A tree-based regressor that adapts to intrinsic dimension

Samory Kpotufe[a,*], Sanjoy Dasgupta[b,**]

[a] *Max Planck Institute for Autonomous Systems*

[b] *UCSD Computer Science and Engineering*

Abstract

We consider the problem of *nonparametric regression*, consisting of learning an arbitrary mapping $f : \mathcal{X} \to \mathcal{Y}$ from a data set of (x, y) pairs in which the y values are corrupted by noise of mean zero. This statistical task is known to to be subject to a severe curse of dimensionality: if $\mathcal{X} \subset \mathbb{R}^D$, and if the only smoothness assumption on f is that it satisfies a Lipschitz condition, it is known that any estimator based on n data points will have an error rate (risk) of $\Omega\left(n^{-2/(2+D)}\right)$.

Here we present a tree-based regressor whose risk depends only on the doubling dimension of $\mathcal{X}$, not on D. This notion of dimension generalizes two cases of contemporary interest: when $\mathcal{X}$ is a low-dimensional manifold, and when $\mathcal{X}$ is sparse. The tree is built using random hyperplanes as splitting criteria, building upon recent work of Dasgupta and Freund [DF08]; and we show that axis-parallel splits cannot achieve the same finite-sample rate of convergence.

Keywords: nonparametric regression, notions of dimension, manifold, sparse data

1. Introduction

Given a set of data points (X, Y), where $Y = f(X) + \text{noise}$ (of mean zero), is it possible to infer the unknown function f? This is the problem of *regression*. When f is a linear function, there are simple solutions such as least-square approximations. But what if f is fairly arbitrary – if all that is assumed about it is simply that it is smooth in some sense, for instance that it satisfies a Lipschitz condition[1]? Several families of statistical estimators have been shown to be consistent for this *nonparametric* problem, including kernel and tree-based methods [GKKW02]. However, it is also known that this statistical task is subject to a severe curse of dimensionality: if X is D-dimensional, then the error rate of any estimator f_n based on n samples is $\Omega\left(n^{-2/(2+D)}\right)$ [Sto80, Sto82]. This means that to halve the error, the

*Principal Corresponding Author

**Corresponding Author

Email addresses: samory@tuebingen.mpg.de (Samory Kpotufe), dasgupta@cs.ucsd.edu (Sanjoy Dasgupta)

[1] That is to say, there exists an (unknown) constant λ such that $\|f(x) - f(x')\| \leq \lambda \|x - x'\|$ for all x, x'.

number of samples needs to grow by a multiplicative factor of about 2^D, which is prohibitive even when D is in the low double digits.

This lower bound would appear to rule out nonparametric approaches for the increasingly high-dimensional data sets that arise in modern applications. In image retrieval, or text classification, or genomic analysis, for instance, the number of features, or dimensions, of X can easily grow to tens of thousands, or more. However, in many of these cases, it is believed that the dimensionality is large only in the superficial sense of there being many coordinates, whereas the true degrees of freedom are much smaller in number. This might occur, for example, because of strong dependencies between the features. It is therefore of interest to identify the *intrinsic dimension* of these data sets as the true measure of their complexity.

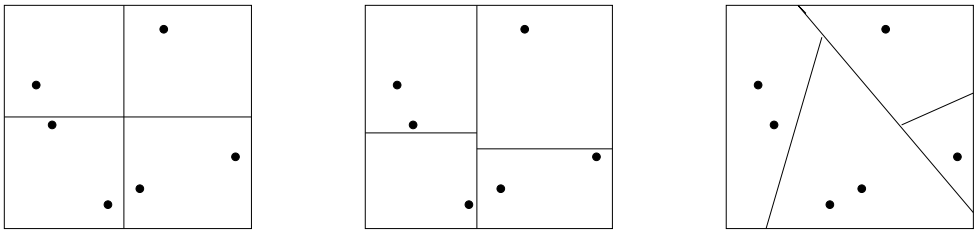
In this paper, we work with a fairly broad such notion, known as *doubling dimension*, and we demonstrate a tree-based regressor whose convergence rate depends only on this quantity rather than on the ambient dimension of the space in which X happens to lie.

1.1. Intrinsic dimension

In what sense might a set of data points in $\mathbb{R}^D$ have an intrinsic dimensionality less than D? To take an example, a speech signal is typically represented by a high-dimensional time series: the signal is broken into overlapping windows, and a variety of filters is applied within each window. Even richer representations can be obtained by using more filters, or by concatenating vectors corresponding to consecutive windows. In this way, the dimensionality D can be made arbitrarily high. However, the physical system can alternatively be described by just a few ($d \ll D$) parameters specifying the configuration of the speaker's vocal apparatus. These are the true degrees of freedom of the data, and as they vary, the high-dimensional representation traces out a d-dimensional submanifold of $\mathbb{R}^D$. A recent trend in machine learning and statistics has been to design algorithms for data that lie on a manifold. Usually the goal is to recover the manifold, or else to obtain a mapping into a lower-dimensional space that preserves key quantities like interpoint distances.

A different type of low-dimensional structure arises in document classification. The most common way of representing a document is as a vector with one coordinate per word, which describes whether or not that word appears in the document (or the number of times the word appears, or some function thereof). The dimensionality D is therefore the size of the vocabulary, which is typically in the tens of thousands. However, any given document only contains a few hundred (or so) words, and thus most of its vector is zero: it is sparse. In a sense, the intrinsic dimension d of the data is the average number of non-zero entries, which is much smaller than D.

There are many ways to formalize intrinsic dimension. We adopt a particular notion called the *doubling dimension*, which is defined for any set of data points in $\mathbb{R}^D$ (or in fact, in any metric space). What makes it particularly attractive is that it generalizes both the notion of manifold dimension and that of sparsity, while at the same time being amenable to the kinds of analysis that arise in algorithm design.



(a) Dyadic tree (b) k-d tree (c) RP tree

Figure 1: Spatial partitioning induced by various splitting rules. Two levels of the tree are shown for each.

1.2. Tree-based regression

A tree-based regression scheme takes as input a data set of n pairs (X, Y), with $X \in \mathbb{R}^D$, and then works (typically) in two phases.

1. It builds a tree T each of whose nodes corresponds to a *cell* (region) of $\mathbb{R}^D$.

   The root node is all of $\mathbb{R}^D$; and each internal node's cell is the disjoint union of the cells of its two children.

2. It prunes the trees to some T', and fits a simple (e.g. constant, or at any rate continuous) function to the data in each leaf of T'.

   The cells corresponding to the leaves of T' are a partition of $\mathbb{R}^D$, and the collection of these local estimates, one per cell, form a piecewise continuous function f_n.

An attractive property of this estimator is that $f_n(x)$ can be evaluated by simply navigating down to the leaf containing x, which takes time proportional to the height of the tree, often just $O(\log n)$. This computational efficiency, and an overall ease of use, have motivated a variety of tree partition methods (Figure 1) such as CART, dyadic trees, and k-d trees [GN05, SN06, DGL96], but none of these has been shown to adapt to intrinsic dimension in its regression risk.

In this paper we build upon the recently-proposed *random projection tree* (RP tree), which uses random hyperplanes to partition space (Figure 1(c)). Previous work has analyzed RP trees for unsupervised learning, and established that they are adaptive to intrinsic dimension when used in this way [DF08, GLZ08, VKD09]. Here we explore their use in regression.

A random projection tree is built using successive hyperplane splits that yield increasingly fine partitions of $\mathbb{R}^D$ into convex cells. At what point should the splitting process be stopped, and the resulting cells used to fit local regressors (step 2 of the template above)? The right granularity is one that properly balances the bias of the estimator (which favors smaller cells) with the variance (which favors large cells). Traditional methods grow a tree fully (to the point where the leaves contain a single data point, say) and then use penalized empirical risk minimization over all possible partitions induced by the tree. Our approach can be more efficient in practice. We grow the tree in blocks, rather than just one node or level at a

time; this severely limits the number of candidate partitions. We then provide two options for selecting the final partition:

(a) An automatic rule for when to stop growing the tree.

This is based entirely on observable quantities like the diameters of cells.

(b) Cross-validation over the candidate partitions.

This chooses the partition with lowest regression risk on a held-out data set.

The former method is computationally cheaper, while the latter gives a slightly better risk bound. In both cases, we show that the excess risk of the RP tree regressor depends only on the (unknown) doubling dimension of the input space, no matter what the distribution of data.

We introduce novel tools for the analysis of bias. In the literature, the bias of a tree estimator is typically analyzed in terms of the physical diameter of its cells (see, for instance, Chapter 20 of [DGL96]). However, this can worked out only when the cells have simple shapes like hyper-rectangles. The cells of an RP tree are irregular convex polytopes, and their diameters might not systematically decrease while moving down the tree. What we do instead is to track the diameter of the *data* within each cell, and we develop new techniques to relate these empirical *data diameters* to the estimator's bias. Our method takes the focus away from the cells' physical diameters, opening the door to richer partitioning rules with nontrivial cell structure.

1.3. Related work

There are many types of high-dimensional data, like the speech example above, that are likely to lie near a low-dimensional manifold because of either physical or geometric constraints. This realization has galvanized the field of *manifold learning*, which seeks to transform data from $\mathbb{R}^D$ to a lower-dimensional space while preserving important structure; key early results are [RS00, BN03, TSL00]. These embedding methods can be used as a prelude to regression: the regressor will then operate in the lower-dimensional space where it might perform better. However, this approach does not easily yield theoretical guarantees for distribution-free regression. Our interest is in circumventing the embedding step and automatically adapting to low intrinsic dimension while operating in the original space $\mathbb{R}^D$.

Recent work of Bickel and Li [BL07] has shown that local kernel regressors are adaptive to manifold structure. Specifically, they obtain a bandwidth setting under which the asymptotic risk at any given point in $\mathbb{R}^D$ depends only on the manifold dimension and on the behavior of the kernel in the vicinity of that point. The appropriate bandwidth can be found either by estimating the manifold dimension or by cross-validating over possible values of this dimension [BL07, LW07].

Earlier work of Kulkarni and Posner [KP95], although not treating the topic of adaptivity to intrinsic dimension, expresses the risk of nearest neighbor regression in terms of the *box dimension* [Cla05] of the data, which is related to the doubling dimension.

A disadvantage of kernel and nearest neighbor regressors is that they are expensive to evaluate on a new data point. Either kernel weights must be computed at many training

points, resulting in an $\Omega(n)$ evaluation time, or the k_n nearest neighbors of a query point must be located, where k_n is optimally chosen as a root of n [GKKW02]. This sort of time complexity can be a burden in practice considering that nonparametric regression usually depends upon large data sizes n for accuracy. Hence the appeal of an adaptive tree-based regressor that can be evaluated in $O(\log n)$ time.

For classification problems, Scott and Nowak [SN06] have shown that dyadic decision trees (Figure 1(a)) achieve convergence rates depending only on (something like) the box dimension, under smoothness conditions on the input distribution and the Bayes decision boundary. We show later in the paper that the risk of a regressor based on dyadic partitioning does depend on D, but that this dependence appears in a leading constant (exponential in D) rather than in the exponent of n.

The random regression graph of Caponnetto and Smale [CS07] is similar in spirit to an RP tree since it also partitions space using random hyperplanes. However, its regression risk has only been analyzed in terms of a quantity that is different from the kind of intrinsic dimension we consider here: the norm of the regression function in the reproducing kernel Hilbert space induced by a specific kernel. In particular, our notion of dimension involves only the predictor variable X and not the response Y.

2. Detailed overview of results

Suppose each data point is of the form (X, Y), where the predictor X lies in a space $\mathcal{X} \subset \mathbb{R}^D$ and the response Y lies in a space $\mathcal{Y} \subset \mathbb{R}^{D'}$. The distance measure we will use in these spaces is the ℓ_2 (Euclidean) norm. Our rates of convergence depend upon the diameter of $\mathcal{X}$ and of $\mathcal{Y}$; to quantify these we assume that the two spaces lie within balls of (unknown) diameter $\Delta_{\mathcal{X}}$ and $\Delta_{\mathcal{Y}}$, respectively.

2.1. Doubling dimension

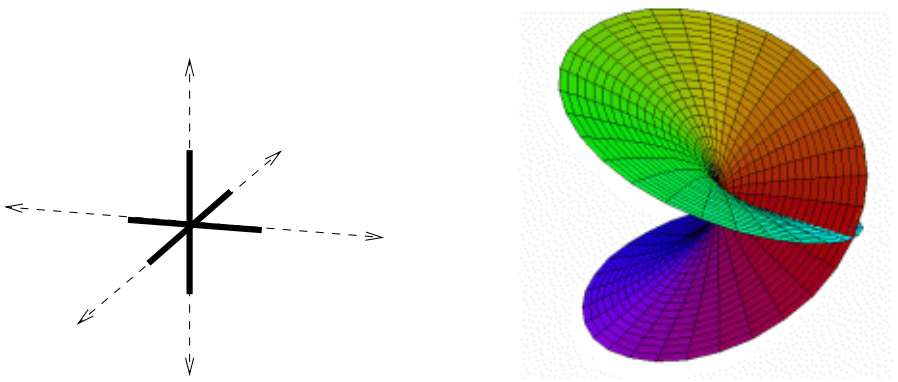
We capture the intrinsic dimensionality of $\mathcal{X}$ by its doubling dimension [GKL03], which is defined for any metric space, but is here specialized to Euclidean spaces.

