Escaping the curse of dimensionality with a tree-based regressor

Samory Kpotufe
UCSD CSE
skpotufe@cs.ucsd.edu

Abstract

We present the first tree-based regressor whose convergence rate depends only on the intrinsic dimension of the data, namely its Assouad dimension. The regressor uses the RPtree partitioning procedure, a simple randomized variant of k-d trees.

1 Introduction

Non-parametric learning algorithms tend to suffer from what is referred to as the curse of dimensionality, namely that prediction performance deteriorates dramatically as the number of features increases. This phenomenon is quantifiable in the case of regression algorithms: as initially shown by Stone [Sto80, Sto82], if we only assume that the regression function $f(x)$ is Lipschitz [1] in $\mathbb{R}^D$, then no non-parametric estimator can achieve a convergence rate faster than $n^{-2/(2+D)}$. In other words, the number of points required to attain a low risk may be exponential in D, and this is infeasible even for moderate values of D.

However, it is often the case that data which appears high dimensional, actually conforms to a structure of low intrinsic dimensionality (interpreted broadly). Examples of such situations are traditional continuous settings where the data is close to a low dimensional submanifold of $\mathbb{R}^D$, and discrete settings such as when the data is sparse. These are all examples of data with low Assouad dimension (see definition 1); this notion of dimension thus offers a natural and broad model of intrinsic data complexity.

We show that, for any input data distribution, the risk of a regressor based on RPtree (a variant of k-d tree) depends just on the unknown Assouad dimension of the data, regardless of the ambient dimension D. This is the first such result for tree-based regression.

1.1 Tree-based regression

Tree-based regression consists of first building a hierarchy of nested partitions of the data space (the tree), and then learning a piecewise continuous function f_n over the cells of some chosen partition in the hierarchy. Future evaluations of $f_n(x)$ can be done in time just $O(\log n)$ by navigating

---

[1]Stone's result concerns a much larger class of regression functions; here we focus on Lipschitz conditions.
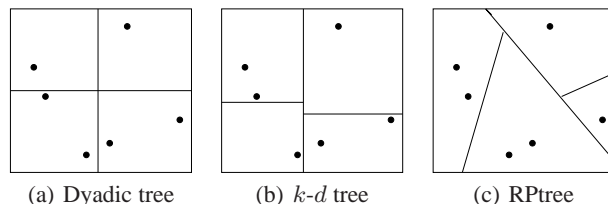

Figure 1: Spatial partitioning induced by various splitting rules. Two levels or the tree are shown for each.

the usually shallow tree down to an appropriate cell. These methods are popular due to their ease of use and computational efficiency (e.g. CART, dyadic trees, k-d tree, see [GN05, SN06, LGL96]), but none has been shown to adapt to intrinsic dimensionality in terms of their regression risk. See figure 1 for some examples.

The Random Projection tree (RPtree) is a hierarchical partitioning procedure which recursively bisects the data space with random hyperplanes (see figure 1(c)). Although RPtree's connections to intrinsic data dimensionality has been studied in unsupervised settings ([DF08, GLZ08]), its use for regression has not been explored.

Using RPtrees for regression requires a method for selecting a partition on which to learn the regressor f_n. Selecting a good partition from the hierarchy is essential to balancing the bias and variance of the regressor. Traditional methods 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 careful steps that enable us to quickly identify a small set of candidate partitions. We then provide a couple of options for selecting the final partition: one is to use cross-validation over the candidate partitions, another is a criterion which allows to automatically stop growing the tree when a good partition is attained. The latter method is computationally cheaper, while the former method results in a slightly better risk. In both cases, the excess risk of the RPTree regressor depends just on the unknown Assouad dimension of the input space, for all distributions.

On the technical side, RPtree regression requires novel techniques for analyzing the bias of the estimator. Estimator bias is well understood to decrease with the diameters of the partition's cells. Unfortunately these *physical diameters* are hard to assess for RPtrees given the random and irregu-

lar shapes of the cells, and in fact they may not decrease at all. However, we can track the diameters of the data within the cells, and we develop new techniques to relate these empirical *data diameters* to the estimator's bias. We believe these techniques are of independent interest as they take focus away from the cells' physical diameters, thus opening the door to richer partitioning rules whose cell diameters are hard to control.

1.2 Background and related work

The realization that data is often less complex than indicated by the ambient dimension has spurred a significant body of work (referred to as manifold learning) that aim to embed the data into a low dimensional euclidean space (see e.g. [RS00, BN03, TSL00]). A possible approach to regression on high dimensional data is to first reduce dimension using manifold learning and learn the regressor in the new space. Unfortunately, this approach is not guaranteed to work since pertinent information may be lost by the embedding. This raises the following natural question: can learning methods such as regression adapt automatically to data that has low intrinsic dimensionality while operating in the original space $\mathbb{R}^D$?

An important result in the direction of adaptive regression is the realization by Bickel and Li [BL06] that standard kernel regressors are adaptive in the following sense: there exists an appropriate bandwidth setting such that the asymptotic pointwise risk at $x \in \mathbb{R}^D$ depends just on the manifold dimension and on the behavior of the kernel in a neighborhood of x. One then has to search for the appropriate bandwidth setting, either by estimating the manifold dimension or through cross validation over all possible values of this dimension (see e.g. [BL06, LW07]).

Kernel regressors can be expensive in practice: the kernel weights must be computed anew at each training point in order to evaluate the regressor on a new data point. This translates into an evaluation time of $\Omega(n)$ which is often a burden given large samples. Contrast this with the $O(\log n)$ evaluation time of tree-based regressors.

In the case of classification, a recent result by Scott and Nowak ([SN06]) for dyadic decision trees is related: they show that if the input data is drawn from an approximately uniform measure on a manifold, and the Bayes decision boundary is sufficiently smooth, DDTs achieve classification rates that depend just on the manifold dimension. It is unclear whether their result will apply in a distribution free regression setting.

2 Detailed overview of results

We're given i.i.d training data $(\mathbf{X}, \mathbf{Y}) = \{(X_i, Y_i)\}_{i=1}^n \in (\mathcal{X} \times \mathcal{Y})^n$, where the input space $\mathcal{X} \subset \mathbb{R}^D$ is contained in a ball[2] of (unknown) diameter $\Delta_{\mathcal{X}}$, and the output space $\mathcal{Y} \subset \mathbb{R}^{D'}$ is contained in a ball of (unknown) diameter $\Delta_{\mathcal{Y}}$.

2.1 Assouad dimension

We model the intrinsic dimensionality of the space $\mathcal{X}$ using the notion of Assouad dimension defined below.

---
[2]We assume a Euclidean l_2 norm in this work.

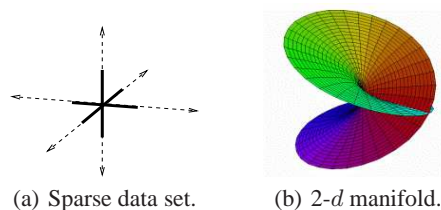(a) Sparse data set.   (b) 2-d manifold.

Figure 2: Examples of data with low Assouad dimension.

