Convergence Rates of Oblique Regression Trees for Flexible Function Libraries*

Matias D. Cattaneo† Rajita Chandak† Jason M. Klusowski†

October 25, 2022

Abstract

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 need 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.

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 authors would like to thank Jonathan Siegel, William Underwood, and Bartolomeo Stellato for insightful discussions. MDC was supported in part by the National Science Foundation through grant SES-2019432. JMK was supported in part by the National Science Foundation through grants DMS-2054808 and CCF-1934924.

†Department of Operations Research and Financial Engineering, Princeton University.

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 (Murdoch et al., 2019; Rudin, 2019). 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 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 (Buciluundefined et al., 2006) is a quick and easy way to enhance the fidelity of an interpretable model, without degenerating 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 (Frosst and Hinton, 2017). 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.

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 (Breiman et al., 1984), 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)

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

with respect to $(b, \mathbf{a})$, with $\mathbb{1}(\cdot)$ denoting the indicator function and $\bar{y}_{\mathrm{t}}$ 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 (Breiman et al., 1984, 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 hyper-rectangles. On the other hand, the *oblique CART* algorithm (Breiman et al., 1984, 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 $\mathrm{t}_L = \{\mathbf{x} \in \mathrm{t} : \hat{\mathbf{a}}^{\mathrm{T}}\mathbf{x} \leq \hat{b}\}$ 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 tree with depth $K = 2$ is shown in Figure 1.

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 $\bar{y}_{\mathrm{t}}$, i.e., if T is a decision tree, then

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

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, $\hat{y}_{\mathrm{t}} \in \operatorname{argmin}_{h \in \operatorname{span}(\mathcal{H})} \sum_{\mathbf{x}_i \in \mathrm{t}} (y_i - h(\mathbf{x}_i))^2$.

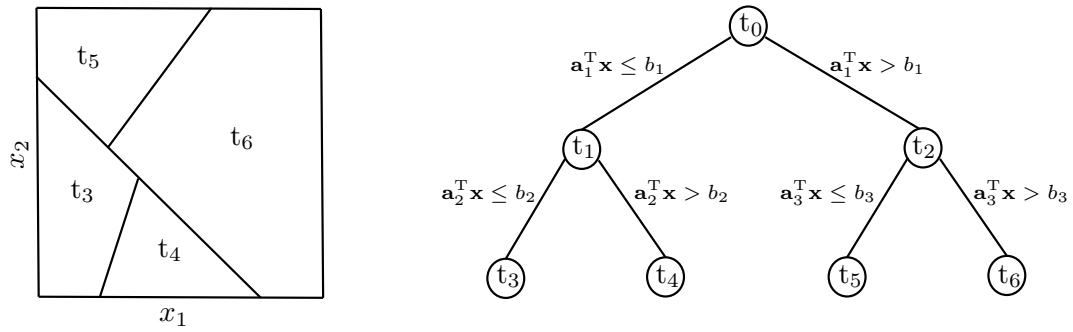

Figure 1: A maximal oblique decision tree with depth $K = 2$ in $p = 2$ dimensions. Splits occur along hyperplanes of the form $a_1 x_1 + a_2 x_2 = b$.

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 (Murthy et al., 1994).

To deal with these computational demands, Breiman et al. (1984) 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 et al. (1993) 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 et al. (1994) use a combination of deterministic hill-climbing and random perturbations in an attempt to find a good hyperplane. See (Brodley and Utgoff, 1995) for additional variations on these algorithms. Other works employ statistical techniques like linear discriminant analysis (LDA) (López-Chau et al., 2013; Li et al., 2003; Loh and Shih, 1997), principle components analysis (PCA) (Menze et al., 2011; Rodriguez et al., 2006), and random projections (Tomita et al., 2020).

While not the focus of the present paper, regarding non-greedy training, other researchers have attempted to find globally optimal tree solutions using linear programming (Bennett, 1994) or mixed-integer linear programming (Bertsimas and Dunn, 2017; Bertsimas et al., 2021). 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 datasets 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 exihibit bounded rationality and therefore make decisions in a more sequential (rather than anticipatory) manner (Hüllermeier et al., 2021, and references therein). Relatedly, another training technique is based on constructing deep neural networks that realize oblique decision trees (Lee and Jaakkola, 2020; Yang et al., 2018) 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 (Heath et al., 1993; Murthy et al., 1994) and can often be comparable, in terms of performance, to neural networks (Bertsimas et al., 2018; Bertsimas and Dunn, 2019; Bertsimas and Stellato, 2021). 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 linear combinations of ridge functions, i.e., the library

$$\mathcal{G} = \left\{ g(\mathbf{x}) = \sum_k g_k(\mathbf{x}^{\mathrm{T}}\mathbf{a}_k),\ \mathbf{a}_k \in \mathbb{R}^p,\ g_k : \mathbb{R} \mapsto \mathbb{R} \right\}.$$

This library encompasses the functions produced from projection pursuit regression, and, more specifically, by taking all g_k to be a fixed activation function such as a sigmoid function or ReLU, single-hidden layer feed-forward neural networks.

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 $f \in \mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$, let $\|f\|^2 = \int_{\mathcal{X}} (f(\mathbf{x}))^2 d\mathbb{P}_{\mathbf{x}}(\mathbf{x})$ be the squared $\mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$ norm and let $\|f\|_n^2 = \frac{1}{n}\sum_{i=1}^n (f(\mathbf{x}_i))^2$ denote the squared norm with respect to the empirical measure on the data. Let $\langle f, g\rangle_n = \frac{1}{n}\sum_{i=1}^n f(\mathbf{x}_i)g(\mathbf{x}_i)$ denote the squared inner product with respect to the empirical measure on the data. The response data vector $\mathbf{y} = (y_1, y_2, \ldots, y_n)^{\mathrm{T}} \in \mathbb{R}^n$ is viewed as a relation, defined on the design matrix $\mathbf{X} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n)^{\mathrm{T}} \in \mathbb{R}^{n\times p}$, that associates $\mathbf{x}_i$ with y_i. Thus, for example, $\|y - f\|_n^2 = \frac{1}{n}\sum_{i=1}^n (y_i - f(\mathbf{x}_i))^2$ and $\langle y, f\rangle_n = \frac{1}{n}\sum_{i=1}^n y_i f(\mathbf{x}_i)$. Furthermore, we use $\{\mathrm{t} : \mathrm{t} \in [T]\}$ to denote the collection of internal (non-terminal) 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 $\mathbf{x} \mapsto h(\mathbf{x}^{\mathrm{T}}\mathbf{a})$ with $\mathbf{a} \in \mathbb{R}^p$ and $h : \mathbb{R} \mapsto \mathbb{R}$ in the node t as

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

where the supremum is over all partitions $\mathcal{P} = \{z_0, z_1, \ldots, z_{|\mathcal{P}|}\}$ of the interval $I(\mathbf{a}, \mathrm{t}) = [\min_{\mathbf{x}\in\mathrm{t}} \mathbf{x}^{\mathrm{T}}\mathbf{a},\ \max_{\mathbf{x}\in\mathrm{t}} \mathbf{x}^{\mathrm{T}}\mathbf{a}] \subset \mathbb{R}$ (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 $\int_{I(\mathbf{a},\mathrm{t})} |h'(z)|\mathrm{d}z$. 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

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

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 (1993) shows that any function $f : \mathbb{R}^p \mapsto \mathbb{R}$ with finite $\int \|\boldsymbol{\theta}\|_{\ell_1} |\tilde{f}(\boldsymbol{\theta})| \mathrm{d}\boldsymbol{\theta}$, where $\tilde{f}$ is the Fourier transform of f and $\|\cdot\|_{\ell_1}$ is the usual ℓ_1 norm of a vector in $\mathbb{R}^p$, belongs to $\mathcal{F}$ when $\mathcal{X} = [0, 1]^p$ and $\mathcal{G}$ consists of linear combinations of sigmoidal ridge functions. Furthermore, we have the bound $\|f\|_{\mathcal{L}_1} \lesssim \int \|\boldsymbol{\theta}\|_{\ell_1} |\tilde{f}(\boldsymbol{\theta})| \mathrm{d}\boldsymbol{\theta}$.

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 unlikely to find the global maximum, we theoretically measure their

success by specifying a sub-optimality (slackness) parameter $\kappa \in (0, 1]$ and considering the probability $P_{\mathcal{A}_t}(\kappa)$ that the maximum of $\widehat{\Delta}(b, \mathbf{a}, t)$ over $\mathbf{a} \in \mathcal{A}_t \subseteq \mathbb{R}^p$ is within a factor κ of the maximum of $\widehat{\Delta}(b, \mathbf{a}, t)$ on the unrestricted parameter space, $\mathbf{a} \in \mathbb{R}^p$. That is, to theoretically quantify the sub-optimality of the chosen hyperplane, we measure

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

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

Another way of thinking about $P_{\mathcal{A}_t}(\kappa)$ is that it represents the degree of optimization mis-specificity of $\mathcal{A}_t$ for the form of the global optimum $\hat{\mathbf{a}}$. For example, if $\mathcal{A}_t = \{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p\}$ is the collection of standard basis vectors in $\mathbb{R}^p$, then $d = 1$ and we believe that the true optimal solution $\hat{\mathbf{a}} \in \mathcal{A}_t$ corresponds to axis-aligned CART. More formally, the event in the definition of $P_{\mathcal{A}_t}(\kappa)$ is equivalent to failing to reject the null hypothesis $H_0 : \hat{\mathbf{a}} \in \mathcal{A}_t$, under the regression model $y = \beta_1 \mathbb{1}(\mathbf{a}^T\mathbf{x} \leq b) + \beta_2 \mathbb{1}(\mathbf{a}^T\mathbf{x} > b) + \varepsilon$ with independent Gaussian noise $\varepsilon \sim N(0, \sigma^2)$, using the likelihood ratio test with threshold value proportional to $1 - \kappa$. The smaller κ is, the more likely we will reject the null hypothesis that $\hat{\mathbf{a}}$ belongs to $\mathcal{A}_t$.

