The Annals of Statistics
2024, Vol. 52, No. 2, 466–490
https://doi.org/10.1214/24-AOS2354

CONVERGENCE RATES OF OBLIQUE REGRESSION TREES FOR FLEXIBLE FUNCTION LIBRARIES

BY MATIAS D. CATTANEO[a], RAJITA CHANDAK[b] AND JASON M. KLUSOWSKI[c]

Department of Operations Research and Financial Engineering, Princeton University,
[a]*cattaneo@princeton.edu,* [b]*rchandak@princeton.edu,* [c]*jason.klusowski@princeton.edu*

We develop a theoretical framework for the analysis of oblique decision trees, where the splits at each decision node occur at linear combinations of the covariates (as opposed to conventional tree constructions that force axis-aligned splits involving only a single covariate). While this methodology has garnered significant attention from the computer science and optimization communities since the mid-80s, the advantages they offer over their axis-aligned counterparts remain only empirically justified, and explanations for their success are largely based on heuristics. Filling this long-standing gap between theory and practice, we show that oblique regression trees (constructed by recursively minimizing squared error) satisfy a type of oracle inequality and can adapt to a rich library of regression models consisting of linear combinations of ridge functions and their limit points. This provides a quantitative baseline to compare and contrast decision trees with other less interpretable methods, such as projection pursuit regression and neural networks, which target similar model forms. Contrary to popular belief, one needs not always trade-off interpretability with accuracy. Specifically, we show that, under suitable conditions, oblique decision trees achieve similar predictive accuracy as neural networks for the same library of regression models. To address the combinatorial complexity of finding the optimal splitting hyperplane at each decision node, our proposed theoretical framework can accommodate many existing computational tools in the literature. Our results rely on (arguably surprising) connections between recursive adaptive partitioning and sequential greedy approximation algorithms for convex optimization problems (e.g., orthogonal greedy algorithms), which may be of independent theoretical interest. Using our theory and methods, we also study oblique random forests.

1. Introduction. Decision trees and neural networks are conventionally seen as two contrasting approaches to learning. The popular belief is that decision trees compromise accuracy for being easy to use and understand, whereas neural networks are more accurate, but at the cost of being less transparent. We challenge the *status quo* by showing that, under suitable conditions, oblique decision trees (also known as multivariate decision trees) achieve similar predictive accuracy as neural networks on the same library of regression models. Of course, while it is somewhat subjective as to what one regards as being transparent, it is generally agreed upon that neural networks are less interpretable than decision trees [35, 41]. Indeed, trees are arguably more intuitive in their construction, which makes it easier to understand how an output is assigned to a given input, including which predictor variables were relevant in its determination. For example, in clinical, legal or business contexts, it may be desirable to build a predictive model that mimics the way a human user thinks and reasons, especially if the results (of scientific or evidential value) are to be communicated to a statistical lay audience. Even though it may be sensible to deploy estimators that more directly target the functional form of the model, predictive accuracy is not the only factor the

Received October 2022; revised September 2023.
MSC2020 subject classifications. Primary 62G08; secondary 62L12.
Key words and phrases. Decision trees, neural networks, projection pursuit regression, CART, random forest.

modern researcher must consider when designing and building an automated system. Facilitating human-machine interaction and engagement is also an essential part of this process. To this end, the technique of knowledge distillation [13] is a quick and easy way to enhance the fidelity of an interpretable model, without degrading the out-of-sample performance too severely. In the context of decision trees and neural networks, one distills the knowledge acquired by a neural network—which relies on nontransparent, distributed hierarchical representations of the data—and expresses similar knowledge in a decision tree that consists of, in contrast, easier to understand hierarchical decision rules [20]. This is accomplished by first training a neural network on the observed data, and then, in turn, training a decision tree on data generated from the fitted neural network model.

In this paper, we show that oblique regression trees (constructed by recursively minimizing squared error) satisfy a type of oracle inequality and can adapt to a rich library of regression models consisting of linear combinations of ridge functions. This provides a quantitative baseline to compare and contrast decision trees with other less interpretable methods, such as projection pursuit regression, neural networks and boosting machines, which directly target similar model forms. When neural network and decision tree models are used in tandem to enhance generalization and interpretability, our theory allows one to measure the knowledge distilled from a neural network to a decision tree. Using our theory and methods, we also study oblique random forests.

1.1. *Background and prior work.* Let $(y_1, \mathbf{x}_1^{\mathrm{T}}), \ldots, (y_n, \mathbf{x}_n^{\mathrm{T}})$ be a random sample from a joint distribution $\mathbb{P}_{(y,\mathbf{x})} = \mathbb{P}_{y|\mathbf{x}}\mathbb{P}_{\mathbf{x}}$ supported on $\mathcal{Y} \times \mathcal{X}$. Here, $\mathbf{x} = (x_1, \ldots, x_p)^{\mathrm{T}}$ is a vector of p predictor variables supported on $\mathcal{X} \subseteq \mathbb{R}^p$ and y is a real-valued outcome variable with range $\mathcal{Y} \subseteq \mathbb{R}$. Our objective is to compute an estimate of the conditional expectation, $\mu(\mathbf{x}) = \mathbb{E}[y \mid \mathbf{x}]$, a target which is optimal for predicting y from some function of $\mathbf{x}$ in mean squared error. One estimation scheme can be constructed by dividing the input space $\mathcal{X}$ into subgroups based on shared characteristics of y—something decision trees can do well.

A decision tree is a hierarchically organized data structure constructed in a top down, greedy manner through recursive binary splitting. According to CART methodology [11], a parent node t (i.e., a region in $\mathcal{X}$) in the tree is divided into two child nodes, t_L and t_R, by maximizing the decrease in sum-of-squares error (SSE)

$$\widehat{\Delta}(b, \mathbf{a}, \mathrm{t}) = \frac{1}{n}\sum_{\mathbf{x}_i \in \mathrm{t}}(y_i - \overline{y}_{\mathrm{t}})^2 - \frac{1}{n}\sum_{\mathbf{x}_i \in \mathrm{t}}\big(y_i - \overline{y}_{\mathrm{t}_L}\mathbb{1}(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i \le b) - \overline{y}_{\mathrm{t}_R}\mathbb{1}(\mathbf{a}^{\mathrm{T}}\mathbf{x}_i > b)\big)^2, \tag{1}$$

with respect to $(b, \mathbf{a})$, with $\mathbb{1}(\cdot)$ denoting the indicator function and $\overline{y}_{\mathrm{t}}$ denoting the sample average of the y_i data whose corresponding $\mathbf{x}_i$ data lies in the node t. In the conventional *axis-aligned* (or, *univariate*) *CART* algorithm [11], Section 2.2, splits occur along values of a single covariate, and so the search space for $\mathbf{a}$ is restricted to the set of standard basis vectors in $\mathbb{R}^p$. In this case, the induced partition of the input space $\mathcal{X}$ is a set of hyperrectangles. On the other hand, the *oblique CART* algorithm [11], Section 5.2, allows for linear combinations of covariates, extending the search space for $\mathbf{a}$ to be all of $\mathbb{R}^p$. Such a procedure generates regions in $\mathbb{R}^p$ that are convex polytopes.

The solution of (1) yields estimates $(\hat{b}, \hat{\mathbf{a}})$, and the refinement of t produces child nodes $\mathrm{t}_L = \{\mathbf{x} \in \mathrm{t} : \hat{\mathbf{a}}^{\mathrm{T}}\mathbf{x} \le \hat{b}\}$ and $\mathrm{t}_R = \{\mathbf{x} \in \mathrm{t} : \hat{\mathbf{a}}^{\mathrm{T}}\mathbf{x} > \hat{b}\}$. These child nodes become new parent nodes at the next level of the tree and can be further refined in the same manner until a desired depth is reached. To obtain a maximal decision tree T_K of depth K, the procedure is iterated K times or until either (i) the node contains a single data point $(y_i, \mathbf{x}_i^{\mathrm{T}})$ or (ii) all input values $\mathbf{x}_i$ and/or all response values y_i within the node are the same. The maximal decision tree with maximum depth is denoted by $T_{\max}$. An illustration of a maximal oblique decision

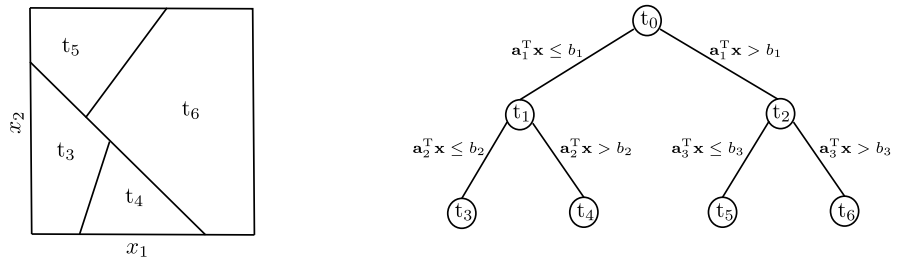

FIG. 1. *A maximal oblique decision tree with depth $K = 2$ in $p = 2$ dimensions. Splits occur along hyperplanes of the form $a_1x_1 + a_2x_2 = b$.*

tree with depth $K = 2$ is shown in Figure 1. For contrast, in Figure 2, we show a maximal axis-aligned decision tree with depth $K = 2$.

In a conventional regression problem, where the goal is to estimate the conditional mean response $\mu(\mathbf{x})$, the canonical tree output for $\mathbf{x} \in \mathrm{t}$ is $\overline{y}_{\mathrm{t}}$, that is, if T is a decision tree, then

$$\widehat{\mu}(T)(\mathbf{x}) = \overline{y}_{\mathrm{t}} = \frac{1}{n(\mathrm{t})} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i, \tag{2}$$

where $n(\mathrm{t})$ denotes the number of observations in the node t. However, one can aggregate the data in each node in a number of ways, depending on the form of the target estimand. In the most general setting, under weak assumptions, all of our forthcoming theory holds when the node output is the result of a least squares projection onto the linear span of a finite dictionary $\mathcal{H}$ that includes the constant function (e.g., polynomials, splines), that is, $\hat{y}_{\mathrm{t}} \in \operatorname{argmin}_{h \in \operatorname{span}(\mathcal{H})} \sum_{\mathbf{x}_i \in \mathrm{t}} (y_i - h(\mathbf{x}_i))^2$.

One of the main practical issues with oblique CART is that the computational complexity of minimizing the squared error in (1) in each node is extremely demanding (in fact, it is NP-hard). For example, if we desire to split a node t with $n(\mathrm{t})$ observations for axis-aligned CART, an exhaustive search would require at most $p \cdot n(\mathrm{t})$ evaluations, whereas oblique CART would require a prodigious $2^p \binom{n(t)}{p}$ evaluations [36].

To deal with these computational demands, Breiman et al. [11] first suggested a method for inducing oblique decision trees. They use a fully deterministic hill-climbing algorithm to search for the best oblique split. A backward feature elimination process is also carried out to delete irrelevant features from the split. Heath, Kasif and Salzberg [24] propose a simulated annealing optimization algorithm, which uses randomization to search for the best split to potentially avoid getting stuck in a local optimum. Murthy, Kasif and Salzberg [36] use a combination of deterministic hill-climbing and random perturbations in an attempt to find a good hyperplane. See Brodley and Utgoff [12] for additional variations on these algorithms.

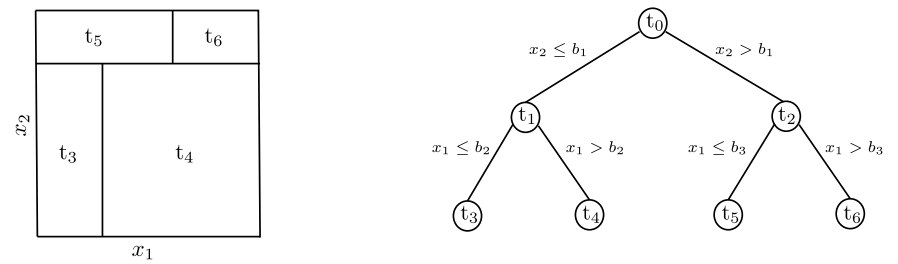

FIG. 2. *A maximal axis-aligned decision tree with depth $K = 2$ in $p = 2$ dimensions. Splits occur along individual covariates of the form $x_j = b$ for $j = 1, 2$.*

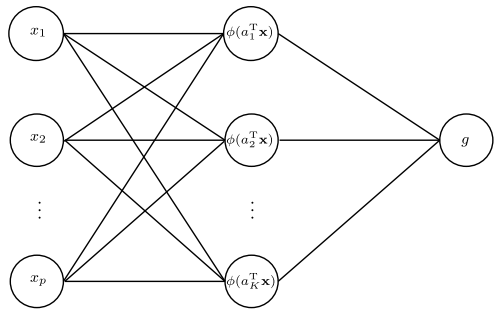

FIG. 3. *A single hidden layer neural network with K hidden nodes.*

Other works employ statistical techniques like linear discriminant analysis (LDA) [30–32], principle components analysis (PCA) [33, 40] and random projections [44].

While not the focus of the present paper, regarding nongreedy training, other researchers have attempted to find globally optimal tree solutions using linear programming [4] or mixed-integer linear programming [5, 7]. It should be clear that all of our results hold verbatim for optimal trees, as greedy implementations belong to the same feasible set. While usually better than greedy trees in terms of predictive performance, scalability to large data sets is the most salient obstacle with globally optimal trees. Moreover, on a qualitative level, a globally optimal tree arguably detracts from the interpretability, as humans, in contrast, often exhibit bounded rationality and, therefore, make decisions in a more sequential (rather than anticipatory) manner [26, and references therein]. Relatedly, another training technique is based on constructing deep neural networks that realize oblique decision trees [29, 46] and then utilizing tools designed for training neural networks.

While there has been a plethora of greedy algorithms over the past 30 years for training oblique decision trees, the literature is essentially silent on their statistical properties. For instance, assuming one can come close to optimizing (1), what types of regression functions can greedy oblique trees estimate and how well?

1.2. *Ridge expansions.* Many empirical studies reveal that oblique trees generally produce smaller trees with better accuracy compared to axis-aligned trees [24, 36] and can often be comparable, in terms of performance, to neural networks [6, 8, 9]. Intuitively, allowing a tree-building system to use both oblique and axis-aligned splits broadens its flexibility. To theoretically showcase these qualities and make comparisons with other procedures (such as neural networks and projection pursuit regression), we will consider modeling μ with finite linear combinations of ridge functions, that is, the library

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \sum_{k=1}^{M} g_k\big(\mathbf{a}_k^{\mathrm{T}}\mathbf{x}\big), \mathbf{a}_k \in \mathbb{R}^p, g_k : \mathbb{R} \mapsto \mathbb{R}, k = 1, \ldots, M, M \geq 1, \|g\|_{\mathcal{L}_1} < \infty \right\},$$

where $\|\cdot\|_{\mathcal{L}_1}$ is a total variation norm that is defined in Section 2.1. This library encompasses the functions produced from projection pursuit regression, and more specifically—by taking $g_k(z) = \phi(z - b_k)$, where ϕ is a fixed activation function, such as a sigmoid function or ReLU, and $b_k \in \mathbb{R}$ is a bias parameter—single hidden layer feed-forward neural networks. A graphical representation of such a neural network is provided in Figure 3. A neural network forms predictions according to distributed hierarchical representations of the data, whereas a decision tree uses hierarchical decision rules (cf., Figures 1 and 2).

