Sparse Learning with CART

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 constructing a prior distribution on the split points and solving a nonlinear optimization problem. We leverage this connection between the training error and Pearson correlation to show that CART with cost-complexity pruning achieves an optimal complexity/goodness-of-fit tradeoff when the depth scales with the logarithm of the sample size. Data dependent quantities, which adapt to the dimensionality and latent structure of the regression model, 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 [1] and random forests [2]). In addition, each iteration of gradient tree boosting (e.g., TreeBoost [6]) 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 adapt to the sparsity of a predictive model?

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) [3] 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 [7] provides oracle-type inequalities for the CART pruning algorithm proposed by [3], 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 unique advantages of tree learning with CART over other unstructured regression procedures, like vanilla k-NN or other kernel based estimators. Filling this theoretical gap, our main message is that, in certain settings, CART can identify low dimensional, latent structure in the data and adapt accordingly. We illustrate the adaptive properties of CART when the model is *sparse*, namely, when the output depends only on a small, unknown subset of the input variables—thereby circumventing the curse of dimensionality.

Arguably the most difficult technical aspect of studying decision trees (and for that matter, any adaptive partitioning-based predictor) is understanding their approximation error, or at least pinning down conditions on the data that enable such an endeavor. Indeed, most existing convergence results [14] for decision trees or ensembles thereof bound the expected (over the training data) prediction error by the size (i.e., the diameter) of the terminal nodes and show that they vanish with the depth of the tree, ensuring that the approximation error does so also.[1] While this technique can be useful to prove consistency statements, it is not generally delicate enough to capture the adaptive properties of the tree on the data. It also often requires making strong assumptions about the tree construction. To address this shortcoming, in contrast, we use the fact that the prediction error is with high-probability (over the training data) bounded by the training error plus a complexity term. One of our crucial insights is that we can avoid using the node diameters as a proxy for the approximation error and, instead, directly bound the training error in terms of data dependent quantities (like the Pearson correlation coefficient) that are more transparent and interpretable, thereby facilitating our analysis and allowing us to prove more fine-grained results.

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 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 the jth coordinate index, denoted by X_j, or additionally on the ith data point, 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 and mean squared prediction error are, respectively,

$$\overline{\mathrm{err}}(\widehat{Y}) := \frac{1}{n} \sum_{i=1}^{n} (Y_i - \widehat{Y}(\mathbf{X}_i))^2 \quad \text{and} \quad \mathrm{Err}(\widehat{Y}) := \mathbb{E}_{(\mathbf{X}', Y')}[(Y' - \widehat{Y}(\mathbf{X}'))^2],$$

where $(\mathbf{X}', Y')$ denotes an independent copy of $(\mathbf{X}, Y)$. For data $\{(\mathbf{X}_1, U_1, V_1), \ldots, (\mathbf{X}_n, U_n, V_n)\}$, we let

$$\widehat{\rho}(U, V \mid \mathbf{X} \in A) := \frac{\frac{1}{N} \sum_{\mathbf{X}_i \in A} (U_i - \overline{U})(V_i - \overline{V})}{\sqrt{\frac{1}{N} \sum_{\mathbf{X}_i \in A} (U_i - \overline{U})^2 \times \frac{1}{N} \sum_{\mathbf{X}_i \in A} (V_i - \overline{V})^2}},$$

(A is a subset, $N = \#\{\mathbf{X}_i \in A\}$, $\overline{U} = \frac{1}{N} \sum_{\mathbf{X}_i \in A} U_i$, and $\overline{V} = \frac{1}{N} \sum_{\mathbf{X}_i \in A} V_i$) denote the empirical Pearson product-moment correlation coefficient, given $\mathbf{X} \in A$, and let $\rho(U, V \mid \mathbf{X} \in A)$ be its infinite sample counterpart. If $U_i = g(X_{ij})$ for a univariate function $g(\cdot)$ of a coordinate X_j, we write $\widehat{\rho}(g(X_j), V \mid \mathbf{X} \in A)$ or $\rho(g(X_j), V \mid \mathbf{X} \in A)$. For brevity, we let $\widehat{\sigma}_Y^2$ denote the sample variance of the response values $Y_1, Y_2, \ldots, Y_n$ in the training data. The rth derivative of a real valued function $g(\cdot)$ is denoted by $g^{(r)}(\cdot)$. 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) [3] 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

---

[1]Because CART outputs the average of the response values in a node, if the regression function is smooth, the approximation error is at most a constant multiple of the largest node diameter. Thus, one can use the node diameters as a proxy for the approximation error.

best overall partition, CART decision 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 generic variable X that splits t into left and right daughter nodes t_L and t_R according to whether $X \leq s$ or $X > s$. These two nodes will be denoted by $\mathrm{t}_L = \{\mathbf{X} \in \mathrm{t} : X \leq s\}$ and $\mathrm{t}_R = \{\mathbf{X} \in \mathrm{t} : X > s\}$. As mentioned previously, a tree is grown by recursively reducing node impurity. Impurity for regression trees is determined by the within node sample variance

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

where $\overline{Y}_{\mathrm{t}} = \frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} Y_i$ is the sample mean for t and $N(\mathrm{t}) = \#\{\mathbf{X}_i \in \mathrm{t}\}$ is the number of data points in t. Similarly, the within node 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 [3, 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.

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 the input coordinate index j. The output $\widehat{Y}(T)$ 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

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

As shown in [3, Section 10.2], the smallest minimizing subtree for the temperature α,

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

3

exists and is unique (smallest in the sense that if $T' \in \arg\min_{T \preceq T_{\max}} R_\alpha(\widehat{Y}(T))$, then $\widehat{T} \preceq T'$). For a fixed α, the optimal subtree $\widehat{T}$ can be found efficiently by weakest link pruning [3, 8], i.e., by successively collapsing the internal node that increases $\overline{\text{err}}(\widehat{Y}(T))$ the least, until we arrive at the single-node 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 $Y = f(\mathbf{X})$, $B = \sup_{\mathbf{x}} |f(\mathbf{x})| < \infty$, $n > (d+1)/2$, and $\alpha > \frac{27B^2(d+1)\log(2en/(d+1))}{n}$. Then, with probability at least $1 - \delta$ over the training sample $\mathcal{D}_n$,*

$$Err(\widehat{Y}(\widehat{T})) \leq 4 \min_{T \preceq T_{max}} R_\alpha(\widehat{Y}(T)) + \frac{54B^2 \log(2/\delta)}{n}.$$

Similar bounds hold for the Bayes risk for binary classification, i.e., $Y \in \{0, 1\}$, since, in this case, the squared error impurity (1) equals one-half of the so-called *Gini* impurity used for classification trees. See also [13] for results of a similar flavor when the penalty is proportional to $\sqrt{|T|}$.

In what follows, we let $T_K \preceq T_{\max}$ denote a fully grown binary tree of depth $K = \Theta(\log_2(n))$, 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 Bounding 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 $\widehat{\Delta}(s, \text{t})$ with averages in the terminal nodes permits further simplifications of the formula (2) above. For example, using the sum of squares decomposition, $\widehat{\Delta}(s, \text{t})$ can equivalently be expressed as [3, Section 9.3]

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

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 $\widehat{\Delta}(s, \text{t})$, which does not appear to have been considered in past literature and will prove to be useful for our purposes, is via its equivalent representation as $\widehat{\Delta}(\text{t}) \times \widehat{\rho}^2(\widetilde{Y}, Y \mid \mathbf{X} \in \text{t})$, where

$$\widehat{\rho}(\widetilde{Y}, Y \mid \mathbf{X} \in \text{t}) := \frac{\frac{1}{N(\text{t})} \sum_{\mathbf{X}_i \in \text{t}} (\widetilde{Y}_i - \overline{Y}_\text{t})(Y_i - \overline{Y}_\text{t})}{\sqrt{\frac{1}{N(\text{t})} \sum_{\mathbf{X}_i \in \text{t}} (\widetilde{Y}_i - \overline{Y}_\text{t})^2 \times \frac{1}{N(\text{t})} \sum_{\mathbf{X}_i \in \text{t}} (Y_i - \overline{Y}_\text{t})^2}} \geq 0 \tag{5}$$

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

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

and response variable Y within t (for the proof, see Lemma A.1 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.,

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

4

Equivalently, CART splits along variables with decision stumps that are most correlated with the residuals $Y_i - \overline{Y}_t$ of the current fit $\overline{Y}_t$. As with simple linear regression, we see that the squared correlation $\widehat{\rho}^2(\widetilde{Y}, Y \mid \mathbf{X} \in t)$ equals the *coefficient of determination* R^2, i.e., it describes the fraction of the variance in Y that is explained by a decision stump in X. Though, unlike r^2 for linear regression, the decision stump need not necessarily be a least squares fit.

Definition 1. *We let $\widehat{Y}$ denote a decision stump $\widetilde{Y}$ with an optimal direction $\hat{\jmath} \in \arg\max_{j=1,2,\ldots,d} \widehat{\Delta}(\hat{s}, t)$ and corresponding optimal split $\hat{s}$.*

It should be stressed that the alternative characterization of the splitting criterion (2) in terms of a correlation is unique to the squared error impurity with (constant) averages in the terminal nodes of the tree.

We now introduce a data dependent quantity that will play a central role in determining the rates of convergence of the prediction error. For a univariate function class $\mathcal{G}$, we let $\widehat{\rho}_{\mathcal{G}}$ be the largest Pearson correlation between the response data Y and a function in $\mathcal{G}$ of a single input coordinate for a worst-case node, i.e.,

$$\widehat{\rho}_{\mathcal{G}} := \min_{t} \sup_{g(\cdot) \in \mathcal{G}, \, j=1,2,\ldots,d} |\widehat{\rho}(g(X_j), Y \mid \mathbf{X} \in t)|, \tag{8}$$

where the minimum runs over all internal nodes t in T_K. We will specifically focus on classes $\mathcal{G}$ that consist of decision stumps, and more generally, monotone functions.