Definition 1 *The Assouad dimension (or 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.*

The Assouad dimension has proved useful in capturing the intrinsic complexity of data spaces as shown in various works on data analysis (see e.g. [IN07, BKL06, Cla05]). It coincides with the natural notions of dimension of various geometric objects: it is easy to see that d-dimensional cubes, spheres, all have Assouad dimension $O(d)$ (see e.g. [Cla05]). It also captures notions of data complexity that are standard in the machine learning and statistics communities; this is stated in the following remarks for emphasis.

Remark 1 *A d-dimensional hyperplane in $\mathbb{R}^D$ has Assouad dimension $O(d)$ (see [Cla05]).*

Remark 2 *A d-dimensional Riemannian submanifold of $\mathbb{R}^D$ has Assouad dimension $O(d)$, subject to a bound on its curvature (see theorem 22 of [DF08]).*

Remark 3 *A d-sparse data space in $\mathbb{R}^D$, i.e. one where each data point has at most d non zero coordinates, has Assouad dimension $O(d \log D)$: it can be described by $\binom{D}{d} \leq D^d$ hyperplanes of dimension d.*

2.2 Notions of diameter

Let $\mathbf{A}$ be some partition of $\mathcal{X}$. Traditionally, bias analysis revolves around the *physical diameters* $\Delta(A) \doteq \max_{x,x' \in A} \|x - x'\|$ of cells $A \in \mathbf{A}$ (see e.g. [GN05, SN06, LGL96]). In this work we instead relate bias to the *data diameters* of the cells, that is $\Delta_n(A) \doteq \max_{x,x' \in A \cap \mathbf{X}} \|x - x'\|$ or 0 if $A \cap \mathbf{X} = \emptyset$.

Cell $A \in \mathbf{A}$
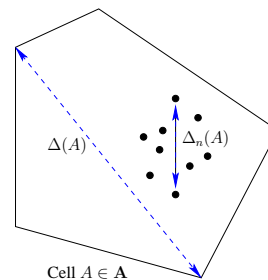
Focusing on data diameter has the following advantage. We never need to evaluate the physical diameters of the cells, and these need not decrease. Consequently, we don't have to

constrain the partition to regular shaped cells (e.g. axis parallel hyper-rectangles) whose physical diameters are easily controlled. In particular, it opens the door to richer partitioning rules such as RPtree which adapt better to the data complexity at the expense of creating irregular cells. We expand on this last point in the example below.

Consider a data space of the following form:

$$\cup_{i \neq j}\{te_i \pm \varepsilon e_j : t \in [-1, 1]\},\ i, j \in [D], \text{for a fixed } \varepsilon \ll 1.$$

This is an extreme case of a noisy sparse data set of Assouad dimension $O(\log D)$, depicted in figure 2(a). We'd like to partition this space in a way that reduces the data diameters of the cells (for low estimator bias) while achieving a small partition size (for low estimator variance). Axis parallel splitting rules such as k-d trees or dyadic trees would require a number of cells exponential in D in order to halve the diameters. Yet, the set itself can be partitioned into at most $2D^2$ cells of half its radius. The richness of random splits allows us to achieve a partitioning just a bit larger than this, even in the worst case over distributions on the set. In fact, given any data set of Assouad dimension d, RPtrees are guaranteed to achieve a partition of size at most $2^{\tilde{O}(d)}$, such that the data diameters of each cell is at most half of the diameter of the full data set. We refer the reader to [DF08] for a detailed analysis.

We'll soon see that, for low estimator bias, we don't need every cell of a partition to have small data diameter, but rather that these diameters are small in an average sense. Given a collection $\mathbf{A}$ of disjoint subsets of $\mathcal{X}$, we define the following notion of average data diameter:

$$\Delta_n(\mathbf{A}) \ \doteq\ \left(\frac{\sum_{A \in \mathbf{A}} \mu_n(A)\Delta_n^2(A)}{\sum_{A \in \mathbf{A}} \mu_n(A)}\right)^{1/2},$$

where μ_n is the empirical measure over $\mathbf{X}$ (we'll let μ denote the marginal measure over $\mathcal{X}$).

2.3 Regression setup

We assume that the regression function $f(x) = \mathbb{E}[Y|X = x]$ is λ-Lipschitz, for an unknown parameter λ:

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

For any function $g(x) : \mathcal{X} \mapsto \mathcal{Y}$, the l_2 pointwise risk at x satisfies

$$R(g(x)) \doteq \mathbb{E}_Y \|Y - g(x)\|^2 = R(f(x)) + \|f(x) - g(x)\|^2,$$

and the integrated risk can then be written as

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

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

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

2.4 Choosing a good partition for regression

A tree-based regressor works in two phases. The partitioning phase returns a partition $\mathbf{A}$ of the data space $\mathcal{X}$ and a final regressor is learned as a piecewise continuous function over the cells of $\mathbf{A}$. In this work we'll consider a piecewise constant regressor over the returned partition $\mathbf{A}$ defined as follows:

For $x \in \mathcal{X}$, let $\mathbf{A}(x)$ be the cell of $\mathbf{A}$ to which x belongs. If $\mu_n(\mathbf{A}(x)) > 0$, the regressor is obtained as

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

otherwise use a default setting $f_{n,\mathbf{A}}(x) = y_0 \in \mathcal{Y}$ whenever $\mathbf{A}(x)$ is empty of training points. We'll often refer to the final regressor as $f_n(\cdot)$ as long as the partition used for the estimate is clear from context.

Procedure adaptiveRPtree makes calls to the the subprocedure coreRPtree which implements the basic RPtree splits. We defer the complete treatment of this subprocedure to section 5.1 since most of the analysis will concern adaptiveRPtree. For now, note that the call to coreRPtree returns a subtree rooted at A with the following property: let $\mathbf{A}$ be the collection of subsets of $\mathbb{R}^D$ defined by the leaves of this subtree, we have $\Delta_n(\mathbf{A}) \leq \Delta_n(A)/2$. Also, the implementation of coreRPtree ensures that the final tree built by adaptiveRPtree has height at most $6 \log n$.

Procedure adaptiveRPtree grows the tree in steps $\mathbf{A}^0$, $\mathbf{A}^1, \ldots$, where $\Delta_n\left(\mathbf{A}^{i+1}\right) \leq \Delta_n\left(\mathbf{A}^i\right)/2$, and eventually returns one of the partitions $\mathbf{A}^i$ for some i. We present a couple of options for selecting a good partition to return. The first option uses cross-validation: grow a large tree and prune it back by minimizing empirical risks over an i.i.d test sample $(\mathbf{X}', \mathbf{Y}')$ of size n. The other option is that of automatic stopping: we return a partition as soon as some stopping condition is met.

The two options for selecting the return partition are outlined in procedure adaptiveRPtree. The empirical risk in the cross-validation option is defined as

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

The automatic stopping option returns one of two partitions and requires no test sample. It is a computationally faster option and, as we'll see, the resulting bounds are only marginally worsened.