Since the first version of our manuscript was released on `arXiv`, several subsequent papers have employed our novel theoretical and methodological statistical framework to derive consistency results for decision trees and related methods. For example, [47] applies our core

ideas and proof techniques to deduce a consistency result for oblique decision trees in low-dimensional settings (cf., Corollary 2.4 below), but under stronger assumptions on the target function class $\mathcal{G}$ and without accounting for the underlying optimization constraints (cf., our novel optimization framework in Section 2.2). [39] also applies our core ideas and proof techniques to deduce a consistency result for axis-aligned decision trees within an alternative computation framework, but under stronger assumptions on the target function class $\mathcal{G}$. Finally, [37] and [17], among others (see their references), study consistency of deep neural network methods using similar notions of Hilbert function spaces and total variation norms as our paper does for adaptive decision trees and shallow neural networks, but without accounting for the underlying optimization constraints. In particular, [37] also shows that neural networks are able to adapt to sparsity in the data (cf., Section 3 below).

2. Main results. We first introduce notation and assumptions that are used throughout the remainder of the paper.

2.1. *Notation and assumptions.* For a function $f : \mathbb{R}^p \to \mathbb{R}$, we define $\mathbf{f} = (f(\mathbf{x}_1), \dots, f(\mathbf{x}_n))^{\mathrm{T}}$ to be the $n \times 1$ vector of f evaluated at the design points $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n \times p}$. Likewise, we use $\widehat{\boldsymbol{\mu}}(T_K)$ to denote the $n \times 1$ vector of fitted values of $\widehat{\mu}(T_K)$. For functions $f, g \in \mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$, let $\|f\|^2 = \int_{\mathcal{X}} (f(\mathbf{x}))^2 \, \mathrm{d}\mathbb{P}_{\mathbf{x}}(\mathbf{x})$ be the squared $\mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$ norm and let $\|\mathbf{f}\|_n^2 = \frac{1}{n} \sum_{i=1}^n (f(\mathbf{x}_i))^2$ denote the squared norm with respect to the empirical measure on the data. Let $\langle \mathbf{f}, \mathbf{g} \rangle_n = \frac{1}{n} \sum_{i=1}^n f(\mathbf{x}_i) g(\mathbf{x}_i)$ denote the inner product with respect to the empirical measure on the data. The response data vector $\mathbf{y} = (y_1, \dots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ is viewed as a relation, defined on the design matrix $\mathbf{X}$, that associates $\mathbf{x}_i$ with y_i. Thus, for example, $\|\mathbf{y} - \mathbf{f}\|_n^2 = \frac{1}{n} \sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$ and $\langle \mathbf{y}, \mathbf{f} \rangle_n = \frac{1}{n} \sum_{i=1}^n y_i f(\mathbf{x}_i)$. We use $[T]$ to denote the collection of internal (nonterminal) nodes and $\{\mathrm{t} : \mathrm{t} \in T\}$ to denote the terminal nodes of the tree. The cardinality of a set A is denoted by $|A|$.

We define the total variation of a ridge function $\mathbf{x} \mapsto h(\mathbf{a}^{\mathrm{T}}\mathbf{x})$ with $\mathbf{a} \in \mathbb{R}^p$ and $h : \mathbb{R} \to \mathbb{R}$ in the node t as

$$V(h, \mathbf{a}, \mathrm{t}) = \sup_{\mathcal{P}} \sum_{\ell=0}^{|\mathcal{P}|-1} \big| h(z_{\ell+1}) - h(z_\ell) \big|,$$

where the supremum is over all partitions $\mathcal{P} = \{z_0, z_1, \dots, z_{|\mathcal{P}|}\}$ of the interval $I(\mathbf{a}, \mathrm{t}) = [\min_{\mathbf{x} \in \mathrm{t}} \mathbf{a}^{\mathrm{T}}\mathbf{x}, \max_{\mathbf{x} \in \mathrm{t}} \mathbf{a}^{\mathrm{T}}\mathbf{x}] \subset \mathbb{R}$ (we allow for the possibility that one or both of the endpoints is infinite). If the function h is smooth, then $V(h, \mathbf{a}, \mathrm{t})$ admits the familiar integral representation $\int_{I(\mathbf{a}, \mathrm{t})} |h'(z)| \, \mathrm{d}z$. We can then define the $\mathcal{L}_1$ norm of an additive function $h(\mathbf{x}) = \sum_{k=1}^M h_k(\mathbf{x})$ as

$$\|h\|_{\mathcal{L}_1} = \sum_{k=1}^M V(h_k, \mathbf{a}_k, \mathrm{t}).$$

Central to our results is the $\mathcal{L}_1$ total variation norm of $f \in \mathcal{F} = \mathrm{cl}(\mathcal{G})$ in the node t, the closure being taken in $\mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$. This quantity captures the local capacity of a function in $\mathcal{F}$. It is defined as

$$\|f\|_{\mathcal{L}_1(\mathrm{t})} := \lim_{\varepsilon \downarrow 0} \inf_{g \in \mathcal{G}} \left\{ \sum_{k=1}^M V(g_k, \mathbf{a}_k, \mathrm{t}) : g(\mathbf{x}) = \sum_{k=1}^M g_k(\mathbf{a}_k^{\mathrm{T}}\mathbf{x}), \, \|f - g\| \le \varepsilon \right\}.$$

For simplicity, we write $\|f\|_{\mathcal{L}_1}$ for $\|f\|_{\mathcal{L}_1(\mathcal{X})}$. This norm may be thought of as an ℓ_1 norm on the coefficients in a representation of the function f by elements of a normalized dictionary of ridge functions. A classic result of Barron [1] shows that, for any function f defined on $\mathcal{X} = [0, 1]^p$, we have the bound $\|f\|_{\mathcal{L}_1} \lesssim \int \|\boldsymbol{\theta}\|_{\ell_1} |\tilde{f}(\boldsymbol{\theta})| \, \mathrm{d}\boldsymbol{\theta}$, where $\tilde{f}$ is the Fourier transform

of f and $\|\cdot\|_{\ell_1}$ is the usual ℓ_1 norm of a vector in $\mathbb{R}^p$. Furthermore, there exists an M-term linear combination of sigmoidal ridge functions in $\mathcal{G}$ whose $\mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$ distance from f is $O(\|f\|_{\mathcal{L}_1}/\sqrt{M})$.

2.2. *Computational framework.* As mentioned earlier, it is challenging to find the direction $\hat{\mathbf{a}}$ that optimizes $\widehat{\Delta}(b, \mathbf{a}, \mathrm{t})$. Many of the aforementioned computational papers address the problem by restricting the search space to a more tractable subset of candidate directions $\mathcal{A}_\mathrm{t}$ with sparsity

$$\sup\{\|\mathbf{a}\|_{\ell_0} : \mathbf{a} \in \mathcal{A}_\mathrm{t}\} \le d,$$

for some positive integer d, where $\|\mathbf{a}\|_{\ell_0}$ counts the number of nonzero coordinates of $\mathbf{a}$. Because such search strategies are sometimes unlikely to find the global maximum, we theoretically measure their success by specifying a suboptimality (slackness) parameter $\kappa \in (0, 1]$ and considering the probability $P_{\mathcal{A}_\mathrm{t}}(\kappa)$ that the maximum of $\widehat{\Delta}(b, \mathbf{a}, \mathrm{t})$ over $\mathbf{a} \in \mathcal{A}_\mathrm{t} \subseteq \mathbb{R}^p$ is within a factor κ of the maximum of $\widehat{\Delta}(b, \mathbf{a}, \mathrm{t})$ on the unrestricted parameter space, $\mathbf{a} \in \mathbb{R}^p$. That is, to theoretically quantify the suboptimality of the chosen hyperplane, we measure

$$P_{\mathcal{A}_\mathrm{t}}(\kappa) = \mathbb{P}_{\mathcal{A}_\mathrm{t}}\Big(\max_{(b,\mathbf{a})\in\mathbb{R}\times\mathcal{A}_\mathrm{t}} \widehat{\Delta}(b, \mathbf{a}, \mathrm{t}) \ge \kappa \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, \mathrm{t})\Big),$$

where $\mathbb{P}_{\mathcal{A}_\mathrm{t}}$ denotes the probability with respect to the randomness in the search spaces $\mathcal{A}_\mathrm{t}$, conditional on the data. The maximum of $\widehat{\Delta}(b, \mathbf{a}, \mathrm{t})$ over $(b, \mathbf{a})$ is achieved because the number of distinct values of $\widehat{\Delta}(b, \mathbf{a}, \mathrm{t})$ is finite (at most the number of ways of dividing n observations into two groups, or, $2^n - 1$).

Another way of thinking about $P_{\mathcal{A}_\mathrm{t}}(\kappa)$ is that it represents the degree of optimization misspecificity of $\mathcal{A}_\mathrm{t}$ for the form of the global optimum $\hat{\mathbf{a}}$. For example, if $\mathcal{A}_\mathrm{t} = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p\}$ is the collection of standard basis vectors in $\mathbb{R}^p$, then $d = 1$ and we believe that the true optimal solution $\hat{\mathbf{a}} \in \mathcal{A}_\mathrm{t}$ corresponds to axis-aligned CART, then $P_{\mathcal{A}_t}(\kappa) = 1$ for all values of κ.

The definition of $P_{\mathcal{A}_\mathrm{t}}(\kappa)$ can also be understood as a hypothesis test. Consider the regression model $y = \beta_1 \mathbb{1}(\mathbf{a}^\mathrm{T}\mathbf{x} \le b) + \beta_2 \mathbb{1}(\mathbf{a}^\mathrm{T}\mathbf{x} > b) + \varepsilon$ with independent Gaussian noise $\varepsilon \sim N(0, \sigma^2)$. Set the null hypothesis $H_0 : \hat{\mathbf{a}} \in \mathcal{A}_\mathrm{t}$. Then, using the likelihood ratio test with threshold proportional to $1 - \kappa$, $P_{\mathcal{A}_\mathrm{t}}(\kappa)$, is the likelihood of failing to reject the null hypothesis. It follows that the smaller κ is, the more likely it is that we will reject the null hypothesis that $\hat{\mathbf{a}}$ belongs to $\mathcal{A}_\mathrm{t}$.

The collection $\mathcal{A}_\mathrm{t}$ of candidate directions can be chosen in many different ways; we discuss some examples next.

- *Deterministic.* If $\mathcal{A}_\mathrm{t}$ is nonrandom, then $P_{\mathcal{A}_\mathrm{t}}(\kappa)$ is either zero or one for any $\mathcal{A}_\mathrm{t} \subset \mathbb{R}^p$, and if $\mathcal{A}_\mathrm{t} = \mathbb{R}^p$, then $P_{\mathcal{A}_\mathrm{t}}(\kappa) = 1$ for all $\kappa \in (0, 1]$. For the latter case, one can use strategies based on mixed-integer optimization (MIO) Zhu et al. [49], Dunn [18], Bertsimas and Dunn [5]. In particular, Dunn [18] presents a global MIO formulation for regression trees with squared error that can also be implemented greedily within each node. Separately, in order to improve interpretability, it may be of interest to restrict the coordinates of $\hat{\mathbf{a}}$ to be integers. Using the hyperplane separation theorem and the fact that constant multiples of vectors in $\mathbb{Z}^p$ are dense in $\mathbb{R}^p$, it can easily be shown that if $\mathcal{A}_\mathrm{t} = \mathbb{Z}^p$, then $P_{\mathcal{A}_\mathrm{t}}(\kappa) = 1$ for all $\kappa \in (0, 1]$. An integer-valued search space may also lend itself to optimization strategies based on integer programming.
- *Purely random.* The most naïve and agnostic way to construct $\mathcal{A}_\mathrm{t}$ is to generate the directions uniformly at random. For example, with axis-aligned CART where the global search space consists of the p standard basis vectors $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p\}$, if $\mathcal{A}_\mathrm{t}$ is generated by selecting $m(\le p)$ standard basis vectors uniformly at random without replacement (as

is done with random forests [10]), then $P_{\mathcal{A}_t}(\kappa) \geq \binom{p-1}{m-1}/\binom{p}{m} = m/p$ for all $\kappa \in (0,1]$. For more complex global search spaces (e.g., oblique), it is quite likely that a purely random selection will yield very small $P_{\mathcal{A}_t}(\kappa)$. For example, if the global search space is $\{\mathbf{a} \in \mathbb{R}^p : \|\mathbf{a}\|_{\ell_0} = d\}$ and $\mathcal{A}_t$ is generated by selecting m (distinct) sets $S_k \subset \{1, 2, \ldots, p\}$ with $|S_k| = d$ uniformly at random without replacement and setting $\mathcal{A}_t = \bigcup_k \{\mathbf{a} \in \mathbb{R}^p : a_j = 0,\ j \notin S_k\}$, then $P_{\mathcal{A}_t}(\kappa) \geq m/\binom{p}{d} \approx 0$ for all $\kappa \in (0,1]$. This has direct consequences for the predictive performance, since, as we shall see (Section 2.4), the expected risk is inflated by the reciprocal probability $1/P_{\mathcal{A}_t}(\kappa)$. Thus, generating $\mathcal{A}_t$ in a principled manner is important for producing small risk.

- *Data-dependent.* Perhaps the most interesting and useful way of generating informative candidate directions in $\mathcal{A}_t$ is to take a data-driven approach. One possibility is to use dimensionality reduction techniques, such as PCA, LDA and Lasso, on a separate sample $\{(\tilde{y}_i, \tilde{\mathbf{x}}_i^{\mathrm{T}}) : \tilde{\mathbf{x}}_i \in \mathrm{t}\}$. The search space $\mathcal{A}_t$ can then be defined in terms of the top principle components produced by PCA or LDA, or similarly, in terms of the relevant coordinates selected by Lasso. Additional randomization can also be introduced by incorporating, for example, sparse random projections or random rotations [44]. On an intuitive level, we expect these statistical methods that aim to capture variance in the data to produce good optimizers of the objective function. Indeed, empirical studies with similar constructions provide evidence for their efficacy over purely random strategies [21, 33, 40].

In order to control the predictive performance of the decision tree theoretically, we assume the researcher has chosen a meaningful method for selecting candidate directions $\mathcal{A}_t$, either with prior knowledge based on the context of the problem, or with an effective data-driven strategy.

2.3. *Orthogonal tree expansions.* We now present a technical result about the construction of trees that is crucial in proving our main results. While Lemma 2.1 below focuses on the special case of constant fit at the terminal nodes for concreteness, all proofs (see Section 6 and the Supplementary Material [14]) are given in full generality. To be more precise, our results in the Appendix allow for any finite-dimensional least squares fit at the terminal nodes, and thus give a general orthogonal tree expansion in the function space for adaptive oblique decision trees, covering canonical adaptive axis-aligned decision trees as a special case.