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 coordinates to split and their split points, i.e.,

$$\widehat{\Delta}(s, t) \underset{n \to \infty}{\to} \Delta(s, t) := \Delta(t) - [P(t_L)\Delta(t_L) + P(t_R)\Delta(t_R)], \tag{9}$$

where quantities without hats are the population level counterparts of the empirical quantities defined previously in (2). A decision stump (6) with an optimal theoretical direction j^* and corresponding optimal theoretical split s^* is denoted by $\widehat{Y}^*$. Now, if the number of data points within t is large and $\Delta(s, t)$ 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, [9, 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, [4, 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 an optimal theoretical split s^* and infinite sample correlation $\rho(\widehat{Y}, Y \mid \mathbf{X} \in t) \stackrel{\text{a.s.}}{:=} \lim_n \widehat{\rho}(\widehat{Y}^*, Y \mid \mathbf{X} \in t)$ for a deterministic node t (the limit exists by the uniform law of large numbers). The proof is based on the first-order optimality condition, namely, $\frac{\partial}{\partial s}\Delta(s, t)\big|_{s=s^*} = 0$.

Theorem 2. *Suppose $\mathbf{X}$ is uniformly distributed on $[0, 1]^d$ and $\Delta(s^*, t) > 0$. For a deterministic parent node t, an optimal theoretical split $s^* \in [a, b]$ along variable X has the form*

$$\frac{a+b}{2} \pm \frac{b-a}{2}\sqrt{\frac{v}{v + \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in t)}}, \tag{10}$$

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

Expression (10) in Theorem 2 reveals that an optimal theoretical split s^* is a perturbation of the median $(a + b)/2$ of the conditional distribution $X \mid \mathbf{X} \in t$, where the gap is governed by the correlation $\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in t)$. These correlations control the local and directional granularity of the partition of the input domain. Splits along input coordinates that contain a strong signal, i.e., $\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in t) \gg 0$, tend to be further away from the parent node edges, thereby producing

side lengths $[a, b]$ 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 $s^* \approx a + (b - a)\rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})/(4v)$ or $s^* \approx b - (b - a)\rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})/(4v)$—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 (10), either $s^* = \Theta(1/w)$ or $s^* = 1 - \Theta(1/w)$ (see Lemma A.2 in the supplement). This phenomenon, where optimal splits concentrate at the endpoints of the node along noisy directions, has been dubbed 'end-cut preference' in the literature and has been known empirically since the inception of CART [9], [3, Section 11.8]. The theory above is also consistent with empirical studies on the adaptive properties of Breiman's random forests which use CART [11, Section 4].

3.3 Training error and Pearson correlation

In addition to determining the location of the splits, the correlation is also directly 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, yielding an optimal decision stump $\widehat{Y}$, the training error in t is reduced by a constant factor, namely, $\exp(-\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}))$ or, uniformly, by $\exp(-\widehat{\rho}_{\mathcal{H}}^2)$, where $\mathcal{H}$ is the collection of all decision stumps (i.e., a step function with two constant pieces) and $\widehat{\rho}_{\mathcal{H}}$ is the quantity defined in (8). Recursing this contraction inequality over nodes at each level of the tree leads to the conclusion that the training error should be exponentially small in the depth K, provided the correlation at each node is large.

Lemma 1. *Almost surely,*

$$\frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} (Y_i - \widehat{Y}_i)^2 \leq \frac{1}{N(\mathrm{t})} \sum_{\mathbf{X}_i \in \mathrm{t}} (Y_i - \overline{Y}_\mathrm{t})^2 \times \exp(-\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t})), \qquad (11)$$

and hence

$$\overline{err}(\widehat{Y}(T_K)) \leq \widehat{\sigma}_Y^2 \exp(-K \times \widehat{\rho}_{\mathcal{H}}^2), \qquad (12)$$

where $\mathcal{H}$ is the collection of all decision stumps.

3.4 Size of Pearson correlation via comparison inequalities

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 model. That is, we fit an arbitrary univariate function $g(X)$ of a generic coordinate 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 an 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 not at all straightforward to disentangle the dependence on the data. Even if this could be done and a target population level quantity could be identified, it is difficult to rely on concentration of measure when t contains very few data points; 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

$$\int \widehat{\rho}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) d\Pi(j, s), \qquad (13)$$

which is much less burdensome to analyze. Importantly, the prior can involve unknown quantities from the distribution of $(\mathbf{X}, Y)$. For a special choice of prior Π, (13) can be further lower bounded by

$$\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \geq \text{constant} \times |\widehat{\rho}(g(X), Y \mid \mathbf{X} \in \mathrm{t})|. \qquad (14)$$

The constant in (14) depends on $g(\cdot)$, though importantly it is invariant to the scale of $g(\cdot)$. If $g(\cdot)$ belongs to a univariate model class $\mathcal{G}$, this constant can either be studied directly for the specific $g(\cdot)$ or minimized over $g(\cdot) \in \mathcal{G}$ to yield a more insightful lower bound. For certain model classes $\mathcal{G}$, the minimization problem turns out to be equivalent to a quadratic program, and the solution can be

obtained explicitly and used to prove the next set of results. Our first application of this technique shows that, despite fitting the data to a decision stump with *one* degree of freedom (i.e., the location of the split), CART behaves almost as if it fit the data to a monotone function with $N(t) - 1$ degrees of freedom, at the expense of a sublogarithmic factor in $N(t)$. For example, the correlation between the response variable and the decision stump is, up to a sub-logarithmic factor, at least as strong as the correlation between the response variable and a linear or isotonic fit.

Fact 1. *Almost surely, uniformly over all monotone functions $g(\cdot)$ of X in the node, we have*

$$\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in t) \geq \frac{1}{\sqrt{1 + \log(2N(t))}} \times |\widehat{\rho}(g(X), Y \mid \mathbf{X} \in t)|. \tag{15}$$

The previous fact also suggests that CART is quite good at fitting response values that have a local, low-dimensional, monotone relationship with the input variables. Note that because correlation is merely a measure of linear association, $|\widehat{\rho}(g(X), Y \mid \mathbf{X} \in t)|$ 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. On the other hand, if Y has no signal in X, then [10, Lemma 1, Supplement] show that, with high probability, $\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in t)$ is $\mathcal{O}(\sqrt{(\log N(t))/N(t)})$.

4 Main Results

In this section, we use the training error bound (12) and the device (13) for obtaining correlation comparison inequalities (à la Fact 1) to give bounds on the prediction error of CART. We first outline the high-level strategy. 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 $R_\alpha(\widehat{Y}(T_K)) = \overline{\mathrm{err}}(\widehat{Y}(T_K)) + \alpha |T_K|$ over all fully grown trees T_K of depth K with $|T_K| \leq 2^K$, i.e.,

$$\inf_{K \geq 1} \{\overline{\mathrm{err}}(\widehat{Y}(T_K)) + \alpha 2^K\}. \tag{16}$$

Coupled with an informative bound on $\overline{\mathrm{err}}(\widehat{Y}(T_K))$, (16) can then be further bounded and solved. The proofs reveal that a good balance between the tree size and its goodness of fit occurs when K is logarithmic in the sample size.
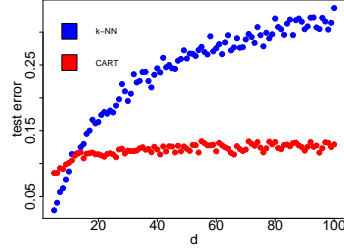
4.1 Asymptotic consistency rates for sparse additive models

Applying the training error bound (12) to (16) with $K = \lceil (\widehat{\rho}_{\mathcal{H}}^2 + \log 2)^{-1} \log(\widehat{\sigma}_Y^2/\alpha) \rceil$, we have from Theorem 1 that with probability at least $1 - \delta$,
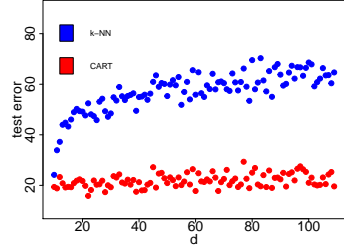
$$\mathrm{Err}(\widehat{Y}(\widehat{T})) = \mathcal{O}\Big(\widehat{\sigma}_Y^2 \Big(\frac{d \log(n/d)}{n\widehat{\sigma}_Y^2}\Big)^{\frac{\widehat{\rho}_{\mathcal{H}}^2}{\widehat{\rho}_{\mathcal{H}}^2 + \log 2}} + \frac{\log(1/\delta)}{n}\Big). \tag{17}$$

It turns out that if $\mathbf{X}$ is uniformly distributed and Y is a sparse additive model with d_0 component functions $g_j(\cdot)$, then $\widehat{\rho}_{\mathcal{H}}^2$ is asymptotically lower bounded by a constant multiple of $1/d_0$.
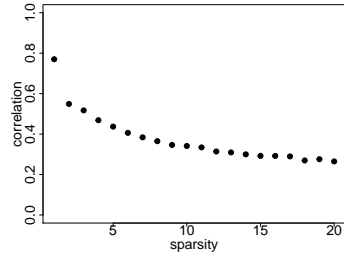
(a) Synthetic data. Prediction error of CART vs. k-NN as d varies.

(b) Boston housing data. Prediction error of CART vs. k-NN as d varies.

(c) Minimum correlation $\widehat{\rho}_{\mathcal{H}}$ (averaged over 10 independent replications) of all nodes of pruned CART tree as sparsity d_0 varies.

Thus, we find from (17) that if d_0 is fixed, then $\lim_n \mathrm{Err}(\widehat{Y}(\widehat{T})) \overset{\text{a.s.}}{=} 0$ even when the ambient dimension grows as $d = o(n)$. Note that such a statement is not possible for vanilla k-NN or other kernel based regression methods with nonadaptive weights, unless feature selection is performed beforehand. In fact, we show next that the prediction error rate that CART achieves is the same as what would be achieved by a standard kernel predictor *if* one had a priori knowledge of the locations of the d_0 relevant input variables that determine the output. A routine computer experiment

on synthetic or real data easily confirms this theory. In Fig. 1a and Fig. 1c, we generate 1000 samples from the model $Y = \sum_{j=1}^{d_0} g_j(X_j)$, where each $g_j(X_j)$ equals $\pm X_j^2$ (alternating signs) and $\mathbf{X} \sim \text{Uniform}([0,1]^d)$. In Fig. 1a, we plot the test error, averaged over 10 independent replications, of pruned CART vs. k-NN (with cross-validated k) as d ranges from 5 to 100 with $d_0 = 5$ fixed. A similar experiment is performed in Fig. 1b on the Boston housing dataset [3, Section 8.2] ($d_0 = 10$ and $n = 506$), where we scale the inputs to be in $[0,1]$ and add $d - d_0$ noisy $\text{Uniform}([0,1])$ input variables. According to Theorem 3, the convergence rate of CART depends primarily on the sparsity d_0 and therefore its performance should not be adversely affected by growing d. Consistent with our theory, the prediction error of CART remains stable as d increases, whereas k-NN does not adapt to the sparsity. Furthermore, Fig. 1c illustrates how $\widehat{\rho}_{\mathcal{H}}^2$ decays with d_0 if $d = 20$, thus corroborating with the aforementioned asymptotic behavior of $\Omega(1/d_0)$.