2.5 Main Results

Definition 2 *Given a sample* $\mathbf{X}$*, we say that* adaptiveRPtree *attains a diameter decrease rate of k on $\mathbf{X}$ for $k \geq d$, if every call to the subprocedure* coreRPtree $(A, \Delta_n(A)/2, \delta, l)$ *in the second loop of the procedure returns a tree rooted at A of depth at most k.*

Theorem 3 *Assume that $\mathcal{X}$ has Assouad dimension d. There exist constants C, C' independent of d and $\mu(\mathcal{X})$, such that the following holds.*

Suppose the cross-validation option is used. Define

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

and assume $n \geq \max\left\{(\lambda\Delta_{\mathcal{X}}/\Delta_{\mathcal{Y}})^2, \alpha(n)\right\}$. With probability at least $1 - \delta$, the algorithm attains a diameter decrease rate of $k \leq C'd\log d$, and the excess risk of the regressor

```
Procedure adaptiveRPtree(sample X, confidence parameter δ)
    A⁰ ← {𝒳};
    for i ← 1 to ∞ do
        foreach cell A ∈ Aⁱ⁻¹ do
            // Create a subtree rooted at A:
            l ← level(A) in the current tree ; // Root is at level 0
            (subtree rooted at A) ← coreRPtree(A, Δₙ(A)/2, δ, l);
        end
        Aⁱ ← partition of 𝒳 defined by the leaves of the current tree;
        level(Aⁱ) ← max_{A∈Aⁱ} level(A) ;
        // At this point we have two options for stopping and returning a partition.
        ┌─────────────────────────────────┐
        │ Option 1:  Cross-validation     │
        └─────────────────────────────────┘
        if Δₙ(Aⁱ) = 0 or level(Aⁱ) ≥ log n² then
            Draw test sample (X', Y') of size n and define R'ₙ(·) as the empirical risk over the test sample;
            A* ←    argmin      R'ₙ(f_{n,Aʲ});
                 Aʲ∈{A⁰,...,Aⁱ}
            return fₙ ≐ f_{n,A*};
        end
        ┌─────────────────────────────────┐
        │ Option 2:  Automatic stopping   │
        └─────────────────────────────────┘
        α(n) ← (log² n) log log(n/δ) + log(1/δ);
        if level(Aⁱ) ≥ log (n · Δₙ²(Aⁱ)/α(n)Δₙ²(𝒳)) then
            A* ←    argmin     ( α(n)/n · |Aʲ| + Δₙ²(Aʲ) );
                 Aʲ∈{Aⁱ⁻¹, Aⁱ}
            return fₙ ≐ f_{n,A*};
        end
    end
```

satisfies

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

$$+ 2\Delta_{\mathcal{Y}}^2 \sqrt{\frac{\ln\log n^6 + \ln 3/\delta}{2n}}.$$

Theorem 4 *Assume that $\mathcal{X}$ has Assouad dimension d. There exist constants C, C' independent of d and $\mu(\mathcal{X})$, such that the following holds.*

Suppose the automatic stopping option is used. Define

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

With probability at least $1 - \delta$, the algorithm attains a diameter decrease rate of $k \leq C'd\log d$, and the excess risk of the regressor satisfies

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

Analysis outline

We start in section 3 by laying out the necessary tools for the rest of the analysis.

The theorems are then proved in two parts. First we bound the excess risk of the algorithm in terms of the observed diameter decrease rates in section 4 (lemma 13 for the cross-validation option, and lemma 15 for the automatic stopping). We subsequently argue that these decrease rates depend just on the intrinsic dimensionality of the data (corollary 17 of section 5).

Theorem 3 results from lemma 13 and corollary 17, while theorem 4 results from lemma 15 and corollary 17.

3 Proof preliminaries: risk bound for $f_{n,\mathbf{A}}$

In this section we develop the necessary tools to bound the excess risk of $f_{n,\mathbf{A}}$, where $\mathbf{A}$ is an RPtree partition, i.e. $\mathbf{A}$ is defined by the leaves of some subtree of the tree returned by `adaptiveRPtree`.

3.1 Generic decomposition of excess risk

We start the analysis with a standard decomposition of the excess risk into bias and variance terms. Let $\mathbf{A}$ be any partition of $\mathcal{X}$. The following function of $x \in \mathcal{X}$ provides a bridge between the regressor $f_{n,\mathbf{A}}$ and the regression function f:

$$\widetilde{f}_{n,\mathbf{A}}(x) \;\doteq\; \mathbb{E}_{\mathbf{Y}|\mathbf{X}} f_{n,\mathbf{A}}(x) = \frac{\sum_{i=1}^n f(X_i)\mathbb{1}_{X_i\in\mathbf{A}(x)}}{n\mu_n(\mathbf{A}(x))},$$

if $\mu_n(\mathbf{A}(x)) \neq 0$, otherwise we set $\widetilde{f}_{n,\mathbf{A}}(x) = y_0 \in \mathcal{Y}$.

The pointwise excess risk can be bounded as

$$\|f_{n,\mathbf{A}}(x) - f(x)\|^2 \;\leq\; 2\left\|f_{n,\mathbf{A}}(x) - \widetilde{f}_{n,\mathbf{A}}(x)\right\|^2$$

$$+ 2\left\|\widetilde{f}_{n,\mathbf{A}}(x) - f(x)\right\|^2. \quad (1)$$

We therefore proceed by bounding each term on the r.h.s separately in the following two lemmas.

Lemma 5 (Variance) *Let* $\mathbf{A}$ *be a partition of* $\mathcal{X}$. *The following inequality holds for all* $x \in \mathcal{X}$ *s.t.* $\mu_n(\mathbf{A}(x)) > 0$, *with probability at least* $1 - \delta'$ *over the random choice of* $\mathbf{Y}$ *for* $\mathbf{X}$ *fixed:*

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

Proof: Fix $\mathbf{X}$. Now fix $A \in \mathbf{A}$, and let $x \in A$. We'll consider $\mathbf{Y}_A \doteq \{Y_i \in \mathbf{Y} \text{ s.t. } X_i \in A\}$. Write:

$$\begin{aligned}
\psi(\mathbf{Y}_A) &\doteq \left\| f_{n,\mathbf{A}}(x) - \widetilde{f}_{n,\mathbf{A}}(x) \right\| \\
&= \left\| \frac{\sum_{i=1}^n (Y_i - f(X_i)) \mathbb{1}_{X_i \in A}}{n\mu_n(A)} \right\|.
\end{aligned}$$

We can now apply McDiarmid's inequality to $\psi(\cdot)$, as it is easy to verify that, changing one of the Y values in $\mathbf{Y}_A$ changes the value of $\psi(\cdot)$ by at most $\frac{\Delta_{\mathcal{Y}}}{n\mu_n(A)}$. We then have that,

$$\psi(\mathbf{Y}_A) \leq \mathbb{E}\,\psi(\mathbf{Y}_A) + \Delta_{\mathcal{Y}} \cdot \sqrt{\frac{\ln(|\mathbf{A}|/\delta')}{2n\mu_n(A)}}$$

with probability at least $1 - \delta'/|\mathbf{A}|$ over the random choice of $\mathbf{Y}_A$.

