}&\int_0^1 g'(s)\left(\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} (\mathbf{1}_{\{s<X_i\}} - \widehat{P}(\mathrm{t}_R))(Y_i - \overline{Y}_{\mathrm{t}})\right)ds \\ &= \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i\in\mathrm{t}} (g(X_i) - \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_{i'}\in\mathrm{t}} g(X_{i'}))(Y_i - \overline{Y}_{\mathrm{t}}) \\ &= \widehat{\mathrm{COV}}(g(X), Y \mid \mathbf{X} \in \mathrm{t}). \end{aligned} \tag{28}$$

Using the inequality (26) together with the identities (27) and (28) and Jensen's inequality for the square function, we have

$$\begin{aligned}\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) &\geq \int_0^1 \frac{\widehat{\Delta}(s,\mathrm{t})}{\Delta(\mathrm{t})} d\Pi(s) \\ &\geq \frac{\widehat{\mathrm{VAR}}(g(X) \mid \mathbf{X} \in \mathrm{t})}{\big(\int_0^1 |g'(s)| \sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)} ds\big)^2} \widehat{\rho}^2(g(X), Y \mid \mathbf{X} \in \mathrm{t}). \end{aligned} \tag{29}$$

Therefore, from (29), we are led to determine how small the ratio

$$\frac{\widehat{\mathrm{VAR}}(g(X) \mid \mathbf{X} \in \mathrm{t})}{\big(\int_0^1 |g'(s)| \sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)} ds\big)^2}. \tag{30}$$

can be, ideally in terms of some simple structural characteristics of $g(\cdot)$. Our next task is to rewrite (30) in terms of the successive differences of $g(\cdot)$ at the (ordered) input data points. To this end, let τ be a permutation of the data in the node such that $X_{\tau(1)} \leq X_{\tau(2)} \leq \cdots \leq X_{\tau(N(\mathrm{t}))}$ and observe that

$$\begin{aligned} &\int_0^1 |g'(s)| \sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)} ds \\ &= \sum_{i=0}^{N(\mathrm{t})} \int_{N(\mathrm{t})\widehat{P}(\mathrm{t}_L)=i} |g'(s)| \sqrt{\frac{i}{N(\mathrm{t})}\left(1 - \frac{i}{N(\mathrm{t})}\right)} ds \\ &= \sum_{i=1}^{N(\mathrm{t})-1} \int_{X_{\tau(i)}}^{X_{\tau(i+1)}} |g'(s)| ds \sqrt{\frac{i}{N(\mathrm{t})}\left(1 - \frac{i}{N(\mathrm{t})}\right)} \\ &= \frac{1}{N(\mathrm{t})} \sum_{i=1}^{N(\mathrm{t})-1} |g(X_{\tau(i+1)}) - g(X_{\tau(i)})| \sqrt{i(N(\mathrm{t}) - i)}, \end{aligned} \tag{31}$$

where the penultimate equality follows from the fact that $\widehat{P}(\mathrm{t}_L) = i/N(\mathrm{t})$ if and only if $X_{\tau(i)} \leq s < X_{\tau(i+1)}$. Moreover, we can rewrite the variance $\widehat{\mathrm{VAR}}(g(X) \mid \mathbf{X} \in \mathrm{t})$ in a similar form, viz.,

$$\begin{aligned} \widehat{\mathrm{VAR}}(g(X) \mid \mathbf{X} \in \mathrm{t}) &= \frac{1}{2N^2(\mathrm{t})} \sum_{\mathbf{X}_{i'}, \mathbf{X}_i \in \mathrm{t}} (g(X_i) - g(X_{i'}))^2 \\ &= \frac{1}{N^2(\mathrm{t})} \sum_{i' \leq i} (g(X_{\tau(i)}) - g(X_{\tau(i')}))^2 \\ &= \frac{1}{N^2(\mathrm{t})} \sum_{i' \leq i} \Big(\sum_{k=i'}^{i-1} (g(X_{\tau(k+1)}) - g(X_{\tau(k)}))\Big)^2. \end{aligned} \tag{32}$$