Theorem 3. *Suppose $\mathbf{X}$ is uniformly distributed on $[0,1]^d$ and $Y = \sum_j g_j(X_j)$ is a sparse additive model with $d_0 \ll d$ smooth component functions $g_j(\cdot)$, where each function is not too locally 'flat' in the sense that*

$$\sup_x \inf\{r \geq 1 : g_j^{(r)}(\cdot) \text{ exists, continuous, and nonzero at } x\} < \infty. \qquad (18)$$

Then there exists a constant $C > 0$ that is independent of d_0 such that, almost surely, $\liminf_n \widehat{\rho}_{\mathcal{H}}^2 \geq C/d_0$, and

$$\limsup_n \frac{Err(\widehat{Y}(\widehat{T}))}{((d/n)\log(n/d))^{\Omega(1/d_0)}} \overset{a.s.}{=} \mathcal{O}(1). \qquad (19)$$

Remark 1. *For independent, continuous marginal input variables X_j, there is no loss of generality in assuming uniform distributions in Theorem 3. Indeed, CART decision trees are invariant to strictly monotone transformations of X_j. One such transformation is the marginal cumulative distribution function $F_{X_j}(\cdot)$, for which $F_{X_j}(X_j) \sim \text{Uniform}([0,1])$—and so the problem can immediately be reduced to the uniform case.*

Any nonconstant component function $g_j(\cdot)$ that admits a power series representation satisfies the hypothesis of Theorem 3, though, in general, the condition (18) accommodates functions that are not analytic or infinitely differentiable. In fact, even differentiability is not necessary—similar results hold if the $g_j(\cdot)$ are *step functions*, as we now show. To this end, assume that $Y = \sum_j g_j(X_j)$ is an additive model, where each component function $g_j(\cdot)$ is a bounded step function and the total number of constant pieces of Y is V. We show in the supplement that each optimal split $\hat{s}$ in a node t satisfies

$$\max\{X_{ij} \in I : \mathbf{X}_i \in \text{t}\} \leq \hat{s} \leq \min\{X_{ij} \in I' : \mathbf{X}_i \in \text{t}\}, \qquad (20)$$

for some direction j and successive intervals I and I' on which $g_j(\cdot)$ is constant. For example, if $Y = c_1 \mathbf{1}_{\{X_1 < s_1\}} + c_2 \mathbf{1}_{\{X_2 < s_2\}}$, then the first split separates the data along X_1 at s_1 (resp. X_2 at s_2), and at the next level down, CART separates the data in both daughter nodes along X_2 at s_2 (resp. X_1 at s_1). Thus, in general, each empirical split always perfectly separates the data in the node between two adjacent constant values of a component function. A CART decision tree grown in this way will achieve zero training error once it has at least V terminal nodes, i.e., $|T| \geq V$. This is in fact the same training error that would result from the global least squares projection onto the space of all step functions with V constant pieces. From Theorem 1, we immediately obtain the following performance bound, which is the optimal $\mathcal{O}(1/n)$ parametric rate for prediction, up to a logarithmic factor in the sample size [16]. Notice that we do not make any assumptions on input distribution.

Theorem 4. *Suppose $Y = \sum_j g_j(X_j)$, where each component function $g_j(\cdot)$ is a bounded step function and the total number of constant pieces of Y is V. With probability at least $1 - \delta$,*

$$Err(\widehat{Y}(\widehat{T})) = \mathcal{O}\left(\frac{Vd\log(n/d)}{n} + \frac{\log(1/\delta)}{n}\right). \qquad (21)$$

4.2 Finite sample consistency rates and general sparse models

Using Fact 1, we now provide results of a similar flavor for more general regression models under a mild assumption on the largest number of data points in a node at level k in T_K, denoted by N_k. Importantly, our theory only requires that each $N(\text{t})$ is *upper* bounded at every level of the tree. This condition still 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 (including tree ensembles like random forests) that requires each $N(\text{t})$ to be *lower* bounded [3, Section 12.2], [12, 15, 5].

Assumption 1. *For some constants $a \geq 0$ and $A > 0$, the largest number of data points in a node at level k in T_K satisfies $N_k \leq Ank^a/2^k$, for $k = 1, 2, \ldots, K = \Theta(\log_2(n))$.*

Recall the quantity $\widehat{\rho}_{\mathcal{G}}$ defined in (8), namely, the largest correlation between the response data Y and a function in $\mathcal{G}$ for a worst-case node. Our next theorem shows that if $\mathcal{M}$ is the collection of all monotone (i.e., increasing or decreasing) functions, then

$$\widehat{\rho}_{\mathcal{M}} = \min_{\text{t}} \quad \sup_{g(\cdot) \text{ monotone}, \; j=1,2,\ldots,d} |\widehat{\rho}\left(g(X_j), Y \mid \mathbf{X} \in \text{t}\right)|$$

governs the rate at which the training error and prediction error decrease. Both errors are small if the local monotone dependence between $\mathbf{X}$ and Y is high; that is, if CART partitions the input domain into pieces where the response variable is locally monotone in a few of the input coordinates.

Theorem 5. *Let $Y = f(\mathbf{X})$, where $f(\cdot)$ is a bounded function. Under Assumption 1, almost surely,*

$$\overline{err}(\widehat{Y}(T_K)) \leq \widehat{\sigma}_Y^2 \left(1 - \frac{K}{\log_2(4K^a An)}\right)^{\widehat{\rho}_{\mathcal{M}}^2}. \tag{22}$$

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

$$Err(\widehat{Y}(\widehat{T})) = \mathcal{O}\left(\widehat{\sigma}_Y^2 \left(\frac{\log((d/\widehat{\sigma}_Y^2)\log^{2+a}(n))}{\log(n)}\right)^{\widehat{\rho}_{\mathcal{M}}^2} + \frac{\log(1/\delta)}{n}\right). \tag{23}$$

We will now argue that $\widehat{\rho}_{\mathcal{M}}$ is an empirical measure of the local dimensionality of Y. More specifically, we argue that if CART effectively partitions the input domain so that, in each node, Y is locally correlated with sparse additive models with $d_0 \ll d$ monotone component functions, then $\widehat{\rho}_{\mathcal{M}}^2 = \Omega(1/d_0)$. To see why this assertion is true, suppose $g_1(X_1), g_2(X_2), \ldots, g_d(X_d)$ is an arbitrary collection of d univariate functions from $\mathcal{M}$. However, suppose that only a subset of d_0 of the input variables $X_1, X_2, \ldots, X_d$ locally affect Y in each node. Then, it can be shown (see Lemma A.3 in the supplement) that there is some node t and sparse additive model Y_0 with d_0 component functions of the form $\pm g_j(X_j)$, corresponding to the d_0 input variables that locally affect Y, such that, almost surely,

$$\widehat{\rho}_{\mathcal{M}}^2 \geq \frac{\widehat{\rho}^2(Y_0, Y \mid \mathbf{X} \in \text{t})}{d_0} = \Omega(1/d_0). \tag{24}$$

The above statement is reminiscent of Theorem 3 in which $\widehat{\rho}_{\mathcal{H}}^2 = \Omega(1/d_0)$ controls the convergence rate of the prediction error when the underlying regression model is additive. Though, in contrast, note that (24) holds regardless of the dependence structure between the d_0 input coordinates that matter and the $d - d_0$ input coordinates that do not. Thus, (24) and Theorem 5 together suggest that it is possible to achieve rates of the form $(\log(d)/\log(n))^{\Omega(1/d_0)}$ in fairly general settings.

5 Extensions to Tree Ensembles

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. Establishing similar relationships is not as easy with trees that are constructed from bootstrap samples or random selections of input variables. Nevertheless, we mention a few ideas for future work. Suppose $\widehat{Y} = (1/m) \sum_T \widehat{Y}(T)$ is the output of an ensemble of m decision trees. By convexity of the squared error loss [2, Section 11] or [1, Section 4.1], we have $Err(\widehat{Y}) \leq (1/m) \sum_T Err(\widehat{Y}(T))$, where the prediction error is averaged with respect to the tree randomization mechanism. Using the law of total expectation by conditioning on each realization of the (random) tree, $Err(\widehat{Y}(T))$ can be decomposed into quantities that involve the prediction error of a fixed (non-random) tree, for which our previous results can be applied. We will leave the exact details of these extensions for future work.

6 Conclusion

A key strength of CART decision trees is that they can exploit local, low dimensionality of the model—via a built-in, automatic dimension reduction mechanism. This is particularly useful since many real-world input/output systems are locally approximated by simple model forms with only a few variables. Adaptivity with CART 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. To illustrate this ability, we identified settings where CART adapts to the unknown sparsity of the model. To the best of our knowledge, the consistency rates given here are the first of their kind for CART decision trees.

Broader Impact

Who may benefit from this research. There are at least two groups of people who will benefit—either directly or indirectly—from this research.

1. Decision makers across a variety of domains, especially those with limited training in statistics. CART has enabled data-driven decision making in multiple high-stakes domains (e.g., business, medicine, and policy) over the past three decades. In particular, those who do not have a formal quantitative background will benefit from the intuitive and interpretable nature of CART and its quick and easy implementation.

2. Members of the society who may be have faced ethical/fairness concerns associated with their data and its use. As this paper has demonstrated, CART forms predictions by emphasizing variables that are more relevant to the output. In a social science context, this suggests that CART may focus more on key diagnostic information (e.g., education, income) without being influenced by potentially non-diagnostic variables that other methods may have falsely deemed relevant (e.g., gender, race).

Who may be put at a disadvantage from this research. There is no foreseeable population who may be put at a disadvantage from this research.

What are the consequences of failure of the system. Overreliance on any prediction method can have obvious, negative real-world consequences, particularly when the prediction method is prone to failure. CART suffers from a couple of pitfalls: instability (i.e., small perturbations in the training samples may significantly change the structure of an optimal tree and consequent predictions) and difficulty in accurately approximating certain simple models, such as linear or, more generally, additive, if given insufficient or low quality data.

