Universal algorithms for learning theory Part I : piecewise constant functions *

Peter Binev, Albert Cohen, Wolfgang Dahmen, Ronald DeVore and Vladimir Temlyakov

October 21, 2004

Abstract

This paper is concerned with the construction and analysis of a universal estimator for the regression problem in supervised learning. Universal means that the estimator does not depend on any a priori assumptions about the regression function to be estimated. The universal estimator studied in this paper consists of a least-square fitting procedure using piecewise constant functions on a partition which depends adaptively on the data. The partition is generated by a splitting procedure which differs from those used in CART algorithms. It is proven that this estimator performs at the optimal convergence rate for a wide class of priors on the regression function. Namely, as will be made precise in the text, if the regression function is in any one of a certain class of approximation spaces (or smoothness spaces of order not exceeding one - a limitation resulting because the estimator uses piecewise constants) measured relative to the marginal measure, then the estimator converges to the regression function (in the least squares sense) with an optimal rate of convergence in terms of the number of samples. The estimator is also numerically feasible and can be implemented on-line.

Key words: Regression, universal estimators, optimal convergence rates in probability, adaptive partitioning, on-line updates

AMS Subject Classification: 62G08, 62G20, 41A25

1 Introduction

This paper addresses the problem of using empirical samples to derive probabilistic or expectation error estimates for the regression function of some unknown probability measure ρ on a product space $Z := X \times Y$. It will be assumed here that X is a bounded

*This research was supported in part by the Office of Naval Research Contracts ONR-N00014-03-1-0051, ONR-N00014-03-1-0675 and ONR N00014-00-1-0470; the Army Research Office Contract DAAD 19-02-1-0028; the AFOSR Contract UF/USAF F49620-03-1-0381; the NSF contracts DMS-0221642 and DMS-0200187 and EEC Human Potential Programme under contract HPRN-CT-2002-00286, "Breaking Complexity".

domain of $\mathbb{R}^{\mathrm{d}}$ and $Y = \mathbb{R}$. Given the data $\mathbf{z} = \{z_1, \ldots, z_m\} \subset Z$ of m independent random observations $z_i = (x_i, y_i)$, $i = 1, \ldots, m$, identically distributed according to ρ, we are interested in estimating the *regression function* $f_\rho(x)$ defined as the conditional expectation of the random variable y at x:

$$f_\rho(x) := \int_Y y d\rho(y|x) \tag{1.1}$$

with $\rho(y|x)$ the conditional probability measure on Y with respect to x. In this paper, it is assumed that this probability measure is supported on an interval $[-M, M]$:

$$|y| \leq M, \tag{1.2}$$

almost surely. It follows in particular that $|f_\rho| \leq M$.

We denote by ρ_X the marginal probability measure on X defined by

$$\rho_X(S) := \rho(S \times Y). \tag{1.3}$$

We shall assume that ρ_X is a Borel measure on X. We have

$$d\rho(x, y) = d\rho(y|x) d\rho_X(x). \tag{1.4}$$

It is easy to check that f_ρ is the minimizer of the risk functional

$$\mathcal{E}(f) := \int_Z (y - f(x))^2 d\rho, \tag{1.5}$$

over $f \in L_2(X, \rho_X)$ where this space consists of all functions from X to Y which are square integrable with respect to ρ_X. In fact one has

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|^2, \tag{1.6}$$

where

$$\|\cdot\| := \|\cdot\|_{L_2(X,\rho_X)}. \tag{1.7}$$

Our objective is therefore to find an *estimator* $f_{\mathbf{z}}$ for f_ρ based on $\mathbf{z}$ such that the quantity $\|f_{\mathbf{z}} - f_\rho\|$ is small.

A common approach to this problem is to choose an hypothesis (or *model*) class $\mathcal{H}$ and then to define $f_{\mathbf{z}}$, in analogy to (1.5), as the minimizer of the empirical risk

$$f_{\mathbf{z}} := \operatorname*{argmin}_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f), \quad \text{with} \quad \mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))2. \tag{1.8}$$

Typically, $\mathcal{H} = \mathcal{H}_m$ depends on a finite number $N = N(m)$ of parameters. In many cases, the number N is chosen using an a priori assumption on f_ρ. In other procedures, the number N is adapted to the data and thereby avoids any a priori assumptions. We shall be interested in estimators of the latter type.

The usual way of evaluating the performance of the estimator $f_{\mathbf{z}}$ is by studying its convergence either in probability or in expectation, i.e. the rate of decay of the quantities

$$\text{Prob}\{\|f_\rho - f_{\mathbf{z}}\| \geq \eta\}, \quad \eta > 0 \quad \text{or} \quad E(\|f_\rho - f_{\mathbf{z}}\|^2) \tag{1.9}$$

as the sample size m increases. Here both the expectation and the probability are taken with respect to the product measure ρ^m defined on Z^m. An estimation of the above probability will automatically give an estimate in expectation by integrating with respect to η. Estimates for the decay of the quantities in (1.9) are usually obtained under certain assumptions (called *priors*) on f_ρ.

It is important to note that the measure ρ_X which appears in the norm (1.7) is unknown and that we want to avoid any assumption on this measure. This type of regression problem is referred to as *random design* or *distribution-free*. A recent survey on distribution free regression theory is provided in the book [20], which includes most existing approaches as well as the analysis of their rate of convergence in the expectation sense.

Priors on f_ρ are typically expressed by a condition of the type $f_\rho \in \Theta$ where Θ is a class of functions that necessarily must be contained in $L_2(X, \rho_X)$. If we wish the error, as measured in (1.9), to tend to zero as the number m of samples tends to infinity then we necessarily need that Θ is a compact subset of $L_2(X, \rho_X)$. There are three common ways to measure the compactness of a set Θ: (i) minimal coverings, (ii) smoothness conditions on the elements of Θ, (iii) the rate of approximation of the elements of Θ by a specific approximation process. In the learning problem, each of these approaches has to deal with the fact that ρ_X is unknown.

To describe approach (i), for a given Banach space $\mathcal{B}$ which contains Θ, we define the entropy number $\varepsilon_n(\Theta, \mathcal{B})$, $n = 1, 2 \ldots$ as the minimal ε such that W can be covered by at most 2^n balls of radius ε in $\mathcal{B}$. The set Θ is compact in $L_2(X, \rho_X)$ if and only if $\varepsilon_n(\Theta, L_2(X, \rho_X))$ tends to zero as $n \to \infty$. One can therefore quantify the level of compactness of Θ by an assumption on the rate of decay of $\varepsilon_n(\Theta, L_2(X, \rho_X))$. A typical prior condition would be to assume that the the entropy numbers satisfy [1]

$$\varepsilon_n(\Theta, \mathcal{B}) \lesssim n^{-r}, \quad n = 1, 2, \cdots \tag{1.10}$$

for some $r > 0$.

Coverings and entropy numbers has a long history in statistics for deriving optimal bounds for the rate of decay in statistical estimation (see e.g. [4]). Several recent works [8, 12, 22] have used this technique to bound the error for the regression problem in learning. It has been communicated to us by Lucien Birgé that one can derive from one of his forthcoming papers [3] that for any class Θ satisfying (1.10) with $\mathcal{B} = L_2(X, \rho_X)$, there is an estimator $f_{\mathbf{z}}$ satisfying

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \lesssim m^{-\frac{2r}{2r+1}}, \quad m = 1, 2, \ldots \tag{1.11}$$

whenever $f_\rho \in \Theta$. Lower bounds which match (1.11) have been given in [12] using a slightly different type of entropy.

[1] Throughout this paper, we use the notation $A \lesssim B$ to mean that there exists a constant C such that $A \leq CB$ independently of the primary variables.

The estimators constructed using this approach are made through ε nets and are more of theoretical interest (in giving the best possible bounds) but are not practical since ρ_X is unknown and therefore these ε nets are also unknown. Another deficiency in this approach is that the estimator typically requires the knowledge of the prior class Θ. One would like to avoid knowledge of Θ in the construction of an estimator since we do not know f_ρ and hence would generally not have any information about Θ. One can also use ε nets to give bounds for $\text{Prob}(\|f_\rho - f_{\mathbf{z}}\|)$. This is one of the main points in [8] and is carried further in [12, 21, 22].

One way to circumvent the problem of not knowing the marginal ρ_X is to is to use coverings in $C(X)$ rather than $L_2(X, \rho_X)$ since a good covering for Θ in $C(X)$ gives bounds for the covering in $L_2(X, \rho_X)$. In this approach one would assume that Θ satisfies (1.10) for $\mathcal{B} = C(X)$ and then build estimators which satisfy (1.11) using ε nets for $C(X)$. Again this does not lead to practical estimators. But the main deficiency of this approach is that the assumption that Θ is a compact subset of $C(X)$ is too severe and does not give a full spectrum of compact subsets of $L_2(X, \rho_X)$.

There is no general approach to defining smoothness spaces with respect to general Borel measures which precludes the direct use of classification according to (ii). One way to circumvent this is to define smoothness in $C(X)$ but then this suffers from the same deficiency of not giving a full array of compact subsets in $L_2(X, \rho_X)$.

The classification of compactness according to approximation properties (iii) begins with a specific method of approximation and then defines the classes Θ in terms of a rate of approximation by the specified method. The simplest example is to take a sequence (X_n) of linear spaces of dimension n and define Θ as the class of all functions f in $L_2(X, \rho_X)$ which satisfy

$$\inf_{g \in X_n} \|f - g\| \le C\alpha_n \tag{1.12}$$

where C is a fixed constant and (α_n) is a sequence of positive real numbers tending to zero. Natural choices for this sequence are $\alpha_n = n^{-r}$, where $r > 0$. Classes defined in such a way will not give a full spectrum of compact subsets in $L_2(X, \rho_X)$. But this deficiency can be removed by using nonlinear spaces Σ_n in place of the linear spaces X_n (see the discusssion in [12]). An illustrative example is approximation by piecewise polynomials on partitions. If the partitions are set in advance this corresponds to the linear space approximation above. In nonlinear methods the partitions are allowed to vary but their size is specfied. We discuss this in more detail later in this paper.

We should mention that in classical settings, for example when ρ_X is Lebesgue measure then the three approaches to measuring compactness are closely related and in a certain sense equivalent. This is the main chapter of approximation theory.