Lemma 2.1 shows that the tree output $\widehat{\mu}(T)(\mathbf{x})$ is equal to the empirical orthogonal projection of $\mathbf{y}$ onto the linear span of orthonormal decision stumps, defined as

$$\psi_{\mathrm{t}}(\mathbf{x}) = \frac{\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) n(\mathrm{t}_R) - \mathbb{1}(\mathbf{x} \in \mathrm{t}_R) n(\mathrm{t}_L)}{\sqrt{w(\mathrm{t}) n(\mathrm{t}_L) n(\mathrm{t}_R)}}, \tag{3}$$

for internal nodes $\mathrm{t} \in [T]$, where $w(\mathrm{t}) = n(\mathrm{t})/n$ denotes the proportion of observations that are in t. By slightly expanding the notion of an internal node to include the empty node (i.e., the empty set), we define $\psi_{\mathrm{t}}(\mathbf{x}) \equiv 1$ if t is the empty node, in which case the tree outputs the grand mean of all the response values. The decision stump ψ_{t} in (3) is produced from the Gram–Schmidt orthonormalization of the functions $\{\mathbb{1}(\mathbf{x} \in \mathrm{t}), \mathbb{1}(\mathbf{x} \in \mathrm{t}_L)\}$ with respect to the empirical inner product space:

$$\left\{ \frac{\mathbb{1}(\mathbf{x} \in \mathrm{t})}{\|\mathbb{1}(\mathbf{x} \in \mathrm{t})\|_n}, \frac{\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) - \frac{\langle \mathbb{1}(\mathbf{x} \in \mathrm{t}_L), \mathbb{1}(\mathbf{x} \in \mathrm{t}) \rangle_n}{\|\mathbb{1}(\mathbf{x} \in \mathrm{t})\|_n^2} \mathbb{1}(\mathbf{x} \in \mathrm{t})}{\|\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) - \frac{\langle \mathbb{1}(\mathbf{x} \in \mathrm{t}_L), \mathbb{1}(\mathbf{x} \in \mathrm{t}) \rangle_n}{\|\mathbb{1}(\mathbf{x} \in \mathrm{t})\|_n^2} \mathbb{1}(\mathbf{x} \in \mathrm{t})\|_n} \right\}$$
$$= \left\{ \frac{\mathbb{1}(\mathbf{x} \in \mathrm{t})}{\sqrt{w(\mathrm{t})}}, \frac{\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) n(\mathrm{t}_R) - \mathbb{1}(\mathbf{x} \in \mathrm{t}_R) n(\mathrm{t}_L)}{\sqrt{w(\mathrm{t}) n(\mathrm{t}_L) n(\mathrm{t}_R)}} \right\}.$$

We refer the reader to Section 6 for an orthonormal decomposition of the tree output that holds in a much more general setting (i.e., when the node output is the least squares projection onto the linear span of a finite dictionary).

LEMMA 2.1. *If T is a decision tree constructed with CART methodology (either axis-aligned or oblique), then its output (2) admits the orthogonal expansion*

$$\widehat{\mu}(T)(\mathbf{x}) = \sum_{\mathrm{t}\in[T]} \langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n \psi_{\mathrm{t}}(\mathbf{x}), \tag{4}$$

where $\boldsymbol{\psi}_{\mathrm{t}} = (\psi_{\mathrm{t}}(\mathbf{x}_1), \ldots, \psi_{\mathrm{t}}(\mathbf{x}_n))^{\mathrm{T}}$. By construction, $\|\boldsymbol{\psi}_{\mathrm{t}}\|_n = 1$ and $\langle \boldsymbol{\psi}_{\mathrm{t}}, \boldsymbol{\psi}_{\mathrm{t}'} \rangle_n = 0$ for distinct internal nodes t and t' in $[T]$. In other words, $\widehat{\boldsymbol{\mu}}(T)$ is the empirical orthogonal projection of $\mathbf{y}$ onto the linear span of $\{\boldsymbol{\psi}_{\mathrm{t}}\}_{\mathrm{t}\in[T]}$. Furthermore,

$$\big|\langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n\big|^2 = \widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, \mathrm{t}). \tag{5}$$

REMARK 1 (Connection to sieve estimation literature). Another way of thinking about CART is through the lens of least squares sieve estimation. For example, for a fixed but otherwise arbitrary ordering of the internal nodes of T, suppose $\boldsymbol{\Psi}$ is the $n \times |[T]|$ data matrix $[\psi_{\mathrm{t}}(\mathbf{x}_i)]_{1\le i\le n, \mathrm{t}\in[T]}$ and $\boldsymbol{\Psi}(\mathbf{x})$ is the $|[T]| \times 1$ feature vector $(\psi_{\mathrm{t}}(\mathbf{x}))_{\mathrm{t}\in[T]}$. Then

$$\widehat{\mu}(T)(\mathbf{x}) = \boldsymbol{\Psi}(\mathbf{x})^{\mathrm{T}}\big(\boldsymbol{\Psi}^{\mathrm{T}}\boldsymbol{\Psi}\big)^{-1}\boldsymbol{\Psi}^{\mathrm{T}}\mathbf{y} = \boldsymbol{\Psi}(\mathbf{x})^{\mathrm{T}}\boldsymbol{\Psi}^{\mathrm{T}}\mathbf{y}.$$

From this perspective, standard sieve estimation and inference theory [15, 25] cannot be applied to studying the statistical properties of $\widehat{\mu}(T)(\mathbf{x})$ because the implied (random) basis functions depend on the entire sample $(\mathbf{y}, \mathbf{X})$ through the adaptive (recursive) split regions underlying the decision tree construction (i.e., the induced random partitioning).

Lemma 2.1 suggests that there may be some connections between oblique CART and sequential greedy optimization in Hilbert spaces. Indeed, our analysis of the oblique CART algorithm suggests that it can be viewed as a local orthogonal greedy procedure in which one iteratively projects the data onto the space of all constant predictors within a greedily obtained node. The algorithm also has similarities to forward-stepwise regression because, at each current node t, it grows the tree by selecting a feature, $\boldsymbol{\psi}_{\mathrm{t}}$, most correlated with the residuals, $(y_i - \overline{y}_{\mathrm{t}})\mathbb{1}(\mathbf{x}_i \in \mathrm{t})$, per (5) and (1), and then adding that chosen feature along with its coefficient back to the tree output in (4).

The proofs show that this local greedy approach has a very similar structure to standard global greedy algorithms in Hilbert spaces. Indeed, the reader familiar with greedy algorithms in Hilbert spaces for overcomplete dictionaries will recognize some similarities in the analysis (see the *orthogonal greedy algorithm* [3] in which one iteratively projects the data onto the linear span of a finite collection of greedily obtained dictionary elements). As with all orthogonal expansions, the decomposition of $\widehat{\boldsymbol{\mu}}(T_K)$ in Lemma 2.1 allows one to write down a recursive expression for the training error. That is, from $\widehat{\boldsymbol{\mu}}(T_K) = \widehat{\boldsymbol{\mu}}(T_{K-1}) + \sum_{\mathrm{t}\in T_{K-1}} \langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n \boldsymbol{\psi}_{\mathrm{t}}$, one obtains the identity

$$\big\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\big\|_n^2 = \big\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_{K-1})\big\|_n^2 - \sum_{\mathrm{t}\in T_{K-1}} \big|\langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n\big|^2. \tag{6}$$

Furthermore, using the fact that $\langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n$ is the result of a local maximization, namely the equivalence (5) in Lemma 2.1, one can construct an empirical probability measure Π on $(b, \mathbf{a})$ and lower bound $|\langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n|^2$ by $\int \widehat{\Delta}(b, \mathbf{a}, \mathrm{t})\, \mathrm{d}\Pi(b, \mathbf{a})$, which is itself further lower bounded by an appropriately scaled squared nodewise excess training error (see Lemma 6.1). Combining this with (6), we can establish a useful training error bound. We formalize this result next.

2.4. *Training error bound for oblique CART.* Applying the techniques outlined earlier, we can show the following result (Lemma 2.2) on the training error of the tree. Our result provides an algorithmic guarantee, namely that the expected excess training error of a depth K tree constructed with oblique CART methodology decays like $1/K$, and with additional assumptions (see Section 3), like $4^{-K/q}$ for some $q > 2$. To the best of our knowledge, this result is the first of its kind for oblique CART. The math behind it is surprisingly simple; in particular, unlike past work on axis-aligned decision trees, there is no need to directly analyze the partition that is induced by recursively splitting, which often entails showing that certain local (i.e., node-specific) empirical quantities concentrate around their population level versions [16, 42, 43, 45].

For the following statements, the output of a depth K tree T_K constructed with oblique CART methodology using the search spaces $\{\mathcal{A}_t : t \in [T]\}$ is denoted $\widehat{\boldsymbol{\mu}}(T_K)$. Throughout the paper, we use $\mathbb{E}$ to denote the expectation with respect to the joint distribution of the (possibly random) search spaces $\{\mathcal{A}_t : t \in [T_K]\}$ and the data.

LEMMA 2.2 (Training error bound for oblique CART). *Let* $\mathbb{E}[y^2 \log(1 + |y|)] < \infty$ *and* $g \in \mathcal{F}$ *with* $\|g\|_{\mathcal{L}_1} < \infty$. *Then, for any* $K \geq 1$,

$$\mathbb{E}\big[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2\big] \leq \mathbb{E}\big[\|\mathbf{y} - \mathbf{g}\|_n^2\big] + \frac{\|g\|_{\mathcal{L}_1}^2 \mathbb{E}[\max_{t \in [T_K]} P_{\mathcal{A}_t}^{-1}(\kappa)]}{\kappa K}. \tag{7}$$

For this result to be nonvacuous, the only additional assumption needed is that the largest of the reciprocal probabilities, $P_{\mathcal{A}_t}^{-1}(\kappa)$, are integrable with respect to the data and (possibly random) search spaces. A simple sufficient condition is that the splitting probabilities are almost surely bounded away from zero, which we record in the following assumption for future reference.

ASSUMPTION 1 (Nonzero splitting probabilities). The splitting probabilities are uniformly bounded away from zero. That is,

$$\inf_{n \geq 1} \inf_{t \in [T_{\max}]} P_{\mathcal{A}_t}(\kappa) > 0 \quad \text{a.s.}$$

Section 2.2 discusses optimization algorithms/approaches that would satisfy Assumption 1, and, more generally, that would guarantee $\mathbb{E}[\max_{t \in [T_K]} P_{\mathcal{A}_t}^{-1}(\kappa)] < \infty$.

2.5. *Pruning.* Without proper tuning of the depth K, the tree T_K can very easily become overly complicated, causing its output $\widehat{\mu}(T_K)(\mathbf{x})$ to generalize poorly to unseen data. While one could certainly select good choices of K via a holdout method, in practice, complexity modulation is often achieved through pruning. We first introduce some additional concepts, and then go on to 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 iteratively merging 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 $T_{\max}$. Recall that the number of terminal nodes in a tree T is denoted $|T|$. As shown in Breiman et al. [11, Section 10.2], the smallest minimizing subtree for the penalty coefficient $\lambda = \lambda_n \geq 0$,

$$T_{\text{opt}} \in \operatorname*{argmin}_{T \preceq T_{\max}} \big\{\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T)\|_n^2 + \lambda |T|\big\}, \tag{8}$$

exists and is unique (smallest in the sense that if T_{opt} optimizes the penalized risk of (8), then $T_{\text{opt}} \preceq T$ for every pruned subtree T of $T_{\max}$). For a fixed λ, the optimal subtree T_{opt} can be found efficiently by weakest link pruning, that is, by successively collapsing the internal

node that decreases $\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T)\|_n^2$ the most, until we arrive at the single-node tree consisting of the root node. This method enumerates a finite list of trees for which the objective function can then be evaluated to find the optimal subtree. Good values of λ can be selected using cross-validation on a holdout subset of data, for example. See Mingers [34] for a description of various pruning algorithms.

We now present our main consistency and convergence rate results for both pruned and un-pruned oblique trees.

2.6. *Oracle inequality for oblique CART.* Our main result establishes an adaptive prediction risk bound (also known as an *oracle inequality*) for oblique CART under model misspecification; that is, when the true model may not belong to $\mathcal{F}$. Essentially, the result shows that oblique CART performs almost as if it was finding the best approximation of the true model with ridge expansions, while accounting for the goodness-of-fit and descriptive complexity relative to sample size. To bound the integrated mean squared error (IMSE), the training error bound from Lemma 2.2 is coupled with tools from empirical process theory [22] for studying partition-based estimators. Our results rely on the following assumption regarding the data generating process.

ASSUMPTION 2 (Exponential tails of the conditional response variable). The conditional distribution of y given $\mathbf{x}$ has exponentially decaying tails. That is, there exist positive constants c_1, c_2, γ and M, such that for all $\mathbf{x} \in \mathcal{X}$,

$$\mathbb{P}\big(|y| > B + M \mid \mathbf{x}\big) \le c_1 \exp\big(-c_2 B^{\gamma}\big), \quad B \ge 0.$$

In particular, note that $\gamma = 1$ for subexponential data, $\gamma = 2$ for sub-Gaussian data and $\gamma = \infty$ for bounded data. Using the layer cake representation for expectations, that is, $|\mu(\mathbf{x})| \le \mathbb{E}[|y| \mid \mathbf{x}] = \int_0^\infty \mathbb{P}(|y| \ge z \mid \mathbf{x})\,\mathrm{d}z$, Assumption 2 implies that the conditional mean is uniformly bounded:

$$\sup_{\mathbf{x}\in\mathcal{X}} \big|\mu(\mathbf{x})\big| \le M + c_1 \int_0^\infty \exp\big(-c_2 z^{\gamma}\big)\,\mathrm{d}z = M' < \infty.$$

THEOREM 2.3 (Oracle inequality for oblique trees). *Let Assumption 2 hold. Then, for any $K \ge 1$,*

$$\begin{aligned}(9)\qquad &\mathbb{E}\big[\|\mu - \widehat{\mu}(T_K)\|^2\big] \\ &\le 2 \inf_{f\in\mathcal{F}} \left\{ \|\mu - f\|^2 + \frac{\|f\|_{\mathcal{L}_1}^2 \mathbb{E}[\max_{\mathrm{t}\in[T_K]} P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)]}{\kappa K} + C\frac{2^K d \log(np/d)\log^{4/\gamma}(n)}{n} \right\},\end{aligned}$$

where $C = C(c_1, c_2, \gamma, M)$ is a positive constant. Furthermore, if the penalty coefficient satisfies $\lambda_n \asymp (d/n)\log(np/d)\log^{4/\gamma}(n)$, then

$$\begin{aligned}(10)\qquad \mathbb{E}\big[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2\big] \le 2 \inf_{K\ge 1, f\in\mathcal{F}} \Bigg\{ &\|\mu - f\|^2 + \frac{\|f\|_{\mathcal{L}_1}^2 \mathbb{E}[\max_{\mathrm{t}\in[T_K]} P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)]}{\kappa K} \\ &+ C\frac{2^K d \log(np/d)\log^{4/\gamma}(n)}{n} \Bigg\}.\end{aligned}$$

Consistency of oblique trees follows from Theorem 2.3 under the additional assumption that the splitting probabilities are bounded away from zero (Assumption 1) and that the depth K grows appropriately with the sample size.

COROLLARY 2.4 (Consistency for fixed dimension). *Let Assumptions 1 and 2 hold. If* $K \asymp \log n$, *then*

$$\lim_{n\to\infty} \mathbb{E}\big[\|\mu - \widehat{\mu}(T_K)\|^2\big] = 0,$$

and if the penalty coefficient satisfies $\lambda_n \asymp (d/n)\log(np/d)\log^{4/\gamma}(n)$, *then*

$$\lim_{n\to\infty} \mathbb{E}\big[\|\mu - \widehat{\mu}(T_{\text{opt}})\|^2\big] = 0.$$

While Corollary 2.4 shows that oblique trees are consistent for fixed dimension p, it does not provide a rate of convergence. Under a few additional assumptions, however, Theorem 2.3 implies that the oblique tree is consistent with a logarithmic rate of convergence even when the dimension grows with the sample size.