Whether the task/method leverages biases in the data. While CART is not impervious to all pre-existing biases in the data (e.g., those arising from systematic measurement errors at the data collection stage), as we have shown, it is less susceptible to the presence of additional, non-diagnostic variables in the data. Consequently, CART has the potential to mitigate the negative consequences of biasing information that is inevitable with most datasets.

Acknowledgments and Disclosure of Funding

References

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

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

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

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

[5] Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *International Conference on Machine Learning (ICML)*, 2014.

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

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

[8] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition*. Springer Series in Statistics. Springer New York, 2009. ISBN 9780387848587. URL https://books.google.com/books?id=tVIjmNS3Ob8C.

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

[10] Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A debiased MDI feature importance measure for random forests. In *Advances in Neural Information Processing Systems 32*, pages 8049–8059. Curran Associates, Inc., 2019.

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

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

[13] Andrew B Nobel. Analysis of a complexity-based pruning scheme for classification trees. *IEEE Transactions on Information Theory*, 48(8):2362–2368, 2002.

[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.

[16] Yi-Ching Yao and Siu-Tong Au. Least-squares estimation of a step function. *Sankhyā: The Indian Journal of Statistics, Series A*, pages 370–381, 1989.

Supplementary Material

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

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. Similar conventions apply to an optimal empirical and theoretical split coordinate index, $\hat{j}$ and j^*, respectively.

A Main Proofs

Lemma A.1 (Equivalence between the decrease in impurity and Pearson correlation from Section 3.1)**.**

$$\widehat{\rho}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) = \sqrt{\widehat{\Delta}(s, \mathrm{t})/\widehat{\Delta}(\mathrm{t})} \geq 0.$$

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,

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

$$= \frac{\widehat{P}(\mathrm{t}_L)(\overline{Y}_{\mathrm{t}_L})^2 + \widehat{P}(\mathrm{t}_R)(\overline{Y}_{\mathrm{t}_R})^2 - (\overline{Y}_{\mathrm{t}})^2}{\sqrt{(\widehat{P}(\mathrm{t}_L)(\overline{Y}_{\mathrm{t}_L})^2 + \widehat{P}(\mathrm{t}_R)(\overline{Y}_{\mathrm{t}_R})^2 - (\overline{Y}_{\mathrm{t}})) \times \widehat{\Delta}(\mathrm{t})}}$$

$$= \sqrt{\frac{\widehat{P}(\mathrm{t}_L)(\overline{Y}_{\mathrm{t}_L})^2 + \widehat{P}(\mathrm{t}_R)(\overline{Y}_{\mathrm{t}_R})^2 - (\overline{Y}_{\mathrm{t}})^2}{\widehat{\Delta}(\mathrm{t})}}$$

$$= \sqrt{\widehat{\Delta}(s, \mathrm{t})/\widehat{\Delta}(\mathrm{t})}.$$

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 (A.1) for the Pearson correlation. $\square$

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

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

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 thus $\Delta(s^*, \mathrm{t}) = \Theta(1/w)$. Since $\Delta(\mathrm{t}) = 1/2$, we have from the infinite sample analog of Lemma A.1 that $\rho(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t}) = \sqrt{\Delta(s^*, \mathrm{t})/\Delta(\mathrm{t})} = \Theta(1/\sqrt{w})$. $\square$

Lemma A.3 (Inequality (24) from Section 4.2)**.** *Let $g_1(X_1), g_2(X_2), \ldots, g_d(X_d)$ be univariate functions and let $Y_0 = \sum_j w_j g_j(X_j)$ consist of a subset of d_0 component functions $g_j(\cdot)$, where $w_j \in \{-1, +1\}$, and $\mathbf{w} = (w_j)_j$. Then,*

$$\max_{j=1,2,\ldots,d} \widehat{\rho}^2(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}) \geq \frac{\min_{\mathbf{w}} \widehat{\rho}^2(Y_0, Y \mid \mathbf{X} \in \mathrm{t})}{d_0}. \tag{A.2}$$

Furthermore, if each $g_j(\cdot)$ has nonnegative Pearson correlation with the others in the node, then

$$\max_{j=1,2,\ldots,d} \widehat{\rho}^2(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}) \geq \frac{\widehat{\rho}^2(Y_0, Y \mid \mathbf{X} \in \mathrm{t})}{d_0}, \tag{A.3}$$

where $Y_0 = \sum_j g_j(X_j)$.

Proof. Before we proceed with proving the lemma, 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., $\widehat{\sigma}_h^2(\mathrm{t}) = \widehat{\mathrm{VAR}}(h(\mathbf{X}) \mid \mathbf{X} \in \mathrm{t})$. Define the discrete prior $\pi(j, \mathbf{w})$ on the component function index j and sign vector $\mathbf{w}$ of Y_0 by

$$\pi(j, \mathbf{w}) = \frac{\widehat{\sigma}_{w_j g_j}(\mathrm{t})}{2^{d_0} \sum_{j'} \widehat{\sigma}_{w_{j'} g_{j'}}(\mathrm{t})} = \frac{\widehat{\sigma}_{g_j}(\mathrm{t})}{2^{d_0} \sum_{j'} \widehat{\sigma}_{g_{j'}}(\mathrm{t})}.$$

We are now in a position to prove (A.2). Since a maximum is greater than an average (with respect to the coordinate index j and sign vector $\mathbf{w}$), we have

$$\max_{j=1,2,\ldots,d} \widehat{\rho}^2(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}) = \max_{j=1,2,\ldots,d} \widehat{\rho}^2(w_j g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})$$
$$\geq \sum_{(j,\mathbf{w})} \pi(j, \mathbf{w}) \widehat{\rho}^2(w_j g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}).$$

Jensen's inequality for the square function yields

$$\sum_{(j,\mathbf{w})} \pi(j, \mathbf{w}) \widehat{\rho}^2(w_j g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}) \geq \sum_{\mathbf{w}} \pi(\mathbf{w}) \left| \sum_{j} \pi(j \mid \mathbf{w}) \widehat{\rho}(w_j g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}) \right|^2$$
$$= \sum_{\mathbf{w}} \pi(\mathbf{w}) \frac{\widehat{\sigma}_{Y_0}^2(\mathrm{t})}{(\sum_{j'} \widehat{\sigma}_{g_{j'}}(\mathrm{t}))^2} \widehat{\rho}^2(Y_0, Y \mid \mathbf{X} \in \mathrm{t})$$
$$\geq \frac{\sum_{\mathbf{w}} \pi(\mathbf{w}) \widehat{\sigma}_{Y_0}^2(\mathrm{t})}{(\sum_{j'} \widehat{\sigma}_{g_{j'}}(\mathrm{t}))^2} \min_{\mathbf{w}} \widehat{\rho}^2(Y_0, Y \mid \mathbf{X} \in \mathrm{t}) \qquad \text{(A.4)}$$

Next, note that $\sum_{\mathbf{w}} \pi(\mathbf{w}) \widehat{\sigma}_{Y_0}^2(\mathrm{t}) = \sum_j \widehat{\sigma}_{g_j}^2(\mathrm{t})$, since the covariance terms of $\widehat{\sigma}_{Y_0}^2(\mathrm{t})$ have mean zero with respect to $\pi(\mathbf{w}) \equiv 2^{-d_0}$; that is,

$$\sum_{\mathbf{w}} \pi(\mathbf{w}) \widehat{\sigma}_{Y_0}^2(\mathrm{t})$$
$$= \sum_{\mathbf{w}} \sum_j \pi(\mathbf{w}) \widehat{\sigma}_{w_j g_j}^2(\mathrm{t}) + \sum_{\mathbf{w}} \sum_j \pi(\mathbf{w}) \widehat{\mathrm{COV}}(w_j g_j(X_j), w_{j'} g_{j'}(X_{j'}) \mid \mathbf{X} \in \mathrm{t})$$
$$= \sum_j \widehat{\sigma}_{g_j}^2(\mathrm{t}) \sum_{\mathbf{w}} \pi(\mathbf{w}) + \sum_{j,j'} \widehat{\mathrm{COV}}(g_j(X_j), g_{j'}(X_{j'}) \mid \mathbf{X} \in \mathrm{t}) \sum_{\mathbf{w}} \pi(\mathbf{w}) w_j w_{j'}$$
$$= \sum_j \widehat{\sigma}_{g_j}^2(\mathrm{t}).$$

Combining this with (A.4) shows that $\max_{j=1,2,\ldots,d} \widehat{\rho}^2(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})$ is at least

$$\frac{\sum_j \widehat{\sigma}_{g_j}^2(\mathrm{t})}{(\sum_{j'} \widehat{\sigma}_{g_{j'}}(\mathrm{t}))^2} \min_{\mathbf{w}} \widehat{\rho}^2(Y_0, Y \mid \mathbf{X} \in \mathrm{t}) \geq \frac{\min_{\mathbf{w}} \widehat{\rho}^2(Y_0, Y \mid \mathbf{X} \in \mathrm{t})}{d_0},$$

where the last inequality follows from the Cauchy-Schwarz inequality. If each $g_j(\cdot)$ has nonnegative Pearson correlation with the others in the node, then $\widehat{\sigma}_{Y_0}^2(\mathrm{t}) \geq \sum_j \widehat{\sigma}_{g_j}^2(\mathrm{t})$ and thus the same argument as above can be repeated with $Y_0 = \sum_j g_j(X_j)$ to prove (A.3). $\qquad \square$

Proof of Theorem 1. Let $\overline{\mathrm{Err}}(\widehat{Y}) = \frac{1}{n} \sum_{i=1}^n (Y_i' - \widehat{Y}(T, \mathbf{X}_i'))^2$ denote the test error of $\widehat{Y}(T)$ on a test sample $\mathcal{D}_n' = \{(\mathbf{X}_i', Y_i')\}_{i=1}^n$ of size n. Let $\mathcal{T}_{\mathbf{X}, \mathbf{X}'}$ 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 the VC-dimension of the collection of all half-spaces, namely, $d + 1$. In this case, Lemma B.2 in [1] shows that the number of trees in $\mathcal{T}_{\mathbf{X}, \mathbf{X}'}$ with exactly $|T|$ nodes is at most $(2ne/(d + 1))^{|T|(d+1)}$. Using this, we have

$$\sum_{T \in \mathcal{T}_{\mathbf{X}, \mathbf{X}'}} e^{-L(T)} \leq \sum_{k:|T|=k \geq 1} \exp\left(-L(T) + |T|(d+1) \log(2ne/(d+1)) \right) \leq 1,$$

