A cautionary tale on fitting decision trees to data from additive models: generalization lower bounds

Yan Shuo Tan University of California, Berkeley
Abhineet Agarwal University of California, Berkeley
Bin Yu University of California, Berkeley

Abstract

Decision trees are important both as interpretable models amenable to high-stakes decision-making, and as building blocks of ensemble methods such as random forests and gradient boosting. Their statistical properties, however, are not well understood. The most cited prior works have focused on deriving pointwise consistency guarantees for CART in a classical nonparametric regression setting. We take a different approach, and advocate studying the generalization performance of decision trees with respect to different generative regression models. This allows us to elicit their *inductive bias*, that is, the assumptions the algorithms make (or do not make) to generalize to new data, thereby guiding practitioners on when and how to apply these methods. In this paper, we focus on sparse additive generative models, which have both low statistical complexity and some nonparametric flexibility. We prove a sharp squared error generalization lower bound for a large class of decision tree algorithms fitted to sparse additive models with C^1 component functions. This bound is surprisingly much worse than the minimax rate for estimating such sparse additive models. The inefficiency is due not to greediness, but to the loss in power for detecting global structure when we average responses solely over each leaf, an observation that suggests opportunities to improve tree-based algorithms, for example, by hierarchical shrinkage. To prove these bounds, we develop new technical machinery, establishing a novel connection between decision tree estimation and rate-distortion theory, a sub-field of information theory.

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.

1 Introduction

Using decision trees for supervised learning has a long and storied history. First introduced by Morgan and Sonquist (1963), the idea is simple: recursively split your covariate space along coordinate directions, and fit a piecewise constant model on the resulting partition. The *adaptivity* of splits to structure in the data improves conciseness and statistical efficiency of tree models. Meanwhile, the *greedy splitting* principle followed by most algorithms, including Breiman et al. (1984)'s Classification and Regression Trees (CART), ensures computational tractability. More recently, there has also been growing interest in fitting optimal decision trees using mathematical programming or dynamic programming techniques (Lin et al., 2020; Aghaei et al., 2021).

Decision tree models are important for two main reasons. First, shallow decision trees are interpretable models (Rudin et al., 2021): They can be implemented by hand, and they are easily described and visualized. While the precise definition and utility of interpretability has been a subject of much debate (Murdoch et al., 2019; Doshi-Velez and Kim, 2017; Rudin, 2019), all agree that it is an important supplement to prediction accuracy in high-stakes decision-making such as medical risk assessment and criminal justice. For this reason, decision trees have been widely applied in both areas (Steadman et al., 2000; Kuppermann et al., 2009; Letham et al., 2015; Angelino et al., 2018). Second, CART trees are used as the basic building blocks of ensemble machine learning algorithms such as random forests (RF) and gradient boosting (Breiman, 2001; Friedman, 2001). These algorithms are recognized as having state-of-the-art performance over a wide class of prediction problems (Caruana and Niculescu-Mizil,

Code and documentation for easily reproducing the results are provided at `https://github.com/aagarwal1996/additive_trees`.

2006; Caruana et al., 2008; Fernández-Delgado et al., 2014; Olson et al., 2018), and receive widespread use, given their implementation in popular machine learning packages such as `ranger` (Wright et al., 2017), `scikit-learn` (Pedregosa et al., 2011) and `xgboost` (Chen and Guestrin, 2016). Random forests in particular have also shown promise in scientific applications, for example in discovering interactions in genomics (Boulesteix et al., 2012; Basu et al., 2018).

Because of the centrality of decision trees in the machine learning edifice, it is all the more surprising that there has been relatively little theory on their statistical properties. In the regression setting, some of the most cited prior works have focused on deriving pointwise *consistency* guarantees for CART when assuming that the conditional mean function is Lipschitz continuous (Biau, 2012; Wager and Athey, 2018). Unfortunately, each is forced to modify the splitting criterion in the algorithm to ensure that the mesh of the learnt partition shrinks to zero. Scornet et al. (2015) proved the first consistency result for the unmodified CART algorithm by replacing the fully nonparametric regression model with an additive regression model (Friedman et al., 2001). This generative assumption simplifies calculations by avoiding some of the complex dependencies between splits that may accumulate during recursive splitting. Moreover, it prevents the existence of locally optimal trees that are not globally optimal, which would otherwise trip up greedy methods such as CART. Klusowski (2020, 2021) has extended this analysis to sparse additive models, showing that when the true conditional mean function depends only on a fixed subset of s covariates, CART is still consistent even when the total number of covariates is allowed to grow exponentially in the sample size. This adaptivity to sparsity somewhat alleviates the curse of dimensionality, and partially explains why CART and RF are often preferred in practice to k-nearest neighbors.

As natural generalizations of linear models, additive models simultaneously have low statistical complexity and yet sufficient nonparametric flexibility required to describe some real world datasets well. Moreover, if the component functions are not too complex, additive models have aspects of interpretability (Rudin et al., 2021). Unsurprisingly, they have accumulated a rich statistical literature (Hastie and Tibshirani, 1986; Sadhanala and Tibshirani, 2019). While the previously discussed works have proved consistency for CART on additive regression models, it is also important to compute *rate upper and lower bounds* for the generalization error of CART and other decision tree algorithms. This would allow us to compare their performance with that of specially tailored algorithms such as backfitting (Breiman and Friedman, 1985), and hence understand whether the inductive biases of decision trees are able to fully exploit the structure present in additive models.

1.1 Main contributions

In this paper, we provide generalization lower bounds for a large class of decision trees, which we call ALA, when fitted to data generated from sparse additive models. We define an ALA tree as one that learns an **a**xis-aligned partition of the covariate space, and makes predictions by averaging the responses over each leaf. We call this second aspect ***l**eaf-only **a**veraging*. In addition, we will assume for analytical reasons that our trees are *honest*, which means that one sample is used to learn the partition, and a separate sample is used to estimate the averages over each leaf (Athey and Imbens, 2016). CART is an example of an ALA tree, and so are most (but not all) decision tree algorithms used in practice. The reason we consider this level of generality is to remove the effect of greediness that has dominated the analysis of CART thus far, and to argue that leaf-only averaging subtly introduces its own inductive bias.

We show that when the true conditional mean function is a sparse additive model with s C^1 univariate component functions, no honest ALA tree, even one that has oracle access to the true conditional mean function, can perform better than $\Omega\left(n^{-\frac{2}{s+2}}\right)$ in expected ℓ_2 risk. This is the ℓ_2 minimax rate for nonparametric estimation of C^1 functions in s dimensions (Stone, 1982). In contrast, if each univariate component function in the model is assumed to be C^1, the minimax rate for sparse additive models scales as $\max\left\{\frac{s\log(d/s)}{n}, \frac{s}{n^{2/3}}\right\}$ (Raskutti et al., 2012). As such, while it is possible to achieve a prescribed error tolerance with $\tilde{O}\left(s^{3/2}\right)$ samples via convex programming, ALA trees have a sample complexity that is at least exponential in s, which means that they needlessly suffer from the curse of dimensionality. The intuitive explanation for this inefficiency is that by ignoring information from other leaves when making a prediction, leaf-only averaging creates an inductive bias *against* global structure.

As far as we know, this paper is the first to establish algorithm-specific lower bounds for CART or any other decision tree algorithm. More broadly, algorithm-specific lower bounds can be challenging in the machine learning literature because they require specialized techniques instead of relying on a general recipe (as is the case with minimax lower bounds). Additionally, we show that the rate lower bound is achievable using an oracle partition. We also obtain a lower bound for additive models over Boolean features, which surprisingly, has a very different form. We note

that Tang et al. (2018) proved sufficient conditions under which honest random forest estimators are inconsistent for special regression functions using Stone (1977)'s adversarial construction. This construction does not produce additive functions. Theirs is the only other work we know of that provides negative results for tree-based estimators. On the other hand, they do not compute lower bounds, and their conditions either involve unrealistic choices of hyperparameters, or pertain to properties of trees after they are grown, such as upper bounds on the rate of shrinkage of leaf diameters. It is not clear if or when these conditions hold in practice.

Our results are obtained using novel technical machinery, which are based on two simple insights: First, we show that the variance term of the expected ℓ_2 risk scales *linearly* with the number of leaf nodes. Second, a tree model can be thought of as a *lossy code*, in which the number of leaves is the *size* of the code, while the bias term of the risk is simply its *distortion*. This link to rate-distortion theory, a sub-field of information theory (Cover and Thomas, 2012), allows us to compute the optimal trade-off between bias and variance to obtain lower bounds. As a happy by-product of our analysis, the first insight yields a better understanding of cost-complexity pruning and minimum impurity decrease procedures that are commonly used with CART to prevent overfitting.

1.2 Other related work

Here, we discuss some other theoretical work on CART that is less directly related to this paper. Syrgkanis and Zampetakis (2020) proved generalization upper bounds for CART in a different setting. They considered Boolean features, and imposed some type of submodularity assumption on the conditional mean function. While this subsumes additive models, the authors did not give concrete examples of other models satisfying this assumption. Scornet (2020) returned to the additive model setting, and was able to compute explicit asymptotic formulas for the popular mean impurity decrease (MDI) feature importance score. Behr et al. (2021) formulated a biologically-inspired discontinuous nonlinear regression model, and showed that CART trees can be used to do inference for the model. We refer the reader to several excellent survey papers for a fuller description of the literature (Loh, 2014; Biau and Scornet, 2016; Hooker and Mentch, 2021).

2 Preliminaries

We work with the standard regression framework in supervised learning, and assume a generative model

$$y = f(\mathbf{x}) + \epsilon \tag{1}$$

where the feature vector $\mathbf{x}$ is drawn from a distribution ν on a subset $\mathcal{X} \subset \mathbb{R}^d$, while the responses y are real-valued, and ϵ is a noise variable that is mean zero when conditioned on $\mathbf{x}$. We assume that the noise is homoskedastic, and denote $\sigma^2 \coloneqq \mathbb{E}\{\epsilon^2 \mid \mathbf{x}\}$. In this paper, $\mathcal{X}$ will either be the unit-length cube $[0,1]^d$ or the hypercube $\{0,1\}^d$. An *additive model* is one in which we can decompose the conditional mean function as the sum of univariate functions along each coordinate direction:

$$f(\mathbf{x}) = \sum_{j=1}^{d} \phi_j(x_j). \tag{2}$$

We are given a training set $\mathcal{D}_n = \{(\mathbf{x}^{(1)}, y^{(1)}), \dots (\mathbf{x}^{(n)}, y^{(n)})\}$ comprising independent samples that are drawn according to the model (1).[1]

A *cell* $\mathcal{C} \subset \mathcal{X}$ is a rectangular subset. If $\mathcal{X} = [0,1]^d$, this means that it can be written as a product of intervals: $\mathcal{C} = [a_1, b_1] \times [a_2, b_2] \times \cdots \times [a_d, b_d]$. If $\mathcal{X} = \{0,1\}^d$, this means that it is a subcube of the form $\mathcal{C}(S, \mathbf{z}) = \{\mathbf{x} \in \{0,1\}^d : x_j = z_j \text{ for } j \in S\}$ where $S \subset [d]$ is a subset of coordinate indices. Given a cell $\mathcal{C}$ and a training set $\mathcal{D}_n$, let $N(\mathcal{C}) \coloneqq \big|\{i : \mathbf{x}^{(i)} \in \mathcal{C}\}\big|$ denote the number of samples in the cell.