COROLLARY 2.5 (Consistency for possibly growing dimension). *Let Assumptions 1 and 2 hold and suppose* $\{\boldsymbol{\mu}_n\}$ *is a sequence of regression functions that belong to* $\mathcal{F}$ *with* $\sup_n \|\boldsymbol{\mu}_n\|_{\mathcal{L}_1} < \infty$. *Assume furthermore that* $d = p = O(n^{1-\xi})$ *for some* $\xi \in (0,1)$. *If* $K \asymp \log n$, *then*

$$\mathbb{E}\big[\|\mu_n - \widehat{\mu}(T_K)\|^2\big] = O\big((\log n)^{-1}\big),$$

and if the penalty coefficient satisfies $\lambda_n \asymp (d/n)\log(np/d)\log^{4/\gamma}(n)$, *then*

$$\mathbb{E}\big[\|\mu_n - \widehat{\mu}(T_{\text{opt}})\|^2\big] = O\big((\log n)^{-1}\big).$$

The results also hold trivially if d *and* p *are fixed.*

REMARK 2 (Connection to adaptive axis-aligned decision trees). By considering elements of $\mathcal{G}$ with $\mathbf{a}_k = \mathbf{e}_k$ (the standard basis vectors in $\mathbb{R}^p$) and $M = p$, we recover the additive library

$$\mathcal{F}^{\text{add}} = \left\{ f(\mathbf{x}) = \sum_{j=1}^{p} f_j(x_j) : f_j : \mathbb{R} \mapsto \mathbb{R} \right\}.$$

Additive models have played an important role in the development of theory for CART. For example, [42] show consistency of axis-aligned CART for fixed-dimensional additive models. More recent work has tried to illustrate the adaptive properties of axis-aligned CART on sparse additive models with growing dimensionality [16, 27, 28, 43], some of which can be recovered as a special case of our more general theory. To see this, note that global optimization of the splitting criterion (1) is feasible with axis-aligned CART ($d = 1$), and hence $\kappa = 1$ and $P_{\mathcal{A}_t}(\kappa) = 1$. Then, according to (10), since $d = 1$, the pruned tree estimator is consistent for regression functions in the class $\mathcal{F}^{\text{add}}$ even in the so-called NP-dimensionality regime, where $\log(p) = O(n^{1-\xi})$ for some $\xi \in (0,1)$. This result was previously established in [28] for axis-aligned CART.

These sort of high-dimensional consistency guarantees are not possible with nonadaptive procedures that do not automatically adjust the amount of smoothing along a particular dimension according to how much the covariate affects the response variable. Such procedures perform local estimation at a query point using data that are close in every single dimension, making them prone to the curse of dimensionality even if the true model is sparse (typical minimax rates [22] necessitate that p must grow at most logarithmically in the sample size to ensure consistency). This is the case with conventional multivariate (Nadaraya–Watson or local polynomial) kernel regression in which the bandwidth is the same for all directions, or k-nearest neighbors with Euclidean distance.

3. Fast convergence rates. When the model is well specified and the response values are bounded (i.e., $\gamma = \infty$), as Corollary 2.5 illustrates, the oracle inequality in (9) yields relatively slow rates of convergence. Because shallow oblique trees often compete empirically with wide neural networks [6, 8, 9], a proper mathematical theory should reflect such qualities. It is therefore natural to compare these rates with the significantly better $r_n = \sqrt{(p/n)\log(n)}$ rates for similar function libraries, achieved by neural networks [2]. In both cases, the prediction risk converges to zero if $p = o(n/\log(n))$ (or equivalently, if $r_n = o(1)$), but the speed differs from logarithmic to polynomial. It is unclear whether the logarithmic rate for oblique CART is optimal in general. We can, however, obtain comparable rates to neural networks by granting two assumptions. Importantly, these assumptions only need to hold on average (with respect to the joint distribution of the data and the search sets) and *not* almost surely for all realizations of the trees. Because most papers that study the convergence rates of neural network estimators proceed without regard for computational complexity, to ensure a fair comparison, we will likewise assume here that $d = p$, $\kappa = 1$ and $P_{\mathcal{A}_t}(\kappa) = 1$ (i.e., direct optimization of (1)).

Our first additional assumption puts a global ℓ_q constraint on the local $\mathcal{L}_1$ total variations of the regression function μ across all terminal nodes of T_K. This is a type of regularity condition on both the tree partition of $\mathcal{X}$ and the regression function μ. It ensures a degree of compatibility between the nonadditive tree model and the additive form of the regression function. In particular, if there existed an (oblique) tessellation of the input space such that the target function is piecewise constant, then the following assumption would hold trivially (i.e., the approximation model is correctly specified). The assumption more generally disciplines the degree of misspecification in globally approximating the unknown target conditional expectation function when employing adaptive oblique tree methods.

ASSUMPTION 3 (Aggregated ℓ_q variation). The regression function μ belongs to $\mathcal{F}$ and there exist positive numbers V and $q > 2$ such that, for any $K \geq 1$,

$$\mathbb{E}\Big[\sum_{t\in T_K} \|\mu\|^q_{\mathcal{L}_1(t)}\Big] \leq V^q. \tag{11}$$

For fixed K and finite $\|\mu\|_{\mathcal{L}_1}$, there is always some choice of V and q for which (11) is satisfied since

$$\limsup_{q\to\infty}\Big(\mathbb{E}\Big[\sum_{t\in T_K} \|\mu\|^q_{\mathcal{L}_1(t)}\Big]\Big)^{1/q} \leq \mathbb{E}\Big[\max_{t\in T_K} \|\mu\|_{\mathcal{L}_1(t)}\Big] \leq \|\mu\|_{\mathcal{L}_1},$$

and hence, for example, $\mathbb{E}[\sum_{t\in T_K} \|\mu\|^q_{\mathcal{L}_1(t)}] \leq (2\|\mu\|_{\mathcal{L}_1})^q$ for q large enough, but finite. However, this alone is not enough to validate Assumption 3 because q may depend on the sample size through its dependence on the depth $K = K_n$. Hence, it is important that (11) hold for the same q *uniformly* over all depths.

It turns out that Assumption 3 can be verified to hold for $V = \|\mu\|_{\mathcal{L}_1}$ and all $q > 2$ when $p = 1$. To see this, recall that $I(\mathbf{a}, t) = [\min_{\mathbf{x}\in t} \mathbf{a}^T\mathbf{x}, \max_{\mathbf{x}\in t} \mathbf{a}^T\mathbf{x}]$. Because the collection of terminal nodes $\{t : t \in T_K\}$ forms a partition of $\mathcal{X}$, when $p = 1$, so does $\{I(\mathbf{a}, t) : t \in T_K\}$ for $I(\mathbf{a}, \mathcal{X}) = [\min_{\mathbf{x}\in\mathcal{X}} \mathbf{a}^T\mathbf{x}, \max_{\mathbf{x}\in\mathcal{X}} \mathbf{a}^T\mathbf{x}]$. Thus, the $\mathcal{L}_1$ total variation is additive over the nodes, that is, $\sum_{t\in T_K} \|\mu\|_{\mathcal{L}_1(t)} = \|\mu\|_{\mathcal{L}_1}$, in which case

$$\sum_{t\in T_K} \|\mu\|^q_{\mathcal{L}_1(t)} \leq \|\mu\|^q_{\mathcal{L}_1}, \quad q \geq 1.$$

In general, for $p > 1$, a crude and not very useful bound is $\sum_{t\in T_K} \|\mu\|^q_{\mathcal{L}_1(t)} \leq 2^K \|\mu\|^q_{\mathcal{L}_1}$; however, the average size of $\sum_{t\in T_K} \|\mu\|^q_{\mathcal{L}_1(t)}$ will often be smaller because it depends on the

specific geometry of the tree partition of $\mathcal{X}$, which captures heterogeneity in the regression function μ. More specifically, the size will depend on how the intervals $I(\mathbf{a}, \mathrm{t})$ overlap across $\mathrm{t} \in T_K$ as well as how much μ varies within each terminal node. We do not expect q to exceed the dimension p, provided that μ is smooth. This is because, by smoothness, $\|\mu\|_{\mathcal{L}_1(\mathrm{t})}$, a proxy for the oscillation of μ in the node is also a proxy for the diameter of the node. Then, because the nodes are disjoint convex polytopes, on average, we expect $\|\mu\|^p_{\mathcal{L}_1(\mathrm{t})}$ to be a proxy for their volume (i.e., their p-dimensional Lebesgue measure), in which case $\mathbb{E}[\sum_{\mathrm{t}\in T_K} \|\mu\|^p_{\mathcal{L}_1(\mathrm{t})}]$ is a constant multiple of the volume of $\mathcal{X}$.

Our final additional assumption puts a moment bound on the maximum number of observations that any one node can contain. Essentially, it says that the $\mathscr{L}_\nu$ norm of $\max_{\mathrm{t}\in T_K} n(\mathrm{t})$ is bounded by a multiple of the average number of observations per node.

ASSUMPTION 4 (Node size moment bound). Let $q > 2$ be the positive number from Assumption 3. There exist positive numbers A and $\nu \geq 1 + 2/(q-2)$ such that, for any $K \geq 1$,

$$\Big(\mathbb{E}\Big[\Big(\max_{\mathrm{t}\in T_K} n(\mathrm{t})\Big)^\nu\Big]\Big)^{1/\nu} \leq \frac{An}{2^K}.$$

Our risk bounds below show that $A = A_n$ is permitted to grow polylogarithmically with the sample size, without affecting the rate of convergence. Because

$$\mathbb{E}\Big[\max_{\mathrm{t}\in T_K} n(\mathrm{t})\Big] \leq \Big(\mathbb{E}\Big[\Big(\max_{\mathrm{t}\in T_K} n(\mathrm{t})\Big)^\nu\Big]\Big)^{1/\nu},$$

and there are at most 2^K disjoint regions t in the partition of $\mathcal{X}$ induced by the tree at depth K such that $\sum_{\mathrm{t}\in T_K} n(\mathrm{t}) = n$, Assumption 4 implies that, on average, no region contains disproportionately more observations than the average number of observations per region, that is, $n/2^K$. Importantly, it still allows for situations where some regions contain very few observations, which does tend to happen in practice. For example, if $n = 1000$, $K = 2$ and T_2 has four terminal nodes with $n(\mathrm{t}) \in \{5, 5, 495, 495\}$, then $\max_{\mathrm{t}\in T_K} n(\mathrm{t}) \leq An/2^K$ holds with $A = 2$.

Previous work [6, 8] has shown that feed-forward neural networks with Heaviside activations can be transformed into oblique decision trees with the same training error. While these tree representations of neural networks require significant depth (the depth of the tree in their construction is at least the width of the target network), they nonetheless demonstrate a proof-of-concept that supports their extensive empirical investigations showing that the modeling power of oblique decision trees is similar to neural networks, even if the trees have modest depth ($K \leq 8$). Our work not only complements these past studies, it also addresses some of the scalability issues associated with global optimization by theoretically validating greedy implementations.

LEMMA 3.1. *Let $d = p$, $\kappa = 1$ and $P_{\mathcal{A}_\mathrm{t}}(\kappa) = 1$, and let Assumptions 3 and 4 hold, and assume $\mathbb{E}[y^2 \log(1 + |y|)] < \infty$. Then, for any $K \geq 1$,*

$$\mathbb{E}[\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2] \leq \mathbb{E}[\|\mathbf{y} - \boldsymbol{\mu}\|_n^2] + \frac{AV^2}{4^{(K-1)/q}}. \tag{12}$$

THEOREM 3.2. *Let $d = p$, $\kappa = 1$ and $P_{\mathcal{A}_\mathrm{t}}(\kappa) = 1$, and let Assumptions 2, 3 and 4 hold. Then, for any $K \geq 1$,*

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_K)\|^2] \leq \frac{2AV^2}{4^{(K-1)/q}} + C\frac{2^{K+1} p \log^{4/\gamma+1}(n)}{n},$$

where $C = C(c_1, c_2, \gamma, M)$ *is a positive constant. Furthermore, if the penalty coefficient satisfies* $\lambda_n \asymp (p/n)\log^{4/\gamma+1}(n)$, *then*

$$\mathbb{E}\big[\|\mu - \widehat{\mu}(T_{\text{opt}})\|^2\big] \le 2(2+q)\left(\frac{AV^2}{q}\right)^{q/(2+q)}\left(\frac{Cp\log^{4/\gamma+1}(n)}{n}\right)^{2/(2+q)}. \tag{13}$$

As mentioned earlier, we see from (13) that $A = A_n$ (as well as $V = V_n$) is allowed to grow polylogarithmically without affecting the convergence rate. When the response values are bounded (i.e., $\gamma = \infty$), the pruned tree estimator $\widehat{\mu}(T_{\text{opt}})$ achieves the rate $r_n^{2/(2+q)} = ((p/n)\log(n))^{2/(2+q)}$, which when $q \approx 2$, is nearly identical to the $\sqrt{r_n}$ rate in Barron [2] for neural network estimators of regression functions $\mu \in \mathcal{F}$ with $\|\mu\|_{\mathcal{L}_1} < \infty$. While we make two additional assumptions (Assumptions 3 and 4) in order for oblique CART to achieve full modeling power on par with neural networks, our theory suggests that decision trees might be preferred in applications where interpretability is valued, without suffering a major loss in predictive accuracy. We also see from these risk bounds that q plays the role of an effective dimension, since it—and not the ambient dimension p—governs the convergence rates. As we have argued above, if μ is smooth, then q should be at most p, and so the convergence rate in (13) should always be at least as fast as the minimax optimal rate $(1/n)^{2/(2+p)}$ for smooth functions in p dimensions.

4. Oblique random forests. A random forest is a randomized ensemble of trees. While traditional random forests use axis-aligned trees, it is also possible to work with oblique trees.

The randomization mechanism in a random forest affects the way each tree is constructed, and consists of two parts. The first part generates a subsample without replacement of size $N < n$ from the original training data, on which the tree is trained, and the second part generates a random collection of candidate splitting directions at each node, from which the optimal one is chosen (see the discussion under the *purely random* heading in Section 2 for generating $\mathcal{A}_t$).

Let Θ denote the random variable whose law governs the aforementioned randomization mechanism and let $T_K(\Theta)$ be the associated maximal tree of depth K. Let $\boldsymbol{\Theta} = (\Theta_1, \ldots, \Theta_B)^{\mathrm{T}}$ denote B independent copies of Θ, corresponding to B trees $T_K(\Theta_b)$, for $b = 1, \ldots, B$. The output of the random forest at a point $\mathbf{x}$ is obtained by averaging the predictions of all B trees in the forest, namely

$$\widehat{\mu}(\boldsymbol{\Theta})(\mathbf{x}) = \frac{1}{B}\sum_{b=1}^{B}\widehat{\mu}\big(T_K(\Theta_b)\big)(\mathbf{x}).$$

By convexity of squared error loss, the expected risk can be bounded as follows:

$$\mathbb{E}\big[\|\mu - \widehat{\mu}(\boldsymbol{\Theta})\|^2\big] \le \frac{1}{B}\sum_{b=1}^{B}\mathbb{E}\big[\|\mu - \widehat{\mu}\big(T_K(\Theta_b)\big)\|^2\big] = \mathbb{E}\big[\|\mu - \widehat{\mu}\big(T_K(\Theta)\big)\|^2\big].$$

The above bound, although crude, tells us that we should expect the random forest to perform no worse than a single (random) tree.