Concrete algorithms have been constructed for the regression problem in learning by using approximation from specific linear spaces such as piecewise polynomial on uniform partitions, convolution kernels, and spline functions. The rate of convergence of the estimators built from such a linear approximation process is related to the approximation rate of the corresponding process on the class Θ. A very useful method for bounding the performance of such estimators was provided in [20] (see Theorem 11.3). For example, if $\mathcal{H}$ is taken a linear space of dimension N and if the least-square estimator (1.8) is post-processed by application of the truncation operator $y \mapsto T_M(y) = \text{sign}(y) \min\{|y|, M\}$,

then

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \lesssim \frac{N\log(m)}{m} + \inf_{g\in\mathcal{H}} \|f_\rho - g\|^2. \tag{1.13}$$

From this, one can derive specific rates of convergence in expectation by balancing both terms. For example, if Θ is a ball of $W^r(L_\infty)$ and $\mathcal{H}$ is taken as a space of piecewise polynomials functions of degree larger than $r-1$ on uniform partitions of X, one derives

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \lesssim (\frac{m}{\log m})^{-\frac{2r}{d+2r}}. \tag{1.14}$$

This estimate is optimal for this class Θ, up to the logarithmic factor.

The deficiency in this approach is twofold. First, it usually chooses the hypothesis classes in advance and typically assumes knowledge of the prior for this choice. Secondly, it uses linear methods of approximation and therefore misses our goal of giving an estimator which performs optimally for the full range of smoothness spaces in $L_2(X,\rho_X)$. To obtain the full range would necessarily require the use of nonlinear methods as noted above. An example in this second approach would be to use piecewise polynomials on partitions for the hypothesis class. It requires choosing the partitions in advance (e.g. uniform partitions) and therefore does not give optimal rates for general compact subsets of $C(X)$ and certainly not for general compact subsets of $L_2(X,\rho_X)$.

An in depth discussion of the approximation theory approach to building estimators for the regression problem in learning is given in [12] and the follow up papers [21] and [22].

In summary, the deficiency in the current array of practical estimators for learning the regression function lies in three directions: (i) they use knowledge of Θ in building the estimator, (ii) they circumvent the absence of knowledge of ρ_X by assuming Θ is compace in $C(X)$ rather than $L_2(X,\rho_X)$, (iii) if they use linear methods than they do not give the full array of compact subsets of $L_2(X,\rho_x)$.

The motivations for our work is to **construct practical estimators which address these drawbacks by (i) not requiring the knowledge of any prior, (ii) being optimal for a full range of relevant compact subsets Θ of $L_2(X,\rho_X)$, even though the marginal is unknown** . In the case where the marginal ρ_X is Lebesgue measure, the estimator would necessarily have to be optimal for all Besov classes which compactly embed into $L_2(X,\rho_X)$. These Besov spaces correspond to smoothness spaces of order s in L_p whenever $s > \frac{d}{p} - \frac{d}{2}$ (see [10]). One can view this problem in another way. We want to construct estimators which perform optimally on the widest class of priors. Thus, we take the viewpoint of the *maxiset* theory formalized for statistical estimation [7, 14].

To obtain estimators which satisfy (i) and (ii), we utilize the notion of *adaptivity* or *universality*: the estimation algorithm should be able to exhibit the optimal rate without the knowledge of the exact amount of smoothness r in the regression function f_ρ. A classical way to reach this goal is to perform model selection using a complexity penalty term in the empirical risk minimization, see [1, 4], Chapter 12 in [20], and [12]. In particular, one can construct one estimator which simultaneously obtains the optimal rate (1.14) for all finite balls in each of the class $W^r(L_\infty)$, $0 < r \le k$ where k is arbitrary but fixed. Of course, as we have stressed before, this class of priors is not a full spectrum of compact sets in $L_2(X,\rho_X)$.

Let us also note that the penalty approach is not always compatible with the practical requirement of *on-line* computations, by which we mean that the estimator for the sample size m can be derived by a simple update of the estimator for the sample size $m-1$, since the optimization problem needs to be globally re-solved when adding a new sample.

Finally, we are interested in deriving optimal estimates in probability, rather than only in expectation. Such estimates would in turn allow us to derive more general expectation estimates of the type $E(\|f_{\mathbf{z}} - f_\rho\|^p)$.

In the present paper we propose a class of concrete estimation schemes with the following properties:

(i) They rely on fast algorithms, which may be implemented by simple on-line updates when the sample size m is increased.

(ii) The error estimates do not require any regularity in $C(X)$ but only in the natural space $L_2(X, \rho_X)$.

(iii) The proven rates are optimal in probability and expectation for the largest possible range of smoothness classes in $L_2(X, \rho_X)$.

(iv) The scheme is universal in that it does not involve any a-priori knowledge concerning the regularity of f_ρ.

In two slightly different contexts, namely density estimation and denoising on a fixed design, it is well known that estimation procedures based on *wavelet thresholding* fullfill these requirements [15, 16, 17, 18]. In the learning theory context, the wavelet thresholding has also been used in [11] for estimation of a modification of the regression function f_ρ, namely, for estimating $(d\rho_X/dx)f_\rho$, where ρ_X is assumed to be absolutely continuous with regard to the Lebesgue measure. The main difficulty in generalizing such procedures to the distribution-free regression context is due to the presence of the marginal probability ρ_X in the $L_2(X, \rho_X)$ norm. This typically leads to the need of using wavelet-type bases which are orthogonal (or biorthogonal) with respect to this inner product. Such bases might be not easy to handle numerically and cannot be constructed exactly since ρ_X is unknown.

In this paper, we propose an approach which allows us to circumvent these difficulties, while staying in spirit close to the ideas of wavelet thresholding. In our approach, the hypothesis classes $\mathcal{H}$ are spaces of piecewise constant functions associated to partitions Λ. The key to realizing universality lies in the choice of Λ and $\mathcal{H}$ which are not simply fixed depending on the number of samples m and some a-priori knowledge on the smoothness properties of f_ρ. Rather, Λ is chosen *adaptively* based on the data $\mathbf{z}$. The partition is chosen within a set of admissible partitions based on a tree structured splitting rule.

Our partitions have the same tree structure as those used in a CART algorithm [5], yet the selection of the appropriate partition is operated quite differently: while the CART algorithm will typically minimize the empirical risk with a complexity penalty over all partitions, our algorithm selects the partition through a thresholding procedure applied to empirical quantities computed at each node of the tree which play a role similar to wavelet coefficients. While the equivalence between CART and thresholding in one or several orthonormal bases is well understood in a fixed design context [13], it is not

clear to us that our main convergence result - Theorem 2.5 - is obtainable with a CART algorithm (see in particular [19] for risk bounds obtained for CART in the distribution free bounded regression context, also with piecewise constant functions).

The present choice of piecewise constant functions limits the optimal rate to classes of low or no pointwise regularity. While the extension of our method to higher order piecewise polynomial approximations is almost straightforward, its analysis in this more general context becomes significantly more difficult and will be given in a forthcoming paper.

Our paper is organized as follows. The learning algorithm as well as the convergence results are described in section 2. The next two sections 3 and 4 are devoted to the proofs of the two main results which deal respectively with the error estimates for non-adaptive and adaptive partitions.

2 The basic strategy and the main results

2.1 Partitions and adaptive approximation

We say that a finite collection Λ of Borel subsets of X is a *partition* if the sets in Λ are pairwise disjoint and their union is all of X. The typical way of generating such partitions is through a refinement strategy. We first describe the prototypical example of dyadic partitions. For this, we assume that $X = [0,1]^d$ and denote by $\mathcal{D}_j = \mathcal{D}_j(X)$ the collection of dyadic subcubes of X of sidelength 2^{-j} and $\mathcal{D} := \cup_{j=0}^{\infty} \mathcal{D}_j$. These cubes are naturally aligned on a tree $\mathcal{T} = \mathcal{T}(\mathcal{D})$. Each node of the tree $\mathcal{T}$ is a cube $I \in \mathcal{D}$. If $I \in \mathcal{D}_j$, then its children are the 2^d dyadic cubes of $J \subset \mathcal{D}_{j+1}$ with $J \subset I$. We denote the set of children of I by $\mathcal{C}(I)$. We call I the parent of each such child J and write $I = P(J)$. A *proper* subtree $\mathcal{T}_0$ of $\mathcal{T}$ is a collection of nodes of $\mathcal{T}$ with the properties: (i) the root node $I = X$ is in $\mathcal{T}_0$, (ii) if $I \neq X$ is in $\mathcal{T}_0$ then its parent and all of its siblings are also in $\mathcal{T}_0$.

We obtain (dyadic) partitions Λ of X from finite proper subtrees $\mathcal{T}_0$ of $\mathcal{T}$. Given any such $\mathcal{T}_0$ the *outer leaves* of $\mathcal{T}_0$ consist of all $J \in \mathcal{T}$ such that $J \notin \mathcal{T}_0$ but $P(J)$ is in $\mathcal{T}_0$. The collection $\Lambda = \Lambda(\mathcal{T}_0)$ of outer leaves of $\mathcal{T}_0$ is a partition of X into dyadic cubes.

A uniform partition of X into dyadic cubes consists of all dyadic cubes in $\mathcal{D}_j(X)$ for some $j \geq 0$. Thus, each cube in a uniform partition has the same measure 2^{-jd}. Another way of generating partitions is through some refinement strategy. One begins at the root X and decides whether to refine X (i.e. subdivide X) based on some refinement criteria. If X is subdivided then one examines each child and decides whether or not to refine such a child based on the refinement strategy. Partitions obtained this way are called *adaptive*.

The results given in this paper can be described for more general refinement. We shall work in the following setting. We assume that $a \geq 2$ is a fixed integer. We assume that if X is to be refined then its children consist of a subsets of X which are a partition of X. Similarly, for each such child there is a rule which spells out how this child is refined. We assume that the child is also refined into a sets which form a partition of the child. (We could actually work with more generality and allow the number of children to depend on the cell to be refined.) Such a refinement strategy also results in a tree $\mathcal{T}$ (called the *master tree*) and children, parents, and partitions are defined as above for the special case of dyadic partitions. The refinement level j of a node is the smallest number

of refinements (starting at root) to create this node. We denote by $\mathcal{T}_j$ the proper subtree consisting of all nodes with level $\leq j$ and we denote by Λ_j the partition corresponding to $\mathcal{T}_j$

Given a partition Λ, let us denote by $\mathcal{S}_\Lambda$ the space of piecewise constant functions subordinate to Λ. Each $S \in \mathcal{S}_\Lambda$ can be written

$$S = \sum_{I \in \Lambda} a_I \chi_I, \tag{2.1}$$

where for $G \subset X$ we denote by χ_G the indicator function, i.e. $\chi_G(x) = 1$ for $x \in G$ and $\chi_G(x) = 0$ for $x \notin G$. We shall consider approximation of a given function $f \in L_2(X, \rho_X)$ by the elements of S_Λ. The best approximation to f in this space is given by

$$P_\Lambda f := \sum_{I \in \Lambda} c_I \chi_I \tag{2.2}$$

where $c_I = c_I(f)$ is given by

$$c_I := \frac{\alpha_I}{\rho_I}, \text{ with } \alpha_I := \int_I f d\rho_X \text{ and } \rho_I := \rho_X(I). \tag{2.3}$$

In the case where $\rho_I = 0$, both f_ρ and its projection are undefined on I. For notational reasons, we set in this case $c_I := 0$.