The expectation can be bounded as follows

$$\begin{aligned}
\mathbb{E}\,\psi(\mathbf{Y}_A) &\leq \left(\mathbb{E}\,(\psi(\mathbf{Y}_A))^2 \right)^{1/2} \\
&= \left(\mathbb{E} \left\| \frac{\sum_{i=1}^n (Y_i - f(X_i)) \mathbb{1}_{X_i \in A}}{n\mu_n(A)} \right\|^2 \right)^{1/2} \\
&\leq \left(\frac{\sum_{i=1}^n \mathbb{E} \left\| Y_i - f(X_i) \right\|^2 \mathbb{1}_{X_i \in A}}{(n\mu_n(A))^2} \right)^{1/2} \\
&\leq \left(\frac{\sum_{i=1}^n \Delta_{\mathcal{Y}}^2 \mathbb{1}_{X_i \in A}}{(n\mu_n(A))^2} \right)^{1/2} \\
&= \frac{\Delta_{\mathcal{Y}}}{\sqrt{n\mu_n(A)}}.
\end{aligned}$$

The first inequality above is an application of Jensen's inequality. The second inequality results from the fact that, for independent random vectors v_i with null expectation, we have $\mathbb{E} \left\| \sum_i v_i \right\|^2 = \sum_i \mathbb{E} \left\| v_i \right\|^2$; here we just take v_i to be $(Y_i - f(X_i))\mathbb{1}_{X_i \in A}/(n\mu_n(A))$.

Combining the above yields the desired bound on $\psi(\mathbf{Y}_A)$ with probability at least $1 - \delta'/|\mathbf{A}|$. We then conclude with a union bound over all $A \in \mathbf{A}$. ∎

Lemma 6 (Bias) *Let* $\mathbf{A}$ *be a partition of* $\mathcal{X}$. *The following inequality holds for all* $x \in \mathcal{X}$ *s.t.* $\mu_n(\mathbf{A}(x)) > 0$:

$$\left\| \widetilde{f}_{n,\mathbf{A}}(x) - f(x) \right\|^2 \leq \lambda^2 \Delta^2(\mathbf{A}(x)). \quad (3)$$

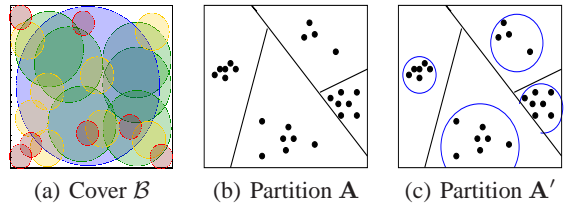(a) Cover $\mathcal{B}$ (b) Partition $\mathbf{A}$ (c) Partition $\mathbf{A}'$

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

Proof: Fix $A \in \mathbf{A}$ and let $x \in A$. Now write

$$\begin{aligned}
\left\| \widetilde{f}_{n,\mathbf{A}}(x) - f(x) \right\|^2 &= \left\| \frac{\sum_{i=1}^n (f(X_i) - f(x)) \mathbb{1}_{X_i \in A}}{n\mu_n(A)} \right\|^2 \\
&\leq \left(\frac{\sum_{i=1}^n \left\| f(X_i) - f(x) \right\| \mathbb{1}_{X_i \in A}}{n\mu_n(A)} \right)^2 \\
&\leq \left(\frac{\sum_{i=1}^n \lambda \left\| X_i - x \right\| \mathbb{1}_{X_i \in A}}{n\mu_n(A)} \right)^2 \\
&\leq \lambda^2 \Delta^2(A),
\end{aligned}$$

where the second inequality results from the Lipschitz condition on $f(\cdot)$. ∎

In lemma 6 above, the bias is bounded in terms of the *physical diameters* $\Delta(A)$. However, for an RPtree partition $\mathbf{A}$ (i.e. $\mathbf{A}$ is defined by the leaves of some subtree), the physical diameters $\{\Delta(A), A \in \mathbf{A}\}$ could be as large as $\Delta_{\mathcal{X}}$, the diameter of the whole space. As previously discussed, RPtree focuses on decreasing the *data diameters* $\Delta_n(A)$, and we'll argue that this is sufficient to decrease the bias of the estimator. For this purpose, we will replace RPtree partitions $\mathbf{A}$ with alternate partitions $\mathbf{A}'$ as explained in the next section.

3.2 Alternate partitions

Given a partition $\mathbf{A}$ built by RPtree, we will consider an alternate partition $\mathbf{A}'$ which will serve to analyze the bias of the regressor $f_{n,\mathbf{A}}$ (see above discussion of lemma 6). Each cell of $\mathbf{A}'$ will either contain no data point, or has physical diameter roughly the same as its data diameter. This is done by intersecting the cells of $\mathbf{A}$ with balls or complements of balls from a fixed collection $\mathcal{B}$ defined below (see figure 3). We'll see that $\mathbf{A}'$ approximately maintains key properties of $\mathbf{A}$, namely partition size and average data diameters.

Definition 7 *We define* $\mathcal{B}$ *as the following collection of balls in* $\mathbb{R}^D$. *Let* $I = \lfloor \log n^{2/(2+d)} \rfloor$. *For each* $i = 0$ *to* I, *consider a minimal* $(2^{-i}\Delta_{\mathcal{X}})$-*cover of* $\mathcal{X}$; *let* $\mathcal{B}_i$ *be the set of all balls* $B(z, 2^{-(i-2)}\Delta_{\mathcal{X}})$ *centered at points* z *in the cover. We set* $\mathcal{B} \doteq \cup_{i=0}^I \mathcal{B}_i$.

Every cell $A \in \mathbf{A}$ such that $A \cap \mathbf{X} \neq \emptyset$ will be replaced in $\mathbf{A}'$ by two cells A_1', A_2' obtained as follows.

Consider the smallest $i \in \{0, \ldots, I\}$ such that $2^{-i}\Delta_{\mathcal{X}} \leq \max\{\Delta_n(A), 2^{-I}\Delta_{\mathcal{X}}\}$, i.e. $i = \min\left\{I, \lceil \log \frac{\Delta_{\mathcal{X}}}{\Delta_n(A)} \rceil\right\}$.

There exists a ball $B \in \mathcal{B}_i$ which covers $A \cap \mathbf{X}$: pick any $x \in A \cap \mathbf{X}$, and pick the ball B in $\mathcal{B}_i$ whose center z is closest to x; we have $\forall x' \in A \cap \mathbf{X}$, that $x' \in B = B\left(z, 2^{-(i-2)}\Delta_{\mathcal{X}}\right)$ since by a triangle inequality

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

We define $A_1' = B \cap A$ and $A_2' = A \setminus A_1'$ for all $A \in \mathbf{A}$, $A \cap \mathbf{X} \neq \emptyset$; on the other hand we let $A_1' = A$, $A_2' = \emptyset$ for all $A \in \mathbf{A}$, $A \cap \mathbf{X} = \emptyset$. We finally define $\mathbf{A}'$ to be the collection of all such A_1', A_2' over $A \in \mathbf{A}$.

In the following lemma we relate diameters of cells of $\mathbf{A}'$ to the data diameters of cells of $\mathbf{A}$.