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 $L(T) := 2|T|(d+1) \log(2en/(d+1)) \geq |T|(\log(2) + (d+1) \log(2ne/(d+1)))$ satisfies

Kraft's inequality, i.e., $\sum_{T \in \mathcal{T}_{\mathbf{X},\mathbf{x}'}} e^{-L(T)} \le 1$. Observe also that $\mathcal{T}_{\mathbf{X},\mathbf{x}'}$ is symmetric in the pairs $(\mathbf{X}_i, \mathbf{X}'_i)$. By Lemma 2.1 in [3], for all $\gamma > 0$,

$$\mathbb{P}\left(\max_{T \in \mathcal{T}_{\mathbf{X},\mathbf{x}'}} \frac{\overline{\mathrm{Err}}(\widehat{Y}(T)) - \overline{\mathrm{err}}(\widehat{Y}(T))}{\frac{1}{n\gamma^2}(L(T) + \log(2/\delta)) + \frac{1}{2}S^2(\widehat{Y}(T))} < \gamma \right) \ge 1 - \delta/2, \qquad (A.5)$$

where $S^2(\widehat{Y}(T)) = \frac{1}{n}\sum_{i=1}^n ((Y'_i - \widehat{Y}(\mathbf{X}'_i))^2 - (Y_i - \widehat{Y}(\mathbf{X}_i))^2)^2$. Using the fact that $S^2(\widehat{Y}(T)) \le 8B^2(\overline{\mathrm{Err}}(\widehat{Y}(T)) + \overline{\mathrm{err}}(\widehat{Y}(T)))$ and $\widehat{T} \in \mathcal{T}_{\mathbf{X},\mathbf{x}'}$, and choosing $\gamma^{-1} = 12B^2$, we find that

$$\overline{\mathrm{Err}}(\widehat{Y}(\widehat{T})) \le 2\overline{\mathrm{err}}(\widehat{Y}(\widehat{T})) + \frac{18B^2 L(\widehat{T})}{n} + \frac{18B^2 \log(2/\delta)}{n} \qquad (A.6)$$

occurs with probability at least $1 - \delta/2$. Next, using Lemma 9 from [2] together with the bound $\overline{\mathrm{Err}}(\widehat{Y}(T)) \le 4B^2$ and the Kraft summability of the penalty $L(T)$, we have that for all $\gamma > 0$,

$$\mathbb{P}\left(\max_{T \in \mathcal{T}_{\mathbf{X}}} \frac{\mathrm{Err}(\widehat{Y}(T)) - \overline{\mathrm{Err}}(\widehat{Y}(T))}{\frac{4B^2}{n\gamma^2}(L(T) + \log(2/\delta)) + \mathrm{Err}(\widehat{Y}(T)) + \overline{\mathrm{Err}}(\widehat{Y}(T))} < \gamma \right) \ge 1 - \delta/2,$$

where $\mathcal{T}_{\mathbf{X}} \subset \mathcal{T}_{\mathbf{X},\mathbf{x}'}$ is the set of all tree-structured partitions constructed using the grid $\{\mathbf{X}_i\}_{i=1}^n$. Choose $\gamma = 1/3$. Since $\widehat{T} \in \mathcal{T}_{\mathbf{X}}$, with probability at least $1 - \delta/2$,

$$\mathrm{Err}(\widehat{Y}(\widehat{T})) \le 2\overline{\mathrm{Err}}(\widehat{Y}(\widehat{T})) + \frac{18B^2 L(\widehat{T})}{n} + \frac{18B^2 \log(2/\delta)}{n}. \qquad (A.7)$$

Combining (A.6) and (A.7), we have that with probability at least $1 - \delta$,

$$\mathrm{Err}(\widehat{Y}(\widehat{T})) \le 4R_\alpha(\widehat{Y}(\widehat{T})) + \frac{54B^2 \log(2/\delta)}{n},$$

provided $d > (n+1)/2$ and $\alpha > \frac{27B^2(d+1)\log(2en/(d+1))}{n}$. The conclusion of the theorem follows from the definition of $\widehat{T}$ as a minimizer of $R_\alpha(\widehat{Y}(T))$. $\qquad \square$

Proof of Theorem 2. The identity (10) is shown by first noting that, in the special case of uniform $\mathbf{X}$, the probability $\mathbb{P}(X \le s^* \mid \mathbf{X} \in \mathrm{t})$ from Lemma B.1 in Appendix B is equal to $(s^* - a)/(b - a)$. Rearranging the resulting expression yields the desired identity. $\qquad \square$

Proof of Lemma 1. We first prove (11) for a general decision stump $\widetilde{Y}$. The training error in t after splitting is

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

$$= \widehat{\Delta}(\mathrm{t})\left(1 - \frac{\widehat{\Delta}(s,\mathrm{t})}{\widehat{\Delta}(\mathrm{t})}\right)$$

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

where the last equality follows from Lemma A.1. Finally, $1 - \widehat{\rho}^2(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \le \exp(-\widehat{\rho}^2(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t}))$ follows from $1 - z \le e^{-z}$ for $z \ge 0$. To show (12), we use (11) with $\widetilde{Y} = \widehat{Y}$ recursively together with the identity

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

where the sum extends over all terminal nodes t of T_K. We stop once we reach the root node, at which point the training error is simply $\widehat{\sigma}_Y^2$. $\qquad \square$

Proof of Fact 1. Fact 1 is a special case of the following lemma. In order to state the lemma, we will need to introduce the concept of stationary intervals. We define a *stationary interval* of a univariate function $g(\cdot)$ to be a maximal interval I such that $g(I) = c$, where c is a local extremum of $g(\cdot)$ (I is maximal in the sense that there does not exist an interval I' such that $I \subset I'$ and $g(I') = c$). In particular, note that a monotone function does not have any stationary intervals.

14

Lemma A.4. *Almost surely, uniformly over all step functions $g(\cdot)$ of X that have at most V constant pieces and M stationary intervals in the node, we have*

$$\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \geq \frac{1}{\sqrt{D^{-1}MN(\mathrm{t}) + (V - M - 1) \wedge (1 + \log(2N(\mathrm{t})))}} \times |\widehat{\rho}(g(X), Y \mid \mathbf{X} \in \mathrm{t})|.$$
(A.8)

where $D \geq 1$ is the smallest number of data points in a stationary interval of $g(\cdot)$ that contains at least one data point.[2]

Proof of Lemma A.4. Let $g(\cdot)$ be any function of a generic coordinate X and assume that the data points in the node are labeled for simplicity as $\{X_i : \mathbf{X} \in \mathrm{t}\} = \{X_1, X_2, \ldots, X_{N(\mathrm{t})}\}$ and ordered such that $X_1 \leq X_2 \leq \cdots \leq X_{N(\mathrm{t})}$. 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 with density

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

where we remind the reader that $\widehat{P}(\mathrm{t}_L) = 1 - \widehat{P}(\mathrm{t}_R) = \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}} \mathbf{1}_{\{X_i \leq s\}}$. Here, $g'(s)$ equals the divided difference $\frac{g(X_{i+1}) - g(X_i)}{X_{i+1} - X_i}$ when $X_i \leq s < X_{i+1}$, $i = 1, 2, \ldots N(\mathrm{t}) - 1$. Accordingly, observe that Π has a piecewise constant density with knots at the data points and supported between the minimum and maximum of the data X_i. Since, by definition, $\widehat{Y}$ maximizes $s \mapsto \widehat{\rho}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t})$ and a maximum is larger than an average, we have

$$\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) = \max_s \widehat{\rho}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t})$$

$$\geq \int \widehat{\rho}(\widetilde{Y}, Y \mid \mathbf{X} \in \mathrm{t})d\Pi(s) = \int \sqrt{\frac{\widehat{\Delta}(s, \mathrm{t})}{\Delta(\mathrm{t})}}d\Pi(s),$$
(A.9)

where the last equality follows from Lemma A.1. Next, working from the representation (4), note that the reduction in impurity admits the form

$$\widehat{\Delta}(s, \mathrm{t}) = \left(\frac{1}{\sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)}}\left(\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}(\mathbf{1}_{\{s < X_i\}} - \widehat{P}(\mathrm{t}_R))(Y_i - \overline{Y}_{\mathrm{t}})\right)\right)^2,$$
(A.10)

and, hence, integrating inside the square in (A.10) against $g'(s)\sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)}$, we have

$$\int g'(s)\left(\frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}(\mathbf{1}_{\{s < X_i\}} - \widehat{P}(\mathrm{t}_R))(Y_i - \overline{Y}_{\mathrm{t}})\right)ds$$

$$= \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_i \in \mathrm{t}}\left(g(X_i) - \frac{1}{N(\mathrm{t})}\sum_{\mathbf{X}_{i'} \in \mathrm{t}}g(X_{i'})\right)(Y_i - \overline{Y}_{\mathrm{t}})$$

$$= \widehat{\mathrm{COV}}(g(X), Y \mid \mathbf{X} \in \mathrm{t}).$$
(A.11)

Using the inequality (A.9) together with the identities (A.10) and (A.11), we have

$$\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \geq \int \sqrt{\frac{\widehat{\Delta}(s, \mathrm{t})}{\Delta(\mathrm{t})}}d\Pi(s)$$

$$\geq \frac{\sqrt{\widehat{\mathrm{VAR}}(g(X) \mid \mathbf{X} \in \mathrm{t})}}{\int |g'(s)|\sqrt{\widehat{P}(\mathrm{t}_L)\widehat{P}(\mathrm{t}_R)}ds} \times |\widehat{\rho}(g(X), Y \mid \mathbf{X} \in \mathrm{t})|.$$
(A.12)

---

[2]More precisely, if $I_1, \ldots, I_M$ are the stationary intervals of $g(\cdot)$ and $D_k = \#\{X_i \in I_k\}$, then $D = \min_k\{D_k : D_k \geq 1\}$.