We shall be interested in two types of approximation corresponding to uniform refinement and adaptive refinement. We first discuss uniform refinement. Let

$$E_n(f) := \|f - P_{\Lambda_n} f\|, \quad n = 0, 1, \ldots \tag{2.4}$$

which is the error for uniform refinement. The decay of this error to zero is connected with the smoothness of f as measured in $L_2(X, \rho_X)$. We shall denote by $\mathcal{A}^s$ the approximation class consisting of all functions $f \in L_2(X, \rho_X)$ such that

$$E_n(f) \leq M_0 a^{-ns}, \quad n = 0, 1, \ldots. \tag{2.5}$$

Notice that $\#(\Lambda_n) = a^n$ so that the decay in (2.5) is like N^{-s} with N the number of elements in the partition. The smallest M_0 for which (2.5) holds serves to define the semi-norm $|f|_{\mathcal{A}^s}$ on $\mathcal{A}^s$. The space $\mathcal{A}^s$ can be viewed as a smoothness space of order $s > 0$ with smoothness measured with respect to ρ_X.

For example, if ρ_X is the Lebesgue measure and we use dyadic partitioning then $\mathcal{A}^{s/d} = B^s_\infty(L_2)$, $0 < s \leq 1$, with equivalent norms. Here $B^s_\infty(L_2)$ is the Besov space which can be described in terms of differences as

$$||f(\cdot + h) - f(\cdot)\|_{L_2} \leq M_0 |h|^s, \quad x, h \in X. \tag{2.6}$$

Instead of working with a-priori fixed partitions there is a second kind of approximation where the partition is generated adaptively and will vary with f. Adaptive partitions are typically generated by using some refinement criterion that determines whether or not to subdivide a given cell. We shall use a refinement criteria that is motivated by adaptive

wavelet constructions such as those given in [6] for image compression. The criteria we shall use to decide when to refine is analogous to thresholding wavelet coefficients. Indeed, it would be exactly this criteria if we were to construct a wavelet (Haar like) bases for $L_2(X, \rho_X)$.

For each cell I in the master tree $\mathcal{T}$ and any $f \in L_2(X, \rho_X)$ we define

$$\varepsilon_I(f)^2 := \sum_{J \in \mathcal{C}(I)} \frac{\left(\int_J f d\rho_X\right)^2}{\rho_J} - \frac{\left(\int_I f d\rho_X\right)^2}{\rho_I}, \tag{2.7}$$

which describes the amount of $L_2(X, \rho_X)$ energy which is increased in the projection of f_ρ onto $\mathcal{S}_\Lambda$ when the element I is refined. It is also accounts for the decreased projection error when I is refined. In fact, one easily verifies that

$$\varepsilon_I(f)^2 = \|f - c_I\|^2_{L_2(I,\rho_X)} - \sum_{J \in \mathcal{C}(I)} \|f - c_J\|^2_{L_2(J,\rho_X)}. \tag{2.8}$$

If we were in a classical situation of Lebesgue measure and dyadic refinement, then $\varepsilon_I(f)^2$ would be exactly the sum of squares of the Haar coefficients of f corresponding to I.

We can use $\varepsilon_I(f)$ to generate an adaptive partition. Given any $\eta > 0$, we let $\mathcal{T}(f, \eta)$ be the smallest proper tree that contains all $I \in \mathcal{T}$ for which $\varepsilon_I(f) > \eta$. Corresponding to this tree we have the partition $\Lambda(f, \eta)$ consisting of the outer leaves of $\mathcal{T}(f, \eta)$. We shall define some new smoothness spaces $\mathcal{B}^s$ which measure the regularity of a given function f by the size of the tree $\mathcal{T}(f, \eta)$. These spaces are related to Besov spaces in the case that ρ_X is Lebesgue measure.

Given $s > 0$, we let $\mathcal{B}^s$ be the collection of all $f \in L_2(X, \rho_X)$ such that the following is finite

$$|f|^p_{\mathcal{B}^s} := \sup_{\eta > 0} \eta^p \#(\mathcal{T}(f, \eta)), \quad \text{where } p := (s + 1/2)^{-1} \tag{2.9}$$

We obtain the norm for $\mathcal{B}^s$ by adding $\|f\|$ to $|f|_{\mathcal{B}^s}$. One can show that

$$\|f - P_{\Lambda(f,\eta)}\| \le C_s |f|_{\mathcal{B}^s} \eta^{\frac{2s}{2s+1}} \le C_s |f|_{\mathcal{B}^s} N^{-s}, \quad N := \#(\mathcal{T}(f, \eta)), \tag{2.10}$$

where the constant C_s depends only on s. For the proof of this fact we refer the reader to [6] where a similar result is proven for dyadic partitioning. It follows that every function $f \in \mathcal{B}^s$ can be approximated to order $O(N^{-s})$ by $P_\Lambda f$ for some partition Λ with $\#(\Lambda) = N$. This should be contrasted with $\mathcal{A}^s$ which has the same approximation order for the uniform partition. It is easy to see that $\mathcal{B}^s$ is larger than $\mathcal{A}^s$. In classical settings, the class $\mathcal{B}^s$ is well understood. For example, in the case of Lebesgue measure and dyadic partitions we know that each Besov space $B^s_q(L_\tau)$ with $\tau > (s/d + 1/2)^{-1}$ and $0 < q \le \infty$ arbitrary, is contained in $\mathcal{B}^{s/d}$ (see [6]). This should be compared with the $\mathcal{A}^s$ where we know that $\mathcal{A}^{s/d} = B^s_\infty(L_2)$ as we have noted earlier.

The distinction between these two forms of approximation is that in the first, the partitions are fixed in advance regardless of f but in the second form the partition can adapt to f.

We have chosen here one particular refinement strategy (based on the size of $\varepsilon_I(f)$) in generating our adaptive partitions. According to (2.10), it provides optimal convergence rates for the class $\mathcal{B}^s$. There is actually a slightly better strategy described in [2] which is guaranteed to give near optimal adaptive partitions (independent of the refinement strategy and hence not necessarily of the above form) for each individual f. We have chosen to stick with the present refinement strategy since it extends easily to empirical data (see §2.2) and it is much easier to analyze the convergence properties of this empirical scheme.

2.2 Least-squares fitting on partitions

We now return to the problem of estimation f_ρ from the given data. We shall use the functions in $\mathcal{S}_\Lambda$ for this purpose. Let us first observe that

$$P_\Lambda f_\rho = \operatorname*{argmin}_{f\in\mathcal{S}_\Lambda} \mathcal{E}(f) = \operatorname*{argmin}_{f\in\mathcal{S}_\Lambda} \int_Z (y-f(x))^2 d\rho. \tag{2.11}$$

Indeed, for all $f \in L_2(X,\rho_X)$ we have

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f-f_\rho\|^2 \tag{2.12}$$

so that minimizing $\mathcal{E}(f)$ over $\mathcal{S}_\Lambda$ is the same as minimizing $\|f_\rho - f\|$ over $f\in\mathcal{S}_\Lambda$. Note that $P_\Lambda f_\rho$ is obtained by solving N independent problems $\min_{c\in\mathbb{R}} \int_I (f_\rho - c)^2 d\rho_X$ for each element $I\in\Lambda$.

As in (1.8) we define the estimator $f_{\mathbf{z},\Lambda}$ of f_ρ on $\mathcal{S}_\Lambda$ as the empirical counterpart of $P_\Lambda f_\rho$ obtained as the solution of the least-squares problem

$$f_{\mathbf{z},\Lambda} := \operatorname*{argmin}_{f\in\mathcal{S}_\Lambda} \mathcal{E}_{\mathbf{z}}(f) = \operatorname*{argmin}_{f\in\mathcal{S}_\Lambda} \frac{1}{m}\sum_{i=1}^m (y_i - f(x_i))^2. \tag{2.13}$$

We can view our data as a multivalued function y with $y(x_i)=y_i$. Then in analogy to $P_\Lambda f_\rho$, we can view $f_{\mathbf{z},\Lambda}$ as an orthogonal projection of y onto $\mathcal{S}_\Lambda$ with respect to the empirical norm

$$\|y\|^2_{L_2(X,\delta_{\mathcal{X}})} := \frac{1}{m}\sum_{i=1}^m |y(x_i)|^2, \tag{2.14}$$

and we can compute it by solving $\#(\Lambda)$ independent problems

$$\min_{c\in\mathbb{R}} \frac{1}{m}\sum_{i=1}^m (y_i-c)^2 \chi_I(x_i), \tag{2.15}$$

for each element $I\in\Lambda$. The minimizer $c_I(\mathbf{z})$ is now given by the empirical average

$$c_I(\mathbf{z}) = \frac{\alpha_I(\mathbf{z})}{\rho_I(\mathbf{z})}, \quad \text{where } \alpha_I(\mathbf{z}) := \frac{1}{m}\sum_{i=1}^m y_i\chi_I(x_i), \qquad \rho_I(\mathbf{z}) := \frac{1}{m}\sum_{i=1}^m \chi_I(x_i). \tag{2.16}$$

Thus, we can rewrite the estimator as

$$f_{\mathbf{z},\Lambda} = \sum_{I\in\Lambda} c_I(\mathbf{z})\chi_I. \tag{2.17}$$

In the case where I contains no sample x_i (which may happen even if $\rho_I > 0$), we set $c_I(\mathbf{z}) := 0$.