To obtain the best lower bound on the ratio (30), we attempt to solve the program

$$\min_{g \in \mathcal{G}} \frac{\widehat{\mathrm{VAR}}(g(X) \mid \mathbf{X} \in \mathrm{t})}{\big(\int_0^1 |g'(s)| \sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)} ds\big)^2}, \tag{33}$$

where $\mathcal{G}$ is a collection of functions. In light of the expressions (31) and (32), the program (33) is equivalent to the following. Let $\{c_i\}$ be a concave sequence of positive

numbers, i.e., $c_i \geq (c_{i-1} + c_{i+1})/2$. Consider:

$$\begin{aligned} \min_{\mathbf{a} \in \mathcal{A}} \quad & \sum_{1 \leq i' \leq i \leq N} \Big(\sum_{k=i'}^{i-1} a_k \Big)^2 \\ \text{s.t.} \quad & \sum_{i=1}^{N-1} c_i |a_i| = 1, \end{aligned} \tag{34}$$

where $\mathcal{A}$ is a collection of vectors in $\mathbb{R}^{N(\mathrm{t})-1}$. We associate c_i with $\sqrt{i(N(\mathrm{t})-i)}$ and a_i with $g(X_{\tau(i+1)}) - g(X_{\tau(i)})$, or conversely, we associate $g(X_{\tau(i)}) - g(X_{\tau(1)})$ with $\sum_{1 \leq i' \leq i-1} a_{i'}$. In order to incorporate structural properties of $g(\cdot)$, we will need to impose conditions on $\mathcal{G}$, or equivalently, on $\mathcal{A}$. However, not all specifications make the program tractable to solve, or even convex. As a compromise between the two, we fix the sparsity set and coordinate signs of $\mathbf{a}$ in advance. That is, we specify two sets where $a_i = 0$, $a_i > 0$, and $a_i < 0$—corresponding to locations where $g(\cdot)$ is constant, increasing, and decreasing, respectively—and solve the resulting (quadratic) program. It turns out that we can obtain the exact minimizer in closed form and use it to solve the program—see (56) from Lemma 8 in Appendix B. In the notation of Lemma 8, N is the number of data points in the node, V is the number of times $g(\cdot)$ is changes from increasing to decreasing or vice versa, $M+1$ is the number of constant pieces of $g(\cdot)$, and D is the smallest number of data points contained in each constant piece with at least one data point. Lemma 3 follows immediately from (57) by noting that, in this case, $V = 0$. Lemma 2 follows similarly since, $D \geq 1$ and $M \leq N$. □

Proof of Theorem 5. Before we proceed with proving the correlation lower bound (18), we first establish some shorthand notation. Let $\widehat{\sigma}_h^2(\mathrm{t})$ denote the empirical variance of a function $h(\mathbf{X})$ in t, i.e., $\widehat{\sigma}_h^2(\mathrm{t}) = \widehat{\mathrm{VAR}}(h(\mathbf{X}) \mid \mathbf{X} \in \mathrm{t})$. Let $g(\mathbf{X}) = \sum_{j \in S} g_j(X_j)$, where each g_j is monotone. Define the discrete prior $\pi(j)$ on the index j by

$$\pi(j) = \frac{\widehat{\sigma}_{g_j}(\mathrm{t})}{\sum_{j' \in S} \widehat{\sigma}_{g_{j'}}(\mathrm{t})}, \quad j = 1, 2, \ldots, d.$$

We are now in a position to prove (18). Since a maximum is greater than an average (with respect to the coordinate index j), we have

$$\begin{aligned} |\widehat{\rho}\,(\hat{g}(X_j), Y \mid \mathbf{X} \in \mathrm{t})| &= \sup_{g \in \mathcal{G},\ j=1,2,\ldots,d} |\widehat{\rho}\,(g(X_j), Y \mid \mathbf{X} \in \mathrm{t})| \\ &\geq \sum_{j \in S} \pi(j) |\widehat{\rho}\,(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})|. \end{aligned}$$

Jensen's inequality for the absolute value function yields

$$\begin{aligned} \sum_{j \in S} \pi(j) |\widehat{\rho}\,(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})| &\geq |\sum_{j \in S} \pi(j) \widehat{\rho}\,(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})| \\ &= \frac{\widehat{\sigma}_g(\mathrm{t})}{\sum_{j' \in S} \widehat{\sigma}_{g_{j'}}(\mathrm{t})} |\widehat{\rho}\,(g(\mathbf{X}), Y \mid \mathbf{X} \in \mathrm{t})|. \end{aligned} \tag{35}$$