Therefore, from (A.12), we are led to determine how small the ratio

$$\frac{\sqrt{\widehat{\text{VAR}}(g(X) \mid \mathbf{X} \in \mathfrak{t})}}{\int |g'(s)|\sqrt{\widehat{P}(\mathfrak{t}_L)\widehat{P}(\mathfrak{t}_R)}ds}. \tag{A.13}$$

can be, ideally in terms of some simple structural characteristics of $g(\cdot)$. Our next task is to simplify (A.13) so that its numerator and denominator can be more easily compared. To this end, observe that

$$\begin{aligned}
&\int |g'(s)|\sqrt{\widehat{P}(\mathfrak{t}_L)\widehat{P}(\mathfrak{t}_R)}ds \\
&= \sum_{i=0}^{N(\mathfrak{t})} \int_{N(\mathfrak{t})\widehat{P}(\mathfrak{t}_L)=i} |g'(s)|\sqrt{\frac{i}{N(\mathfrak{t})}\left(1 - \frac{i}{N(\mathfrak{t})}\right)}ds \\
&= \sum_{i=1}^{N(\mathfrak{t})-1} \int_{X_i}^{X_{i+1}} |g'(s)|ds\sqrt{\frac{i}{N(\mathfrak{t})}\left(1 - \frac{i}{N(\mathfrak{t})}\right)} \\
&= \frac{1}{N(\mathfrak{t})} \sum_{i=1}^{N(\mathfrak{t})-1} |g(X_{i+1}) - g(X_i)|\sqrt{i(N(\mathfrak{t})-i)}, \tag{A.14}
\end{aligned}$$

where the penultimate equality follows from the fact that $\widehat{P}(\mathfrak{t}_L) = i/N(\mathfrak{t})$ if and only if $X_i \le s < X_{i+1}$. Next, we further simplify the above expression (A.14) using summation by parts, that is,

$$\frac{1}{N(\mathfrak{t})} \sum_{i=1}^{N(\mathfrak{t})-1} |g(X_{i+1}) - g(X_i)|\sqrt{i(N(\mathfrak{t})-i)} = -\frac{1}{N(\mathfrak{t})} \sum_{i=1}^{N(\mathfrak{t})} g(X_i)(b_i - b_{i-1}), \tag{A.15}$$

where $b_i = \text{sgn}(g(X_{i+1}) - g(X_i)) \times \sqrt{i(N(\mathfrak{t})-i)}$ with $b_0 = b_{N(\mathfrak{t})} = 0$. Next, since $\sum_{i=1}^{N(\mathfrak{t})}(b_i - b_{i-1}) = b_{N(\mathfrak{t})} - b_0 = 0$, (A.15) can be written as

$$-\frac{1}{N(\mathfrak{t})} \sum_{i=1}^{N(\mathfrak{t})} (g(X_i) - \frac{1}{N(\mathfrak{t})} \sum_{\mathbf{X}_{i'} \in \mathfrak{t}} g(X_{i'}))(b_i - b_{i-1}). \tag{A.16}$$

Moreover, we can express the variance $\widehat{\text{VAR}}(g(X) \mid \mathbf{X} \in \mathfrak{t})$ in a similar form, viz.,

$$\widehat{\text{VAR}}(g(X) \mid \mathbf{X} \in \mathfrak{t}) = \frac{1}{N(\mathfrak{t})} \sum_{i=1}^{N(\mathfrak{t})} (g(X_i) - \frac{1}{N(\mathfrak{t})} \sum_{\mathbf{X}_{i'} \in \mathfrak{t}} g(X_{i'}))^2. \tag{A.17}$$

To obtain the best lower bound on the ratio (A.13), we attempt to solve the program

$$\min_{g(\cdot) \in \mathcal{G}} \frac{\widehat{\text{VAR}}(g(X) \mid \mathbf{X} \in \mathfrak{t})}{\left(\int |g'(s)|\sqrt{\widehat{P}(\mathfrak{t}_L)\widehat{P}(\mathfrak{t}_R)}ds\right)^2}, \tag{A.18}$$

where $\mathcal{G}$ is a collection of functions. In light of the expressions (A.16) and (A.17), the program (A.18) is equivalent to the following program:

$$\min_{\mathbf{a} \in \mathcal{A}} \sum_{i=1}^{N(\mathfrak{t})} |a_i|^2 \quad \text{s.t.} \quad \frac{1}{\sqrt{N(\mathfrak{t})}} \sum_{i=1}^{N(\mathfrak{t})} a_i(b_i - b_{i-1}) = 1, \quad \sum_{i=1}^{N(\mathfrak{t})} a_i = 0. \tag{A.19}$$

where $b_i = \text{sgn}(a_{i+1} - a_i)\sqrt{i(N(\mathfrak{t})-i)}$ and $\mathcal{A}$ is a collection of vectors in $\mathbb{R}^{N(\mathfrak{t})}$. In order to incorporate structural and/or regularity properties of $g(\cdot)$, we will need to impose conditions on $\mathcal{G}$ or, since we associate a_i with $g(X_i) - \frac{1}{N(\mathfrak{t})} \sum_{\mathbf{X}_{i'} \in \mathfrak{t}} g(X_{i'})$, on $\mathcal{A}$. However, not all specifications make the program tractable to solve, or even convex. As a compromise, we fix the signs of the b_i in advance. That is, we specify two sets where $b_i = 0$, $b_i > 0$, and $b_i < 0$—corresponding to locations where $g(\cdot)$ is constant, increasing, and decreasing, respectively—and solve the resulting (quadratic) program. More formally, let V and M respectively denote the number of constant pieces and stationary intervals

16

of $g(\cdot)$ and let $S = \{i_k\}_{1 \le k \le V-1}$ and $S' \subset S$ be two subsets of $\{1, 2, \ldots, N(\mathrm{t}) - 1\}$ with $i_0 = 0$ and $i_V = N(\mathrm{t})$. Let $\mathcal{A} = \{\mathbf{a} \in \mathbb{R}^{N(\mathrm{t})} : b_i = 0 \text{ for } i \notin S,\ b_i > 0 \text{ for } i \in S',\ b_i < 0 \text{ for } i \notin S'\}$, and $D_k = i_k - i_{k-1}$. (Note that M can be regarded as the number of times $g(\cdot)$ changes from strictly increasing to decreasing (or vice versa) and hence $b_{i-1} b_i < 0$ at most M times.) With these specifications fixed, the program (A.19) becomes

$$\min_{\mathbf{a} \in \mathcal{A}} \quad \sum_{k=1}^{V} |a_{i_k}|^2 D_k \quad \text{s.t.} \quad \frac{1}{\sqrt{N(\mathrm{t})}} \sum_{k=1}^{V} a_{i_k}(b_{i_k} - b_{i_{k-1}}) = 1, \quad \sum_{k=1}^{V} a_{i_k} D_k = 0. \tag{A.20}$$

Using the method of Lagrange multipliers, it is easy to see that the solution to (A.20) is

$$a_{i_k}^* = \frac{\sqrt{N(\mathrm{t})}(b_{i_k} - b_{i_{k-1}})/D_k}{\sum_{k=1}^{V}(b_{i_k} - b_{i_{k-1}})^2/D_k}, \quad k = 1, 2, \ldots, V, \tag{A.21}$$

and the value of the program is

$$\frac{N(\mathrm{t})}{\sum_{k=1}^{V}(b_{i_k} - b_{i_{k-1}})^2/D_k}. \tag{A.22}$$

Lemma B.3 in Appendix B shows that (A.22) is at least

$$\frac{1}{D^{-1}MN(\mathrm{t}) + (V - M - 1) \wedge (1 + \log(2N(\mathrm{t})))},$$

where D is the smallest number of data points in a stationary interval of $g(\cdot)$ that contains at least one data point. Hence by (A.12), we obtain the desired (A.8). $\qquad\square$

Fact 1 follows immediately from (A.8) by noting that, in this case, $M = 0$. $\qquad\square$

Remark A.1. *Another candidate prior Π is*

$$\frac{d\Pi(j, s)}{d(j, s)} := \frac{|g_j'(s)|\sqrt{\widehat{P}_j(\mathrm{t}_L)\widehat{P}_j(\mathrm{t}_R)}}{\sum_j \int |g_j'(s')|\sqrt{\widehat{P}_j(\mathrm{t}_L)\widehat{P}_j(\mathrm{t}_R)}ds'},$$

which, akin to (A.12), leads to the correlation inequality

$$\widehat{\rho}(\widehat{Y}, Y \mid \mathbf{X} \in \mathrm{t}) \ge \frac{\sqrt{\widehat{VAR}(\sum_j g_j(X_j) \mid \mathbf{X} \in \mathrm{t})}}{\sum_j \int |g_j'(s)|\sqrt{\widehat{P}_j(\mathrm{t}_L)\widehat{P}_j(\mathrm{t}_R)}ds} \times |\widehat{\rho}(\sum_j g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})|.$$

While this enables comparisons with additive models via $\widehat{\rho}(\sum_j g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})$, the factor $\frac{\sqrt{\widehat{VAR}(\sum_j g_j(X_j)|\mathbf{X}\in\mathrm{t})}}{\sum_j \int |g_j'(s)|\sqrt{\widehat{P}_j(\mathrm{t}_L)\widehat{P}_j(\mathrm{t}_R)}ds}$ is less amenable to analysis.