A natural way of assessing the error $\|f_\rho - f_{\mathbf{z},\Lambda}\|$ is by splitting it into a bias and stochastic part : since $f_\rho - P_\Lambda f_\rho$ is orthogonal to S_Λ,

$$\|f_\rho - f_{\mathbf{z},\Lambda}\|^2 = \|f_\rho - P_\Lambda f_\rho\|^2 + \|P_\Lambda f_\rho - f_{\mathbf{z},\Lambda}\|^2 =: e_1 + e_2. \tag{2.18}$$

Concerning the variance term e_2, we shall establish the following probability estimate.

Theorem 2.1 *For any partition Λ and any $\eta > 0$,*

$$\text{Prob}\left\{\|P_\Lambda f_\rho - f_{\Lambda,\mathbf{z}}\| > \eta\right\} \le 4Ne^{-c\frac{m\eta^2}{N}}, \tag{2.19}$$

where $N := \#(\Lambda)$ and c depends only on M.

As will be explained later in detail, the following estimate of the variance term in expectation is obtained by integration of over $\eta > 0$.

Corollary 2.2 *If Λ is any partition, the mean square error is bounded by*

$$E\Big(\|P_\Lambda f_\rho - f_{\Lambda,\mathbf{z}}\|^2\Big) \le C\frac{N\log N}{m}, \tag{2.20}$$

where $N := \#(\Lambda)$ and the constant C depends only on M.

Let us consider now the case of uniform refinement. We can equilibrate the bias term with the variance term described by Theorem 2.1 and Corollary 2.2 and obtain the following result.

Theorem 2.3 *Assume that $f_\rho \in \mathcal{A}^s$ and define the estimator $f_{\mathbf{z}} := f_{\Lambda_j,\mathbf{z}}$ with j chosen as the smallest integer such that $a^{j(1+2s)} \ge \frac{m}{\log m}$. Then, given any $\beta > 0$, there is a constant $\tilde{c} = \tilde{c}(M,\beta,a)$ such that*

$$\text{Prob}\left\{\|f_\rho - f_{\mathbf{z}}\| > (\tilde{c} + |f_\rho|_{\mathcal{A}^s})\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\right\} \le Cm^{-\beta}, \tag{2.21}$$

and

$$E\Big(\|f_\rho - f_{\mathbf{z}}\|^2\Big) \le (C + |f_\rho|^2_{\mathcal{A}^s})\Big(\frac{\log m}{m}\Big)^{\frac{2s}{2s+1}}. \tag{2.22}$$

where C depends only on a and M.

Remark 2.4 *It is also possible to prove Corollary 2.2 using Theorem C* of [8]. The expectation estimate (2.22) in Theorem 2.3 can also be obtained as a consequence of Theorem 11.3 in [20] quoted in our introduction. In order to prepare for the subsequent developments direct proofs of these results are given later in §4.*

Theorem 2.3 is satisfactory in the sense that it is obtained under no assumption on the measure ρ_X and the assumption $f_\rho \in \mathcal{A}^s$ is measuring smoothness (and hence compactness) in $L_2(X, \rho_X)$, i.e. the compactness assumption is done in $L_2(\rho_X)$ rather than in L_∞. Moreover, the rate $(\frac{m}{\log m})^{-\frac{s}{2s+1}}$ is known to be optimal (or minimax) over the class $\mathcal{A}^s$ save for the logarithmic factor. However, it is unsatisfactory in the sense that the estimation procedure requires the a-priori knowledge of the smoothness parameter s which appears in the choice of the resolution level j. Moreover, as noted before, the smoothness assumption $f_\rho \in \mathcal{A}^s$ is too severe.

In the context of density estimation or denoising, it is well known that adaptive methods based on wavelet thresholding [15, 16, 17, 18] allow one to treat both defects. Our next goal is to define similar strategies in our learning context, in which two specific features have to be taken into account : the error is measured in the norm $L_2(X, \rho_X)$ and the marginal probability measure ρ_X is unknown.

2.3 A universal algorithm based on adaptive partitions

The main feature of our algorithm is to adaptively choose a partition $\Lambda = \Lambda(\mathbf{z})$ depending on the data $\mathbf{z}$. It will not require a priori knowledge of the smoothness of f_ρ but rather will learn the smoothness from the data. Thus, it will automatically choose the right size for the partition Λ.

Our starting point is the adaptive procedure introduced in §2.1 applied to the function f_ρ. We use the notation $\varepsilon_I := \varepsilon_I(f_\rho)$ in this case. Then, by (2.7),

$$\varepsilon_I^2 := \sum_{J \in \mathcal{C}(I)} \frac{\alpha_J^2}{\rho_J} - \frac{\alpha_I^2}{\rho_I}. \tag{2.23}$$

The selection of the partition Λ in our learning scheme will be based on the empirical coefficients

$$\varepsilon_I^2(\mathbf{z}) := \sum_{J \in \mathcal{C}(I)} \frac{\alpha_J^2(\mathbf{z})}{\rho_J(\mathbf{z})} - \frac{\alpha_I^2(\mathbf{z})}{\rho_I(\mathbf{z})}. \tag{2.24}$$

We define the threshold

$$\tau_m := \kappa \sqrt{\frac{\log m}{m}}, \tag{2.25}$$

where the constant κ is absolute and will be fixed later in the proof of Theorem 2.5 stated below. Let $\gamma > 0$ be an arbitrary but fixed constant. We define $j_0 = j_0(m, \gamma)$ as the largest integer j such that $a^j \leq \tau_m^{-1/\gamma}$. We next consider the smallest tree $\mathcal{T}(\mathbf{z}, m)$ which contains the set

$$\Sigma(\mathbf{z}, m) := \{I \in \cup_{j \leq j_0} \Lambda_j \ ; \ \varepsilon_I(\mathbf{z}) \geq \tau_m\}. \tag{2.26}$$

We then define the partition $\Lambda = \Lambda(\mathbf{z}, m)$ associated to this tree and the corresponding estimator $f_{\mathbf{z}} := f_{\Lambda, \mathbf{z}}$. In summary, our algorithm consists in the following steps:

(i) Compute the $\varepsilon_I(\mathbf{z})$ for $I \in \cup_{j \leq j_0} \Lambda_j$.

(ii) Threshold these quantities at level τ_m to obtain the set $\Sigma(\mathbf{z}, m)$.

(iii) Complete $\Sigma(\mathbf{z}, m)$ to the tree $\mathcal{T}(\mathbf{z}, m)$.

(iv) Compute the estimator $f_{\mathbf{z}}$ by empirical risk minimization on the partition $\Lambda(\mathbf{z}, m)$.

Further comments on the implementation will be given in the next section. The main result of this paper is the following theorem.

Theorem 2.5 *Let $\beta, \gamma > 0$ be arbitrary. Then, there exists $\kappa_0 = \kappa_0(\gamma)$ such that if $\kappa \geq \kappa_0$, then whenever $f_\rho \in \mathcal{A}^\gamma \cap \mathcal{B}^s$ for some $s > 0$, the following concentration estimate holds*

$$\text{Prob}\left\{\|f_\rho - f_{\mathbf{z}}\| \geq \tilde{c}\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\right\} \leq Cm^{-\beta}, \tag{2.27}$$

as well as the following expectation bound

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \leq C\Big(\frac{\log m}{m}\Big)^{\frac{2s}{2s+1}}, \tag{2.28}$$

where the constants $\tilde{c}$ and C are independent of m.

Theorem 2.5 is more satisfactory than Theorem 2.3 in two respects: (i) the optimal rate $(\frac{\log m}{m})^{\frac{s}{2s+1}}$ is now obtained under weaker smoothness assumptions on the regression function, namely, $f_\rho \in \mathcal{B}^s$ in place of $f_\rho \in \mathcal{A}^s$, with the extra assumption of $f_\rho \in \mathcal{A}^\gamma$ smoothness with $\gamma > 0$ arbitrarily small, (ii) the algorithm is universal. Namely, the value of s does not enter the definition of the algorithm. Indeed, the algorithm automatically exploits this unknown smoothness through the samples $\mathbf{z}$. We note however that the algorithm does require the knowledge of the parameter γ which can be arbitrarily small. It is actually possible to build an algorithm without assuming knowledge of a $\gamma > 0$ by using the adaptive tree algorithm in [2]. However, the implementation of such an algorithm would involve complications we wish to avoid in this presentation.

2.4 Remarks on algorithmic aspects and on-line implementation

Our first remarks concern the construction of the adaptive partition $\Lambda(\mathbf{z}, m)$ for a fixed m which requires the computation of the numbers $\varepsilon_I(\mathbf{z})$ for $I \in \Lambda_j$ when j satisfies $a^j \leq \tau_m^{-1/\gamma}$. This would require the computation of $O(m \ln m)$ coefficients. One can actually save a substantial amount of computation by remarking that by definition we always have

$$\varepsilon_I(\mathbf{z})^2 \leq \mathcal{E}_I(\mathbf{z}) \tag{2.29}$$

with $\mathcal{E}_I(\mathbf{z}) := \|y - c_I(\mathbf{z})\|^2_{L_2(\delta_{\mathcal{X}}, I)}$ the least-square error on I. In contrast to $\varepsilon_I(\mathbf{z})$, the quantity $\mathcal{E}_I(\mathbf{z})$ is monotone with respect to inclusion:

$$J \subset I \;\Rightarrow\; \mathcal{E}_J(\mathbf{z}) \leq \mathcal{E}_I(\mathbf{z}). \tag{2.30}$$

This allows one to organize the search for those I satisfying $\varepsilon_I(\mathbf{z}) \geq \tau_m$ from coarse to fine elements. In particular, one no longer has to check those descendents of an element I for which $\mathcal{E}_I(\mathbf{z})$ is less than τ_m.

Our next remarks concern the on-line implementation of the algorithm. Suppose that we have computed $\rho_I(\mathbf{z})$, $\alpha_I(\mathbf{z})$ and the $\varepsilon_I(\mathbf{z})$ where $\mathbf{z}$ contains m samples. If we now add a new sample (x_{m+1}, y_{m+1}) to $\mathbf{z}$ to obtain $\mathbf{z}^+$, the new ρ_I and α_I are

$$\rho_I(\mathbf{z}^+) = \frac{m}{m+1}(\rho_I(\mathbf{z}) + \chi_I(x_{m+1})) \tag{2.31}$$

and

$$\alpha_I(\mathbf{z}^+) = \frac{m}{m+1}(\alpha_I(\mathbf{z}) + y_{m+1}\chi_I(x_{m+1})). \tag{2.32}$$

In particular, we see that at each level j, only one I is affected by the new sample. Therefore, if we store the quantities $\rho_I(\mathbf{z})$ and $\alpha_I(\mathbf{z})$ in the current partition, then this new step requires at most j_0 additional computations in the case where j_0 is not increased. In the case where j_0 is increased to j_0+1 (this may happen because τ_m is decreased), the computations of the quantities $\rho_I(\mathbf{z})$ and $\alpha_I(\mathbf{z})$ need to be performed, of course, for all the elements in the newly added level.