Lemma 8 (Diameters of $\mathbf{A}'$) *Let $\mathbf{A}$ be some partition of $\mathcal{X}$ and let $\mathbf{A}'$ as defined above. We have that*

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

Proof: Let $A \in \mathbf{A}$, $A \cap \mathbf{X} \neq \emptyset$. We have $\mu_n(A_1') = \mu_n(A)$ and $\mu_n(A_2') = 0$. Also, given the smallest $i \in \{0, \ldots, I\}$ such that $2^{-i}\Delta_{\mathcal{X}} \leq \max\{\Delta_n(A), 2^{-I}\Delta_{\mathcal{X}}\}$, we have that

- $\Delta_n(A) > 2^{-I}\Delta_{\mathcal{X}}$ implies $\Delta(A_1') \leq 2 \cdot 2^{-(i-2)}\Delta_{\mathcal{X}} \leq 8\Delta_n(A)$,

- $\Delta_n(A) \leq 2^{-I}\Delta_{\mathcal{X}}$ implies $\Delta(A_1') \leq 2 \cdot 2^{-(I-2)}\Delta_{\mathcal{X}} \leq 16n^{-2/(2+d)} \cdot \Delta_{\mathcal{X}}$.

Therefore, let $\mathbf{A}_+ = \{A \in \mathbf{A}, \Delta_n(A) > 2^{-I}\Delta_{\mathcal{X}}\}$, we have

$$
\begin{aligned}
\sum_{A' \in \mathbf{A}'} \mu_n(A')\Delta^2(A') &= \sum_{A \in \mathbf{A}_+} \mu_n(A)\Delta^2(A_1') \\
&\quad + \sum_{A \in \mathbf{A} \setminus \mathbf{A}_+} \mu_n(A)\Delta^2(A_1') \\
&\leq \sum_{A \in \mathbf{A}_+} 64\mu_n(A)\Delta_n^2(A) \\
&\quad + \sum_{A \in \mathbf{A} \setminus \mathbf{A}_+} 256\mu_n(A)n^{-\frac{4}{2+d}} \cdot \Delta_{\mathcal{X}}^2 \\
&\leq 64\Delta_n^2(\mathbf{A}) + 256n^{-\frac{4}{2+d}} \cdot \Delta_{\mathcal{X}}^2.
\end{aligned}
$$

∎

In order to bound the integrated excess risk, we'll need the empirical mass of cells of $\mathbf{A}'$ to be close to their true mass. In particular, this will allow us to effectively discard cells that are empty of data since they will have little effect on the integrated excess risk. The following lemma from VC theory will come in handy.

Lemma 9 (Relative VC bounds -[VC71]) *Let $\mathcal{C}$ be a class of subsets of $\mathbb{R}^D$, and let its $2n$-shatter coefficient be given by $\mathcal{S}(\mathcal{C}, 2n)$. With probability at least $1 - \delta'$ over the choice of $\mathbf{X}$, all $A' \in \mathcal{C}$ satisfy*

$$
\begin{aligned}
\mu(A') &\leq \mu_n(A') + 2\sqrt{\mu_n(A')\frac{\ln \mathcal{S}(\mathcal{C}, 2n) + \ln(4/\delta')}{n}} \\
&\quad + 4\frac{\ln \mathcal{S}(\mathcal{C}, 2n) + \ln(4/\delta')}{n}. \tag{4}
\end{aligned}
$$

The next lemma establishes the convergence of empirical masses of cells of $\mathbf{A}'$.

Lemma 10 (Mass of cells of $\mathbf{A}'$) *With probability at least $1 - \delta'$ over $\mathbf{X}$ and the randomness in the algorithm, we have for all RPtree partitions $\mathbf{A}$, for all $A' \in \mathbf{A}'$ that*

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

Proof: Suppose w.l.o.g that the RPtree is built by picking random directions from a fixed collection $\mathcal{P}$ without replacement. How big should $\mathcal{P}$ be so we have enough directions to choose from? The implementation of coreRPtree ensures that $|\mathcal{P}| \leq 2n^6 \log\left(6n^2/\delta\right)$ is sufficient (see remark 4 of section 5.1). Now fix such a collection $\mathcal{P}$ and let $\mathcal{H}_{\mathcal{P}}$ be the union of $\{\mathcal{X}\}$ and the class of half spaces of $\mathbb{R}^D$ defined by hyperplanes normal to the directions in $\mathcal{P}$. For an RPtree partition $\mathbf{A}$, each cell of $\mathbf{A}$ is the intersection of at most $6 \log n$ elements of $\mathcal{H}_{\mathcal{P}}$ since the tree is guaranteed to have height at most $6 \log n$ (remark 4). Each cell of $\mathbf{A}'$ is the intersection of a ball or the complement of a ball in $\mathcal{B}$ with a cell of $\mathbf{A}$.

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

$$
\mathcal{C} = \left\{ h : h = h_0 \cap \left(\bigcap_{l=1}^{6 \log n} h_l\right), h_0 \text{ or } h_0^{\complement} \text{ is in } \mathcal{B}, h_l \in \mathcal{H}_{\mathcal{P}} \right\}.
$$

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

Given $2n$ sample points, every direction $v \in \mathcal{P}$ defines at most $2(2n + 1)$ equivalent choices of half-spaces in $\mathbb{R}^D$. We therefore have

$$
\begin{aligned}
\mathcal{S}(\mathcal{C}, 2n) &\leq 2|\mathcal{B}|\left((4n + 2)|\mathcal{P}| + 1\right)^{6 \log n} \\
&\leq 2|\mathcal{B}|\left(n^6(8n + 4)\log\left(6n^2/\delta\right) + 1\right)^{6 \log n}.
\end{aligned}
$$

Since $\mathcal{X}$ has Assouad 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 calling on lemma 9. ∎

Lemma 11 (Excess risk) *There exists a constant C_1 independent of d and $\mu(\mathcal{X})$ 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.*

Define $\alpha(n) \doteq \left(\log^2 n\right)\log\log(1/\delta) + \log(1/\delta)$. Let $\mathbf{A}^i$ be the final partition reached by adaptiveRPtree. For all partitions $\mathbf{A} \in \{\mathbf{A}^j\}_{j=0}^{i}$, we have

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

Proof: Let the partition $\mathbf{A} \in \{\mathbf{A}^j\}_{j=0}^{i}$ and the sample $\mathbf{X}$ be fixed. By lemma 10 we have, with probability at least $1 - \delta'$, that equation (5) holds for all $A' \in \mathbf{A}'$ with $\mathcal{V} \leq O(\log n)(\log n + \log\log(1/\delta))$.

The excess risk decomposes over $\mathbf{A}'$ as

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

We next divide the cells of $\mathbf{A}'$ into two groups:

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

and $\mathbf{A}'_< \doteq \mathbf{A}' \setminus \mathbf{A}'_>$.

It's easy to see that from equation (4), we have $\forall A' \in \mathbf{A}'_>$, $\mu(A') \leq 7\mu_n(A')$, and $\forall A' \in \mathbf{A}'_<$, $\mu(A') \leq 7\frac{\mathcal{V}+\ln(4/\delta')}{n}$.

