



HAL
open science

A theory of shape regularity for local regression maps

Jérémy Bettinger, François Portier, Adrien Saumard

► **To cite this version:**

Jérémy Bettinger, François Portier, Adrien Saumard. A theory of shape regularity for local regression maps. 2025. hal-04913380

HAL Id: hal-04913380

<https://hal.science/hal-04913380v1>

Preprint submitted on 29 Jan 2025

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

A theory of shape regularity for local regression maps

Jérémy Bettinger, François Portier, Adrien Saumard

jeremy.bettinger@ensai.fr ; francois.portier@ensai.fr ; adrien.saumard@ensai.fr

Department of Statistics,

University of Rennes, ENSAI, CNRS, CREST-UMR 9194, F-35000 Rennes, France

January 29, 2025

We introduce the concept of shape-regular regression maps as a framework to derive optimal rates of convergence for various non-parametric local regression estimators. Using Vapnik-Chervonenkis theory, we establish upper and lower bounds on the pointwise and the sup-norm estimation error, even when the localization procedure depends on the full data sample, and under mild conditions on the regression model. Our results demonstrate that the shape regularity of regression maps is not only sufficient but also necessary to achieve an optimal rate of convergence for Lipschitz regression functions. To illustrate the theory, we establish new concentration bounds for many popular local regression methods such as nearest neighbors algorithm, CART-like regression trees and several purely random trees including Mondrian trees.

1 Introduction

Consider the standard regression problem where the goal is to estimate the regression function of a random variable $Y \in \mathbb{R}$ given the covariates vector $X \in \mathbb{R}^d$, defined as $g(x) := \mathbb{E}[Y|X = x]$, $x \in \mathbb{R}^d$. One leading approach, called *local regression* or *local averaging*, consists in averaging the observed response variables, restricted to covariates that lie in a small region of the domain \mathbb{R}^d . Local regression methods include kernel smoothing regression [Nad64], nearest neighbors algorithm [FH51, Cov68] and regression trees or more generally partitioning regression estimator [BFSO84, Nob96]. We refer to the books [DGL96, GKKW06] for a complete view on local regression methods and to [BD15] for a precise account on the theory of nearest neighbors algorithm.

Concerning the estimation problem, when the error is measured in terms of the mean squared error (L_2 -error), optimal rates of convergence are known [Sto82] and they depend on the smoothness of the regression function g . Achieving or not these convergence rates often serves as a theoretical baseline to evaluate the accuracy of local regression methods. For instance, a Lipschitz function g can only be approximated at the rate $n^{-1/(d+2)}$ in general, when n independent observations are given. Many of the above estimators are known to achieve optimal convergence rates. The nearest neighbors, the Nadaraya Watson and the fixed partitioning (histogram) regression estimators are all optimal for Lipschitz functions (as well as for twice differentiable functions for the first two listed methods), as explained in [BD15], [Tsy08] and Chapter 4 of [GKKW06], respectively. Furthermore, the Nadaraya-Watson [EM00, GG02] and the nearest

neighbors [Jia19, Por21] estimators are both known to achieve a rate of convergence in sup-norm that is of the same order as the L_2 -rate (up to a logarithmic term).

Other local regression estimators that are based on purely random trees are of interest [BS16], despite the independence of the leaves with respect to the original data, because of their ability to explain certain patterns in the success or failure of different tree constructions and to illustrate the success of random forest regression. Among purely random trees, Mondrian trees, as introduced in [LRT14], indeed achieve the optimal rate of convergence [MGS17] for Lipschitz functions while, in contrast, centered trees fail to reach it [Bia12, Klu21]. We also highlight that partitioning the space with Voronoi cells allows to attain the optimal rate of convergence [GW21] (see also [HKS21] and the reference therein for more results).

Despite the many existing results available for the Nadaraya-Watson and the nearest neighbors regression estimators, and also fixed or purely random partitioning regression rules, only few is known about local regression based on data-dependent partitions such as the well-known CART regression tree [BFSO84]. Such an algorithm is indeed much harder to precisely analyze mathematically. First results on data dependent partitions can be found in [Sto77], but they are restricted to cases where the partition depends only on the covariates, as in nearest neighbors regression or for statistically equivalent blocks [And66]. More advanced results, that are valid for general data dependent partitioning estimators, are obtained in [GO80, BFSO84, Nob96], where conditions are given to ensure almost sure L_2 -consistency. The typical assumptions that are required in the previous works include (i) large enough points in each partition element and (ii) small diameter, while having (iii) a reduced complexity on the partition elements. Note also that Theorem 1 in [SBV15] can be applied to CART regression algorithm and gives sufficient conditions for the L_2 -consistency.

Beyond consistency, few is known about the convergence rates of data-dependent, CART-like regression tree estimators. Recent studies [CVFL22, MW24] have obtained convergence rates for the L_2 -error under the so-called *sufficient impurity decrease* (SID) condition, that is directly linked to the behavior of the precise splitting rule of CART in the regression context. The rate of convergence obtained depends on a parameter - denoted λ in [MW24] - quantifying the strength of the SID condition, and it is not *a priori* easy to discuss the rate optimality. Nonetheless, it is shown in [MW24] that for a univariate linear regression function, the rate obtained through the SID condition is actually optimal. A specific class of additive regression functions achieving a particular smoothness assumption called the "locally reverse Poincaré inequality" is provided in [MW24], satisfying the SID condition. In another direction, the recent negative results in [CKT22] show that CART regression can be sub-optimal, and even inconsistent, for the pointwise - and also uniform - estimation error. Such phenomenon does not occur when focusing on the L_2 -error, but as highlighted in [CKT22], pointwise convergence of decision trees is also essential for reliability of the methodologies developed in some causal inference and multi-step semi-parametric settings for instance.

In this work, we develop a theory for obtaining pointwise and uniform rates of convergence for a large class of local regression estimators, that includes previously mentioned partitioning estimators. More precisely, in a random design regression with heteroscedastic sub-Gaussian noise framework, the theory allows the localization method to be general as it may depend on a different source of randomness (as for the purely random tree) or on the covariates sample (as for nearest neighbors) and even on the full regression sample (as in CART). Instead of studying the integrated L_2 -error, our approach deals with the pointwise and uniform estimation errors recently put forward in the literature [CKT22], for which we obtain a sharp probability upper bound (Theorem 4). To prove such a result, we proceed with a decomposition of the pointwise estimation error into the sum of a variance term (scaling as the inverse of the square root of the

number of covariates in the partition elements) and a bias term (scaling as the diameter of the partition elements). We point out that the major advantage of focusing on the pointwise error, compared to the L_2 -error, is that it allows the control of the variance and bias terms through the use of the Vapnik dimension of a class containing the *elements* of the random partition, instead of having to control the combinatorial size of the class of the *entire* partitions themselves as in [LN96].

Next, we introduce the notion of *shape regularity* by imposing a simple relationship between the Lebesgue volume and the diameter of the localizing set. The major interest of this simple geometric property is that it turns out to be necessary (Proposition 10) and sufficient (Theorem 12) for obtaining optimal rates in pointwise and uniform estimation error. We then discuss how several tree constructions, including purely random trees such as uniform and Mondrian trees, satisfy - or not - the shape regularity condition, allowing to obtain - or not - optimal rates of convergence. In addition, the shape regularity allows to recover and slightly extend some results pertaining to the nearest neighbors literature [Jia19, Por21].

Finally, we obtain a deviation inequality on the uniform estimation error of CART-like regression trees, grown by ensuring a minimal number of covariates in the tree leaves and by following a simple rule which maintains the shape regularity of the resulting localizing sets. In the case of partitions made of hyper-rectangles, such as for CART-like algorithms, the shape-regularity condition reduces to a control of the largest side length of the localizing set by its smallest side length. Recent results obtained in [CKT22] indeed tend to indicate that such rules addition is unavoidable to ensure good pointwise convergence rates for CART regression.

It is worth noting that our approach substantially differs from the use of the SID condition described earlier. The latter indeed ensures convergence rates for the L_2 -error and is highly linked to the precise cost in the splitting rule of CART, defined through the so-called impurity gain. Moreover, the SID condition is expressed through the behavior of the unknown regression function and covariates distribution, and cannot hold for any regression function. In contrast, our shape-regularity condition does not depend on the regression function g , neither on the covariates distribution, and only imposes restriction that may be effective with any cost function involved in the splitting rule. This makes our shape regularity condition easy to guarantee in practice as illustrated in Algorithm 1 (see Section 5.2), where a general cost function is used to build the tree.

The outline is as follows. We state in Section 2 some necessary background and formulate the setting of local regression map estimators. Section 3 then gives a first deviation inequality for local regression map estimators. We introduce in Section 4 the shape regularity conditions and their properties. Section 5 is dedicated to pointwise and uniform convergence bounds for data-dependent regression maps, namely nearest neighbors and CART-like trees. Finally, Section 6 includes new positive and negative results about some classical purely random trees. All the mathematical proofs are given in the Appendix.

2 Mathematical background

2.1 Regression set-up

Let (X, Y) be a random vector with probability distribution \mathbb{P} on $\mathbb{R}^d \times \mathbb{R}$, where $d \geq 1$ is the dimension of covariates vector X . Consider the standard regression framework where the random variable $X \in S_X \subset \mathbb{R}^d$ is called the covariates and $Y \in \mathbb{R}$ is the output variable. We aim at estimating the conditional expectation $x \mapsto g(x) = \mathbb{E}[Y|X = x]$, $x \in S_X$. The quality of the estimation of the function g by an estimator \hat{g} will be assessed with the help of the uniform

norm defined as $\sup_{x \in S_X} |\hat{g}(x) - g(x)|$. For a fixed $x \in S_X$, we also address the estimation error of the value $g(x)$ through the analysis of the deviations of the quantity $|\hat{g}(x) - g(x)|$.

The following assumption on \mathbb{P} will be key in this work and, roughly speaking, amounts to assume that the noise $\epsilon = Y - g(X)$ in the regression model is light tailed.

- (E) The random variable ϵ is sub-Gaussian conditionally on X with parameter σ^2 . That is, $\mathbb{E}[\epsilon|X] = 0$ and for all $\lambda \in \mathbb{R}$,

$$\mathbb{E}[\exp(\lambda\epsilon)|X] \leq \exp\left(\frac{\lambda^2}{2\sigma^2}\right).$$

Note that under assumption (E), the noise term ϵ is squared integrable and it is allowed to depend on the covariates X . In particular, the noise is *heteroscedastic*, with a uniform upper bound on its conditional variance: $\mathbb{E}[\epsilon^2|X] \leq \sigma^2$ *a.s.*. A more restrictive assumption is when ϵ is independent from X and sub-Gaussian with parameter σ^2 .

Associated to the regression model, we consider a sample $\mathcal{D}_n = \{(X_i, Y_i) : i = 1, \dots, n\}$ of independent and identically distributed pairs of random variables with distribution $\mathbb{P}^{\otimes n}$. Let also $\varepsilon_i = Y_i - g(X_i)$ for each $i = 1, \dots, n$.

2.2 Local regression maps

We consider general local regression estimators using the concept of local maps so as to include regression trees and partitioning estimators but also the nearest neighbors regression rule. Let $\mathcal{B}(S_X)$ denote the Borel σ -algebra on S_X .

Definition 1. A local map for a variable X is a mapping $\mathcal{V} : S_X \rightarrow \mathcal{B}(S_X)$ such that for all $x \in S_X$, $x \in \mathcal{V}(x)$.

We emphasize here that this work deals with continuous covariates and therefore all local maps will have their images to sets with positive Lebesgue measure. Also we stress out that similar maps were introduced in [Nob96], where they are however restricted to partition based estimator. For any local map \mathcal{V} , the associated regression estimator is given by

$$\forall x \in S_X, \quad \hat{g}_{\mathcal{V}}(x) = \frac{\sum_{i=1}^n Y_i \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sum_{i=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_i)},$$

with the convention $0/0 = 0$. Local maps \mathcal{V} depending on the sample $(X_1, Y_1), \dots, (X_n, Y_n)$ are of particular interest. This is indeed the case for certain adaptive tree constructions as well as for nearest neighbors regression.

The local regression map framework is particularly interesting because it includes a variety of different methods, e.g., fixed partitioning, purely random trees, nearest neighbors, and CART-like constructions, and each method induces a particular dependence structure when creating the partition.

Example 1 (fixed hyper-rectangles partition). The most simple case for the dependence structure of the local map is when the partition is fixed, not random. Suppose $S_X = (0, 1]^d$. For each coordinate $k = 1, \dots, d$, consider the collection $0 = u_0^{(k)} < u_1^{(k)} < \dots < u_{N_k}^{(k)} = 1$. This allows to introduce a partition of S_X made of $\prod_{k=1}^d N_k$ elements defined as $V_{i_1, \dots, i_d} = \prod_{k=1}^d (u_{i_k}^{(k)}, u_{i_k+1}^{(k)})$ for each d-uplet (i_1, \dots, i_d) satisfying $i_\ell \in \{0, \dots, N_\ell - 1\}$ for $\ell \in \{1, \dots, d\}$. Note that each V_{i_1, \dots, i_d} has a positive Lebesgue measure $\prod_{k=1}^d (u_{i_k+1}^{(k)} - u_{i_k}^{(k)})$.

Example 2 (purely random trees). In contrast with Example 1, *purely random tree* construction, as described in [AG14] and initially introduced in [Bre00], follows from using some randomness that is independent from the observed sample. It includes centered (resp. uniform) trees, for which the split direction is uniformly distributed along the space coordinates and the split location of the selected side is at the center (resp. uniformly distributed). It also includes Mondrian trees [LRT14] where the split direction is selected at random depending on the shape - i.e. side lengths - of the leaf.

Example 3 (nearest neighbors regression). Nearest neighbors algorithm induces a Voronoi-like partition which dependence structure is different from the one of purely random trees, since the resulting partition depends here on the data through the location of the covariates in the space. The k -nearest neighbors (k -NN) estimator (see [BD15] for a recent textbook) is defined, for each $x \in S_X$, as the average responses among the k -nearest neighbors to point x . As such, we have

$$\hat{g}_{NN}(x) = \frac{1}{k} \sum_{i=1}^n Y_i \mathbb{1}_{B(x, \hat{\tau}_k(x))}(X_i),$$

where $\hat{\tau}_k(x)$ is the so-called k -NN radius defined as the smallest radius $\tau > 0$ such that $\sum_{i=1}^n \mathbb{1}_{B(x, \tau)}(X_i) \geq k$. Note that here the local map is $\mathcal{V}(x) = B(x, \hat{\tau}_k(x))$ and therefore depends on X_1, \dots, X_n .

Example 4 (CART-like trees). Regression trees are a class of partition based estimators where the partition is recursively built, and made of hyper-rectangles. Therefore, they are part of the local map framework, just as example 1 and 2 above. Usual regression trees are grown sequentially by splitting stage-wise each (adult) leaf into two (children) leaves. In most cases, as in CART regression [BFSO84], each cell division results from splitting along one single variable according to a data-based criterion. This precise step is crucial as it allows to adapt the partition to the prediction problem. For instance, if one variable is not significant then it must be better not to split with respect to it. This enables to obtain a flexible regression estimator which behaves well in many problems even when the dimension d is rather large. The fact that the resulting partition depends on the full data (including the response) is however problematic for the theory since in this case, the local averaging estimator is not a sum over independent random variables, thus prohibiting a direct application of concentration inequalities for sums of independent observations. Finally, it is worth mentioning that CART regression trees are the ones that are usually combined in the standard random forest regression algorithm as introduced in [Bre01].

3 A deviation bound for local map estimators

The section is divided into two parts. We first give some preliminary concentration bounds that are free from any restriction on the probability distribution of the covariates. Then we use them in the regression framework in order to have some concentration bounds on the estimation error.

3.1 A deviation bound for the variance term

The *shattering coefficient*, as introduced in Vapnik's seminal work [VC15] and detailed for instance in [VDVW96], is key to obtain upper bounds on certain empirical sums indexed by sets or functions. Let \mathcal{A} be a collection of subsets of a set S . Given an arbitrary collection

$z = (z_1, \dots, z_n)$ of distinct points in S , consider the collection of \mathbb{R}^n -points $\mathbb{1}_{\mathcal{A}}(z)$ defined as $\{(\mathbb{1}_{\mathcal{A}}(z_1), \dots, \mathbb{1}_{\mathcal{A}}(z_n)) : A \in \mathcal{A}\} \subset \{0, 1\}^n$. We have that $|\mathbb{1}_{\mathcal{A}}(z)| \leq 2^n$ and when $|\mathbb{1}_{\mathcal{A}}(z)| = 2^n$ we say that z is shattered by \mathcal{F} . An important quantity is then

$$\mathbb{S}_{\mathcal{A}}(n) := \sup_{z \in \mathbb{R}^n} |\mathbb{1}_{\mathcal{A}}(z)|$$

which is called the shattering coefficient.