3 Proof of the results on non-adaptive partitions

We first give the proof of Theorem 2.1. Let Λ be any partition. By (2.2) and (2.17), we can write

$$\|P_\Lambda f_\rho - f_{\Lambda,\mathbf{z}}\|^2 = \sum_{I\in\Lambda} |c_I - c_I(\mathbf{z})|^2 \rho_I. \tag{3.1}$$

According to their definitions (2.3), (2.16), both c_I and $c_I(\mathbf{z})$ are bounded in modulus by M. Therefore, given $\eta > 0$, if we define

$$\Lambda^- := \{I \in \Lambda \,:\, \rho_I \le \frac{\eta^2}{8NM^2}\}, \tag{3.2}$$

we clearly have

$$\sum_{I\in\Lambda^-} |c_I - c_I(\mathbf{z})|^2 \rho_I \le \frac{\eta^2}{2}. \tag{3.3}$$

We next consider the complement set $\Lambda^+ = \Lambda \setminus \Lambda^-$. In order to prove (2.19), it now suffices to establish that for all $I \in \Lambda^+$

$$\text{Prob}\left\{|c_I(\mathbf{z}) - c_I|^2 \ge \frac{\eta^2}{2N\rho_I}\right\} \le 4e^{-c\frac{m\eta^2}{N}}. \tag{3.4}$$

To see this, we write $\rho_I(\mathbf{z}) = (1+\mu_I)\rho_I$ and remark that if $|\mu_I| \le 1/2$ we have

$$\begin{aligned} |c_I(\mathbf{z}) - c_I| &= \left|\frac{\alpha_I(\mathbf{z})}{\rho_I(\mathbf{z})} - \frac{\alpha_I}{\rho_I}\right| = \frac{1}{\rho_I(1+\mu_I)}|\alpha_I(\mathbf{z}) - \alpha_I - \mu_I\alpha_I| \\ &\le 2\rho_I^{-1}(|\alpha_I(\mathbf{z}) - \alpha_I| + |\alpha_I\mu_I|). \end{aligned} \tag{3.5}$$

It follows that $|c_I(\mathbf{z}) - c_I| \le \frac{\eta}{\sqrt{2N\rho_I}}$ provided that we have jointly

$$|\alpha_I(\mathbf{z}) - \alpha_I| \le \frac{\eta\sqrt{\rho_I}}{4\sqrt{2N}}, \tag{3.6}$$

and (since $\alpha_I \mu_I = \alpha_I(\rho_I(\mathbf{z}) - \rho_I)/\rho_I$)

$$|\rho_I(\mathbf{z}) - \rho_I| \leq \min\left\{ \frac{1}{2}\rho_I, \frac{\eta \rho_I^{3/2}}{4\sqrt{2N}|\alpha_I|} \right\} \tag{3.7}$$

and therefore

$$\begin{aligned} \text{Prob}\left\{ |c_I(\mathbf{z}) - c_I|^2 \geq \frac{\eta^2}{2N\rho_I} \right\} &\leq \text{Prob}\left\{ |\alpha_I(\mathbf{z}) - \alpha_I| \geq \frac{\eta\sqrt{\rho_I}}{4\sqrt{2N}} \right\} \\ &+ \text{Prob}\left\{ |\rho_I(\mathbf{z}) - \rho_I| \geq \min\left\{ \frac{1}{2}\rho_I, \frac{\eta\rho_I^{3/2}}{4\sqrt{2N}|\alpha_I|} \right\} \right\}. \end{aligned}$$

In order to estimate these probabilities, we shall use Bernstein's inequality which says that for m independent realizations ζ_i of a random variable ζ such that $|\zeta(z) - E(\zeta)| \leq M_0$ and $\text{Var}(\zeta) = \sigma^2$, one has for any $\varepsilon > 0$

$$\text{Prob}\left\{ \left| \frac{1}{m}\sum_{i=1}^{m} \zeta(z_i) - E(\zeta) \right| \geq \varepsilon \right\} \leq 2e^{-\frac{m\varepsilon^2}{2(\sigma^2 + M_0\varepsilon/3)}}. \tag{3.8}$$

In our context, we apply this inequality to $\zeta = y\chi_I(x)$ for which $E(\zeta) = \alpha_I$, $M_0 \leq 2M$ and $\sigma^2 \leq M^2\rho_I$, and to $\zeta = \chi_I(x)$ for which $E(\zeta) = \rho_I$, $M_0 \leq 1$, and $\sigma^2 \leq \rho_I$.

We first obtain that

$$\begin{aligned} \text{Prob}\left\{ |\alpha_I(\mathbf{z}) - \alpha_I| \geq \frac{\eta\sqrt{\rho_I}}{4\sqrt{2N}} \right\} &\leq 2e^{-\frac{m\eta^2\rho_I}{64N(M^2\rho_I + 2M\eta\sqrt{\rho_I/2N}/12)}} \\ &\leq 2e^{-\frac{m\eta^2\rho_I}{512NM^2\rho_I}} \\ &\leq 2e^{-c\frac{m\eta^2}{N}}, \end{aligned}$$

with $c = [512M^2]^{-1}$, where we have used in the second line that $I \in \Lambda^+$ to bound the second term in the denominator of the exponential by the first term in the denominator. We next obtain in the case where $\frac{1}{2}\rho_I \leq \frac{\eta\rho_I^{3/2}}{4\sqrt{2N}|\alpha_I|}$

$$\text{Prob}\left\{ |\rho_I(\mathbf{z}) - \rho_I| \geq \frac{1}{2}\rho_I \right\} \leq 2e^{-\frac{m\rho_I^2}{8(\rho_I + \rho_I/6)}} = 2e^{-\frac{3}{28}m\rho_I} \leq 2e^{-c\frac{m\eta^2}{N}}$$

with $c = [\frac{224}{3}M^2]^{-1}$ where we have used in the last line that $I \in \Lambda^+$. Finally, in the case where $\frac{1}{2}\rho_I \geq \frac{\eta\rho_I^{3/2}}{4\sqrt{2N}|\alpha_I|}$, we obtain

$$\text{Prob}\left\{ |\rho_I(\mathbf{z}) - \rho_I| \geq \frac{\eta\rho_I^{3/2}}{4\sqrt{2N\rho_I}|\alpha_I|} \right\} \leq 2e^{-\frac{m\eta^2\rho_I^3}{64N\rho_I|\alpha_I|^2(7\rho_I/6)}} \leq 2e^{-c\frac{m\eta^2}{N}}$$

with $c = [\frac{448}{6}M^2]^{-1}$ since $|\alpha_I| \leq M\rho_I$. Therefore, we obtain (3.4) with the smallest of the three values of c namely $c = [526M^2]^{-1}$, which concludes the proof of Theorem 2.1.

Remark 3.1 *The constant c in the estimate behaves like $1/M^2$ and therefore degenerates to 0 as $M \to +\infty$. This is due to the fact that we are using Bernstein's estimate as a concentration inequality since we are lacking any other information on the conditional law $\rho(y|x)$. For more specific models where we have more information on the conditional law $\rho(y|x)$, one can avoid the limitation $|y| \leq M$. For instance, in the Gaussian regression problem $y_i = f_\rho(x_i) + g_i$ where g_i are i.i.d. Gaussian (and therefore unbounded) variables $\mathcal{N}(0, \sigma^2)$, the probabilistic estimate (2.19) can be obtained by a direct use of the concentration property of the gaussian.*

The proof of Corollary 2.2 follows by integration of (2.19) over η:

$$\begin{aligned} E\Big(\|P_\Lambda f_\rho - f_{\Lambda,\mathbf{z}}\|^2_{L_2(X,\rho_X)}\Big) &= \int_0^{+\infty} \eta \,\mathrm{Prob}\left\{\|P_\Lambda f_\rho - f_{\Lambda,\mathbf{z}}\|_{L_2(\rho_X)} > \eta\right\} d\eta \\ &\leq \int_0^{+\infty} \eta \min\{1, 4Ne^{-c\frac{m\eta^2}{N}}\} d\eta \\ &= \int_0^{\eta_0} \eta \, d\eta + \int_{\eta_0}^{+\infty} 4N\eta \, e^{-c\frac{m\eta^2}{N}} d\eta \\ &= \frac{\eta_0^2}{2} + \frac{2N^2}{cm} e^{-c\frac{m\eta_0^2}{N}}, \end{aligned}$$

where η_0 is such that $4Ne^{-c\frac{m\eta_0^2}{N}} = 1$, or equivalently $\eta_0^2 = \frac{N\log(4N)}{cm}$. This proves the estimate (2.20).

Finally, to prove the estimates in Theorem 2.3, we first note that, by assumption, $N = \#(\Lambda_j) \leq a^{j+1} \leq a^2 \left(\frac{m}{\log m}\right)^{\frac{1}{2s+1}}$. Further, from the definition of $\mathcal{A}^s$, we have

$$\|f_\rho - P_{\Lambda_j} f_\rho\| \leq |f_\rho|_{\mathcal{A}^s} a^{-js} \leq |f_\rho|_{\mathcal{A}^s} \left(\frac{\log m}{m}\right)^{\frac{s}{2s+1}}. \tag{3.9}$$

Hence, using Theorem 2.1, we see that the probability on the left of (2.21) is bounded from above by

$$\mathrm{Prob}\left\{\|P_\Lambda f_\rho - f_{\Lambda,\mathbf{z}}\| > \tilde{c}\left(\frac{\log m}{m}\right)^{\frac{s}{2s+1}}\right\} \leq 4a^2 m e^{-\frac{c\tilde{c}^2 \log m}{a^2}} \tag{3.10}$$

which does not exceed $Cm^{-\beta}$ provided $\tilde{c}^2 c > a^2(1+\beta)$. The proof of (2.22) follows in a similar way from Corollary 2.2.