4.1. *Oracle inequality for oblique forests.* We can now establish an oracle inequality for oblique forests similar to that of Theorem 2.3. Conditional on the randomness due to the indices $\mathcal{I} \subset \{1, \ldots, n\}$ of the original training data that belong to the subsampled training data, $\widehat{\mu}(T_K(\Theta_b))$ is a depth K oblique tree (with randomized splits) trained on N samples for each draw $b = 1, \ldots, B$. This means that $\mathbb{E}[\|\mu - \widehat{\mu}(T_K(\Theta_b))\|^2 \mid \mathcal{I}]$ enjoys the *exact* same bounds in Theorem 2.3 but with n replaced by the effective sample size $|\mathcal{I}| = N$. We formalize this notion in Theorem 4.1.

THEOREM 4.1 (Oracle inequality for oblique forests). *Suppose Assumptions 2 holds. Let* $\widehat{\mu}(\boldsymbol{\Theta})$ *be the output of the oblique random forest constructed with oblique trees of depth* K. *Then*

$$\mathbb{E}[\|\mu - \widehat{\mu}(\boldsymbol{\Theta})\|^2] \leq 2 \inf_{f \in \mathcal{F}} \left\{ \|\mu - f\|^2 + \frac{\|f\|_{\mathcal{L}_1}^2 \mathbb{E}[\max_{\mathrm{t} \in [T_K]} P_{\mathcal{A}_\mathrm{t}}^{-1}(\kappa)]}{\kappa K} + C \frac{2^K d \log(Np/d) \log^{4/\gamma}(N)}{N} \right\},$$

where C *is some positive constant and* N *is the subsample size.*

While the efficacy of forests is not reflected in these risk bounds, they do show that forests of oblique trees inherit the same desirable properties as single trees. It should be noted that the expectation in the second term of the bound in Theorem 4.1 is over the subsampled data (instead of over the entire data set as in Theorem 2.3). As such, for consistency results similar to those in Corollaries 2.4 and 2.5, the splitting probabilities would need to be almost surely bounded away from zero (Assumption 1) for any realization of the subsampled data. Additionally, with the stronger assumptions analogous results to Theorem 3.2 can also be derived for oblique forests. We omit details to conserve space.

5. Conclusion and future work. We explored how oblique decision trees—which output constant averages over polytopal partitions of the feature space—can be used for predictive modeling with ridge expansions, sometimes achieving the same convergence rates as neural networks. The theory presented here is encouraging as it implies that interpretable models can exhibit provably good performance similar to their black-box counterparts such as neural networks. The computational bottleneck still remains the main obstacle for practical implementation. Crucially, however, our risk bounds show that favorable performance can occur even if the optimization is only done approximately. We conclude with a discussion of some directions for potential future research.

5.1. *Multilayer networks.* We can go beyond approximating single-hidden layer neural networks if instead the split boundaries of the oblique trees have the form $\mathbf{a}^\mathrm{T}\boldsymbol{\Phi}(\mathbf{x}) = b$, where $\boldsymbol{\Phi}$ is a multidimensional feature map, such as the output layer of a neural network. For example, if $\boldsymbol{\Phi}_k(\mathbf{x}) = \phi(\mathbf{a}_k^\mathrm{T}\mathbf{x} - b_k)$, where ϕ is some activation function, then this additional flexibility allows us to approximate two-hidden layer networks, that is, functions of the form $\sum_{k_2} c_{k_2} \phi(\sum_{k_1} c_{k_1,k_2} \phi(\mathbf{a}_{k_1,k_2}^\mathrm{T}\mathbf{x} - b_{k_1,k_2}))$.

5.2. *Classification.* While we have focused on regression trees, oblique decision trees are commonly applied to the problem of binary classification, that is, $y_i \in \{-1, 1\}$. In this case, because Gini impurity [11, 23] is equivalent to the squared error criterion (1), our results also directly apply to the classification setting provided the conditional class probability $\eta(\mathbf{x}) = \mathbb{P}(y = 1 \mid \mathbf{x})$ belongs to $\mathcal{F}$ and has finite $\|\eta\|_{\mathcal{L}_1}$. A more natural assumption when modeling *probabilities*, however, would be to have the log-odds $f(\mathbf{x}) = \log(\eta(\mathbf{x})/(1 - \eta(\mathbf{x})))$ belong to $\mathcal{F}$ and have finite $\|f\|_{\mathcal{L}_1}$. In this case, we must use another widely used splitting criterion, the *information gain*, namely the amount by which the binary entropy of the class probabilities in the node can be reduced from splitting the parent node [23, 38]:

$$\mathrm{IG}(b, \mathbf{a}, \mathrm{t}) = H(\mathrm{t}) - \frac{n(\mathrm{t}_L)}{n(\mathrm{t})} H(\mathrm{t}_L) - \frac{n(\mathrm{t}_R)}{n(\mathrm{t})} H(\mathrm{t}_R),$$

where $H(\mathrm{t}) = \eta(\mathrm{t}) \log(1/\eta(\mathrm{t})) + (1 - \eta(\mathrm{t})) \log(1/(1 - \eta(\mathrm{t})))$ and $\eta(\mathrm{t}) = \frac{1}{n(\mathrm{t})} \sum_{\mathbf{x}_i \in \mathrm{t}} \mathbb{1}(y_i = 1)$. Interestingly, maximizing the information gain in the node is equivalent to minimizing the

node-wise logistic loss with respect to the family of log-odds models of the form $\theta_{\mathrm{t}}(\mathbf{x}) = \beta_1 \mathbb{1}(\mathbf{a}^{\mathrm{T}}\mathbf{x} \le b) + \beta_2 \mathbb{1}(\mathbf{a}^{\mathrm{T}}\mathbf{x} > b)$; that is,

$$(\hat{b}, \hat{\mathbf{a}}) \in \operatorname*{argmax}_{(b,\mathbf{a})} \mathrm{IG}(b, \mathbf{a}, \mathrm{t}) \quad \Longleftrightarrow \quad (\hat{\beta}_1, \hat{\beta}_2, \hat{b}, \hat{\mathbf{a}}) \in \operatorname*{argmin}_{(\beta_1,\beta_2,b,\mathbf{a})} \sum_{\mathbf{x}_i \in \mathrm{t}} \log\bigl(1 + \exp\bigl(-y_i \theta_{\mathrm{t}}(\mathbf{x}_i)\bigr)\bigr).$$

One can use techniques from Klusowski and Tian [28], which exploits connections to sequential greedy algorithms for other convex optimization problems [48] (e.g., LogitBoost), to establish a training error bound (with respect to logistic loss) akin to Lemma 2.2.

6. Proofs. Our discussion so far has focused on oblique trees that output a constant (sample average) at each node. Fortunately, most of our results hold in a much more general setting. In particular, we can allow for the nodes to output $\hat{y}_{\mathrm{t}} \in \operatorname{argmin}_{h \in \operatorname{span}(\mathcal{H})} \sum_{\mathbf{x}_i \in \mathrm{t}} (y_i - h(\mathbf{x}_i))^2$, where $\mathcal{H}$ is a finite-dimensional dictionary that contains the constant function. The proofs here deal with the general case.

In what follows, we assume without loss of generality that the infimum in the definition of $\|f\|_{\mathcal{L}_1}$ for $f \in \mathcal{F}$ is achieved at some element $g \in \mathcal{G}$, since otherwise there exists $g \in \mathcal{G}$ with $\|f - g\|$ arbitrarily small and $\|g\|_{\mathcal{L}_1}$ arbitrarily close to $\|f\|_{\mathcal{L}_1}$. We denote the supremum norm of a function $f : \mathcal{X} \mapsto \mathbb{R}$ by $\|f\|_\infty = \sup_{\mathbf{x} \in \mathcal{X}} |f(\mathbf{x})|$. Additionally, we slightly abuse notation by taking $\mathbf{y} - \hat{y}_{\mathrm{t}}$ to mean $\mathbf{y} - \hat{y}_{\mathrm{t}}\mathbf{1}$, where $\mathbf{1} = (1, \dots, 1)^{\mathrm{T}}$ is the $n \times 1$ vector of ones.

PROOF OF LEMMA 2.1. Set $\mathcal{U}_{\mathrm{t}} = \{u(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) + v(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_R) : u, v \in \operatorname{span}(\mathcal{H})\}$ and consider the closed subspace $\mathcal{V}_{\mathrm{t}} = \{v(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}) : v \in \operatorname{span}(\mathcal{H})\}$. By the orthogonal decomposition property of Hilbert spaces, we can express $\mathcal{U}_{\mathrm{t}}$ as the direct sum $\mathcal{V}_{\mathrm{t}} \oplus \mathcal{V}_{\mathrm{t}}^{\perp}$, where $\mathcal{V}_{\mathrm{t}}^{\perp} = \{u \in \mathcal{U}_{\mathrm{t}} : \langle u, v \rangle_n = 0, \text{ for all } v \in \mathcal{V}_{\mathrm{t}}\}$. Let Ψ_{t} be any orthonormal basis for $\mathcal{V}_{\mathrm{t}}$ that includes $w^{-1/2}(\mathrm{t})\mathbb{1}(\mathbf{x} \in \mathrm{t})$, where we remind the reader that $w(\mathrm{t}) = n(\mathrm{t})/n$. Let $\Psi_{\mathrm{t}}^{\perp}$ be any orthonormal basis for $\mathcal{V}_{\mathrm{t}}^{\perp}$ that includes the decision stump (3). We will show that

$$\widehat{\mu}(T)(\mathbf{x}) = \sum_{\mathrm{t} \in [T]} \sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} \langle \mathbf{y}, \boldsymbol{\psi} \rangle_n \psi(\mathbf{x}), \tag{14}$$

where $\{\psi \in \Psi_{\mathrm{t}}^{\perp} : \mathrm{t} \in [T]\}$ is an orthonormal dictionary and, furthermore, that

$$\sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2 = \widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, \mathrm{t}). \tag{15}$$

These identities are the respective generalizations of (4) and (5). Because $\hat{y}_{\mathrm{t}}(\mathbf{x})$ is the projection of $\mathbf{y}$ onto $\mathcal{V}_{\mathrm{t}}$, it follows that $\hat{y}_{\mathrm{t}}(\mathbf{x}) = \sum_{\psi \in \Psi_{\mathrm{t}}} \langle \mathbf{y}, \boldsymbol{\psi} \rangle_n \psi(\mathbf{x})$. For similar reasons, $\hat{y}_{\mathrm{t}_L}(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) + \hat{y}_{\mathrm{t}_R}(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_R) = \sum_{\psi \in \Psi_{\mathrm{t}} \cup \Psi_{\mathrm{t}}^{\perp}} \langle \mathbf{y}, \boldsymbol{\psi} \rangle_n \psi(\mathbf{x})$.

To prove the identity in (14) (and, as a special case, (4)), using the above expansions, observe that for each internal node t,

$$\sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} \langle \mathbf{y}, \boldsymbol{\psi} \rangle_n \psi(\mathbf{x}) = \bigl(\hat{y}_{\mathrm{t}_L}(\mathbf{x}) - \hat{y}_{\mathrm{t}}(\mathbf{x})\bigr)\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) + \bigl(\hat{y}_{\mathrm{t}_R}(\mathbf{x}) - \hat{y}_{\mathrm{t}}(\mathbf{x})\bigr)\mathbb{1}(\mathbf{x} \in \mathrm{t}_R). \tag{16}$$

For each $\mathbf{x} \in \mathcal{X}$, let $\mathrm{t}_0, \mathrm{t}_1, \dots, \mathrm{t}_{K-1}, \mathrm{t}_K = \mathrm{t}$ be the unique path from the root node t_0 to the terminal node t that contains $\mathbf{x}$. Next, sum (16) over all internal nodes and telescope the successive internal node outputs to obtain

$$\sum_{k=0}^{K-1} \bigl(\hat{y}_{\mathrm{t}_{k+1}}(\mathbf{x}) - \hat{y}_{\mathrm{t}_k}(\mathbf{x})\bigr) = \hat{y}_{\mathrm{t}_K}(\mathbf{x}) - \hat{y}_{\mathrm{t}_0}(\mathbf{x}) = \hat{y}_{\mathrm{t}}(\mathbf{x}) - \hat{y}(\mathbf{x}), \tag{17}$$

where $\hat{y} \in \operatorname{argmin}_{h \in \mathcal{H}} \sum_{i=1}^n (y_i - h(\mathbf{x}_i))^2$. Combining (16) and (17), we have

$$\sum_{\mathrm{t} \in T} \hat{y}_{\mathrm{t}}(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}) = \hat{y}(\mathbf{x}) + \sum_{\mathrm{t} \in [T] \setminus \{\mathrm{t}_0\}} \sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} \langle \mathbf{y}, \boldsymbol{\psi} \rangle_n \psi(\mathbf{x}) = \sum_{\mathrm{t} \in [T]} \sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} \langle \mathbf{y}, \boldsymbol{\psi} \rangle_n \psi(\mathbf{x}),$$

where we recall that the null node t_0 is an internal node of T. Next, we show that $\{\psi \in \Psi_{\mathrm{t}}^{\perp} : \mathrm{t} \in [T]\}$ is orthonormal. The fact that each ψ has unit norm, $\|\boldsymbol{\psi}\|_n^2 = 1$, is true by definition. If $\psi, \psi' \in \Psi_{\mathrm{t}}^{\perp}$, then by definition, $\langle \boldsymbol{\psi}, \boldsymbol{\psi}' \rangle_n = 0$. Let t and t′ be two distinct internal nodes and suppose $\psi \in \Psi_{\mathrm{t}}^{\perp}$ and $\psi' \in \Psi_{\mathrm{t}'}^{\perp}$. If $\mathrm{t} \cap \mathrm{t}' = \varnothing$, then orthogonality between ψ and ψ' is immediate, since $\psi(\mathbf{x}) \cdot \psi'(\mathbf{x}) \equiv 0$. If $\mathrm{t} \cap \mathrm{t}' \neq \varnothing$, then due to the nested property of the nodes, either $\mathrm{t} \subseteq \mathrm{t}'$ or $\mathrm{t}' \subseteq \mathrm{t}$. Assume without loss of generality that $\mathrm{t} \subseteq \mathrm{t}'$. Then ψ', when restricted to $\mathbf{x} \in \mathrm{t}$, belongs to $\mathcal{V}_{\mathrm{t}}$, which also implies that ψ and ψ' are orthogonal, since $\psi \in \mathcal{V}_{\mathrm{t}}^{\perp}$.