Definition 1. *The doubling dimension of $\mathcal{X} \subset \mathbb{R}^D$ is the smallest d such that for any ball $B \subset \mathbb{R}^D$, the set $B \cap \mathcal{X}$ can be covered by 2^d balls of half the radius of B.*

Consider, for instance, a line S in a high-dimensional space $\mathbb{R}^D$. For any ball $B \subset \mathbb{R}^D$, the intersection of S and B is a line segment, and this segment can be covered by two balls whose radii are half that of B. Therefore the doubling dimension of S is 1. Something similar holds for any affine subspace of $\mathbb{R}^D$:

Lemma 2. *[Cla05] There is a universal constant $c_o < 3$ such that for any $d < D$, a d-dimensional affine subspace of $\mathbb{R}^D$ has doubling dimension at most $c_o d$.*

A set of n points can always be covered by n balls, and therefore has doubling dimension at most $\log n$ (where the logarithm is taken base two). This is easily generalized:



(a) Sparse data set. (b) 2-dimensional manifold.

Figure 2: Examples of data with low doubling dimension.

Lemma 3. *Suppose sets $S_1, \ldots, S_n$ each have doubling dimension $\leq d$. Then $S_1 \cup \cdots \cup S_n$ has doubling dimension at most $d + \log n$.*

Proof. Pick any ball B; by hypothesis $B \cap S_i$ can be covered by 2^d balls of half the radius. Therefore $B \cap (S_1 \cup \cdots \cup S_n)$ can be covered by $n \cdot 2^d$ such balls. □

The previous two lemmas yield a bound on the doubling dimension of any sparse set.

Lemma 4. *Suppose that $S \subset \mathbb{R}^D$ is k-sparse: that is, each point in S has at most k nonzero coordinates. Then S has doubling dimension at most $c_o k + k \log D$.*

Proof. S is contained within the union of $\binom{D}{k} \leq D^k$ subspaces of dimension k: pick which k coordinates, out of D, will be nonzero, and consider the subspace in which the remaining coordinates are forced to zero. By Lemma 2, each of these subspaces has doubling dimension at most $c_o k$. Lemma 3 then bounds the increase in dimension from taking the union of the subspaces. □

Thus the doubling dimension captures sparse data, a subject of significant contemporary interest. What about manifold data? Here the situation is slightly more subtle. Although it is intuitively sensible in many situations to suppose that data lie on (or close to) a low-dimensional manifold, this is not of much help, algorithmically or statistically, unless the manifold has bounded curvature; a space-filling 1-dimensional curve, for instance, is just as bad as a full-dimensional data set. Recent work [NSW08] has identified a clean way to capture curvature by a single value called the *condition number* of the manifold. When this is bounded, neighborhoods of the manifolds are sufficiently flat that they can be shown to have low doubling dimension.

Lemma 5. *[DF08] If a d-dimensional Riemannian submanifold of $\mathbb{R}^D$ has bounded condition number $\kappa < \infty$, then its neighborhoods of radius $< 1/\kappa$ have doubling dimension $O(d)$.*

2.2. Other notions of dimension

The problem of identifying the intrinsic dimension of a set has arisen in many different scientific communities, and has produced a variety of definitions. Where does doubling dimension lie in this panorama? Some insight can be obtained by situating it with respect to three other successful notions of dimensionality, arranged here in decreasing order of generality: covering dimension, manifold dimension, and affine dimension. It turns out that the doubling dimension lies somewhere between the first two.

The most general is the *covering dimension* of a set $\mathcal{X}$: the smallest d for which there is a constant $C > 0$ such that for any $\epsilon > 0$, $\mathcal{X}$ has an ϵ-cover of size $C(1/\epsilon)^d$. This notion lies at the heart of much of empirical process theory. Although it permits many kinds of analysis and is very general, for our purposes it falls short on one count: in nonparametric estimation, we need small covering numbers not just for $\mathcal{X}$, but also for individual *neighborhoods* of $\mathcal{X}$. Thus we would like this same covering condition (with the same constant C) to hold for all Euclidean balls within $\mathcal{X}$. This additional stipulation yields the *doubling dimension* as defined above. The following two trivial lemmas summarize this connection.

Lemma 6. *If $\mathcal{X}$ has diameter C and doubling dimension d, then for any $\epsilon > 0$, it has an ϵ-cover of size at most $(2C/\epsilon)^d$.*

Proof. Applying the doubling condition recursively, $\mathcal{X}$ can be covered by one ball of radius C, 2^d balls of radius $C/2$, 2^{2d} balls of radius $C/4$, and so on. □

Lemma 7. *If $\mathcal{X}$ has doubling dimension d, then so does any subset of $\mathcal{X}$.*

At the bottom end of the spectrum is the *affine dimension*, which is simply the smallest d such that $\mathcal{X}$ is contained in a d-dimensional affine subspace of $\mathbb{R}^D$. It is a tall order to expect this to be smaller than D, although we may hope that $\mathcal{X}$ lies close to such a subspace. A more general hope is that $\mathcal{X}$ lies on (or close to) a d-dimensional Riemannian submanifold of $\mathbb{R}^D$. As we have seen, under suitable curvature conditions, this notion is less general than doubling dimension (at least when limited to small enough neighborhoods). In fact, the containment is strict: there is a substantial gap between manifolds of bounded curvature and sets of low doubling dimension, on account of the smoothness properties of the former. This divide is not just a technicality but has important algorithmic implications. For instance, a variant of the Johnson-Lindenstrauss lemma [JL84] states that when a d-dimensional manifold (of bounded curvature) is projected onto a random subspace of dimension $O(d/\epsilon^2)$, then all interpoint distances are preserved within $1 \pm \epsilon$ [Cla05]. For sets of doubling dimension d, however, no such guarantee can be given: an arbitrarily high-dimensional range space might be needed [IN07].

Recent work [DF08] introduced a notion called the *local covariance dimension*, meant specifically for data analysis. The definition formalizes a type of "local flatness" and tries to capture the intuition that a data set might have low intrinsic dimension only when limited to neighborhoods that are small enough (Figure 4). We consider a similar extension of our results in the appendix.

2.3. Statistical setup

Our results are in the framework of *statistical learning theory*, which posits an (unknown) underlying distribution over joint predictor-response space $\mathcal{X} \times \mathcal{Y}$. All data points are drawn independently at random from this distribution. Let μ be the marginal distribution over $\mathcal{X}$. Recall we are assuming $\mathcal{X} \subset \mathbb{R}^D$ and $\mathcal{Y} \subset \mathbb{R}^{D'}$.

In nonparametric regression, the target function is

$$f(x) \doteq \mathbb{E}\left[Y|X=x\right].$$

The rate of convergence to f depends inevitably on how smooth it is, and there are a variety of ways in which this can be quantified. Here we simply assume that f is λ-Lipschitz for some unknown parameter λ:

$$\forall x, x' \in \mathcal{X},\ \|f(x) - f(x')\| \leq \lambda \|x - x'\|.$$

Suppose $g : \mathcal{X} \to \mathcal{Y}$ is some estimate of f. We define its l_2 *pointwise risk at* x to be $R(g,x) \doteq \mathbb{E}_{Y|X=x} \|Y - g(x)\|^2$ and its *integrated risk* to be $R(g) \doteq \mathbb{E}_X R(g,X)$. Standard manipulations show that

$$\begin{aligned} R(g,x) &= R(f,x) + \|f(x) - g(x)\|^2 \\ R(g) &= R(f) + \mathbb{E}_X \|f(X) - g(X)\|^2. \end{aligned}$$

Thus, the pointwise excess risk of $g(x)$ over $f(x)$ is simply $\|f(x) - g(x)\|^2$. In this paper we are interested in the *integrated excess risk*

$$\|f - g\|^2 \doteq R(g) - R(f) = \mathbb{E}_X \|f(X) - g(X)\|^2.$$

Suppose the training set consists of n points $(X_1, Y_1), \ldots, (X_n, Y_n)$; denote these collectively by $(\mathbf{X}, \mathbf{Y})$. This data set defines an *empirical distribution* which assigns mass $1/n$ to each of these n support points. Let μ_n be the marginal empirical distribution over $\mathcal{X}$.

2.4. Notions of diameter

Based on the training set, we will construct a partition $\mathcal{A}$ of $\mathcal{X}$ (or more precisely, of $\mathbb{R}^D$, since $\mathcal{X}$ is unknown), and we will build a piecewise-constant estimator $f_{n,\mathcal{A}}$ on the cells of this partition. It is standard to decompose the error of the estimator into two parts.

$$\begin{aligned} \text{bias} &\equiv \text{how much does } f \text{ vary within a single cell?} \\ \text{variance} &\equiv \text{what is the error in estimating the mean value of } f \text{ within a cell?} \end{aligned}$$

The bias can controlled by making sure cells are small. The variance can be controlled by making sure cells are large enough that they contain many data points. A lot of the novelty of our approach arises from the particular way in which we define the size of a cell.

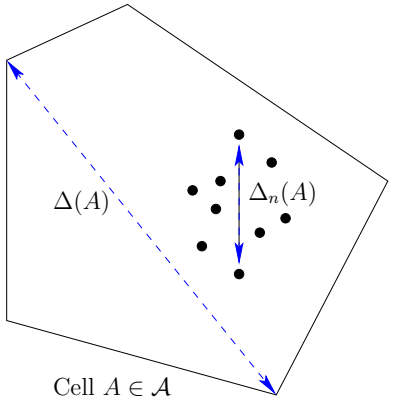
Traditionally, the analysis of bias is based on the *physical diameters* of cells $A \in \mathcal{A}$,

$$\Delta(A) \doteq \max_{x,x' \in A} \|x - x'\|$$

(see, for instance, [GN05, SN06, DGL96]). In this work we instead relate bias to the *data diameters* of the cells,

$$\Delta_n(A) \doteq \max_{x,x' \in A \cap \mathbf{X}} \|x - x'\|$$

(or 0 if $A \cap \mathbf{X}$ is empty); recall that $\mathbf{X}$ is the set of data points.



We'll see, in fact, that in order to bound the bias of the estimator, we don't need all cells of a partition to have small data diameter, but merely for these diameters to be small in an average sense. For a collection $\mathcal{A}$ of disjoint subsets of $\mathcal{X}$, we use the following notion of average data diameter:

$$\Delta_n(\mathcal{A}) \doteq \sqrt{\frac{\sum_{A \in \mathcal{A}} \mu_n(A) \Delta_n^2(A)}{\sum_{A \in \mathcal{A}} \mu_n(A)}}.$$

By focusing on data diameter, we are no longer constrained to the kinds of highly regular cells (like hyper-rectangles) whose physical diameters are amenable to analysis. Instead, we move towards richer partitioning schemes which may adapt better to intrinsic dimension.

2.5. Limitations of axis parallel splitting rules

Consider the data space depicted in Figure 2(a),

$$S = \bigcup_{i=1}^{D} \{te_i : -1 \leq t \leq 1\}$$

where e_i is the unit vector in the ith coordinate direction. S is an extreme case of a sparse data set: each point in it has at most one nonzero coordinate. It is not hard to see that tree structures with axis-parallel splits (such as k-d trees and dyadic trees) would require at least D levels to halve the diameter of S; that is, any tree with fewer levels would contain leaf cells of diameter greater than one. Thus halving the diameter would require 2^D data points, which is prohibitive for large D.

However, by using a richer class of splits, cell size can be decreased a lot quicker. By Lemma 3, S has doubling dimension $d \leq 1 + \log D$, and it is shown in [DF08] that an RP tree halves the diameter in just $O(d \log d)$ levels, no matter what the distribution over the data space.

This example suggests that, depending on the distribution μ on $\mathcal{X}$, regression based on axis-parallel cells might require a data size (n) exponential in D in order to attain low risk, whereas regression based on RP splits might do better, requiring resources that depend just on the intrinsic dimension d. However, there is an interesting subtlety. We show in Theorem 24 (appendix) that the excess risk of a dyadic tree regressor depends on D only in the form of a leading constant 2^D, and not in the exponent of n. That is, for $n \geq 2^D$, the risk looks like $O\left(n^{-2/(2+d)}\right)$. This is a curse of dimensionality that emerges in a finite-sample analysis but not necessarily in an asymptotic analysis. All our results on RP tree regression in this paper are finite-sample convergence rates which depend just on d even for small n.

2.6. Building the regression tree

A tree-based regressor works in two phases.

1. The data space is split into some partition $\mathcal{A}$.
2. A regressor is learned as a piecewise continuous function over the cells of $\mathcal{A}$.