4 Proof of Theorem 2.5

The remainder of this paper is devoted to a proof of Theorem 2.5. We begin with our notation. Recall that the tree $\mathcal{T}(f_\rho, \eta)$ is the smallest tree which contains all I for which $\varepsilon_I = \varepsilon_I(f_\rho)$ is larger than η. $\Lambda(f_\rho, \eta)$ is the partition induced by the outer leaves of $\mathcal{T}(f_\rho, \eta)$. We use τ_m as defined in (2.25) and $j_0 = j_0(m)$ is the largest integer such that $a^{j_0} \leq \tau_m^{-1/\gamma}$. For any partition Λ we write $f_{\mathbf{z},\Lambda} = \sum_{I\in\Lambda} c_I(\mathbf{z})\chi_I$.

If Λ_0 and Λ_1 are two adaptive partitions respectively associated to trees $\mathcal{T}_0$ and $\mathcal{T}_1$ we denote by $\Lambda_0 \vee \Lambda_1$ and $\Lambda_0 \wedge \Lambda_1$ the partitions associated to the trees $\mathcal{T}_0 \cup \mathcal{T}_1$ and $\mathcal{T}_0 \cap \mathcal{T}_1$, respectively. Given any $\eta > 0$, we define the partitions $\Lambda(\eta) := \Lambda(f_\rho, \eta) \wedge \Lambda_{j_0}$ and $\Lambda(\eta, \mathbf{z})$ associated with the smallest trees containing those I

such that $\varepsilon_I \geq \eta$ and $\varepsilon_I(\mathbf{z}) \geq \eta$, respectively, and such that the refinement level j of any I in either one of these two partitions satisfies $j \leq j_0$. In these terms our estimator $f_{\mathbf{z}}$ is given by

$$f_{\mathbf{z}} = f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})}. \tag{4.1}$$

With this notation in hand, we begin now with the proof of the Theorem. Using the triangle inequality, we have

$$\|f_\rho - f_{\mathbf{z},m}\| \leq e_1 + e_2 + e_3 + e_4 \tag{4.2}$$

with each term defined by

$$\begin{aligned}
e_1 &:= \|f_\rho - P_{\Lambda(\tau_m,\mathbf{z})\vee\Lambda(b\tau_m)} f_\rho\|, \\
e_2 &:= \|P_{\Lambda(\tau_m,\mathbf{z})\vee\Lambda(b\tau_m)} f_\rho - P_{\Lambda(\tau_m,\mathbf{z})\wedge\Lambda(\tau_m/b)} f_\rho\|, \\
e_3 &:= \|P_{\Lambda(\tau_m,\mathbf{z})\wedge\Lambda(\tau_m/b)} f_\rho - f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})\wedge\Lambda(\tau_m/b)}\|, \\
e_4 &:= \|f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})\wedge\Lambda(\tau_m/b)} - f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})}\|,
\end{aligned}$$

with $b := 2\sqrt{a-1} > 1$.

The first term e_1 can be treated by a deterministic estimate. Namely, since $\Lambda(\tau_m, \mathbf{z}) \vee \Lambda(b\tau_m)$ is a finer partition than $\Lambda(b\tau_m)$, we have with probability one

$$\begin{aligned}
e_1 &\leq \|f_\rho - P_{\Lambda(b\tau_m)} f_\rho\| \leq \|f_\rho - P_{\Lambda(f_\rho,b\tau_m)} f_\rho\| + \|P_{\Lambda(f_\rho,b\tau_m)} f_\rho - P_{\Lambda(b\tau_m)} f_\rho\| \\
&\leq \|f_\rho - P_{\Lambda(f_\rho,b\tau_m)} f_\rho\| + \|f_\rho - P_{\Lambda_{j_0}} f_\rho\| \\
&\leq C_s (b\tau_m)^{\frac{2s}{2s+1}} |f_\rho|_{\mathcal{B}^s} + a^{-\gamma j_0} |f_\rho|_{\mathcal{A}^\gamma} \\
&\leq C_s (b\tau_m)^{\frac{2s}{2s+1}} |f_\rho|_{\mathcal{B}^s} + a^\gamma \tau_m |f_\rho|_{\mathcal{A}^\gamma}.
\end{aligned}$$

Therefore we conclude that

$$e_1 \leq C_s (\kappa^{\frac{2s}{2s+1}} + a^\gamma \kappa) \max\{|f_\rho|_{\mathcal{A}^\gamma}, |f_\rho|_{\mathcal{B}^s}\} \Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}, \tag{4.3}$$

whenever $f \in \mathcal{B}^s \cap \mathcal{A}^\gamma$.

The third term e_3 can be treated by the estimate (2.19) of Theorem 2.1:

$$\text{Prob}\{e_3 > \eta\} \leq 4N e^{-c\frac{m\eta^2}{N}}, \tag{4.4}$$

with

$$N = \#(\Lambda(\tau_m, \mathbf{z}) \wedge \Lambda(\tau_m/b)) \leq \#(\Lambda(\tau_m/b)) \leq \#(\Lambda(f_\rho, \tau_m/b)).$$

Hence we infer from (2.9) that

$$N \leq b^p \tau_m^{-p} |f_\rho|^p_{\mathcal{B}^s} = b^p \tau_m^{-\frac{2}{2s+1}} |f_\rho|^p_{\mathcal{B}^s} = b^p \kappa^{-\frac{2}{2s+1}} |f_\rho|^p_{\mathcal{B}^s} \Big(\frac{m}{\log m}\Big)^{\frac{1}{2s+1}}, \tag{4.5}$$

where we have used that $1/p = 1/2 + s$.

Concerning the two remaining terms e_2 and e_4, we shall prove that for a fixed but arbitrary $\beta > 0$, we have

$$\text{Prob}\{e_2 > 0\} + \text{Prob}\{e_4 > 0\} \leq Cm^{-\beta}, \tag{4.6}$$

whenever $\kappa \geq \kappa_0$ with κ_0 depending on β and γ and with C depending only on κ and γ.

Before proving this result, let us show that the combination (4.3) , (4.4) and (4.6) imply the validity of the estimates (2.27) and (2.28) in Theorem 2.5. We fix the value of β and we fix any constant κ for which (4.6) holds. Let $\eta_1 := \tilde{c}(\frac{\log m}{m})^{\frac{s}{2s+1}}$ with $\tilde{c}$ from (2.27) and $\eta_2 := c_0(\frac{\log m}{m})^{\frac{s}{2s+1}}$ with $c_0 := C_s(\kappa^{\frac{2s}{2s+1}} + a^\gamma\kappa)\max\{|f_\rho|_{\mathcal{A}^\gamma}, |f_\rho|_{\mathcal{B}^s}\}$. From (4.3) it follows that for $\tilde{c} > c_0$ we have $\text{Prob}\{\|f_\rho - f_{\mathbf{z},m}\| > \eta_1\} \leq \text{Prob}\{e_2 + e_3 + e_4 > \eta_1 - \eta_2\}$. Hence, defining $\eta = (\tilde{c} - c_0)(\frac{\log m}{m})^{\frac{s}{2s+1}}$, the probability on the left side of (2.27) does not exceed

$$\text{Prob}\{e_2 > 0\} + \text{Prob}\{e_3 > \eta\} + \text{Prob}\{e_4 > 0\} \leq \text{Prob}\{e_3 > \eta\} + Cm^{-\beta},$$

Moreover, on account of (4.4) and (4.5), we can estimate $\text{Prob}\{e_3 > \eta\}$ by

$$\begin{aligned}
\text{Prob}\{e_3 > \eta\} &\leq C\Big(\frac{m}{\log m}\Big)^{\frac{1}{2s+1}} e^{-cm\eta^2 b^{-p}\kappa^{-\frac{2}{2s+1}}|f_\rho|_{\mathcal{B}^s}^{-p}\left(\frac{\log m}{m}\right)^{\frac{1}{2s+1}}} \\
&= C\Big(\frac{m}{\log m}\Big)^{\frac{1}{2s+1}} e^{-cD^2 m\left(\frac{\log m}{m}\right)} \\
&= C\Big(\frac{m}{\log m}\Big)^{\frac{1}{2s+1}} m^{-cD^2} \\
&\leq Cm^{1-cD^2}
\end{aligned}$$

where $D := \frac{(\tilde{c}-c_0)^2}{\kappa^{\frac{2}{2s+1}} b^p |f|_{\mathcal{B}^s}^p}$. The concentration estimate (2.27) follows now by taking $\tilde{c}$ large enough so that $1 - cD^2 + \beta \leq 0$.

For the expectation estimate (2.28), we recall that according to Corollary 2.2, we have

$$E(e_3^2) \leq C\frac{N\log N}{m} \leq C\frac{\left(\frac{m}{\log m}\right)^{\frac{1}{2s+1}}\log m}{m} = C\Big(\frac{\log m}{m}\Big)^{\frac{2s}{1+2s}} \tag{4.7}$$

We then remark that we always have $e_2^2 \leq 4M^2$, and therefore

$$E(e_2^2) \leq 4M^2\text{Prob}\{e_2 > 0\} \leq Cm^{-\beta} \leq C\Big(\frac{m}{\log m}\Big)^{-\frac{2s}{2s+1}}, \tag{4.8}$$

by choosing β larger than $2s/(2s+1)$, for example $\beta = 1$. The same holds for e_4 and therefore we obtain (2.28).

It remains to prove (4.6). The main tool here is a probabilistic estimate of how the empirical coefficient $\varepsilon_I(\mathbf{z})$ may differ from ε_I with respect to the threshold. This is expressed by the following lemma.

Lemma 4.1 *For any $\eta > 0$ and any element $I \in \Lambda_{j_0}$, one has*

$$\text{Prob}\{\varepsilon_I(\mathbf{z}) \leq \eta \text{ and } \varepsilon_I \geq b\eta\} \leq Ce^{-cm\eta^2} \tag{4.9}$$

and

$$\text{Prob}\{\varepsilon_I \leq \eta \text{ and } \varepsilon_I(\mathbf{z}) \geq b\eta\} \leq Ce^{-cm\eta^2} \tag{4.10}$$

where the constant c depends only on M and the constant C depends only on κ and γ.