We now provide a VC-type inequality tailored to local regression maps. This will be key to the analysis of the variance term of local map estimators.

Theorem 2. *Let $n \geq 1$ and $\delta \in (0, 1)$. Suppose that (E) is fulfilled and that $\{\mathcal{V}(x) : x \in \mathbb{R}^d\} \subset \mathcal{A}$, a deterministic collection of sets in \mathbb{R}^d . The following inequality holds with probability at least $1 - \delta$,*

$$\sup_{x \in \mathbb{R}^d} \frac{\sum_{i=1}^n \varepsilon_i \mathbb{1}_{\mathcal{V}(x)}(X_i)}{\sqrt{\sum_{j=1}^n \mathbb{1}_{\mathcal{V}(x)}(X_j)}} \leq \sqrt{2\sigma^2 \log \left(\frac{\mathbb{S}_{\mathcal{A}}(n)}{\delta} \right)}.$$

Note that in Theorem 2 above, only an upper bound is given but a lower bound is also valid, since the same holds true when each ε_i are replaced by $-\varepsilon_i$. Moreover, combining such inequalities through a union bound gives a result for the supremum of the absolute value.

3.2 A pointwise error bound

We now state a general deviation bound on the uniform error of local regression map estimators with finite Vapnik-Chervonenkis (VC) dimension. The VC dimension is defined as

$$vc(\mathcal{A}) = \max\{n \geq 1 : \mathbb{S}_{\mathcal{A}}(n) = 2^n\}.$$

As a consequence, the fact that all given z_1, \dots, z_{v+1} points cannot be shattered is equivalent to the fact that the VC dimension is smaller than v . The reason why the VC dimension is appropriate to control the complexity of classes of sets is perhaps explained by the Sauer's lemma (see [Lug02] for a proof) which states that $\mathbb{S}_{\mathcal{A}}(n) \leq \sum_{i=0}^{vc(\mathcal{A})} \binom{n}{i}$. An interesting consequence of Sauer's lemma is that $\mathbb{S}_{\mathcal{A}}(n) \leq (n+1)^{vc(\mathcal{A})}$.

As established in [WD81], previous examples include the class of cells $(-\infty, t] \subset \mathbb{R}^d$, having VC dimension equal to d , or the class $(s, t]$, $s, t \in \mathbb{R}^d$, of VC dimension equal to $2d$. In addition, the class of balls in \mathbb{R}^d has dimension equal to $d+1$.

Definition 3. *A local map \mathcal{V} is said to be VC whenever $\{\mathcal{V}(x) : x \in S_X\} \subset \mathcal{A}$, a fixed VC collection of sets in \mathbb{R}^d .*

Let us further define some quantities that will be instrumental in our analysis. For any set V , its diameter is given by the formula

$$\text{diam}(V) = \sup_{(x,y) \in V \times V} \|x - y\|_2,$$

where $\|x\|_2^2 = \sum_{k=1}^d x_k^2$. A real function g on S_X is called L -Lipschitz as soon as $|g(x) - g(y)| \leq L\|x - y\|_2$ for all $(x, y) \in S_X^2$. Define also the local Lipschitz constant $L(V)$ of g on $V \subset S_X$ as the smallest constant $L > 0$ such that, for all (x, y) in V^2 ,

$$|g(x) - g(y)| \leq L\|x - y\|_2.$$

For a L -Lipschitz function, it holds $L(V) \leq L$ for any set $V \subset S_X$.

The next probability error bound is valid for local map estimators, with a general VC local map, that may for instance depend on the sample.

Theorem 4. Let $n \geq 1$ and $\delta \in (0, 1/2)$. Under (E), suppose that g is Lipschitz on S_X and that the local map is VC with dimension v . We have, with probability at least $1 - 2\delta$, for all $x \in S_X$,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{2\sigma^2 \log\left(\frac{(n+1)^v}{\delta}\right)}{n\mathbb{P}_n(\mathcal{V}(x))}} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

An alternative approach proposed in [LN96, Nob96] as well as in [DGL96], see Theorem 21.2 therein, follows from a uniform control on all resulting partitions, implying consistency results for sums over all partition elements. In Theorem 4, our approach is substantially different, since by considering the pointwise or sup-norm error, the complexity term comes from the elements of the partition only. In addition, Theorem 4 above might be compared with Theorem 6.1 in [DGL96], which is suitable to either non-random or purely random (i.e., independent from the sample) data partitioning ([BDL08]). While Theorem 4 is valid for data dependent partitions, we recover almost sure consistency by imposing two conditions that are similar to those required in Theorem 6.1 of [DGL96], namely $\text{diam}(\mathcal{V}(x)) \rightarrow 0$ and $n\mathbb{P}_n(\mathcal{V}(x))/\log(n) \rightarrow 0$. Depending on whether the previous conditions hold uniformly in x or for a given x , the consistency, uniform or pointwise, of the local map regression estimator can thus be obtained.

4 Shape regularity

We describe here the minimal mass assumption, which deals with the distribution of the covariates \mathbb{P} . We then introduce the concept of shape regularity for local maps.

4.1 Minimal mass assumption

The next minimal mass assumption allows to obtain an estimate for $\mathbb{P}_n(\mathcal{V}(x))$, which appears in the upper bound stated in Theorem 4.

- (X) For a local map \mathcal{V} on S_X , there exist a constant $\kappa > 0$ and a density function f_X such that, almost surely, for all $x \in S_X$,

$$\mathbb{P}(\mathcal{V}(x)) \geq \kappa f_X(x) \lambda(\mathcal{V}(x)),$$

where λ stands for Lebesgue measure on \mathbb{R}^d .

The minimal mass assumption is quite general as it allows to include k -NN regression estimators as well as partitioning estimators on $[0, 1]^d$. In the first case, $\mathcal{V}(x)$ is a (small enough) ball with positive (random) radius. As shown in [Jia19], such assumption is satisfied by bounded from below density on smooth sets S_X , as well as Gaussian or Laplace variables on \mathbb{R}^d ([GKM16]). Some further details will be given in Section 5.1. In the second example $\mathcal{V}(x)$ is an hyper-rectangle included in $S_X = [0, 1]^d$ and therefore lower bounded densities on S_X easily satisfy (X).

The following definition is now required to ensure that enough points are lying within each element of the local map.

Definition 5. A VC local map $x \mapsto \mathcal{V}(x)$ with dimension $v > 0$ is called (δ, n) -large whenever, for all $x \in S_X$, almost surely,

$$n \max(\mathbb{P}_n(\mathcal{V}(x)), \mathbb{P}(\mathcal{V}(x))) \geq 8 \log\left(\frac{4(2n+1)^v}{\delta}\right).$$

Note in particular that the latter inequality is easy to check in practice as it suffices to make sure that enough points are in each element of the local map.

Theorem 6. *Let $n \geq 1$ and $\delta \in (0, 1/3)$. Under (E) and (X), suppose that g is L -Lipschitz on S_X , that the local map is VC with dimension v and is (δ, n) -large, then we have with probability at least $1 - 3\delta$, for all $x \in S_X$,*

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{3\sigma^2 \log\left(\frac{(n+1)^v}{\delta}\right)}{n\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

The previous result is different from the one of Theorem 4 as the bound no longer depends on the number of points in the associated sets but instead on the volume. Together with the diameter, these two quantities will appear in the definition of the γ -shape regularity, so as to minimize the latter upper bound and therefore, to attain optimal rates of convergence for the underlying regression problem.

4.2 Shape-regular sets

A key concept, which will help us to characterize the rates of convergence of the local map regression estimators, is now introduced. As established in Theorem 6, under the minimal mass assumption, the quantity $|\hat{g}_{\mathcal{V}}(x) - g(x)|$ is bounded by $\sqrt{1/(n\lambda(\mathcal{V}(x)))} + \text{diam}(\mathcal{V}(x))$, up to constants and log terms. Theorem 6 allows to understand that a trade-off between the volume and the diameter must be achieved in order to reach optimal rates. In this regard, first note that the volume cannot be larger than the diameter to the power d , as $\lambda(\mathcal{V}(x)) \leq \text{diam}(\mathcal{V}(x))^d$. Then optimizing the previous bound with respect to $\lambda(\mathcal{V}(x))$ and $\text{diam}(\mathcal{V}(x))$ under constraint that $\lambda(\mathcal{V}(x)) \leq \text{diam}(\mathcal{V}(x))^d$ leads to $\lambda(\mathcal{V}(x)) = \text{diam}(\mathcal{V}(x))^d = n^{-d/(d+2)}$ which leads to optimal rate. On the contrary, if $\text{diam}(\mathcal{V}(x))^d = \gamma_n \lambda(\mathcal{V}(x))$ with $\gamma_n \rightarrow \infty$, then the rate of convergence is suboptimal. This reasoning motivates the introduction of the following notion of shape-regularity.

Definition 7. *For $\gamma > 0$, a set V is called γ -shape-regular (γ -SR) if $\text{diam}(V)^d \leq \gamma \lambda(V)$.*

The previous condition can be interpreted as a volume condition: the volume of V should be of the same order as the volume of the smallest ball containing V . Roughly speaking, the shape of V is not that different from that of a ball. Moreover, it does not depend on the covariates density, making it easy to check in practice.

We provide now an alternative to Definition 7, specifically designed for local maps valued in the set of hyper-rectangles. For any hyper-rectangle $A \subset S_X$, let $h_-(A)$ and $h_+(A)$ denote the smallest and largest side length, respectively.

Definition 8. *For $\beta > 0$, a hyper-rectangle A is called β -shape-regular (β -SR) if $h_+(A) \leq \beta h_-(A)$.*

It is easily seen that when a set V is an hyper-rectangle, the γ -SR property is related to β -SR. This is the subject of the following proposition.

Proposition 9. *A β -SR hyper-rectangle is γ -SR with $\gamma = \beta^d d^{d/2}$. Conversely, a γ -SR hyper-rectangle is β -SR with $\beta = \gamma$.*

The two definitions of shape regularity, γ and β , are therefore equivalent in the case of hyper-rectangles. More precisely, the first implication from above will be of particular interest for us, as it will allow to show that some regression trees are γ -shape-regular. In practice, one way to obtain a β -SR (and therefore γ -SR) tree is to allow only for β -SR splits when growing the tree, i.e., valid splits in light of Definition 8. This is easily imposed as it only requires to restrict the optimization domain when finding the optimal split. We further develop this aspect in Section 5.2 below.

Note that, in dimension $d = 1$, trees are necessarily shape-regular for $\gamma = \beta = 1$ as $h_- = h_+$. From this perspective, dimension 1 plays a special role and might exhibit convergence properties that would not generalize to larger dimensions.

Let us now formalize a bit more on the idea that a non-shape-regular set would lead to a suboptimal convergence rate, at least for some regression functions that are sufficiently varying. Consider indeed the function $g : x \mapsto \sum_{k=1}^d x_k$ defined on $[0, 1]^d$. Set $d \geq 1$ and assume that $X \sim \mathcal{U}[0, 1]^d$. Since g is Lipschitz - note that each partial derivative of g is actually *equal* to one pointwise - optimal rates are of order $n^{-1/(d+2)}$. Consider estimating g at 0 using a rectangular cell such that $\text{diam}(\mathcal{V})^d/\lambda(\mathcal{V}) \geq \bar{\gamma}$ where $\bar{\gamma} > 0$. Next we show that, under standard conditions, the optimal rate cannot be achieved when $\bar{\gamma}$ grows with n . This is important as it means that the optimal rate cannot be attained except when $\bar{\gamma}$ is bounded, meaning that trees need to be shape-regular for being optimal.

Proposition 10. *Let $n \geq 1$ and $d \geq 1$. Suppose that $X \sim \mathcal{U}[0, 1]^d$ and that (E) is fulfilled with $g(x) = \sum_{k=1}^d x_k$. Consider a localizing map \mathcal{V} such that $\mathcal{V}(0) = \prod_k [0, h_k]$. Let $\bar{\gamma}$ be such that $\text{diam}(\mathcal{V})^d/\lambda(\mathcal{V}) \geq \bar{\gamma}$. Whenever $2^{d+4} \log(2) \leq n \prod_{k=1}^d h_k$, there exists a constant $C_d > 0$ depending only on d such that*

$$\mathbb{E}[(\hat{g}_{\mathcal{V}}(0) - g(0))^2]^{1/2} \geq C_d \left(\frac{\bar{\gamma} \sigma^2}{n} \right)^{1/(d+2)}.$$

More generally, the latter result still holds $g(x) \geq \sum_{k=1}^d x_k - g(0)$ on $\mathcal{V}(0)$. An example of such function is for instance g differentiable, $\nabla g(0) = (1, \dots, 1)^T$ and g convex. But many non-convex functions satisfy this condition of course. Note also that the previous result can be extended to covariates X having a density uniformly bounded from above and from below. Note finally, that the lower bound provided in Proposition 10 is also valid when for the $L_2(P^{\otimes n})$ -norm of the sup-norm estimation error $\|g_{\mathcal{V}} - g\|_{\infty}$, since the latter quantity is bounded from below by the pointwise estimation error.

4.3 Shape regularity of local maps

Let us now introduce the following definition which requires that all elements of the localizing map are γ -SR.

Definition 11. *A localization map $x \mapsto \mathcal{V}(x)$ is γ -SR if all elements in $\{\mathcal{V}(x) : x \in S_X\}$ are γ -SR.*

We stress that for most trees, the randomness of the construction will require to study the stochastic variability of γ (see Section 6 where uniform, centered and Mondrian tree are considered). To validate the γ -SR condition, we now provide an optimal error rate for such γ -SR local maps when choosing a good value for the volume. In the next statement, we use the notation $f \lesssim g$ when there exists a universal constant $a > 0$ such that $f \leq ag$. We write $f \asymp g$ whenever $f \lesssim g$ and $g \lesssim f$.

Theorem 12. *Under the assumptions of Theorem 6, if the local map is γ -SR and if for all $x \in S_X$, $\lambda(\mathcal{V}(x)) \asymp (\log((n+1)^v/\delta)/n)^{d/(d+2)}$, we have, with probability at least $1 - 3\delta$, for all $x \in S_X$,*

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \lesssim c \left(\frac{\log\left(\frac{(n+1)^v}{\delta}\right)}{n} \right)^{1/(d+2)}$$

where $c = \sqrt{3\sigma^2/(\kappa f_X(x))} + L(\mathcal{V}(x))\gamma^{1/d}$. In addition, whenever S_X is bounded and $f_X(x) \geq b > 0$ for all $x \in S_X$, we have, with probability at least $1 - 3\delta$,

$$\sup_{x \in S_X} |\hat{g}_{\mathcal{V}}(x) - g(x)| \lesssim c \left(\frac{\log\left(\frac{(n+1)^v}{\delta}\right)}{n} \right)^{1/(d+2)}$$

where $c = \sqrt{3\sigma^2/(\kappa b)} + L\gamma^{1/d}$.

5 Data-dependent local regression maps

In this section, we show that shape regularity is useful to analyse local regression maps that are data-dependent. The first example is the nearest neighbors regression estimator and the second one is a modified version of the CART-like algorithm.

5.1 Nearest neighbors regression

Nearest neighbors regression estimators are local maps estimators for which $\mathcal{V}(x) = B(x, \hat{\tau}_{n,k}(x))$ where $\hat{\tau}_{n,k}(x)$ has been defined in Section 2, Example 3. Using the standard index order to break possible ties, we have $\mathbb{P}_n(\mathcal{V}(x)) = k/n$ and, by relying on Theorem 4, we obtain the following result.

Theorem 13. *Let $\delta \in (0, 1/3)$, $n \geq 1$, $d \geq 1$ and $k \geq 8 \log(4(2n+1)^{(d+1)}/\delta)$. Let \mathcal{V} be obtained from nearest neighbors algorithm as detailed in Example 3. Suppose that (X) is valid for all balls with radius smaller than $T_0 > 0$. Suppose that (E) is fulfilled, and that g is L -Lipschitz on S_X . We have with probability at least $1 - 3\delta$, for all $x \in S_X$ such that $2k \leq T_0^d n \kappa f_X(x)$,*

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{2\sigma^2 \log((n+1)^{d+1}/\delta)}{k}} + 2 \left(\frac{2k}{n \kappa f_X(x)} \right)^{1/d} L(\mathcal{V}(x)).$$

Note that the conditions on the value of k are satisfied for n sufficiently large and $k \asymp n^a$, for any $a \in (0, 1)$. The assumption that (X) is valid for all balls with radius smaller than $T_0 > 0$ is easy to check when the density is lower bounded by a constant $b > 0$ on S_X (in which case S_X must be compact) and when S_X satisfies $\int_{S_X \cap B(x, \tau)} \geq \kappa_0 \tau^d$ for all $\tau \leq T_0$, for some $\kappa_0 > 0$ and $T_0 > 0$. This is done in [Jia19, Por21] and extended to unbounded sets S_X in [GKM16].