In this work we'll consider a piecewise constant regressor over $\mathcal{A}$, defined as follows: for any $x \in \mathcal{X}$, let $\mathcal{A}(x)$ be the cell of $\mathcal{A}$ to which x belongs, and set

$$f_{n,\mathcal{A}}(x) \doteq \frac{\sum_{i=1}^{n} Y_i \cdot \mathbf{1}\left(X_i \in \mathcal{A}(x)\right)}{n \cdot \mu_n(\mathcal{A}(x))}$$

if $\mu_n(\mathcal{A}(x)) > 0$ (that is, if the cell $\mathcal{A}(x)$ contains at least one training point). If $\mathcal{A}(x) \cap \mathbf{X}$ is empty, then a default setting $f_{n,\mathcal{A}}(x) = y_o$ is used instead, for some $y_o \in \mathcal{Y}$. We will often refer to the final regressor as f_n when the partition $\mathcal{A}$ used for the estimate is clear from context.

The first phase of the regression algorithm implicitly builds a tree, each of whose nodes corresponds to a region of $\mathbb{R}^D$. Each node has two children whose regions are a partition of its own. We will also associate each such cell A with the data points $A \cap \mathbf{X}$ that happen to fall in it.

All the splitting is done by random hyperplanes, and thus each cell is a convex region of $\mathbb{R}^D$. The precise details are deferred to section 5.1; all we need to know at present is that there is a subroutine `coreRPtree` that operates as follows:

- It takes as input a region $A \subset \mathbb{R}^D$ (or more precisely, the data points that fall in this region).
- By recursive splits, it builds a tree whose root corresponds to A and whose leaves all have data diameter at most half that of A.

Procedure `adaptiveRPtree`(*sample* $\mathbf{X} \subset \mathbb{R}^D$*, confidence parameter* δ)

$\mathcal{A}^0 \leftarrow \mathbb{R}^D$;
for $i \leftarrow 1$ **to** ∞ **do**
  foreach cell $A \in \mathcal{A}^{i-1}$ **do**
    (subtree rooted at A) $\leftarrow$ `coreRPtree` $(A,\ \Delta_n(A)/2, \delta)$;
  end
  $\mathcal{A}^i \leftarrow$ partition of $\mathbb{R}^D$ defined by the leaves of the current tree;
  level $\left(\mathcal{A}^i\right) \leftarrow \max_{A \in \mathcal{A}^i}$ level (A) ; `// level = depth in tree`
  `// There are two options for stopping and returning a partition.`
  `Option 1: Cross-validation`
  if $\Delta_n\left(\mathcal{A}^i\right) = 0$ or level $\left(\mathcal{A}^i\right) \geq 2 \log n$ **then**
    Define $R'_n(\cdot)$ as the empirical risk on a validation sample $(\mathbf{X}', \mathbf{Y}')$ of size n;
    $\mathcal{A}^* \leftarrow \underset{\mathcal{A} \in \{\mathcal{A}^0, \ldots, \mathcal{A}^i\}}{\operatorname{argmin}} R'_n(f_{n,\mathcal{A}})$;
    return $f_n \doteq f_{n,\mathcal{A}^*}$;
  end
  `Option 2: Automatic stopping`
  $\alpha(n) \leftarrow \left(\log^2 n\right) \log\log(n/\delta) + \log(1/\delta)$;
  if $\Delta_n^2\left(\mathcal{A}^i\right) \leq \Delta_n^2\left(\mathcal{A}^0\right) \cdot (\alpha(n)/n) \cdot 2^{\text{level}(A^i)}$ **then**
    $\mathcal{A}^* \leftarrow \underset{\mathcal{A} \in \{\mathcal{A}^{i-1}, \mathcal{A}^i\}}{\operatorname{argmin}} \left(\frac{\alpha(n)}{n} \cdot |\mathcal{A}| + \Delta_n^2(\mathcal{A})\right)$;
    return $f_n \doteq f_{n,\mathcal{A}^*}$;
  end
end

- If A has zero diameter (for instance, if it contains one point), then the procedure leaves it untouched. Otherwise, a tree is returned whose leaves contain at most $\lceil |A \cap \mathbf{X}|/2 \rceil$ points.

The main tree building algorithm is Procedure `adaptiveRPtree`. It starts with a single node $\mathcal{A}^0$ for all of $\mathbb{R}^D$, and then grows a tree in measured steps. At each stage, the current set of leaves constitute a partition $\mathcal{A}^i$ of $\mathbb{R}^D$, whose cells have diameter $\Delta_n(\mathcal{A}^i) \leq 2^{-i}\Delta_n(\mathbb{R}^D)$. Then the subroutine `coreRPtree` is called on each leaf to yield an even finer partition $\mathcal{A}^{i+1}$.

This process is stopped when each cell of the current partition is sufficiently small that the bias is controlled, but also has sufficiently many data points in it that the variance is controlled. How can the right stopping point be identified? We present two options.

1. *Automatic stopping.* We return a partition as soon as the data diameters of cells are small enough relative to tree size.

2. *Cross-validation.* Here, we grow a large tree and then prune it using a separate validation sample $(\mathbf{X}', \mathbf{Y}')$, also of size n, drawn from the same underlying distribution.

To prune, the intermediate partition $\mathcal{A}^i$ is chosen which minimizes the empirical risk

$$R'_n(g) \doteq \frac{1}{n} \sum_{i \in [n]} \|Y'_i - g(X'_i)\|^2 .$$

The automatic stopping option requires no validation sample and is computationally faster. As we'll see, its risk bound is only slightly worse than that of the cross-validation option.

Regardless of which stopping rule is employed, it follows from the properties of `coreRPtree` that the final tree has height at most $2 \log 2n$ and the number of partitions $\mathcal{A}^i$ generated is at most $\log 2n$.

2.7. Main Results

The excess risk of the tree-based regressor can be expressed in terms of the rate at which diameters decrease from the root down. We have the following definition.

Definition 8. *Given a sample* $\mathbf{X}$*, we say that* `coreRPtree` *attains a diameter decrease rate of* k *on* $\mathbf{X}$ *for* $k \geq d$*, if every call to it in the second loop of the main procedure* `adaptiveRPtree` *returns a tree of depth at most* k*.*

Recent work [DF08] shows that by using RP trees, a diameter decrease rate of $O(d \log d)$ can be achieved, where d is the doubling dimension of $\mathcal{X}$. Building upon this result, we have the following main theorem.

Theorem 9. *Assume that* $\mathcal{X}$ *has doubling dimension* d*. There exist constants* C, C' *independent of* d *and* μ*, such that the following hold. Pick any* $\delta > 0$ *and define*

$$\alpha(n) \doteq \left(\log^2 n\right) \log\log(n/\delta) + \log(1/\delta).$$

With probability at least $1 - \delta$*:*

(a) `coreRPtree` *attains a diameter decrease rate of* $k \leq C' d \log d$*.*

(b) If the automatic stopping option is used, the excess risk of the regressor is

$$\|f_n - f\|^2 \leq C \cdot \left(\Delta_{\mathcal{Y}}^2 + \lambda^2\right) \left(\Delta_{\mathcal{X}}^2 + 1\right) \cdot \left(\frac{\alpha(n)}{n}\right)^{2/(2+k)} .$$

(c) If the cross-validation option is used and $n \geq \max\left\{\left(\lambda \Delta_{\mathcal{X}} / \Delta_{\mathcal{Y}}\right)^2, \alpha(n)\right\}$*, then the excess risk of the regressor is*

$$\|f_n - f\|^2 \quad \leq \quad C \cdot \left(\lambda \Delta_{\mathcal{X}}\right)^{2k/(2+k)} \left(\frac{\Delta_{\mathcal{Y}}^2 \cdot \alpha(n)}{n}\right)^{2/(2+k)} + 2\Delta_{\mathcal{Y}}^2 \sqrt{\frac{\log\log n + \log 8/\delta}{2n}}.$$

The two stopping options yield similar bounds in terms of the dependence on n and d; however the cross-validation bound has a better dependence on λ, $\Delta_{\mathcal{X}}$, and $\Delta_{\mathcal{Y}}$.

In section 3, we lay out the key tools for the rest of the analysis, culminating in a risk bound in terms of data diameter. In section 4, we investigate the two stopping rules, and bound the excess risk of the final regressor in terms of the observed diameter decrease rate. And in section 5, we show that these decrease rates depend only on the intrinsic dimensionality of the data.

The algorithm takes an input a permissible failure probability δ. There are three sources of failure, and we apportion each of them a $\delta/3$ probability: failure to build a tree with the desired height and diameter decrease rate; an $(\mathbf{X}, \mathbf{Y})$ sampling failure in which either the empirical masses of cells do not accurately represent their true masses or the y-values within cells have non-representative averages; and an $(\mathbf{X}', \mathbf{Y}')$ sampling failure in the cross-validation step.

Parts (a), (b), and (c) of Theorem 9 result from Corollary 23, Lemma 20, and Lemma 18 respectively.

3. Risk bound for $f_{n,\mathcal{A}}$

In this section we develop the necessary tools to bound the excess risk of $f_{n,\mathcal{A}}$, where $\mathcal{A}$ is an RP tree partition, that is, $\mathcal{A}$ is defined by the leaves of the tree returned by `adaptiveRPtree`.

3.1. Generic decomposition of pointwise excess risk

We start the analysis with a standard decomposition of the excess risk into bias and variance terms. Let $\mathcal{A}$ be any partition of $\mathcal{X}$, on which the regressor $f_{n,\mathcal{A}}$ is defined. Recall that we denote by $\mathcal{A}(x)$ the cell of $\mathcal{A}$ containing x.

A useful intermediary between $f_{n,\mathcal{A}}$ and the target f is the following function on $\mathcal{X}$:

$$\widetilde{f}_{n,\mathcal{A}}(x) \doteq \frac{\sum_{i=1}^{n} f(X_i)\mathbf{1}\left(X_i \in \mathcal{A}(x)\right)}{n\mu_n(\mathcal{A}(x))}$$

if $\mu_n(\mathcal{A}(x)) \neq 0$; otherwise $\widetilde{f}_{n,\mathcal{A}}(x) = y_o \in \mathcal{Y}$. Notice that both $f_{n,\mathcal{A}}$ and $\widetilde{f}_{n,\mathcal{A}}$ are constant within any cell $A \in \mathcal{A}$; we will therefore overload notation and occasionally write these quantities as $f_{n,\mathcal{A}}(A)$ and $\widetilde{f}_{n,\mathcal{A}}(A)$, respectively.

The pointwise excess risk at x can be bounded as

$$\begin{aligned} \left\|f_{n,\mathcal{A}}(x) - f(x)\right\|^2 &\leq \left(\left\|f_{n,\mathcal{A}}(x) - \widetilde{f}_{n,\mathcal{A}}(x)\right\| + \left\|\widetilde{f}_{n,\mathcal{A}}(x) - f(x)\right\|\right)^2 \\ &\leq 2\underbrace{\left\|f_{n,\mathcal{A}}(\mathcal{A}(x)) - \widetilde{f}_{n,\mathcal{A}}(\mathcal{A}(x))\right\|^2}_{\text{variance}} + 2\underbrace{\left\|\widetilde{f}_{n,\mathcal{A}}(x) - f(x)\right\|^2}_{\text{bias}^2}. \end{aligned} \quad (1)$$

In the next two lemmas, we separately bound the variance and the bias.

Lemma 10 (Variance)**.** *Fix any partition $\mathcal{A}$ and any set of n points $\mathbf{X} = \{X_1, \ldots, X_n\} \subset \mathcal{X}$. Suppose the Y_i are now drawn according to their conditional distribution given X_i. Pick any $\delta > 0$. Then with probability at least $1 - \delta$, for every cell $A \in \mathcal{A}$ with $\mu_n(A) > 0$:*

$$\left\| f_{n,\mathcal{A}}(A) - \widetilde{f}_{n,\mathcal{A}}(A) \right\|^2 \leq \Delta_{\mathcal{Y}}^2 \cdot \frac{2 + \ln(|\mathcal{A}|/\delta)}{n\mu_n(A)}.$$

Proof. For any cell $A \in \mathcal{A}$, let $I(A) = \{1 \leq i \leq n : X_i \in A\}$ be the indices of points falling in that cell. Then $\mu_n(A) = |I(A)|/n$, and

$$\left\| f_n(A) - \widetilde{f}_n(A) \right\| \;=\; \left\| \frac{1}{|I(A)|} \sum_{i \in I(A)} (Y_i - f(X_i)) \right\|.$$

Changing any single Y_i value alters this expression by at most $\Delta_{\mathcal{Y}}/|I(A)|$. We can therefore use McDiarmid's inequality to assert that with probability at least $1 - \delta/|\mathcal{A}|$ over the choice of the Y_i's,

$$\left\| f_n(A) - \widetilde{f}_n(A) \right\| \leq \mathbb{E} \left\| f_n(A) - \widetilde{f}_n(A) \right\| + \Delta_{\mathcal{Y}} \cdot \sqrt{\frac{\ln(|\mathcal{A}|/\delta)}{2|I(A)|}}.$$

The expectation can be bounded as follows:

$$\begin{aligned}
\mathbb{E} \left\| f_n(A) - \widetilde{f}_n(A) \right\| &\leq \left(\mathbb{E} \left\| f_n(A) - \widetilde{f}_n(A) \right\|^2 \right)^{1/2} \\
&= \frac{1}{|I(A)|} \left(\mathbb{E} \left\| \sum_{i \in I(A)} (Y_i - f(X_i)) \right\|^2 \right)^{1/2} \\
&= \frac{1}{|I(A)|} \left(\sum_{i \in I(A)} \mathbb{E} \left\| Y_i - f(X_i) \right\|^2 \right)^{1/2} \\
&\leq \frac{1}{|I(A)|} \left(|I(A)| \Delta_{\mathcal{Y}}^2 \right)^{1/2} \;=\; \frac{\Delta_{\mathcal{Y}}}{\sqrt{|I(A)|}}.
\end{aligned}$$

The first line uses Jensen's inequality. The third uses the fact that the vectors $v_i = Y_i - f(X_i)$ are independent random vectors with zero expectation, so that $\mathbb{E} \left\| \sum_i v_i \right\|^2 = \sum_i \mathbb{E} \left\| v_i \right\|^2$.