Before proving Lemma 4.1, let us show how this results implies (4.6). We first consider the second term e_2. Clearly $e_2 = 0$ if $\Lambda(\tau_m, \mathbf{z}) \vee \Lambda(b\tau_m) = \Lambda(\tau_m, \mathbf{z}) \wedge \Lambda(\tau_m/b)$ or equivalently $\mathcal{T}(\tau_m, \mathbf{z}) \cup \mathcal{T}(b\tau_m) = \mathcal{T}(\tau_m, \mathbf{z}) \cap \mathcal{T}(\tau_m/b)$. Now if the inclusion $\mathcal{T}(\tau_m, \mathbf{z}) \cap \mathcal{T}(\tau_m/b) \subset \mathcal{T}(\tau_m, \mathbf{z}) \cup \mathcal{T}(b\tau_m)$ is strict, then one either has $\mathcal{T}(\tau_m, \mathbf{z}) \not\subset \mathcal{T}(\tau_m/b)$ or $\mathcal{T}(b\tau_m) \not\subset \mathcal{T}(\tau_m, \mathbf{z})$. Thus, there either exists an I such that both $\varepsilon_I(\mathbf{z}) \leq \tau_m$ and $\varepsilon_I \geq b\tau_m$ or there exists an I such that both $\varepsilon_I(\mathbf{z}) \geq \tau_m$ and $\varepsilon_I < \tau_m/b$. It follows that

$$\begin{aligned} \text{Prob}\{e_2 > 0\} \quad \leq \quad & \sum_{I \in \Lambda_{j_0}} \text{Prob}\{\varepsilon_I(\mathbf{z}) \leq \tau_m \text{ and } \varepsilon_I \geq b\tau_m\} \\ & + \sum_{I \in \Lambda_{j_0}} \text{Prob}\{\varepsilon_I(\mathbf{z}) \geq \tau_m \text{ and } \varepsilon_I \leq b\tau_m\}. \end{aligned} \tag{4.11}$$

Using (4.9) with $\eta = \tau_m$ yields

$$\begin{aligned} \textstyle\sum_{I \in \Lambda_{j_0}} \text{Prob}\{\varepsilon_I(\mathbf{z}) \leq \tau_m \text{ and } \varepsilon_I \geq b\tau_m\} \quad & \leq \#(\Lambda_{j_0}) e^{-cm\tau_m^2} \\ & \leq \#(\Lambda_0) a^{j_0} e^{-c\kappa^2 \log m} \\ & \leq \#(\Lambda_0) \tau_m^{-1/\gamma} m^{-c\kappa^2} \\ & \leq C m^{1/\gamma - c\kappa^2}. \end{aligned}$$

We can treat the second sum in (4.11) the same way and obtain the same bound. By similar considerations, we obtain

$$\text{Prob}\{e_4 > 0\} \leq \sum_{I \in \Lambda_J} \text{Prob}\{\varepsilon_I(\mathbf{z}) \geq \tau_m \text{ and } \varepsilon_I \leq \tau_m/b\}, \tag{4.12}$$

and we use (4.10) with $\eta = \tau_m/b$ which yields $\text{Prob}\{e_4 > 0\} \leq Cm^{1/\gamma - c\kappa^2/b^2}$. We therefore obtain (4.6) by choosing $\kappa \geq \kappa_0$ with $c\kappa_0^2 = 4b^2/\gamma$.

We are left with the proof of Lemma 4.1. As a first step, we show that the proof can be reduced to the particular case $a = 2$. To this end, we remark that the splitting of I into its a children $\{J_1, \cdots, J_a\}$ can be decomposed into $a - 1$ steps consisting of splitting an element into a pair of elements: defining $I_n := I \setminus (J_1 \cup \cdots \cup J_n)$ we start from $I = I_0$ and refine iteratively I_{n-1} into the two elements I_n and J_n, for $n = 1, \cdots, a-1$. By orthogonality, we can write

$$\varepsilon_I^2 := \sum_{n=0}^{a-2} \varepsilon_{I_n}^2, \tag{4.13}$$

where $\varepsilon_{I_n}^2$ is the amount of $L_2(X,\rho_X)$ energy which is increased in the projection of f_ρ when I_{n+1} is refined into I_n and J_n. In a similar way, we can write for the observed quantities

$$\varepsilon_I^2(\mathbf{z}) := \sum_{n=0}^{a-2} \varepsilon_{I_n}^2(\mathbf{z}), \tag{4.14}$$

Now if $\varepsilon_I^2 \le \eta^2$ and $\varepsilon_I(\mathbf{z})^2 \ge b^2\eta^2 = 4(a-1)\eta^2$, it follows that there exist $n \in \{0,\cdots,a-2\}$ such that $\varepsilon_{I_n}^2 \le \eta^2$ and $\varepsilon_{I_n}(\mathbf{z})^2 \ge 4\eta^2$. Therefore,

$$\mathrm{Prob}\{\varepsilon_I \le \eta \ \text{ and } \ \varepsilon_I(\mathbf{z}) \ge b\eta\} \le \sum_{n=0}^{a-2} \mathrm{Prob}\{\varepsilon_{I_n} \le \eta \ \text{ and } \ \varepsilon_{I_n}(\mathbf{z}) \ge 2\eta\}, \tag{4.15}$$

and similarly

$$\mathrm{Prob}\{\varepsilon_I(\mathbf{z}) \le \eta \ \text{ and } \ \varepsilon_I \ge b\eta\} \le \sum_{n=0}^{a-2} \mathrm{Prob}\{\varepsilon_{I_n}(\mathbf{z}) \le \eta \ \text{ and } \ \varepsilon_{I_n} \ge 2\eta\}, \tag{4.16}$$

so that the estimates (4.9) and (4.10) for $a > 2$ follow from the same estimates established for $a = 2$ in which case $b = 2$.

In the case $a = 2$, we denote by I^+ and I^- the two children of I. Note that if $\rho_J = 0$ for $J = I^+$ or for $J = I^-$, there is nothing to prove, since in this case we find that $\varepsilon_I = \varepsilon_I(\mathbf{z}) = 0$ with probability one. We therefore assume that $\rho_J > 0$ for $J = I^+$ and I^-. We first rewrite ε_I as follows

$$\begin{aligned}
\varepsilon_I^2 &= \frac{\alpha_{I^+}^2}{\rho_{I^+}} + \frac{\alpha_{I^-}^2}{\rho_{I^-}} - \frac{\alpha_I^2}{\rho_I} = \rho_{I^+}c_{I^+}^2 + \rho_{I^-}c_{I^-}^2 - \rho_I c_I^2 \\
&= \rho_{I^+}c_{I^+}^2 + \rho_{I^-}c_{I^-}^2 - \rho_I((\rho_{I^+}c_{I^+} + \rho_{I^-}c_{I^-})/\rho_I)^2 \\
&= \frac{\rho_{I^+}\rho_{I^-}}{\rho_I}(c_{I^+} - c_{I^-})^2,
\end{aligned}$$

and therefore $\varepsilon_I = |\beta_I|$ with

$$\beta_I := \sqrt{\frac{\rho_{I^+}\rho_{I^-}}{\rho_I}}(c_{I^+} - c_{I^-}). \tag{4.17}$$

In a similar way we obtain $\varepsilon_I(\mathbf{z}) = |\beta_I(\mathbf{z})|$ with

$$\beta_I(\mathbf{z}) := \sqrt{\frac{\rho_{I^+}(\mathbf{z})\rho_{I^-}(\mathbf{z})}{\rho_I(\mathbf{z})}}(c_{I^+}(\mathbf{z}) - c_{I^-}(\mathbf{z})). \tag{4.18}$$

Introducing the quantities $a_{I^+} = \sqrt{\frac{\rho_{I^-}}{\rho_I\rho_{I^+}}}$ and $a_{I^-} = \sqrt{\frac{\rho_{I^+}}{\rho_I\rho_{I^-}}}$ and their empirical counterpart $a_{I^+}(\mathbf{z})$ and $a_{I^-}(\mathbf{z})$ we can rewrite β_I and $\beta_I(\mathbf{z})$ as

$$\beta_I = a_{I^+}\alpha_{I^+} - a_{I^-}\alpha_{I^-} \tag{4.19}$$

and

$$\beta_I(\mathbf{z}) = a_{I^+}(\mathbf{z})\alpha_{I^+}(\mathbf{z}) - a_{I^-}(\mathbf{z})\alpha_{I^-}(\mathbf{z}). \tag{4.20}$$

It follows that

$$|\varepsilon_I - \varepsilon_I(\mathbf{z})| \leq |a_{I^+}\alpha_{I^+} - a_{I^+}(\mathbf{z})\alpha_{I^+}(\mathbf{z})| + |a_{I^-}\alpha_{I^-} - a_{I^-}(\mathbf{z})\alpha_{I^-}(\mathbf{z})|. \tag{4.21}$$

We next introduce the numbers δ_J defined by the relation $\rho_J(\mathbf{z}) = (1+\delta_J)\rho_J$, for $J = I^+, I^-$ or I. It is easily seen that if $|\delta_J| \leq \delta \leq 1/4$ for $J = I^+, I^-$ and I, one has

$$a_{I^+}(\mathbf{z}) = (1+\mu_I^+)a_{I^+} \tag{4.22}$$

with $|\mu_I^+| \leq 3\delta$. This follows indeed from the basic inequalities

$$1 - 3\delta \leq \sqrt{\frac{(1-\delta)}{(1+\delta)^2}} \leq \sqrt{\frac{(1+\delta)}{(1-\delta)^2}} \leq 1 + 3\delta \tag{4.23}$$

which hold for $0 \leq \delta \leq 1/4$. Therefore if $|\delta_J| \leq \delta \leq 1/4$ for $J = I^+, I^-$ and I, we have

$$\begin{aligned} |a_{I^+}\alpha_{I^+} - a_{I^+}(\mathbf{z})\alpha_{I^+}(\mathbf{z})| &\leq a_{I^+}(\mathbf{z})|\alpha_{I^+} - \alpha_{I^+}(\mathbf{z})| + |\alpha_{I^+}(a_{I^+} - a_{I^+}(\mathbf{z}))| \\ &\leq 2a_{I^+}|\alpha_{I^+} - \alpha_{I^+}(\mathbf{z})| + 3\delta a_{I^+}|\alpha_{I^+}|. \end{aligned}$$

By similar considerations, we obtain the estimate

$$|a_{I^-}\alpha_{I^-} - a_{I^-}(\mathbf{z})\alpha_{I^-}(\mathbf{z})| \leq 2a_{I^-}|\alpha_{I^-} - \alpha_{I^-}(\mathbf{z})| + 3\delta a_{I^-}|\alpha_{I^-}|,$$

and therefore

$$|\varepsilon_I - \varepsilon_I(\mathbf{z})| \leq \sum_{K=I^+,I^-} 2a_K|\alpha_K - \alpha_K(\mathbf{z})| + 3\delta a_K|\alpha_K|. \tag{4.24}$$