To our knowledge the above result is new among the nearest neighbors literature, in which uniform deviation inequalities are provided, to our best knowledge, for densities uniformly bounded away from 0. Such results have been investigated recently in [Jia19] and [Por21] for compactly supported covariates. In contrast, the above upper bound is valid for all x in any domain S_X , at the price of accounting for regions with low density value that may deteriorate the accuracy. We have the following corollary in which we consider an optimal choice for k as well as a lower bounded assumption on the density.

Corollary 14. *Assuming that n is sufficiently large, then choosing the integer $k \asymp n^{2/(d+2)} \log((n+1)^{d+1}/\delta)^{d/(d+2)}$ in Theorem 13 yields the following inequality with probability at least $1 - 3\delta$, for all $x \in S_X$,*

$$|\hat{g}_V(x) - g(x)| \lesssim c \left(\frac{\log((n+1)^{d+1}/\delta)}{n} \right)^{1/(d+2)},$$

where $c = \sqrt{2\sigma^2} + 2L(\mathcal{V}(x)) [2/(\kappa f_X(x))]^{1/d}$. In addition, whenever S_X is bounded and $f_X(x) \geq b > 0$ for all $x \in S_X$, we have, when n is sufficiently large, with probability at least $1 - 3\delta$,

$$\sup_{x \in S_X} |\hat{g}_V(x) - g(x)| \lesssim c \left(\frac{\log((n+1)^{d+1}/\delta)}{n} \right)^{1/(d+2)},$$

where $c = \sqrt{2\sigma^2} + 2L [2/(\kappa b)]^{1/d}$.

We note that the convergence rate is the same as in Theorem 12. However, the constant c differs significantly in the two results as when $f_X(x)$ is small, the constant in Theorem 12 is of order $(\kappa f_X(x))^{-1/2}$, whereas in Theorem 13 it scales as $(\kappa f_X(x))^{-1/d}$.

5.2 CART-like regression tree

We now consider general local regression trees for which each split is selected using a general cost function. In particular, the deviation inequality obtained below is valid for local regression maps that may depend on the whole dataset $(X_i, Y_i)_{i=1, \dots, n}$ and not only on the covariates as in nearest neighbor algorithm. We call these trees "CART-like", since the CART algorithm is arguably the most important instance of such data-dependent regression trees, due to its wide fame and use in practice.

Let us introduce a general class of recursive data dependent trees. For a given cell V , a split is characterized by two parameters $(p, u) \in S := \{1, \dots, d\} \times (0, 1)$. The resulting left and right child cells, $V(l)$ and $V(r)$, are such that for any $k \neq p$, $h_k(V(l)) = h_k(V(r)) = h_k(V)$, and for $k = p$, $h_k(V(l)) = h_k(V)u$ and $h_k(V(r)) = h_k(V)(1 - u)$. We also recall that $h_-(V) = \min_{k=1, \dots, d} h_k(V)$ and $h_+(V) = \max_{k=1, \dots, d} h_k(V)$. With this notation, the split condition for V to be β -shape regular can be expressed with the help of a restriction on the set of valid splits. Given V , let us define the set of β -shape regular splits as

$$S_\beta(V) := \{(p, u) \in S : h_+(V(s)) \leq \beta h_-(V(s)), \forall s \in \{l, r\}\}.$$

We note that when $\beta \geq 2$, the $S_\beta(V)$ cannot be empty. Splitting the largest side in the middle is always in $S_\beta(V)$. Another restriction on the splits is needed to ensure a sufficient number of points. It is given by

$$S_m(V) := \{(p, u) \in S : n\mathbb{P}_n(V(s)) \geq m, \forall s \in \{l, r\}\}.$$

We do not need to fully specify the splitting criterion. When $S_m(V) \neq \emptyset$, the split in the cell V is defined as a minimizer on $S_\beta(V) \cap S_m(V)$, of a cost function M_n , given by

$$\begin{aligned} M_n : S \times \mathcal{R}([0, 1]^d) &\rightarrow \mathbb{R} \\ ((p, u), V) &\mapsto M_n((p, u), V), \end{aligned}$$

where $\mathcal{R}([0, 1]^d)$ is the set of hyper-rectangles in S_X . In case $S_m(V) = \emptyset$, no split is performed and the cell V remains unchanged. The main strength of our analysis lies in the generality of

the cost function, which can actually be any function that may depend or not on the sample. For instance, in CART-regression, the cost function depends on the sample and is defined as

$$M_n((p, u), V) = \frac{\sum_{i: X_i \in V(l)} (Y_i - \bar{Y}(V(l)))^2}{n\mathbb{P}_n(V(l))} + \frac{\sum_{i: X_i \in V(r)} (Y_i - \bar{Y}(V(r)))^2}{n\mathbb{P}_n(V(r))}$$

where $\bar{Y}(V) = \sum_{i: X_i \in V} Y_i / (n\mathbb{P}_n(V))$ for any cell V .

Algorithm 1 CART-like regression tree

Input: Sample $(X_i, Y_i)_{i=1, \dots, n} \subset [0, 1]^d \times \mathbb{R}$, minimal number of points $m \in \{1, \dots, n\}$, shape-regularity $\beta \geq 2$, cost function $M_n : S \times \mathcal{R}([0, 1]^d) \rightarrow \mathbb{R}$. Let $V^{(0)} = \{[0, 1]^d\}$ be the initial partition, made of one element (i.e. $|V^{(0)}| = 1$).

for $j = 0, 1, \dots$ **do**

Let $V^{(j+1)} = \emptyset$ denote the partition at step $j + 1$. The update is as follows:

for $k = 1, 2, \dots, |V^{(j)}|$ **do**

(a) Whenever $S_m(V_k^{(j)}) \neq \emptyset$, define two children, $V(l)$ and $V(r)$, according to

$$\arg \min_{(p, u) \in S_\beta(V_k^{(j)}) \cap S_m(V_k^{(j)})} M_n((p, u), V_k^{(j)})$$

If the above optimization problem has no solution, just pick p as the largest side and $u = 1/2$. Set

$$V^{(j+1)} = \{V^{(j+1)}, V(l), V(r)\}$$

(b) Whenever $S_m(V_k^{(j)}) = \emptyset$, child is same as parent. Set

$$V^{(j+1)} = \{V^{(j+1)}, V_k^{(j)}\}$$

end for

STOP if $V^{(j+1)} = V^{(j)}$ (no valid split exists)

end for

Return the final partition elements $V^{(j+1)}$

By splitting on S_β and S_m , Algorithm 1 ensures that two conditions are met when growing the tree. The first growing condition, that is the β -shape regularity of the cell, may not constitute a stopping criterion. Indeed, because $\beta \geq 2$, one can always split at the middle the largest side of the considered cell. The other growing condition on m is easy to check in practice since it amounts to keep a cell as a leaf if and only if the number of data points belonging to that cell is greater than m and strictly smaller than $2m$. As a consequence, one might modify classical algorithms, in the case precisely where the split proposed by the algorithm does not respect the β -shape-regularity condition for a prescribed value of β or the other growing condition asking for sufficiently many points in the cells. The next theorem gives a deviation inequality for the associated regression map.

Theorem 15. *Let $S_X = [0, 1]^d$, $\delta \in (0, 1/3)$, $n \geq 1$, $d \geq 1$, $\beta \geq 2$ and $m \in \{1, \dots, n\}$ such that $m \geq 4 \log(4(2n + 1)^{2d}/\delta)$. Suppose that (X) and (E) are fulfilled and that g is L -Lipschitz. Let \mathcal{V} be the local regression map obtained from a CART-like tree with input parameters β , m and cost function M_n , then we have, with probability $1 - 3\delta$, for all $x \in S_X$,*

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{2\sigma^2 \log((n + 1)^{2d}/\delta)}{m}} + L(\mathcal{V}(x))\beta\sqrt{d} \left(\frac{5m}{nf_X(x)\kappa} \right)^{1/d}.$$

Note that the conditions on the value of m are satisfied whenever n is sufficiently large and $m \asymp n^a$, for any $a \in (0, 1)$. Notice that taking $m = n^{2/(d+2)}$ in the estimation bound of Theorem 15 gives the optimal convergence rate $n^{-1/(d+2)}$, up a multiplicative logarithmic term. Moreover, such a value of m allows the bound to be valid with a probability that grows to one polynomially in n , since the constraint $m \geq 4 \log(4(2n+1)^{2d}/\delta)$ will be then satisfied. In addition, such results and comments remain valid for the rate of convergence in sup-norm whenever the density f is uniformly bounded from below by a positive constant independent of n . This is stated in the subsequent corollary.

Corollary 16. *In Theorem 15, if the integer m is chosen as $m \asymp n^{2/(d+2)} \log((n+1)^{2d}/\delta)^{d/(d+2)}$, then we have the following inequality for n sufficiently large with probability at least $1 - 3\delta$, for all $x \in S_X$,*

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \lesssim c \left(\frac{\log((n+1)^{2d}/\delta)}{n} \right)^{1/(d+2)},$$

where $c = \sqrt{2\sigma^2} + \beta L(\mathcal{V}(x))\sqrt{d} [5/(\kappa f_X(x))]^{1/d}$. In addition, whenever S_X is bounded and $f_X(x) \geq b > 0$ for all $x \in S_X$, we have for n sufficiently large and with probability at least $1 - 3\delta$,

$$\sup_{x \in S_X} |\hat{g}_{\mathcal{V}}(x) - g(x)| \lesssim c \left(\frac{\log((n+1)^{2d}/\delta)}{n} \right)^{1/(d+2)},$$

where $c = \sqrt{2\sigma^2} + \beta L\sqrt{d} [5/(\kappa b)]^{1/d}$.

The previous result shows that CART-like regression trees are able to attain the optimal rate of convergence as soon as a simple constraint - restricting acceptable splits by a simple rule - is imposed during the tree construction.

Interestingly, results presented in [CKT22] tend to indicate that such modifications are in general necessary for the classical CART algorithm to achieve a good pointwise - or uniform - behavior. More precisely, it is shown in [CKT22] that the use of CART is problematic for the estimation of a constant regression function, measured with the sup-norm error. Indeed, its rate of convergence in dimension one is slower than any polynomial of the sample size n , with non-vanishing probability. In addition, the honest version of CART - i.e. when the prediction values among the cells use data that are independent of those used to construct the partition, see Definition 5.1 in [CKT22] -, is proved to be inconsistent with positive probability as soon as the tree depth is of order at least $\log(\log(n))$. This is due to the fact that the splitting criterion produces leaves that are too small.

Our results complete the picture drawn in [CKT22] by putting forward the fact that producing too small cells is *the only problem* that can occur with the use of CART in dimension one. Indeed, any cell being β -shape-regular in dimension one, with $\beta = 1$, Theorem 15 shows that the only problem must come from the amount of data m in the least populated cell. Indeed, if m is of order $\log(n)$, then our deviation does not converge to zero when δ is fixed and the sample size goes to infinity. This is basically what happens in [CKT22]. In such case, we are indeed not able to prove the consistency of CART. On the contrary, when m is of order $n^{2/(d+2)}$, Corollary 16 shows that our modified version of CART is rate optimal in sup-norm.

6 Purely random trees

We consider now purely random trees (PRT), that are built by successively refining a partition of the space, in a way that is independent from the initial sample \mathcal{D}_n . Before considering

uniform, centered and Mondrian trees, we start by studying a key property of Lebesgue volume invariance which will be satisfied for the trees of interest. In this section we assume, for clarity, that $S_X = [0, 1]^d$ and we always take $x \in S_X$.

6.1 Lebesgue volume invariance

To set up notations, let us describe a PRT locally around a point x using the local maps framework introduced before. The tree is generated iteratively, and at each step i , for the cell $\mathcal{V}(x)$ containing x , a coordinate is selected according to a random variable $D_i \in \{1, \dots, d\}$ and then the side of the cell in direction D_i , that we write (a, b) , $a < b$, is split into two intervals $(a, a + (b - a)S_i)$ and $(a + (b - a)S_i, b)$, thus defining two new cells C_1 and C_2 . Consequently, each step i consists in splitting a cell and depends on a pair of random variables (D_i, S_i) , that is independent from the dataset \mathcal{D}_n . After N steps, we denote $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$.

In the following proposition, we state the remarkable fact that the Lebesgue volume of the cell can be expressed independently from the successive coordinate choices. We denote \bar{S}_i the length reduction of the side D_i of the considered cell at step i , that is either equal to S_i or $1 - S_i$ according to the fact that the coordinate x_{D_i} is smaller or greater than $a + (b - a)S_i$ respectively.

Proposition 17 (Lebesgue Volume Invariance). *With the notations above, the following formula holds*

$$\lambda(\mathcal{V}(x, (D_i, S_i)_{i=1}^N)) = \prod_{i=1}^N \bar{S}_i.$$

Assume in addition that for any i , the distribution of S_i is symmetric around $1/2$, that is, $S_i \sim 1 - S_i$. Then we get the following equality in distribution,

$$\lambda(\mathcal{V}(x, (D_i, S_i)_{i=1}^N)) \sim \prod_{i=1}^N S_i.$$

Note that for centered or uniform random trees, the distributions of the S_i 's are indeed symmetric around $1/2$. Furthermore, for centered trees, $S_i = 1/2$ almost surely, the Lebesgue volume of the cell containing x after N steps is equal to $1/2^N$.

It is worth also noticing that actually, the formulas of Proposition 17 are valid even if the random variables D_i and S_i depend on the dataset. Thus, the Lebesgue volume of the cell containing x is independent from the direction choices as soon as the random vectors $(D_i)_{i=1}^N$ and $(S_i)_{i=1}^N$ are independent from each other, but not necessarily from the dataset.

6.2 Uniform random trees

Let us first provide some deviation bounds for the diameter and volume of the localizing map built with uniform random trees.

Proposition 18. *Consider that $S_i = U_i$ are independent and uniformly distributed over $(0, 1)$ and that D_i are independent from each other and from the U_i 's and uniformly distributed over $\{1, \dots, d\}$. Then, for $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ and for any $\beta \geq 0$,*

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \geq \sqrt{d}e^{-N/d+N\beta}) \leq de^{-Nd\beta^2/4}.$$

Moreover, for all $\beta \in (0, 2/d)$ we have,

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \leq \sqrt{d}e^{-N/d-N\beta}) \leq de^{-Nd\beta^2/8}.$$

Proposition 19. Consider that $S_i = U_i$ are independent and uniformly distributed over $(0, 1)$ and that D_i are independent from each other and from the U_i 's and uniformly distributed over $\{1, \dots, d\}$. Then, for $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ and for any $\alpha > 1$,

$$\mathbb{P}(\lambda(\mathcal{V}(x)) \leq e^{-\alpha N}) \leq (\alpha e^{1-\alpha})^N.$$

In addition, for any $\alpha \in (0, 1)$,

$$\mathbb{P}(\lambda(\mathcal{V}(x)) \geq e^{-\alpha N}) \leq (\alpha e^{1-\alpha})^N.$$

Corollary 20. When the number of splits goes to infinity, it holds that, almost surely, there exists $n_0 \geq 1$ such that for all $N \geq n_0$,

$$\sqrt{d}e^{-N/d-4\sqrt{N \log(N)/d}} \leq \text{diam}(\mathcal{V}(x)) \leq \sqrt{d}e^{-N/d+2\sqrt{2N \log(N)/d}}$$

and

$$e^{-N-2\sqrt{N \log(N)}} \leq \lambda(\mathcal{V}(x)) \leq e^{-N+2\sqrt{N \log(N)}}.$$

Moreover, if we denote the normalized diameter $\text{diam}^\#(\mathcal{V}(x)) := \text{diam}(\mathcal{V}(x))/\sqrt{d}$, we obtain for N large enough,

$$e^{-2\sqrt{N \log(N)}(1+2\sqrt{d})} \leq \frac{\text{diam}^\#(\mathcal{V}(x))^d}{\lambda(\mathcal{V}(x))} \leq e^{2\sqrt{N \log(N)}(1+\sqrt{2d})}.$$

The previous results are valid in any dimension $d \geq 1$, but it is worth noting that in dimension one, the (normalized) diameter of any cell is always equal to its Lebesgue volume, $\text{diam}(\mathcal{V}(x)) = \text{diam}^\#(\mathcal{V}(x)) = \lambda(\mathcal{V}(x))$.

We deduce the following high probability upper bound on the pointwise error of the resulting local map regression estimator.