We conclude with a union bound over all nonempty $A \in \mathcal{A}$. □

Lemma 11 (Bias)**.** *Fix any partition $\mathcal{A}$ and any set of n points $\mathbf{X} = \{X_1, \ldots, X_n\} \subset \mathcal{X}$. For any $x \in \mathcal{X}$ with $\mu_n(\mathcal{A}(x)) > 0$,*

$$\left\| \widetilde{f}_{n,\mathcal{A}}(x) - f(x) \right\| \leq \lambda \cdot \Delta\left(\mathcal{A}(x)\right).$$

Proof. Let $A = \mathcal{A}(x)$, so that

$$\begin{aligned}\left\|\widetilde{f}_{n,\mathcal{A}}(x) - f(x)\right\| &= \left\|\frac{\sum_{i=1}^n (f(X_i) - f(x))\mathbf{1}\left(X_i \in A\right)}{n\mu_n(A)}\right\| \\ &\leq \frac{\sum_{i=1}^n \|f(X_i) - f(x)\|\,\mathbf{1}\left(X_i \in A\right)}{n\mu_n(A)} \\ &\leq \frac{\sum_{i=1}^n \lambda\,\|X_i - x\|\,\mathbf{1}\left(X_i \in A\right)}{n\mu_n(A)} \leq \lambda \cdot \Delta\left(A\right),\end{aligned}$$

where the second inequality uses the Lipschitz condition on $f(\cdot)$. □

What we have at this point is a fairly standard bias-variance decomposition of the risk. It contains two quantities that are nontrivial to bound in our context: the empirical weights of cells, $\mu_n(A)$; and, more importantly, their physical diameters $\Delta(A)$.

To relate the empirical masses $\mu_n(A)$ to their true values $\mu(A)$, we could use a uniform large deviation bound. A naive such bound would involve terms in D, since each cell is an intersection of hyperplanes. To avoid such a dependency, we make heavy use of the fact that the *directions* of the hyperplanes are chosen at random, independent of the data points, and that the data are consulted only to determine the *displacements* of the boundaries along these directions.

The bigger challenge is to handle cell diameters. The bound on bias involves the physical diameters $\Delta(A)$ of cells, and these might not decrease gracefully down the tree. So we create an alternate partition $\mathcal{A}'$ with the following properties:

- Each cell of $\mathcal{A}$ is the union of two cells of $\mathcal{A}'$.
- Every cell in $\mathcal{A}'$ is either void of data points (and thus likely has low probability under μ and can be disregarded) or else has a physical diameter that is roughly the same as its data diameter.

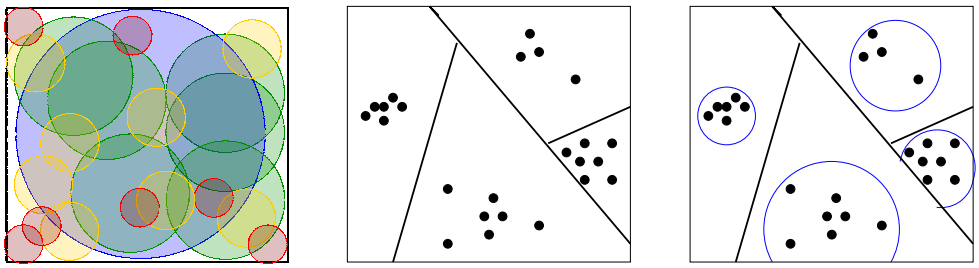
This construction lets us upper-bound the bias in terms of the data diameters $\Delta_n(A)$ of cells, which are easier to quantify and to control.

3.2. An alternate partition

Although the algorithm works with a partition $\mathcal{A}$ built from recursive hyperplane splits, and the regressor is defined using this partition, for purposes of the analysis only we will also consider an alternate, related partition $\mathcal{A}'$. This $\mathcal{A}'$ will be designed so that $f_{n,\mathcal{A}'}$ is equivalent to $f_{n,\mathcal{A}}$ on most of $\mathcal{X}$, but has the advantage that its cells are well-behaved as explained at the end of the previous section.

$\mathcal{A}'$ is obtained by intersecting the cells of $\mathcal{A}$ with balls or complements-of-balls from a fixed, pre-defined collection $\mathcal{B}$ (Figure 3). Specifically, let $\mathcal{B}_i$ be a cover of $\mathcal{X}$ by balls of radius $\Delta_{\mathcal{X}}/2^i$. Take a variety of scales: $i = 0, 1, 2, \ldots, I = \lfloor \log n^{2/(2+d)} \rfloor$. Then $\mathcal{B}$ is the union of all these balls of different sizes, blown up by a factor of 4:

$$\mathcal{B} = \bigcup_{i=0}^{I} \{4B : B \in \mathcal{B}_i\}.$$



(a) Cover $\mathcal{B}$ (b) Partition $\mathcal{A}$ (c) Partition $\mathcal{A}'$

Figure 3: We start with a cover $\mathcal{B}$ of $\mathcal{X}$ with balls of different size; then, we see the data and obtain a partition $\mathcal{A}$; and finally we substitute $\mathcal{A}$ with an alternate partition $\mathcal{A}'$, by intersecting the cells of $\mathcal{A}$ with balls of $\mathcal{B}$.

The partition $\mathcal{A}'$ is created by replacing each cell $A \in \mathcal{A}$ by two cells A'_1, A'_2 as follows:

- If $A \cap \mathbf{X} = \emptyset$, then set $A'_1 = A$ and $A'_2 = \emptyset$.
- Otherwise, set $i = \min\{I, \lceil \log(\Delta_{\mathcal{X}}/\Delta_n(A))\rceil\}$; we'll find a ball $B \in \mathcal{B}_i$ such that $A \cap \mathbf{X}$ is contained in $4B$. To this end, pick any $x \in A \cap \mathbf{X}$, and pick the ball $B \in \mathcal{B}_i$ whose center z is closest to x. Then $A \cap \mathbf{X} \subset 4B$ because $\forall x' \in A \cap \mathbf{X}$,

$$\begin{aligned} \|z - x'\| &\leq \|z - x\| + \|x - x'\| \\ &\leq 2^{-i}\Delta_{\mathcal{X}} + \Delta_n(A) \\ &\leq 2^{-i}\Delta_{\mathcal{X}} + 2^{-(i-1)}\Delta_{\mathcal{X}} \quad \leq \quad 4 \cdot 2^{-i}\Delta_{\mathcal{X}} \end{aligned}$$

  (we've used the fact that $i - 1 \leq \log(\Delta_{\mathcal{X}}/\Delta_n(A))$, whereby $\Delta_n(A) \leq 2^{-(i-1)}\Delta_{\mathcal{X}}$). Define $A'_1 = A \cap 4B$ and $A'_2 = A \setminus A'_1$.

$\mathcal{A}'$ is the collection of all such A'_1, A'_2, over $A \in \mathcal{A}$. What makes this refined partition valuable is that the average physical diameter of its cells can be upper-bounded by the empirical data diameters of cells in $\mathcal{A}$.

Lemma 12 (Diameters of $\mathcal{A}'$)**.** *Let $\mathcal{A}$ be a partition of $\mathcal{X}$ and define $\mathcal{A}'$ as above. Then*

$$\sum_{A' \in \mathcal{A}'} \mu_n(A')\Delta^2(A') \leq 64\Delta_n^2(\mathcal{A}) + 256n^{-4/(2+d)}\Delta_{\mathcal{X}}^2.$$

Proof. Pick any cell $A \in \mathcal{A}$ for which $A \cap \mathbf{X} \neq \emptyset$. This cell is broken into two pieces in $\mathcal{A}'$: a set A'_1 with $\mu_n(A'_1) = \mu_n(A)$ and a set A'_2 with $\mu_n(A'_2) = 0$. Specifically, $A'_1 = A \cap 4B$, where B is a ball of radius $2^{-i}\Delta_{\mathcal{X}}$, for $i = \min\{I, \lceil \log(\Delta_{\mathcal{X}}/\Delta_n(A))\rceil\}$. It follows that A'_1 has diameter at most $8 \cdot 2^{-i}\Delta_{\mathcal{X}} \leq 8\max\{2^{-I}\Delta_{\mathcal{X}}, \Delta_n(A)\}$.

This bound makes it natural to divide the cells of $\mathcal{A}$ into two groups: $\mathcal{A}_+ = \{A \in \mathcal{A} : \Delta_n(A) > 2^{-I}\Delta_{\mathcal{X}}\}$; and $\mathcal{A} \setminus \mathcal{A}_+$. Then

$$\begin{aligned}\sum_{A' \in \mathcal{A}'} \mu_n(A')\Delta^2(A') &= \sum_{A \in \mathcal{A}_+} \mu_n(A)\Delta^2(A'_1) + \sum_{A \in \mathcal{A} \setminus \mathcal{A}_+} \mu_n(A)\Delta^2(A'_1) \\ &\leq \sum_{A \in \mathcal{A}_+} 64\mu_n(A)\Delta_n^2(A) + \sum_{A \in \mathcal{A} \setminus \mathcal{A}_+} 64\mu_n(A)2^{-2I}\Delta_{\mathcal{X}}^2 \\ &\leq 64\Delta_n^2(\mathcal{A}) + 256n^{-4/(2+d)}\Delta_{\mathcal{X}}^2.\end{aligned}$$

□

3.3. Bounding the empirical masses of cells

In order to bound the integrated excess risk, we will need the empirical masses of cells, $\mu_n(A')$, to be close to their true masses, $\mu(A')$. In particular, this will allow us to disregard cells that are empty of data since they will have little effect on the integrated excess risk.

The uniform convergence bounds we use are based on the following standard notion of *shatter coefficient*, which describes the complexity of a (potentially infinite) collection of subsets of $\mathbb{R}^D$. In our case, each such subset is a cell.

Definition 13. *Let n be some positive integer, and let $\mathcal{C}$ be a class of subsets of $\mathbb{R}^D$. The n-shatter coefficient of $\mathcal{C}$, denoted $\mathcal{S}(\mathcal{C}, n)$, is the largest possible size of a collection of sets obtained by intersecting sets of $\mathcal{C}$ with a sample $\mathbf{X}$ of size n. That is,*

$$\mathcal{S}(\mathcal{C}, n) \doteq \max_{|\mathbf{X}|=n} |\{C \cap \mathbf{X} : C \in \mathcal{C}\}|.$$

For example, suppose $D = 1$ and $\mathcal{C}$ is the set of all half lines, that is, intervals of the form $(-\infty, t]$ or $[t, +\infty)$. For any set of n distinct points $\mathbf{X} = \{x_1, \ldots, x_n\}$ where (without loss of generality) $x_1 < \cdots < x_n$, the intersection of these points with half lines consists of all subsets of the form $\{x_1, \ldots, x_i\}$ or $\{x_i, \ldots, x_n\}$. Therefore $\mathcal{S}(\mathcal{C}, n) = 2n$.

The following theorem of Vapnik and Chervonenkis gives uniform rates of convergence for empirical masses over a class $\mathcal{C}$, using the $2n$-shattering coefficient of $\mathcal{C}$.

Lemma 14 (Relative VC bounds [VC71])**.** *Let $\mathcal{C}$ be a class of subsets of $\mathbb{R}^D$. Pick any $\delta > 0$. Suppose a sample of size n is drawn independently at random from a distribution μ over $\mathbb{R}^D$, with resulting empirical distribution μ_n. Then with probability at least $1 - \delta$ over the choice of sample, all $C \in \mathcal{C}$ satisfy*

$$\mu(C) \leq \mu_n(C) + 2\sqrt{\mu_n(C)\frac{\ln \mathcal{S}(\mathcal{C}, 2n) + \ln(4/\delta)}{n}} + 4\frac{\ln \mathcal{S}(\mathcal{C}, 2n) + \ln(4/\delta)}{n}.$$

where $\mathcal{S}(\mathcal{C}, 2n)$ is the $2n$-shatter coefficient of $\mathcal{C}$.

Recall that in our algorithm, we use the data sample $\mathbf{X}$ to generate a tree that contains various candidate partitions $\mathcal{A}^i$, and that eventually one of these partitions is chosen, and a

regressor is defined on it. We would like to argue that for any $\mathcal{A} = \mathcal{A}^i$, the empirical mass of each cell of $\mathcal{A}'$ is close to its true mass. How should the class $\mathcal{C}$ in lemma 14 be defined?