Next, by assumption, the components of $g(\cdot)$ have nonnegative empirical correlation with each other in the node, which ensures that $\widehat{\sigma}_g^2(\mathrm{t}) \geq \sum_{j \in S} \widehat{\sigma}_{g_j}^2(\mathrm{t})$. Combining this with (35) shows that $\widehat{\rho}^{\,2}(\hat{g}(X_j), Y \mid \mathbf{X} \in \mathrm{t})$ is at least

$$\frac{\sum_{j \in S} \widehat{\sigma}_{g_j}^2(\mathrm{t})}{(\sum_{j' \in S} \widehat{\sigma}_{g_{j'}}(\mathrm{t}))^2} \widehat{\rho}^{\,2}(g(\mathbf{X}), Y \mid \mathbf{X} \in \mathrm{t}) \geq \frac{\widehat{\rho}^{\,2}(g(\mathbf{X}), Y \mid \mathbf{X} \in \mathrm{t})}{\#S}, \tag{36}$$

where the last inequality follows from the Cauchy-Schwarz inequality. □

Proof of Theorem 4. We first show that

$$\overline{\mathrm{err}}_n(\widehat{Y}(T_K)) \le \widehat{\sigma}_Y^2 \exp\Big(- \widehat{\rho}^{\,2} \sum_{k=1}^{K} (\log_2(4N_k))^{-1} \Big). \tag{37}$$

By (10) in Lemma 1, the training error in the node is decreased by a factor of $\exp(-\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}))$ each time the node is split. By Lemma 3, with probability one, $\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \ge \frac{1}{1+\log(2N(\mathrm{t}))} \times \widehat{\rho}^{\,2} \ge \frac{1}{\log_2(4N(\mathrm{t}))} \times \widehat{\rho}^{\,2} \ge \frac{1}{\log_2(4N_k)} \times \widehat{\rho}^{\,2}$, if t is a node at level k. Thus, the training error at level $k+1$ is at most $\exp(-\widehat{\rho}^2(\log_2(4N_k))^{-1})$ times the training error at level k—in other words, the training error is geometrically decreasing. The proof can then be completed using an induction argument, noting that the training error at the root node is simply $\widehat{\sigma}_Y^2$.

For the training error bound (16), we use the inequality $\sum_{k=1}^{K} \frac{1}{\log_2(4An)-k} \ge \log\Big(\frac{\log_2(4An)}{\log_2(4An)-K}\Big)$ for integers $K \ge 1$. By (37), if T_K is a fully grown tree of depth K, we have

$$\begin{aligned}
\overline{\mathrm{err}}_n(\widehat{Y}(T_K)) &\le \widehat{\sigma}_Y^2 \exp\Big(- \widehat{\rho}^{\,2} \sum_{k=1}^{K} (\log_2(4N_k))^{-1} \Big) \\
&\le \widehat{\sigma}_Y^2 \exp\Big(- \widehat{\rho}^{\,2} \sum_{k=1}^{K} \frac{1}{\log_2(4An) - k} \Big) \\
&\le \widehat{\sigma}_Y^2 \Big(1 - \frac{K}{\log_2(4An)} \Big)^{\widehat{\rho}^{\,2}}.
\end{aligned} \tag{38}$$

Next, we show (17), i.e., the bound on the prediction error. By Theorem 1, with high probability, the leading behavior of the test error $\mathrm{Err}(\widehat{Y}(\widehat{T}))$ is governed by

$$\inf_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T)), \tag{39}$$

where the temperature α is $\Theta((d/n)\log(n/d))$. Note that (39) is smaller than the minimum of $R_\alpha(\widehat{Y}(T_K)) = \overline{\mathrm{err}}_n(\widehat{Y}(T_K)) + \alpha|T_K|$ over all fully grown trees T_K of depth K with $|T_K| \le 2^K$, i.e.,

$$\inf_{K \ge 1} \{\overline{\mathrm{err}}_n(\widehat{Y}(T_K)) + \alpha 2^K\}. \tag{40}$$