Corollary 21. Let $n \geq 1$, $d \geq 1$, $x \in S_X$ and $N = d \log(n)/(d+2)$. Suppose that $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ is obtained from a uniform random tree as described in Proposition 18. Under (E) and (X), suppose that g is Lipschitz on S_X . Then there exists $C > 0$, that only depends on the parameters of the problem but not on n , such that with probability 1, there is n_0 such that for all $n \geq n_0$,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq Cn^{-1/(d+2)} \sqrt{\log(n)} e^{2\sqrt{\log(n) \log \log(n)}}.$$

From Corollary 20, we see that the local regression map estimator based on the uniform random partition achieves with probability tending to one an estimation error on a fixed point of the covariates space that is close to - but *a priori* greater than - the optimal one, in the sense that for any $\varepsilon > 0$, the estimation error is negligible compared to $n^{-1/(d+2)+\varepsilon}$ for n sufficiently large.

The following negative result establishes that uniform trees are not shape-regular, thus indicating that the optimal rate of convergence may indeed not be achieved by the local estimator based on a uniform random tree.

Proposition 22. Uniform trees are not β -SR, i.e., for any $N \geq d$ and any hyper-rectangle $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ obtained from a uniform random tree as described in Proposition 18, we have, with probability at least $1/11$,

$$\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} \geq e^{\sqrt{N/d}}.$$

While, in the above, the value $1/11$ can certainly be improved, we stress that our result implies that shape-regularity fails to happen on an event having positive probability (independent from n).

6.3 Centered random trees

In the case of centered random trees, the volume of the cell $\mathcal{V}(x)$ after N steps is simply $\lambda(\mathcal{V}(x)) = (1/2)^N$. The diameter of the localizing map behaves as follows.

Proposition 23. *Let $d \geq 2$ be an integer. Consider that $S_i = 1/2$ almost surely and that D_i are independent from each other and uniformly distributed over $\{1, \dots, d\}$. Then, for $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ and for any $\alpha \in (0, 1/d)$,*

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \geq \sqrt{d}2^{-\alpha N}) \leq d \left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N},$$

where $\beta = (d-1)\alpha/(1-\alpha)$. Moreover, for any $\alpha \in (1/d, 1)$, we have for the same β

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \leq \sqrt{d}2^{-\alpha N}) \leq d \left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N}.$$

We have the following corollary about the shape regularity of centered random trees.

Corollary 24. *When the number of splits goes to infinity, it holds that, almost surely, there exists $n_0 \geq 1$ such that for all $N \geq n_0$,*

$$\sqrt{d}2^{-N/d-2\sqrt{(d-1)N \log(N)/d^2}} \leq \text{diam}(\mathcal{V}(x)) \leq \sqrt{d}2^{-N/d+2\sqrt{(d-1)N \log(N)/d^2}}.$$

Moreover, if we denote the normalized diameter by $\text{diam}^\#(\mathcal{V}(x)) := \text{diam}(\mathcal{V}(x))/\sqrt{d}$, we obtain for N large enough,

$$2^{-2\sqrt{(d-1)N \log(N)}} \leq \frac{\text{diam}^\#(\mathcal{V}(x))^d}{\lambda(\mathcal{V}(x))} \leq 2^{2\sqrt{(d-1)N \log(N)}}.$$

In the same spirit as for uniform random tree, we give an upper bound on the pointwise error of the resulting local map regression estimator.

Corollary 25. *Let $n \geq 1$, $d \geq 1$, $x \in S_X$ and $N = d \log(n)/\{(d+2) \log(2)\}$. Suppose that $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ is obtained from a centered random tree as described in Proposition 23. Under (E) and (X), suppose that g is Lipschitz on S_X , there exists $C > 0$, that only depends on the parameters of the problem but not on n , such that with probability 1, there is N_0 such that for all $n \geq N_0$,*

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq Cn^{-1/(d+2)} e^{2\sqrt{\log(n) \log \log(n)}}.$$

As for uniform random trees, the convergence rate is close to - but greater than - the optimal one, in the sense that for any $\varepsilon > 0$, the estimation error is negligible compared to $n^{-1/(d+2)+\varepsilon}$ for n sufficiently large.

We also include a negative result which establishes that centered trees are not shape-regular.

Proposition 26. *Centered trees are not β -SR, i.e., for any $N \geq d$ and any hyper-rectangle $\mathcal{V}(x) = \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ obtained from a centered random tree as described in Proposition 23, we have, with probability at least $1/14$,*

$$\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} \geq 2\sqrt{N/d}.$$

As for Proposition 22 above, the precise value $1/14$ does not play any crucial role here. The important fact is that Proposition 26 shows that shape-regularity is violated on an event having positive probability (independent from n).

6.4 Mondrian trees

A Mondrian process MP is a process that generates infinite tree partitions of $S_X = [0, 1]^d$. These partitions are built by iteratively splitting the different cells at random times, where both the timing and the position of the splits are determined randomly. Additionally, the probability that a cell is split depends on the length of its sides, and the probability of splitting a particular side is proportional to the length of that side. Once a side is selected, the exact position of the split is chosen uniformly along that side. We can then define the pruned Mondrian process $MP(\lambda)$. This version introduces a pruning mechanism that removes splits occurring after a specific time $\lambda > 0$, which is referred to as the lifetime.

Mondrian trees are studied in details in the paper [MGS19], along with various properties that help demonstrate that Mondrian trees are β -SR in probability.

Proposition 27. *Mondrian trees are β -SR in probability. More precisely, for any $x \in [0, 1]^d$, let $\mathcal{V}(x)$ be the hyper-rectangle containing x obtained from a $MP(\lambda)$ tree. For $\delta \leq 1 - (1 - e^{-1})^d$, we then have, with probability at least $1 - 2\delta$,*

$$\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} \leq \frac{5d \log(\delta/d)}{\log(1 - \delta)}.$$

This implies that there is a constant $K_\delta > 0$ such that $h_+(\mathcal{V}(x)) \leq K_\delta h_-(\mathcal{V}(x))$ and thus $h_+(\mathcal{V}(x))/h_-(\mathcal{V}(x)) = O_{\mathbb{P}}(1)$. As a consequence, we obtain optimal rates in probability - that is, that hold with a fixed positive probability - for Mondrian trees.

7 Proofs

Proof of Theorem 2

Let $\mathbb{P}_{X_1^n}$ denote the conditional probability given X_1, \dots, X_n . Let $\mathcal{V} = \{\mathcal{V}(x) : x \in \mathbb{R}^d\}$ and define

$$\mathcal{G} = \{(\mathbf{1}_A(X_1), \dots, \mathbf{1}_A(X_n)) : A \in \mathcal{A}\}.$$

With this notation we have

$$\sup_{x \in \mathbb{R}^d} \frac{\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sqrt{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)}} \leq \sup_{(g_1, \dots, g_n) \in \mathcal{G}} \frac{\sum_{i=1}^n \varepsilon_i g_i}{\sqrt{\sum_{j=1}^n g_j}}$$

Consequently, for all $t > 0$,

$$\begin{aligned} \mathbb{P}_{X_1^n} \left(\sup_{x \in \mathbb{R}^d} \frac{\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sqrt{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)}} > t \right) &\leq \mathbb{P}_{X_1^n} \left(\bigcup_{(g_1, \dots, g_n) \in \mathcal{G}} \left\{ \frac{\sum_{i=1}^n \varepsilon_i g_i}{\sqrt{\sum_{j=1}^n g_j}} > t \right\} \right) \\ &\leq \sum_{(g_1, \dots, g_n) \in \mathcal{G}} \mathbb{P}_{X_1^n} \left(\frac{\sum_{i=1}^n \varepsilon_i g_i}{\sqrt{\sum_{j=1}^n g_j}} > t \right). \end{aligned}$$

Moreover, since the conditional law of ε_i with respect to X_1, \dots, X_n is sub-Gaussian with parameter σ^2 , then $\varepsilon_i g_i$ is sub-Gaussian under $\mathbb{P}_{X_1^n}$, with parameter $\sigma^2 g_i^2$. Hence, $\sum_{i=1}^n \varepsilon_i g_i / \sqrt{\sum_{j=1}^n g_j}$ is sub-Gaussian with parameter $\sigma^2 \sum_{i=1}^n g_i^2 / \sum_{j=1}^n g_j$ by independence. Moreover, $\sum_{i=1}^n g_i^2 =$

$\sum_{i=1}^n g_i$ because $g_i \in \{0, 1\}$. Hence, $\sum_{i=1}^n \varepsilon_i g_i / \sqrt{\sum_{j=1}^n g_j}$ is sub-Gaussian with parameter σ^2 under $\mathbb{P}_{X_1^n}$. Therefore, we obtain

$$\mathbb{P}_{X_1^n} \left(\sup_{x \in \mathbb{R}^d} \frac{\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sqrt{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)}} > t \right) \leq \sum_{(g_1, \dots, g_n) \in \mathcal{G}} \exp\left(\frac{-t^2}{2\sigma^2}\right) \leq \mathbb{S}_{\mathcal{V}}(n) \exp\left(\frac{-t^2}{2\sigma^2}\right).$$

If we set $\delta = \mathbb{S}_{\mathcal{A}}(n) \exp(-t^2/(2\sigma^2))$, we have $t = \sqrt{2\sigma^2 \log(\mathbb{S}_{\mathcal{A}}(n)/\delta)}$. Finally, with probability $\mathbb{P}_{X_1^n}$ at least equal to $1 - \delta$, we get

$$\sup_{x \in \mathbb{R}^d} \frac{\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sqrt{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)}} \leq \sqrt{2\sigma^2 \log\left(\frac{\mathbb{S}_{\mathcal{A}}(n)}{\delta}\right)}.$$

Since δ is independent of (X_1, \dots, X_n) , we obtain the result by integrating with respect to (X_1, \dots, X_n) . \square

Proof of Theorem 4

Let $x \in S_X$. We write the bias-variance decomposition $\hat{g}_{\mathcal{V}}(x) - g(x) = V + B$, where

$$V := \frac{\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)} \quad \text{and} \quad B := \frac{\sum_{i=1}^n (g(X_i) - g(x)) \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)}.$$

The inequality from Theorem 2 gives, with probability at least $1 - 2\delta$, for all $x \in S_X$,

$$|V| \leq \left(\sqrt{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)} \right)^{-1} \sup_{x \in \mathbb{R}^d} \left| \frac{\sum_{i=1}^n \varepsilon_i \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sqrt{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)}} \right| \leq \frac{1}{\sqrt{n\mathbb{P}_n(\mathcal{V}(x))}} \sqrt{2\sigma^2 \log\left(\frac{\mathbb{S}_{\mathcal{A}}(n)}{\delta}\right)}.$$

Using the inequality $\mathbb{S}_{\mathcal{A}}(n) \leq (n+1)^v$ we recover the first term of the stated bound. Furthermore, using the triangle inequality, we obtain that

$$\begin{aligned} |B| &\leq \frac{\sum_{i=1}^n |g(X_i) - g(x)| \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)} \\ &\leq \frac{\sum_{i=1}^n \sup_{y \in \mathcal{V}(x)} |g(y) - g(x)| \mathbf{1}_{\mathcal{V}(x)}(X_i)}{\sum_{j=1}^n \mathbf{1}_{\mathcal{V}(x)}(X_j)} = \sup_{y \in \mathcal{V}(x)} |g(y) - g(x)|. \end{aligned}$$

Moreover, using the Lipschitz assumption, it follows that

$$|g(y) - g(x)| \leq L(\mathcal{V}(x)) \|x - y\|_2 \leq L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)),$$

which concludes the proof. \square

Proof of Theorem 6

Assume that the maximum of $\mathbb{P}_n(\mathcal{V}(x))$ and $\mathbb{P}(\mathcal{V}(x))$ is $\mathbb{P}(\mathcal{V}(x))$. We have, by assumption, for all $x \in S_X$,

$$\frac{n\mathbb{P}(\mathcal{V}(x))}{2} \geq 4 \log\left(\frac{4(2n+1)^v}{\delta}\right).$$

Using that $1 - 1/\sqrt{2} \geq 2/3$, we deduce that

$$\frac{2}{3} \leq 1 - \sqrt{\frac{4 \log \left(\frac{4(2n+1)^v}{\delta} \right)}{n\mathbb{P}(\mathcal{V}(x))}}.$$

Hence, using Theorem 29, we obtain that with probability $1 - \delta$, for all $x \in S_X$,

$$\mathbb{P}_n(\mathcal{V}(x)) \geq \mathbb{P}(\mathcal{V}(x)) \left(1 - \sqrt{\frac{4 \log(4(2n+1)^v/\delta)}{n\mathbb{P}(\mathcal{V}(x))}} \right) \geq \frac{2}{3} \mathbb{P}(\mathcal{V}(x)) \geq \frac{2}{3} \kappa f_X(x) \lambda(\mathcal{V}(x)).$$

Now, if the maximum of $\mathbb{P}_n(\mathcal{V}(x))$ and $\mathbb{P}(\mathcal{V}(x))$ is $\mathbb{P}_n(\mathcal{V}(x))$, then we have

$$\mathbb{P}_n(\mathcal{V}(x)) \geq \mathbb{P}(\mathcal{V}(x)) \geq \kappa f_X(x) \lambda(\mathcal{V}(x)).$$

Using Theorem 4 and the previous inequality on $\mathbb{P}_n(\mathcal{V}(x))$ yields the result. \square

Proof of Proposition 9

Let A be a hyper-rectangle. We use the shortcut h_- and h_+ for $h_-(A)$ and $h_+(A)$, respectively. The first statement is a consequence of $\text{diam}(A) \leq \sqrt{d}h_+$ and $\lambda(A) \geq h_-^d$, as using β -shape regularity, we obtain

$$\text{diam}(A) \leq \sqrt{d}\beta h_- \leq \sqrt{d}\beta \lambda(A)^{1/d}.$$

The second statement can be obtained as follows. Since $\text{diam}(A) \geq h_+$ and $\lambda(A)^{1/d} \leq h_+^{1-1/d} h_-^{1/d}$ we find

$$\gamma^{1/d} \geq \frac{\text{diam}(A)}{\lambda(A)^{1/d}} \geq \frac{h_+}{h_+^{1-1/d} h_-^{1/d}} = \left(\frac{h_+}{h_-} \right)^{1/d}.$$

\square

Proof of Proposition 10

Let $V_0 = \mathcal{V}(0)$. Define

$$W = \frac{\sum_{i=1}^n (Y_i - g(X_i)) \mathbf{1}_{V_0}(X_i)}{\sum_{i=1}^n \mathbf{1}_{V_0}(X_i)}$$

and

$$B = \frac{\sum_{i=1}^n g(X_i) \mathbf{1}_{V_0}(X_i)}{\sum_{i=1}^n \mathbf{1}_{V_0}(X_i)}.$$

We have, since $g(0) = 0$,

$$\hat{g}_\nu(0) - g(0) = W + B,$$

and by conditional independence

$$\mathbb{E}[(\hat{g}_\nu(0) - g(0))^2] = \mathbb{E}[W^2] + \mathbb{E}[B^2].$$

The lower bound for W can be obtained relying on (E). We have

$$\mathbb{E}[W^2 | X_1, \dots, X_n] = \frac{\sigma^2}{\sum_{i=1}^n \mathbf{1}_{V_0}(X_i)},$$

and then, taking the expectation and using Jensen's inequality, we get

$$\mathbb{E}[W^2] \geq \sigma^2 \mathbb{E} \left[\sum_{i=1}^n \mathbb{1}_{V_0}(X_i) \right]^{-1} = \sigma^2 (n\lambda(V_0))^{-1}.$$

To obtain the lower bound on B , we make use of (X). Let $V_1 = \prod_{k=1}^d [h_k/2, h_k] \subset V_0$. We have, $a_0 = \sum_{i=1}^n \mathbb{1}_{V_0}(X_i) \geq \sum_{i=1}^n \mathbb{1}_{V_1}(X_i) = a_1$. We are looking for a constant $c > 0$ such that $a_1 \geq ca_0$ in order to have

$$\begin{aligned} B &\geq \frac{\sum_{i=1}^n g(X_i) \mathbb{1}_{V_1}(X_i)}{\sum_{i=1}^n \mathbb{1}_{V_0}(X_i)} \geq \frac{1}{2} (h_1 + \dots + h_d) \frac{a_1}{a_0} \\ &\geq \frac{c}{2} \text{diam}(V_0). \end{aligned}$$

This implies that

$$\mathbb{E}[B^2] \geq \mathbb{E}[\mathbb{1}_{a_1 \geq ca_0} B^2] \geq \frac{c^2}{4} \text{diam}_1(V_0)^2 \mathbb{P}(a_1 \geq ca_0).$$