Since the tree has height at most $2\log 2n$ (remark 21 of section 5.1) and the splits are by hyperplanes, one option is to let $\mathcal{C}$ consist of all convex sets that are intersections of at most $2\log 2n$ halfspaces. This works, but yields a bound that depends on the ambient dimension D. Instead, we exploit the fact that the *directions* of the hyperplanes used in the tree are chosen at random independently of the sample $\mathbf{X}$, whereas their *displacements* depend on the sample $\mathbf{X}$. We can therefore condition on these directions being fixed before the choice of $\mathbf{X}$. This allows us to define a less complex class $\mathcal{C}$ containing the cells, and therefore yields a tighter bound independent of D.

Lemma 15 (Masses of cells of $\mathcal{A}'$)**.** *There is a constant C_0 such that the following holds. Pick any $\delta > 0$. With probability at least $1-\delta$ over the choice of $\mathbf{X}$ and the randomness in the algorithm, we have that for any partition $\mathcal{A} = \mathcal{A}^i$ generated during the construction of the tree, every cell $A' \in \mathcal{A}'$ satisfies*

$$\begin{aligned} \mu(A') &\leq \mu_n(A') + 2\sqrt{\mu_n(A')\frac{\mathcal{V}+\ln(4/\delta)}{n}} + 4\frac{\mathcal{V}+\ln(4/\delta)}{n}, \text{ where} \\ \mathcal{V} &\leq C_0 \log n(\log n + \text{loglog}(1/\delta)). \end{aligned} \tag{2}$$

Proof. Suppose without loss of generality that during the construction of the tree, all random directions (for hyperplane splits) are picked from a fixed collection $\mathcal{P}$ without replacement. How big should $\mathcal{P}$ be so that there are enough directions to choose from? The implementation of `coreRPtree` ensures that $|\mathcal{P}| \leq 8n^2 \log(3n/\delta)$ is sufficient (see remark 21 of section 5.1). Now fix such a collection $\mathcal{P}$ and let $\mathcal{H}_\mathcal{P}$ be the class of half spaces of $\mathbb{R}^D$ defined by hyperplanes normal to the directions in $\mathcal{P}$. For any tree partition $\mathcal{A}$, each cell of $\mathcal{A}$ is the intersection of at most $2\log 2n$ elements of $\mathcal{H}_\mathcal{P}$ since the tree is guaranteed to have height at most $2\log 2n$ (remark 21). Each cell of $\mathcal{A}'$ is the intersection of a ball or the complement of a ball in $\mathcal{B}$ with a cell of $\mathcal{A}$.

All such cells therefore belong to the following class of subsets of $\mathbb{R}^D$:

$$\mathcal{C} = \left\{ h : h = h_0 \cap \left(\bigcap_{l=1}^{2\log 2n} h_l \right),\ h_0 \text{ or } h_0^\complement \text{ is in } \mathcal{B}, h_l \in \mathcal{H}_\mathcal{P} \right\}.$$

We now proceed to bounding $\mathcal{S}(\mathcal{C}, 2n)$, the $2n$-shatter coefficient of $\mathcal{C}$.

Given $2n$ sample points and a direction $v \in \mathcal{P}$, there are at most $4n$ possible intersections of the sample with halfspaces normal to v. Therefore

$$\begin{aligned} \mathcal{S}(\mathcal{C}, 2n) &\leq 2|\mathcal{B}|\left(4n|\mathcal{P}|+1\right)^{2\log 2n} \\ &\leq 2|\mathcal{B}|\left(32n^3\log(3n/\delta)+1\right)^{2\log 2n}. \end{aligned}$$

Since $\mathcal{X}$ has doubling dimension d, we have $|\mathcal{B}| \leq \sum_{i=0}^{I} 2^{di} \leq 2n^{2d/(2+d)}$. The proof is completed by letting $\mathcal{V} = \log \mathcal{S}(\mathcal{C}, 2n)$ for $\mathcal{P}$ fixed, and then appealing to Lemma 14. □

3.4. A bound on the integrated excess risk in terms of data diameters

Lemma 16 (Integrated excess risk)**.** *There exists a constant C_1 independent of d and μ such that the following holds. Define $\alpha(n) \doteq \left(\log^2 n\right) \mathrm{loglog}(1/\delta) + \log(1/\delta)$. With probability at least $1-\delta/3$ over the choice of $(\mathbf{X}, \mathbf{Y})$ and the randomness in the algorithm, for all partitions $\mathcal{A} = \mathcal{A}^i$ obtained during the execution of* `adaptiveRPtree`*,*

$$\|f_{n,\mathcal{A}} - f\|^2 \quad \leq \quad C_1 \left(\Delta_{\mathcal{Y}}^2 \, |\mathcal{A}| \, \frac{\alpha(n)}{n} + \lambda^2 \left(\Delta_n^2 \left(\mathcal{A} \right) + n^{-4/(2+d)} \Delta_{\mathcal{X}}^2 \right) \right).$$

Proof. Define $\delta' = \delta/(6 \log 2n)$. By Lemma 15 we have that with probability at least $1 - \delta'$ over the randomness in the algorithm and the choice of $\mathbf{X}$, equation (2) — with δ' substituted for δ — holds for all cells $A' \in \mathcal{A}'$, where $\mathcal{A} = \mathcal{A}^i$ is any partition obtained during the construction of the tree and $\mathcal{V} \leq C_0 \log n (\log n + \mathrm{loglog}(1/\delta'))$. Let's assume that this condition holds, and fix $\mathbf{X}$. Henceforth we will randomize only over the choice of $\mathbf{Y}$.

Pick any partition $\mathcal{A} = \mathcal{A}^i$ obtained by `adaptiveRPtree`. The integrated excess risk can be decomposed over $\mathcal{A}'$ as follows:

$$\|f_{n,\mathcal{A}} - f\|^2 = \sum_{A' \in \mathcal{A}'} \int_{A'} \|f_{n,\mathcal{A}}(x) - f(x)\|^2 \, \mu(dx).$$

We next divide the cells of $\mathcal{A}'$ into two groups: those of significant mass, whose bias and variance must be controlled, and those of negligible mass, whose contribution to the overall risk can be ignored even if it is the worst possible. Specifically, set

$$\mathcal{A}'_{>} \doteq \left\{ A' \in \mathcal{A}', \mu_n(A') \geq \frac{\mathcal{V} + \ln(4/\delta')}{n} \right\}, \text{ and } \mathcal{A}'_{<} \doteq \mathcal{A}' \setminus \mathcal{A}'_{>}.$$

From equation (2), every $A' \in \mathcal{A}'_{>}$ satisfies $\mu(A') \leq 7\mu_n(A')$ while every $A' \in \mathcal{A}'_{<}$ has $\mu(A') \leq 7(\mathcal{V} + \ln(4/\delta'))/n$.

Given this upper bound on the masses of cells in $\mathcal{A}'_{<}$, their integrated risk is

$$\sum_{A' \in \mathcal{A}'_{<}} \int_{A'} \|f_{n,\mathcal{A}}(x) - f(x)\|^2 \, \mu(dx) \quad \leq \quad \sum_{A' \in \mathcal{A}'_{<}} \Delta_{\mathcal{Y}}^2 \cdot \mu(A') \quad \leq \quad 7\Delta_{\mathcal{Y}}^2 \cdot |\mathcal{A}'| \cdot \frac{\mathcal{V} + \ln(4/\delta')}{n}. \quad (3)$$

Now for the integration over $\mathcal{A}'_{>}$. Each cell $A' \in \mathcal{A}'_{>}$ holds exactly the same data points as its counterpart $A \in \mathcal{A}$; thus $f_{n,\mathcal{A}}$ and $f_{n,\mathcal{A}'}$ coincide on A'. We first apply (1), and then use

Lemmas 10 and 11 to assert that with probability at least $1-\delta'$ over the choice of $\mathbf{Y}$,

$$\begin{aligned}
&\sum_{A' \in \mathcal{A}'_>} \int_{A'} \|f_{n,\mathcal{A}}(x) - f(x)\|^2 \, \mu(dx) \\
&= \sum_{A' \in \mathcal{A}'_>} \int_{A'} \|f_{n,\mathcal{A}'}(x) - f(x)\|^2 \, \mu(dx) \\
&\leq \sum_{A' \in \mathcal{A}'_>} 2\lambda^2 \Delta^2(A') \cdot \mu(A') + \sum_{A' \in \mathcal{A}'_>} 2\Delta_{\mathcal{Y}}^2 \cdot \frac{2 + \ln(|\mathcal{A}'|/\delta')}{n\mu_n(A')} \cdot \mu(A') \\
&\leq \sum_{A' \in \mathcal{A}'_>} 2\lambda^2 \Delta^2(A') \cdot 7\mu_n(A') + \sum_{A' \in \mathcal{A}'_>} 2\Delta_{\mathcal{Y}}^2 \cdot \frac{2 + \ln(|\mathcal{A}'|/\delta')}{n\mu_n(A')} \cdot 7\mu_n(A') \\
&\leq 14\lambda^2 \sum_{A' \in \mathcal{A}'_>} \mu_n(A')\Delta^2(A') + 14\Delta_{\mathcal{Y}}^2 |\mathcal{A}'| \cdot \frac{2 + \ln(|\mathcal{A}'|/\delta')}{n}. \qquad (4)
\end{aligned}$$

We can simplify $\ln|\mathcal{A}'|$ to $O(\log n)$ since the tree has at most n leaves. By combining the bounds in (3) and (4), and absorbing various constants into a single C_o, we get

$$\|f_{n,\mathcal{A}} - f\|^2 \leq C_o \left(\Delta_{\mathcal{Y}}^2 |\mathcal{A}| \frac{\log^2 n + \log n \log\log 1/\delta' + \log(1/\delta')}{n} + \lambda^2 \sum_{A' \in \mathcal{A}'} \mu_n(A')\Delta^2(A') \right).$$

To finish up, we call on lemma 12 to bound the summation on the right, and then take a union bound over the $\leq \log 2n$ possible partitions $\mathcal{A} = \mathcal{A}^i$. □

4. Risk of final regressor $f_n \doteq f_{n,\mathcal{A}^*}$

Recall that the `adaptiveRPtree` procedure starts with a partition $\mathcal{A}^0$ that has a single cell containing all the data, and then grows the tree to get increasingly finer partitions $\mathcal{A}^1, \mathcal{A}^2, \ldots$, where the data diameter of each $\mathcal{A}^i$ is half that of $\mathcal{A}^{i-1}$. Recall also that the *diameter decrease rate*, denoted k, is defined to be the maximum increase in tree depth during each of these individual growth spurts.

The tree is not grown indefinitely. To see this, note that the implementation of `coreRPtree` ensures that all cells at some level down the hierarchy would eventually have a single data point in them (see remark 21). In other words, $\Delta_n\left(\mathcal{A}^i\right) = 0$ eventually, at which point either of the two stopping criteria would hold.

Once the tree is constructed, a partition $\mathcal{A}^* = \mathcal{A}^i$ is chosen and a regressor is built on it. We now bound the excess risk of $f_n \doteq f_{n,\mathcal{A}^*}$ in terms of the diameter decrease rate achieved during `adaptiveRPtree`.

To get some insight into the form of the final risk bound, pretend for a moment that $\Delta_{\mathcal{X}}$, $\Delta_{\mathcal{Y}}$, and λ are all 1. Consider a partition $\mathcal{A}$ induced by the tree. If $\Delta_n(\mathcal{A}) = \zeta$, we would expect that the data diameter has been halved roughly $\log(1/\zeta)$ times. Since each halving grows the tree by $\leq k$ levels, $\mathcal{A}$ has depth at most $k\log(1/\zeta)$ in the tree,

implying also that $|\mathcal{A}| \leq (1/\zeta)^k$. Plugging these values into the bound of Lemma 16, we get $\|f_{n,\mathcal{A}} - f\|^2 \lesssim \zeta^{-k}/n + \zeta^2$. Setting $\zeta = n^{-1/(2+k)}$ then gives the optimal bound $\|f_{n,\mathcal{A}^*} - f\|^2 \lesssim n^{-2/(2+k)}$.

In the analysis, a few basic facts will repeatedly be used. First, because such successive partition halves the data diameter,

$$\Delta_n(\mathcal{A}^i) \quad \leq \quad 2^{-i}\Delta_n(\mathcal{A}^0). \tag{5}$$

Second, by definition of diameter decrease rate, each halving grows the tree by $\leq k$ levels:

$$\text{level}\left(\mathcal{A}^i\right) \quad \leq \quad ki. \tag{6}$$

4.1. Risk bound for cross-validation option

For the cross-validation option, we begin by arguing that the tree contains at least one good partition $\mathcal{A}^i$, such that both $\Delta_n(\mathcal{A}^i)$ and $|\mathcal{A}^i|$ are reasonably small. The shrinkage in diameter, $\Delta_n(\mathcal{A}^i)/\Delta_n(\mathcal{A}^0)$, is roughly

$$\zeta \doteq \left(\frac{\Delta_{\mathcal{Y}}^2}{\lambda^2\Delta_{\mathcal{X}}^2} \cdot \frac{\alpha(n)}{n}\right)^{1/(2+k)}$$

(recall $\alpha(n) = (\log^2 n)\log\log(n/\delta) + \log(1/\delta)$.) The analysis requires an unusual, albeit benign, lower bound on the number of samples, n, the purpose of which is to ensure that n^2 exceeds both $(1/\zeta)^k$ and $(1/\zeta)^{2+d}$.