Finally, the decrease in impurity identity (15) (and, as a special case, (5)) can be shown as follows:

$$\begin{aligned}
\widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, \mathrm{t}) &= \frac{1}{n} \sum_{\mathbf{x}_i \in \mathrm{t}} \big(y_i - \hat{y}_{\mathrm{t}}(\mathbf{x}_i)\big)^2 - \frac{1}{n} \sum_{\mathbf{x}_i \in \mathrm{t}} \big(y_i - \hat{y}_{\mathrm{t}_L}(\mathbf{x}_i) \mathbb{1}(\mathbf{x}_i \in \mathrm{t}_L) - \hat{y}_{\mathrm{t}_R}(\mathbf{x}_i) \mathbb{1}(\mathbf{x}_i \in \mathrm{t}_R)\big)^2 \\
&= \left(\frac{1}{n} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i^2 - \sum_{\psi \in \Psi_{\mathrm{t}}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2 \right) - \left(\frac{1}{n} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i^2 - \sum_{\psi \in \Psi_{\mathrm{t}} \cup \Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2 \right) \\
&= \sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2.
\end{aligned}$$

□

Throughout the remaining proofs, we will assume that there exists a positive constant $Q \geq 1$ such that $\sup_{\mathbf{x} \in \mathcal{X}} |\widehat{\mu}(T)(\mathbf{x})| \leq Q \cdot \sqrt{\max_{1 \leq i \leq n} \frac{1}{i} \sum_{\ell=1}^{i} y_\ell^2}$, almost surely. This assumption is drawn from the bound

$$|\hat{y}_{\mathrm{t}}(\mathbf{x})| \leq \sqrt{\max_{1 \leq i \leq n} \frac{1}{i} \sum_{1 \leq \ell \leq i} y_\ell^2} \sqrt{w(\mathrm{t}) \sum_{\psi \in \Psi_{\mathrm{t}}} \psi^2(\mathbf{x})},$$

which is established by first using the basis expansion for $\hat{y}_{\mathrm{t}}$ provided in the proof of Lemma 2.1 and the Cauchy–Schwarz inequality,

$$|\hat{y}_{\mathrm{t}}(\mathbf{x})| = \left| \sum_{\psi \in \Psi_{\mathrm{t}}} \langle \mathbf{y}, \psi \rangle_n \psi(\mathbf{x}) \right| \leq \sqrt{\sum_{\psi \in \Psi_{\mathrm{t}}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2} \sqrt{\sum_{\psi \in \Psi_{\mathrm{t}}} \psi^2(\mathbf{x})},$$

and then, because $\{\psi : \psi \in \Psi_{\mathrm{t}}\}$ is orthonormal, employing Bessel's inequality to obtain $\sum_{\psi \in \Psi_{\mathrm{t}}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2 \leq n^{-1} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i^2 \leq w(\mathrm{t}) \max_{1 \leq i \leq n} \frac{1}{i} \sum_{\ell=1}^{i} y_\ell^2$. Thus, Q could be taken to equal (or be an almost sure bound on) $\sup_{\mathbf{x} \in \mathcal{X}} \max_{\mathrm{t} \in [T]} \sqrt{w(\mathrm{t}) \sum_{\psi \in \Psi_{\mathrm{t}}} \psi^2(\mathbf{x})}$. In the conventional case where the tree outputs a constant in each node, $\Psi_{\mathrm{t}} = \{w^{-1/2}(\mathrm{t}) \mathbb{1}(\mathbf{x} \in \mathrm{t})\}$, and hence $Q = 1$. To ensure that $\widehat{\mu}(T)(\mathbf{x})$ is square-integrable, that is, $\mathbb{E}[\sup_{\mathbf{x} \in \mathcal{X}} |\widehat{\mu}(T)(\mathbf{x})|^2] < \infty$, we merely need to check that $\mathbb{E}[\max_{1 \leq i \leq n} \frac{1}{i} \sum_{\ell=1}^{i} y_\ell^2] < \infty$. This follows easily from Doob's maximal inequality for positive submartingales [19, Theorem 5.4.4], since $\mathbb{E}[y^2 \log(1 + |y|)] < \infty$ by assumption.

PROOF OF LEMMAS 2.2 AND 3.1. Define the excess training error as $R_K = \|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2 - \|\mathbf{y} - \mathbf{g}\|_n^2$. Define the squared nodewise norm and nodewise inner product as $\|\mathbf{f}\|_{\mathrm{t}}^2 = \frac{1}{n(\mathrm{t})} \sum_{\mathbf{x}_i \in \mathrm{t}} (f(\mathbf{x}_i))^2$ and $\langle \mathbf{f}, \mathbf{g} \rangle_{\mathrm{t}} = \frac{1}{n(\mathrm{t})} \sum_{\mathbf{x}_i \in \mathrm{t}} f(\mathbf{x}_i) g(\mathbf{x}_i)$, respectively. We define the nodewide excess training error as $R_K(\mathrm{t}) = \|\mathbf{y} - \hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|\mathbf{y} - \mathbf{g}\|_{\mathrm{t}}^2$. We use this to rewrite the total excess training error as a weighted combination of the nodewide excess train errors: $R_K = \sum_{\mathrm{t} \in T_K} w(\mathrm{t}) R_K(\mathrm{t})$, where $w(\mathrm{t}) = n(\mathrm{t})/n$, and $\mathrm{t} \in T_K$ means t is a terminal node of T_K. From the orthogonal decomposition of the tree, as given in (14), we have

$$\|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_K)\|_n^2 = \|\mathbf{y} - \widehat{\boldsymbol{\mu}}(T_{K-1})\|_n^2 - \sum_{\mathrm{t} \in T_{K-1}} \sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2. \tag{18}$$

Subtracting $\|\mathbf{y}-\mathbf{g}\|_n^2$ on both sides of (18), and using the definition of R_K, we obtain

$$R_K = R_{K-1} - \sum_{\mathrm{t}\in T_{K-1}} \sum_{\psi\in\Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi}\rangle_n|^2. \tag{19}$$

Henceforth, we adopt the notation $\mathbb{E}_{T_K}[R_K]$ to mean that the expectation is taken with respect to the joint distribution of $\{\mathcal{A}_{\mathrm{t}} : \mathrm{t}\in[T_K]\}$, conditional on the data. We can assume $\mathbb{E}[R_K]>0$ for all $K\geq 1$, since otherwise, by definition of R_K, $\mathbb{E}[R_K] = \mathbb{E}[\|\mathbf{y}-\widehat{\boldsymbol{\mu}}(T_K)\|_n^2 - \|\mathbf{y}-\mathbf{g}\|_n^2] \leq 0$, which directly gives the desired result. Using the law of iterated expectations and the recursive relationship obtained in (19),

$$\begin{aligned}\mathbb{E}_{T_K}[R_K] &= \mathbb{E}_{T_{K-1}}\big[\mathbb{E}_{T_K|T_{K-1}}[R_K]\big] \\ &= \mathbb{E}_{T_{K-1}}[R_{K-1}] - \mathbb{E}_{T_{K-1}}\Bigg[\mathbb{E}_{T_K|T_{K-1}}\Bigg[\sum_{\mathrm{t}\in T_{K-1}}\sum_{\psi\in\Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi}\rangle_n|^2\Bigg]\Bigg].\end{aligned} \tag{20}$$

By (15) and the suboptimality probability, $P_{\mathcal{A}(\mathrm{t})}(\kappa)$, we can rewrite the term inside the iterated expectation in (20) as

$$\begin{aligned}&\sum_{\mathrm{t}\in T_{K-1}}\sum_{\psi\in\Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi}\rangle_n|^2 \\ &\quad= \sum_{\mathrm{t}\in T_{K-1}} \widehat{\Delta}(\hat{b},\hat{\mathbf{a}},\mathrm{t}) \geq \sum_{\mathrm{t}\in T_{K-1}} \mathbb{1}\Big(\widehat{\Delta}(\hat{b},\hat{\mathbf{a}},\mathrm{t}) \geq \kappa \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t})\Big)\widehat{\Delta}(\hat{b},\hat{\mathbf{a}},\mathrm{t}) \\ &\quad\geq \kappa \sum_{\mathrm{t}\in T_{K-1}} \mathbb{1}\Big(\widehat{\Delta}(\hat{b},\hat{\mathbf{a}},\mathrm{t}) \geq \kappa \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t})\Big) \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t}).\end{aligned} \tag{21}$$

Taking expectations of both sides of (21) with respect to the conditional distribution of T_K given T_{K-1}, we have

$$\begin{aligned}&\mathbb{E}_{T_K|T_{K-1}}\Bigg[\sum_{\mathrm{t}\in T_{K-1}}\sum_{\psi\in\Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi}\rangle_n|^2\Bigg] \\ &\quad\geq \kappa \sum_{\mathrm{t}\in T_{K-1}} \mathbb{E}_{T_K|T_{K-1}}\Big[\mathbb{1}\Big(\widehat{\Delta}(\hat{b},\hat{\mathbf{a}},\mathrm{t}) \geq \kappa \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t})\Big) \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t})\Big].\end{aligned} \tag{22}$$

By definition of $P_{\mathcal{A}(\mathrm{t})}$,

$$\begin{aligned}&\sum_{\mathrm{t}\in T_{K-1}} \mathbb{E}_{T_K|T_{K-1}}\Big[\mathbb{1}\Big(\widehat{\Delta}(\hat{b},\hat{\mathbf{a}},\mathrm{t}) \geq \kappa \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t})\Big) \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t})\Big] \\ &\quad= \sum_{\mathrm{t}\in T_{K-1}} P_{\mathcal{A}_{\mathrm{t}}}(\kappa) \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t}) \geq \sum_{\mathrm{t}\in T_{K-1}: R_{K-1}(\mathrm{t})>0} P_{\mathcal{A}_{\mathrm{t}}}(\kappa) \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t}).\end{aligned} \tag{23}$$

In turn, by Lemma 6.1,

$$\sum_{\mathrm{t}\in T_{K-1}: R_{K-1}(\mathrm{t})>0} P_{\mathcal{A}_{\mathrm{t}}}(\kappa) \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t}) \geq \sum_{\mathrm{t}\in T_{K-1}: R_{K-1}(\mathrm{t})>0} w(\mathrm{t}) \frac{R_{K-1}^2(\mathrm{t})}{P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(\mathrm{t})}^2}, \tag{24}$$

and by Lemma A.1 (see the Supplemental Material [14] for statement and proof),

$$\begin{aligned}\sum_{\mathrm{t}\in T_{K-1}: R_{K-1}(\mathrm{t})>0} w(\mathrm{t}) \frac{R_{K-1}^2(\mathrm{t})}{P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(\mathrm{t})}^2} &\geq \frac{(\sum_{\mathrm{t}\in T_{K-1}: R_{K-1}(\mathrm{t})>0} w(\mathrm{t}) R_{K-1}(\mathrm{t}))^2}{\sum_{\mathrm{t}\in T_{K-1}: R_{K-1}(\mathrm{t})>0} w(\mathrm{t}) P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(\mathrm{t})}^2} \\ &\geq \frac{(R_{K-1}^+)^2}{\sum_{\mathrm{t}\in T_{K-1}} w(\mathrm{t}) P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(\mathrm{t})}^2},\end{aligned} \tag{25}$$

where $R_{K-1}^+ = \sum_{t \in T_{K-1}: R_{K-1}(t) > 0} w(t) R_{K-1}(t) \geq R_{K-1}$. Combining (22), (23), (24) and (25) and plugging the result into (20), we obtain

$$\mathbb{E}_{T_K}[R_K] \leq \mathbb{E}_{T_{K-1}}[R_{K-1}] - \kappa \mathbb{E}_{T_{K-1}}\left[\frac{(R_{K-1}^+)^2}{\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2}\right].$$

Using [14, Lemma A.1], again, we have

$$\mathbb{E}_{T_{K-1}}\left[\frac{(R_{K-1}^+)^2}{\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2}\right] \geq \frac{(\mathbb{E}_{T_{K-1}}[R_{K-1}^+])^2}{\mathbb{E}_{T_{K-1}}[\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2]}.$$

We have therefore derived the recursion

$$\mathbb{E}_{T_K}[R_K] \leq \mathbb{E}_{T_{K-1}}[R_{K-1}] - \kappa \frac{(\mathbb{E}_{T_{K-1}}[R_{K-1}^+])^2}{\mathbb{E}_{T_{K-1}}[\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2]}. \tag{26}$$

Next, let us take the expectation of both sides of (26) with respect to the data, apply [14, Lemma A.1], and use the fact that $R_{K-1}^+ \geq R_{K-1}$ and $\mathbb{E}[R_{K-1}] > 0$ to obtain

$$\begin{aligned}
\mathbb{E}[R_K] &\leq \mathbb{E}[R_{K-1}] - \kappa \mathbb{E}\left[\frac{(\mathbb{E}_{T_{K-1}}[R_{K-1}^+])^2}{\mathbb{E}_{T_{K-1}}[\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2]}\right] \\
&\leq \mathbb{E}[R_{K-1}] - \kappa \frac{(\mathbb{E}[R_{K-1}^+])^2}{\mathbb{E}[\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2]} \\
&\leq \mathbb{E}[R_{K-1}] - \kappa \frac{(\mathbb{E}[R_{K-1}])^2}{\mathbb{E}[\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2]}.
\end{aligned}$$

We have therefore obtained a recursion for $\mathbb{E}[R_K]$, which we can now solve thanks to Lemma 6.2. Setting $a_k = \mathbb{E}[R_k]$ and $b_k = \kappa / \mathbb{E}[\sum_{t \in T_{k-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2]$ in Lemma 6.2, we have

$$\mathbb{E}[R_K] \leq \frac{1}{\kappa \sum_{k=1}^K 1 / \mathbb{E}[\sum_{t \in T_{k-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2]}. \tag{27}$$

The next part of the proof depends on the assumptions we make about $w(t)$, $P_{\mathcal{A}_t}(\kappa)$ and $\|g\|_{\mathcal{L}_1(t)}^2$ and how they enable us to upper bound

$$\mathbb{E}\left[\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2\right].$$

For Lemma 2.2: In this case, we do not impose any assumptions on $w(t)$. We can use the fact that $\sum_{t \in T_{K-1}} w(t) = 1$ and $\|g\|_{\mathcal{L}_1(t)}^2 \leq \|g\|_{\mathcal{L}_1}^2$ for all $t \in T_{K-1}$ to get

$$\begin{aligned}
&\mathbb{E}\left[\sum_{t \in T_{K-1}} w(t) P_{\mathcal{A}_t}^{-1}(\kappa) \|g\|_{\mathcal{L}_1(t)}^2\right] \\
&\leq \|g\|_{\mathcal{L}_1}^2 \mathbb{E}\left[\max_{t \in T_{K-1}} P_{\mathcal{A}_t}^{-1}(\kappa) \sum_{t \in T_{K-1}} w(t)\right] = \|g\|_{\mathcal{L}_1}^2 \mathbb{E}\left[\max_{t \in T_{K-1}} P_{\mathcal{A}_t}^{-1}(\kappa)\right] \\
&\leq \|g\|_{\mathcal{L}_1}^2 \mathbb{E}\left[\max_{t \in [T_K]} P_{\mathcal{A}_t}^{-1}(\kappa)\right].
\end{aligned}$$

Plugging this bound into (27), we obtain the desired inequality in (7) on the expected excess training error, namely

$$\mathbb{E}[R_K] \leq \frac{\|g\|_{\mathcal{L}_1}^2 \mathbb{E}[\max_{t \in [T_K]} P_{\mathcal{A}_t}^{-1}(\kappa)]}{\kappa K}.$$

For Lemma 3.1: If we grant Assumptions 3 and 4, and take $g = \mu \in \mathcal{G}$, we can arrive at a stronger bound. Recall that we also assume that $\kappa = 1$ and $P_{\mathcal{A}_t}(\kappa) = 1$. Since $q > 2$, by two successive applications of Hölder's inequality, we have

$$\sum_{t \in T_{K-1}} w(t)\|\mu\|^2_{\mathcal{L}_1(t)} \le \Big(\sum_{t \in T_{K-1}} (w(t))^{q/(q-2)}\Big)^{1-2/q} \Big(\sum_{t \in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(t)}\Big)^{2/q}, \tag{28}$$

and

$$\begin{aligned} &\mathbb{E}\Big[\Big(\sum_{t \in T_{K-1}} (w(t))^{q/(q-2)}\Big)^{1-2/q} \Big(\sum_{t \in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(t)}\Big)^{2/q}\Big] \\ &\quad \le \Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} (w(t))^{q/(q-2)}\Big]\Big)^{1-2/q} \Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(t)}\Big]\Big)^{2/q}. \end{aligned} \tag{29}$$

Combining the two inequalities (28) and (29), we obtain

$$\mathbb{E}\Big[\sum_{t \in T_{K-1}} w(t)\|\mu\|^2_{\mathcal{L}_1(t)}\Big] \le \Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} (w(t))^{q/(q-2)}\Big]\Big)^{1-2/q} \Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(t)}\Big]\Big)^{2/q}.$$

Assumptions 3 and 4 provide further upper bounds, since

$$\begin{aligned} &\Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} (w(t))^{q/(q-2)}\Big]\Big)^{1-2/q} \Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(t)}\Big]\Big)^{2/q} \\ &\quad \le \Big(2^{K-1}\mathbb{E}\Big[\Big(\max_{t \in T_{K-1}} w(t)\Big)^{q/(q-2)}\Big]\Big)^{1-2/q} \Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(t)}\Big]\Big)^{2/q} \\ &\quad \le 2^{(K-1)(1-2/q)} \Big(\mathbb{E}\Big[\Big(\max_{t \in T_{K-1}} w(t)\Big)^{\nu}\Big]\Big)^{1/\nu} \Big(\mathbb{E}\Big[\sum_{t \in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(t)}\Big]\Big)^{2/q} \le \frac{AV^2}{4^{(K-1)/q}}. \end{aligned}$$

Plugging this bound into (27), we obtain the desired inequality (12) on the expected excess training error, namely $\mathbb{E}[R_K] \le \frac{AV^2}{4^{(K-1)/q}}$. □

PROOF OF THEOREMS 2.3 AND 3.2. See the Supplemental Material [14]. □

PROOF OF THEOREM 4.1. See the Supplemental Material [14]. □

PROOF OF COROLLARY 2.4. See the Supplemental Material [14]. □

PROOF OF COROLLARY 2.5. See the Supplemental Material [14]. □

6.1. *Additional lemmas*. First, we state and prove a lemma that establishes an important connection between the decrease in impurity and the empirical nodewise excess risk.