Combining the training error bound (38) with (40), we are led to optimize

$$\widehat{\sigma}_Y^2 \Big(1 - \frac{K}{\log_2(4An)} \Big)^{\widehat{\rho}^{\,2}} + \alpha 2^K, \tag{41}$$

over $K \ge 1$, although suboptimal choices of K will suffice for our purposes. Choosing K to equal $\log_2\big(\frac{\widehat{\sigma}_Y^2 (\log_2(4An))^{-\widehat{\rho}^{\,2}}}{\alpha}\big)$, we find that (41) is equal to

$$\begin{aligned}
&\widehat{\sigma}_Y^2 \Big(\frac{\log_2(4An\alpha(\log_2(4An))^{\widehat{\rho}^{\,2}}/\widehat{\sigma}_Y^2)}{\log_2(4An)} \Big)^{\widehat{\rho}^{\,2}} + \widehat{\sigma}_Y^2 \Big(\frac{1}{\log_2(4An)} \Big)^{\widehat{\rho}^{\,2}} \\
&= \mathcal{O}\Big(\widehat{\sigma}_Y^2 \Big(\frac{\log((d/\widehat{\sigma}_Y^2)(\log(n))^2)}{\log(n)} \Big)^{\widehat{\rho}^{\,2}} \Big).
\end{aligned}$$

Finally, by Theorem 1, we have the test error bound (17). □

Proof of Theorem 6. As with the proofs of Lemma 3 and Lemma 2, we use (29), the equivalence between the programs (33), and (34), and the solution to (34) in Lemma 8 from Appendix B. Our assumption on $g(\cdot)$ corresponds to, in the notation of Lemma 8, $D \geq q \times N(\mathrm{t})$ and $V < M \leq 1/q - 1$. Therefore, in each node t of T_K,

$$|\widehat{\rho}\,(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})| \geq q \times |\widehat{\rho}\,(g(X), Y \mid \mathbf{X} \in \mathrm{t})| = q.$$

Next, using a similar inductive argument to the proof of (37), we have that

$$\overline{\mathrm{err}}_n(\widehat{Y}(T_K)) \leq \widehat{\sigma}_Y^2 \exp(-Kq^2),$$

which proves (20). To show (21), we use the same strategy as the proof of Theorem 4; that is, we optimize

$$\widehat{\sigma}_Y^2 \exp(-Kq^2) + \alpha 2^K, \tag{42}$$

which upper bounds $\inf_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T))$ for all $K \geq 1$. Choosing $K = \frac{1}{q^2 + \log 2} \log(\widehat{\sigma}_Y^2/\alpha)$ sets the value of (42) at

$$2\widehat{\sigma}_Y^2 (\alpha/\widehat{\sigma}_Y^2)^{\frac{q^2}{q^2+\log 2}}.$$

Thus, using that $\alpha = \Theta((d/n)\log(n/d))$, by Theorem 1, with probability at least $1 - \delta$,

$$\mathrm{Err}(\widehat{Y}(\widehat{T})) = \mathcal{O}\bigg(\widehat{\sigma}_Y^2 \bigg(\frac{d \log(n/d)}{n \widehat{\sigma}_Y^2}\bigg)^{\frac{q^2}{q^2+\log 2}} + \frac{\log(1/\delta)}{n} + \delta\bigg),$$

which proves the second claim (21). □

B Supplementary lemmas

Henceforth, we let $p(\mathrm{t}_L) = \frac{\partial}{\partial s}\mathbb{P}(X \leq s \mid \mathbf{X} \in \mathrm{t})$ and $G(s) = \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X = s] - \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t}]$.

Lemma 6. *Suppose the density of* $\mathbf{X}$ *never vanishes and* $\Delta(s^*, \mathrm{t}) > 0$. *Then the conditional probability of the left daughter node along the splitting variable, i.e.,* $\mathbb{P}(X \leq s^* \mid \mathbf{X} \in \mathrm{t})$*, has the form*

$$\frac{1}{2} \ \pm \ \frac{1}{2}\sqrt{\frac{v}{v + \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})}}, \tag{43}$$

where $v = \frac{(\mathbb{E}[Y|\mathbf{X}\in \mathrm{t},\ X=s^*] - \mathbb{E}[Y|\mathbf{X}\in\mathrm{t}])^2}{VAR(Y|\mathbf{X}\in\mathrm{t})}$.