Integrating over $\mathbf{A}'_<$, we have

$$\sum_{A' \in \mathbf{A}'_<} \int_{A'} \|f_{n,\mathbf{A}}(x) - f(x)\|^2 \, \mu(dx)$$

$$\leq \sum_{A' \in \mathbf{A}'_<} \Delta_{\mathcal{Y}}^2 \cdot \mu(A')$$

$$\leq \sum_{A' \in \mathbf{A}'_<} \Delta_{\mathcal{Y}}^2 \cdot 7\frac{\mathcal{V}+\ln(4/\delta')}{n}$$

$$\leq 7\Delta_{\mathcal{Y}}^2 \cdot |\mathbf{A}'| \cdot \frac{\mathcal{V}+\ln(4/\delta')}{n}. \tag{5}$$

For the integration over $\mathbf{A}'_>$, we first apply (1), and recall lemmas 6 and 5 to have that with probability at least $1 - \delta'$ over $\mathbf{Y}$,

$$\sum_{A' \in \mathbf{A}'_>} \int_{A'} \|f_{n,\mathbf{A}}(x) - f(x)\|^2 \, \mu(dx)$$

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

$$\leq \sum_{A' \in \mathbf{A}'_>} 2\lambda^2\Delta^2(A') \cdot \mu(A')$$

$$+ \sum_{A' \in \mathbf{A}'_>} 2\Delta_{\mathcal{Y}}^2 \cdot \frac{2 + \ln(|\mathbf{A}'|/\delta')}{n\mu_n(A')} \cdot \mu(A')$$

$$\leq \sum_{A' \in \mathbf{A}'_>} 2\lambda^2\Delta^2(A') \cdot 7\mu_n(A')$$

$$+ \sum_{A' \in \mathbf{A}'_>} 2\Delta_{\mathcal{Y}}^2 \cdot \frac{2 + \ln(|\mathbf{A}'|/\delta')}{n\mu_n(A')} \cdot 7\mu_n(A')$$

$$\leq 14\lambda^2 \sum_{A' \in \mathbf{A}'_>} \mu_n(A')\Delta^2(A')$$

$$+ 14\Delta_{\mathcal{Y}}^2 |\mathbf{A}'| \cdot \frac{2 + \ln(|\mathbf{A}'|/\delta')}{n}. \tag{6}$$

Note that the term $\ln|\mathbf{A}'|$ in (6) is at most $O(\ln n)$ since the entire tree has height at most $6\log n$. Combining the bounds in (5) and (6), we get that there exists a constant C_0

such that $\|f_{n,\mathbf{A}} - f\|^2$ is at most

$$C_0 \left(\Delta_{\mathcal{Y}}^2 \cdot |\mathbf{A}| \frac{\log^2 n \log\log 1/\delta + \log(1/\delta')}{n} \right.$$

$$\left. + \lambda^2 \sum_{A' \in \mathbf{A}'} \mu_n(A')\Delta^2(A') \right),$$

with probability at least $1 - 2\delta'$.

Setting $\delta' = \delta/36\log n$, the lemma follows by a union bound over at most $6\log n$ partitions in $\{\mathbf{A}^j\}_{j=0}^{i}$, and then calling on lemma 8. ∎

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

In this section we bound the excess risk of the final regressor $f_n \doteq f_{n,\mathbf{A}^*}$ in terms of the diameter decrease rate attained when `adaptiveRPtree` stops.

To see that the stopping criteria eventually hold, note that the implementation of `coreRPtree` ensures that all cells at some level down the hierarchy have a single data point in them (see remark 4). In other words, we have $\Delta_n(\mathbf{A}^i) = 0$ eventually, forcing either stopping criterion to hold.

We now outline the arguments in this section. For simplicity, assume $\Delta_{\mathcal{X}}$, $\Delta_{\mathcal{Y}}$, and λ are all 1. Consider some RPtree partition $\mathbf{A}$ and let $\Delta_n(\mathbf{A}) \approx \zeta$ for some scalar ζ, we then have $|\mathbf{A}| \lesssim \zeta^{-k}$ where k is the diameter decrease rate attained by the algorithm. From lemma 11 above, we roughly have $\|f_{n,\mathbf{A}} - f\|^2 \lesssim \zeta^{-k}/n + \zeta^2$, and the best bound is obtained by setting $\zeta \approx n^{-1/(2+k)}$. Provided we pick an appropriate partition which optimizes ζ, the final bound would then take the form $\|f_{n,\mathbf{A}^*} - f\|^2 \lesssim n^{-2/(2+k)}$.

4.1 Risk bound for cross-validation option

Lemma 12 (Existence of a good pruning) *Suppose the cross-validation option is used, and* `adaptiveRPtree` *attains a diameter decrease rate of k on $\mathbf{X}$. Define*

$$\alpha(n) \doteq (\log^2 n) \log\log(n/\delta) + \log(1/\delta),$$

and $\zeta \doteq \left(\frac{\Delta_{\mathcal{Y}}^2 \cdot \alpha(n)}{\lambda^2 \Delta_{\mathcal{X}}^2 \cdot n} \right)^{1/(2+k)}$. Let $n \geq \max\left\{ \left(\frac{\lambda\Delta_{\mathcal{X}}}{\Delta_{\mathcal{Y}}}\right)^2, \alpha(n) \right\}$, and for $i \geq 0$, let $\mathbf{A}^i$ as defined in `adaptiveRPtree`. *Then there exists $i_0 \geq 0$ such that $\Delta_n(\mathbf{A}^{i_0}) \leq 2\zeta \cdot \Delta_n(\mathcal{X})$ and $|\mathbf{A}^{i_0}| \leq \zeta^{-k}$.*

Proof: Let $i \geq 0$. We have by definition that $\Delta_n(\mathbf{A}^i) \leq 2^{-i}\Delta_n(\mathcal{X})$, while it follows from the assumption on diameter decrease rate that $\text{level}(\mathbf{A}^i) \leq ki$. Now let $\mathbf{A}^i$ be the last partition of $\mathcal{X}$ achieved by `adaptiveRPtree` when the stopping criteria holds. We have either that $\Delta_n(\mathbf{A}^i) = 0 < \zeta \cdot \Delta_n(\mathcal{X})$, or

$$ki \geq \text{level}(\mathbf{A}^i) \geq \log n^2 \geq k\log n^{2/(k+2)} \geq k\log 1/\zeta,$$

implying that $\Delta_n(\mathbf{A}^i) \leq 2^{-i} \cdot \Delta_n(\mathcal{X}) \leq \zeta \cdot \Delta_n(\mathcal{X})$.

Now, let $j \in 1, \ldots, i$ be the first j such that $\Delta_n(\mathbf{A}^j) \leq \zeta \cdot \Delta_n(\mathcal{X})$. We consider the following two cases:

- Either $\text{level}(\mathbf{A}^j) \leq \log \zeta^{-k}$, and we get $|\mathbf{A}^j| \leq \zeta^{-k}$.