LEMMA 6.1 (Impurity bound). *Define* $R_{K-1}(t) = \|\mathbf{y} - \hat{y}_t\|^2_t - \|\mathbf{y} - \mathbf{g}\|^2_t$. *Let* t *be a terminal node of* T_{K-1}, *and assume* $R_{K-1}(t) > 0$. *Then, if* $g \in \mathcal{G}$,

$$\max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t) \ge \frac{w(t) R^2_{K-1}(t)}{\|g\|^2_{\mathcal{L}_1(t)}}.$$

PROOF OF LEMMA 6.1. Assume that $g \in \mathcal{G}$, $g(\mathbf{x}) = \sum_{k=1}^{M} g_k(\mathbf{a}_k^{\mathrm{T}}\mathbf{x})$ and that $g(\mathbf{x}_i)$ is not constant across $\mathbf{x}_i \in \mathrm{t}$, the result being trivial otherwise. We use g_k' to denote the divided difference of g_k of successive ordered datapoints in the $\mathbf{a}_k$ direction in node t. That is, if the data $\{(y_i, \mathbf{x}_i^{\mathrm{T}}) : \mathbf{x}_i \in \mathrm{t}\}$ is reindexed so that $\mathbf{a}_k^{\mathrm{T}}\mathbf{x}_1 \le \mathbf{a}_k^{\mathrm{T}}\mathbf{x}_2 \le \cdots \le \mathbf{a}_k^{\mathrm{T}}\mathbf{x}_{n(\mathrm{t})}$, then

$$g_k'(b) = \frac{g_k(\mathbf{a}_k^{\mathrm{T}}\mathbf{x}_{i+1}) - g_k(\mathbf{a}_k^{\mathrm{T}}\mathbf{x}_i)}{\mathbf{a}_k^{\mathrm{T}}\mathbf{x}_{i+1} - \mathbf{a}_k^{\mathrm{T}}\mathbf{x}_i} \quad \text{for } \mathbf{a}_k^{\mathrm{T}}\mathbf{x}_i \le b < \mathbf{a}_k^{\mathrm{T}}\mathbf{x}_{i+1} \text{ and } i = 1, \ldots, n(\mathrm{t}) - 1, \tag{30}$$

where $g_k'(b) = 0$ if $b = \mathbf{a}_k^{\mathrm{T}}\mathbf{x}_i = \mathbf{a}_k^{\mathrm{T}}\mathbf{x}_{i+1}$. Let

$$\frac{\mathrm{d}\Pi(b, \mathbf{a}_k)}{\mathrm{d}(b, \mathbf{a}_k)} = \frac{|g_k'(b)|\sqrt{\mathbb{P}(\mathrm{t}_L)\mathbb{P}(\mathrm{t}_R)}}{\sum_{k'=1}^{M} \int |g_{k'}'(b')|\sqrt{\mathbb{P}(\mathrm{t}_L')\mathbb{P}(\mathrm{t}_R')}\,\mathrm{d}b'} \tag{31}$$

denote the Radon–Nikodym derivative (with respect to the Lebesgue measure and counting measure) of a probability measure on $(b, \mathbf{a})$ after splitting node t at the decision boundary $\mathbf{a}_k^{\mathrm{T}}\mathbf{x} = b$. Here, $\mathrm{t}_L = \mathrm{t}_L(b, \mathbf{a}_k)$ and $\mathrm{t}_R = \mathrm{t}_R(b, \mathbf{a}_k)$ are the child nodes of t after splitting at $\mathbf{a}_k^{\mathrm{T}}\mathbf{x} = b$, and $\mathbb{P}(\mathrm{t}_L) = n(\mathrm{t}_L)/n(\mathrm{t})$ and $\mathbb{P}(\mathrm{t}_R) = n(\mathrm{t}_R)/n(\mathrm{t})$ are the proportions of observations in node t that is in t_L and t_R, respectively. Similarly, $\mathrm{t}_L' = \mathrm{t}_L'(b', \mathbf{a}_{k'})$ and $\mathrm{t}_R' = \mathrm{t}_R'(b', \mathbf{a}_{k'})$ are the child nodes of t after splitting at $\mathbf{a}_{k'}^{\mathrm{T}}\mathbf{x} = b'$. Additionally, define

$$\tilde{\psi}_{\mathrm{t}}(\mathbf{x}) = \frac{\mathbb{1}(\mathbf{x} \in \mathrm{t}_L)\mathbb{P}(\mathrm{t}_R) - \mathbb{1}(\mathbf{x} \in \mathrm{t}_R)\mathbb{P}(\mathrm{t}_L)}{\sqrt{\mathbb{P}(\mathrm{t}_L)\mathbb{P}(\mathrm{t}_R)}} = \sqrt{w(\mathrm{t})}\psi_{\mathrm{t}}(\mathbf{x}).$$

Note that $\{\tilde{\psi}_{\mathrm{t}} : \mathrm{t} \in [T_K]\}$ is an orthonormal dictionary with respect to the nodewise inner product, $\langle \cdot, \cdot \rangle_{\mathrm{t}}$. Because a maximum is larger than an average, $\max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, \mathrm{t}) \ge \int \widehat{\Delta}(b, \mathbf{a}_k, \mathrm{t})\,\mathrm{d}\Pi(b, \mathbf{a}_k)$. Then, using the identity from (15) and the fact that the decision stump ψ_{t} belongs to $\Psi_{\mathrm{t}}^{\perp}$ (see (3)), we have

$$\max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, \mathrm{t}) \ge \int \sum_{\psi \in \Psi_{\mathrm{t}}^{\perp}} |\langle \mathbf{y}, \boldsymbol{\psi} \rangle_n|^2\,\mathrm{d}\Pi(b, \mathbf{a}_k) \ge \int |\langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n|^2\,\mathrm{d}\Pi(b, \mathbf{a}_k). \tag{32}$$

By the definition of $\tilde{\psi}_{\mathrm{t}}$ and Jensen's inequality,

$$\begin{aligned} \int |\langle \mathbf{y}, \boldsymbol{\psi}_{\mathrm{t}} \rangle_n|^2\,\mathrm{d}\Pi(b, \mathbf{a}_k) &= w(\mathrm{t}) \int |\langle \mathbf{y}, \tilde{\boldsymbol{\psi}}_{\mathrm{t}} \rangle_{\mathrm{t}}|^2\,\mathrm{d}\Pi(b, \mathbf{a}_k) \\ &\ge w(\mathrm{t}) \bigg(\int |\langle \mathbf{y}, \tilde{\boldsymbol{\psi}}_{\mathrm{t}} \rangle_{\mathrm{t}}|\,\mathrm{d}\Pi(b, \mathbf{a}_k) \bigg)^2. \end{aligned} \tag{33}$$

Our next task will be to lower bound the expectation $\int |\langle \mathbf{y}, \tilde{\boldsymbol{\psi}}_{\mathrm{t}} \rangle_{\mathrm{t}}|\,\mathrm{d}\Pi(b, \mathbf{a}_k)$. First, note the following identity: $\mathbb{1}(\mathbf{x} \in \mathrm{t}_L)\mathbb{P}(\mathrm{t}_R) - \mathbb{1}(\mathbf{x} \in \mathrm{t}_R)\mathbb{P}(\mathrm{t}_L) = -(\mathbb{1}(\mathbf{x}^{\mathrm{T}}\mathbf{a} > b) - \mathbb{P}(\mathrm{t}_R))\mathbb{1}(\mathbf{x} \in \mathrm{t})$, which means $\sqrt{\mathbb{P}(\mathrm{t}_L)\mathbb{P}(\mathrm{t}_R)}\langle \mathbf{y}, \tilde{\boldsymbol{\psi}}_{\mathrm{t}} \rangle_{\mathrm{t}} = \sqrt{\mathbb{P}(\mathrm{t}_L)\mathbb{P}(\mathrm{t}_R)}\langle \mathbf{y} - \hat{y}_{\mathrm{t}}, \tilde{\boldsymbol{\psi}}_{\mathrm{t}} \rangle_{\mathrm{t}} = -\langle \mathbf{y} - \hat{y}_{\mathrm{t}}, \mathbb{1}(\mathbf{x}^{\mathrm{T}}\mathbf{a} > b) \rangle_{\mathrm{t}}$. Using this identity together with the empirical measure (defined in (31)), we see that the expectation in (33) is lower bounded by

$$\begin{aligned} \int |\langle \mathbf{y} - \hat{y}_{\mathrm{t}}, \tilde{\boldsymbol{\psi}}_{\mathrm{t}} \rangle_{\mathrm{t}}|\,\mathrm{d}\Pi(b, \mathbf{a}_k) &= \frac{\sum_{k=1}^{M} \int |g_k'(b)|\,|\langle \mathbf{y} - \hat{y}_{\mathrm{t}}, \mathbb{1}(\mathbf{a}_k^{\mathrm{T}}\mathbf{x} > b) \rangle_{\mathrm{t}}|\,\mathrm{d}b}{\sum_{k'=1}^{M} \int |g_{k'}'(b')|\sqrt{\mathbb{P}(\mathrm{t}_L')\mathbb{P}(\mathrm{t}_R')}\,\mathrm{d}b'} \\ &\ge \frac{|\langle \mathbf{y} - \hat{y}_{\mathrm{t}}, \sum_{k=1}^{M} \int g_k'(b)\mathbb{1}(\mathbf{a}_k^{\mathrm{T}}\mathbf{x} > b)\,\mathrm{d}b \rangle_{\mathrm{t}}|}{\sum_{k'=1}^{M} \int |g_{k'}'(b')|\sqrt{\mathbb{P}(\mathrm{t}_L')\mathbb{P}(\mathrm{t}_R')}\,\mathrm{d}b'}. \end{aligned} \tag{34}$$

Then, by the definition of g_k', we have $\sum_{k=1}^{M} \int g_k'(b)\mathbb{1}(\mathbf{a}_k^{\mathrm{T}}\mathbf{x}_i > b)\,\mathrm{d}b = g(\mathbf{x}_i) - g(\mathbf{x}_1)$ for each $i = 1, 2, \ldots, n(\mathrm{t})$, and hence

$$\bigg\langle \mathbf{y} - \hat{y}_{\mathrm{t}}, \sum_{k=1}^{M} \int g_k'(b)\mathbb{1}\big(\mathbf{a}_k^{\mathrm{T}}\mathbf{x} > b\big)\,\mathrm{d}b \bigg\rangle_{\mathrm{t}} = \langle \mathbf{y} - \hat{y}_{\mathrm{t}}, \mathbf{g} \rangle_{\mathrm{t}}. \tag{35}$$

In light of (32), (33), (34) and (35), we obtain

$$\max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t}) \geq \frac{w(\mathrm{t})|\langle \mathbf{y}-\hat{y}_{\mathrm{t}},\mathbf{g}\rangle_{\mathrm{t}}|^2}{(\sum_{k'=1}^{M}\int |g'_{k'}(b')|\sqrt{\mathbb{P}(\mathrm{t}'_L)\mathbb{P}(\mathrm{t}'_R)}\,\mathrm{d}b')^2}. \tag{36}$$

Next, we derive upper and lower bounds on the denominator and numerator of (36), respectively. First, we look at the denominator. Note that for each k', the integral can be decomposed as follows:

$$\int |g'_{k'}(b')|\sqrt{\mathbb{P}(\mathrm{t}'_L)\mathbb{P}(\mathrm{t}'_R)}\,\mathrm{d}b' = \sum_{i=1}^{n(\mathrm{t})-1}\int_{\{b':n(\mathrm{t}'_L)=i\}} |g'_{k'}(b')|\sqrt{(i/n(\mathrm{t}))(1-i/n(\mathrm{t}))}\,\mathrm{d}b'. \tag{37}$$

Then, using the fact that $\sqrt{(i/n(\mathrm{t}))(1-i/n(\mathrm{t}))} \leq 1/2$ for $1 \leq i \leq n(\mathrm{t})$, and that the end points of each integral in the sum of (37) can be explicitly identified from the definition of $g'_{k'}$ in (30),

$$\begin{aligned}\int |g'_{k'}(b')|\sqrt{\mathbb{P}(\mathrm{t}'_L)\mathbb{P}(\mathrm{t}'_R)}\,\mathrm{d}b' &\leq \frac{1}{2}\sum_{i=1}^{n(\mathrm{t})-1}\int_{\{b':n(\mathrm{t}'_L)=i\}} |g'_{k'}(b')|\,\mathrm{d}b' \\ &= \frac{1}{2}\sum_{i=1}^{n(\mathrm{t})-1}\int_{\mathbf{a}_{k'}^{\mathrm{T}}\mathbf{x}_i}^{\mathbf{a}_{k'}^{\mathrm{T}}\mathbf{x}_{i+1}} |g'_{k'}(b')|\,\mathrm{d}b'.\end{aligned} \tag{38}$$

By the definition of $g'_{k'}$ as a divided difference (30) and the definition of total variation, for each k',

$$\sum_{i=1}^{n(\mathrm{t})-1}\int_{\mathbf{a}_{k'}^{\mathrm{T}}\mathbf{x}_i}^{\mathbf{a}_{k'}^{\mathrm{T}}\mathbf{x}_{i+1}} |g'_{k'}(b')|\,\mathrm{d}b' = \sum_{i=1}^{n(\mathrm{t})-1} |g_{k'}(\mathbf{a}_{k'}^{\mathrm{T}}\mathbf{x}_{i+1}) - g_{k'}(\mathbf{a}_{k'}^{\mathrm{T}}\mathbf{x}_i)| \leq V(g_{k'},\mathbf{a}_{k'},\mathrm{t}). \tag{39}$$