Lemma 17 (Existence of a good pruning)**.** *Suppose* `adaptiveRPtree` *is run with the cross-validation option, and yields a sequence of partitions $\mathcal{A}^0, \mathcal{A}^1, \ldots$ with a diameter decrease rate of k. Define*

$$\zeta \doteq \left(\frac{\Delta_{\mathcal{Y}}^2}{\lambda^2\Delta_{\mathcal{X}}^2} \cdot \frac{\alpha(n)}{n}\right)^{1/(2+k)}$$

If $n \geq \max\left\{\alpha(n), \lambda^2\Delta_{\mathcal{X}}^2/\Delta_{\mathcal{Y}}^2, \alpha(n)\Delta_{\mathcal{Y}}^2/\lambda^2\Delta_{\mathcal{X}}^2\right\}$, then there exists $i \geq 0$ such that $\Delta_n\left(\mathcal{A}^i\right) \leq 2\zeta \cdot \Delta_n\left(\mathcal{X}\right)$ and $\left|\mathcal{A}^i\right| \leq (1/\zeta)^k$.

Proof. Consider the largest i at which $\text{level}\left(\mathcal{A}^i\right) < k\log(1/\zeta)$. Then $|\mathcal{A}^i| \leq (1/\zeta)^k$. In bounding $\Delta_n(\mathcal{A}^i)$, there are two cases to consider.

Case 1: $\mathcal{A}^{i+1}$ is part of the tree. Then its level is $\geq k\log(1/\zeta)$, implying that $i+1 \geq \log(1/\zeta)$ (by (6)) and therefore that $i \geq \log(1/2\zeta)$, whereupon (by (5)) $\Delta_n(\mathcal{A}^i) \leq 2\zeta\Delta_n(\mathcal{A}^0)$.

Case 2: $\mathcal{A}^{i+1}$ is not part of the tree; that is, $\mathcal{A}^i$ satisfies one of the two stopping criteria. The lower bound on n ensures that $\text{level}\left(\mathcal{A}^i\right) < k\log(1/\zeta) \leq 2\log n$. Therefore $\Delta_n(\mathcal{A}^i) = 0$. □

Next, we argue that cross-validation will find a partition that isn't too much worse than the $\mathcal{A}^i$ of Lemma 17.

Lemma 18. *There exists an absolute constant C (independent of d and μ), such that the following holds. Under the hypotheses of Lemma 17, with probability at least $1 - 2\delta/3$ over $(\mathbf{X}, \mathbf{Y})$ and the randomness in the algorithm, the excess risk of the final regressor is bounded by*

$$\|f_n - f\|^2 \;\leq\; C \cdot (\lambda \Delta_{\mathcal{X}})^{2k/(2+k)} \left(\Delta_{\mathcal{Y}}^2 \cdot \frac{\alpha(n)}{n} \right)^{2/(2+k)} + 2\Delta_{\mathcal{Y}}^2 \sqrt{\frac{\log\log n + \log 4/\delta}{2n}}.$$

Proof. Let $\mathcal{A}^i$ and ζ be as in Lemma 17. By applying Lemma 16 and then Lemma 17, we have with probability at least $1 - \delta/3$ that

$$\begin{aligned} \left\| f_{n,\mathcal{A}^i} - f \right\|^2 &\leq C_1 \left(\Delta_{\mathcal{Y}}^2 \left| \mathcal{A}^i \right| \frac{\alpha(n)}{n} + \lambda^2 \left(\Delta_n^2 \left(\mathcal{A}^i \right) + n^{-4/(2+d)} \Delta_{\mathcal{X}}^2 \right) \right) \\ &\leq C_1 \left(\Delta_{\mathcal{Y}}^2 \cdot \zeta^{-k} \frac{\alpha(n)}{n} + 5\lambda^2 \zeta^2 \Delta_{\mathcal{X}}^2 \right) \;\leq\; C_2 \lambda^2 \Delta_{\mathcal{X}}^2 \zeta^2. \end{aligned}$$

To analyze the cross validation phase, we fix the partitions $\mathcal{A}^j$ obtained by `adaptiveRPtree`; there at most $\log 2n$ of these. Applying McDiarmid's inequality to the empirical risk, we see that with probability at least $1 - \delta/3$ over the choice of $(\mathbf{X}', \mathbf{Y}')$, each $\mathcal{A}^j$ satisfies

$$\left| R\left(f_{n,\mathcal{A}^j} \right) - R'_n\left(f_{n,\mathcal{A}^j} \right) \right| \leq \Delta_{\mathcal{Y}}^2 \sqrt{\frac{\ln(\log 2n) + \ln 3/\delta}{2n}}.$$

Thus if $f_n \doteq f_{n,\mathcal{A}^*}$ is the empirical risk minimizer,

$$\|f_n - f\|^2 \leq C_2 \lambda^2 \Delta_{\mathcal{X}}^2 \zeta^2 + 2\Delta_{\mathcal{Y}}^2 \sqrt{\frac{\log\log n + \log 4/\delta}{2n}}$$

with probability at least $1 - 2\delta/3$. □

4.2. Risk bound for automatic stopping option

The automatic criterion stops growing the tree as soon as

$$\frac{\Delta_n^2\left(\mathcal{A}^i \right)}{\Delta_n^2\left(\mathcal{A}^0 \right)} \;\leq\; \frac{\alpha(n)}{n} 2^{\text{level}\left(\mathcal{A}^i \right)},$$

at which point either $\mathcal{A}^i$ or $\mathcal{A}^{i-1}$ is chosen as the final partition $\mathcal{A}^*$. The shrinkage in diameter is expected to be roughly a factor of

$$\zeta \doteq \left(\frac{\alpha(n)}{n} \right)^{1/(2+k)},$$

corresponding to a depth of $k \log(1/\zeta)$ in the tree. In particular, if $\text{level}\left(\mathcal{A}^i \right) \geq k \log(1/\zeta)$ then the stopping criterion holds, because then $i \geq \text{level}\left(\mathcal{A}^i \right)/k \geq \log(1/\zeta)$ (recall (6)) and $\Delta_n(\mathcal{A}^i) \leq 2^{-i} \Delta_n(\mathcal{A}^0) \leq \zeta \Delta_n(\mathcal{A}^0)$ (recall (5)), whereupon

$$\frac{\Delta_n^2\left(\mathcal{A}^i \right)}{\Delta_n^2\left(\mathcal{A}^0 \right)} \;\leq\; \zeta^2 \;=\; \frac{\alpha(n)}{n} \left(\frac{1}{\zeta} \right)^k \;\leq\; \frac{\alpha(n)}{n} 2^{\text{level}\left(\mathcal{A}^i \right)}.$$

Lemma 19 (Properties of $\mathcal{A}^*$)**.** *Suppose the automatic stopping option is used, and that* `adaptiveRPtree` *attains a diameter decrease rate of* k *on* $\mathbf{X}$*. Define* $\zeta \doteq \left(\frac{\alpha(n)}{n}\right)^{1/(2+k)}$ *and assume* $n \geq \alpha(n)$*. Then, the final partition* $\mathcal{A}^*$ *retained for regression satisfies*

$$\left(\frac{\alpha(n)}{n} \cdot |\mathcal{A}^*| + \Delta_n^2(\mathcal{A}^*)\right) \leq \left(4\Delta_n^2(\mathcal{X}) + 1\right)\zeta^2.$$

Proof. Let $\mathcal{A}^0, \mathcal{A}^1, \ldots$ be the partitions found by `adaptiveRPtree`, and suppose the stopping criterion holds for $\mathcal{A}^i$. We consider two cases:

Case 1: $\text{level}\left(\mathcal{A}^i\right) \leq k\log(1/\zeta)$. Then $|\mathcal{A}^i| \leq (1/\zeta)^k$ and by the stopping condition

$$\frac{\Delta_n^2\left(\mathcal{A}^i\right)}{\Delta_n^2\left(\mathcal{A}^0\right)} \;\leq\; \frac{\alpha(n)}{n}2^{\text{level}(\mathcal{A}^i)} \;\leq\; \frac{\alpha(n)}{n}\left(\frac{1}{\zeta}\right)^k \;=\; \zeta^2.$$

Case 2: $\text{level}\left(\mathcal{A}^i\right) > k\log(1/\zeta)$. Then $ki \geq \text{level}\left(\mathcal{A}^i\right) \geq k\log(1/\zeta)$, implying that $i - 1 \geq \log(1/2\zeta)$, whereupon $\Delta_n(\mathcal{A}^{i-1}) \leq 2\zeta\Delta_n(\mathcal{A}^0)$ (recall (5)). Moreover, since the stopping condition doesn't hold for $\mathcal{A}^{i-1}$ we have (by the discussion preceding the lemma) that $\text{level}\left(\mathcal{A}^{i-1}\right) < k\log(1/\zeta)$.

In either case at least one of $\mathcal{A}^i$ and $\mathcal{A}^{i-1}$ has size at most $(1/\zeta)^k$ and diameter at most $2\zeta \cdot \Delta_n(\mathcal{A}^0)$. It follows that

$$\min_{j\in\{i-1,\,i\}}\left(\frac{\alpha(n)}{n}\cdot\left|\mathcal{A}^j\right| + \Delta_n^2\left(\mathcal{A}^j\right)\right) \leq \frac{\alpha(n)}{n}\cdot\zeta^{-k} + 4\zeta^2\cdot\Delta_n^2(\mathcal{X}) = \left(4\Delta_n^2(\mathcal{X}) + 1\right)\zeta^2,$$

which concludes the argument. □

Lemma 20. *There exists an absolute constant* C *(independent of* d *and* μ*), such that the following holds. Suppose the automatic stopping option is used and that* `adaptiveRPtree` *achieves a diameter decrease rate of* $k \geq d$ *on* $\mathbf{X}$*. With probability at least* $1 - \delta/3$ *over* $(\mathbf{X}, \mathbf{Y})$ *and the randomness in the algorithm, the excess risk of the regressor is bounded by*

$$\|f_n - f\|^2 \leq C\cdot\left(\Delta_{\mathcal{Y}}^2 + \lambda^2\right)\left(\Delta_{\mathcal{X}}^2 + 1\right)\cdot\left(\frac{\alpha(n)}{n}\right)^{2/(2+k)}.$$

Proof. For $n \leq \alpha(n)$, the bound on the excess risk holds vacuously. We assume henceforth that $n > \alpha(n)$. Let $\zeta \doteq \left(\frac{\alpha(n)}{n}\right)^{1/(2+k)}$. By first applying Lemma 16 then Lemma 19, we have with probability at least $1 - \delta$ that

$$\begin{aligned}
\|f_{n,\mathcal{A}^*} - f\|^2 &\leq C_1\left(\Delta_{\mathcal{Y}}^2|\mathcal{A}^*|\frac{\alpha(n)}{n} + \lambda^2\left(\Delta_n^2(\mathcal{A}^*) + n^{-4/(2+d)}\Delta_{\mathcal{X}}^2\right)\right)\\
&\leq C_1\left(\Delta_{\mathcal{Y}}^2 + \lambda^2\right)\left(|\mathcal{A}^*|\frac{\alpha(n)}{n} + \left(\Delta_n^2(\mathcal{A}^*) + n^{-4/(2+d)}\Delta_{\mathcal{X}}^2\right)\right)\\
&\leq C_1\left(\Delta_{\mathcal{Y}}^2 + \lambda^2\right)\left(\left(4\Delta_{\mathcal{X}}^2 + 1\right)\zeta^2 + \zeta^2\Delta_{\mathcal{X}}^2\right)\\
&\leq C\left(\Delta_{\mathcal{Y}}^2 + \lambda^2\right)\left(\Delta_{\mathcal{X}}^2 + 1\right)\zeta^2,
\end{aligned}$$

which concludes the argument. □

5. The `coreRPtree` procedure and diameter decrease rates

5.1. The `coreRPtree` *procedure*

Procedure `basicRPtree`($A_0 \subset \mathcal{X}$, Δ)

$\mathcal{A}_0 \leftarrow \{A_0\}$;
for $i \leftarrow 1$ **to** ∞ **do**
  if $\Delta_n(\mathcal{A}_{i-1}) \leq \Delta$ and i is odd **then**
    return;
  end
  Choose a random direction $v \sim \mathcal{N}\left(0, \frac{1}{D} I_D\right)$;
  Choose a random $\tau \sim \mathcal{U}[-1, 1] \cdot \frac{6}{\sqrt{D}} \Delta_n(A_0)$;
  foreach cell $A \in \mathcal{A}_{i-1}$ **do**
    if i is odd **then**
      $t \leftarrow \text{median}\{z^\top v : z \in \mathbf{X} \cap A_0\} + \tau$; `// Noisy splits`
    else
      $t \leftarrow \text{median}\{z^\top v : z \in \mathbf{X} \cap A\}$; `// Median splits`
    end
    $A_{\text{left}} \leftarrow \{x \in A,\, x^\top v \leq t\}$;
    $A_{\text{right}} \leftarrow A \setminus A_{\text{left}}$;
    if $(A_{\text{left}} \cap \mathbf{X})$ and $(A_{\text{right}} \cap \mathbf{X})$ are both nonempty **then**
      (children of A) $\leftarrow A_{\text{left}}, A_{\text{right}}$;
    end
  end
  $\mathcal{A}_i \leftarrow$ partition of A_0 defined by the leaves of the current tree;
end

Procedure `coreRPtree`($A \subset \mathcal{X}$, Δ, δ)

Call `basicRPtree`(A, Δ) $\log(3n/\delta)$ times and return the shortest tree.