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

- **Deterministic.** If $\mathcal{A}_t$ is nonrandom, then $P_{\mathcal{A}_t}(\kappa)$ is either zero or one for any $\mathcal{A}_t \subset \mathbb{R}^p$, and if $\mathcal{A}_t = \mathbb{R}^p$, then $P_{\mathcal{A}_t}(\kappa) = 1$ for all $\kappa \in (0, 1]$. For the latter case, one can use strategies based on mixed-integer optimization (MIO) Zhu et al. (2020); Dunn (2018); Bertsimas and Dunn (2017). In particular, Dunn (2018) 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 $\mathcal{A}_t = \mathbb{Z}^p$, then $P_{\mathcal{A}_t}(\kappa) = 1$ 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 $\mathcal{A}_t$ 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 $\mathcal{A}_t$ is generated by selecting m standard basis vectors uniformly at random without replacement (as is done with random forests (Breiman et al., 1984)), then $P_{\mathcal{A}_t}(\kappa) \geq m/p$ 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 $P_{\mathcal{A}_t}(\kappa)$. This has direct consequences for the predictive performance, since, as we shall see, 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 $\mathcal{A}_{\mathrm{t}}$ 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 (Tomita et al., 2020). 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 (Ghosh et al., 2021; Menze et al., 2011; Rodriguez et al., 2006).

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 $\mathcal{A}_{\mathrm{t}}$, either with prior knowledge based on the context of the problem, or with an effective data-driven strategy.

We now present a technical result about the construction of trees that is crucial in proving our main results. All proofs are provided in full generality in Appendix A.

2.3 Orthogonal Tree Expansions

Lemma 1 shows that the tree output $\widehat{\mu}(T)$ 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:

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

We refer the reader to Appendix A 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 1

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

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

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

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

Remark 1 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 $\mathbf{\Psi}$ is the $n \times |[T]|$ data matrix $[\psi_t(\mathbf{x}_i)]_{1\leq i\leq n,\, t\in[T]}$ and $\mathbf{\Psi}(\mathbf{x})$ is the $|[T]| \times 1$ feature vector $(\psi_t(\mathbf{x}))_{t\in[T]}$. Then,

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

From this perspective, the key impediment is that the implied (random) basis functions depend on the entire sample $(\mathbf{y}, \mathbf{X})$, and hence standard sieve estimation and inference theory (Huang, 2003; Cattaneo et al., 2020) does not apply.

Lemma 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, ψ_t, most correlated with the residuals, $y - \overline{y}_t$, per (5), 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 over-complete dictionaries will recognize some similarities in the analysis (see the *orthogonal greedy algorithm* (Barron et al., 2008) 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 $\widehat{\mu}(T_K)$ in Lemma 1 allows one to write down a recursive expression for the training error. That is, from $\widehat{\mu}(T_K) = \widehat{\mu}(T_{K-1}) + \sum_{t\in T_{K-1}} \langle y, \psi_t\rangle_n \psi_t$, one obtains the identity

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

Furthermore, using the fact that $\langle y, \psi_t\rangle_n$ is the result of a local maximization, viz., the equivalence in (5) in Lemma 1, one can construct an empirical probability measure Π on $(b, \mathbf{a})$ and lower bound $|\langle y, \psi_t\rangle_n|^2$ by $\int \widehat{\Delta}(b, \mathbf{a}, t)\mathrm{d}\Pi(b, \mathbf{a})$, which is itself further lower bounded by an appropriately scaled squared node-wise excess training error. These inequalities can be combined with (6) to provide a useful training error bound. We formally present this result next.

2.4 Training Error Bound for Oblique CART

Applying the techniques outlined earlier, we can show the following result (Lemma 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 (Scornet et al., 2015; Wager and Athey, 2018; Syrgkanis and Zampetakis, 2020; Chi et al., 2022).

For the following statements, the output of a depth K tree T_K constructed with oblique CART methodology using the search space $\mathcal{A}_t$ is denoted $\widehat{\mu}(T_K)$. 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 (Training error bound for oblique CART)
Let $\mathbb{E}[y^2 \log(1 + |y|)] < \infty$ *and* $g \in \mathcal{F} \cap \mathscr{L}_2(\mathbb{P}_{\mathbf{x}})$ *with* $\|g\|_{\mathcal{L}_1} < \infty$. *Then, for any* $K \geq 1$,

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

For this result to be non-vacuous, 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 these probabilities are bounded away from zero.

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. (1984, Section 10.2), the smallest minimizing subtree for the penalty coefficient $\lambda = \lambda_n \geq 0$,

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

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 a fixed λ, the optimal subtree T_{opt} can be found efficiently by weakest link pruning, i.e., by successively collapsing the internal node that decreases $\|y - \widehat{\mu}(T)\|_n^2$ the most, until we arrive at the single-node tree consisting of the root node. Good values of λ can be selected using cross-validation on a holdout subset of data, for example. See Mingers (1989) for a description of various pruning algorithms.

We now present our main consistency 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 mis-specification; 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 can be coupled with tools from empirical process theory (Györfi et al., 2002) for studying partition-based estimators.

Our results rely on the following assumption regarding the data generating process.

Assumption 1 (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 $\mathbf{x} \in \mathcal{X}$,

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

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

$$\sup_{\mathbf{x} \in \mathcal{X}} |\mu(\mathbf{x})| \leq M + c_1 \int_0^{\infty} \exp(-c_2 z^{\gamma}) \mathrm{d}z = M' < \infty. \tag{9}$$

Theorem 1
Let Assumption 1 hold. Then, for any $K \geq 1$,

$$\begin{aligned} &\mathbb{E}[\|\mu - \widehat{\mu}(T_K)\|^2] \\ &\quad \leq 2 \inf_{f \in \mathcal{F}} \left\{ \|\mu - f\|^2 + \frac{\|f\|_{\mathcal{L}_1}^2 \mathbb{E}[\max_{\mathrm{t} \in [T_K]} P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)]}{\kappa K} + C \frac{2^K d \log(np/d) \log^{4/\gamma}(n)}{n} \right\}, \end{aligned} \tag{10}$$

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

$$\begin{aligned} &\mathbb{E}[\|\mu - \widehat{\mu}(T_{opt})\|^2] \\ &\quad \leq 2 \inf_{K \geq 1,\, f \in \mathcal{F}} \left\{ \|\mu - f\|^2 + \frac{\|f\|_{\mathcal{L}_1}^2 \mathbb{E}[\max_{\mathrm{t} \in [T_K]} P_{\mathcal{A}_{\mathrm{t}}}^{-1}(\kappa)]}{\kappa K} + C \frac{2^K d \log(np/d) \log^{4/\gamma}(n)}{n} \right\}. \end{aligned} \tag{11}$$

Theorem 1 implies that if the probabilities $P_{\mathcal{A}_{\mathrm{t}}}(\kappa)$ are bounded away from zero, $d = p = o(n/\log(n))$ and the model is well-specified (i.e., $\mu \in \mathcal{F}$), then the pruned tree is consistent:

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

where $\{\mu_n\}$ is a sequence of regression functions that belong to $\mathcal{F}$ with $\sup_n \|\mu_n\|_{\mathcal{L}_1} < \infty$. As a special case, if we are dealing with axis-aligned CART ($d = 1$) and $\mathcal{F}$ is the additive library

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

then $\kappa = 1$ and $P_{\mathcal{A}_{\mathrm{t}}}(\kappa) = 1$. In this case, according to (11), the pruned tree estimator is consistent even in the so-called NP-dimensionality regime, where $\log(p) = o(n)$. A result for axis-aligned decision trees was obtained previously in Klusowski and Tian (2022).

These sort of high dimensional consistency guarantees are not possible with non-adaptive 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 (Györfi et al., 2002) 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, the oracle inequality in (11) yields relatively slow rates of convergence of the order $(\log(1/r_n))^{-1}$, with $r_n = (p/n)\log(n)$, for the IMSE. Because shallow oblique trees often compete empirically with wide neural networks (Bertsimas et al., 2018; Bertsimas and Dunn, 2019; Bertsimas and Stellato, 2021), a proper mathematical theory should reflect such qualities. It is therefore natural to compare these rates with the significantly better $\sqrt{r_n}$ rates for similar function libraries, achieved by neural networks (Barron, 1994). 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 assumption puts an ℓ_q constraint on the $\mathcal{L}_1$ total variations of the regression function μ across all terminal nodes of T_K. This is a type of sparsity (regularity) condition on both the tree partition of $\mathcal{X}$ and the regression function μ. It ensures a degree of compatibility between the non-additive tree model and the additive form of the regression function.

Assumption 2 (Node ℓ_q sparsity)
The regression function μ belongs to $\mathcal{F}$ and there exist positive numbers $V > 0$ and $q > 2$ such that, for any $K \geq 1$,

$$\mathbb{E}\left[\sum_{t\in T_K} \|\mu\|^q_{\mathcal{L}_1(t)}\right] \leq V^q. \tag{12}$$

For fixed K and finite $\|\mu\|_{\mathcal{L}_1}$, there is always some choice of V and q for which (12) is satisfied since

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

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 2 because q may depend on the sample size through its

dependence on the depth $K = K_n$. Hence, it is important that (12) hold for the same q *uniformly* over all depths.