Combining (38) and (39) and plugging the result into the summation in the denominator of (36), we get

$$\sum_{k'=1}^{M}\int |g'_{k'}(b')|\sqrt{\mathbb{P}(\mathrm{t}'_L)\mathbb{P}(\mathrm{t}'_R)}\,\mathrm{d}b' \leq \frac{1}{2}\sum_{k'=1}^{M} V(g_{k'},\mathbf{a}_{k'},\mathrm{t}) = \frac{1}{2}\|g\|_{\mathcal{L}_1(\mathrm{t})}.$$

Next, we lower bound the numerator in (36). Using the Cauchy–Schwarz inequality and the fact that $\langle \mathbf{y}-\hat{y}_{\mathrm{t}},\mathbf{y}\rangle_{\mathrm{t}} = \|\mathbf{y}-\hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2$, we obtain

$$\langle \mathbf{y}-\hat{y}_{\mathrm{t}},\mathbf{g}\rangle_{\mathrm{t}} = \langle \mathbf{y}-\hat{y}_{\mathrm{t}},\mathbf{y}\rangle_{\mathrm{t}} - \langle \mathbf{y}-\hat{y}_{\mathrm{t}},\mathbf{y}-\mathbf{g}\rangle_{\mathrm{t}} \geq \|\mathbf{y}-\hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|\mathbf{y}-\hat{y}_{\mathrm{t}}\|_{\mathrm{t}}\|\mathbf{y}-\mathbf{g}\|_{\mathrm{t}}. \tag{40}$$

By the AM-GM inequality, we know that $\|\mathbf{y}-\hat{y}_{\mathrm{t}}\|_{\mathrm{t}}\|\mathbf{y}-\mathbf{g}\|_{\mathrm{t}} \leq \frac{1}{2}(\|\mathbf{y}-\hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2 + \|\mathbf{y}-\mathbf{g}\|_{\mathrm{t}}^2)$. Plugging this into (40), we get $\langle \mathbf{y}-\hat{y}_{\mathrm{t}},\mathbf{g}\rangle_{\mathrm{t}} \geq \frac{1}{2}(\|\mathbf{y}-\hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|\mathbf{y}-\mathbf{g}\|_{\mathrm{t}}^2)$. Now, squaring both sides and using the assumption that $R_{K-1}(\mathrm{t}) > 0$, we have

$$|\langle \mathbf{y}-\hat{y}_{\mathrm{t}},\mathbf{g}\rangle_{\mathrm{t}}|^2 \geq \frac{1}{4}(\|\mathbf{y}-\hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|\mathbf{y}-\mathbf{g}\|_{\mathrm{t}}^2)^2 = \frac{1}{4}R_{K-1}^2(\mathrm{t}).$$

Now we can put the bounds on the numerator and denominator together to get the desired result:

$$\max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},\mathrm{t}) \geq \frac{w(\mathrm{t})R_{K-1}^2(\mathrm{t})}{\|g\|_{\mathcal{L}_1(\mathrm{t})}^2}.$$

□

Next, we provide a solution to a simple recursive inequality.

LEMMA 6.2. *Let* $\{a_k\}$ *be a decreasing sequence of numbers and* $\{b_k\}$ *be a positive sequence numbers satisfying the following recursive expression*:

$$a_k \le a_{k-1}(1 - b_k a_{k-1}), \quad k = 1, 2, \ldots, K.$$

Then

$$a_K \le \frac{1}{\sum_{k=1}^{K} b_k}, \quad K = 1, 2, \ldots$$

PROOF OF LEMMA 6.2. We may assume without loss of generality that $a_{K-1} > 0$; otherwise the result holds trivially since $a_K \le a_{K-1} \le 0 \le \frac{1}{\sum_{k=1}^{K} b_k}$. For $K = 1$,

$$a_1 \le a_0(1 - b_1 a_0) \le \frac{1}{4b_1} < \frac{1}{b_1}.$$

For $K > 1$, assume $a_{K-1} \le \frac{1}{\sum_{k=1}^{K-1} b_k}$. Then, either $a_{K-1} \le \frac{1}{\sum_{k=1}^{K} b_k}$, in which case we are done since $a_K \le a_{K-1}$, or, $a_{K-1} \ge \frac{1}{\sum_{k=1}^{K} b_k}$, in which case

$$a_K \le a_{K-1}(1 - b_K a_{K-1}) \le \frac{1}{\sum_{k=1}^{K-1} b_k}\left(1 - \frac{b_K}{\sum_{k=1}^{K} b_k}\right) = \frac{1}{\sum_{k=1}^{K} b_k}. \qquad \square$$

Acknowledgments. The authors would like to thank Florentina Bunea, Sameer Deshpande, Jianqing Fan, Yingying Fan, Jonathan Siegel, Bartolomeo Stellato and William Underwood for insightful discussions. The authors are particularly grateful to two anonymous reviewers whose comments improved the quality of the paper.

Funding. MDC was supported in part by the National Science Foundation through SES-2019432 and SES-2241575.

JMK was supported in part by the National Science Foundation through CAREER DMS-2239448, DMS-2054808 and HDR TRIPODS CCF-1934924.

SUPPLEMENTARY MATERIAL

Supplement to "Convergence rates of oblique regression trees for flexible function libraries" (DOI: 10.1214/24-AOS2354SUPP; .pdf). Omitted proofs of theoretical results presented in the main text, namely Theorems 2.3, 3.2, 4.1 and Corollaries 2.4 and 2.5.

REFERENCES

[1] BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inf. Theory* **39** 930–945. MR1237720 https://doi.org/10.1109/18.256500

[2] BARRON, A. R. (1994). Approximation and estimation bounds for artificial neural networks. *Mach. Learn.* **14** 115–133.

[3] BARRON, A. R., COHEN, A., DAHMEN, W. and DEVORE, R. A. (2008). Approximation and learning by greedy algorithms. *Ann. Statist.* **36** 64–94. MR2387964 https://doi.org/10.1214/009053607000000631

[4] BENNETT, K. P. (1994). Global tree optimization: A non-greedy decision tree algorithm. *J. Comput. Sci. Stat.* 156–156.

[5] BERTSIMAS, D. and DUNN, J. (2017). Optimal classification trees. *Mach. Learn.* **106** 1039–1082. MR3665788 https://doi.org/10.1007/s10994-017-5633-9

[6] BERTSIMAS, D. and DUNN, J. (2019). *Machine Learning Under a Modern Optimization Lens*. Dynamic Ideas LLC.

[7] BERTSIMAS, D., DUNN, J. and WANG, Y. (2021). Near-optimal nonlinear regression trees. *Oper. Res. Lett.* **49** 201–206. MR4204496 https://doi.org/10.1016/j.orl.2021.01.002

[8] BERTSIMAS, D., MAZUMDER, R. and SOBIESK, M. (2018). Optimal classification and regression trees with hyperplanes are as powerful as classification and regression neural networks. Unpublished manuscript.
[9] BERTSIMAS, D. and STELLATO, B. (2021). The voice of optimization. *Mach. Learn.* **110** 249–277. MR4207500 https://doi.org/10.1007/s10994-020-05893-5
[10] BREIMAN, L. (2001). Random forests. *Mach. Learn.* **45** 5–32.
[11] BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984). *Classification and Regression Trees*. *Wadsworth Statistics/Probability Series*. Wadsworth Advanced Books and Software, Belmont, CA. MR0726392
[12] BRODLEY, C. E. and UTGOFF, P. E. (1995). Multivariate decision trees. *Mach. Learn.* **19** 45–77.
[13] BUCILUUNDEFINED, C., CARUANA, R. and NICULESCU-MIZIL, A. (2006). Model compression. In *Proceedings of the* 12*th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. *KDD'*06 535–541. Association for Computing Machinery, New York, NY, USA.
[14] CATTANEO, M. D., CHANDAK, R. and KLUSOWSKI, J. M. (2024). Supplement to "Convergence rates of oblique regression trees for flexible function libraries." https://doi.org/10.1214/24-AOS2354SUPP
[15] CATTANEO, M. D., FARRELL, M. H. and FENG, Y. (2020). Large sample properties of partitioning-based series estimators. *Ann. Statist.* **48** 1718–1741. MR4124341 https://doi.org/10.1214/19-AOS1865
[16] CHI, C.-M., VOSSLER, P., FAN, Y. and LV, J. (2022). Asymptotic properties of high-dimensional random forests. *Ann. Statist.* **50** 3415–3438. MR4524502 https://doi.org/10.1214/22-aos2234
[17] DEVORE, R., NOWAK, R. D., PARHI, R. and SIEGEL, J. W. (2023). Weighted variation spaces and approximation by shallow ReLU networks. ArXiv preprint. Available at arXiv:2307.15772.
[18] DUNN, J. W. (2018). Optimal trees for prediction and prescription. Ph.D. thesis, Massachusetts Institute of Technology.
[19] DURRETT, R. (2019). *Probability—Theory and Examples*, 5th ed. *Cambridge Series in Statistical and Probabilistic Mathematics* **49**. Cambridge Univ. Press, Cambridge. MR3930614 https://doi.org/10.1017/9781108591034
[20] FROSST, N. and HINTON, G. (2017). Distilling a neural network into a soft decision tree. Preprint. Available at arXiv:1711.09784.
[21] GHOSH, P., AZAM, S., JONKMAN, M., KARIM, A., SHAMRAT, F. M. J. M., IGNATIOUS, E., SHULTANA, S., BEERAVOLU, A. R. and DE BOER, F. (2021). Efficient prediction of cardiovascular disease using machine learning algorithms with relief and LASSO feature selection techniques. *IEEE Access* **9** 19304–19326.
[22] GYÖRFI, L., KOHLER, M., KRZYŻAK, A. and WALK, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*. *Springer Series in Statistics*. Springer, New York. MR1920390 https://doi.org/10.1007/b97848
[23] HASTIE, T., TIBSHIRANI, R. and FRIEDMAN, J. (2009). *The Elements of Statistical Learning*: *Data Mining*, *Inference*, *and Prediction*, 2nd ed. *Springer Series in Statistics*. Springer, New York. MR2722294 https://doi.org/10.1007/978-0-387-84858-7
[24] HEATH, D., KASIF, S. and SALZBERG, S. (1993). Induction of oblique decision trees. *J. Artificial Intelligence Res.* **2** 1–32.
[25] HUANG, J. Z. (2003). Local asymptotics for polynomial spline regression. *Ann. Statist.* **31** 1600–1635. MR2012827 https://doi.org/10.1214/aos/1065705120
[26] HÜLLERMEIER, E., MOHR, F., TORNEDE, A. and WEVER, M. (2021). Automated machine learning, bounded rationality, and rational metareasoning. Preprint. Available at arXiv:2109.04744.
[27] KLUSOWSKI, J. M. (2020). Sparse learning with CART. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 11612–11622. Curran Associates, Red Hook, NY.
[28] KLUSOWSKI, J. M. and TIAN, P. (2023). Large scale prediction with decision trees. *J. Amer. Statist. Assoc.*
[29] LEE, G.-H. and JAAKKOLA, T. S. (2020). Oblique decision trees from derivatives of ReLU networks. In *International Conference on Learning Representations*.
[30] LI, X.-B., SWEIGART, J. R., TENG, J. T. C., DONOHUE, J. M., THOMBS, L. A. and WANG, S. M. (2003). Multivariate decision trees using linear discriminants and tabu search. *IEEE Trans. Syst. Man Cybern., Part A, Syst. Humans* **33** 194–205.
[31] LOH, W.-Y. and SHIH, Y.-S. (1997). Split selection methods for classification trees. *Statist. Sinica* **7** 815–840. MR1488644
[32] LÓPEZ-CHAU, A., CERVANTES, J., LÓPEZ-GARCÍA, L. and LAMONT, F. G. (2013). Fisher's decision tree. *Expert Syst. Appl.* **40** 6283–6291.
[33] MENZE, B. H., KELM, B. M., SPLITTHOFF, D. N., KOETHE, U. and HAMPRECHT, F. A. (2011). On oblique random forests. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 453–469. Springer, Berlin.

[34] MINGERS, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Mach. Learn.* **4** 227–243.

[35] MURDOCH, W. J., SINGH, C., KUMBIER, K., ABBASI-ASL, R. and YU, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proc. Natl. Acad. Sci. USA* **116** 22071–22080. MR4030584 https://doi.org/10.1073/pnas.1900654116

[36] MURTHY, S. K., KASIF, S. and SALZBERG, S. (1994). A system for induction of oblique decision trees. *J. Artificial Intelligence Res.* **2** 1–32.

[37] PARHI, R. and NOWAK, R. D. (2023). Deep learning meets sparse regularization: A signal processing perspective. Preprint. Available at arXiv:2301.09554.

[38] QUINLAN, J. R. (1993). C4.5, programs for machine learning. In *Proc. of* 10*th International Conference on Machine Learning* 252–259.

[39] RAYMAEKERS, J., ROUSSEEUW, P. J., VERDONCK, T. and YAO, R. (2023). Fast linear model trees by PILOT. Preprint. Available at arXiv:2302.03931.

[40] RODRIGUEZ, J. J., KUNCHEVA, L. I. and ALONSO, C. J. (2006). Rotation forest: A new classifier ensemble method. *IEEE Trans. Pattern Anal. Mach. Intell.* **28** 1619–1630.

[41] RUDIN, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nat. Mach. Intell.* **1** 206–215. https://doi.org/10.1038/s42256-019-0048-x

[42] SCORNET, E., BIAU, G. and VERT, J.-P. (2015). Consistency of random forests. *Ann. Statist.* **43** 1716–1741. MR3357876 https://doi.org/10.1214/15-AOS1321

[43] SYRGKANIS, V. and ZAMPETAKIS, M. (2020). Estimation and inference with trees and forests in high dimensions. In *Proceedings of Thirty Third Conference on Learning Theory* (J. Abernethy and S. Agarwal, eds.). *Proceedings of Machine Learning Research* **125** 3453–3454. PMLR.

[44] TOMITA, T. M., BROWNE, J., SHEN, C., CHUNG, J., PATSOLIC, J. L., FALK, B., PRIEBE, C. E., YIM, J., BURNS, R. et al. (2020). Sparse projection oblique randomer forests. *J. Mach. Learn. Res.* **21** 1–39.

[45] WAGER, S. and ATHEY, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *J. Amer. Statist. Assoc.* **113** 1228–1242. MR3862353 https://doi.org/10.1080/01621459.2017.1319839

[46] YANG, Y., MORILLO, I. G. and HOSPEDALES, T. M. (2018). Deep neural decision trees. In *ICML Workshop on Human Interpretability in Machine Learning* (*WHI*).

[47] ZHAN, H., LIU, Y. and XIA, Y. (2023). Consistency of the oblique decision tree and its random forest. Preprint. Available at arXiv:2211.12653.

[48] ZHANG, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Trans. Inf. Theory* **49** 682–691. MR1967192 https://doi.org/10.1109/TIT.2002.808136

[49] ZHU, H., MURALI, P., PHAN, D., NGUYEN, L. and KALAGNANAM, J. (2020). A scalable MIP-based method for learning optimal multivariate decision trees. In *Advances in Neural Information Processing Systems* (H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan and H. Lin, eds.) **33** 1771–1781. Curran Associates, Red Hook, NY.