Let us look for such a constant $c > 0$. From Theorem 30, one has that with probability at least $1 - 2\delta = 1/2$,

$$\frac{a_1}{a_0} \geq \frac{\mathbb{P}(V_1) \left(1 - \sqrt{2 \log(4)/(n\mathbb{P}(V_1))}\right)}{\mathbb{P}(V_0) \left(1 + \sqrt{3 \log(4)/(n\mathbb{P}(V_0))}\right)}.$$

Furthermore, note that $\lambda(V_1) = \prod_{k=1}^d h_k/2 = 2^{-d} \prod_{k=1}^d h_k = 2^{-d} \lambda(V_0)$ and $\mathbb{P}(V_k) = \lambda(V_k)$ for each $k \in \{0, 1\}$. Note also that we necessarily have $n \prod_{k=1}^d h_k \geq 2^{d+3} \log(4) \geq 3 \times 2^{d+1} \log(4) \geq 3 \times 4 \log(4)$. This ensures also that the numerator is positive. As a consequence, we find that, with probability at least $1/2$

$$\frac{a_1}{a_0} \geq 2^{-d} \frac{1 - \sqrt{\frac{2^{d+1} \log(4)}{n \prod_{k=1}^d h_k}}}{1 + \sqrt{\frac{3 \log(4)}{n \prod_{k=1}^d h_k}}} \geq 2^{-d} \frac{1 - 1/2}{1 + 1/2} = \frac{2^{-d}}{3} := c.$$

Thus, we have obtained that

$$\begin{aligned} \mathbb{E}[(\hat{g}_V(0) - g(0))^2] &= \mathbb{E}[W^2] + \mathbb{E}[B^2] \\ &\geq \frac{\sigma^2}{n\lambda(V_0)} + \frac{c^2}{4} \text{diam}(V_0)^2 \times \frac{1}{2} \\ &\geq \frac{\sigma^2}{n\lambda(V_0)} + \frac{(c\gamma)^2}{8} \lambda(V_0)^{2/d}. \end{aligned}$$

where $\gamma = \bar{\gamma}^{1/d}$. Let a_1 and a_2 be positive real numbers. By studying the function $\psi : x \mapsto a_1 x^{-d} + a_2 x^2$ on $\mathbb{R}_{>0}$, we notice that ψ has global minimum achieved at $x_m = (a_1 d / (2a_2))^{1/(d+2)}$. This implies that

$$\begin{aligned} \min_{x>0} \psi(x) &\geq x_m^2 a_2 \left(\frac{a_1}{a_2 x_m^{d+2}} + 1 \right) \\ &= \left(\frac{a_1 d}{2a_2} \right)^{2/(d+2)} a_2 \left(\frac{2}{d} + 1 \right) \\ &= \left(\frac{a_1^{d/2} a_2}{2} \right)^{2/(d+2)} \left(\frac{2}{d} + 1 \right) \end{aligned}$$

Now, setting $a_1 = \sigma^2 n^{-1}$, $a_2 = (c\gamma)^2/8$, we find

$$\mathbb{E}[(\hat{g}_{\mathcal{V}}(0) - g(0))^2] \geq \psi(\lambda(\mathcal{V})^{1/d}) \geq \left(\frac{\sigma^2 d (c\gamma)^d}{2(2\sqrt{2})^d n} \right)^{2/(d+2)} \left(1 + \frac{2}{d} \right) = C_d^2 \left(\frac{\bar{\gamma}\sigma^2}{n} \right)^{2/(d+2)}$$

where

$$C_d = \sqrt{1 + \frac{2}{d}} \left(\frac{d}{2} \right)^{1/(d+2)} \left(\frac{1}{2^d \sqrt{72}} \right)^{d/(d+2)}.$$

□

Proof of Theorem 12

By assumption, there is (a_-, a_+) such that $0 < a_- \leq 1 \leq a_+ < +\infty$ and for all $x \in S_X$,

$$\lambda(\mathcal{V}(x))a_- \leq \left(\frac{\log((n+1)^v/\delta)}{n} \right)^{d/(d+2)} \leq a_+ \lambda(\mathcal{V}(x)).$$

According to Theorem 6, the γ -SR assumption, we obtain with probability at least $1 - 3\delta$, for all $x \in S_X$

$$\begin{aligned} |\hat{g}_{\mathcal{V}}(x) - g(x)| &\leq \sqrt{\frac{3\sigma^2 \log\left(\frac{(n+1)^v}{\delta}\right)}{n\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)) \\ &\leq \sqrt{\frac{3\sigma^2 \log\left(\frac{(n+1)^v}{\delta}\right)}{n\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x))\gamma^{1/d}\lambda(\mathcal{V}(x))^{1/d} \\ &\leq \sqrt{\frac{3\sigma^2 \lambda(\mathcal{V}(x))^{(d+2)/d} a_+^{(d+2)/d}}{\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x))\gamma^{1/d}\lambda(\mathcal{V}(x))^{1/d} \\ &\leq \left(\sqrt{\frac{3\sigma^2 a_+^{(d+2)/d}}{\kappa f_X(x)}} + L(\mathcal{V}(x))\gamma^{1/d} \right) \lambda(\mathcal{V}(x))^{1/d} \\ &\leq \left(\sqrt{\frac{3\sigma^2 a_+^{(d+2)/d}}{\kappa f_X(x)}} + L(\mathcal{V}(x))\gamma^{1/d} \right) \left(\frac{\log\left(\frac{(n+1)^v}{\delta}\right)}{n} \right)^{1/(d+2)} a_-^{-1/d}. \end{aligned}$$

The result follows by taking care that $a_+^{(d+2)/d} \leq a_+^3$ and $a_-^{-1/d} \leq a_-^{-1}$ which means that the universal constant in the upper bound can be taken as $a_+^{3/2}/a_-$.

□

Proof of Theorem 13

For any $x \in S_X$, define $\tau(x)^d = 2k/(n\kappa f_X(x))$ and check that $\tau(x)^d \leq T_0^d$. Using (X) we obtain

$$\forall x \in S_X, \quad n\mathbb{P}(B(x, \tau(x))) \geq n\kappa f_X(x)\tau(x)^d = 2k.$$

Next from Theorem 29, and using that the set of all balls in \mathbb{R}^d , denoted by \mathcal{A} , has Vapnik dimension $d+1$ so that $\mathbb{S}_{\mathcal{A}}(2n) \leq (2n+1)^{(d+1)}$, we deduce that with probability $1 - \delta$,

$$\forall x \in S_X, \quad n\mathbb{P}_n(B(x, \tau(x))) \geq n\mathbb{P}(B(x, \tau(x))) - \sqrt{n\mathbb{P}(B(x, \tau(x)))4 \log(4(2n+1)^{(d+1)}/\delta)}.$$

Note that $x \mapsto x - \sqrt{x\ell}$ is increasing whenever $x \geq \ell/4$. Since, by assumption on k ,

$$\forall x \in S_X, \quad n\mathbb{P}(B(x, \tau(x))) \geq 2k \geq 16 \log(4(2n+1)^{(d+1)}/\delta) \geq \log(4(2n+1)^{(d+1)}/\delta).$$

We obtain that, with probability $1 - \delta$,

$$\forall x \in S_X, \quad n\mathbb{P}_n(B(x, \tau(x))) \geq 2k - \sqrt{8k \log(4(2n+1)^{(d+1)}/\delta)}.$$

Now using again the assumption on k , $k \geq 8 \log(4(2n+1)^{(d+1)}/\delta)$, which implies that with probability $1 - \delta$

$$\forall x \in S_X, \quad n\mathbb{P}_n(B(x, \tau(x))) \geq k.$$

However, for each $x \in S_X$, $\hat{\tau}_{n,k}(x)$ is defined as the smallest such value of τ . Therefore, we obtain that for all $x \in S_X$, $\hat{\tau}_{n,k}(x) \leq \tau(x)$ then, with probability $1 - \delta$,

$$\forall x \in S_X, \quad \hat{\tau}_{n,k}(x)^d \leq \frac{2k}{n\kappa f_X(x)}.$$

The result then follows from applying Theorem 4. The variance term is obtained just noting that $n\mathbb{P}_n(\mathcal{V}(x)) = k$ and $v = d + 1$ because the local map is valued in the collection of balls which VC dimension is given in [WD81]. For the bias we use the Lipschitz condition and the inequality above since the ℓ^2 -diameter is twice the radius $\hat{\tau}_{n,k}(x)$, which gives the upper bound with probability at least $1 - 3\delta$.

Proof of Corollary 14

By assumption, there is (a_-, a_+) such that $0 < a_- \leq 1 \leq a_+ < +\infty$ and

$$k a_- \leq n^{2/(d+2)} \log((n+1)^{d+1}/\delta)^{d/(d+2)} \leq a_+ k.$$

According to Theorem 13, we have the following inequalities with probability at least $1 - 3\delta$, for all $x \in S_X$,

$$\begin{aligned} & |\hat{g}_{\mathcal{V}}(x) - g(x)| \\ & \leq \sqrt{\frac{2\sigma^2 \log((n+1)^{d+1}/\delta) a_+}{n^{2/(d+2)} \log((n+1)^{d+1}/\delta)^{d/(d+2)}}} + 2L(\mathcal{V}(x)) \left(\frac{2n^{2/(d+2)} \log((n+1)^{d+1}/\delta)^{d/(d+2)}}{n\kappa f_X(x) a_-} \right)^{1/d} \\ & \leq c \left(\frac{\log((n+1)^{d+1}/\delta)}{n} \right)^{1/(d+2)} \frac{\sqrt{a_+}}{a_-} \end{aligned}$$

where $c = \sqrt{2\sigma^2} + 2L(\mathcal{V}(x)) [2/(\kappa f_X(x))]^{1/d}$. This choice of k is valid for applying the theorem as long as $8 \log(4(2n+1)^{(d+1)}/\delta) \leq k = n^{2/(d+2)} \log((n+1)^{d+1}/\delta)^{d/(d+2)} \leq T_0^d n\kappa f_X(x)/2$, which is satisfied for n sufficiently large. Moreover, if f is bounded below uniformly on S_X and n is also sufficiently large, we can uniformly bound the upper bound $\sqrt{2\sigma^2} + 2L(\mathcal{V}(x)) [2/(\kappa f_X(x))]^{1/d}$ by $\sqrt{2\sigma^2} + 2L [2/(\kappa b)]^{1/d}$. \square

Proof of Theorem 15

The proof follows from an easy application of the next result which is stated for general local regression maps.

Theorem 28. Let $S_X = [0, 1]^d$, $\delta \in (0, 1/3)$, $n \geq 1$, $d \geq 1$, and $m \geq 4 \log(4(2n + 1)^{2d}/\delta)$. Suppose that (X) and (E) are fulfilled and that g is L -Lipschitz. Let $\beta \geq 2$ and suppose that \mathcal{V} is a local regression map valued in the set of hyper-rectangles contained in S_X , for all $V \in \{\mathcal{V}(x) : x \in \mathbb{R}^d\}$,

$$h_+(V) \leq \beta h_-(V) \quad \text{and} \quad n\mathbb{P}_n(V) \geq m,$$

then we have, with probability $1 - 3\delta$, for all $x \in S_X$,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{2\sigma^2 \log((n+1)^{2d}/\delta)}{m}} + L(\mathcal{V}(x))\beta\sqrt{d} \left(\frac{5m}{nf_X(x)\kappa} \right)^{1/d}.$$

Note that, when growing the tree, the constraint $h_+(V) \leq \beta h_-(V)$ can never be a stopping criterion because one can always select the largest side and split it in the middle. When the tree is fully grown according to the prescribed rules, acceptable splits are no longer possible. Therefore any V satisfies

$$2m \geq n\mathbb{P}_n(V) \geq m.$$

Since the Vapnik dimension of hyper-rectangles is $v = 2d$, using Assumption (X) and Theorem 29, we obtain with probability at least $1 - \delta$,

$$f_X(x)\kappa h_-^d \leq \mathbb{P}(V) \leq \frac{4}{n} \log \left(\frac{4(2n+1)^{2d}}{\delta} \right) + 2\mathbb{P}_n(V) \leq \frac{m}{n} + \frac{4m}{n} = \frac{5m}{n}.$$

In addition,

$$\text{diam}(V) \leq \sqrt{d}h_+ \leq \sqrt{d}\beta h_- \leq \sqrt{d}\beta \left(\frac{5m}{nf_X(x)\kappa} \right)^{1/d}.$$

It remains to apply Theorem 4 and to use that $n\mathbb{P}_n(V) \geq m$ for the variance term to get the stated result. \square

Proof of Corollary 16

It is sufficient to reason as in Corollary 14 by applying Theorem 15 and by bounding f_X from below by $b > 0$. \square

Proof of Proposition 17

Recall that $S_X = [0, 1]^d$, so each side length of the initial cell is equal to one. For any $k \in \{1, \dots, d\}$, we have the following formula for the length h_k of the side k of the cell $\mathcal{V}(x, (D_i, S_i)_{i=1}^N)$,

$$h_k = \prod_{i=1}^n \tilde{S}_i^{B_i^{(k)}},$$

where $B_i^{(k)} = \mathbb{I}_{D_i=k}$. Taking the product over k gives

$$\prod_{k=1}^d h_k = \prod_{k=1}^d \prod_{i=1}^n \tilde{S}_i^{B_i^{(k)}} = \prod_{i=1}^n \tilde{S}_i^{\sum_{k=1}^d B_i^{(k)}}.$$

The result follows by noting that $\lambda(\mathcal{V}(x, (D_i, S_i)_{i=1}^N)) = \prod_{k=1}^d h_k$ and that for any i , $\sum_{k=1}^d B_i^{(k)} = 1$, the latter identity simply corresponding to the fact that exactly one side of the cell is split at each step. \square

Proof of Proposition 18

First notice that, by a union bound and symmetry in the directions, we have

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \geq t) \leq d\mathbb{P}\left(h_1 \geq \frac{t}{\sqrt{d}}\right).$$

Furthermore, by denoting $B_i^{(1)} = \mathbb{1}_{D_i=1}$ as in the proof of Proposition 17, we get for any $r \in (0, 1)$ and $\lambda > 0$,

$$\begin{aligned} \mathbb{P}(h_1 \geq r^N) &= \mathbb{P}\left(\prod_{i=1}^N U_i^{B_i^{(1)}} \geq r^N\right) \\ &\leq \mathbb{E}\left[\left(\frac{\prod_{i=1}^N U_i^{B_i^{(1)}}}{r^N}\right)^\lambda\right] = \left(\frac{\mathbb{E}\left[U_1^{\lambda B_1^{(1)}}\right]}{r^\lambda}\right)^N. \end{aligned}$$

It holds

$$\mathbb{E}\left[U_1^{\lambda B_1^{(1)}}\right] = \frac{1}{d(1+\lambda)} + 1 - \frac{1}{d}.$$

Hence,

$$\mathbb{P}(h_1 \geq r^N) \leq \left(\frac{1}{d(1+\lambda)} + 1 - \frac{1}{d}\right)^N r^{-\lambda N}.$$

First note that it suffices to optimize the bound for $N = 1$. Let us denote

$$Q(\lambda) = \frac{1}{d(1+\lambda)} + 1 - \frac{1}{d}$$

and

$$h(\lambda) = Q(\lambda)r^{-\lambda}.$$

Denote $r = (1/e)^{1/d-\beta}$ for $\beta > 0$, then

$$\begin{aligned} h(\lambda) &= \exp\left(\lambda\left(\frac{1}{d} - \beta\right) + \log\left(1 - \frac{\lambda}{d(1+\lambda)}\right)\right) \\ &\leq \exp\left(\lambda\left(\frac{1}{d} - \beta\right) - \frac{\lambda}{d(1+\lambda)}\right) \\ &\leq \exp\left(\lambda\left(\frac{1}{d} - \beta\right) - \frac{\lambda(1-\lambda)}{d}\right) \\ &= \exp\left(-\lambda\left(\beta - \frac{\lambda}{d}\right)\right). \end{aligned}$$

By taking $\lambda = d\beta/2$, we get

$$\mathbb{P}(h_1 \geq e^{N(\beta-1/d)}) \leq e^{-d\beta^2 N/4}.$$

Then for all $\beta > 0$, we obtain

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \geq \sqrt{d}e^{N(\beta-1/d)}) \leq d\mathbb{P}\left(h_1 \geq e^{N(\beta-1/d)}\right) \leq de^{-d\beta^2 N/4}.$$

We now consider the lower bound. We proceed in the same way as before. By a union bound and symmetry in the directions, we have

$$\mathbb{P}(\text{diam}(\mathcal{V}(x))) \leq t \leq d\mathbb{P}\left(h_1 \leq \frac{t}{\sqrt{d}}\right).$$