In a random projection (RP) tree [DF08], each cell is split by a random hyperplane; specifically, a random direction is chosen from the surface of the unit sphere, and then the cell is split along that direction, at the median plus a small random perturbation. As a result of this perturbation, the two halves of the cell might not contain an equal number of points, and, in some cases, might be severely imbalanced. In our present setting, we need to get a handle on the data diameters of individual cells – which the RP tree split gives us – but also on the depth of the tree, since this relates to the complexity of the cells (see Lemma 15). To control this latter quantity, we alternate the RP split with another type of bisection that splits exactly at the median. Thus, if the tree is grown to l levels, we are assured that each cell contains at most a $2^{-l/2}$ fraction of the original data set; hence the overall depth of the tree must be $O(\log n)$.

Another complication associated with the RP tree is that the rapid decrease in diameters is assured only with a certain probability. The procedure `coreRPtree` boosts this probability by calling `basicRPtree` multiple times in parallel, and picking the shortest tree obtained.

Remark 21. *Given the implementation of* `coreRPtree`*, the tree returned by* `adaptiveRPtree` *has the following properties:*

- *The number of data points in a cell (node) at level i is at most half the number contained in its ancestor at level $i-2$. Taking rounding effects into consideration, this means that by level $2(1+\log n)$, each cell will contain at most one point. Thus the entire tree built by* `adaptiveRPtree` *has depth at most $2\log 2n$.*
- *By construction, each node contains at least one data point. Therefore, there are at most n leaves and $n-1$ internal nodes.*
- *Since the tree has height at most $2\log 2n = \log 4n^2$, a total of at most $8n^2\log(3n/\delta)$ random directions are required to build the entire tree.*

5.2. Worst case decrease rates

In this section we consider worst case bounds for the diameter decrease rates achieved by `coreRPtree` on data sets of low intrinsic dimension. The following theorem, which is based upon Lemma 9 of [DF08], gives the basic bound we will rely upon.

Theorem 22. *There is an absolute constant C' for which the following holds. Let $A \subset \mathbb{R}^D$ and suppose $A \cap \mathbf{X}$ has doubling dimension d. Then with probability at least $1/2$ over the randomization within the algorithm,* `basicRPtree`$(A, \Delta_n(A)/2)$ *returns a tree of depth at most $C'd\log d$.*

Proof idea. The proof is a direct consequence of Lemma 9 of [DF08] applied to the "noisy" splits at alternating levels in procedure `basicRPtree`.

Let $r = \Delta_n(A)/512\sqrt{d}$ and consider an r-cover of A; now consider pairs of balls $B = B(z,r)$, $B' = B(z',r)$, where z, z' are in the cover and $\|z - z'\| \geq \frac{1}{2}\Delta_n(A) - 2r$. Notice that `basicRPtree` stops if for all such pairs, no leaf of the tree contains points from both $B \cap \mathbf{X}$ and $B' \cap \mathbf{X}$.

Fix such a pair B and B'. By Lemma 9 of [DF08], every "noisy" split has a constant probability of separating $B \cap \mathbf{X}$ and $B' \cap \mathbf{X}$. Thus, the probability that some cell at level i contains points from both $B \cap \mathbf{X}$ and $B' \cap \mathbf{X}$ goes down exponentially with i. A union bound over at most $(O(d)^d)$ such pairs yields the theorem. □

Corollary 23. *Let C' be as in Theorem 22. Suppose $\mathcal{X}$ has doubling dimension d and fix $\mathbf{X} \subset \mathcal{X}$. With probability at least $1 - \delta/3$ over the randomness in the algorithm,* `adaptiveRPtree` *attains a diameter decrease rate $k \leq C'd\log d$ on $\mathbf{X}$.*

Proof. The procedure `adaptiveRPtree` grows the tree in blocks: it starts with a single node (cell) that contains all of $\mathbf{X}$ and then repeatedly expands one of its current leaf nodes A into the subtree that is generated by the call `coreRPtree`$(A, \Delta_n(A), 2)$.

Consider any such A. Since $\mathcal{X}$ has doubling dimension d, so does $A \cap \mathbf{X} \subset \mathcal{X}$; we can therefore apply Theorem 22. Procedure `coreRPtree` calls `basicRPtree` $\log(3n/\delta)$ times

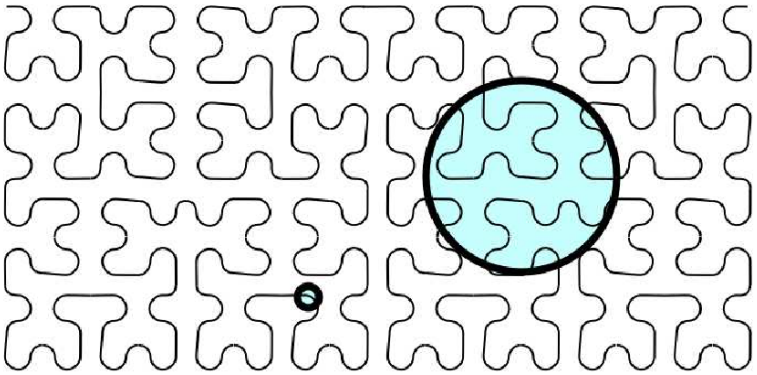


Figure 4: Hilbert space filling curve: the dimension depends on the scale at which the set is examined. Image obtained from [DF08].

and returns the smallest tree; thus the probability that this tree has depth $> C'd \log d$ is at most $\delta/(3n)$.

How many nodes A are expanded in this way? Any A with data diameter zero (for instance, containing just one point) is untouched by `coreRPtree`; on the other hand, any A with nonzero diameter will certainly get expanded (on account of the median split, if nothing else). Thus `coreRPtree` is invoked at most once on each internal node of the tree. There are at most n leaf nodes and thus at most $n-1$ internal nodes. A union bound over them yields an overall probability of failure at most $\delta/3$. □

6. Extensions

We have demonstrated a tree regressor that performs well in scenarios where the data space $\mathcal{X} \subset \mathbb{R}^D$ has low doubling dimension $d \ll D$. In such cases, the integrated excess risk is roughly of the form $n^{-2/(2+k)}$ for $k = O(d \log d)$, and has no dependence on the ambient dimension D. But this still leaves room for improvement: is there an efficient tree-based regressor that achieves the optimal rate, $n^{-2/(2+d)}$?

Some very recent work [Kpo09] uses kernel regression to achieve this rate in general metric spaces. Moreover, in that paper the usual $O(n)$ evaluation time of kernel methods is reduced to $O(2^d \log n)$ using a special tree data structure. This is a significant improvement, though slower than the $O(\log n)$ evaluation time of a tree regressor.

Another set of open questions concerns the data model. doubling dimension is fairly general while at the same time being amenable to analysis. However, it has some shortcomings that motivate exploration into alternative notions of intrinsic dimension. First of all, it is natural to allow the dimensionality of a data set to depend on the *scale* at which it is being examined. The set in Figure 4, for instance, looks two-dimensional from a distance but one-dimensional when restricted to smaller neighborhoods. And realistically, at even

smaller neighborhood sizes, it would be full-dimensional because of white noise. At the very least, we would like to be able to handle data sets that have low intrinsic dimension only when restricted to neighborhoods of a certain radius. In the Appendix, we show how to extend our results to such a setting.

A second shortcoming of doubling dimension is that it seems difficult to efficiently estimate for a given data set. Although our algorithm doesn't need to know the intrinsic dimension, it would be nice to have some concrete reassurance that this quantity is small for a wide range of data. Is there a notion of dimension that is empirically verifiable, and fairly general, and powerful enough to be the key exponent in risk bounds for nonparametric methods? One recent proposal is the *local covariance dimension* [DF08, VKD09], but regression risk has not yet been analyzed in terms of it.

Acknowledgements

This work was supported by the National Science Foundation (under grants IIS-0347646, IIS-0713540, and IIS-0812598) and by a fellowship from the Engineering Institute at the Los Alamos National Laboratory.

References

References

[BL07] P. Bickel and B. Li. Local polynomial regression on unknown manifolds. *Complex Datasets and Inverse Problems: Tomography, Networks, and Beyond, IMS Lecture Notes – Monograph Series*, 54:177–186, 2007.

[BN03] M. Belkin and N. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15:1373–1396, 2003.

[Cla05] K. Clarkson. Nearest-neighbor searching and metric space dimensions. *Nearest-Neighbor Methods for Learning and Vision: Theory and Practice*, 2005.

[CS07] A. Caponnetto and S. Smale. Risk bounds for random regression graphs. *Foundations of Computational Mathematics*, 7:495–528, 2007.

[DF08] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. *Fortieth ACM Symposium on Theory of Computing*, 2008.

[DGL96] L. Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[GKKW02] L. Gyorfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution Free Theory of Nonparametric Regression*. Springer, 2002.

[GKL03] A. Gupta, R. Krauthgamer, and J. R. Lee. Bounded geometries, fractals, and low-distortion embeddings. *Symposium on Foundations of Computer Science*, 2003.

[GLZ08] A. B. Goldberg, M. Li, and X. Zhu. Online manifold regularization: a new learning setting and empirical study. *European Conference on Machine Learning and Practice of Knowledge Discovery in Databases*, 2008.

[GN05] S. Gey and E. Nedelec. Model selection for cart regression trees. *IEEE Transactions on Information Theory*, 51(2):658–670, 2005.

[IN07] P. Indyk and A. Naor. Nearest neighbor preserving embedding. *ACM Transactions on Algorithms*, 3(3), 2007.

[JL84] W. Johnson and J. Lindenstrauss. Extensions of lipschitz maps into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

[KP95] S. Kulkarni and S. Posner. Rates of convergence of nearest neighbor estimation under arbitrary sampling. *IEEE Transactions on Information Theory*, 41(4):1028–1039, 1995.

[Kpo09] S. Kpotufe. Fast, smooth, and adaptive regression in metric spaces. *Neural Information Processing Systems*, 2009.

[LW07] J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. *Neural Information Processing Systems*, 2007.

[NSW08] P. Niyogi, S. Smale, and S. Weinberger. Finding the homology of submanifolds with high confidence from random samples. *Discrete and Computational Geometry*, 39(1):419–441, 2008.

[RS00] S. Roweis and L. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290:2323–2326, 2000.

[SN06] C. Scott and R.D. Nowak. Minimax-optimal classification with dyadic decision trees. *IEEE Transactions on Information Theory*, 52(4):1335–1353, 2006.

[Sto80] C. J. Stone. Optimal rates of convergence for non-parametric estimators. *Annals of Statistics*, 8:1348–1360, 1980.

[Sto82] C. J. Stone. Optimal global rates of convergence for non-parametric estimators. *Annals of Statistics*, 10:1340–1353, 1982.

[TSL00] J.B. Tenenbaum, V. De Silva, and J. Langford. A global geometric framework for non-linear dimensionality reduction. *Science*, 290:2319–2323, 2000.

[VC71] V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their expectation. *Theory of probability and its applications*, 16:264–280, 1971.

[VKD09] N.A. Verma, S. Kpotufe, and S. Dasgupta. Which spatial partition trees are adaptive to intrinsic dimension? *Uncertainty in Artificial Intelligence*, 2009.

Appendix A. On the adaptivity of an axis-parallel splitting rule

In this section we show that if the input space $\mathcal{X}$ is a subset of $[-1,1]^D$ of doubling dimension d, then a dyadic tree regressor (Figure 1(a)) achieves a convergence rate of the form $O(n^{-2/(2+d)})$, but with a leading constant that is exponential in D.

The dyadic tree starts with a single cell corresponding to all of $[-1,1]^D$, and then grows one level at a time. In each such expansion, a particular coordinate direction is chosen and every current leaf cell is bisected at its midpoint along that coordinate. There is flexibility in how the coordinate direction is chosen; a common choice is to simply cycle through the D coordinates. The final level of the tree defines a partition $\mathcal{A}$ of $[-1,1]^D$, and a regressor $f_{n,\mathcal{A}}$ is obtained by averaging the Y values in each cell $A \in \mathcal{A}$.