A *partition* $\mathfrak{p} = \{\mathcal{C}_1, \dots, \mathcal{C}_j\}$ is a collection of cells with disjoint interiors, whose union is the entire space $\mathcal{X}$. Given the training set $\mathcal{D}_n$, every partition yields an estimator $\hat{f}(-; \mathfrak{p}, \mathcal{D}_n)$ for f via *leaf-only averaging*: For every input $\mathbf{x}$, the estimator outputs the mean response over the cell containing $\mathbf{x}$. In other words, we define

$$\hat{f}(\mathbf{x}; \mathfrak{p}, \mathcal{D}_n) \coloneqq \sum_{\mathcal{C} \in \mathfrak{p}} \Big(\frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} y^{(i)} \Big) \, \mathbf{1}\{\mathbf{x} \in \mathcal{C}\}.$$

We will use the convention that if $N(\mathcal{C}) = 0$, then we set $\frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} y^{(i)} = 0$. We call such an estimator an *ALA tree*.[2]

Note that decision tree algorithms that make non-axis-aligned splits do not yield partitions, though this is not

[1] Sample indices will be denoted using superscripts, while subscripts will be reserved for coordinate indices.

[2] Certain partitions cannot be obtained by recursive binary partitioning. This distinction is not important for our analysis, so we will slightly abuse terminology in calling these estimators trees.

the case for CART and most other algorithms popularly used today. In this definition, we have also kept the partition fixed, whereas decision tree algorithms learn a data-adaptive partition. Having a fixed partition, however, is in keeping with our setting of honest decision trees: We assume that the partition $\mathfrak{p} = \mathfrak{p}(\mathcal{D}'_m)$ has been learnt using a separate dataset $\mathcal{D}'_m$ that we are conditioning on. Furthermore, we note that any lower bounds that hold conditionally on $\mathcal{D}'_m$ will also hold unconditionally.

The *squared error risk*, or *generalization error* of an estimator $\hat{f}$ for f is defined as

$$\mathcal{R}(\hat{f}) \coloneqq \mathbb{E}_{\mathbf{x}\sim\nu}\Big\{\Big(\hat{f}_n(\mathbf{x}) - f(\mathbf{x})\Big)^2\Big\}.$$

We are interested in the smallest possible risk of an ALA tree. To rule out irregularities that happen when some cell $\mathcal{C}$ does not contain any samples from the training set $\mathcal{D}_n$, we need to ensure that the cells are not too small. We say that a partition $\mathfrak{p}$ is *permissible* if for every cell $\mathcal{C} \in \mathfrak{p}$, we have $\nu\{\mathcal{C}\} \geq \frac{1}{n}$. This is a reasonable assumption, as we should expect each cell to contain at least one sample point. Finally, given a conditional mean function f, we define the *oracle expected risk* for ALA trees to be

$$\mathcal{R}^*(f, \nu, n) \coloneqq \inf_{\mathfrak{p}} \mathbb{E}\Big\{R(\hat{f}(-; \mathfrak{p}, \mathcal{D}_n)\Big\} \tag{3}$$

where the infimum is taken over all permissible partitions.

3 A bias-variance risk decomposition for ALA trees

Our main results rely on two key ingredients: A bias-variance decomposition of the expected risk for ALA trees, and a connection to information theory. We state the former as follows.

Theorem 3.1 (Bias-variance decomposition of expected risk)**.** *Assume the regression model* (1)*. Given a permissible partition* $\mathfrak{p}$ *and a training set* $\mathcal{D}_n$*, the expected risk satisfies the following lower and upper bounds:*

$$\mathbb{E}\mathcal{R}(\hat{f}(-; \mathfrak{p}, \mathcal{D}_n)) \geq \sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} + \frac{|\mathfrak{p}|\sigma^2}{2n}, \tag{4}$$

$$\mathbb{E}\mathcal{R}(\hat{f}(-; \mathfrak{p}, \mathcal{D}_n)) \leq 7\sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} + \frac{6|\mathfrak{p}|\sigma^2}{n} + E(\mathfrak{p}), \tag{5}$$

where

$$E(\mathfrak{p}) = \sum_{\mathcal{C}\in\mathfrak{p}} \mathbb{E}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}^2 \left(1 - \nu\{\mathcal{C}\}\right)^n \nu\{\mathcal{C}\}.$$

We make a few remarks about the above theorem. First, we draw attention to its generality: It holds for any conditional mean function and any distribution ν on $\mathcal{X}$, where $\mathcal{X}$ is allowed to be any measurable subset of $\mathbb{R}^d$. In fact, inspecting the proof shows that we do not even require the partition to be axis-aligned.

Next, observe that the lower and upper bounds match up to constant factors and an additive error term E for the upper bound. This term is due to each cell receiving possibly zero samples from the training set, and thus can be made arbitrarily small in comparison with the main terms by further constraining the minimum volume of cells in the partition.

The first main term can be thought of as the approximation error or bias, and has the following equivalent representations:

$$\begin{aligned}\sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} &= \mathbb{E}\{\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\} \\ &= \mathbb{E}\Big\{\big(f(\mathbf{x}) - \bar{f}_{\mathfrak{p}}(\mathbf{x})\big)^2\Big\},\end{aligned}$$

where $\bar{f}_{\mathfrak{p}}$ is the function that takes the value of the conditional mean of f over each cell. In other words, this term is the expected mean square error of the ALA tree if we had infinite data.

The second main term is the contribution from variance, and can be traced to using empirical averages over each cell to estimate the conditional means. The form of this term is striking: It scales linearly with the size of the partition, in direct analogy with the penalty term in cost-complexity pruning (Friedman et al., 2001). Furthermore, it precisely quantifies the trade off between bias and variance when splitting a cell $\mathcal{C}$ in the partition into two children $\mathcal{C}_L$ and $\mathcal{C}_R$. The gain in variance is of the order $\frac{\sigma^2}{n}$, while the reduction in bias is

$$\begin{aligned}\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} &- \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}_L\}\nu\{\mathcal{C}_L\} \\ &- \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}_R\}\nu\{\mathcal{C}_R\}.\end{aligned}$$

One can check that this is the population version of the weighted impurity decrease of this split, which is the quantity used to determine splits in CART, and also the value compared against a threshold in early stopping with the minimum impurity decrease criterion. These observations show that both these methods for preventing overfitting in CART attempt to optimize an objective function that is a weighted combination of plug-in estimates of the bias and variance terms in the expected risk decomposition for an honest tree.

The proof of Theorem 3.1 and that of a tighter but more complicated version of the decomposition can both be found in Appendix A. In the next two sections, we will see how the decomposition can be used in

conjunction with rate distortion theory to yield lower bounds for additive models.

4 Connecting decision trees to rate-distortion theory

The second ingredient we need is supplied by rate-distortion theory. We start by recalling some definitions from Cover and Thomas (2012). We will use $H(-)$, $h(-)$ and $I(-;-)$ to denote discrete entropy, differential entropy, and mutual information respectively. Let $\mathcal{X}$ be a subset of $\mathbb{R}^d$ as before. Given a vector $\boldsymbol{\beta} \in \mathbb{R}^d$, we denote the associated weighted Euclidean norm on $\mathcal{X}$ via $\|\mathbf{x}-\mathbf{y}\|_{\boldsymbol{\beta}}^2 \coloneqq \sum_{j=1}^d \beta_j^2 (x_j - y_j)^2$. Now let p denote a joint distribution on $\mathcal{X} \times \mathcal{X}$. The *distortion* of p with respect to $\|-\|_{\boldsymbol{\beta}}$ is defined as

$$\delta(p; \boldsymbol{\beta}) \coloneqq \mathbb{E}_{(\mathbf{x},\hat{\mathbf{x}})\sim p}\Big\{\|\mathbf{x} - \hat{\mathbf{x}}\|_{\boldsymbol{\beta}}^2\Big\}.$$

The *rate distortion function* of the marginal $p_{\mathbf{x}}$ is defined by

$$R(D; p_{\mathbf{x}}, \boldsymbol{\beta}) \coloneqq \inf_{p_{\hat{\mathbf{x}}|\mathbf{x}}} I(\mathbf{x}; \hat{\mathbf{x}})$$

where the infimum is taken over all conditional distributions such that $\delta(p_{\mathbf{x}} p_{\hat{\mathbf{x}}|\mathbf{x}}; \boldsymbol{\beta}) \le D$.

In rate-distortion theory, the rate distortion function characterizes the length of a binary code needed to encode a source so that the distortion is not too large. In this paper, it clarifies the trade-off between the bias and variance terms in the decomposition (4). Under some independence conditions, we show that the bias term is equivalent to a distortion, while the size of the partition occurring in the variance term is bounded from below by the rate of this distortion. More precisely, we have the following lemma.

Lemma 4.1 (Rate-distortion bound for oracle expected risk)**.** *Assume the regression model* (1)*, and that* $\mathcal{X} = \{0,1\}^d$ *or* $\mathcal{X} = [0,1]^d$*. Furthermore, assume that the covariates are independent, and that the conditional mean function is linear:* $f(\mathbf{x}) = \boldsymbol{\beta}^T \mathbf{x}$*. Then the oracle expected risk is lower bounded by*

$$\mathcal{R}^*(f, \nu, n) \ge \frac{1}{2} \inf_{D>0}\left\{D + \frac{\sigma^2 2^{R(D;\nu,\boldsymbol{\beta})}}{n}\right\}. \tag{6}$$

Proof. Consider some permissible partition $\mathfrak{p}$. For any cell $\mathcal{C} \in \mathfrak{p}$, notice that the conditional covariate distribution $\nu|_{\mathcal{C}}$ also has independent covariates. Let $\mathbf{x}'$ be an independent copy of $\mathbf{x}$. Using independence, we compute

$$\begin{aligned}
\operatorname{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} &= \frac{1}{2}\mathbb{E}\Big\{\big(\boldsymbol{\beta}^T(\mathbf{x}-\mathbf{x}')\big)^2 \mid \mathbf{x}, \mathbf{x}' \in \mathcal{C}\Big\} \\
&= \frac{1}{2}\mathbb{E}\left\{\sum_{j=1}^d \beta_j^2 \big(x_j - x_j'\big)^2 \mid \mathbf{x}, \mathbf{x}' \in \mathcal{C}\right\} \\
&= \frac{1}{2}\mathbb{E}\big\{\|\mathbf{x}-\mathbf{x}'\|_{\boldsymbol{\beta}}^2 \mid \mathbf{x}, \mathbf{x}' \in \mathcal{C}\big\} \\
&\ge \frac{1}{2}\mathbb{E}\big\{\|\mathbf{x} - \mathbf{z}(\mathcal{C})\|_{\boldsymbol{\beta}}^2 \mid \mathbf{x} \in \mathcal{C}\big\}, \qquad (7)
\end{aligned}$$

where $\mathbf{z}(\mathcal{C}) \coloneqq \arg\min_{\mathbf{x}' \in \mathcal{C}} \mathbb{E}\Big\{\|\mathbf{x}-\mathbf{x}'\|_{\boldsymbol{\beta}}^2 \mid \mathbf{x} \in \mathcal{C}\Big\}$.[3] To define a conditional distribution, for each $\mathbf{x}$, we let $p_{\hat{\mathbf{x}}|\mathbf{x}}$ be a Dirac mass at $\mathbf{z}(\mathcal{C}(\mathbf{x}))$, where $\mathcal{C}(\mathbf{x})$ is the cell in $\mathfrak{p}$ containing $\mathbf{x}$. Then the bias term in (4) can be lower bounded by the distortion for the joint distribution $p = \nu p_{\hat{\mathbf{x}}|\mathbf{x}}$:

$$\sum_{\mathcal{C} \in \mathfrak{p}} \operatorname{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} \ge \frac{\delta(p; \boldsymbol{\beta})}{2}.$$

Meanwhile, notice that $\hat{\mathbf{x}}$ is a discrete distribution on $|\mathfrak{p}|$ elements, so we may use the max entropy property of the uniform distribution to write

$$\log|\mathfrak{p}| \ge H(\hat{\mathbf{x}}) \ge I(\mathbf{x}; \hat{\mathbf{x}}) \ge R(\delta(p; \boldsymbol{\beta}); \nu, \boldsymbol{\beta}).$$

Plugging these formulas into (4) gives the lower bound

$$\begin{aligned}
&\sum_{\mathcal{C} \in \mathfrak{p}} \operatorname{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} + \frac{|\mathfrak{p}|\sigma^2}{2n} \\
&\quad \ge \frac{\delta(p; \boldsymbol{\beta})}{2} + \frac{\sigma^2 2^{R(\delta(p;\boldsymbol{\beta});\nu,\boldsymbol{\beta})}}{2n}.
\end{aligned}$$

Minimizing over all partitions yields (6). □

We remark that the lemma applies to both continuous and discrete distribution, and may be valid for other subsets $\mathcal{X} \subset \mathbb{R}^d$. Next, it is known that rate distortion functions are convex and monotonically decreasing, and one may therefore check that the same applies to the function $D \mapsto 2^{R(D;\nu,\boldsymbol{\beta})}$. As such, the right-hand-side of (6) is the solution to a convex optimization problem, and can be solved efficiently in principle. This is especially significant when $\mathcal{X}$ is the Boolean cube, because it allows us to turn what is a priori a combinatorial optimization problem into a smooth, convex one.