Proof. Recall from (4) (albeit, the infinite sample version) that one can write

$$\Delta(s, \mathrm{t}) = P(\mathrm{t}_L)P(\mathrm{t}_R)(\mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X \leq s] - \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X > s])^2. \tag{44}$$

Next, define

$$\Xi(s) = P(\mathrm{t}_L)P(\mathrm{t}_R)(\mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X \leq s] - \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X > s]),$$

so that

$$\Delta(s,\mathrm{t}) = |\Xi(s)|^2/(P(\mathrm{t}_L)P(\mathrm{t}_R)). \tag{45}$$

An easy calculation shows that

$$\frac{\partial}{\partial s}\Xi(s) = p(\mathrm{t}_L)(\mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X = s] - \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t}]) = p(\mathrm{t}_L)G(s), \tag{46}$$

where $p(\mathrm{t}_L) = \frac{\partial}{\partial s}\mathbb{P}(X \le s \mid \mathbf{X} \in \mathrm{t})$ and $G(s) = \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X = s] - \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t}]$.

Taking the derivative of $\Delta(s,\mathrm{t})$ with respect to s, we find that

$$\frac{\partial}{\partial s}\Delta(s,\mathrm{t}) = \frac{\Xi(s)p(\mathrm{t}_L)(2P(\mathrm{t}_L)P(\mathrm{t}_R)G(s) - \Xi(s)(1-2P(\mathrm{t}_L)))}{(P(\mathrm{t}_L)P(\mathrm{t}_R))^2}. \tag{47}$$

Suppose s^* is a global maximizer of (45) (in general, it need not be unique). Then a necessary condition (first-order optimality condition) is that the derivative of $\Delta(s,\mathrm{t})$ is zero at s^*. That is, from (47), s^* satisfies

$$\Xi(s^*)p(\mathrm{t}_L^*)(2P(\mathrm{t}_L^*)P(\mathrm{t}_R^*)G(s^*) - \Xi(s^*)(1-2P(\mathrm{t}_L^*))) = 0. \tag{48}$$

By assumption, $p(\mathrm{t}_L^*) > 0$ (since the density of $\mathbf{X}$ never vanishes) and $\Delta(s^*,\mathrm{t}) > 0$. It follows from rearranging (48) that

$$P(\mathrm{t}_L^*) = \frac{1}{2} - \frac{\mathrm{sgn}(\Xi(s^*)) \times G(s^*)}{\sqrt{\Delta(s^*,\mathrm{t})}}\sqrt{P(\mathrm{t}_L^*)P(\mathrm{t}_R^*)}. \tag{49}$$

The solution to (49) is obtained by solving a simple quadratic equation of the form $p = 1/2 \pm c\sqrt{p(1-p)}$, $0 \le p \le 1$, and noting from Lemma 4 that $\Delta(s^*,\mathrm{t}) = \Delta(\mathrm{t}) \times \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})$, which proves the identity (43). □

Lemma 7. *Suppose X is uniformly distributed on the unit interval and $R = \inf_{r\ge 1}\{r : g^{(r)}(\cdot)$ is nonzero and continuous at $c\} < \infty$. Then*

$$\liminf_{(a,b)\to(c,c)} \left\{ \frac{VAR(g(X) \mid X \in [a,b])}{\left(\int_a^b |g'(x)|\sqrt{\frac{x-a}{b-a}\frac{b-x}{b-a}}dx\right)^2} \right\} = \Omega(1/R). \tag{50}$$

Proof. Since the distribution of $(X-a)/(b-a)$ given $X \in [a,b]$ is uniform on the unit interval, the ratio in the limit infimum (50) is

$$\frac{\mathrm{VAR}(g(X(b-a)+a))}{\left((b-a)\int_0^1 |g'(x(b-a)+a)|\sqrt{x(1-x)}dx\right)^2}.$$