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

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

In general, for $p > 1$, a crude and not very useful bound is $\sum_{\mathrm{t}\in T_K} \|\mu\|^q_{\mathcal{L}_1(\mathrm{t})} \leq 2^K \|\mu\|^q_{\mathcal{L}_1}$; however, the average size of $\sum_{\mathrm{t}\in T_K} \|\mu\|^q_{\mathcal{L}_1(\mathrm{t})}$ 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) and thus $\mathbb{E}[\sum_{\mathrm{t}\in T_K} \|\mu\|^p_{\mathcal{L}_1(\mathrm{t})}] = O(1)$.

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

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

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

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

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 3 implies that, on average, no region contains disproportionately more observations than the average number of observations per region, i.e., $n/2^K$. 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 by Bertsimas et al. (2018) and Bertsimas and Dunn (2019) showed 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

Let Assumptions 2 and 3 hold, and assume $\mathbb{E}[y^2 \log(1 + |y|)] < \infty$ *and* $\mu \in \mathcal{F}$. *Then, for any* $K \geq 1$,

$$\mathbb{E}[\|y - \widehat{\mu}(T_K)\|_n^2] \leq \mathbb{E}[\|y - \mu\|_n^2] + \frac{AV^2}{4^{(K-1)/q}}. \tag{13}$$

Theorem 2

Let Assumptions 1, 2, and 3 hold. Then, for any $K \geq 1$,

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_K)\|^2] \leq \frac{2AV^2}{4^{(K-1)/q}} + C\frac{2^{K+1} p \log^{4/\gamma+1}(n)}{n}, \tag{14}$$

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

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_{opt})\|^2] \leq 2(2+q)\left(\frac{AV^2}{q}\right)^{q/(2+q)}\left(\frac{Cp\log^{4/\gamma+1}(n)}{n}\right)^{2/(2+q)}. \tag{15}$$

As mentioned earlier, we see from (15) that $A = A_n$ (as well as $V = V_n$) is allowed to grow poly-logarithmically 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 (1994) for neural network estimators of regression functions $\mu \in \mathcal{F}$ with $\|\mu\|_{\mathcal{L}_1} < \infty$. While we make two additional assumptions (Assumptions 2 and 3) 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 (15) should always be at least as fast as the minimax optimal rate $(1/n)^{2/(2+p)}$ for smooth functions in p dimensions.

4 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.

4.1 Multi-layer Networks

We can go beyond single-hidden layer neural networks if instead the split boundaries have the form $\mathbf{a}^{\mathrm{T}}\Phi(\mathbf{x}) = b$, where Φ is a feature map, such as the output layer of a neural network, i.e., $\Phi_k(\mathbf{x}) = \phi(\mathbf{a}_k^{\mathrm{T}}\mathbf{x} - b_k)$, where ϕ is some activation function. Here the functions we can approximate look like $\sum_{k_2} c_{k_2}\phi(\sum_{k_1} c_{k_1,k_2}\phi(\mathbf{a}_{k_1,k_2}^{\mathrm{T}}\mathbf{x} - b_{k_1,k_2}))$, encompassing two-hidden layer networks.

4.2 Oblique Random Forests

Our theory can be readily modified to accommodate ensembles of oblique decision trees through averaging, if the constituent trees are constructed with sub-sampled data (i.e., a fraction of the data is selected at random without replacement). While the efficacy of forests is not reflected in these risk bounds, they show in principle that the methodology does not lead to a loss in performance.

4.3 Classification

While we have focused on regression trees, oblique decision trees are commonly applied to the problem of binary classification, i.e., $y_i \in \{-1, 1\}$. One of the most widely used splitting criterion is 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 (Quinlan, 1993):

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

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,

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

Therefore, there is hope that one could exploit connections to sequential greedy algorithms for other convex optimization problems (Zhang, 2003) (e.g., LogitBoost) and establish a training error bound (with respect to logistic loss) akin to Lemma 2.

A Proofs

The main text presented theory for 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 $\hat{y}_{\mathrm{t}} \in \operatorname{argmax}_{h \in \mathrm{span}(\mathcal{H})} \sum_{\mathbf{x}_i \in \mathrm{t}} (y_i - h(\mathbf{x}_i))^2$, where $\mathcal{H}$ is a finite 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 $\|f\|_\infty = \sup_{\mathbf{x}\in\mathcal{X}} |f(\mathbf{x})|$.

Proof of Lemma 1. Set $\mathcal{U}_\mathrm{t} = \{u(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) + v(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_R) : u, v \in \mathrm{span}(\mathcal{H})\}$ and consider the closed subspace $\mathcal{V}_\mathrm{t} = \{v(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}) : v \in \mathrm{span}(\mathcal{H})\}$. 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 $\mathbf{\Psi}_\mathrm{t}$ 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 $\mathbf{\Psi}_\mathrm{t}^\perp$ be any orthonormal basis for $\mathcal{V}_\mathrm{t}^\perp$ that includes the decision stump (3). We will show that

$$\widehat{\mu}(T)(\mathbf{x}) = \sum_{\mathrm{t}\in[T]} \sum_{\psi\in\mathbf{\Psi}_\mathrm{t}^\perp} \langle y, \psi\rangle_n \psi(\mathbf{x}), \tag{16}$$

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

$$\sum_{\psi\in\mathbf{\Psi}_\mathrm{t}^\perp} |\langle y, \psi\rangle_n|^2 = \widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, \mathrm{t}). \tag{17}$$

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 $\hat{y}_\mathrm{t}(\mathbf{x}) = \sum_{\psi\in\mathbf{\Psi}_\mathrm{t}} \langle y, \psi\rangle_n \psi(\mathbf{x})$. For similar reasons, $\hat{y}_{\mathrm{t}_L}(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) + \hat{y}_{\mathrm{t}_R}(\mathbf{x})\mathbb{1}(\mathbf{x} \in \mathrm{t}_R) = \sum_{\psi\in\mathbf{\Psi}_\mathrm{t}\cup\mathbf{\Psi}_\mathrm{t}^\perp} \langle y, \psi\rangle_n \psi(\mathbf{x})$.

To prove the identity in (16) (and, as a special case, (4)), using the above expansions, observe that for each internal node t,

$$\sum_{\psi\in\mathbf{\Psi}_\mathrm{t}^\perp} \langle y, \psi\rangle_n \psi(\mathbf{x}) = (\hat{y}_{\mathrm{t}_L}(\mathbf{x}) - \hat{y}_\mathrm{t}(\mathbf{x}))\mathbb{1}(\mathbf{x} \in \mathrm{t}_L) + (\hat{y}_{\mathrm{t}_R}(\mathbf{x}) - \hat{y}_\mathrm{t}(\mathbf{x}))\mathbb{1}(\mathbf{x} \in \mathrm{t}_R). \tag{18}$$

For each $\mathbf{x} \in \mathcal{X}$, let $\mathrm{t}_0, \mathrm{t}_1, \ldots, \mathrm{t}_{K-1}, \mathrm{t}_K = \mathrm{t}$ be the unique path from the root node t_0 to the terminal node t that contains $\mathbf{x}$. Next, sum (18) over all internal nodes and telescope the successive internal node outputs to obtain

$$\sum_{k=0}^{K-1} (\hat{y}_{\mathrm{t}_{k+1}}(\mathbf{x}) - \hat{y}_{\mathrm{t}_k}(\mathbf{x})) = \hat{y}_{\mathrm{t}_K}(\mathbf{x}) - \hat{y}_{\mathrm{t}_0}(\mathbf{x}) = \hat{y}_\mathrm{t}(\mathbf{x}) - \hat{y}(\mathbf{x}). \tag{19}$$

Combining (18) and (19), we have

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

where we recall that the null node t_0 is an internal node of T. Next, we show that $\{\psi \in \mathbf{\Psi}_\mathrm{t}^\perp : \mathrm{t} \in [T]\}$ is orthonormal. The fact that each ψ has unit norm, $\|\psi\|_n^2 = 1$, is true by definition. If $\psi, \psi' \in \mathbf{\Psi}_\mathrm{t}^\perp$, then by definition, $\langle \psi, \psi'\rangle_n = 0$. Let t and t′ be two distinct internal nodes and suppose $\psi \in \mathbf{\Psi}_\mathrm{t}^\perp$ and $\psi' \in \mathbf{\Psi}_{\mathrm{t}'}^\perp$. If $\mathrm{t} \cap \mathrm{t}' = \emptyset$, then orthogonality between ψ and ψ' is immediate, since $\psi(\mathbf{x}) \cdot \psi'(\mathbf{x}) \equiv 0$. If $\mathrm{t} \cap \mathrm{t}' \neq \emptyset$, 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 (17) (and, as a special case, (5)) can be shown as follows:

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

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

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

which is established by first using the basis expansion for $\hat{y}_\mathrm{t}$ provided in the proof of Lemma 1 and the Cauchy-Schwarz inequality,