When $\boldsymbol{\beta}$ has at least s large coefficients, the independence of the covariates allows us to use standard calculations to bound $R(D; \nu, \boldsymbol{\beta})$ from below by elementary

[3] $\mathbf{z}(\mathcal{C})$ is the cell centroid when $\mathcal{X}$ is the unit length cube, but not when it is the Boolean cube. Furthermore, equality actually holds without the factor of $1/2$ in the former case.

functions, therefore giving us a closed-form formula for the oracle expected risk. We state the result for the case where $\mathcal{X}$ is the unit length cube, and leave more general versions of the theorem to the next section.

Theorem 4.2 (Lower bound for linear models)**.** *In addition to the assumptions of Lemma 4.1, assume that $\mathcal{X}$ is the unit length cube $[0,1]^d$, and that the covariates each follow some marginal distribution ν_0. Suppose there is some subset of coordinates $S \subset [d]$ of size s such that $|\beta_j| \geq \beta_0$ for all $j \in S$. Then the oracle expected risk is lower bounded by*

$$\mathcal{R}^*(f, \nu, n) \geq Cs\beta_0^2 \left(\frac{\sigma^2}{\beta_0^2 n} \right)^{2/(s+2)}, \tag{8}$$

where $C = \frac{1}{2}\left(2^{2h(\nu_0)}/\pi e\right)^{s/(s+2)}$.

Proof. Combining Lemmas B.1, B.2, and B.3, we know that $R(D; \nu, \boldsymbol{\beta})$ is lower bounded by the value of

$$\inf_{\beta_0^2 \sum_{j \in S} D_j \leq D} \sum_{j \in S} \left(h(\nu_0) - \frac{1}{2} \log(2\pi e D_j) \right) \vee 0. \tag{9}$$

This is a convex optimization problem, and by symmetry over the coordinate indices in S, it is easy to see that the infimum is achieved at $D_j = \frac{D}{s\beta_0^2} \vee \frac{2^{2h(\nu_0)}}{2\pi e}$ for $j \in S$, and $D_j = \frac{2^{2h(\nu_0)}}{2\pi e}$ for $j \notin S$, where $s = |S|$. Plugging these into (9), we get

$$R(D; \nu, \boldsymbol{\beta}) \geq s \left(h(\nu_0) - \frac{1}{2} \log \left(\frac{2\pi e D}{s\beta_0^2} \right) \right). \tag{10}$$

As such, we have

$$D + \frac{\sigma^2 2^{R(D;\nu,\boldsymbol{\beta})}}{n} \geq D + \frac{\sigma^2 2^{sh(\nu_0)}}{n} \left(\frac{s\beta_0^2}{2\pi e D} \right)^{s/2}.$$

Differentiating, we easily see that the minimum is achieved at

$$D = s 2^{\frac{2s}{s+2} h(\nu_0) - 1} \left(\frac{\beta_0^2}{\pi e} \right)^{s/(s+2)} \left(\frac{\sigma^2}{n} \right)^{2/(s+2)}.$$

Using this value in (6) and dropping the second term then completes the proof. □

5 Results for additive models

In the previous section, we saw how the bias-variance risk decomposition and an information theoretic argument can be used to obtain a lower bound for oracle expected risk. With a more sophisticated application of the latter, we can derive more powerful results for additive models over both continuous and Boolean feature spaces. We state these results in this section, deferring all proofs to the appendix because of space constraints.

Theorem 5.1 (Lower bound for additive models on unit length cube)**.** *Assume the regression model (1), with f be defined as in (2), and assume that the covariate space is the unit length cube $[0,1]^d$. Suppose $\phi_j \in C^1([0,1])$ for $j = 1, \ldots, d$. Let $I_1, I_2, \ldots, I_d \subset [0,1]$ be sub-intervals, and suppose there is some subset of indices $S \subset [d]$ of size s such that $\min_{t \in I_j} |\phi_j'(t)| \geq \beta_0 > 0$ for all $j \in S$. Denote $\mathcal{K} = \{\mathbf{x} \colon x_j \in I_j \text{ for } j = 1, \ldots, d\}$. Assume that ν is a continuous distribution with density q, and denote $q_{min} = \min_{\mathbf{x} \in \mathcal{K}} q(\mathbf{x})$. Then the oracle expected risk is lower bounded by*

$$\mathcal{R}^*(f, \nu, n) \geq Cs\beta_0^2 \left(\frac{\sigma^2}{\beta_0^2 n} \right)^{2/(s+2)}, \tag{11}$$

where $C = \mu(\mathcal{K}) q_{min}^{s/(s+2)}/12$.

As mentioned before, the $\Omega(n^{-2/(s+2)})$ rate in (11) is the ℓ_2 minimax rate for nonparametric estimation of C^1 functions in s dimensions (Stone, 1982). This is far worse than the minimax rate for estimating sparse additive models, which scales as $\max\left\{ \frac{s \log(d/s)}{n}, s\epsilon_n^2(\mathcal{H}) \right\}$, where $\epsilon_n(\mathcal{H})$ is a quantity that depends only on $\mathcal{H}$ and the sample size n (Raskutti et al., 2012).

The theorem is more flexible than Theorem 4.2 in the following ways: It allows the component functions ϕ_j to be nonlinear, and even have vanishing derivatives everywhere except on an interval, which means (11) applies to any nontrivial choice of the ϕ_j's. Furthermore, unlike Theorem 4.2, it does not require the covariates to be independent. Finally, a more general version of the lower bound, stated as Theorem C.1 in Appendix C, allows us to provide tighter lower bounds in the case where the β_j's may be decaying in magnitude rather than having a non-zero lower bound.

These improvements require a different information theoretic argument, which roughly works as follows: First, we derive the maximum volume a cell can have under gradient lower bounds and a prescribed variance constraint (see Lemma C.3.) We then use this to compute the number of cells necessary to cover the portion of $[0,1]^d$ over which the the gradient lower bounds hold. As an easy by-product of the above calculations, we also compute the optimal dimensions of a cell under the variance constraint. This allows us to derive matching oracle upper bounds for sparse additive models:

Proposition 5.2 (Upper bound for sparse additive models on unit length cube)**.** *Let f be a sparse additive model, i.e. there is a subset of coordinates $S \subset [d]$ such that $f(\mathbf{x}) = \sum_{j \in S} \phi_j(x_j)$. Assume that the covariate space is the unit length cube $[0,1]^d$. Suppose*

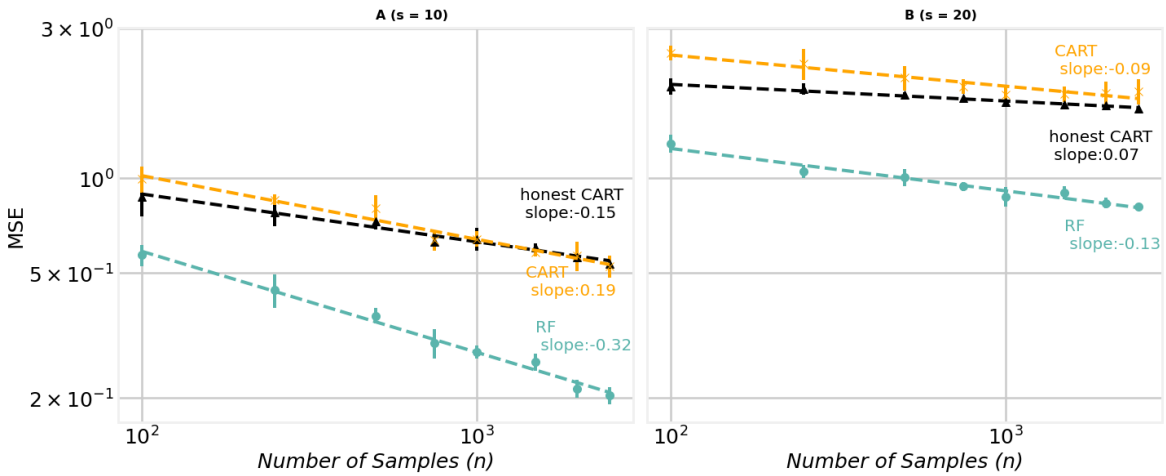

Figure 1: Scaling of the test set error for CART and RF for a sparse sum of squares generative model $y = \sum_j \beta_j x_j^2 + \epsilon$ with $\mathbf{x} \sim \text{Unif}([0,1]^d)$.We show the scaling with respect to n for **(A)** $s = 10$, and **(B)** $s = 20$..

$\phi_j \in C^1([0,1])$, and $\|\phi_j\|_\infty \leq \beta_{max}$ for $j \in S$. Assume ν is a continuous distribution with density q. Then

$$\limsup_{n\to\infty} \frac{\mathcal{R}^*(f,\nu,n)}{n^{-2/(s+2)}} \leq \left(Cs\beta_{max}^2 + 6\right)\sigma^{2/(s+2)}, \quad (12)$$

where $C = 168\|q\|_\infty$.

Our next main result is for additive models over the Boolean cube. Note that all additive models are linear in this setting, and we are thus able to prove this using the original rate-distortion argument.

Theorem 5.3 (Lower bounds for additive Boolean models)**.** *Assume the regression model* (1) *and that the conditional mean function is linear: $f(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x}$. Assume that the covariate space is the hypercube $\{0,1\}^d$, and that the covariates are independent, with $x_j \sim Ber(\pi)$, $0 \leq \pi \leq \frac{1}{2}$, for $j = 1,\ldots,d$. Suppose there is some subset of coordinates $S \subset [d]$ of size s such that $|\beta_j| \geq \beta_0 > 0$ for all $j \in S$. Then the oracle expected risk is lower bounded by*

$$\mathcal{R}^*(f,\nu,n) \geq \frac{s\beta_0^2}{2}\left(1 - \left(\frac{2e^s n\beta_0^2}{2^{sH(\pi)}\sigma^2}\right)^{\frac{1}{s-1}}\right). \quad (13)$$

The form of the lower bound (13) is different from that in Theorem 5.1. This is due to the fact that we can achieve zero approximation error over the Boolean cube with finitely many cells. As a consequence, while the rate-distortion function in the continuous case (10) tends to infinity as D tends to 0, that in the Boolean case (22) tends to a finite number. Our proof of (13) does not actually use the sharp rate bound (22), and instead approximates it by a more computationally tractable bound (23). It is unclear how much slack from this approximation propagates into the final bound (13), but it is reassuring that the general concave down shape of the test error scaling in Figure 3 is consistent with (13). We provide a more general version of the lower bound (Theorem B.5) in Appendix B.