Let $\delta = (c-a)/(b-a)$. By a Taylor expansion of $g'(\cdot)$ and the definition of R, for fixed δ,

$$\lim_{(a,b)\to(c,c)} (b-a)^{-R}\int_0^1 |g'(x(b-a)+a)|\sqrt{x(1-x)}dx = \frac{|g^{(R)}(c)|}{(R-1)!}\int_0^1 |x-\delta|^{R-1}\sqrt{x(1-x)}dx. \tag{51}$$

For the variance, first note that

$$\mathrm{VAR}(g(X(b-a)+a)) = \int_0^1 (g(x(b-a)+a) - \int_0^1 g(x'(b-a)+a)dx')^2 dx.$$

Let $D(x)$ denote the divided difference $\frac{g(x(b-a)+a)-g(c)}{(x(b-a)+a-c)^R}$. Then, we can rewrite $(b-a)^{-R}(g(x(b-a)+a) - \int_0^1 g(x'(b-a)+a)dx')$ as

$$D(x)(x-\delta)^R - \int_0^1 D(x')(x'-\delta)^R dx'. \tag{52}$$

Next, use a Taylor expansion of $g(\cdot)$ about the point c and continuity of $g^{(R)}(\cdot)$ at c to argue that

$$\lim_{(a,b)\to(c,c)} D(x) = \frac{g^{(R)}(c)}{R!},$$

where the convergence is uniform. Therefore, for fixed δ,

$$\begin{aligned}\lim_{(a,b)\to(c,c)} (b-a)^{-2R}\mathrm{VAR}(g(X(b-a)+a)) &= \Big(\frac{g^{(R)}(c)}{R!}\Big)^2 \int_0^1 ((x-\delta)^R - \int_0^1 (x'-\delta)^R dx')^2 dx \\ &= \Big(\frac{g^{(R)}(c)}{R!}\Big)^2 \mathrm{VAR}((X-\delta)^R). \end{aligned} \tag{53}$$

Combining (51) and (53), we have that the limit infimum (50) is at least

$$\inf_\delta \frac{\mathrm{VAR}((X-\delta)^R)}{(R\int_0^1 |x-\delta|^{R-1}\sqrt{x(1-x)}dx)^2}. \tag{54}$$

Tedious calculations show that the infimum is achieved at $\delta = 1/2$ and hence (54) is $\Omega(1/R)$. □

Lemma 8. *Let $S = \{i_k\}_{1\le k\le M-1}$ and $S' \subset S$ be two subsets of $\{1,2,\dots,N-1\}$. Let $\mathcal{A} = \{\mathbf{a}\in\mathbb{R}^{N-1} : a_i = 0 \text{ for } i\notin S,\ a_i > 0 \text{ for } i\in S',\ a_i < 0 \text{ for } i\notin S'\}$. Consider the quadratic program*

$$\begin{aligned} \min_{\mathbf{a}\in\mathcal{A}} \quad & \sum_{1\le i'\le i\le N} \Big(\sum_{k=i'}^{i-1} a_k\Big)^2 \\ s.t. \quad & \sum_{i=1}^{N-1} c_i|a_i| = 1. \end{aligned} \tag{55}$$

Let $D_k = i_k - i_{k-1}$, $D_1 = i_1$, $D_M = N - i_{M-1}$, $c_{j_0} = 0$, $c_{i_M} = 0$, and $b_{i_k} = c_{i_k} sgn(a_{i_k})$. The unique solution $\mathbf{a}^*$ to (55) is proportional to

$$a^*_{i_k} \propto \frac{b_{i_k} - b_{i_{k-1}}}{D_k} - \frac{b_{i_{k+1}} - b_{i_k}}{D_{k+1}}, \quad k = 1,2,\dots,M-1, \tag{56}$$

where the constant of proportionality is $\big(\sum_{k=1}^M \frac{(b_{i_k}-b_{i_{k-1}})^2}{D_k}\big)^{-1}$. Furthermore, if V is the number of sign changes in the successive coordinates of $\mathbf{a}^*$, then

$$\sum_{i'\le i}\Big(\sum_{k=i'}^{i-1} a^*_k\Big)^2 = \Big(\sum_{k=1}^M \frac{(b_{i_k}-b_{i_{k-1}})^2}{D_k}\Big)^{-1} \ge \frac{1}{D^{-1}VN + M\wedge(1+\log(2N))}, \tag{57}$$

where $D = \min_{2\le k\le M-1} D_k$.