- Or level $\left(\mathbf{A}^j\right) > \log\zeta^{-k}$ in which case the following must hold:
  - $\Delta_n\left(\mathbf{A}^{j-1}\right) \leq 2\zeta\cdot\Delta_n\left(\mathcal{X}\right)$, since $kj \geq \text{level}\left(\mathbf{A}^j\right) \geq k\log 1/\zeta$, implying that $j-1 \geq \log(1/2\zeta)$.
  - $\text{level}\left(\mathbf{A}^{j-1}\right) < \log\zeta^{-k}$, for otherwise $j-1 \geq \log 1/\zeta$ implying that $\Delta_n\left(\mathbf{A}^{j-1}\right) \leq \zeta\Delta_n\left(\mathcal{X}\right)$. It follows that $\left|\mathbf{A}^{j-1}\right| \leq \zeta^{-k}$

Thus, either $\mathbf{A}^j$ or $\mathbf{A}^{j-1}$ satisfies the claim. $\blacksquare$

Lemma 13 *There exists a constant C independent of d and $\mu(\mathcal{X})$, 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 procedure adaptiveRPtree *attains a diameter decrease rate of $k \geq d$ on $\mathbf{X}$. Define*

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

and assume $n \geq \max\left\{(\lambda\Delta_{\mathcal{X}}/\Delta_{\mathcal{Y}})^2, \alpha(n)\right\}$. The excess risk of the regressor is then bounded as

$$\|f_n - f\|^2 \leq C\cdot(\lambda\Delta_{\mathcal{X}})^{2k/(2+k)}\left(\Delta_{\mathcal{Y}}^2\cdot\frac{\alpha(n)}{n}\right)^{2/(2+k)}$$
$$+2\Delta_{\mathcal{Y}}^2\sqrt{\frac{\ln\log n^6 + \ln 3/\delta}{2n}}.$$

Proof: Let $\mathbf{A}^{i_0}$ be as in lemma 12, and $\zeta \doteq \left(\frac{\Delta_{\mathcal{Y}}^2\cdot\alpha(n)}{\lambda^2\Delta_{\mathcal{X}}^2\cdot n}\right)^{1/(2+k)}$. By applying lemma 11 and then lemma 12, we have with probability at least $1 - \delta/3$ that

$$\|f_{n,\mathbf{A}^{i_0}} - f\|^2 \leq C_1\left(\Delta_{\mathcal{Y}}^2\left|\mathbf{A}^{i_0}\right|\frac{\alpha(n)}{n}\right.$$
$$\left. +\lambda^2\left(\Delta_n^2\left(\mathbf{A}^{i_0}\right) + n^{-4/(2+d)}\Delta_{\mathcal{X}}^2\right)\right)$$
$$\leq C_1\left(\Delta_{\mathcal{Y}}^2\cdot\zeta^{-k}\frac{\alpha(n)}{n} + 5\lambda^2\zeta^2\Delta_{\mathcal{X}}^2\right)$$
$$\leq C_2\lambda^2\Delta_{\mathcal{X}}^2\zeta^2.$$

To analyze the cross validation phase, we first fix the partition tree and consider the obtained partitions from $\mathbf{A}^0$ to the final partition $\mathbf{A}^i$ when the stopping criteria holds. We have with probability at least $1 - \delta/3$ over the choice of $(\mathbf{X}', \mathbf{Y}')$ that $\forall j \in \{0, \ldots, i\}$

$$\left|R\left(f_{n,\mathbf{A}^j}\right) - R_n'\left(f_{n,\mathbf{A}^j}\right)\right| \leq \Delta_{\mathcal{Y}}^2\sqrt{\frac{\ln\log n^6 + \ln 3/\delta}{2n}}.$$

The above is obtained by applying McDiarmid's to the empirical risk followed by a union bound over at most $6\log n$ regressors $f_{n,\mathbf{A}^j}, j \in \{0, \ldots, i\}$.

Let $f_n \doteq f_{n,\mathbf{A}^*}$ be the empirical risk minimizer, we can then conclude that

$$\|f_n - f\|^2 \leq C_2\lambda^2\Delta_{\mathcal{X}}^2\zeta^2 + 2\Delta_{\mathcal{Y}}^2\sqrt{\frac{\ln\log n^6 + \ln 3/\delta}{2n}}$$

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

4.2 Risk bound for automatic stopping option

Lemma 14 (Properties of $\mathbf{A}^*$) *Suppose the automatic stopping option is used, and that* adaptiveRPtree *attains a diameter decrease rate of k on $\mathbf{X}$. Define*

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

and $\zeta \doteq \left(\frac{\alpha(n)}{n}\right)^{1/(2+k)}$. Finally, assume $n \geq \alpha(n)$. Then, the following holds for the final partition $\mathbf{A}^*$ retained for regression:

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

Proof: For $i \geq 0$, let $\mathbf{A}^i$ as defined in adaptiveRPtree. We have by definition that $\Delta_n\left(\mathbf{A}^i\right) \leq 2^{-i}\Delta_n\left(\mathcal{X}\right)$, while it follows from the assumption on diameter decrease rate that $\text{level}\left(\mathbf{A}^i\right) \leq ki$. Now for some $i \geq 1$, let $\mathbf{A}^i$ be the final partition of $\mathcal{X}$ achieved by adaptiveRPtree when the stopping criteria holds. We consider the following two cases:

- Either $\text{level}\left(\mathbf{A}^i\right) \leq \log\zeta^{-k}$, and we have by the stopping condition that:

$$\Delta_n^2\left(\mathbf{A}^i\right) \leq \frac{\alpha(n)}{n}2^{\text{level}\left(\mathbf{A}^i\right)}\cdot\Delta_n^2\left(\mathcal{X}\right)$$
$$\leq \frac{\alpha(n)}{n}\zeta^{-k}\Delta_n^2\left(\mathcal{X}\right) = \zeta^2\Delta_n^2\left(\mathcal{X}\right).$$

- Or $\text{level}\left(\mathbf{A}^i\right) > \log\zeta^{-k}$, in which case the following must hold:
  - $\Delta_n\left(\mathbf{A}^{i-1}\right) \leq 2\zeta\cdot\Delta_n\left(\mathcal{X}\right)$, since $ki \geq \text{level}\left(\mathbf{A}^i\right) \geq k\log(1/\zeta)$, implying that $i-1 \geq \log(1/2\zeta)$.
  - $\text{level}\left(\mathbf{A}^{i-1}\right) < \log\zeta^{-k}$, for otherwise we would have stopped at $i-1$. To see this, assume instead that $\text{level}\left(\mathbf{A}^{i-1}\right) \geq \log\zeta^{-k}$: we have that $(i-1) \geq \log\frac{1}{\zeta}$ and subsequently that

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

In other words,

$$\text{level}\left(\mathbf{A}^{i-1}\right) \geq \log\left(n\Delta_n^2\left(\mathbf{A}^{i-1}\right)/\alpha(n)\Delta_n^2\left(\mathcal{X}\right)\right).$$

In either case at least one of $\mathbf{A}^i$ and $\mathbf{A}^{i-1}$ has size at most ζ^{-k} and diameter at most $2\zeta\cdot\Delta_{\mathcal{X}}$. It follows that

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

which concludes the argument. $\blacksquare$

Lemma 15 *There exists a constant C independent of d and $\mu(\mathcal{X})$, such that the following holds with probability at least $1 - \delta/3$ over $(\mathbf{X}, \mathbf{Y})$ and the randomness in the algorithm. Suppose the automatic stopping option is used; assume* adaptiveRPtree *attains a diameter decrease rate of $k \geq d$ on $\mathbf{X}$. Define $\alpha(n) \doteq \left(\log^2 n\right) \log \log(n/\delta) + \log(1/\delta)$. The excess risk of the regressor is then bounded as*