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

and then, because $\{\psi : \psi \in \mathbf{\Psi}_\mathrm{t}\}$ is orthonormal, employing Bessel's inequality to obtain $\sum_{\psi \in \mathbf{\Psi}_\mathrm{t}} |\langle y, \psi \rangle_n|^2 \leq n^{-1} \sum_{\mathbf{x}_i \in \mathrm{t}} y_i^2 \leq w(\mathrm{t}) \max_{1 \leq i \leq n} \frac{1}{i}\sum_{\ell=1}^{i} y_\ell^2$. Thus, Q could be taken to equal (or be an almost sure bound on) $\sup_{\mathbf{x} \in \mathcal{X}} \max_{\mathrm{t} \in [T]} \sqrt{w(\mathrm{t}) \sum_{\psi \in \mathbf{\Psi}_\mathrm{t}} \psi^2(\mathbf{x})}$. In the conventional case where the tree outputs a constant in each node, $\mathbf{\Psi}_\mathrm{t} = \{w^{-1/2}(\mathrm{t})\mathbb{1}(\mathbf{x} \in \mathrm{t})\}$, and hence $Q = 1$. To ensure that $\widehat{\mu}(T)(\mathbf{x})$ is square-integrable, i.e., $\mathbb{E}[\sup_{\mathbf{x} \in \mathcal{X}} |\widehat{\mu}(T)(\mathbf{x})|^2] < \infty$, we merely need to check that $\mathbb{E}[\max_{1 \leq i \leq n} \frac{1}{i}\sum_{\ell=1}^{i} y_i^2] < \infty$. This follows easily from Doob's maximal inequality for positive sub-martingales (Durrett, 2019, Theorem 5.4.4), since $\mathbb{E}[y^2 \log(1 + |y|)] < \infty$ by assumption.

Proof of Lemmas 2 and 3. Define the excess training error as

$$R_K = \|y - \widehat{\mu}(T_K)\|_n^2 - \|y - g\|_n^2.$$

Define the squared node-wise norm and node-wise inner product as $\|f\|_\mathrm{t}^2 = \frac{1}{n(\mathrm{t})}\sum_{\mathbf{x}_i \in \mathrm{t}} (f(\mathbf{x}_i))^2$ and $\langle f, g \rangle_\mathrm{t} = \frac{1}{n(\mathrm{t})}\sum_{\mathbf{x}_i \in \mathrm{t}} f(\mathbf{x}_i) g(\mathbf{x}_i)$, respectively. We define the node-wide excess training error as

$$R_K(\mathrm{t}) = \|y - \hat{y}_\mathrm{t}\|_\mathrm{t}^2 - \|y - g\|_\mathrm{t}^2.$$

We use this to rewrite the total excess training error as a weighted combination of the node-wide excess train errors:

$$R_K = \sum_{\mathrm{t} \in T_K} w(\mathrm{t}) R_K(\mathrm{t}), \quad w(\mathrm{t}) = n(\mathrm{t})/n,$$

where $t \in T_K$ means t is a terminal node of T_K. From the orthogonal decomposition of the tree, as given in (16), we have

$$\|y - \widehat{\mu}(T_K)\|_n^2 = \|y - \widehat{\mu}(T_{K-1})\|_n^2 - \sum_{t \in T_{K-1}} \sum_{\psi \in \mathbf{\Psi}_t^\perp} |\langle y, \psi \rangle_n|^2. \tag{20}$$

Subtracting $\|y - g\|_n^2$ on both sides of (20), and using the definition of R_K, we obtain

$$R_K = R_{K-1} - \sum_{t \in T_{K-1}} \sum_{\psi \in \mathbf{\Psi}_t^\perp} |\langle y, \psi \rangle_n|^2. \tag{21}$$

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 $\{\mathcal{A}_t : t \in [T_K]\}$, conditional on the data. We can assume $\mathbb{E}[R_K] > 0$ for all $K \geq 1$, since otherwise, by definition of R_K,

$$\mathbb{E}[R_K] = \mathbb{E}[\|y - \widehat{\mu}(T_K)\|_n^2 - \|y - g\|_n^2] \leq 0,$$

which directly gives the desired result.

Using the law of iterated expectations and the recursive relationship obtained in (21),

$$\mathbb{E}_{T_K}[R_K] = \mathbb{E}_{T_{K-1}}[\mathbb{E}_{T_K|T_{K-1}}[R_K]] = \mathbb{E}_{T_{K-1}}[R_{K-1}] - \mathbb{E}_{T_{K-1}}\left[\mathbb{E}_{T_K|T_{K-1}}\left[\sum_{t \in T_{K-1}} \sum_{\psi \in \mathbf{\Psi}_t^\perp} |\langle y, \psi \rangle_n|^2\right]\right]. \tag{22}$$

By (17) and the sub-optimality probability, $P_{\mathcal{A}(t)}(\kappa)$, we can rewrite the term inside the iterated expectation in (22) as

$$\begin{aligned}
\sum_{t \in T_{K-1}} \sum_{\psi \in \mathbf{\Psi}_t^\perp} |\langle y, \psi \rangle_n|^2 &= \sum_{t \in T_{K-1}} \widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, t) \\
&\geq \sum_{t \in T_{K-1}} \mathbb{1}\Big(\widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, t) \geq \kappa \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t)\Big) \widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, t) \\
&\geq \kappa \sum_{t \in T_{K-1}} \mathbb{1}\Big(\widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, t) \geq \kappa \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t)\Big) \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t).
\end{aligned} \tag{23}$$

Taking expectations of both sides of (23) with respect to the conditional distribution of T_K given T_{K-1}, we have

$$\begin{aligned}
&\mathbb{E}_{T_K|T_{K-1}}\left[\sum_{t \in T_{K-1}} \sum_{\psi \in \mathbf{\Psi}_t^\perp} |\langle y, \psi \rangle_n|^2\right] \\
&\geq \kappa \sum_{t \in T_{K-1}} \mathbb{E}_{T_K|T_{K-1}}\Big[\mathbb{1}\Big(\widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, t) \geq \kappa \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t)\Big) \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t)\Big].
\end{aligned} \tag{24}$$

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

$$\begin{aligned}
&\sum_{t \in T_{K-1}} \mathbb{E}_{T_K|T_{K-1}}\Big[\mathbb{1}\Big(\widehat{\Delta}(\hat{b}, \hat{\mathbf{a}}, t) \geq \kappa \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t)\Big) \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t)\Big] \\
&= \sum_{t \in T_{K-1}} P_{\mathcal{A}_t}(\kappa) \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t) \\
&\geq \sum_{t \in T_{K-1} : R_{K-1}(t) > 0} P_{\mathcal{A}_t}(\kappa) \max_{(b,\mathbf{a}) \in \mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, t).
\end{aligned} \tag{25}$$

In turn, by Lemma 4,

$$\sum_{t\in T_{K-1}:R_{K-1}(t)>0} P_{\mathcal{A}_t}(\kappa) \max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b,\mathbf{a},t) \geq \sum_{t\in T_{K-1}:R_{K-1}(t)>0} w(t)\frac{R_{K-1}^2(t)}{P_{\mathcal{A}_t}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(t)}^2}, \tag{26}$$

and Lemma 6,

$$\begin{aligned}\sum_{t\in T_{K-1}:R_{K-1}(t)>0} w(t)\frac{R_{K-1}^2(t)}{P_{\mathcal{A}_t}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(t)}^2} &\geq \frac{(\sum_{t\in T_{K-1}:R_{K-1}(t)>0} w(t)R_{K-1}(t))^2}{\sum_{t\in T_{K-1}:R_{K-1}(t)>0} w(t)P_{\mathcal{A}_t}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(t)}^2} \\ &\geq \frac{(R_{K-1}^+)^2}{\sum_{t\in T_{K-1}} w(t)P_{\mathcal{A}_t}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(t)}^2},\end{aligned} \tag{27}$$

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 (24), (25), (26), and (27) and plugging the result into (22), 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 Lemma 6 again, we have

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

We have therefore derived the recursion

$$\mathbb{E}_{T_K}[R_K] \leq \mathbb{E}_{T_{K-1}}[R_{K-1}] - \kappa\frac{(\mathbb{E}_{T_{K-1}}[R_{K-1}^+])^2}{\mathbb{E}_{T_{K-1}}\left[\sum_{t\in T_{K-1}} w(t)P_{\mathcal{A}_t}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(t)}^2\right]}. \tag{28}$$

Next, let us take the expectation of both sides of (28) with respect to the data, apply Lemma 6 once again, and use the fact that $R_{K-1}^+ \geq R_{K-1}$ and $\mathbb{E}[R_{K-1}] > 0$ to obtain

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

We have therefore obtained a recursion for $\mathbb{E}[R_K]$, which we can now solve thanks to Lemma 5. 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 5, we have

$$\mathbb{E}[R_K] \leq \frac{1}{\kappa\sum_{k=1}^K 1/\mathbb{E}[\sum_{t\in T_{k-1}} w(t)P_{\mathcal{A}_t}^{-1}(\kappa)\|g\|_{\mathcal{L}_1(t)}^2]}. \tag{29}$$

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

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

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

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

Plugging this bound into (29), we obtain the desired inequality in (7) on the expected excess training error, namely,