We examined the empirical validity of each of our main results by simulating the generalization error of tree-based algorithms fitted to sparse linear models with both continuous and Boolean features, as well as an additive non-linear sum of squares model with continuous features. We present only results for the sum of squares model as seen in Figure 1 in the main text of the paper due to space constraints, while the results and experimental design for other experiments can be found in Appendix D. Details of our experimental design and algorithm settings are shown below.

Experimental design: We simulate data via a sparse sum of squares model $y = \sum_j \beta_j x_j^2 + \epsilon$ with $\mathbf{x} \sim \text{Unif}([0,1]^d)$. We varied n, but fixed $d = 50$, $\sigma^2 = 0.01$, and set $\beta_j = 1$ for $j = 1,\ldots,s$, and $\beta_j = 0$ otherwise, where s is a sparsity parameter. We ran the experiments with both $s = 10$ and $s = 20$, and plotted the results for each setting in panel A and panel B respectively for all of the figures. We computed the generalization error using a test set of size 500, averaging the results over 25 runs.

Algorithm settings: We fit both honest and non-honest versions of CART, as well as the non-honest version of RF using a training set of size n. For the honest version of CART, we use a separate independent sample of size n to compute averages over each leaf in the tree. Furthermore, if a cell contains no samples from the training data used to do averaging, we search for the closest ancestor node that contains at least one sample and use the average over that node

to make a prediction. We use `min_samples_leaf=5` as the stopping condition, although we also ran experiments with cost-complexity pruning and achieved very similar results.

We note that Figures 1 and 2 not only give an empirical validation of the theoretical rates in our lower bound, which are 0.17 and 0.09 for $s = 10$ and 20 respectively, but also indicate that honest CART almost achieves these bounds despite there being no a priori guarantee that CART grows an optimal tree. While the theory does not cover the case of non-honest CART, its test error is worse than that of honest CART, and has a similar rate. An interesting facet of all our simulations is that RF has a markedly faster rate, implying that diverse trees allow the algorithm to pool information across the training samples more efficiently, supporting Breiman (2001)'s original hypothesis.

There are a few interesting additional observations that can be made. First, we remark that although Theorem 5.1 is stated in terms of distributions over the unit cube, we can easily extend it to non-compact distributions over $\mathbb{R}^d$ such as multivariate Gaussians by using marginal quantile transforms. Second, from the formulas (8) and (13), we see that the lower bound decreases with the entropy of the covariate distribution, although the rate in the sample size n remains the same. Third, in Appendix B, we provide more complicated lower bounds (25) and (20), which, for a fixed value of $\|\boldsymbol{\beta}\|_2^2$, are smaller when more of the ℓ_2 energy is concentrated in fewer coordinates, i.e. when the coefficients experience faster decay. This agrees with our intuition that decision trees are adaptive to low-dimensional structure beyond the hard sparsity regime, which has been the focus of recent literature (Syrgkanis and Zampetakis, 2020; Klusowski, 2020, 2021).

6 Discussion

In this paper, we have obtained theoretical lower bounds on the expected risk for honest ALA trees when fitted to additive models, while our simulations suggest that these results should also hold for their non-honest counterparts. These bounds lead us to argue that such estimators, including CART, have an inductive bias against global structure, a bias that arises not from the greedy splitting criterion used by most decision tree algorithms, but from the leaf-only averaging property of this class of estimators. Furthermore, we provide experimental evidence that the bounds do not apply to RF, which supports Breiman (2001)'s original narrative that the diversity of trees in a forest helps to reduce variance and improve prediction performance. Nonetheless, the rates exhibited by RF are still significantly slower than the minimax rates for sparse additive models, hinting at fundamental limits we are yet to understand.

Our results further the conversation about how decision tree algorithms can be improved, and suggest that they should be modified to more easily learn global structure. One natural idea on how to do this is to adopt some type of hierarchical shrinkage or global pooling. Another is to combine tree-based methods with linear or additive methods in a way that incorporates the statistical advantages of both classes of methods, in the vein of Friedman and Popescu (2008)'s RuleFit.[4] Recently, Bloniarz et al. (2016) and Friedberg et al. (2020) suggested using the RF kernel in conjunction with local linear (or polynomial) regression, while Künzel et al. (2019) replaced the constant prediction over leaf with a linear model. These works, however, aim at modifying RF to better exploit smoothness, and do not directly address the loss in power for detecting global structure that comes from partitioning the covariate space. Furthermore, the focus on forests forestalls the possibility of preserving interpretability.

Taking a step back to look at the bigger picture, we believe that it is important to analyze the generalization performance of CART and other decision tree algorithms on other generative regression models in order to further elicit their inductive biases. The same type of analyses can also be applied to other machine learning algorithms. Since real world data sets often present some structure that can be exploited using the right inductive bias, this research agenda will allow us to better identify which algorithm to use in a given application, especially in settings, such as the estimation of heterogeneous treatment effects, where a held out test set is not available. Moreover, as seen in this paper, such investigations can yield inspiration for improving existing algorithms.

The approach we follow is different from the classical paradigm of statistical estimation, which starts with an estimation problem, and then searches for estimation procedures that can achieve some form of optimality. Instead, as is common in machine learning, we take an algorithm as the primitive object of investigation, and seek to analyze its performance under different generative models to elicit its inductive bias. This approach is more aligned with modern data analysis, in which we seldom have a good grasp over the functional form of the data generating process, leading to a hand-

[4] Given the interest of practitioners in using tree-based methods to identify interactions in genomics (Chen and Ishwaran, 2012; Boulesteix et al., 2012; Basu et al., 2018; Behr et al., 2021), it is fair to say that this is a key strength of trees and RF.

ful of general purpose algorithms being used for the vast majority of prediction problems. This embrace of suboptimality is consistent with viewing the models as approximations – an old tradition in the statistical literature (Huber (1967); Box (1979); Grenander (1981); Geman and Hwang (1982); Buja et al. (2019)).

We have only scratched the surface of investigating the inductive biases of decision trees, RF, gradient boosting, and other tree-based methods, and envision an abundant garden for future work.

References

Sina Aghaei, Andrés Gómez, and Phebe Vayanos. Strong optimal classification trees. *arXiv preprint arXiv:2103.15965*, 2021.

Elaine Angelino, Nicholas Larus-Stone, Daniel Alabi, Margo Seltzer, and Cynthia Rudin. Learning certifiably optimal rule lists for categorical data. *Journal of Machine Learning Research*, 18:1–78, 2018.

Susan Athey and Guido Imbens. Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360, 2016.

Sumanta Basu, Karl Kumbier, James B Brown, and Bin Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018.

Merle Behr, Yu Wang, Xiao Li, and Bin Yu. Provable boolean interaction recovery from tree ensemble obtained via random forests. *arXiv preprint arXiv:2102.11800*, 2021.

Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13:1063–1095, 2012.

Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

Adam Bloniarz, Ameet Talwalkar, Bin Yu, and Christopher Wu. Supervised neighborhoods for distributed nonparametric regression. In *Artificial Intelligence and Statistics*, pages 1450–1459. PMLR, 2016.

Anne-Laure Boulesteix, Silke Janitza, Jochen Kruppa, and Inke R König. Overview of random forest methodology and practical guidance with emphasis on computational biology and bioinformatics. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 2(6):493–507, 2012.

George EP Box. Robustness in the strategy of scientific model building. In *Robustness in Statistics*, pages 201–236. Elsevier, 1979.

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

Leo Breiman and Jerome H Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

Andreas Buja, Lawrence Brown, Richard Berk, Edward George, Emil Pitkin, Mikhail Traskin, Kai Zhang, and Linda Zhao. Models as approximations I: Consequences illustrated with linear regression. *Statistical Science*, 34(4):523–544, 2019.

Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine learning*, pages 161–168, 2006.

Rich Caruana, Nikos Karampatziakis, and Ainur Yessenalina. An empirical evaluation of supervised learning in high dimensions. In *Proceedings of the 25th International Conference on Machine learning*, pages 96–103, 2008.

Tianqi Chen and Carlos Guestrin. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

Xi Chen and Hemant Ishwaran. Random forests for genomic data analysis. *Genomics*, 99(6):323–329, 2012.

T.M. Cover and J.A. Thomas. *Elements of Information Theory*. Wiley, 2012. ISBN 9781118585771. URL `https://books.google.com/books?id=VWq5GG6ycxMC`.

Finale Doshi-Velez and Been Kim. Towards a rigorous science of interpretable machine learning. *arXiv preprint arXiv:1702.08608*, 2017.

Manuel Fernández-Delgado, Eva Cernadas, Senén Barro, and Dinani Amorim. Do we need hundreds of classifiers to solve real world classification problems? *The Journal of Machine Learning Research*, 15(1):3133–3181, 2014.

Rina Friedberg, Julie Tibshirani, Susan Athey, and Stefan Wager. Local linear forests. *Journal of Computational and Graphical Statistics*, pages 1–15, 2020.

Jerome Friedman, Trevor Hastie, Robert Tibshirani, et al. *The Elements of Statistical Learning*, volume 1. Springer Series in Statistics New York, 2001.

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

Jerome H Friedman and Bogdan E Popescu. Predictive learning via rule ensembles. *The Annals of Applied Statistics*, 2(3):916–954, 2008.

Stuart Geman and Chii-Ruey Hwang. Nonparametric maximum likelihood estimation by the method of sieves. *The Annals of Statistics*, pages 401–414, 1982.

U. Grenander. *Abstract Inference*. Wiley, 1981.

Trevor Hastie and Robert Tibshirani. Generalized additive models. *Statistical Science*, pages 297–310, 1986.

Giles Hooker and Lucas Mentch. Bridging Breiman's brook: From algorithmic modeling to statistical learning. *Observational Studies*, 7(1):107–125, 2021.

Peter J Huber. The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics*, pages 221–233. University of California Press, 1967.

Jason Klusowski. Sparse learning with cart. In *Advances in Neural Information Processing Systems*, volume 33, pages 11612–11622, 2020.

Jason M. Klusowski. Universal consistency of decision trees in high dimensions. *arXiv preprint arXiv:2104.13881*, 2021.

Sören R Künzel, Theo F Saarinen, Edward W Liu, and Jasjeet S Sekhon. Linear aggregation in tree-based estimators. *arXiv preprint arXiv:1906.06463*, 2019.

Nathan Kuppermann, James F Holmes, Peter S Dayan, John D Hoyle, Shireen M Atabaki, Richard Holubkov, Frances M Nadel, David Monroe, Rachel M Stanley, Dominic A Borgialli, et al. Identification of children at very low risk of clinically-important brain injuries after head trauma: a prospective cohort study. *The Lancet*, 374(9696): 1160–1170, 2009.

Benjamin Letham, Cynthia Rudin, Tyler H McCormick, and David Madigan. Interpretable classifiers using rules and bayesian analysis: Building a better stroke prediction model. *The Annals of Applied Statistics*, 9(3):1350–1371, 2015.

Jimmy Lin, Chudi Zhong, Diane Hu, Cynthia Rudin, and Margo Seltzer. Generalized and scalable optimal sparse decision trees. In *International Conference on Machine Learning*, pages 6150–6160. PMLR, 2020.

Wei-Yin Loh. Fifty years of classification and regression trees. *International Statistical Review*, 82(3): 329–348, 2014.

James N Morgan and John A Sonquist. Problems in the analysis of survey data, and a proposal. *Journal of the American Statistical Association*, 58(302): 415–434, 1963.

W James Murdoch, Chandan Singh, Karl Kumbier, Reza Abbasi-Asl, and Bin Yu. Definitions, methods, and applications in interpretable machine learning. *Proceedings of the National Academy of Sciences*, 116(44):22071–22080, 2019.

Randal S Olson, William La Cava, Zairah Mustahsan, Akshay Varik, and Jason H Moore. Data-driven advice for applying machine learning to bioinformatics problems. In *Biocomputing 2018: Proceedings of the Pacific Symposium*, pages 192–203. World Scientific, 2018.

Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *The Journal of Machine Learning Research*, 12:2825–2830, 2011.

Garvesh Raskutti, Martin J Wainwright, and Bin Yu. Minimax-optimal rates for sparse additive models over kernel classes via convex programming. *Journal of Machine Learning Research*, 13(2), 2012.

Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215, 2019.

Cynthia Rudin, Chaofan Chen, Zhi Chen, Haiyang Huang, Lesia Semenova, and Chudi Zhong. Interpretable machine learning: Fundamental principles and 10 grand challenges. *arXiv preprint arXiv:2103.11251*, 2021.

Veeranjaneyulu Sadhanala and Ryan J Tibshirani. Additive models with trend filtering. *The Annals of Statistics*, 47(6):3032–3068, 2019.

Erwan Scornet. Trees, forests, and impurity-based variable importance. *arXiv preprint arXiv:2001.04295*, 2020.

Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716 – 1741, 2015. doi: 10.1214/15-AOS1321. URL `https://doi.org/10.1214/15-AOS1321`.

Henry J Steadman, Eric Silver, John Monahan, Paul Appelbaum, Pamela Clark Robbins, Edward P Mulvey, Thomas Grisso, Loren H Roth, and Steven Banks. A classification tree approach to the development of actuarial violence risk assessment tools. *Law and Human Behavior*, 24(1):83–100, 2000.

Charles J Stone. Consistent nonparametric regression. *The Annals of Statistics*, pages 595–620, 1977.

Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The Annals of Statistics*, pages 1040–1053, 1982.

Vasilis Syrgkanis and Manolis Zampetakis. Estimation and inference with trees and forests in high dimensions. In *Conference on Learning Theory*, pages 3453–3454. PMLR, 2020.

Cheng Tang, Damien Garreau, and Ulrike von Luxburg. When do random forests fail? In *NeurIPS*, pages 2987–2997, 2018.

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

Marvin N Wright, Andreas Ziegler, et al. ranger: A fast implementation of random forests for high dimensional data in C++ and R. *Journal of Statistical Software*, 77(i01), 2017.

A Proof of Theorem 3.1

We first state and prove a tighter and more complicated version of the bias-variance decomposition of the expected risk. For notational convenience, we denote

$$\hat{\mathbb{E}}_{\mathcal{C}}\{y\} \coloneqq \hat{\mathbb{E}}\{y|\mathbf{x} \in \mathcal{C}\} = \frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} y^{(i)}.$$

Proposition A.1 (Bias-variance decomposition of risk)**.** *Assume the regression model* (1)*. Given a partition* $\mathfrak{p}$ *and a training set* $\mathcal{D}_n$*, the expected squared error risk satisfies the following upper and lower bounds*

$$\mathbb{E}\mathcal{R}(\hat{f}(-;\mathfrak{p},\mathcal{D}_n)) \geq \sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} \left(\nu\{\mathcal{C}\} + \frac{1}{n}\right) + \frac{|\mathfrak{p}|\sigma^2}{n} + E_1 - E_2 \tag{14}$$

$$\mathbb{E}\mathcal{R}(\hat{f}(-;\mathfrak{p},\mathcal{D}_n)) \leq \sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} \left(\nu\{\mathcal{C}\} + \frac{6}{n}\right) + \frac{6|\mathfrak{p}|\sigma^2}{n} + E_1 \tag{15}$$

where

$$E_1 = \sum_{\mathcal{C}\in\mathfrak{p}} \mathbb{E}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}^2 \left(1 - \nu\{\mathcal{C}\}\right)^n \nu\{\mathcal{C}\}.$$

$$E_2 = \frac{1}{n} \sum_{\mathcal{C}\in\mathfrak{p}} \left(\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2\right) \left(1 - \nu\{\mathcal{C}\}\right)^n.$$

Proof. First consider a cell $\mathcal{C}$, with $\mathbf{x} \in \mathcal{C}$. Supposing that $N(\mathcal{C}) \neq 0$, We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left\{\left(f(\mathbf{x}) - \hat{\mathbb{E}}_{\mathcal{C}}\{y\}\right)^2 \mid N(\mathcal{C})\right\} &= \mathbb{E}_{\mathcal{D}_n}\left\{\left(f(\mathbf{x}) - \frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} \left(f(\mathbf{x}^{(i)}) + \epsilon_i\right)\right)^2 \,\middle|\, N(\mathcal{C})\right\} \\
&= \mathbb{E}_{\mathcal{D}_n}\left\{\left(f(\mathbf{x}) - \frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} f(\mathbf{x}^{(i)})\right)^2 \,\middle|\, N(\mathcal{C})\right\} + \frac{\sigma^2}{N(\mathcal{C})}.
\end{aligned}$$

Taking a further conditional expectation with respect to $\mathbf{x} \in \mathcal{C}$, we see that the distribution $\mathbf{x}$ is the same as that of each $\mathbf{x}^{(i)}$. We can therefore compute

$$\mathbb{E}_{\mathbf{x}}\left\{\mathbb{E}_{\mathcal{D}_n}\left\{\left(f(\mathbf{x}) - \frac{1}{N(\mathcal{C})} \sum_{\mathbf{x}^{(i)} \in \mathcal{C}} f(\mathbf{x}^{(i)})\right)^2 \,\middle|\, N(\mathcal{C})\right\} \,\middle|\, \mathbf{x} \in \mathcal{C}\right\} = \left(1 + \frac{1}{N(\mathcal{C})}\right) \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}.$$

Putting these two calculations together, and interchanging the order of expectation, we have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_{\mathbf{x}}\left\{\left(f(\mathbf{x}) - \hat{\mathbb{E}}_{\mathcal{C}}\{y\}\right)^2 \mid \mathbf{x} \in \mathcal{C}\right\} \mid N(\mathcal{C})\right\} &= \mathbb{E}_{\mathbf{x}}\left\{\mathbb{E}_{\mathcal{D}_n}\left\{\left(f(\mathbf{x}) - \hat{\mathbb{E}}_{\mathcal{C}}\{y\}\right)^2 \mid N(\mathcal{C})\right\} \mid \mathbf{x} \in \mathcal{C}\right\} \\
&= \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \frac{\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2}{N(\mathcal{C})}.
\end{aligned}$$

Recall our convention that we set $\hat{\mathbb{E}}_{\mathcal{C}}\{y\} = 0$ if $N(\mathcal{C}) = 0$. We may then write

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n,\mathbf{x}}\left\{\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2 \,\middle|\, \mathbf{x} \in \mathcal{C}\right\} &= \mathbb{E}_{\mathcal{D}_n,\mathbf{x}}\left\{\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2 \mathbf{1}\{N(\mathcal{C}) \neq 0\} \,\middle|\, \mathbf{x} \in \mathcal{C}\right\} \\
&\quad + \mathbb{E}_{\mathcal{D}_n,\mathbf{x}}\left\{\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2 \mathbf{1}\{N(\mathcal{C}) = 0\} \,\middle|\, \mathbf{x} \in \mathcal{C}\right\} \\
&= \mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_{\mathcal{D}_n}\left\{\mathbb{E}_{\mathbf{x}}\left\{\left(f(\mathbf{x}) - \hat{\mathbb{E}}_{\mathcal{C}}\{y\}\right)^2 \,\middle|\, \mathbf{x} \in \mathcal{C}\right\} \,\middle|\, N(\mathcal{C})\right\}\mathbf{1}\{N(\mathcal{C}) \neq 0\}\right\} \\
&\quad + \mathbb{E}\{f(\mathbf{x})^2 \,|\, \mathbf{x} \in \mathcal{C}\}\mathbb{P}\{N(\mathcal{C}) = 0\} \\
&= \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\mathbb{P}\{N(\mathcal{C}) \neq 0\} + \mathbb{E}\{f(\mathbf{x})^2 \,|\, \mathbf{x} \in \mathcal{C}\}\mathbb{P}\{N(\mathcal{C}) = 0\} \\
&\quad + \left(\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2\right)\mathbb{E}_{\mathcal{D}_n}\left\{\frac{\mathbf{1}\{N(\mathcal{C}) \neq 0\}}{N(\mathcal{C})}\right\} \qquad (16)
\end{aligned}$$

Note that

$$\mathbb{E}\{f(\mathbf{x})^2 \,|\, \mathbf{x} \in \mathcal{C}\} = \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \mathbb{E}\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{C}\}^2$$

so that the first two terms on the right hand side of (16) can be rewritten as

$$\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \mathbb{E}\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{C}\}^2\mathbb{P}\{N(\mathcal{C}) = 0\}$$

We continue by providing upper and lower bounds for $\mathbb{E}_{\mathcal{D}_n}\left\{\frac{\mathbf{1}\{N(\mathcal{C})\neq 0\}}{N(\mathcal{C})}\right\}$. For the upper bound, recall that $N(\mathcal{C})$ is a binomial random variable, and so has variance smaller than its expectation. This allows to apply Chebyshev's inequality to get

$$\begin{aligned}
\mathbb{P}\left\{|N(\mathcal{C}) - \mathbb{E}N(\mathcal{C})| \geq \frac{\mathbb{E}N(\mathcal{C})}{2}\right\} &\leq \frac{\mathbb{E}N(\mathcal{C})}{\left(\frac{1}{2}\mathbb{E}N(\mathcal{C})\right)^2} \\
&= \frac{4}{\mathbb{E}N(\mathcal{C})}
\end{aligned}$$

Next, since $\frac{\mathbf{1}\{N(\mathcal{C})\neq 0\}}{N(\mathcal{C})} \leq 1$ we have

$$\begin{aligned}
\mathbb{E}\left\{\frac{\mathbf{1}\{N(\mathcal{C}) \neq 0\}}{N(\mathcal{C})}\right\} &\leq \frac{2}{\mathbb{E}N(\mathcal{C})}\mathbb{P}\left\{N(\mathcal{C}) \geq \frac{\mathbb{E}N(\mathcal{C})}{2}\right\} + \mathbb{P}\left\{N(\mathcal{C}) \leq \frac{\mathbb{E}N(\mathcal{C})}{2}\right\} \\
&\leq \frac{2}{\mathbb{E}N(\mathcal{C})} + \frac{4}{\mathbb{E}N(\mathcal{C})} \\
&= \frac{6}{\mathbb{E}N(\mathcal{C})}.
\end{aligned}$$

The right hand side of (16) is bounded above by

$$\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \mathbb{E}\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{C}\}^2\mathbb{P}\{N(\mathcal{C}) = 0\} + \frac{6}{\mathbb{E}N(\mathcal{C})}\left(\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2\right)$$

Observe that $\mathbb{P}\{N(\mathcal{C}) = 0\} = (1-\nu\{\mathcal{C}\})^n$, and $\mathbb{E}N(\mathcal{C}) = n\nu\{\mathcal{C}\}$. Taking expectation with respect to $\mathbf{x}$ then gives

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}_n,\mathbf{x}}\left\{\left(f(\mathbf{x}) - \hat{f}(\mathbf{x})\right)^2\right\} &\leq \sum_{\mathcal{C}\in\mathfrak{p}}\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} + \sum_{\mathcal{C}\in\mathfrak{p}}\mathbb{E}\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{C}\}^2(1-\nu\{\mathcal{C}\})^n\nu\{\mathcal{C}\} \\
&\quad + \frac{6}{n}\sum_{\mathcal{C}\in\mathfrak{p}}\left(\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2\right).
\end{aligned}$$

Rearranging this gives (15).