Furthermore, for any $(r, \lambda) \in (0, 1)^2$,

$$\begin{aligned} \mathbb{P}(h_1 \leq r^N) &= \mathbb{P}\left(\prod_{i=1}^N U_i^{B_i^{(1)}} \leq r^N\right) = \mathbb{P}\left(\prod_{i=1}^N U_i^{-\lambda B_i^{(1)}} \geq r^{-\lambda N}\right) \\ &\leq \mathbb{E}\left[\left(\frac{\prod_{i=1}^N U_i^{B_i^{(1)}}}{r^N}\right)^{-\lambda}\right] = \left(\frac{\mathbb{E}\left[U_1^{-\lambda B_1^{(1)}}\right]}{r^{-\lambda}}\right)^N. \end{aligned}$$

It holds

$$\mathbb{E}\left[U_1^{-\lambda B_1^{(1)}}\right] = \frac{1}{d(1-\lambda)} + 1 - \frac{1}{d}.$$

Hence,

$$\mathbb{P}(h_1 \leq r^N) \leq \left(\frac{1}{d(1-\lambda)} + 1 - \frac{1}{d}\right)^N r^{\lambda N}.$$

Without loss of generality, we can optimize the bound for $N = 1$. Define

$$Q(\lambda) = \frac{1}{d(1-\lambda)} + 1 - \frac{1}{d} = 1 + \frac{\lambda}{d(1-\lambda)},$$

and

$$h(\lambda) = Q(\lambda)r^\lambda.$$

Set $r = (1/e)^{1/d+\beta}$ for $\beta > 0$, then for all $\lambda \in (0, 1/2)$,

$$\begin{aligned} h(\lambda) &= \exp\left(-\lambda\left(\frac{1}{d} + \beta\right) + \log\left(1 + \frac{\lambda}{d(1-\lambda)}\right)\right) \\ &\leq \exp\left(-\lambda\left(\frac{1}{d} + \beta\right) + \frac{\lambda}{d(1-\lambda)}\right) \\ &\leq \exp\left(-\lambda\left(\frac{1}{d} + \beta\right) + \frac{\lambda(1+2\lambda)}{d}\right) \\ &= \exp\left(-\lambda\left(\beta - \frac{2\lambda}{d}\right)\right), \end{aligned}$$

where in the second inequality we used the fact that $(1-\lambda)^{-1} \leq 1+2\lambda$ for $\lambda \in (0, 1/2)$. By taking $\lambda = d\beta/4 \in (0, 1/2)$, we get

$$\mathbb{P}(h_1 \leq e^{-N(\beta+1/d)}) \leq e^{-d\beta^2 N/8}.$$

Then for all $\beta \in (0, 2/d)$,

$$\mathbb{P}(\text{diam}(\mathcal{V}(x))) \leq \sqrt{d}e^{-N(\beta+1/d)} \leq d\mathbb{P}\left(h_1 \leq e^{-N(\beta+1/d)}\right) \leq de^{-d\beta^2 N/8}.$$

□

Proof of Proposition 19

As in the proof of Proposition 18, we optimize along some polynomial moments controlling the deviation probability of interest. We have $\lambda(\mathcal{V}(x)) = \prod_{i=1}^N U_i$, which gives, for any $\alpha > 1$, $\lambda \in (0, 1)$,

$$\mathbb{P}((\lambda(\mathcal{V}(x)))^{-1} \geq e^{N\alpha}) \leq \mathbb{E} \left[\prod_{i=1}^N U_i^{-\lambda} \right] e^{-N\alpha\lambda} = \left(\frac{e^{-\alpha\lambda}}{1-\lambda} \right)^N.$$

By taking $\lambda = 1 - 1/\alpha$, we get

$$\mathbb{P}((\lambda(\mathcal{V}(x)))^{-1} \geq e^{N\alpha}) \leq (\alpha e^{1-\alpha})^N.$$

Moreover for any $\lambda > 0$, $\alpha \in (0, 1)$,

$$\mathbb{P}(\lambda(\mathcal{V}(x)) \geq e^{-N\alpha}) \leq \mathbb{E} \left[\prod_{i=1}^N U_i^\lambda \right] e^{N\alpha\lambda} = \left(\frac{e^{\alpha\lambda}}{1+\lambda} \right)^N.$$

By taking $\lambda = 1/\alpha - 1$, this gives

$$\mathbb{P}(\lambda(\mathcal{V}(x)) \geq e^{-N\alpha}) \leq (\alpha e^{1-\alpha})^N.$$

□

Proof of Corollary 20

We will use the Borel Cantelli lemma together with the inequalities obtained in theorems 18 and 19. To prove the upper bound on the diameter, we provide values β_N leading to small enough probabilities. More precisely, by taking $\beta_N = 2\sqrt{2 \log(N)/(dN)}$, we get $e^{-Nd\beta_N^2/4} = N^{-2}$. Then

$$\mathbb{P} \left(\text{diam}(\mathcal{V}(x)) \geq \sqrt{d} e^{N(-1/d+\beta_N)} \right) \leq \frac{d}{N^2}.$$

The Borel-Cantelli lemma then gives

$$\mathbb{P} \left(\liminf_{N \rightarrow +\infty} \left\{ \text{diam}(\mathcal{V}(x)) \leq \sqrt{d} e^{N(-1/d+\beta_N)} \right\} \right) = 1.$$

This means that almost surely, beyond a certain rank, we have

$$\text{diam}(\mathcal{V}(x)) \leq \sqrt{d} e^{N(-1/d+\beta_N)}.$$

For the lower bound on the diameter, we proceed in the same way with the choice $\tilde{\beta}_N = 4\sqrt{\log(N)/(dN)}$, or equivalently $e^{-Nd\tilde{\beta}_N^2/8} = N^{-2}$. We deduce that, almost surely, beyond a certain rank,

$$\text{diam}(\mathcal{V}(x)) \geq \sqrt{d} e^{-N(1/d+\tilde{\beta}_N)}.$$

Now regarding the volume, we set $\alpha_N = 1 + 2\sqrt{\log(N)/N}$ and we obtain

$$(\alpha_N e^{1-\alpha_N})^N = \exp \left(-2\sqrt{N \log(N)} + N \log \left(1 + 2\sqrt{\log(N)/N} \right) \right).$$

As $\log \left(1 + 2\sqrt{\log(N)/N} \right) = 2\sqrt{\log(N)/N} - 2 \log(N)/N + O((\log(N)/N)^{3/2})$, we get

$$(\alpha_N e^{1-\alpha_N})^N = \exp \left(-2 \log(N) + O \left(\log(N)^{3/2} / \sqrt{N} \right) \right) \sim 1/N^2.$$

The Borel-Cantelli lemma gives us that almost surely, beyond a certain rank n_0 , we have

$$\forall N \geq n_0, \quad \frac{\lambda(\mathcal{V}(x))}{e^{-N-2\sqrt{N\log(N)}}} \geq 1.$$

For the upper bound on the volume, we set for $N \geq 9$, $\tilde{\alpha}_N = 1 - 2\sqrt{\log(N)/N} \in (0, 1)$ and we obtain

$$(\tilde{\alpha}_N e^{1-\tilde{\alpha}_N})^N = \exp\left(2\sqrt{N\log(N)} + N \log\left(1 - 2\sqrt{\log(N)/N}\right)\right).$$

As $\log\left(1 - 2\sqrt{\log(N)/N}\right) = -2\sqrt{\log(N)/N} - 2\log(N)/N + O((\log(N)/N)^{3/2})$, we get

$$(\tilde{\alpha}_N e^{1-\tilde{\alpha}_N})^N = \exp\left(-2\log(N) + O\left(\log(N)^{3/2}/\sqrt{N}\right)\right) \sim 1/N^2.$$

The Borel-Cantelli lemma gives us that almost surely, beyond a certain rank n_0 , we have

$$\forall N \geq n_0, \quad \frac{\lambda(\mathcal{V}(x))}{e^{-N+2\sqrt{N\log(N)}}} \leq 1.$$

Finally, the last inequality stated in Corollary 20 comes readily by using the two previous inequalities on the diameter and the volume. \square

Proof of Corollary 21

Firstly, since the local regression map is obtained from a tree construction, each element in $\mathcal{V}(x) := \mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ is a rectangle. Hence, in light of [WD81], it holds that the image of the resulting local map is included in the set of rectangles, that has VC dimension $v = 2d$. Hence, the local map is indeed VC.

Secondly, according to Theorem 6, applied pointwise for $x \in S_X$ and with $\delta = (n+1)^{-2}$, we find that whenever $n\mathbb{P}(\mathcal{V}(x))/\log(n) \rightarrow \infty$, it holds that

$$\sum_{n \geq 1} \mathbb{P}(|\hat{g}_{\mathcal{V}}(x) - g(x)| > v_n) < \infty,$$

where

$$v_n = \sqrt{3\sigma^2 \log((n+1)^{v+2}) / (n\kappa f_X(x)\lambda(\mathcal{V}(x)))} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

Applying the Borel Cantelli Lemma, we get that with probability 1, for n large enough,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{n\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

Thirdly, from Corollary 20 and using that $N = d\log(n)/(d+2)$, with probability 1, for a sufficiently large n , we have

$$\lambda(\mathcal{V}(x)) \geq e^{-N-2\sqrt{N\log(N)}} \geq n^{-d/(d+2)} e^{-2\sqrt{\log(n)\log(\log(n))}},$$

where we have used that $N \leq \log(n)$. Using (X), it follows that

$$\mathbb{P}(\mathcal{V}(x)) \geq \kappa f_X(x)\lambda(\mathcal{V}(x)) \geq \kappa f_X(x) n^{-d/(d+2)} e^{-2\sqrt{\log(n)\log(\log(n))}}.$$

As a consequence,

$$n\mathbb{P}(\mathcal{V}(x)) \geq \kappa f_X(x) n^{2/(d+2)} e^{-2\sqrt{\log(n)\log(\log(n))}}.$$

Hence, we get that with probability 1, $n\mathbb{P}(\mathcal{V}(x))/\log(n) \rightarrow \infty$. This ensures the (δ, n) -large hypothesis, in order to apply Theorem 6.

Fourthly, by putting together the second and third point from above, we have the following inequality, with probability 1, for n large enough and $\delta = (n+1)^{-2}$,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{n\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

This gives in virtue of Corollary 20

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{n\kappa f_X(x)e^{-N-2\sqrt{N\log(N)}}}} + L(\mathcal{V}(x))\sqrt{d}e^{-N/d+2\sqrt{2N\log(N)/d}}.$$

Recalling that $N = d\log(n)/(d+2)$, we obtain

$$\begin{aligned} & |\hat{g}_{\mathcal{V}}(x) - g(x)| \\ & \leq n^{-1/(d+2)}e^{\sqrt{N\log(N)}} \sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{\kappa f_X(x)}} + n^{-1/(d+2)}L(\mathcal{V}(x))\sqrt{d}e^{2\sqrt{2N\log(N)/d}}. \end{aligned}$$

Then

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq n^{-1/(d+2)}e^{C_d\sqrt{N\log(N)}} \left(\sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{\kappa f_X(x)}} + L(\mathcal{V}(x))\sqrt{d} \right),$$

where $C_d = \max(1, \sqrt{8/d})$. But since $\log(n+1) \leq 2\log(n)$ for $n \geq 2$, we have

$$\sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{\kappa f_X(x)}} = \sqrt{\frac{3\sigma^2(v+2)\log(n+1)}{\kappa f_X(x)}} \leq \sqrt{\frac{6\sigma^2(v+2)\log(n)}{\kappa f_X(x)}}.$$

Then we set

$$C = \sqrt{\frac{6\sigma^2(v+2)}{\kappa f_X(x)}} + L(\mathcal{V}(x))\sqrt{d} = \sqrt{\frac{6\sigma^2(d+1)}{\kappa f_X(x)}} + L(\mathcal{V}(x))\sqrt{d}.$$

Additionally, since $N \leq \log(n)$, we have

$$N \log(N) \leq \frac{d}{d+2} \log(n) \log(\log(n)),$$

so we set $c_d = \sqrt{\frac{d}{d+2}}C_d = \max\left(\sqrt{\frac{d}{d+2}}, \sqrt{\frac{8}{d+2}}\right) \leq 2$ to obtain the desired inequality. \square

Proof of Proposition 22

At each stage, for each terminal leaf, draw uniformly D_i in $\{1, \dots, d\}$ as well as a uniform random variable U_i . Then we divide the cell according to coordinate $k = D_i$. The corresponding length $h_k(\mathcal{V}(x))$ is then updated into $h_k(\mathcal{V}(x))U_i$ and $h_k(\mathcal{V}(x))(1 - U_i)$. Note that $1 - U_i$ is still uniformly distributed. As a consequence, for a given leaf, after N stages, the k -th length has the following representation

$$h_k(\mathcal{V}(x)) = U_1^{B_1^{(k)}} \times \dots \times U_N^{B_N^{(k)}} = \exp\left(\sum_{i=1}^N B_i^{(k)} \log(U_i)\right)$$

where $B_i^{(k)} = \mathbf{1}_{D_i=k}$. It follows that

$$\begin{aligned} h_+(\mathcal{V}(x)) &= \exp\left(\max_{k=1,\dots,d} \sum_{i=1}^N B_i^{(k)} \log(U_i)\right), \\ h_-(\mathcal{V}(x)) &= \exp\left(\min_{k=1,\dots,d} \sum_{i=1}^N B_i^{(k)} \log(U_i)\right), \end{aligned}$$

and the expression of the ratio is

$$h_+(\mathcal{V}(x))/h_-(\mathcal{V}(x)) = \exp\left(\max_{1 \leq k, j \leq d} \sum_{i=1}^N (B_i^{(k)} - B_i^{(j)}) E_i\right)$$

where $E_i = -\log(U_i)$ follows an exponential distribution with parameter 1.

By denoting $V_i^{k,j} = B_i^{(k)} - B_i^{(j)}$, we get

$$V_i^{k,j} = \begin{cases} 1 & \text{with probability } 1/d \\ 0 & \text{with probability } 1 - 2/d \\ -1 & \text{with probability } 1/d \end{cases}.$$

Note that the variables $(V_i^{k,j})_{i=1}^N$ are mutually independent because the $(D_i)_{i=1}^N$ are independent. Furthermore, since the U_i 's are independent of the V_i 's, the $V_i^{k,j}$'s are independent of the E_i 's. Let $Z_{k,j} = \sum_{i=1}^N V_i^{k,j} E_i$ such that

$$\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} = \exp\left(\max_{1 \leq k, j \leq d} Z_{k,j}\right).$$

Note that $Z_{k,j} = -Z_{j,k}$ and $Z_{k,k} = 0$, which gives

$$\max_{1 \leq k, j \leq d} Z_{k,j} = \max_{1 \leq k < j \leq d} |Z_{k,j}|$$

and thus the formula

$$\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} = \exp\left(\max_{1 \leq k < j \leq d} |Z_{k,j}|\right).$$

By using the Paley-Zygmund inequality to $Z_{k,j}^2$, we get for all $\theta \in (0, 1)$,

$$\mathbb{P}\left(|Z_{k,j}| \geq \sqrt{\theta} \sqrt{\mathbb{E}(Z_{k,j}^2)}\right) \geq (1 - \theta)^2 \frac{\mathbb{E}(Z_{k,j}^2)^2}{\mathbb{E}(Z_{k,j}^4)}.$$

We therefore seek to calculate the 2nd and 4th moments of $Z_{k,j}$. Since $Z_{k,j}^2 = \sum_{i \neq \ell} V_i^{k,j} V_\ell^{k,j} E_i E_\ell + \sum_{i=1}^N V_i^{k,j}{}^2 E_i^2$, $\mathbb{E}(V_i^{k,j}) = 0$ and by independence along the subscripts, we obtain

$$\mathbb{E}(Z_{k,j}^2) = \sum_{i=1}^N \mathbb{E}((V_i^{k,j})^2) \mathbb{E}(E_i^2) = N \times \frac{2}{d} \times 2 = \frac{4N}{d}.$$

Moreover, according to Lemma 31 applied to $M_i := V_i^{k,j} E_i$, we obtain

$$\begin{aligned} \mathbb{E}(Z_{k,j}^4) &= N \mathbb{E}(M^4) + 3N(N-1) \mathbb{E}(M^2)^2 \\ &= N \mathbb{E}((V_i^{k,j})^4) \mathbb{E}(E_i^4) + 3N(N-1) \mathbb{E}((V_i^{k,j})^2)^2 \mathbb{E}(E_i^2)^2. \end{aligned}$$