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

For Lemma 3: If we grant Assumptions 2 and 3, and take $g = \mu \in \mathcal{G}$, we can arrive at a stronger bound. Recall that we also assume that $\kappa = 1$ and $P_{\mathcal{A}_\mathrm{t}}(\kappa) = 1$. Since $q > 2$, by two successive applications of Hölder's inequality, we have

$$\sum_{\mathrm{t}\in T_{K-1}} w(\mathrm{t})\|\mu\|^2_{\mathcal{L}_1(\mathrm{t})} \le \Big(\sum_{\mathrm{t}\in T_{K-1}} (w(\mathrm{t}))^{q/(q-2)}\Big)^{1-2/q}\Big(\sum_{\mathrm{t}\in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(\mathrm{t})}\Big)^{2/q}, \tag{30}$$

and

$$\begin{aligned}&\mathbb{E}\Big[\Big(\sum_{\mathrm{t}\in T_{K-1}} (w(\mathrm{t}))^{q/(q-2)}\Big)^{1-2/q}\Big(\sum_{\mathrm{t}\in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(\mathrm{t})}\Big)^{2/q}\Big]\\ &\qquad\le \Big(\mathbb{E}\Big[\sum_{\mathrm{t}\in T_{K-1}} (w(\mathrm{t}))^{q/(q-2)}\Big]\Big)^{1-2/q}\Big(\mathbb{E}\Big[\sum_{\mathrm{t}\in T_{K-1}} \|\mu\|^q_{\mathcal{L}_1(\mathrm{t})}\Big]\Big)^{2/q}.\end{aligned} \tag{31}$$

Combining the two inequalities (30) and (31), we obtain

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

Assumptions 2 and 3 provide further upper bounds, since

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

Plugging this bound into (29), we obtain the desired inequality (13) on the expected excess training error, namely, $\mathbb{E}[R_K]\leq\frac{AV^2}{4^{(K-1)/q}}$. ■

Proof of Theorems 1 and 2. We begin by splitting the MSE (averaging only with respect to the joint distribution of $\{\mathcal{A}_t : t\in[T_k]\}$) into two terms, $\mathbb{E}_{T_k}[\|\mu-\widehat{\mu}(T_k)\|^2]=E_1+E_2$, where

$$\begin{aligned}
E_1&=\mathbb{E}_{T_k}[\|\mu-\widehat{\mu}(T_k)\|^2]-2(\mathbb{E}_{T_K}[\|y-\widehat{\mu}(T_k)\|_n^2]-\|y-\mu\|_n^2)-\alpha(n,k)-\beta(n)\\
E_2&=2(\mathbb{E}_{T_k}[\|y-\widehat{\mu}(T_k)\|_n^2]-\|y-\mu\|_n^2)+\alpha(n,k)+\beta(n),
\end{aligned}$$

and where $\alpha(n,k)$ and $\beta(n)$ are positive sequences that will be specified later.

To bound $\mathbb{E}[E_1]$, we split our analysis into two cases based on the observed data y_i. Accordingly, we have

$$\mathbb{E}[E_1]=\mathbb{E}[E_1\mathbb{1}(\forall i:|y_i|\leq B)]+\mathbb{E}[E_1\mathbb{1}(\exists i:|y_i|>B)],\quad B\geq 0. \tag{32}$$

A.0.1 Bounded Term

We start by looking at the first term on the right hand side of (32).

Proceeding, we introduce a few useful concepts and definitions for studying data-dependent partitions, due to Nobel (1996). Let

$$\Lambda_{n,k}=\{\mathcal{P}(\{(\tilde{y}_1,\tilde{\mathbf{x}}_1^{\mathrm{T}}),\ldots,(\tilde{y}_n,\tilde{\mathbf{x}}_n^{\mathrm{T}})\}):(\tilde{y}_i,\tilde{\mathbf{x}}_i^{\mathrm{T}})\in\mathbb{R}^{1+p}\}$$

be the family of all achievable partitions $\mathcal{P}$ by growing a depth k oblique decision tree on n data points with split boundaries of the form $\mathbf{x}^{\mathrm{T}}\mathbf{a}=b$, where $\|\mathbf{a}\|_{\ell_0}\leq d$. In particular, note that $\Lambda_{n,k}$ contains all data-dependent partitions. We also define

$$M(\Lambda_{n,k})=\max\{|\mathcal{P}|:\mathcal{P}\in\Lambda_{n,k}\}$$

to be the maximum number of terminal nodes among all partitions in $\Lambda_{n,k}$. Note that $M(\Lambda_{n,k})\leq 2^k$ (this statement does not rely on the specific algorithm used to grow a depth k oblique tree, as long as the tree generates a partition of $\mathcal{X}$ at each level). Given a set $\mathbf{z}^n=\{\mathbf{z}_1,\mathbf{z}_2,\ldots,\mathbf{z}_n\}\subset\mathbb{R}^p$, define $\Gamma(\mathbf{z}^n,\Lambda_{n,k})$ to be the number of distinct partitions of $\mathbf{z}^n$ induced by elements of $\Lambda_{n,k}$, that is, the

number of different partitions $\{\mathbf{z}^n \cap A : A \in \mathcal{P}\}$, for $\mathcal{P} \in \Lambda_{n,k}$. The partitioning number $\Gamma_{n,k}(\Lambda_{n,k})$ is defined by

$$\Gamma_{n,k}(\Lambda_{n,k}) = \max\{\Gamma(\mathbf{z}^n, \Lambda_{n,k}) : \mathbf{z}_1, \mathbf{z}_2, \ldots, \mathbf{z}_n \in \mathbb{R}^p\},$$

i.e., the maximum number of different partitions of any n point set that can be induced by members of $\Lambda_{n,k}$. Finally, let $\mathcal{F}_{n,k}(R)$ denote the collection of all functions (bounded by R) that output an element of span($\mathcal{H}$) on each region from a partition $\mathcal{P} \in \Lambda_{n,k}$.

We can deduce that the partitioning number is bounded by

$$\Gamma_{n,k}(\Lambda_{n,k}) \leq \left(\binom{p}{d} n^d\right)^{2^k} \leq \left(\left(\frac{ep}{d}\right)^d n^d\right)^{2^k} = \left(\frac{enp}{d}\right)^{d2^k}.$$

The bound on $\Gamma_{n,k}$ follows from the maximum number of ways in which n data points can be split by a hyperplane in d dimensions. The $\binom{p}{d}$ factor accounts for the number of ways in which a d-dimensional hyperplane can be constructed in a p-dimensional space. Note that this bound is not derived from the specific algorithm used to select the splitting hyperplanes; it is purely combinatorial. Following the calculations in Györfi et al. (2002, p. 240) and modifying them slightly with Györfi et al. (2002, Lemma 13.1, Theorem 9.4), we have the following bound for the covering number $\mathcal{N}(r, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n}))$ of $\mathcal{F}_{n,k}(R)$ by balls of radius $r > 0$ in $\mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})$ with respect to the empirical discrete measure $\mathbb{P}_{\mathbf{x}^n}$ on $\mathbf{x}^n = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\} \subset \mathbb{R}^p$:

$$\begin{aligned}
\mathcal{N}\left(\frac{\beta(n)}{40R}, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})\right) &\leq \Gamma_{n,k}(\Lambda_{n,k})\left(3\left(\frac{6eR}{\frac{\beta(n)}{40R}}\right)^{2\mathrm{VC}(\mathcal{H})}\right)^{2^k} \\
&\leq \left(\left(\frac{enp}{d}\right)^d\right)^{2^k}\left(3\left(\frac{240eR^2}{\beta(n)}\right)^{2\mathrm{VC}(\mathcal{H})}\right)^{2^k} \\
&= \left(3\left(\frac{enp}{d}\right)^d\right)^{2^k}\left(\frac{240eR^2}{\beta(n)}\right)^{\mathrm{VC}(\mathcal{H})2^{k+1}},
\end{aligned} \tag{33}$$

where we use VC($\mathcal{H}$) to denote the VC dimension of span($\mathcal{H}$). According to (9), we know that the regression function is uniformly bounded, $\|\mu\|_\infty \leq M'$. Let $R = QB$. We assume, without loss of generality, that $R \geq M'$ so that $\|\mu\|_\infty \leq R$ and $\|\widehat{\mu}(T_k)\|_\infty \leq R$ almost surely, if $\max_{1\leq i\leq n} |y_i| \leq B$. By Györfi et al. (2002, Theorem 11.4), with $\varepsilon = 1/2$ (in their notation),

$$\begin{aligned}
&\mathbb{P}\Big(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \geq 2(\|y - f\|_n^2 - \|y - \mu\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B\Big) \\
&\leq 14 \sup_{\mathbf{x}^n} \mathcal{N}\left(\frac{\beta(n)}{40R}, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})\right) \exp\left(-\frac{\alpha(n,k)n}{2568R^4}\right).
\end{aligned}$$

Then, we have the following probability concentration

$$\begin{aligned}
&\mathbb{P}(\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|] \geq 2(\mathbb{E}_{T_K}[\|y - \widehat{\mu}(T_k)\|_n^2] - \|y - \mu\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B) \\
&\leq 14 \sup_{\mathbf{x}^n} \mathcal{N}\left(\frac{\beta(n)}{40R}, \mathcal{F}_{n,k}(R), \mathscr{L}_1(\mathbb{P}_{\mathbf{x}^n})\right) \exp\left(-\frac{\alpha(n,k)n}{2568R^4}\right).
\end{aligned} \tag{34}$$