The lower bound is follows from Cauchy-Schwarz. We have

$$\begin{aligned}\mathbb{P}\{N(\mathcal{C}) \neq 0\} &= \mathbb{E}\left\{ \frac{N(\mathcal{C})^{1/2}\mathbf{1}\{N(\mathcal{C}) \neq 0\}}{N(\mathcal{C})^{1/2}} \right\} \\ &\leq \mathbb{E}\{N(\mathcal{C})\}\mathbb{E}\left\{ \frac{\mathbf{1}\{N(\mathcal{C}) \neq 0\}}{N(\mathcal{C})} \right\}.\end{aligned}$$

Rearranging this gives

$$\begin{aligned}\mathbb{E}\left\{ \frac{\mathbf{1}\{N(\mathcal{C}) \neq 0\}}{N(\mathcal{C})} \right\} &\geq \frac{\mathbb{P}\{N(\mathcal{C}) \neq 0\}}{\mathbb{E}N(\mathcal{C})} \\ &= \frac{1 - \mathbb{P}\{N(\mathcal{C}) = 0\}}{\mathbb{E}N(\mathcal{C})}.\end{aligned}$$

The right hand side of (16) is bounded below by

$$\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \mathbb{E}\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{C}\}^2 \mathbb{P}\{N(\mathcal{C}) = 0\} + \frac{1 - \mathbb{P}\{N(\mathcal{C}) = 0\}}{\mathbb{E}N(\mathcal{C})} \left(\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2 \right).$$

Taking expectation with respect to $\mathbf{x}$ then gives

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_n, \mathbf{x}}\left\{ \left(f(\mathbf{x}) - \hat{f}(\mathbf{x}) \right)^2 \right\} \geq \sum_{\mathcal{C} \in \mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} + \frac{1}{n} \sum_{\mathcal{C} \in \mathfrak{p}} \left(\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2 \right). \\ + \sum_{\mathcal{C} \in \mathfrak{p}} \left(\mathbb{E}\{f(\mathbf{x}) \,|\, \mathbf{x} \in \mathcal{C}\}^2 \nu\{\mathcal{C}\} - \frac{\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \sigma^2}{n} \right) (1 - \nu\{\mathcal{C}\})^n.\end{aligned}$$

Rearranging this gives (14). □

Proof of Theorem 3.1. We use the fact that for any cell $\mathcal{C}$,

$$(1 - \nu\{\mathcal{C}\})^n \leq \left(1 - \frac{1}{n}\right)^n \leq \frac{1}{2}.$$

As such, the term E_2 in (14) is at most

$$E_2 \leq \frac{1}{2n} \sum_{\mathcal{C} \in \mathfrak{p}} \mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} + \frac{|\mathfrak{p}|\sigma^2}{2n}.$$

After performing cancellations, we get (4). The upper bound follows similarly. □

B Proofs for rate-distortion argument

Lemma B.1 (Rates over product distributions)**.** *Suppose $\nu = \nu_1 \times \nu_2 \times \cdots \times \nu_d$ is a product distribution on $\mathcal{X} \subset \mathbb{R}^d$. For all $D > 0$, we have*

$$R(D; \nu, \boldsymbol{\beta}) \geq \inf_{\sum_j \beta_j^2 D_j \leq D} \sum_{j=1}^{d} R(D_j; \nu_j, 1).$$

Proof. Let $\hat{\mathbf{x}}$ follow the conditional distribution achieving the infimum in the definition of $R(D; \nu, \boldsymbol{\beta})$. Following

the calculations in Chapter 10 of Cover and Thomas (2012), we have

$$\begin{aligned} R(D;\nu,\boldsymbol{\beta}) &= I(\mathbf{x};\hat{\mathbf{x}}) \\ &= \sum_{j=1}^{d} h(x_j) - \sum_{j=1}^{d} h(x_j|x_{1:j-1},\hat{\mathbf{x}}) \\ &\geq \sum_{j=1}^{d} h(x_j) - \sum_{j=1}^{d} h(x_j|\hat{x}_j) \\ &= \sum_{j=1}^{d} I(x_j;\hat{x}_j). \end{aligned}$$

Note that $x_j \sim \nu_j$ for $j = 1,\ldots,d$. Denoting $\delta_j \coloneqq \mathbb{E}\{(x_j - \hat{x}_j)^2\}$ for each j, we can therefore write

$$I(x_j;\hat{x}_j) \geq R(D_j;\nu_j,1).$$

Finally, observe that

$$\sum_{j=1}^{d} \beta_j^2 \delta_j = \mathbb{E}\Big\{ \|\mathbf{x} - \hat{\mathbf{x}}\|_{\boldsymbol{\beta}}^2 \Big\} \leq D,$$

so taking the infimum over possible values of δ_j's satisfying this constraint gives us the statement of the lemma. □

Lemma B.2 (Rates for dominated weighted norms)**.** *Let $\boldsymbol{\beta}$ and $\boldsymbol{\beta}'$ be two vectors such as $\beta_j^2 \geq (\beta_j')^2$ for $j = 1,\ldots,d$. Then for all $D > 0$, we have*

$$R(D;\nu,\boldsymbol{\beta}) \geq R(D;\nu,\boldsymbol{\beta}').$$

Proof. Obvious from the definition of the rate as an infimum. □

Lemma B.3 (Univariate rates)**.** *Let ν_0 be a continuous distribution on $\mathbb{R}$. Then we have*

$$R(D;\nu_0,1) \geq \left(h(\nu_0) - \frac{1}{2}\log(2\pi e D) \right) \vee 0.$$

If ν_0 is Bernoulli with parameter $0 < \pi_0 \leq 1/2$, then we have the tighter bound:

$$R(D;\nu_0,1) \geq (H(\pi_0) - H(D)) \vee 0.$$

Proof. We once again follow the calculations in Chapter 10 of Cover and Thomas (2012). Let $\hat{x}$ follow the conditional distribution achieving the infimum in the definition of $R(D;\nu_0,1)$. Then

$$\begin{aligned} R(D;\nu_0,1) &= I(x;\hat{x}) \\ &= h(x) - h(x|\hat{x}) \\ &= h(x) - h(x-\hat{x}|\hat{x}) \\ &\geq h(x) - h(x-\hat{x}). \end{aligned}$$

Next, we use the maximum entropy property of the normal distribution to write

$$h(x-\hat{x}) \leq \frac{1}{2}\log(2\pi e D).$$

Combining this with the observation that mutual information is non-negative completes the proof of the first statement. For the second statement, we repeat the same arguments with discrete entropy, and observe that

$$H(x-\hat{x}) = H(D).$$

□

Lemma B.4 (Rate bound for Boolean covariates)**.** *Assume the conditions of Theorem B.5. The rate distortion function may be lowered bounded as*

$$R(D;\nu,\boldsymbol{\beta}) \geq \sum_{j\colon \beta_j^2 \geq \colon m^{-1}_{\boldsymbol{\beta},\boldsymbol{\pi}}(D)\log((1-\pi_j)/(\pi_j))} H(\pi_j) - H\left(\frac{1}{1+e^{\beta_j^2/m^{-1}_{\boldsymbol{\beta},\boldsymbol{\pi}}(D)}}\right) \tag{17}$$

Proof. Combining Lemmas B.1 and B.3, we get

$$R(D;\nu,\boldsymbol{\beta}) \geq \inf_{\sum_j \beta_j^2 D_j \leq D} \sum_{j=1}^d (H(\pi_j) - H(D_j)) \vee 0. \tag{18}$$

The right hand side is equivalent to the solution of the following convex optimization program:

$$\min \sum_{j=1}^d H(\pi_j) - H(\delta_j) \quad \text{s.t.} \quad \sum_{j=1}^d \beta_j^2 \delta_j \leq D, \quad \delta_j \leq \pi_j \text{ for } j = 1,2,\ldots,d.$$

The Lagrangian of this program is

$$L(\boldsymbol{\delta},\boldsymbol{\lambda}) = \sum_{j=1}^d H(\pi_j) - H(\delta_j) + \lambda_0\left(\sum_{j=1}^d \beta_j^2\delta_j - D\right) + \sum_{j=1}^d \lambda_j(\delta_j - \pi_j).$$

Differentiating with respect to δ_j, we get

$$\frac{dL}{d\delta_j} = \log\left(\frac{\delta_j}{1-\delta_j}\right) + \lambda_0\beta_j^2 + \lambda_j.$$

Let δ_j^*, $j = 1,\ldots,d$ and λ_j^*, $j = 0,\ldots d$ denote the solution to KKT conditions. The above equation yields

$$\delta_j^* = \frac{1}{1+e^{\lambda_j^* + \lambda_0^*\beta_j^2}}.$$

By complementary slackness, we have either $\lambda_j^* = 0$ or $\delta_j^* = \pi_j$ for each j. It is easy to see that this implies

$$\delta_j^* = \pi_j \wedge \frac{1}{1+e^{\beta_j^2/\alpha}}$$

where α is chosen so that

$$D = \sum_{j=1}^d \beta_j^2\delta_j^* = m_{\boldsymbol{\beta},\boldsymbol{\pi}}(\alpha).$$

Plugging these values of δ_j into (18) gives us (17). □

Theorem B.5 (Lower bounds for additive Boolean models)**.** *Assume the regression model* (1) *and that the conditional mean function is linear: $f(\mathbf{x}) = \boldsymbol{\beta}^T\mathbf{x}$. Assume that the covariate space is the hypercube $\{0,1\}^d$, and that the covariates are independent, with $x_j \sim Ber(\pi_j)$, $0 \leq \pi_j \leq \frac{1}{2}$, for $j = 1,\ldots,d$. Define the function $m_{\boldsymbol{\beta},\boldsymbol{\pi}}\colon \left(0, \max_j \frac{\beta_j^2}{\log((1-\pi_j)/\pi_j)}\right) \to \mathbb{R}$ via the formula*

$$m_{\boldsymbol{\beta},\boldsymbol{\pi}}(\alpha) = \sum_{j=1}^d \beta_j^2\left(\pi_j \wedge \frac{1}{1+e^{\beta_j^2/\alpha}}\right) \tag{19}$$

and notice that it is strictly increasing and hence invertible on its domain. Then we have

$$\mathcal{R}^*(f,\nu,n) \geq \frac{1}{2}\inf_{D>0}\left\{D + \frac{\sigma^2 2^{R(D)}}{n}\right\}, \tag{20}$$

where

$$R(D) = \sum_{j\colon \beta_j^2 \geq \colon m^{-1}_{\boldsymbol{\beta},\boldsymbol{\pi}}(D)\log((1-\pi_j)/(\pi_j))} H(\pi_j) - H\left(\frac{1}{1+e^{\beta_j^2/m^{-1}_{\boldsymbol{\beta},\boldsymbol{\pi}}(D)}}\right).$$

In particular, if $\pi_j = \pi$ *for* $j = 1 \dots d$, *and* $\min_{j\in S}|\beta_j| \geq \beta_0 > 0$ *for some subset of indices* S *of size* s, *then we have*

$$\mathcal{R}^*(f,\nu,n) \geq \frac{s\beta_0^2}{2}\left(1 - \left(\frac{2e^s n\beta_0^2}{2^{sH(\pi)}\sigma^2}\right)^{\frac{1}{s-1}}\right). \tag{21}$$

Proof. The first statement in the theorem follows immediately from plugging in the bound from Lemma B.4 into Lemma 4.1. For the second statement, we first use Lemma B.2 to see that it suffices to bound $R(D;\nu,\tilde{\boldsymbol{\beta}})$, where $\tilde{\beta}_j = \beta_0$ for $j \in S$, and $\tilde{\beta}_j = 0$ for $j \notin S$. One can check that

$$m_{\tilde{\boldsymbol{\beta}},\boldsymbol{\pi}}(\alpha) = s\beta_0^2\left(\pi \wedge \frac{1}{1+e^{\beta_0^2/\alpha}}\right),$$

and so

$$\frac{1}{1+e^{\beta_j^2/m^{-1}_{\tilde{\boldsymbol{\beta}},\boldsymbol{\pi}}(D)}} = \frac{D}{s\beta_0^2} \wedge \pi.$$

Plugging this formula into (17), we get

$$R(D;\nu,\tilde{\boldsymbol{\beta}}) \geq s\left(H(\pi) - H\left(\frac{D}{s\beta_0^2}\right)\right) \vee 0. \tag{22}$$

For $\frac{D}{s\beta_0^2} \leq \pi$, we expand

$$\begin{aligned} 2^{sR(D;\nu,\boldsymbol{\beta})} &\geq \left(\frac{D}{s\beta_0^2}\right)^{\frac{D}{\beta_0^2}}\left(1-\frac{D}{s\beta_0^2}\right)^{s\left(1-\frac{D}{s\beta_0^2}\right)} \\ &\geq e^{-s}\left(1-\frac{D}{s\beta_0^2}\right)^s, \end{aligned} \tag{23}$$

where the second inequality comes from applying Lemma B.6. Optimizing the expression

$$D + \frac{\sigma^2}{n}\left(\frac{2^{H(\pi)}}{e}\right)^s\left(1-\frac{D}{s\beta_0^2}\right)^s$$

in D, we see that the minimum is achieved at

$$D = s\beta_0^2\left(1 - \left(\frac{2e^s n\beta_0^2}{2^{sH(\pi)}\sigma^2}\right)^{\frac{1}{s-1}}\right).$$

Finally, plugging this into equation (6) completes the proof.