Proof. Let $u(\mathbf{a}) = \sum_{i' \le i} \big(\sum_{k=i'}^{i-1} a_k\big)^2$ and $v(\mathbf{a}) = \sum_{i=1}^{N-1} c_i |a_i|$. Then $\frac{\partial}{\partial a_{i_k}} u(\mathbf{a}) = u_{i_k}(\mathbf{a}) = 2\sum_{i=1}^{N-1} \min\{i_k, i\}(N - \max\{i_k, i\})a_i = 2(N-i_k)\sum_{i \le i_k} i a_i + 2i_k \sum_{i \ge i_k} (N-i)a_i$ and $\frac{\partial}{\partial a_{i_k}} v(\mathbf{a}) = c_{i_k} \mathrm{sgn}(a_{i_k})$. Hence, using the method of Lagrange multipliers, we need to solve the system

$$u_{i_k} - \lambda c_{i_k} \mathrm{sgn}(a_{i_k}) = 0, \quad k = 1, 2, \ldots, M-1, \ \lambda \in \mathbb{R}.$$

Note that

$$u_{i_k} = 2(N - i_k) \sum_{i \le i_k} i a_i + 2 i_k \sum_{i \ge i_{k+1}} (N-i) a_i,$$

and

$$\begin{aligned}
u_{i_{k+1}} &= 2(N - i_{k+1}) \sum_{i \le i_{k+1}} i a_i + 2 i_{k+1} \sum_{i \ge i_{k+2}} (N-i) a_i \\
&= 2(N - i_{k+1}) \sum_{j \le i_k} i a_i + 2 i_{k+1} \sum_{i \ge i_{k+1}} (N-i) a_i \\
&= u_{i_k} - 2D_{k+1} \sum_{i \le i_k} i a_i + 2D_{k+1} \sum_{i \ge i_{k+1}} (N-i) a_i \\
&= u_{i_k}(1 + \frac{D_{k+1}}{i_k}) - \frac{2ND_{k+1}}{i_k} \sum_{i \le i_k} i a_i \\
&= u_{i_k}(1 - \frac{D_{k+1}}{N - i_k}) + \frac{2ND_{k+1}}{N - i_k} \sum_{i \ge i_{k+1}} (N-i) a_i.
\end{aligned}$$

This means that

$$\sum_{i \le i_k} i a_i = \frac{i_k}{2ND_{k+1}} (u_{i_k}(1 + \frac{D_{k+1}}{i_k}) - u_{i_{k+1}});$$

$$\sum_{i \ge i_{k+1}} (N-i) a_i = \frac{N - i_k}{2ND_{k+1}} (u_{i_{k+1}} - u_{i_k}(1 - \frac{D_{k+1}}{N - i_k})).$$

Hence, for $k = 1, 2, \ldots, M-1$,

$$\begin{aligned}
a^*_{i_k} &= \frac{1}{i_k} \Big(\sum_{i \le j_k} i a^*_i - \sum_{i \le i_{k-1}} i a^*_i\Big) \\
&= \frac{1}{N - i_k} \Big(\sum_{i \ge i_k} (N-i) a^*_i - \sum_{i \ge i_{k+1}} (N-i) a^*_i\Big) \\
&\propto \frac{b_{i_k} - b_{i_{k-1}}}{D_k} - \frac{b_{i_{k+1}} - b_{i_k}}{D_{k+1}},
\end{aligned}$$

where the constant of proportionality is $\big(\sum_{k=1}^{M} \frac{(b_{i_k} - b_{i_{k-1}})^2}{D_k}\big)^{-1}$, which follows from $u_{i_k}(\mathbf{a}^*) = \lambda b_{i_k}$. Since $\{c_i\}$ is a concave sequence of numbers, i.e., $c_i \ge (c_{i-1} + c_{i+1})/2$, the sign of $a^*_{i_k}$ is consistent with the assumed form of b_{i_k}. Hence, the claimed $\mathbf{a}^*$ from (56) solves the program.