This inequality follows from the fact that, on the event $\{\forall i : |y_i| \leq B\}$, if

$$\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|^2 - 2\|y - \widehat{\mu}(T_k)\|_n^2] \geq -2\|y - \mu\|_n^2 + \alpha(n,k) + \beta(n)$$

holds, then there exists a realization $\widehat{\mu}(T_k') \in \mathcal{F}_{n,k}(R)$ such that

$$\|\mu - \widehat{\mu}(T_k')\|^2 - 2\|y - \widehat{\mu}(T_k')\|_n^2 \geq -2\|y - \mu\|_n^2 + \alpha(n,k) + \beta(n),$$

and hence

$$\begin{aligned}\mathbb{P}\Big(\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|] \geq 2(\mathbb{E}_{T_K}[\|y - \widehat{\mu}(T_k)\|_n^2] - \|y - \mu\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B\Big) \\ \leq \mathbb{P}\Big(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \geq 2(\|y - f\|_n^2 - \|y - \mu\|_n^2) + \alpha(n,k) + \beta(n),\ \forall i : |y_i| \leq B\Big).\end{aligned}$$

We can now plug in the result of (33) into (34) to obtain

$$\mathbb{P}(E_1 \geq 0,\ \forall i : |y_i| \leq B) \leq 14\left(3\left(\frac{enp}{d}\right)^d\right)^{2^k}\left(\frac{240eR^2}{\beta(n)}\right)^{\mathrm{VC}(\mathcal{H})2^{k+1}} \exp\left(-\frac{\alpha(n,k)n}{2568R^4}\right). \tag{35}$$

We choose

$$\begin{aligned}\alpha(n,k) &= \frac{2568R^4\Big(2^k d\log(enp/d) + 2^k\log(3) + \mathrm{VC}(\mathcal{H})2^{k+1}\log(\frac{240eR^2}{\beta(n)}) + \log(14n^2)\Big)}{n} \\ \beta(n) &= \frac{240eR^2}{n^2}\end{aligned}$$

so that $\mathbb{P}(E_1 \geq 0,\ \forall i : |y_i| \leq B) \leq 1/n^2$. Thus,

$$E_1\mathbb{1}(\forall i : |y_i| \leq B) \leq (\mathbb{E}_{T_K}[\|\mu - \widehat{\mu}(T_k)\|^2] + 2\|y - \mu\|_n^2)\mathbb{1}(\forall i : |y_i| \leq B) \leq 12R^2,$$

and so we have

$$\mathbb{E}[E_1\mathbb{1}(\forall i : |y_i| \leq B)] \leq 12R^2\mathbb{P}(E_1 \geq 0,\ \forall i : |y_i| \leq B) \leq \frac{12R^2}{n^2} = \frac{12Q^2B^2}{n^2}. \tag{36}$$

A.0.2 Unbounded Term

We now look at the second term on the right hand side of (32). Because we have $\|\widehat{\mu}(T_k)\|_\infty \leq Q \cdot \sqrt{\max_{1\leq i\leq n}\frac{1}{i}\sum_{\ell=1}^i y_\ell^2}$ almost surely, we can bound

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_k)\|^2\mathbb{1}(\exists i : |y_i| > B)] \leq (Q+1)^2\mathbb{E}\Big[\max_{1\leq i\leq n}\max\{y^2, y_i^2\}\mathbb{1}(\exists i : |y_i| > B)\Big].$$

Using the fact that the sum of non-negative variables upper bounds their maximum, and the exponential concentration of the conditional distribution of y given $\mathbf{x}$ (Assumption 1) together with a union bound, we can then apply Cauchy-Schwarz to obtain

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_k)\|^2\mathbb{1}(\exists i : |y_i| > B)] \leq (Q+1)^2\sqrt{(n+1)\mathbb{E}[y^4]}\sqrt{nc_1\exp(-c_2(B-M)^\gamma)}.$$

Setting $B = B_n = M + ((6/c_2)\log(n+1))^{1/\gamma} \geq M'$, we have that

$$\mathbb{E}[\|\mu - \widehat{\mu}(T_k)\|^2\mathbb{1}(\exists i : |y_i| > B)] \leq \frac{(Q+1)^2\sqrt{c_1\mathbb{E}[y^4]}}{n^2}. \tag{37}$$

Thus combining (36) and (37), we have

$$\begin{aligned}\mathbb{E}[E_1] &= \mathbb{E}[E_1 \mathbb{1}(\forall i : |y_i| \le B)] + \mathbb{E}[E_1 \mathbb{1}(\exists i : |y_i| > B)] \\ &\le \frac{12Q^2B^2}{n^2} + \frac{(Q+1)^2\sqrt{c_1\mathbb{E}[y^4]}}{n^2} = O\left(\frac{\log^{2/\gamma}(n)}{n^2}\right).\end{aligned} \tag{38}$$

Next, we turn our attention to $\mathbb{E}[E_2]$. Since

$$\mathbb{E}[\|y - \widehat{\mu}(T_k)\|_n^2 - \|y - \mu\|_n^2] = \|\mu - g\|^2 + \mathbb{E}[\|y - \widehat{\mu}(T_k)\|_n^2 - \|y - g\|_n^2],$$

it follows that

$$\mathbb{E}[E_2] = 2\|\mu - g\|^2 + 2\mathbb{E}[\|y - \widehat{\mu}(T_k)\|_n^2 - \|y - g\|_n^2] + \alpha(n,k) + \beta(n). \tag{39}$$

Finally, combining the bounds (38) and (39) and simplifying $\alpha(n,k)$ and $\beta(n)$,

$$\begin{aligned}&\mathbb{E}[\|\mu - \widehat{\mu}(T_K)\|^2] \\ &\quad\le 2\|\mu - g\|^2 + 2\mathbb{E}[\|y - \widehat{\mu}(T_K)\|_n^2 - \|y - g\|_n^2] + C\frac{2^K(d + \mathrm{VC}(\mathcal{H}))\log(np/d)\log^{4/\gamma}(n)}{n},\end{aligned} \tag{40}$$

for some positive constant $C = C(c_1, c_2, \gamma, M, Q)$.

A.0.3 Pruned Tree

We now consider the pruned tree, T_{opt}. Let $\mathbb{E}_{T_{\mathrm{opt}}}[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2] = E_1' + E_2'$, where

$$\begin{aligned}E_1' &= \mathbb{E}_{T_{\mathrm{opt}}}[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2] - 2(\mathbb{E}_{T_{\mathrm{opt}}}[\|y - \widehat{\mu}(T_{\mathrm{opt}})\|_n^2] - \|y - \mu\|_n^2) - 2\lambda|T_{\mathrm{opt}}| \\ E_2' &= 2(\mathbb{E}_{T_{\mathrm{opt}}}[\|y - \widehat{\mu}(T_{\mathrm{opt}})\|_n^2] - \|y - \mu\|_n^2) + 2\lambda|T_{\mathrm{opt}}|.\end{aligned}$$

Note that, for each $k = 1, 2, \ldots, n-1$,

$$\|y - \widehat{\mu}(T_{\mathrm{opt}})\|_n^2 + \lambda|T_{\mathrm{opt}}| \le \|y - \widehat{\mu}(T_k)\|_n^2 + \lambda 2^k,$$

and hence, for each $k \ge 1$,

$$\mathbb{E}[E_2'] \le 2\|\mu - g\|^2 + 2\mathbb{E}[\|y - \widehat{\mu}(T_k)\|_n^2 - \|y - g\|_n^2] + \lambda 2^{k+1}. \tag{41}$$

Choose $\lambda = \lambda_n$ such that $\alpha(n,k) + \beta(n) \le \lambda_n 2^{k+1}$. This implies that $\lambda_n \gtrsim \frac{(d+\mathrm{VC}(\mathcal{H}))\log(np/d)\log^{4/\gamma}(n)}{n}$. For each realization of T_{opt}, there exists k such that $|T_{\mathrm{opt}}| \ge 2^k$. By a union bound and the result established in (35), we have

$$\begin{aligned}P(E_1' \ge 0) &\le \mathbb{P}(\mathbb{E}_{T_{\mathrm{opt}}}[\|\mu - \widehat{\mu}(T_{\mathrm{opt}})\|^2] \ge 2(\mathbb{E}_{T_{\mathrm{opt}}}[\|y - \widehat{\mu}(T_{\mathrm{opt}})\|_n^2] - \|y - \mu\|_n^2) + 2\lambda_n|T_{\mathrm{opt}}|) \\ &\le \sum_{1\le k\le n-1} \mathbb{P}(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \ge 2(\|y - f\|_n^2 - \|y - \mu\|_n^2) + \lambda_n 2^{k+1}) \\ &\le \sum_{1\le k\le n-1} \mathbb{P}(\exists f \in \mathcal{F}_{n,k}(R) : \|\mu - f\|^2 \ge 2(\|y - f\|_n^2 - \|y - \mu\|_n^2) + \alpha(n,k) + \beta(n)) \\ &\le \sum_{1\le k\le n-1} n^{-2} \le 1/n.\end{aligned}$$

Once again, we split the expectation, $\mathbb{E}[E_1']$ into two cases, as in (32), and bound each case separately. The argument is identical to that for the un-pruned tree so we omit details here. Combining this bound on $\mathbb{E}[E_1']$ with (41) gives as an analogous result to (40), namely, for all $K \geq 1$,

$$\begin{aligned}&\mathbb{E}[\|\mu - \widehat{\mu}(T_{\text{opt}})\|^2] \\ &\quad \leq 2\|\mu - g\|^2 + 2\mathbb{E}[\|y - \widehat{\mu}(T_K)\|_n^2 - \|y - g\|_n^2] + C\frac{2^K(p + \text{VC}(\mathcal{H}))\log^{1+4/\gamma}(n)}{n},\end{aligned} \tag{42}$$

for some positive constant $C = C(c_1, c_2, \gamma, M, Q)$.