Indeed, it is easily checked that the variables $(M_i)_{i=1}^N$ are centered and independent, due to the independence between the elements of the collections $(V_i^{k,j})_{i=1}^N$ and $(E_i)_{i=1}^N$ and the fact that the $V_i^{k,j}$ are centered. Basic calculations then give

$$\begin{aligned}\mathbb{E}(Z_{k,j}^4) &= N\mathbb{E}(V_i^{k,j^4})\mathbb{E}(E_i^4) + 3N(N-1)\mathbb{E}(V_i^{k,j^2})^2\mathbb{E}(E_i^2)^2 \\ &= N \times \frac{2}{d} \times 4! + 3N(N-1) \left(\frac{2}{d}\right)^2 \times 2^2 \\ &= \frac{48N}{d^2}(d+N-1).\end{aligned}$$

Consequently, we get

$$\frac{\mathbb{E}(Z_{k,j}^2)^2}{\mathbb{E}(Z_{k,j}^4)} = \frac{16N^2}{d^2} \times \frac{d^2}{48N(d+N-1)} = \frac{N}{3(d+N-1)}$$

and thus, for all $\theta \in (0, 1)$,

$$\mathbb{P}\left(|Z_{k,j}| \geq \sqrt{\theta} \sqrt{\frac{4N}{d}}\right) \geq (1-\theta)^2 \frac{N}{3(d+N-1)}.$$

In particular, for $N \geq d$, we have $3(d+N-1) \leq 6N$, which gives

$$\mathbb{P}\left(|Z_{k,j}| \geq \sqrt{\theta} \sqrt{\frac{4N}{d}}\right) \geq (1-\theta)^2/6.$$

With the choice $\theta = 1/4$, it holds

$$\mathbb{P}\left(|Z_{k,j}| \geq \sqrt{\frac{N}{d}}\right) \geq \frac{9}{16} \times \frac{1}{6} = \frac{3}{32} \geq \frac{1}{11}.$$

Finally, by the following lower bound,

$$\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} = \exp\left(\max_{1 \leq k < j \leq d} |Z_{k,j}|\right) \geq \exp(|Z_{1,2}|),$$

we get, for any $N \geq d$,

$$\mathbb{P}\left(\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} \geq \exp\left(\sqrt{\frac{N}{d}}\right)\right) \geq \mathbb{P}\left(\exp(|Z_{1,2}|) \geq \exp\left(\sqrt{\frac{N}{d}}\right)\right) = \mathbb{P}\left(|Z_{1,2}| \geq \sqrt{\frac{N}{d}}\right) \geq \frac{1}{11}.$$

□

Proof of Proposition 23

As in the proof of Proposition 18, notice that

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \geq t) \leq d\mathbb{P}\left(h_1 \geq \frac{t}{\sqrt{d}}\right).$$

Then, for any $r \in (0, 1)$ and $\lambda > 0$,

$$\begin{aligned}\mathbb{P}(h_1 \geq r^N) &= \mathbb{P}\left(\prod_{i=1}^N 2^{-B_i^{(1)}} \geq r^N\right) \\ &\leq \mathbb{E}\left[\left(\frac{\prod_{i=1}^N 2^{-B_i^{(1)}}}{r^N}\right)^\lambda\right] = \left(\frac{\mathbb{E}\left[2^{-\lambda B_1^{(1)}}\right]}{r^\lambda}\right)^N.\end{aligned}$$

It holds

$$\mathbb{E}\left[2^{-\lambda B_1^{(1)}}\right] = \frac{1}{d2^\lambda} + 1 - \frac{1}{d}.$$

Hence,

$$\mathbb{P}(h_1 \geq r^N) \leq \left(\frac{1}{d2^\lambda} + 1 - \frac{1}{d}\right)^N r^{-\lambda N}.$$

Let us set $r = 2^{-\alpha}$ and define

$$h(\lambda) = Q(\lambda)2^{\lambda\alpha}$$

with

$$Q(\lambda) = \frac{1}{d2^\lambda} + 1 - \frac{1}{d}.$$

By differentiating in λ , we get

$$h'(\lambda) = \log(2)2^{\lambda\alpha} \left(\alpha Q(\lambda) - \frac{1}{d2^\lambda}\right).$$

Hence, $h'(\lambda_0) = 0$ for λ_0 such that $2^{-\lambda_0} = \beta = \alpha(d-1)/(1-\alpha)$ and $\alpha \in (0, 1/d)$. With this choice of λ ,

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \geq \sqrt{d}2^{-\alpha N}) \leq d \left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N}.$$

We proceed in the same way as before for the diameter upper bound. By a union bound and symmetry in the directions, we have

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \leq t) \leq d\mathbb{P}\left(h_1 \leq \frac{t}{\sqrt{d}}\right).$$

Then, for any $r \in (0, 1)$ and $\lambda > 0$,

$$\begin{aligned}\mathbb{P}(h_1 \leq r^N) &= \mathbb{P}\left(\prod_{i=1}^N 2^{B_i^{(1)}} \geq r^{-N}\right) \\ &\leq \mathbb{E}\left[\left(\frac{\prod_{i=1}^N 2^{B_i^{(1)}}}{r^{-N}}\right)^\lambda\right] = \left(\frac{\mathbb{E}\left[2^{\lambda B_1^{(1)}}\right]}{r^{-\lambda}}\right)^N.\end{aligned}$$

It holds

$$\mathbb{E}\left[2^{\lambda B_1^{(1)}}\right] = \frac{2^\lambda}{d} + 1 - \frac{1}{d}.$$

Hence,

$$\mathbb{P}(h_1 \leq r^N) \leq \left(\frac{2^\lambda}{d} + 1 - \frac{1}{d}\right)^N r^{\lambda N}.$$

Let us set $r = 2^{-\alpha}$ and denote

$$h(\lambda) = Q(\lambda)2^{-\lambda\alpha}$$

with

$$Q(\lambda) = \frac{2^\lambda}{d} + 1 - \frac{1}{d}.$$

By differentiating in λ , we get

$$h'(\lambda) = \log(2)2^{-\lambda\alpha} \left(\frac{2^\lambda}{d} - \alpha Q(\lambda) \right).$$

Hence, $h'(\lambda_0) = 0$ for λ_0 such that $2^{\lambda_0} = \beta = \alpha(d-1)/(1-\alpha)$ and $\alpha \in (1/d, 1)$. With this choice of λ ,

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \leq \sqrt{d}2^{-\alpha N}) \leq d \left(1 - \frac{1-\beta}{d} \right)^N \beta^{-\alpha N}.$$

□

Proof of Corollary 24

According to Proposition 23, for any $\alpha \in (1/d, 1)$, we have for $\beta = \alpha(d-1)/(1-\alpha)$ the inequality

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \leq \sqrt{d}2^{-\alpha N}) \leq d \left(1 - \frac{1-\beta}{d} \right)^N \beta^{-\alpha N}.$$

Take now $\alpha = \alpha_N = 1/d + \gamma_N$, with $\gamma_N \rightarrow_{N \rightarrow +\infty} 0$. In this case,

$$\beta = \beta_N = \frac{\alpha_N(d-1)}{1-\alpha_N} = (d-1) \frac{1+d\gamma_N}{d-1-d\gamma_N} = 1 + a_d\gamma_N + b_d\gamma_N^2 + O(\gamma_N^3),$$

where $a_d = d^2/(d-1)$ and $b_d = d^3/(d-1)^2$. This gives

$$\log \left(1 - \frac{1-\beta}{d} \right) = \frac{a_d}{d}\gamma_N + \frac{b_d}{d}\gamma_N^2 - \frac{a_d^2}{2d^2}\gamma_N^2 + O(\gamma_N^3).$$

Futhermore

$$\log(\beta) = -a_d\gamma_N + b_d\gamma_N^2 - \frac{a_d^2}{2}\gamma_N^2 + O(\gamma_N^3)$$

so

$$\alpha \log(\beta) = \frac{a_d}{d}\gamma_N + \frac{b_d}{d}\gamma_N^2 - \frac{a_d^2}{2d}\gamma_N^2 + a_d\gamma_N^2 + O(\gamma_N^3).$$

Then

$$\begin{aligned} \log \left(1 - \frac{1-\beta}{d} \right) - \alpha \log(\beta) &= -\frac{a_d^2}{2d^2}\gamma_N^2 + \frac{a_d^2}{2d}\gamma_N^2 - a_d\gamma_N^2 + O(\gamma_N^3) \\ &= -a_d\gamma_N^2 \left(1 + \frac{a_d}{2d^2} - \frac{a_d}{2d} \right) + O(\gamma_N^3). \end{aligned}$$

Moreover

$$1 + \frac{a_d}{2d^2} - \frac{a_d}{2d} = 1 + \frac{1}{2(d-1)} - \frac{d}{2(d-1)} = 1 - \frac{1}{2} = \frac{1}{2}.$$

Finally

$$\begin{aligned} \left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N} &= \exp\left(N \log\left(1 - \frac{1-\beta}{d}\right) - N\alpha \log(\beta)\right) \\ &= \exp\left(-\frac{a_d}{2}N\gamma_N^2 + O(N\gamma_N^3)\right). \end{aligned}$$

Choosing $\gamma_N = 2\sqrt{\log(N)/(a_d N)} \in (0, 1 - 1/d)$ for N large enough, gives

$$\left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N} = \exp\left(-2\log(N) + O\left(\log(N)^{3/2}/\sqrt{N}\right)\right) \underset{N \rightarrow +\infty}{\sim} N^{-2}$$

and concludes the proof via the Borel-Cantelli lemma. Moreover for the upper bound of the diameter we use also Proposition 23. For any $\alpha \in (0, 1/d)$, we have for $\beta = \alpha(d-1)/(1-\alpha)$,

$$\mathbb{P}(\text{diam}(\mathcal{V}(x)) \geq \sqrt{d}2^{-\alpha N}) \leq d \left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N}.$$

Let us take here $\alpha = \alpha_N = 1/d - \gamma_N$, with $\gamma_N \rightarrow_{N \rightarrow +\infty} 0$. In this case,

$$\beta = \beta_N = \frac{\alpha_N(d-1)}{1-\alpha_N} = (d-1) \frac{1-d\gamma_N}{d-1+d\gamma_N} = 1 - a_d\gamma_N + b_d\gamma_N^2 + O(\gamma_N^3),$$

where $a_d = d^2/(d-1)$ and $b_d = d^3/(d-1)^2$. This gives

$$\log\left(1 - \frac{1-\beta}{d}\right) = -\frac{a_d}{d}\gamma_N + \frac{b_d}{d}\gamma_N^2 - \frac{a_d^2}{2d^2}\gamma_N^2 + O(\gamma_N^3).$$

Futhermore

$$\log(\beta) = -a_d\gamma_N + b_d\gamma_N^2 - \frac{a_d^2}{2}\gamma_N^2 + O(\gamma_N^3)$$

so

$$\alpha \log(\beta) = -\frac{a_d}{d}\gamma_N + \frac{b_d}{d}\gamma_N^2 - \frac{a_d^2}{2d}\gamma_N^2 + a_d\gamma_N^2 + O(\gamma_N^3).$$

Then

$$\begin{aligned} \log\left(1 - \frac{1-\beta}{d}\right) - \alpha \log(\beta) &= -\frac{a_d^2}{2d^2}\gamma_N^2 + \frac{a_d^2}{2d}\gamma_N^2 - a_d\gamma_N^2 + O(\gamma_N^3) \\ &= -a_d\gamma_N^2 \left(1 + \frac{a_d}{2d^2} - \frac{a_d}{2d}\right) + O(\gamma_N^3). \end{aligned}$$

Moreover

$$1 + \frac{a_d}{2d^2} - \frac{a_d}{2d} = 1 + \frac{1}{2(d-1)} - \frac{d}{2(d-1)} = 1 - \frac{1}{2} = \frac{1}{2}.$$

Finally

$$\begin{aligned} \left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N} &= \exp\left(N \log\left(1 - \frac{1-\beta}{d}\right) - N\alpha \log(\beta)\right) \\ &= \exp\left(-\frac{a_d}{2}N\gamma_N^2 + O(N\gamma_N^3)\right). \end{aligned}$$

Choosing $\gamma_N = 2\sqrt{\log(N)/(a_d N)}$ gives

$$\left(1 - \frac{1-\beta}{d}\right)^N \beta^{-\alpha N} = \exp\left(-2\log(N) + O\left(\log(N)^{3/2}/\sqrt{N}\right)\right) \underset{N \rightarrow +\infty}{\sim} N^{-2}$$

and concludes the proof via the Borel-Cantelli lemma.

The last inequality follows directly by invoking the two previous inequalities on diameter and volume. \square

Proof of Corollary 25

The proof follows the same steps as the one of Corollary 21. First, since the local regression map results from a tree construction, each element in $\mathcal{V}(x, (D_i, S_i)_{i=1}^N)$ is a rectangle. Hence, in light of [WD81], it holds that the resulting local map has dimension $v = 2d$.

Second, according to Theorem 6 (applied pointwise for $x \in S_X$), we find that whenever $n\mathbb{P}(\mathcal{V}(x))/\log(n) \rightarrow \infty$, it holds that $\sum_{n \geq 1} \mathbb{P}(|\hat{g}_{\mathcal{V}}(x) - g(x)| > v_n) < \infty$ where

$$v_n = \sqrt{3\sigma^2 \log((n+1)^{v+2}) / (n\kappa f_X(x)\lambda(\mathcal{V}(x)))} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

Note that we have set $\delta = (n+1)^{-2}$. Making use of Borel Cantelli Lemma, it implies that with probability 1, for n large enough,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{n\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

Third, using (X), it follows that

$$\mathbb{P}(\mathcal{V}(x)) \geq \kappa f_X(x)\lambda(\mathcal{V}(x)) = \kappa f_X(x)2^{-N} = n^{-d/(d+2)}\kappa f_X(x)$$

then

$$n\mathbb{P}(\mathcal{V}(x)) \geq n^{2/(d+2)}\kappa f_X(x).$$

Hence, we get that $n\mathbb{P}(\mathcal{V}(x))/\log(n) \rightarrow \infty$.

Fourth, by putting together the second and third point from above, we have the following inequality, with probability 1, for n large enough,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{n\kappa f_X(x)\lambda(\mathcal{V}(x))}} + L(\mathcal{V}(x)) \text{diam}(\mathcal{V}(x)).$$

Now, from Proposition 24, for a sufficiently large, we have

$$\begin{aligned} \text{diam}(\mathcal{V}(x)) &\leq \sqrt{d}2^{-N/d+2}\sqrt{(d-1)N \log(N)/d^2}, \\ \lambda(\mathcal{V}(x)) &= 2^{-N}. \end{aligned}$$

Hence, we get, with probability 1, for n large enough,

$$|\hat{g}_{\mathcal{V}}(x) - g(x)| \leq \sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{n\kappa f_X(x)2^{-N}}} + L(\mathcal{V}(x))\sqrt{d}2^{-N/d+2}\sqrt{(d-1)N \log(N)/d^2}.$$

Because $N = d \log(n)/(\log(2)(d+2))$, we obtain

$$\begin{aligned} |\hat{g}_{\mathcal{V}}(x) - g(x)| &\leq n^{-1/(d+2)}\sqrt{\frac{3\sigma^2 \log((n+1)^{v+2})}{\kappa f_X(x)}} + n^{-1/(d+2)}L(\mathcal{V}(x))\sqrt{de^2\sqrt{(d-1)N \log(N)/d^2}} \\ &\leq n^{-1/(d+2)}\sqrt{\frac{6\sigma^2(v+2) \log(n)}{\kappa f_X(x)}} + n^{-1/(d+2)}L(\mathcal{V}(x))\sqrt{de^2\sqrt{(d-1)N \log(N)/d^2}} \\ &\leq n^{-1/(d+2)}\sqrt{\frac{12\sigma^2(d+1) \log(n)}{\kappa f_X(x)}} + n^{-1/(d+2)}L(\mathcal{V}(x))\sqrt{de^2\sqrt{N \log(N)}} \end{aligned}$$

where we use $v = 2d$ and the inequality $2\sqrt{(d-1)/d^2} \leq 1$ since $(d-2)^2 \geq 0$.