Proof of Theorem 3. We first employ a technique similar to (A.12) in the proof of Fact 1 (essentially, the infinite sample analog) to lower bound $\rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})$. That is, for each function $g(\cdot)$ of X and node t,

$$\rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t}) \ge \Lambda \times \rho^2(g(X), Y \mid \mathbf{X} \in \mathrm{t}), \tag{A.23}$$

where

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

In contrast with the proof of Fact 1, here we do not attempt to minimize Λ over all $g(\cdot)$ in some function class. Rather, we attempt to lower bound it for a *fixed* $g(\cdot)$. Now, (A.23) is valid for all $g_j(X_j)$ and so we can instead consider the maximum correlation over all $g_j(X_j)$, i.e., $\max_j \rho^2(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t})$, where now Λ is the minimum over all $g_j(X_j)$. By the infinite sample analog of (A.3) in Lemma A.3, we have $\max_j \rho^2(g_j(X_j), Y \mid \mathbf{X} \in \mathrm{t}) \ge \frac{\rho^2(Y, Y|\mathbf{X}\in\mathrm{t})}{d_0} = 1/d_0$, and hence

$$\rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t}) \ge \Lambda/d_0. \tag{A.24}$$

17

Next, we show that Λ can be further lower bounded by a positive constant that is independent of t. To this end, note that Λ is continuous in (a, b) and strictly positive for all $a < b$ and, furthermore by Lemma B.2 in Appendix B,

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

where $R = \sup_{c\in[0,1]} \inf\{r \geq 1 : g^{(r)}(\cdot) \text{ exists and is continuous and nonzero at } c\}$—which means that $\inf_{(a,b)} \Lambda > 0$. Note that, in particular, R is finite if $g(\cdot)$ admits a power series representation. Taking the minimum of $\inf_{(a,b)} \Lambda$ over all $g_j(\cdot)$—each of which has finite R—results in a positive quantity that depends only on each $g_j(\cdot)$ individually. This shows that $\inf_t \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in t) \geq C/d_0$ for some positive constant C that depends only on each $g_j(\cdot)$ individually and not on d_0. Next, we will show that, almost surely,

$$\liminf_n \widehat{\rho}_{\mathcal{H}}^2 = \liminf_n \inf_t \widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in t) \geq \inf_t \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in t),$$

from which the first statement in Theorem 3 will follow, i.e., $\liminf_n \widehat{\rho}_{\mathcal{H}}^2 \geq C/d_0$ almost surely. First, by definition of $\widehat{Y}$ as the optimizer of $(j, s) \mapsto \widehat{\rho}^2(\widetilde{Y}, Y \mid \mathbf{X} \in t)$, almost surely,

$$\liminf_n \inf_t \widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in t) \geq \liminf_n \inf_t \widehat{\rho}^2(\widehat{Y}^*, Y \mid \mathbf{X} \in t),$$

where we remind the reader that $\widehat{Y}^*$ is the decision stump $\widetilde{Y}$ at an optimal theoretical direction j^* and split s^*. Next, note that $\widehat{\rho}(\widehat{Y}^*, Y \mid \mathbf{X} \in t)$ is invariant to scale. Working instead with $\frac{N(t)}{n}\widehat{Y}^*$ and $\frac{N(t)}{n}Y$, we find that the correlation involves terms (empirical processes) of the form $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\mathbf{X}_i\in t'\}}$, $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\mathbf{X}_i\in t'\}}Y_i$, and $\frac{1}{n}\sum_{i=1}^n \mathbf{1}_{\{\mathbf{X}_i\in t\}}Y_i^2$, where t' is either the parent node t or one of the daughter nodes, $t_L^* := \{\mathbf{X} \in t : X \leq s^*\}$ and $t_R^* := \{\mathbf{X} \in t : X > s^*\}$ at an optimal theoretical split s^*. The collection of hyperrectangles in $\mathbb{R}^d$ is a finite VC-class with VC-dimension at most $2d$, and hence these terms converge almost surely, uniformly over all nodes t', to their respective population level counterparts when $d = o(n)$. Thus, $\liminf_n \inf_t \widehat{\rho}^2(\widehat{Y}^*, Y \mid \mathbf{X} \in t) \stackrel{\text{a.s.}}{=} \inf_t \liminf_n \widehat{\rho}^2(\widehat{Y}^*, Y \mid \mathbf{X} \in t) \stackrel{\text{a.s.}}{=} \inf_t \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in t)$.

The almost sure limit (19) in Theorem 3 follows from (17) with $\delta = 1/n^2$ and $\liminf_n \widehat{\rho}_{\mathcal{H}}^2 \geq C/d_0$ (almost surely) together with the Borel-Cantelli lemma. $\square$

Proof of Theorem 4. As mentioned right before the statement of Theorem 4, we need to prove (20). To lighten notation, we consider a generic direction X, write N for $N(t)$, and assume that the data is labeled in the node t so that $X_1 \leq X_2 \leq \cdots \leq X_N$. Let I be one of the intervals on which $g(X)$ is constant and let $X_{i_1} = \min\{X_i \in I : \mathbf{X}_i \in t\}$ and $X_{i_2} = \max\{X_i \in I : \mathbf{X}_i \in t\}$ so that $i_1 \leq i_2$. We will show that if $\widehat{\Delta}(\hat{s}, t) > 0$, then the maximum of $\widehat{\Delta}(s, t)$ for $s \in [X_{i_1}, X_{i_2+1})$ must occur at the boundary, i.e., $[X_{i_1}, X_{i_1+1})$ or $[X_{i_2}, X_{i_2+1})$. Let $\mu_1 = \frac{1}{i_1}\sum_{\mathbf{X}\in t,\, X_i\leq X_{i_1}} Y_i$, $\mu_2 = \frac{1}{N-i_2}\sum_{\mathbf{X}\in t,\, X_i>X_{i_2}} Y_i$, and $\mu = \frac{1}{i_2-i_1}\sum_{X_{i_1}<X_i\leq X_{i_2}} Y_i$. Suppose $X_i \leq s < X_{i+1}$. Then the decrease in impurity equals

$$\widehat{\Delta}(s, t) = \frac{i}{N} \times \frac{N-i}{N} \times \left(\frac{1}{i}(i_1\mu_1 + (i-i_1)\mu) - \frac{1}{N-i}((N-i_2)\mu_2 + (i_2-i)\mu)\right)^2.$$

Viewed as a function of i, $\widehat{\Delta}(s, t) = \widehat{\Delta}(i)$ has two critical values, one of which is a zero solution, namely, $i^* = \frac{(\mu_1-\mu)i_1 N}{\mu_1 i_1 + \mu_2(N-i_2) - \mu(N+i_1-i_2)}$. The other critical value, equal to

$$i^* = \frac{(\mu_1-\mu)i_1 N}{\mu_1 i_1 - \mu_2(N-i_2) + \mu(N-i_1-i_2)},$$

produces the value

$$\widehat{\Delta}(i^*) = \frac{4i_1(N-i_2)(\mu_1-\mu)(\mu-\mu_2)}{N^2}.$$

We will be done if we can show that either

$$\widehat{\Delta}(i_1) = \frac{i_1(\mu_1(N-i_1) - \mu_2(N-i_2) - \mu(i_2-i_2))^2}{N^2(N-i_1)}$$

or

$$\widehat{\Delta}(i_2) = \frac{(N - i_2)(\mu_1 i_1 - \mu_2 i_2 + \mu(i_2 - i_1))^2}{N^2 i_2}$$

are (strictly) greater than $\widehat{\Delta}(i^*)$. After some tedious algebra, we find that $\widehat{\Delta}(i_1) > \widehat{\Delta}(i^*)$ and $\widehat{\Delta}(i_2) > \widehat{\Delta}(i^*)$ with equality if and only if $i^* = i_1$ and $i^* = i_2$, respectively. $\qquad\square$

Proof of Theorem 5. We first show that

$$\overline{\mathrm{err}}(\widehat{Y}(T_K)) \le \widehat{\sigma}_Y^2 \exp\Big(- \widehat{\rho}_{\mathcal{M}}^2 \sum_{k=1}^{K} (\log_2(4N_k))^{-1} \Big). \tag{A.25}$$

By (11) in Lemma 1, the training error in the node is decreased by a factor of $\exp(-\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathfrak{t}))$ each time the node is split. By Fact 1, almost surely, $\widehat{\rho}^2(\widehat{Y}, Y \mid \mathbf{X} \in \mathfrak{t}) \ge \frac{1}{1 + \log(2N(\mathfrak{t}))} \times \widehat{\rho}_{\mathcal{M}}^2 \ge \frac{1}{\log_2(4N(\mathfrak{t}))} \times \widehat{\rho}_{\mathcal{M}}^2 \ge \frac{1}{\log_2(4N_k)} \times \widehat{\rho}_{\mathcal{M}}^2$, if $\mathfrak{t}$ is a node at level k. Thus, the training error at level $k + 1$ is at most $\exp(-\widehat{\rho}_{\mathcal{M}}^2 (\log_2(4N_k))^{-1})$ times the training error at level k—in other words, the training error is geometrically decreasing. The proof of (A.25) 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 (22), we use the inequality $\sum_{k=1}^{K} \frac{1}{\log_2(4Ank^a/2^k)} \ge \log\Big(\frac{\log_2(4K^a An)}{\log_2(4K^a An) - K} \Big)$ for integers $K \ge 1$. By (A.25), if T_K is a fully grown tree of depth K, then under Assumption 1, i.e., $N_k \le Ank^a/2^k$, we have

$$\overline{\mathrm{err}}(\widehat{Y}(T_K)) \le \widehat{\sigma}_Y^2 \exp\Big(- \widehat{\rho}_{\mathcal{M}}^2 \sum_{k=1}^{K} (\log_2(4N_k))^{-1} \Big)$$

$$\le \widehat{\sigma}_Y^2 \exp\Big(- \widehat{\rho}_{\mathcal{M}}^2 \sum_{k=1}^{K} \frac{1}{\log_2(4Ank^a/2^k)} \Big)$$

$$\le \widehat{\sigma}_Y^2 \Big(1 - \frac{K}{\log_2(4K^a An)} \Big)^{\widehat{\rho}_{\mathcal{M}}^2}. \tag{A.26}$$

Next, we show (23), 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

$$\inf_{T \preceq T_{\max}} R_\alpha(\widehat{Y}(T)), \tag{A.27}$$

where the temperature α is $\Theta((d/n)\log(n/d))$. Note that (A.27) is smaller than the minimum of $R_\alpha(\widehat{Y}(T_K)) = \overline{\mathrm{err}}(\widehat{Y}(T_K)) + \alpha|T_K|$ over all fully grown trees T_K of depth K with $|T_K| \le 2^K$, i.e.,

$$\inf_{K \ge 1} \{ \overline{\mathrm{err}}(\widehat{Y}(T_K)) + \alpha 2^K \}. \tag{A.28}$$

Combining the training error bound (A.26) with (A.28), we are led to optimize

$$\widehat{\sigma}_Y^2 \Big(1 - \frac{K}{\log_2(4K^a An)} \Big)^{\widehat{\rho}_{\mathcal{M}}^2} + \alpha 2^K, \tag{A.29}$$

over $K \ge 1$, although suboptimal choices of K will suffice for our purposes. Choosing K to satisfy $K = \lceil \log_2\big(\frac{\widehat{\sigma}_Y^2 (\log_2(4K^a An))^{-\widehat{\rho}_{\mathcal{M}}^2}}{\alpha} \big) \rceil < \lceil \log_2(\widehat{\sigma}_Y^2/\alpha) \rceil$, we find that (A.29) is equal to

$$\widehat{\sigma}_Y^2 \Big(\frac{\log_2(4K^a An\alpha(\log_2(4K^a An))^{\widehat{\rho}_{\mathcal{M}}^2}/\widehat{\sigma}_Y^2)}{\log_2(4K^a An)} \Big)^{\widehat{\rho}_{\mathcal{M}}^2} + \widehat{\sigma}_Y^2 \Big(\frac{1}{\log_2(4K^a An)} \Big)^{\widehat{\rho}_{\mathcal{M}}^2}$$

$$= \mathcal{O}\Big(\widehat{\sigma}_Y^2 \Big(\frac{\log((d/\widehat{\sigma}_Y^2)\log^{2+a}(n))}{\log(n)} \Big)^{\widehat{\rho}_{\mathcal{M}}^2} \Big).$$