□

Lemma B.6. *For any* $0 < p \leq \frac{1}{2}$, *we have* $\left(\frac{p}{1-p}\right)^p \geq e^{-1}$.

Proof. We compute

$$\begin{aligned} \log\left(\frac{p}{1-p}\right) &= -\log\left(\frac{1-p}{p}\right) \\ &= -\log\left(\frac{1}{p}-1\right) \\ &\geq -\frac{1}{p}+2. \end{aligned}$$

As such, we have

$$\left(\frac{p}{1-p}\right)^p = \exp\left(p\log\left(\frac{p}{1-p}\right)\right) \geq e^{-1}.$$

□

C Proofs for covering argument

The primary goal of this section is to prove the following more general version of Theorem 5.1.

Theorem C.1 (Lower bound for additive models on unit length cube)**.** *Assume the regression model* (1)*, with f be defined as in* (2)*, and assume that the covariate space is the unit length cube $[0,1]^d$. Suppose $\phi_j \in C^1([0,1])$ for $j = 1, \dots, d$. Let $I_1, I_2, \dots, I_d \subset [0,1]$ be sub-intervals, and suppose that $\boldsymbol{\beta} \in \mathbb{R}^d$ is a vector of non-negative values such that for each $j = 1, \dots, d$,*

$$\min_{t \in I_j} \left|\phi_j'(t)\right| \geq \beta_j.$$

Denote $\mathcal{K} = \{\mathbf{x} \colon x_j \in I_j \text{ for } j = 1, \dots, d\}$. Assume that ν is a continuous distribution with density q, and denote $q_{min} = \min_{\mathbf{x} \in \mathcal{K}} q(\mathbf{x})$. Define the function $g_{\boldsymbol{\beta}} \colon [0, \max_j \beta_j] \to \mathbb{R}$ via the formula

$$g_{\boldsymbol{\beta}}(\alpha) = \alpha^2 |\{j \colon \beta_j \geq \alpha\}| + \sum_{j \colon \beta_j < \alpha} \beta_j^2, \tag{24}$$

and notice that it is strictly increasing and thus invertible on its domain. Then the oracle expected risk is lower bounded as

$$\mathcal{R}^*(f, \nu, n) \geq \inf_{D > 0} \left\{ D + \frac{\mu(\mathcal{K})\sigma^2}{4n} \prod_{j \colon \beta_j \geq g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))} \left(\frac{\beta_j}{g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))} \right) \right\}, \tag{25}$$

with the convention that $g_{\boldsymbol{\beta}}^{-1}(t) = \infty$ whenever t is out of the range of $g_{\boldsymbol{\beta}}$. In particular, if $\min_{j \in S} \min_{t \in I_j} \left|\phi_j'(t)\right| \geq \beta_0 > 0$ for some subset of indices $S \subset [d]$ of size s, then we have

$$\mathcal{R}^*(f, \nu, n) \geq s\mu(\mathcal{K}) \left(\frac{\beta_0^2 q_{min}}{12} \right)^{s/(s+2)} \left(\frac{\sigma^2}{4n} \right)^{2/(s+2)}. \tag{26}$$

We first work with the uniform measure, and relate the conditional variance over a cell with the weighted sum of its squared side lengths. This is the equivalent of (7) in the rate-distortion argument.

Lemma C.2 (Variance and side lengths)**.** *Let μ be the uniform measure on $[0,1]^d$. Let $\mathcal{C} \subset [0,1]^d$ be a cell. Let f be an additive model as in* (2)*, and assume that each component function ϕ_j is continuously differentiable with $\beta_j \coloneqq \min_{a_j \leq t \leq b_j} \left|\phi_j'(t)\right|$, where a_j and b_j are the lower and upper limits respectively of $\mathcal{C}$ with respect to coordinate j. Then we have*

$$\mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} \geq \frac{1}{6} \sum_{j=1}^d \beta_j^2 (b_j - a_j)^2. \tag{27}$$

Proof. Note that $\phi_1(x_1), \dots, \phi(x_j)$ are independent given the uniform distribution on $\mathcal{C}$. As such, we have

$$\mathrm{Var}\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} = \sum_{j=1}^d \mathrm{Var}\{\phi_j(x_j) \mid \mathbf{x} \in \mathcal{C}\}.$$

We can then further write

$$\mathrm{Var}\{\phi_j(x_j) \mid \mathbf{x} \in \mathcal{C}\} = \frac{1}{2}\mathbb{E}\Big\{ \left(\phi_j(t) - \phi_j(t')\right)^2 \Big\}$$

where t, t' are independent random variables drawn uniformly from $[a_j, b_j]$. For fixed t, t', we use the mean value theorem together with our lower bound on $|\phi_j'|$ to write

$$\left(\phi_j(t) - \phi_j(t')\right)^2 = \phi_j'(\tilde{t})^2 (t - t')^2 \geq \beta_j^2 (t - t')^2.$$

Since the expectation of the right hand side satisfies

$$\mathbb{E}\Big\{ \beta_j^2 (t - t')^2 \Big\} = \frac{\beta_j^2 (b_j - a_j)^2}{6},$$

we immediately obtain (27). □

Next, we use this to compute the maximum volume of a cell under a constraint on its conditional variance. This is similar to the argument in the proof of Theorem 4.2.

Lemma C.3 (Variance and volume)**.** *Assume the conditions of Lemma C.2, and that*

$$\mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} \le D.$$

Then the volume of $\mathcal{C}$ satisfies the upper bound

$$\mu(\mathcal{C}) \le \prod_{j\colon \beta_j \ge g_{\boldsymbol{\beta}}^{-1}(6D)} \left(\frac{g_{\boldsymbol{\beta}}^{-1}(6D)}{\beta_j}\right). \tag{28}$$

Proof. Let $l_j = b_j - a_j$ denote the side length of $\mathcal{C}$ along coordinate j. By Lemma C.2, we have $\sum_{j=1}^d \beta_j^2 (b_j - a_j)^2 \le 6D$. Since $\mu(\mathcal{C}) = \prod_{j=1}^d l_j$, we therefore need to solve the convex optimization problem:

$$\min \prod_{j=1}^d l_j^{-1} \quad \text{s.t.} \quad \sum_{j=1}^d \beta_j^2 l_j^2 \le 6D, \quad l_j \le 1 \text{ for } j = 1, 2, \ldots, d.$$

The Lagrangian of this program is

$$L(l, \lambda) = \prod_{j=1}^d l_j^{-1} + \lambda_0 \left(\sum_{j=1}^d \beta_j^2 l_j^2 - 6D\right) + \sum_{j=1}^d \lambda_j (l_j - 1).$$

Differentiating with respect to l_j, we get

$$\frac{dL}{dl_j} = -l_j^{-2} \prod_{k \ne j} l_k^{-1} + 12\lambda_0 \beta_j^2 l_j + \lambda_j.$$

Let l_j^*, $j = 1, \ldots, d$ and λ_j^*, $j = 0, \ldots, d$ denote the solution to the KKT conditions. Our goal is to solve for the l_j^*'s. The above equation yields

$$(l_j^*)^2 = \frac{1}{12\lambda_0^* \beta_j^2 \prod_{j=1}^d l_j^*} - \frac{\lambda_j^* l_j^*}{12\lambda_0^* \beta_j^2}.$$

Notice that the first term is proportional to β_j^{-2}. Furthermore, by complementary slackness, we have $\lambda_j^* = 0$ or $l_j^* = 1$ for each j. If the former holds, then the second term is equal to 0, so that $l_j^* = \alpha/\beta_j$ for some constant α. Putting everything together, we get

$$l_j^* = \frac{\alpha}{\beta_j} \wedge 1 \tag{29}$$

where α is chosen so that

$$6D = \sum_{j=1}^d \beta_j^2 (l_j^*)^2 = g_{\boldsymbol{\beta}}(\alpha).$$

This implies that

$$\mu(\mathcal{C}) \le \prod_{j=1}^d l_j^* = \prod_{j\colon \beta_j \ge g_{\boldsymbol{\beta}}^{-1}(6D)} \left(\frac{g_{\boldsymbol{\beta}}^{-1}(6D)}{\beta_j}\right)$$

as we wanted. □

The proof of this lemma allows us to interpret the function $g_{\boldsymbol{\beta}}$: The value $g_{\boldsymbol{\beta}}(\alpha)$ is a bound for the weighted sum of optimal squared side lengths, given the choice of α in (29). Hence, for any $D > 0$, $g_{\boldsymbol{\beta}}^{-1}(D)$ is the value of α that ensures that this weighted sum, and hence the conditional variance, is bounded by the value D.

We are now ready to combine these ingredients to prove Theorem C.1. The proof proceeds by first using Lemmas C.4 and C.5 to reduce to the case of uniform measure on a subset. Equipped with Lemma C.3, we can then lower bound the size of the partition using a volumetric argument. More detailed calculations involving the $g_{\boldsymbol{\beta}}$ function yield the second statement.

Proof of Theorem C.1. Let $\mathfrak{p}$ be any permissible partition. Define

$$D \coloneqq \sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}_\nu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\}.$$

The goal is now to find a lower bound on $|\mathfrak{p}|$ in terms of D. To do this, we first transform the above bound via Lemma C.4 and Lemma C.5 to get

$$D \geq q_{min} \sum_{\mathcal{C}\in\mathfrak{p}} \mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\cap\mathcal{K}\}\mu\{\mathcal{C}\cap\mathcal{K}\}.$$

Dividing both sides by $q_{min}\mu(\mathcal{K})$, we get the expression

$$\tilde{D} \geq \sum_{\mathcal{C}\in\mathfrak{p}} \tilde{p}(\mathcal{C})\mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\cap\mathcal{K}\}$$

where $\tilde{D} \coloneqq \frac{D}{q_{min}\mu(\mathcal{K})}$, and the weights $\tilde{p}(\mathcal{C}) \coloneqq \frac{\mu(\mathcal{C}\cap\mathcal{K})}{\mu(\mathcal{K})}$ satisfy $\sum_{\mathcal{C}\in\mathfrak{p}}\tilde{p}(\mathcal{C}) = 1$. By Markov's inequality, we can therefore find a subcollection $\mathfrak{C}\subset\mathfrak{p}$ such that the following two conditions hold:

$$\sum_{\mathcal{C}\in\mathfrak{C}} \tilde{p}(\mathcal{C}) \geq \frac{1}{2} \tag{30}$$

and

$$\mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\cap\mathcal{K}\} \leq 2\tilde{D} \tag{31}$$

for $\mathcal{C}\in\mathfrak{C}$. We now proceed as follows. First, we rewrite (30) as

$$\sum_{\mathcal{C}\in\mathfrak{C}} \mu\{\mathcal{C}\cap\mathcal{K}\} \geq \frac{\mu(\mathcal{K})}{2}.$$

Next, noting that each $\mathcal{C}\cap\mathcal{K}$, being the intersection of two cells, is itself a cell, we can make use of Lemma C.3 and (31) to get

$$\mu(\mathcal{C}\cap\mathcal{K}) \leq \prod_{j\colon \beta_j \geq g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))} \left(\frac{g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))}{\beta_j}\right).$$

Combining the last two statements and rearranging gives

$$|\mathfrak{C}| \geq \frac{\mu(\mathcal{K})}{2} \prod_{j\colon \beta_j \geq g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))} \left(\frac{\beta_j}{g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))}\right). \tag{32}$$

Since the right hand side of (32) is also a lower bound for $|\mathfrak{p}|$, we may plug it into (4) to get the first statement of the theorem.