The next part of the proof entails bounding $\mathbb{E}[\|y-\widehat{\mu}(T_K)\|_n^2 - \|y-g\|_n^2]$, depending on the assumptions we make. Note that for the constant output $\hat{y}_t(\mathbf{x}) \equiv \bar{y}_t$, we have $Q = 1$ and $\text{VC}(\mathcal{H}) = 1$.

For Theorem 1: We bound $\mathbb{E}[\|y-\widehat{\mu}(T_K)\|_n^2 - \|y-g\|_n^2]$ using Lemma 2. The inequality (10) follows directly from (40) and the inequality (11) follows directly from (42).

For Theorem 2: Taking $g = \mu \in \mathcal{G}$ and $d = p$, we bound $\mathbb{E}[\|y-\widehat{\mu}(T_K)\|_n^2 - \|y-g\|_n^2]$ using Lemma 3. The inequality (14) follows directly from (40). To show (15), we use (14) and

$$\begin{aligned}&\inf_{K\geq 1}\left\{\frac{2AV^2}{4^{(K-1)/q}} + C\frac{2^{K+1}p\log^{4/\gamma+1}(n)}{n}\right\} \\ &\qquad = 2(2+q)\left(\frac{AV^2}{q}\right)^{q/(2+q)}\left(\frac{Cp\log^{4/\gamma+1}(n)}{n}\right)^{2/(2+q)}.\end{aligned}$$

This completes the proof of both Theorem 1 and Theorem 2. ∎

B Technical Lemmas

In this section, we present some technical lemmas that aid in the proof of our main results and may also be of independent interest.

B.1 Impurity Bound

Our first lemma establishes an important connection between the decrease in impurity and the empirical node-wide excess risk.

Lemma 4 (Impurity bound)
Define $R_{K-1}(\text{t}) = \|y - \hat{y}_\text{t}\|_\text{t}^2 - \|y - g\|_\text{t}^2$. *Let* t *be a terminal node of* T_{K-1}, *and assume* $R_{K-1}(\text{t}) > 0$. *Then, if* $g \in \mathcal{G}$,

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

Proof of Lemma 4. Assume that $g \in \mathcal{G}$, $g(\mathbf{x}) = \sum_k g_k(\mathbf{x}^\text{T}\mathbf{a}_k)$, is not the constant function (the result is trivial for constant g). 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 re-indexed so that $\mathbf{x}_1^{\mathrm{T}}\mathbf{a}_k \le \mathbf{x}_2^{\mathrm{T}}\mathbf{a}_k \le \cdots \le \mathbf{x}_{n(\mathrm{t})}^{\mathrm{T}}\mathbf{a}_k$, then

$$g_k'(b) = \frac{g_k(\mathbf{x}_{i+1}^{\mathrm{T}}\mathbf{a}_k) - g_k(\mathbf{x}_i^{\mathrm{T}}\mathbf{a}_k)}{\mathbf{x}_{i+1}^{\mathrm{T}}\mathbf{a}_k - \mathbf{x}_i^{\mathrm{T}}\mathbf{a}_k}, \quad \text{for } \mathbf{x}_i^{\mathrm{T}}\mathbf{a}_k \le b < \mathbf{x}_{i+1}^{\mathrm{T}}\mathbf{a}_k \quad \text{and } i = 1, \ldots, n(\mathrm{t}) - 1. \tag{43}$$

Let

$$\frac{\mathrm{d}\Pi(b, \mathbf{a}_k)}{\mathrm{d}(b, \mathbf{a}_k)} = \frac{|g_k'(b)|\sqrt{\mathbb{P}(\mathrm{t}_L)\mathbb{P}(\mathrm{t}_R)}}{\sum_{k'}\int |g_{k'}'(b')|\sqrt{\mathbb{P}(\mathrm{t}_L')\mathbb{P}(\mathrm{t}_R')}\mathrm{d}b'} \tag{44}$$

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 $\mathbf{x}^{\mathrm{T}}\mathbf{a} = b$. 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 node-wise 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 (17) and the fact that the decision stump ψ_{t} (see (3)) belongs to $\mathbf{\Psi}_{\mathrm{t}}^{\perp}$, we have

$$\max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, \mathrm{t}) \ge \int \sum_{\psi\in\mathbf{\Psi}_{\mathrm{t}}^{\perp}} |\langle y, \psi\rangle_n|^2 \mathrm{d}\Pi(b, \mathbf{a}_k) \ge \int |\langle y, \psi_{\mathrm{t}}\rangle_n|^2 \mathrm{d}\Pi(b, \mathbf{a}_k). \tag{45}$$

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

$$\int |\langle y, \psi_{\mathrm{t}}\rangle_n|^2 \mathrm{d}\Pi(b, \mathbf{a}_k) = w(\mathrm{t}) \int |\langle y, \tilde{\psi}_{\mathrm{t}}\rangle_{\mathrm{t}}|^2 \mathrm{d}\Pi(b, \mathbf{a}_k) \ge w(\mathrm{t}) \left(\int |\langle y, \tilde{\psi}_{\mathrm{t}}\rangle_{\mathrm{t}}| \mathrm{d}\Pi(b, \mathbf{a}_k) \right)^2. \tag{46}$$

Our next task will be to lower bound the expectation $\int |\langle y, \tilde{\psi}_{\mathrm{t}}\rangle_{\mathrm{t}}| \mathrm{d}\Pi(b, \mathbf{a}_k)$. 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

$$\sqrt{\mathbb{P}(\mathrm{t}_L)\mathbb{P}(\mathrm{t}_R)}\langle y, \tilde{\psi}_{\mathrm{t}}\rangle_{\mathrm{t}} = \sqrt{\mathbb{P}(\mathrm{t}_L)\mathbb{P}(\mathrm{t}_R)}\langle y - \hat{y}_{\mathrm{t}}, \tilde{\psi}_{\mathrm{t}}\rangle_{\mathrm{t}} = -\langle y - \hat{y}_{\mathrm{t}}, \mathbb{1}(\mathbf{x}^{\mathrm{T}}\mathbf{a} > b)\rangle_{\mathrm{t}}.$$

Using this identity together with the empirical measure (defined in (44)), we see that the expectation in (46) is lower bounded by

$$\begin{aligned} \int |\langle y - \hat{y}_{\mathrm{t}}, \tilde{\psi}_{\mathrm{t}}\rangle_{\mathrm{t}}| \mathrm{d}\Pi(b, \mathbf{a}_k) &= \frac{\sum_k \int |g_k'(b)| |\langle y - \hat{y}_{\mathrm{t}}, \mathbb{1}(\mathbf{x}^{\mathrm{T}}\mathbf{a}_k > b)\rangle_{\mathrm{t}}| \mathrm{d}b}{\sum_{k'}\int |g_{k'}'(b')|\sqrt{\mathbb{P}(\mathrm{t}_L')\mathbb{P}(\mathrm{t}_R')}\mathrm{d}b'} \\ &\ge \frac{|\langle y - \hat{y}_{\mathrm{t}}, \sum_k \int g_k'(b)\mathbb{1}(\mathbf{x}^{\mathrm{T}}\mathbf{a}_k > b)\mathrm{d}b\rangle_{\mathrm{t}}|}{\sum_{k'}\int |g_{k'}'(b')|\sqrt{\mathbb{P}(\mathrm{t}_L')\mathbb{P}(\mathrm{t}_R')}\mathrm{d}b'}. \end{aligned} \tag{47}$$

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

$$\Big\langle y - \hat{y}_{\mathrm{t}}, \sum_k \int g'_k(b)\mathbb{1}(\mathbf{x}^{\mathrm{T}}\mathbf{a}_k > b)\mathrm{d}b \Big\rangle_{\mathrm{t}} = \langle y - \hat{y}_{\mathrm{t}}, g\rangle_{\mathrm{t}}. \tag{48}$$

In light of (45), (46), (47), and (48), we obtain

$$\max_{(b,\mathbf{a})\in\mathbb{R}^{1+p}} \widehat{\Delta}(b, \mathbf{a}, \mathrm{t}) \geq \frac{w(\mathrm{t})|\langle y - \hat{y}_{\mathrm{t}}, g\rangle_{\mathrm{t}}|^2}{(\sum_{k'} \int |g'_{k'}(b')| \sqrt{\mathbb{P}(\mathrm{t}'_L)\mathbb{P}(\mathrm{t}'_R)}\mathrm{d}b')^2}. \tag{49}$$

Next, we derive upper and lower bounds on the denominator and numerator of (49), respectively. First, we look at the denominator. Note that for each k', the integral can be decomposed as follows:

$$\int |g'_{k'}(b')| \sqrt{\mathbb{P}(\mathrm{t}'_L)\mathbb{P}(\mathrm{t}'_R)}\mathrm{d}b' = \sum_{i=1}^{n(\mathrm{t})-1} \int_{\{b':n(\mathrm{t}'_L)=i\}} |g'_{k'}(b')| \sqrt{(i/n(\mathrm{t}))(1 - i/n(\mathrm{t}))}\mathrm{d}b'. \tag{50}$$

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 (50) can be explicitly identified from the definition of $g'_{k'}$ in (43),

$$\int |g'_{k'}(b')| \sqrt{\mathbb{P}(\mathrm{t}'_L)\mathbb{P}(\mathrm{t}'_R)}\mathrm{d}b' \leq \frac{1}{2} \sum_{i=1}^{n(\mathrm{t})-1} \int_{\{b':n(\mathrm{t}'_L)=i\}} |g'_{k'}(b')|\mathrm{d}b' = \frac{1}{2} \sum_{i=1}^{n(\mathrm{t})-1} \int_{\mathbf{x}_i^{\mathrm{T}}\mathbf{a}_{k'}}^{\mathbf{x}_{i+1}^{\mathrm{T}}\mathbf{a}_{k'}} |g'_{k'}(b')|\mathrm{d}b'. \tag{51}$$