Unlike an RP tree, the dyadic tree is not data-dependent. In such cases, a generic risk bound applies. If the cells of $\mathcal{A}$ have diameter $\leq \zeta$, and if $\mathcal{A}_\mathcal{X}$ is the subset of cells intersecting $\mathcal{X}$, then it is implicit, for instance, in the proof of Theorem 4.3 of [GKKW02], that

$$\mathbb{E}\,\|f_{n,\mathcal{A}} - f\|^2 \leq C\left(\Delta_\mathcal{Y}^2 \frac{|\mathcal{A}_\mathcal{X}|}{n} + \lambda^2\zeta^2\right). \tag{A.1}$$

The result in this section is obtained by noticing that most cells of $\mathcal{A}$ will be empty if $\mathcal{X}$ has doubling dimension much smaller than D. Think for instance of $\mathcal{X}$ as a line curving slowly through the cube $[-1,1]^D$.

Theorem 24. *There are absolute constants C_1, C_2, and C_3 for which the following holds. Consider an input space $\mathcal{X} \subset [-1,1]^D$ of diameter 1 and doubling dimension d. Let $\mathcal{A}$ be a dyadic partition where each cell has diameter $\zeta < 1$, that is, cells have side lengths $\zeta/\sqrt{D}$. If $\zeta = C_1\left(\Delta_\mathcal{Y}^2 \cdot 2^{C_3 D\log D}/(\lambda^2 n)\right)^{1/(2+d)}$, we have*

$$\mathbb{E}\,\|f_{n,\mathcal{A}} - f\|^2 \leq C_2\lambda^{2d/(2+d)}\left(\frac{\Delta_\mathcal{Y}^2 \cdot 2^{C_3 D\log D}}{n}\right)^{2/(2+d)}.$$

Proof. Let $\mathcal{A}_\mathcal{X} \subset \mathcal{A}$ be the cells of $\mathcal{A}$ that intersect $\mathcal{X}$. We'll first show that $|\mathcal{A}_\mathcal{X}| \leq 2^{O(D\log D)}(1/\zeta)^d$. By the doubling assumption, $\mathcal{X}$ has a $(\zeta/2)$-cover of size $N \leq (2/\zeta)^d$; call it $\{z_i\}_1^N \subset \mathcal{X}$. Now consider the (closed) balls $B(z_i,\zeta)$. By a triangle inequality, the center of each hypercube $A \in \mathcal{A}_\mathcal{X}$ is contained in some ball $B(z_i,\zeta)$ (the center of each A is within $\zeta/2$ of all $x \in A \cap \mathcal{X}$ and each such x is within $\zeta/2$ of some z_i). Therefore, if M is the maximum number of such centers in a single ball $B(z_i,\zeta)$, then $|\mathcal{A}_\mathcal{X}| \leq M \cdot N$.

To bound M, notice that the centers of the hypercubes $A \in \mathcal{A}_\mathcal{X}$ are at least $\zeta/\sqrt{D}$ away from each other. In other words, the centers contained in any $B(z_i,\zeta)$ form a $(\zeta/\sqrt{D})$-packing of it. By a standard duality, any r-packing of a space is of size at most that of the minimum $(r/2)$-cover of the space. In this case the ball $B(z_i,\zeta) \subset \mathbb{R}^D$ has a minimum $(\zeta/2\sqrt{D})$-cover of size at most $(2\sqrt{D})^{c_o D}$ (recall from Lemma 2 that $\mathbb{R}^D$ has doubling dimension $\leq c_o D$ for some constant $c_o < 3$).

Thus $|\mathcal{A}_\mathcal{X}| \leq M \cdot N \leq 2^{C_3 D\log D}(1/\zeta)^d$ (for some constant C_3) and we conclude by plugging this value into (A.1). □

Appendix B. A more general setting

Finally, we consider a more general setting where the space $\mathcal{X} \subset \mathbb{R}^D$ has low doubling dimension $d \ll D$ only in sufficiently small neighborhoods (as in Figure 4). In this case, an RP tree might initially decrease diameter slowly; but when its cells are small enough, further splits will rapidly decrease diameter. We will show that the higher dimensionality of large regions of space do not tremendously affect the final excess risk, provided n is large enough for the tree to arrive at well populated regions of sufficiently small diameter.

Appendix B.1. Result for the general case

The next definition of decrease rate is made more general by allowing for a good rate k to be attained only later down the tree; in other words we allow for speedups to occur only in smaller regions of $\mathcal{X}$, of diameter at most $2r < \Delta_{\mathcal{X}}$. The algorithm remains unchanged except that we now need $\alpha(n) \geq \left(\log^2 n\right) \log\log(n/\delta) + \log(\mathcal{N}_r/\delta)$, where $\mathcal{N}_r$ is the size of a minimal r-cover of $\mathcal{X}$. Note that $\mathcal{N}_r \leq (\Delta_{\mathcal{X}}/r)^{O(D)}$.

Definition 25. *Given a sample* $\mathbf{X}$*, we say that* `adaptiveRPtree` *attains a* ***diameter decrease rate*** *of* (k, γ) *on* $\mathbf{X}$*, for* $k \geq d$ *and* $\gamma \leq \frac{n}{\alpha(n)}$*, if the following holds:*
`adaptiveRPtree` *arrives at an intermediate partition* $\mathcal{A}^{i_\gamma}$*,* $\left|\mathcal{A}^{i_\gamma}\right| = \gamma$*, such that any subsequent call to* `coreRPtree`$(A, \Delta_n(A)/2, \delta)$ *over cells* A *with ancestor in* $\mathcal{A}^{i_\gamma}$*, returns a tree rooted at* A *of height at most* k*.*

Theorem 26. *Assume that for every ball* $B \in \mathbb{R}^D$ *of radius* r*,* $B \cap \mathcal{X}$ *has doubling dimension* d*. There exist constants* C*,* C' *independent of* d *and* μ*, and* $C'' = C''(\mu, r)$ *such that the following holds.*

Suppose the cross-validation option is used with $\alpha(n) \geq \left(\log^2 n\right) \log\log(n/\delta) + \log(\mathcal{N}_r/\delta)$*. Assume* $n \geq \max\left\{(\lambda\Delta_{\mathcal{X}}/\Delta_{\mathcal{Y}})^2, C''\alpha(n)\right\}$*. With probability at least* $1-\delta$*, the algorithm attains a diameter decrease rate of* (k, γ) *where* $k \leq C' d \log d$ *and* $\gamma \leq C''$*, and the excess risk of the regressor satisfies*

$$\|f_n - f\|^2 \leq C \cdot (\lambda\Delta_{\mathcal{X}})^{2k/(2+k)} \left(\frac{\Delta_{\mathcal{Y}}^2 \cdot \gamma \cdot \alpha(n)}{n}\right)^{2/(2+k)} + 2\Delta_{\mathcal{Y}}^2 \sqrt{\frac{\ln\log n^6 + \ln 1/\delta}{2n}}.$$

Appendix B.2. Proof of theorem 26

The proof of theorem 26 closely mirrors that of theorem 9. We'll therefore only show the key lemmas whose statement change. We assume in what follows that the cross-validation option is used.

The proof proceeds also by first bounding the risks in terms of the observed diameter decrease rate (lemma 30 of section Appendix B.2.1), and then bounding the worst case decrease rates (lemma 32 of section Appendix B.2.2).

Appendix B.2.1. Risk bound in terms of observed diameter decrease rate

Lemma 27 (Mass of cells of $\mathcal{A}'$)**.** *With probability at least $1-\delta'$ over $\mathcal{X}$ and the randomness in the algorithm, we have for all partitions $\mathcal{A} = \mathcal{A}^0, \mathcal{A}^1, \ldots$ found by* `adaptiveRPtree`*, for all $A' \in \mathcal{A}'$ that*

$$\begin{aligned} \mu(A') &\leq \mu_n(A') + 2\sqrt{\mu_n(A')\frac{\mathcal{V}+\ln(4/\delta')}{n}} + 4\frac{\mathcal{V}+\ln(4/\delta')}{n}, \text{ where} \\ \mathcal{V} &\leq O\left(\log n(\log n + \text{loglog}(1/\delta)) + \log \mathcal{N}_r\right). \end{aligned} \tag{B.1}$$

Proof. Follow the outline of lemma 15, the only difference being that the bound on $|\mathcal{B}|$ introduces the term $\mathcal{N}_r$. □

Lemma 28 (Excess risk)**.** *There exists a constant C_1 independent of d and μ such that the following holds with probability at least $1-\delta/3$ over the choice of $(\mathbf{X}, \mathbf{Y})$ and the randomness in the algorithm.*

Let $\alpha(n) \geq \left(\log^2 n\right)\log\log(n/\delta) + \log(\mathcal{N}_r/\delta)$. Let $\mathcal{A}^i$ be the final partition reached by `adaptiveRPtree`*. For all partitions $\mathcal{A} \in \left\{\mathcal{A}^j\right\}_{j=0}^{i}$, we have*

$$\|f_{n,\mathcal{A}} - f\|^2 \leq C_1\left(\Delta_{\mathcal{Y}}^2 |\mathcal{A}| \frac{\alpha(n)}{n} + \lambda^2\left(\Delta_n^2(\mathcal{A}) + n^{-4/(2+d)}\Delta_{\mathcal{X}}^2\right)\right).$$

Proof. The proof is identical to that of lemma 16, using lemma 27 in place of lemma 15. □

Lemma 29 (Existence of a good pruning)**.** *Suppose the cross-validation option is used, and* `adaptiveRPtree` *attains a diameter decrease rate of (k, γ) on $\mathbf{X}$. Let $\alpha(n) \geq \left(\log^2 n\right)\log\log(n/\delta) + \log(\mathcal{N}_r/\delta)$, and $\zeta \doteq \left(\frac{\Delta_{\mathcal{Y}}^2 \cdot \gamma \cdot \alpha(n)}{\lambda^2 \Delta_{\mathcal{X}}^2 \cdot n}\right)^{1/(2+k)}$. Finally, assume $n \geq \max\left\{\left(\lambda\Delta_{\mathcal{X}}/\Delta_{\mathcal{Y}}\right)^2, \gamma \cdot \alpha(n)\right\}$. Then there exists an RPtree partition $\mathcal{A}$ such that $|\mathcal{A}| \leq \gamma \cdot \zeta^{-k}$ and $\Delta_n(\mathcal{A}) \leq 2\zeta \cdot \Delta_n(\mathcal{X})$.*

Proof. Follow the outline of lemma 17, while noticing that now we have for all $i \geq 1$, $\text{level}\left(\mathcal{A}^i\right) \leq ki + \log\gamma$ and $\Delta_n\left(A^i\right) \leq 2^{-i}\Delta_n(\mathcal{X})$. □

Lemma 30. *There exists a constant C independent of d and μ such that the following holds with probability at least $1 - 2\delta/3$ over $(\mathbf{X}, \mathbf{Y})$ and the randomness in the algorithm.*

Suppose the cross-validation option is used, and `adaptiveRPtree` *attains a diameter decrease rate of (k, γ) on $\mathbf{X}$. Let $\alpha(n) \geq \left(\log^2 n\right)\log\log(n/\delta) + \log(\mathcal{N}_r/\delta)$, and assume $n \geq \max\left\{\left(\lambda\Delta_{\mathcal{X}}/\Delta_{\mathcal{Y}}\right)^2, \gamma \cdot \alpha(n)\right\}$. The excess risk of the regressor is then bounded as*

$$\|f_n - f\|^2 \leq C \cdot \left(\lambda\Delta_{\mathcal{X}}\right)^{2k/(2+k)}\left(\frac{\Delta_{\mathcal{Y}}^2 \cdot \gamma \cdot \alpha(n)}{n}\right)^{2/(2+k)} + 2\Delta_{\mathcal{Y}}^2\sqrt{\frac{\ln\log n^6 + \ln 1/\delta}{2n}}.$$

Proof. Follow the outline of lemma 18. □

Appendix B.2.2. Worst case decrease rates

Lemma 31. *Assume that for every ball $B \in \mathbb{R}^D$ of radius r, $B \cap \mathcal{X}$ has doubling dimension d and consider the tree built by* `adaptiveRPtree`*. There exists a constant $C'' = C''(\mu, r)$, such that with probability at least $1 - \delta/3$ over the randomness in the algorithm, we have $\Delta_n(A) \leq r$ for all cells A of the tree at level at least $\log C''$.*

Proof outline. This is a consequence of the fact that $\mathcal{X}$ has finite doubling dimension at most $O(D)$. By theorem 22 and the fact that `basicRPtree` is called multiple times to boost the probability of obtaining a small tree (see proof of corollary 23) we have the following: with probability at least $1 - \delta/3$, and independently of the distribution, it takes at most a constant number of levels to get the data diameter within the cells below r.

The number of levels needed for each particular distribution is therefore just a constant. □

Lemma 32. *Assume that for every ball $B \in \mathbb{R}^D$ of radius r, $B \cap \mathcal{X}$ has doubling dimension d. There exist constants C independent of $\mathcal{X}$ and d, and $C'' = C''(\mu, r)$, such that with probability at least $1 - \delta/3$, the algorithm attains a diameter decrease rate of (k, γ) where $k \leq C' d \log d$ and $\gamma \leq C''$.*

Proof. This results from lemma 31 and theorem 22. □