For n large enough, we have $N = d\log(n)/(\log(2)(d+2)) \geq 8$. Thus, this implies that $\log(n) = \frac{d+2}{d}\log(2)N \leq 3\log(2)N = \log(8)N \leq \log(N)N$. Moreover $N \leq 2\log(n)$, and for n large enough $N \leq \log(n)^2$ then $\log(N) \leq 2\log\log(n)$. Finally $\log(n) \leq \log(N)N \leq 4\log(n)\log(\log(n))$. We conclude by using the inequality $\sqrt{x} \leq e^{\sqrt{x}}$ for $x = \log(n)$ and setting $C = \sqrt{12\sigma^2(d+1)\log(n)/(\kappa f_X(x))} + L(\mathcal{V}(x))\sqrt{d}$. \square

Proof of Proposition 26

We follow the proof of Proposition 22, with similar notation, but this time the variable $E_i := -\log(U_i)$ is replaced by $E_i := \log(2)$. By performing the calculations again, we find the moments with lemma 31,

$$\begin{aligned} \mathbb{E}(Z_{k,j}^2) &= 2N\log(2)^2/d. \\ \mathbb{E}(Z_{k,j}^4) &= N\mathbb{E}(V_i^{k,j^4})\mathbb{E}(E_i^4) + 3N(N-1)\mathbb{E}(V_i^{k,j^2})^2\mathbb{E}(E_i^2)^2 \\ &= N \times \frac{2}{d} \times \log(2)^4 + 3N(N-1) \left(\frac{2}{d}\right)^2 \times \log(2)^4 \\ &= \frac{2N}{d^2} \log(2)^4 (6N - 6 + d). \end{aligned}$$

The Paley-Zygmund bound becomes

$$\frac{\mathbb{E}(Z_{k,j}^2)^2}{\mathbb{E}(Z_{k,j}^4)} = \frac{4\log(2)^4 N^2}{d^2} \times \frac{d^2}{2N\log(2)^4(6N-6+d)} = \frac{2N}{6N-6+d}.$$

Thus, for all $\theta \in (0, 1)$ and $N \geq d$,

$$\mathbb{P}\left(|Z_{k,j}| \geq \sqrt{\theta} \sqrt{\frac{2N\log(2)^2}{d}}\right) \geq (1-\theta)^2 \frac{2N}{6N-6+d} \geq \frac{2(1-\theta)^2}{7}.$$

Let us choose $\theta = 1/2$ to obtain

$$\mathbb{P}\left(|Z_{k,j}| \geq \log(2)\sqrt{\frac{N}{d}}\right) \geq \frac{1}{14}$$

and thus for $N \geq d$,

$$\mathbb{P}\left(\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} \geq 2\sqrt{N/d}\right) \geq \frac{1}{14}.$$

Thus, with probability at least $1/14$, the ratio $h_+(\mathcal{V}(x))/h_-(\mathcal{V}(x))$ is bounded below by a quantity that grows exponentially towards infinity. This means that uniform trees are not regular. \square

Proof of Proposition 27

According to the paper by [MGS19] (see proposition 1), we know the distribution of the largest and the smallest side. In fact, we have $h_-(\mathcal{V}(x)) \sim \min(X_1, \dots, X_d)$ and $h_+(\mathcal{V}(x)) \sim \max(X_1, \dots, X_d)$, where the X_i are i.i.d. and follow the Gamma distribution $X \sim \Gamma(2, \lambda)$. We have for all $u \geq 0$

$$\mathbb{P}(X \geq u) = \int_u^{+\infty} \lambda^2 t e^{-\lambda t} dt = e^{-\lambda u} (1 + u\lambda) \geq e^{-\lambda u}.$$

Moreover,

$$\mathbb{P}(h_-(\mathcal{V}(x)) \geq u) = \mathbb{P}(X \geq u)^d \geq e^{-\lambda u d} = 1 - \delta$$

for $u = -\log(1 - \delta)/(\lambda d)$. Then with probability at least $1 - \delta$,

$$h_-(\mathcal{V}(x)) \geq -\frac{\log(1 - \delta)}{\lambda d}.$$

We focus now on $h_+(\mathcal{V}(x))$. We have for all $t \geq 0$,

$$\mathbb{P}(h_+(\mathcal{V}(x)) \leq t) = \mathbb{P}(X \leq t)^d.$$

Let $Y := X - \mathbb{E}(X)$. Since X follows a Gamma distribution, Y is Sub-Gamma. According to [BLM13] (page 29),

$$\forall t > 0, \quad \mathbb{P}(\lambda Y \geq 2\sqrt{t} + t) \leq e^{-t}.$$

Thus,

$$\begin{aligned} \mathbb{P}(\lambda h_+ \leq 2\sqrt{t} + t + \lambda \mathbb{E}(X)) &= \mathbb{P}(\lambda Y \leq 2\sqrt{t} + t)^d = \left(1 - \mathbb{P}(\lambda Y > 2\sqrt{t} + t)\right)^d \\ &\geq (1 - e^{-t})^d = 1 - \delta \end{aligned}$$

with $t = -\log(1 - (1 - \delta)^{1/d})$. Therefore, with probability at least $1 - \delta$,

$$h_+(\mathcal{V}(x)) \leq \frac{2 + 2\sqrt{-\log(1 - (1 - \delta)^{1/d})} - \log(1 - (1 - \delta)^{1/d})}{\lambda}.$$

In particular, for $\delta \leq 1 - (1 - e^{-1})^d$,

$$h_+(\mathcal{V}(x)) \leq \frac{-5 \log(1 - (1 - \delta)^{1/d})}{\lambda} \leq \frac{-5 \log(\delta/d)}{\lambda}$$

where the last inequality comes from the inequality $\delta/d \leq 1 - (1 - \delta)^{1/d}$. Hence, with probability at least $1 - 2\delta$ for $\delta \leq 1 - (1 - e^{-1})^d$

$$\frac{h_+(\mathcal{V}(x))}{h_-(\mathcal{V}(x))} \leq \frac{5d \log(\delta/d)}{\log(1 - \delta)}.$$

□

8 Auxiliary results

Let us state the following Vapnik-type inequality [VC15], which involves some standard-error normalization. The first inequality in the next theorem is Theorem 2.1 in [AST93] (see also Theorem 1.11 in [Lug02]). The second inequality can be obtained from the first one.

Theorem 29 (normalized Vapnik inequality). *Let (Z, Z_1, \dots, Z_n) is a collection of random variables independent and identically distributed with common distribution \mathbb{P} on (S, \mathcal{S}) . For any class $\mathcal{A} \subset \mathcal{S}$, $\delta > 0$ and $n \geq 1$, it holds with probability at least $1 - \delta$, for all $A \in \mathcal{A}$,*

$$\mathbb{P}_n(A) \geq \mathbb{P}(A) \left(1 - \sqrt{\frac{4 \log(4\mathbb{S}_{\mathcal{A}}(2n)/\delta)}{n\mathbb{P}(A)}}\right).$$

In particular, with probability at least $1 - \delta$ we have, for all $A \in \mathcal{A}$,

$$\mathbb{P}(A) \leq \frac{4}{n} \log \left(\frac{4\mathbb{S}_{\mathcal{A}}(2n)}{\delta} \right) + 2\mathbb{P}_n(A).$$

Proof. The first statement is proved in [AST93]. Let us prove the second statement. According to the first point, with probability at least $1 - \delta$, we have for all $A \in \mathcal{A}$

$$n\mathbb{P}_n(A) - n\mathbb{P}(A) \geq -\sqrt{4n\mathbb{P}(A) \log(4\mathbb{S}_{\mathcal{A}}(2n)/\delta)},$$

equivalently,

$$n\mathbb{P}(A) - \sqrt{4n\mathbb{P}(A) \log(4\mathbb{S}_{\mathcal{A}}(2n)/\delta)} - n\mathbb{P}_n(A) \leq 0.$$

Setting $x = \sqrt{n\mathbb{P}(A)}$, $\alpha = \sqrt{4 \log(4\mathbb{S}_{\mathcal{A}}(2n)/\delta)}$ and $\beta = n\mathbb{P}_n(A)$, we have that $x^2 - \alpha x - \beta \leq 0$. Solving the inequality, we find

$$(\alpha - \sqrt{\alpha^2 + 4\beta})/2 \leq x \leq (\alpha + \sqrt{\alpha^2 + 4\beta})/2.$$

By using the fact that x is positive and squaring both sides, it follows that $x^2 \leq (\alpha + \sqrt{\alpha^2 + 4\beta})^2/4$. And by using the inequality $(a + b)^2 \leq 2(a^2 + b^2)$, we obtain $n\mathbb{P}(A) = x^2 \leq \alpha^2 + 2\beta = 4 \log(4\mathbb{S}_{\mathcal{A}}(2n)/\delta) + 2n\mathbb{P}_n(A)$ which is the desired result by dividing each side of the inequality by n . \square

For more details, one can also refer to the book by [BLM13], especially chapters 12 and 13, as well as [DGL96].

The following result is standard and known as the multiplicative Chernoff bound for empirical processes. The following version can be found in [HR90].

Theorem 30. *Let A be a set in \mathbb{R}^d . For any $\delta \in (0, 1)$ and all $n \geq 1$, we have with probability at least $1 - \delta$*

$$\mathbb{P}_n(A) \geq \left(1 - \sqrt{\frac{2 \log(1/\delta)}{n\mathbb{P}(A)}}\right) \mathbb{P}(A).$$

In addition, for any $\delta \in (0, 1)$ and $n \geq 1$, we have with probability at least $1 - \delta$

$$\mathbb{P}_n(A) \leq \left(1 + \sqrt{\frac{3 \log(1/\delta)}{n\mathbb{P}(A)}}\right) \mathbb{P}(A).$$

This lemma is useful for proving Propositions 22 and 26.

Lemma 31. *Let $M, (M_i)_{i=1, \dots, N}$ be a collection of independent and identically distributed random variables such that $\mathbb{E}[M^4] < \infty$. It holds*

$$\mathbb{E} \left[\left(\sum_{i=1}^N M_i \right)^4 \right] = N\mathbb{E}(M^4) + 3N(N-1)\mathbb{E}(M^2)^2.$$

Proof. We have

$$\left(\sum_{i=1}^N M_i \right)^4 = \sum_{i,p,q,r=1}^N M_i M_p M_q M_r.$$

Since the M_i are independent and centered, the expectation of each product $M_i M_p M_q M_r$ will be zero if at least one of the indices is distinct. This restricts the analysis to cases where all indices are identical or two pairs of indices are identical. If all indices are identical, i.e. $i = p = q = r$,

then the expectation of M_i^4 contributes to the sum: $\sum_{i=1}^N \mathbb{E}(M_i^4) = N \mathbb{E}(M^4)$. When two indices are identical and the other two are also identical, i.e. $i = p \neq q = r$, we get a product of the form $M_i^2 M_q^2$. We have 3 choices either i is equal to p , q , or r . The remaining two indices must necessarily be equal. This yields: $3 \sum_{i \neq q} \mathbb{E}(M_i^2) \mathbb{E}(M_q^2) = 3N(N-1) \mathbb{E}(M^2)^2$. Combining the two terms, this proves the result. \square

References

- [AG14] Sylvain Arlot and Robin Genuer. Analysis of purely random forests bias. *arXiv preprint arXiv:1407.3939*, 2014.
- [And66] T. W. Anderson. *Some nonparametric multivariate procedures based on statistically equivalent blocks*. Multivariate Analysis (P. R. Krishnaiah, ed), 5-27, Academic Press, New York., 1966.
- [AST93] Martin Anthony and John Shawe-Taylor. A result of Vapnik with applications. *Discrete Appl. Math.*, 47(3):207–217, 1993.
- [BD15] Gérard Biau and Luc Devroye. *Lectures on the nearest neighbor method*. Springer Series in the Data Sciences. Springer, Cham, 2015.
- [BDL08] Gérard Biau, Luc Devroye, and Gábor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(66):2015–2033, 2008.
- [BFSO84] Leo Breiman, Jerome Friedman, Charles J Stone, and RA Olshen. *Classification and Regression Trees*. CRC Press, 1984.
- [Bia12] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.
- [BLM13] Stéphane Boucheron, Gábor Lugosi, and Pascal Massart. *Concentration inequalities. A nonasymptotic theory of independence*. Oxford University Press, Oxford, 2013.
- [Bre00] Leo Breiman. Some infinity theory for predictor ensembles. Technical report, Cite-seer, 2000.
- [Bre01] Leo Breiman. Random forests. *Machine learning*, 45:5–32, 2001.
- [BS16] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25:197–227, 2016.
- [CKT22] Matias D Cattaneo, Jason M Klusowski, and Peter M Tian. On the pointwise behavior of recursive partitioning and its implications for heterogeneous causal effect estimation. *arXiv preprint arXiv:2211.10805*, 2022.
- [Cov68] T Cover. Estimation by the nearest neighbor rule. *IEEE Trans. Inform. Theory*, 14(1):50–55, 1968.
- [CVFL22] Chien-Ming Chi, Patrick Vossler, Yingying Fan, and Jinchi Lv. Asymptotic properties of high-dimensional random forests. *The Annals of Statistics*, 50(6):3415–3438, 2022.
- [DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*, volume 31. Springer Science & Business Media, 1996.
- [EM00] Uwe Einmahl and David M. Mason. An empirical process approach to the uniform consistency of kernel-type function estimators. *J. Funct. Anal.*, 13(1):1–37, 2000.

- [FH51] Evelyn Fix and Joseph Lawson Hodges. Discriminatory analysis. nonparametric discrimination: Consistency properties. *Int. Stat. Rev.*, 57(3):238–247, 1951.
- [GG02] Evarist Giné and Armelle Guillou. Rates of strong uniform consistency for multivariate kernel density estimators. *Ann. Inst. Henri Poincaré Probab. Stat.*, 38(6):907–921, 2002. En l’honneur de J. Bretagnolle, D. Dacunha-Castelle, I. Ibragimov.
- [GKKW06] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.
- [GKM16] Sébastien Gadat, Thierry Klein, and Clément Marteau. Classification in general finite dimensional spaces with the k -nearest neighbor rule. *Ann. Statist.*, 44(3):982–1009, 2016.
- [GO80] Louis Gordon and Richard A Olshen. Consistent nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 10(4):611–627, 1980.
- [GW21] László Györfi and Roi Weiss. Universal consistency and rates of convergence of multiclass prototype algorithms in metric spaces. *Journal of Machine Learning Research*, 22(151):1–25, 2021.
- [HKS21] Steve Hanneke, Aryeh Kontorovich, Sivan Sabato, and Roi Weiss. Universal bayes consistency in metric spaces. *The Annals of Statistics*, 49(4):2129–2150, 2021.
- [HR90] Torben Hagerup and Christine Rüb. A guided tour of chernoff bounds. *Information processing letters*, 33(6):305–308, 1990.
- [Jia19] Heinrich Jiang. Non-asymptotic uniform rates of consistency for k -NN regression. In *AAAI proceedings*, volume 33, pages 3999–4006, 2019.
- [Klu21] Jason Klusowski. Sharp analysis of a simple model for random forests. In *International Conference on Artificial Intelligence and Statistics*, pages 757–765. PMLR, 2021.
- [LN96] Gabor Lugosi and Andrew Nobel. Consistency of data of data-driven histogram methods for density estimation and classification. *The Annals of Statistics*, 1996.
- [LRT14] Balaji Lakshminarayanan, Daniel M Roy, and Yee Whye Teh. Mondrian forests: Efficient online random forests. *Advances in neural information processing systems*, 27, 2014.
- [Lug02] Gábor Lugosi. Pattern classification and learning theory. In *Principles of nonparametric learning*, pages 1–56. Springer, 2002.
- [MGS17] Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Universal consistency and minimax rates for online mondrian forests. *Advances in Neural Information Processing Systems 30*, 2017.
- [MGS19] Jaouad Mourtada, Stéphane Gaïffas, and Erwan Scornet. Minimax optimal rates for mondrian trees and forests. *Annals of Statistics*, 2019.

- [MW24] Rahul Mazumder and Haoyue Wang. On the convergence of CART under sufficient impurity decrease condition. *Advances in Neural Information Processing Systems*, 36, 2024.
- [Nad64] Elizbar A Nadaraya. On estimating regression. *Theory Probab. Appl.*, 9(1):141–142, 1964.
- [Nob96] Andrew Nobel. Histogram regression estimation using data-dependent partitions. *The Annals of Statistics*, 1996.
- [Por21] François Portier. Nearest neighbor process: weak convergence and non-asymptotic bound. *arXiv preprint arXiv:2110.15083*, 2021.
- [SBV15] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *Annals of Statistics*, 43(4):1716–1741, 2015.
- [Sto77] Charles J Stone. Consistent nonparametric regression. *Ann. Statist.*, 5(1):595–620, 1977.
- [Sto82] Charles J Stone. Optimal global rates of convergence for nonparametric regression. *The annals of statistics*, pages 1040–1053, 1982.
- [Tsy08] Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008.
- [VC15] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, Cham., 2015. Reprint of Theor. Probability Appl. 16 (1971), 264–280.
- [VDVW96] Aad W. Van Der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes. With Applications to Statistics*. Springer Series in Statistics. Springer-Verlag, New York, 1996.
- [WD81] Roberta S Wenocur and Richard M Dudley. Some special Vapnik-Chervonenkis classes. *Discrete Math.*, 33(3):313–318, 1981.