By the definition of $g'_{k'}$ as a divided difference (43) and the definition of total variation, for each k',

$$\sum_{i=1}^{n(\mathrm{t})-1} \int_{\mathbf{x}_i^{\mathrm{T}}\mathbf{a}_{k'}}^{\mathbf{x}_{i+1}^{\mathrm{T}}\mathbf{a}_{k'}} |g'_{k'}(b')|\mathrm{d}b' = \sum_{i=1}^{n(\mathrm{t})-1} |g_{k'}(\mathbf{x}_{i+1}^{\mathrm{T}}\mathbf{a}_{k'}) - g_{k'}(\mathbf{x}_i^{\mathrm{T}}\mathbf{a}_{k'})| \leq V(g_{k'}, \mathbf{a}_{k'}, \mathrm{t}). \tag{52}$$

Combining (51) and (52) and plugging the result into the summation in the denominator of (49), we get

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

Next, we lower bound the numerator in (49). Using the Cauchy-Schwarz inequality and the fact that $\langle y - \hat{y}_{\mathrm{t}}, y\rangle_{\mathrm{t}} = \|y - \hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2$, we obtain

$$\langle y - \hat{y}_{\mathrm{t}}, g\rangle_{\mathrm{t}} = \langle y - \hat{y}_{\mathrm{t}}, y\rangle_{\mathrm{t}} - \langle y - \hat{y}_{\mathrm{t}}, y - g\rangle_{\mathrm{t}} \geq \|y - \hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2 - \|y - \hat{y}_{\mathrm{t}}\|_{\mathrm{t}}\|y - g\|_{\mathrm{t}}. \tag{53}$$

By the AM-GM inequality, we know that $\|y - \hat{y}_{\mathrm{t}}\|_{\mathrm{t}}\|y - g\|_{\mathrm{t}} \leq \frac{1}{2}(\|y - \hat{y}_{\mathrm{t}}\|_{\mathrm{t}}^2 + \|y - g\|_{\mathrm{t}}^2)$. Plugging this into (53), we get

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

Now, squaring both sides and using the assumption that $R_{K-1}(\mathrm{t}) > 0$, we have

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

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

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

■

B.2 Recursive Inequality

Here we provide a solution to a simple recursive inequality.

Lemma 5
Let $\{a_k\}$ be a decreasing sequence of numbers and $\{b_k\}$ be a positive sequence numbers satisfying the following recursive expression:

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

Then,

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

Proof of Lemma 5. We may assume without loss of generality that $a_{K-1} > 0$; otherwise the result holds trivially since $a_K \leq a_{K-1} \leq 0 \leq \frac{1}{\sum_{k=1}^{K} b_k}$. For $K = 1$,

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

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

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

B.3 Sedrakyan's Inequality

For completeness, we reproduce Sedrakyan's inequality (Sedrakyan, 1997) in its generalized form below.

Lemma 6 (Sedrakyan's inequality (Sedrakyan, 1997))
Let U and V be two nonnegative random variables with $V > 0$ almost surely. Then

$$\mathbb{E}\left[\frac{U}{V}\right] \geq \frac{(\mathbb{E}[\sqrt{U}])^2}{\mathbb{E}[V]}.$$

Proof of Lemma 6. By the Cauchy-Schwarz inequality,

$$\mathbb{E}[\sqrt{U}] = \mathbb{E}\left[\sqrt{\frac{U}{V}}\sqrt{V}\right] \leq \sqrt{\mathbb{E}\left[\frac{U}{V}\right]}\sqrt{\mathbb{E}[V]}.$$

Rearranging the above inequality gives the desired result. ■

References

Barron, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Trans. Inform. Theory*, 39(3):930–945.

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

Barron, A. R., Cohen, A., Dahmen, W., and DeVore, R. A. (2008). Approximation and learning by greedy algorithms. *Annals of Statistics*, 36(1):64–94.

Bennett, K. P. (1994). Global tree optimization: A non-greedy decision tree algorithm. *Computing Science and Statistics*, pages 156–156.

Bertsimas, D. and Dunn, J. (2017). Optimal classification trees. *Machine Learning*, 106(7):1039–1082.

Bertsimas, D. and Dunn, J. (2019). *Machine learning under a modern optimization lens*. Dynamic Ideas LLC.

Bertsimas, D., Dunn, J., and Wang, Y. (2021). Near-optimal nonlinear regression trees. *Operations Research Letters*, 49(2):201–206.

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*.

Bertsimas, D. and Stellato, B. (2021). The voice of optimization. *Machine Learning*, 110(2):249–277.

Breiman, L., Friedman, J., Olshen, R., and Stone, C. (1984). *Classification and Regression Trees*. Belmont, Calif.: Wadsworth International Group, c1984.

Brodley, C. E. and Utgoff, P. E. (1995). Multivariate decision trees. *Machine Learning*, 19(1):45–77.

Buciluundefined, C., Caruana, R., and Niculescu-Mizil, A. (2006). Model compression. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '06, pages 535–541, New York, NY, USA. Association for Computing Machinery.

Cattaneo, M. D., Farrell, M. H., and Feng, Y. (2020). Large sample properties of partitioning-based series estimators. *The Annals of Statistics*, 48(3):1718–1741.

Chi, C.-M., Vossler, P., Fan, Y., and Lv, J. (2022). Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*.

Dunn, J. W. (2018). *Optimal trees for prediction and prescription*. PhD thesis, Massachusetts Institute of Technology.

Durrett, R. (2019). *Probability: Theory and Examples*. Cambridge Series in Statistical and Probabilistic Mathematics. Cambridge University Press, 5 edition.

Frosst, N. and Hinton, G. (2017). Distilling a neural network into a soft decision tree. *arXiv preprint arXiv:1711.09784*.

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.

Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression*, volume 1. Springer.

Heath, D., Kasif, S., and Salzberg, S. (1993). Induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2(2):1–32.

Huang, J. Z. (2003). Local asymptotics for polynomial spline regression. *The Annals of Statistics*, 31(5):1600–1635.

Hüllermeier, E., Mohr, F., Tornede, A., and Wever, M. (2021). Automated machine learning, bounded rationality, and rational metareasoning. *arXiv preprint arXiv:2109.04744*.

Klusowski, J. M. and Tian, P. (2022). Large scale prediction with decision trees. *Journal of the American Statistical Association*.

Lee, G.-H. and Jaakkola, T. S. (2020). Oblique decision trees from derivatives of relu networks. In *International Conference on Learning Representations*.

Li, X.-B., Sweigart, J., Teng, J., Donohue, J., Thombs, L., and Wang, S. (2003). Multivariate decision trees using linear discriminants and tabu search. *IEEE Transactions on Systems, Man, and Cybernetics - Part A: Systems and Humans*, 33(2):194–205.

Loh, W.-Y. and Shih, Y.-S. (1997). Split selection methods for classification trees. *Statistica Sinica*, 7(4):815–840.

López-Chau, A., Cervantes, J., López-García, L., and Lamont, F. G. (2013). Fisher's decision tree. *Expert Systems with Applications*, 40(16):6283–6291.

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*, pages 453–469. Springer.

Mingers, J. (1989). An empirical comparison of pruning methods for decision tree induction. *Machine learning*, 4(2):227–243.

Murdoch, W. J., Singh, C., Kumbier, K., Abbasi-Asl, R., and Yu, B. (2019). Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080.

Murthy, S. K., Kasif, S., and Salzberg, S. (1994). A system for induction of oblique decision trees. *Journal of Artificial Intelligence Research*, 2(1):1–32.

Nobel, A. (1996). Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 24(3):1084 – 1105.

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

Rodriguez, J., Kuncheva, L., and Alonso, C. (2006). Rotation forest: A new classifier ensemble method. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(10):1619–1630.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215.

Scornet, E., Biau, G., and Vert, J.-P. (2015). Consistency of random forests. *The Annals of Statistics*, 43(4):1716 – 1741.

Sedrakyan, N. (1997). About the applications of one useful inequality. *Kvant Journal*, 97(2):42–44.

Syrgkanis, V. and Zampetakis, M. (2020). Estimation and inference with trees and forests in high dimensions. In Abernethy, J. and Agarwal, S., editors, *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 3453–3454. PMLR.

Tomita, T. M., Browne, J., Shen, C., Chung, J., Patsolic, J. L., Falk, B., Priebe, C. E., Yim, J., Burns, R., Maggioni, M., and Vogelstein, J. T. (2020). Sparse projection oblique randomer forests. *Journal of Machine Learning Research*, 21(104):1–39.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Yang, Y., Morillo, I. G., and Hospedales, T. M. (2018). Deep neural decision trees. In *ICML Workshop on Human Interpretability in Machine Learning (WHI)*.

Zhang, T. (2003). Sequential greedy approximation for certain convex optimization problems. *IEEE Transactions on Information Theory*, 49(3):682–691.

Zhu, H., Murali, P., Phan, D., Nguyen, L., and Kalagnanam, J. (2020). A scalable mip-based method for learning optimal multivariate decision trees. In Larochelle, H., Ranzato, M., Hadsell, R., Balcan, M., and Lin, H., editors, *Advances in Neural Information Processing Systems*, volume 33, pages 1771–1781. Curran Associates, Inc.