Combining this bound with Theorem 1 proves (23). $\qquad\square$

B Auxiliary Lemmas

Lemma B.1. *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*

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

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

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

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

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

$$\Delta(s, \mathrm{t}) = |\Xi(s)|^2 / (P(\mathrm{t}_L)P(\mathrm{t}_R)). \tag{B.3}$$

An easy calculation shows that

$$\frac{\partial}{\partial s}\Xi(s) = p(\mathrm{t}_L)(\mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t},\ X = s] - \mathbb{E}[Y \mid \mathbf{X} \in \mathrm{t}]) = p(\mathrm{t}_L)G(s), \tag{B.4}$$

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

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

$$\frac{\partial}{\partial s}\Delta(s, \mathrm{t}) = \frac{\Xi(s)p(\mathrm{t}_L)(2P(\mathrm{t}_L)P(\mathrm{t}_R)G(s) - \Xi(s)(1 - 2P(\mathrm{t}_L)))}{(P(\mathrm{t}_L)P(\mathrm{t}_R))^2}. \tag{B.5}$$

Suppose s^* is a global maximizer of (B.3) (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 (B.5), s^* satisfies

$$\Xi(s^*)p(\mathrm{t}_L^*)(2P(\mathrm{t}_L^*)P(\mathrm{t}_R^*)G(s^*) - \Xi(s^*)(1 - 2P(\mathrm{t}_L^*))) = 0, \tag{B.6}$$

where we denote the daughter nodes with an optimal theoretical split s^* by t_L^* and t_R^*, i.e., $\mathrm{t}_L^* = \{\mathbf{X} \in \mathrm{t} : X \leq s^*\}$ and $\mathrm{t}_R^* = \{\mathbf{X} \in \mathrm{t} : X > s^*\}$. 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 (B.6) and using the identity (B.3) that

$$P(\mathrm{t}_L^*) = \frac{1}{2} - \frac{\mathrm{sgn}(\Xi(s^*)) \times G(s^*)}{\sqrt{\Delta(s^*, \mathrm{t})}}\sqrt{P(\mathrm{t}_L^*)P(\mathrm{t}_R^*)}. \tag{B.7}$$

The solution to (B.7) is obtained by solving a simple quadratic equation of the form $p = 1/2 \pm c\sqrt{p(1 - p)}$, $0 \leq p \leq 1$, and noting from Lemma A.1 that $\Delta(s^*, \mathrm{t}) = \Delta(\mathrm{t}) \times \rho^2(\widehat{Y}^*, Y \mid \mathbf{X} \in \mathrm{t})$, which proves the identity (B.1). $\square$

Lemma B.2. *Suppose X is uniformly distributed on the unit interval and $R = \inf\{r \geq 1 : g^{(r)}(\cdot)$ exists and is continuous and nonzero at $c\} < \infty$, where $c \in [0, 1]$. Then*

$$\liminf_{(a,b) \to (c,c)} \left\{ \frac{VAR(g(X) \mid X \in [a, b])}{\left(\int_a^b |g'(x)|\sqrt{\frac{x-a}{b-a}\frac{b-x}{b-a}}dx\right)^2} \right\} = \Omega(1/R). \tag{B.8}$$

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 (B.8) is

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

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

$$\lim_{(a,b)\to(c,c)} (b-a)^{-R} \int_0^1 |g'(x(b-a)+a)|\sqrt{x(1-x)}dx \tag{B.9}$$

$$= \frac{|g^{(R)}(c)|}{(R-1)!} \int_0^1 |x-\delta|^{R-1}\sqrt{x(1-x)}dx. \tag{B.10}$$

For the variance, first note that

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

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

$$D(x)(x-\delta)^R - \int_0^1 D(x')(x'-\delta)^R dx'. \tag{B.11}$$

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 and the limit is nonzero by definition of R. Therefore, for fixed δ,

$$\lim_{(a,b)\to(c,c)} (b-a)^{-2R}\mathrm{VAR}(g(X(b-a)+a)) \tag{B.12}$$

$$= \left(\frac{g^{(R)}(c)}{R!}\right)^2 \int_0^1 \left((x-\delta)^R - \int_0^1 (x'-\delta)^R dx'\right)^2 dx$$

$$= \left(\frac{g^{(R)}(c)}{R!}\right)^2 \mathrm{VAR}((X-\delta)^R). \tag{B.13}$$

Combining (B.9) and (B.13), we have that the limit infimum (B.8) is at least

$$\inf_\delta \frac{\mathrm{VAR}((X-\delta)^R)}{(R\int_0^1 |x-\delta|^{R-1}\sqrt{x(1-x)}dx)^2}. \tag{B.14}$$

Tedious calculations show that the infimum is achieved at $\delta = 1/2$ and hence (B.14) is $\Omega(1/R)$. $\quad\square$

Lemma B.3. *Consider the expression (A.22). Then,*

$$\frac{N(\mathrm{t})}{\sum_{k=1}^V (b_{i_k} - b_{i_{k-1}})^2/D_k} \geq \frac{1}{D^{-1}MN(\mathrm{t}) + (V-M-1)\wedge(1+\log(2N(\mathrm{t})))}, \tag{B.15}$$

where M, V, and D are defined in Lemma A.4.

Proof. For brevity, we omit dependent on t and write N instead of $N(\mathrm{t})$.

Suppose that b_i changes sign at index i_k (one of the M many indices such that $b_{i_{k-1}}b_{i_k} < 0$). Then, since $b_{i_k} = \mathrm{sgn}(a_{i_k} - a_{i_{k-1}})\sqrt{i_k(N-i_k)}$, we have

$$\sum_{k:b_{i_{k-1}}b_{i_k}<0} \frac{(b_{i_k} - b_{i_{k-1}})^2}{ND_k} = \sum_{k:b_{i_{k-1}}b_{i_k}<0} \frac{(|b_{i_k}| + |b_{i_{k-1}}|)^2}{ND_k}$$

$$\leq \sum_{k:b_{i_{k-1}}b_{i_k}<0} \frac{(|b_{i_k}| + |b_{i_{k-1}}|)^2}{ND}$$

$$\leq D^{-1}MN,$$

where the last line is from $(|b_{i_k}| + |b_{i_{k-1}}|)^2 = (\sqrt{i_k(N - i_k)} + \sqrt{i_{k-1}(N - i_{k-1})})^2 \leq N^2$. Next, for the remaining $V - M$ indices such that $b_{i_{k-1}}b_{i_k} > 0$ we have,

$$\sum_{k:b_{i_{k-1}}b_{i_k}>0} \frac{(|b_{i_k}| - |b_{i_{k-1}}|)^2}{ND_k} \leq \sum_{k:b_{i_{k-1}}b_{i_k}>0} \frac{|N - i_k - i_{k-1}|}{N}$$
$$\leq V - M - 1,$$

where the last line follows from the fact there is always one index such that $|N - i_k - i_{k-1}| + |N - i_{k+1} - i_k| = |i_{k+1} - i_{k-1}|$, namely, at $k^* := \min\{k : i_k + i_{k-1} \geq N\}$. Thus, it follows that $\frac{N}{\sum_{k=1}^{V}(b_{i_k} - b_{i_{k-1}})^2/D_k}$ is at least

$$\frac{1}{D^{-1}MN + (V - M - 1) \wedge \sum_{k=1}^{V} \frac{(|b_{i_k}| - |b_{i_{k-1}}|)^2}{ND_k}}. \tag{B.16}$$

We now obtain an upper bound for

$$\sum_{k=1}^{V} \frac{(|b_{i_k}| - |b_{i_{k-1}}|)^2}{ND_k} = \sum_{k=1}^{V} \frac{D_k(N - i_k - i_{k-1})^2}{N(\sqrt{i_k(N - i_k)} + \sqrt{i_{k-1}(N - i_{k-1})})^2}. \tag{B.17}$$

Now, $(\sqrt{i_k(N - i_k)} + \sqrt{i_{k-1}(N - i_{k-1})})^2 \geq (2N - i_k - i_{k-1})(i_k + i_{k-1} - N)$ for all $k \geq k^*$. Thus, the sum $\sum_{k \geq k^*} \frac{D_k(N - i_k - i_{k-1})^2}{N(\sqrt{i_k(N - i_k)} + \sqrt{i_{k-1}(N - i_{k-1})})^2}$ is at most

$$\sum_{k \geq k^*} \frac{D_k}{2N - i_k - i_{k-1}}\left(\frac{i_{k-1} + i_k}{N} - 1\right) \leq \sum_{k \geq k^*} \frac{i_k - i_{k-1}}{2N - i_k - i_{k-1}}, \tag{B.18}$$

where we used the fact that $D_k = i_k - i_{k-1}$. Next, $(\sqrt{i_k(N - i_k)} + \sqrt{i_{k-1}(N - i_{k-1})})^2 \geq (i_k + i_{k-1})(N - i_k - i_{k-1})$ 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

$$\sum_{k < k^*} \frac{D_k}{i_k + i_{k-1}}\left(1 - \frac{i_{k-1} + i_k}{N}\right) \leq \sum_{k < k^*} \frac{i_k - i_{k-1}}{i_k + i_{k-1}}. \tag{B.19}$$

Combining (B.18) and (B.19), we have shown that (B.17) is at most

$$\sum_{k < k^*} \frac{i_k - i_{k-1}}{i_k + i_{k-1}} + \sum_{k \geq k^*} \frac{i_k - i_{k-1}}{2N - i_k - i_{k-1}}. \tag{B.20}$$

The sum (B.20) is largest when $V = N$, yielding

$$\sum_{i=1}^{(N-1)/2} \frac{1}{2i - 1} + \sum_{i=1}^{(N+1)/2} \frac{1}{2i - 1} \leq 1 + \log(2N). \tag{B.21}$$

Combining (B.20) and (B.21) with (B.16) proves (B.15). $\qquad\square$

References

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

[2] David Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, 1992.

[3] 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.