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

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 11 then lemma 14, we have with probability at least $1 - \delta/3$ that

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

which concludes the argument. ∎

5 Core RPtree and diameter decrease rates

5.1 Core RPTree procedures

Procedure basicRPtree($A_0 \subset \mathcal{X}$, Δ, *level l*)

$\mathbf{A}_0 \leftarrow \{A_0\}$;
for $i \leftarrow 1$ **to** ∞ **do**
    if $\Delta_n\left(\mathbf{A}_{i-1}\right) \leq \Delta$ **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* $A \in \mathbf{A}_{i-1}$ **do**
        if $(l + i)$ *is odd* **then**
            // Noisy splits.
            $t \leftarrow \text{median}\{z^\top v : z \in \mathbf{X} \cap A_0\} + \tau$;
        else
            // Median splits.
            $t \leftarrow \text{median}\{z^\top v : z \in \mathbf{X} \cap A\}$;
        end
        $A_{\text{left}} \leftarrow \{x \in A, x^\top v \leq t\}$;
        $A_{\text{right}} \leftarrow A \setminus A_{\text{left}}$;
    end
    $\mathbf{A}_i \leftarrow$ partition of A_0 defined by the leaves of the current tree;
end

Procedure coreRPtree($A_0 \subset \mathcal{X}$, Δ, δ, *level l*)

Call basicRPtree(A_0, Δ, l) $\log\left(6n^2/\delta\right)$ times and return the shortest tree.

RPtree consists of hierarchically bisecting the data space with random hyperplanes. In basicRPtree we alternate between two types of bisections: we split exactly at the median in order to balance the tree, while we split at the median + noise to improve the rate at which the data diameters are reduced down the tree. Notice that for the "noisy" split we use the same hyperplane to bisect all nodes $A \in \mathbf{A}_{i-1}$.

The procedure coreRPtree serves to boost the probability that we get a small tree. The many calls to basicRPtree can be done in parallel so that we don't keep growing the trees that are to be discarded once the smallest tree is identified.

Remark 4 *Given the implementation of* coreRPtree*, the tree returned by* adaptiveRPtree *has the following properties:*

- *Any node at level $6 \log n$ has at most 1 data point: the data is split at the exact median at every other level so that the number of points per nodes decreases exponentially from the root down. If n were a power of 2, we'd need at most $2 \log n$ levels to get to 1 point per node. For general n, notice that the number of points in a node at level $i \geq 2$ is at most $\frac{3}{4}$ of that of its ancestor at level $i - 2$. In other words we need at most $2 \log n / \log(4/3) \leq 6 \log n$ levels to get down to 1 point per node.*

- *As a consequence, the entire tree reaches depth at most $6 \log n$ under either stopping criteria, and therefore has at most $2n^6$ nodes.*

- *Another consequence is that at most $2n^6 \log(6n^2/\delta)$ 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 attainable by the algorithm over supports of low intrinsic dimension.

The following theorem, adapted from Dasgupta and Freund [DF08], is the core of the argument.

Theorem 16 *Let $A \subset \mathbb{R}^D$ and suppose $A \cap \mathbf{X}$ has Assouad dimension d. There exists a constant C' independent of the sample $\mathbf{X}$ and d, with the following property. We have with probability at least $\frac{1}{2}$ that the tree rooted at A returned by the call* basicRPtree($A, \Delta_n\left(A\right)/2, l$) *has 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 $B' \cap \mathbf{X}$.
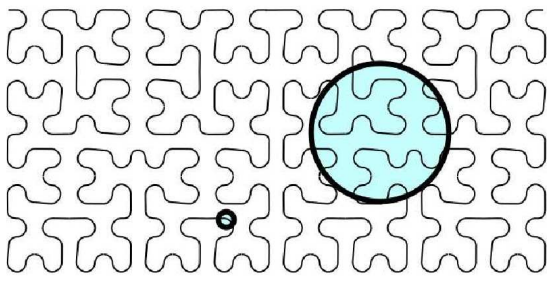
Figure 4: Hilbert space filling curve, balls of smaller radius have lower Assouad dimension.

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 17 *Suppose $\mathcal{X}$ has Assouad dimension d. Let C' be as in theorem 16. Fix $\mathbf{X}$. With probability at least $1 - \delta/3$ over the randomness in the algorithm, the procedure* adaptiveRPtree *attains a diameter decrease rate of $k \leq C' d \log d$ on $\mathbf{X}$.*

Proof: Consider a subtree rooted at A returned by the call coreRPtree$(A, \Delta_n (A)/2, \delta, l)$ in the second loop of procedure adaptiveRPtree. Since $\mathcal{X}$ has Assouad dimension d, $A \cap \mathbf{X}$ also has Assouad dimension d by definition so theorem 16 holds.

Procedure coreRPtree calls basicRPtree as many as $\log \left(6n^2/\delta \right)$ times and returns the smallest tree; thus the probability that the subtree rooted at A has depth over $C' d \log d$ is at most $\delta/6n^2$. Now, under both stopping conditions, coreRPtree is only called on nodes at level at most $\log n^2$; a union bound over all such nodes (at most $2n^2$) yield a probability of failure at most $\delta/3$. ∎

6 Final Remarks

We have shown in this paper that an RPtree regressor will perform well in a scenario where the data space $\mathcal{X}$ has low Assouad dimension $d \ll D$.

Our results are easily extended to other settings. We can for example consider a scenario where the data has low Assouad dimension d at small resolution but "fills" up space at higher resolution. One may think for instance of a Hilbert space filling curve where balls of small enough radius have low Assouad dimension relative to the entire space. (see figure 4). RPtree in this case would initially decrease diameter at a slow rate till it arrives at small enough neighborhoods, at which time the diameter decrease rates speed up. Even in this case, the complexity of the data in larger regions of space has little effect on the final excess risk, provided n is large enough for the tree to arrive at well populated regions with sufficiently small diameter.

References

[BKL06] A Beygelzimer, S. Kakade, and J. Langford. Cover trees for nearest neighbors. *ICML*, 2006.

[BL06] P. Bickel and B. Li. Local polynomial regression on unknown manifolds. *Tech. Re. Dep. of Stats. UC Berkley*, 2006.

[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.

[DF08] S. Dasgupta and Y. Freund. Random projection trees and low dimensional manifolds. *STOC*, 2008.

[GLZ08] A. B. Goldberg, M. Li, and X. Zhu. Online manifold regularization: a new learning setting and empirical study. *ECML PKDD*, 2008.

[GN05] S. Gey and E. Nedelec. Model selection for cart regression trees. *IEEE Transactions on Information Theory*, 51, 2005.

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

[LGL96] L.Devroye, L. Gyorfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[LW07] J. Lafferty and L. Wasserman. Statistical analysis of semi-supervised regression. *NIPS*, 2007.

[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, 2006.

[Sto80] C. J. Stone. Optimal rates of convergence for nonparametric estimators. *Ann. Statist.*, 8:1348–1360, 1980.

[Sto82] C. J. Stone. Optimal global rates of convergence for non-parametric estimators. *Ann. Statist.*, 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.