For the second statement, it is easy to check that $g_{\boldsymbol{\beta}}(\alpha) \geq s\alpha^2$ for $\alpha < \beta_0$, and so we have

$$g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K})) \geq \left(\frac{12D}{sq_{min}\mu(\mathcal{K})}\right)^{1/2}.$$

This implies that

$$\prod_{j\colon \beta_j \geq g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))} \left(\frac{\beta_j}{g_{\boldsymbol{\beta}}^{-1}(12D/q_{min}\mu(\mathcal{K}))}\right) = \left(\frac{s\beta_0^2 q_{min}\mu(\mathcal{K})}{12D}\right)^{s/2}.$$

We plug this into the right hand side of (25) and differentiate to get

$$1 - \frac{s\mu(\mathcal{K})\sigma^2}{4n}\left(\frac{s\beta_0^2 q_{min}\mu(\mathcal{K})}{12}\right)^{s/2} D^{-s/2-1}.$$

Setting this to be zero and solving for D gives

$$D = s\mu(\mathcal{K})\left(\frac{\beta_0^2 q_{min}}{12}\right)^{s/(s+2)}\left(\frac{\sigma^2}{4n}\right)^{2/(s+2)},$$

which gives us the bound we want. □

Proof of Proposition 5.2. Similar to the proof of Lemma C.2, we can show that

$$\mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} \le \frac{\beta_{max}^2}{6}\sum_{j\in S} l_j(\mathcal{C})^2$$

where $l_j(\mathcal{C})$ is the length of $\mathcal{C}$ along coordinate j. As such, if we set

$$l_j = \left(\frac{D}{6s\|q\|_\infty \beta_{max}^2}\right)^{1/2} \wedge 1$$

for $j \in S$ and $l_j = 1$ for $j \notin S$, then we have

$$\mathrm{Var}_\nu\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\} \le D.$$

Tessellating the unit cube with cells of these dimensions gives us a valid partition, whose approximation error is upper bounded by D. We count how many cells are in this partition using a volumetric argument. Each cell has Euclidean volume at least $\left(\frac{D}{6s\|q\|_\infty\beta_{max}^2}\right)^{s/2}$. Meanwhile, it is easy to see that the union of the cells is contained in a rectangular region with side lengths equal to 1 in the coordinates with index $j \notin S$, and equal to 2 in the coordinates with index $j \in S$. This means that there are at most $\left(\frac{24s\|q\|_\infty\beta_{max}^2}{D}\right)^{s/2}$ cells.

Choose

$$D = 24s\|q\|_\infty\beta_{max}^2\left(\frac{\sigma^2}{n}\right)^{2/(s+2)}.$$

Then, the second term in (5) is bounded by

$$\frac{6\sigma^2}{n}\left(\frac{24s\|q\|_\infty\beta_{max}^2}{D}\right)^{s/2} \le \frac{6\sigma^2}{n}\left(\frac{n}{\sigma^2}\right)^{\frac{2}{s+2}\cdot\frac{s}{2}} = 6\left(\frac{\sigma^2}{n}\right)^{2/(s+2)}.$$

Finally, to take care of the error term $E(\mathfrak{p})$, we compute

$$\begin{aligned}(1-\nu\{\mathcal{C}\})^n &= \left(1-\left(\frac{D}{24s\beta_{max}}\right)^{s/2}\right)^n \\ &= \left(1-\left(\frac{\sigma^2}{n}\right)^{s/(s+2)}\right)^n \\ &\le \exp\left(-n\left(\frac{\sigma^2}{n}\right)^{s/(s+2)}\right) \\ &= \exp\left(-\sigma^{2s/(s+2)}n^{2/(s+2)}\right).\end{aligned}$$

This allows us to bound

$$E(\mathfrak{p}) \le \|f\|_\infty^2 \exp\left(-\sigma^{2s/(s+2)}n^{2/(s+2)}\right).$$

Plugging these values into (5) completes the proof. □

Lemma C.4 (Change of measure)**.** *Let ν be a distribution on $[0,1]^d$. Let $\mathcal{C} \subset [0,1]^d$ be a cell such that ν has density $q(x)$ on $\mathcal{C}$ satisfying $\min_{\mathbf{x}\in\mathcal{C}} q(\mathbf{x}) \ge q_{min}$. Then we have*

$$\mathrm{Var}_\nu\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\nu\{\mathcal{C}\} \ge q_{min}\mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x} \in \mathcal{C}\}\mu\{\mathcal{C}\}. \tag{33}$$

Proof. We compute

$$\begin{aligned}\mathrm{Var}_\nu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\nu\{\mathcal{C}\} &= \int_{\mathbf{x}\in\mathcal{C}} (f(\mathbf{x}) - \mathbb{E}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\})^2 q(\mathbf{x})d\mathbf{x} \\ &\geq q_{min}\int_{\mathbf{x}\in\mathcal{C}} (f(\mathbf{x}) - \mathbb{E}\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\})^2 d\mathbf{x} \\ &= q_{min}\mathrm{Var}_\mu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}\}\mu\{\mathcal{C}\}.\end{aligned}$$

□

Lemma C.5 (Restricting to sub-rectangle)**.** *Let ν be a distribution on $[0,1]^d$. Let $\mathcal{C}_1 \subset \mathcal{C}_2 \subset [0,1]^d$ be nested cells. Then*

$$\mathrm{Var}_\nu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}_1\}\nu\{\mathcal{C}_1\} \leq \mathrm{Var}_\nu\{f(\mathbf{x}) \mid \mathbf{x}\in\mathcal{C}_2\}\nu\{\mathcal{C}_2\} \tag{34}$$

Proof. This is proved similarly to the previous lemma. □

D Numerical Experiments

Details of experimental design and algorithm settings for our numerical experiments with tree-based methods fitted to sparse linear models with continuous and Boolean features is shown below. Results are displayed in Figures 2 and 3 for the continuous and Boolean case respectively.

Experimental design: For Figures 2 and 3, we simulate data from a sparse linear generative model $y = \boldsymbol{\beta}^T\mathbf{x}+\epsilon$ with $\mathbf{x} \sim \mathrm{Unif}\big([0,1]^d\big)$ and $\mathbf{x} \sim \mathrm{Unif}\big(\{0,1\}^d\big)$ respectively. In all of the experiments, we varied n, but fixed $d = 50$, $\sigma^2 = 0.01$, and set $\beta_j = 1$ for $j = 1,\ldots,s$, and $\beta_j = 0$ otherwise, where s is a sparsity parameter. We ran the experiments with both $s = 10$ and $s = 20$, and plotted the results for each setting in panel A and panel B respectively for all of the figures. We computed the generalization error using a test set of size 500, averaging the results over 25 runs.

Algorithm settings: The algorithm settings are identical to those as described in Section 5.

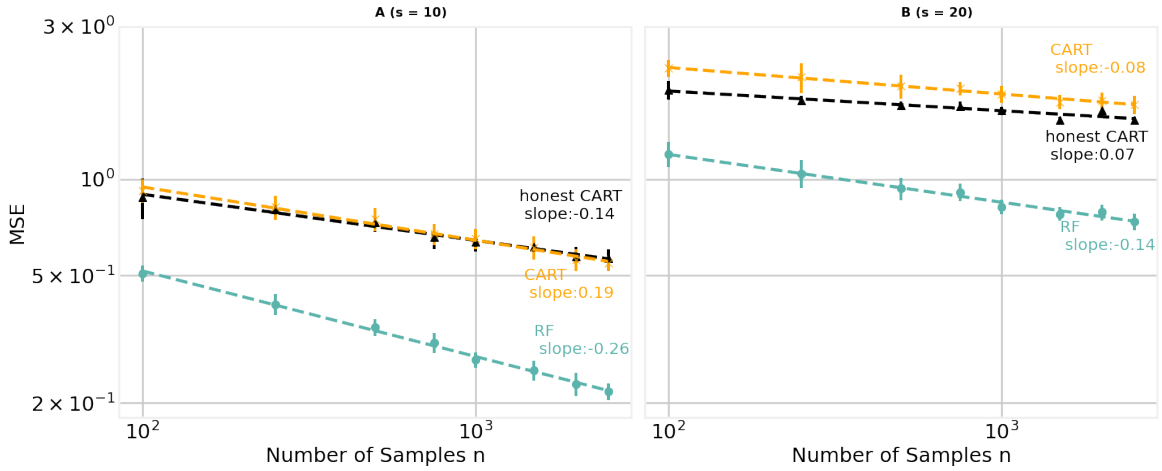

Figure 2: Scaling of the test set error for CART and RF for a sparse linear generative model $y = \boldsymbol{\beta}^T\mathbf{x} + \epsilon$ with $\mathbf{x} \sim \mathrm{Unif}\big([0,1]^d\big)$.We show the scaling with respect to n for **(A)** $s = 10$, and **(B)** $s = 20$.

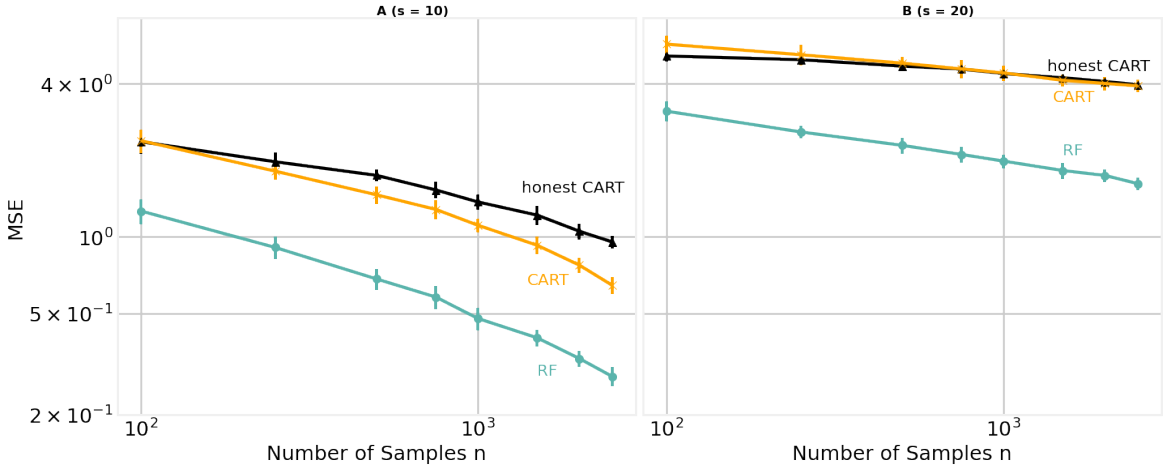

Figure 3: Scaling of the test set error for CART and RF for a sparse linear generative model $y = \boldsymbol{\beta}^T \mathbf{x} + \epsilon$ with $\mathbf{x} \sim \{0,1\}^d$ and each $x_j \sim \text{Ber}(\frac{1}{2})$.We show the scaling with respect to n for **(A)** $s = 10$, and **(B)** $s = 20$.