Next, using summation by parts, we have

$$\sum_{i' \le i} \Big(\sum_{k=i'}^{i-1} a^*_k\Big)^2 = \frac{1}{\sum_{k=1}^{M} \frac{(b_{i_k} - b_{i_{k-1}})^2}{ND_k}}. \tag{58}$$

We now specialize our analysis to $c_{i_k} = \sqrt{i_k(N-i_k)}$. Suppose that $\mathbf{a}^*$ changes sign at index i_k, i.e., $a^*_{i_{k-1}} a^*_{i_k} < 0$. Then, since $b_{i_k} = c_{i_k}\mathrm{sgn}(a_{i_k})$, we have

$$\begin{aligned}\frac{(b_{i_k} - b_{i_{k-1}})^2}{ND_k} &= \frac{(c_{i_k} + c_{i_{k-1}})^2}{ND_k} \\ &\le \frac{(c_{i_k} + c_{i_{k-1}})^2}{ND} \\ &\le \frac{N}{D},\end{aligned}$$

where the last line is from $(c_{i_k} + c_{i_{k-1}})^2 \le N^2$. Thus, it follows that (58) is at least

$$\frac{1}{D^{-1}VN + \sum_{k=1}^{M} \frac{(c_{i_k} - c_{i_{k-1}})^2}{ND_k}}, \tag{59}$$

V is the number of sign changes in the successive coordinates of $\mathbf{a}^*$. We now obtain an upper bound for

$$\sum_{k=1}^{M} \frac{(c_{i_k} - c_{i_{k-1}})^2}{ND_k} = \sum_{k=1}^{M} \frac{D_k(N - i_k - i_{k-1})^2}{N(\sqrt{i_k(N-i_k)} + \sqrt{i_{k-1}(N-i_{k-1})})^2}. \tag{60}$$

Let $k^* = \min\{k : i_k + i_{k-1} \ge N\}$. Then, $(\sqrt{i_k(N-i_k)} + \sqrt{i_{k-1}(N-i_{k-1})})^2 \ge (2N - i_k - i_{k-1})(i_k + i_{k-1} - N)$ for all $k \ge k^*$. Thus, the sum $\sum_{k \ge k^*} \frac{D_k(N-i_k-i_{k-1})^2}{N(\sqrt{i_k(N-i_k)}+\sqrt{i_{k-1}(N-i_{k-1})})^2}$ is at most

$$\sum_{k \ge k^*} \frac{D_k}{2N - i_k - i_{k-1}} \left(\frac{i_{k-1} + i_k}{N} - 1\right) \le \sum_{k \ge k^*} \frac{i_k - i_{k-1}}{2N - i_k - i_{k-1}}, \tag{61}$$

where we used the fact that $D_k = i_k - i_{k-1}$. Next, $(\sqrt{i_k(N-i_k)} + \sqrt{i_{k-1}(N-i_{k-1})})^2 \ge (i_k + i_{k-1})(N - i_k - i_{k-1})$ for all $k < k^*$ and hence the sum $\sum_{k < k^*} \frac{D_k(N-i_k-i_{k-1})^2}{N(\sqrt{i_k(N-i_k)}+\sqrt{i_{k-1}(N-i_{k-1})})^2}$ is at most

$$\sum_{k < k^*} \frac{D_k}{i_k + i_{k-1}} \left(1 - \frac{i_{k-1} + i_k}{N}\right) \le \sum_{k < k^*} \frac{i_k - i_{k-1}}{i_k + i_{k-1}}. \tag{62}$$

Combining (61) and (62), we have shown that (60) is at most

$$\sum_{k < k^*} \frac{i_k - i_{k-1}}{i_k + i_{k-1}} + \sum_{k \ge k^*} \frac{i_k - i_{k-1}}{2N - i_k - i_{k-1}} \le M. \tag{63}$$

The sum (63) is largest when $M = N$, yielding

$$\sum_{k=1}^{(N-1)/2} \frac{1}{2k-1} + \sum_{k=1}^{(N+1)/2} \frac{1}{2k-1} \le 1 + \log(2N). \tag{64}$$

Combining (63) and (64) with (59) proves (57). □