We first turn to (4.9), which corresponds to the case where $\varepsilon_I \geq 2\eta$ and $\varepsilon_I(\mathbf{z}) \leq \eta$. In this case, we remark that we have

$$\eta^2 \leq \frac{\varepsilon_I^2}{4} = \frac{\rho_{I^+}\rho_{I^-}}{\rho_I}\frac{(c_{I^+} - c_{I^-})^2}{4} \leq M^2\rho_L, \tag{4.25}$$

for $L = I^+, I^-$ and I. Combining (4.24) and (4.25), we estimate the probability by

$$\text{Prob}\{\varepsilon_I(\mathbf{z}) \leq \eta \text{ and } \varepsilon_I \geq 2\eta\} \leq \sum_{K=I^+,I^-}\Big(p_K + \sum_{J=I^-,I^+,I} q_{K,J}\Big), \tag{4.26}$$

with

$$p_K := \text{Prob}\{|\alpha_K - \alpha_K(\mathbf{z})| \geq [8a_K]^{-1}\eta \text{ given } \rho_K \geq \frac{\eta^2}{M^2}\}, \tag{4.27}$$

and

$$q_{K,J} := \text{Prob}\{|\rho_J - \rho_J(\mathbf{z})| \geq \rho_J \min\{\frac{1}{4}, \eta[12a_K|\alpha_K|]^{-1}\} \text{ given } \rho_J \geq \frac{\eta^2}{M^2}\}. \tag{4.28}$$

Using Bernstein's inequality, we can estimate p_K as follows

$$p_K \leq 2e^{-\frac{m\eta^2}{2(64a_K^2M^2\rho_K + 8a_K\eta M/3)}} \leq 2e^{-\frac{m\eta^2}{2(64a_K^2M^2\rho_K + 8a_K\sqrt{\rho_K}M^2/3)}} \leq 2e^{-cm\eta^2},$$

with $c = [(128 + 16/3)M^2]^{-1}$, where we have used $\eta^2 \le \rho_K M^2$ in the second line and the fact that $a_K^2 \rho_K \le 1$ in the third line.

In the case where $12 a_K |\alpha_K| \le 4\eta$, we estimate $q_{K,J}$ by

$$q_{K,J} \le 2e^{-\frac{m\rho_J}{2(16+4/3)}} \le 2e^{-cm\eta^2},$$

with $c = [(32 + 8/3)M^2]^{-1}$, where we have used $\rho_J \ge \eta^2/M^2$.

In the opposite case $12 a_K |\alpha_K| \ge 4\eta$, we estimate $q_{K,J}$ by

$$q_{K,J} \le 2e^{-m\frac{\left(\frac{\rho_J \eta}{12 a_K |\alpha_K|}\right)^2}{2\left(\rho_J + \frac{\rho_J \eta}{36 a_K |\alpha_K|}\right)}} \le 2e^{-\frac{m\rho_J \eta^2}{312 a_K^2 |\alpha_K|^2}}$$

where in the last inequaltiy we used $3 a_K |\alpha_K| \ge \eta$ to bound the second term in the denominator. Since $|\alpha_K| \le M\rho_K$, we have $a_K^2 \alpha_K^2 \le M^2(\rho_{I^-}\rho_{I^+}/\rho_I) \le M^2 \min\{\rho_{I^-}, \rho_{I^+}\}$ so that $\rho_J \ge a_K^2 \alpha_K^2 / M^2$. Therefore, we obtain

$$q_{K,J} \le e^{-cm\eta^2} \tag{4.29}$$

with $c = [312M^2]^{-1}$.

Using these estimates for p_K and $q_{K,J}$ back in (4.26), we obtain (4.9).

We next turn to (4.10), which corresponds to the opposite case where $\varepsilon_I \le \eta$ and $\varepsilon_I(\mathbf{z}) \ge 2\eta$. In this case, we remark that we have

$$\eta^2 \le \frac{\varepsilon_I^2(\mathbf{z})}{4} = \frac{\rho_{I^+}(\mathbf{z})\rho_{I^-}(\mathbf{z})}{\rho_I(\mathbf{z})} \frac{(c_{I^+}(\mathbf{z}) - c_{I^-}(\mathbf{z}))^2}{4} \le M^2 \rho_L(\mathbf{z}), \tag{4.30}$$

for $L = I^+, I^-$ and I. In this case, we do not have $\eta^2 \le M^2 \rho_L$, but we shall use the fact that $\eta^2 \le 2M^2 \rho_L$ with high probability, by writing

$$\text{Prob}\{\varepsilon_I \le \eta \ \text{ and } \ \varepsilon_I(\mathbf{z}) \ge 2\eta\} \le \sum_{K=I^+,I^1} \Big(p_K + \tilde{p}_K + \sum_{J=I^-,I^+,I} (q_{K,J} + \tilde{p}_J) \Big), \tag{4.31}$$

where now

$$p_K := \text{Prob}\{|\alpha_K - \alpha_K(\mathbf{z})| \ge [8a_K]^{-1}\eta; \ \text{given} \ \ \rho_K \ge \frac{\eta^2}{2M^2}\}, \tag{4.32}$$

and

$$q_{K,J} := \text{Prob}\{|\rho_J - \rho_J(\mathbf{z})| \ge \rho_J \min\{\frac{1}{4}, \eta[12 a_K |\alpha_K|]^{-1}\} \ \text{given} \ \ \rho_J \ge \frac{\eta^2}{2M^2}\} \tag{4.33}$$

and the additional probability is given by

$$\tilde{p}_J := \text{Prob}\{\eta^2 \le M^2 \rho_J(\mathbf{z}) \ \text{ given } \ \eta^2 \ge 2M^2 \rho_J\}. \tag{4.34}$$

Clearly, p_K and $q_{K,J}$ are estimated as in the proof of (4.9). The additional probability is estimated by

$$\begin{aligned}
\tilde{p}_J &\le \text{Prob}\{\eta^2 \ge M^2 \rho_J \ \text{ and } \ |\rho_J - \rho_J(\mathbf{z})| \ge (\eta/M)^2\} \\
&\le 2e^{-\frac{m\eta^4}{2(\rho_J M^4 + M^2 \eta/3)}} \\
&\le 2e^{-\frac{m\eta^4}{2(\eta^2 M^2 + M^2 \eta^2/3)}} \\
&\le 2e^{-cm\eta^2},
\end{aligned}$$

with $c = (8M^2/3)^{-1}$. Using these estimates in (4.31), we obtain (4.10), which concludes the proof of the lemma. □

References

[1] Baraud, Y. (2002) *Model selction for regression on a random design*, ESAIM Prob. et Stats. **6**, 127-146.

[2] Binev, P. and R. DeVore (2004) *Fast computation in adaptive tree approximation*, Numerische Math. **97**, 193–217.

[3] Birgé, L. (2004) *Model selection via testing : an alternative to (penalized) maximum likelihood estimators*, preprint, submitted to Ann. IHP.

[4] Birgé, L. and P. Massart (2001) *Gaussian model selection* J. Eur. Math. Soc. **3**, 203-268.

[5] Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984) *Classification and regression trees*, Wadsworth international, Belmont, CA.

[6] Cohen, A., W. Dahmen, I. Daubechies and R. DeVore (2001) *Tree-structured approximation and optimal encoding*, App. Comp. Harm. Anal. **11**, 192–226.

[7] Cohen, A., R. DeVore, G. Kerkyacharian and D. Picard (2001) *Maximal spaces with given rate of convergence for thresholding algorithms*, App. Comp. Harm. Anal. **11**, 167–191.

[8] Cucker, S. and S. Smale (2001) *On the mathematical foundations of learning*, Bulletin of AMS **39**, 1-49.

[9] Daubechies, I. (1992) *Ten Lectures on Wavelets*, SIAM, Philadelphia.

[10] DeVore, R. (1998) *Nonlinear approximation*, Acta Numerica **7**, 51-150.

[11] DeVore, R., G. Kerkyacharian, D. Picard and V. Temlyakov (2004) *On mathematical methods of learning*, IMI Preprint 2004:10, 1–24.

[12] DeVore, R., G. Kerkyacharian, D. Picard and V. Temlyakov (2004) *Lower bounds in learning theory*, Preprint dept. of math., U. of South Carolina.

[13] Donoho, D.L (1997) *CART and best-ortho-basis : a connection*, Ann. Stat. **25**, 1870-1911.

[14] Kerkyacharian, G. and D. Picard (2002) *Minimax or Maxisets ?*, Bernouilli **8**, no. 2, 219-253.

[15] Donoho, D.L. and I. M. Johnstone (1998) *Minimax Estimation via Wavelet shrinkage*, Annals of Statistics **26**, no. 3, 879-921.

[16] Donoho, D.L. and I.M. Johnstone (1995) *Adapting to unknown smoothness via Wavelet shrinkage*, J. Amer. Statist. Assoc. **90**, no. 432, 1200-1224.

[17] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) *Wavelet Shrinkage: Asymptopia?* , J. Royal Statistical Soc. **57**, 301-369.

[18] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) *Density estimation by wavelet thresholding*, Annals of Statistics **24**, 508-539.

[19] Gey, S. and Nedelec, E. (2001) *Model selection for CART regression trees*, preprint département de mathématiques, Université Paris XI, to appear in IEEE Transactions on Information Theory.

[20] Györfy, L., M. Kohler, A. Krzyzak, A. and H. Walk (2002) *A distribution-free theory of nonparametric regression*, Springer, Berlin.

[21] Konyagin S.V., Temlyakov V.N. (2004) *Some error estimates in learning theory*, IMI Preprints 05, 1–18.

[22] Konyagin S.V., Temlyakov V.N. (2004) *The entropy in learning theory. Error estimates*, IMI Preprints 09, 1–25.

Peter Binev, Industrial Mathematics Institute, University of South Carolina, Columbia, SC 29208, binev @math.sc.edu

Albert Cohen, Laboratoire Jacques-Louis Lions, Université Pierre et Marie Curie 175, rue du Chevaleret, 75013 Paris, France, cohen@ann.jussieu.fr

Wolfgang Dahmen, Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Templergraben 55, D-52056 Aachen Germany, dahmen@igpm.rwth-aachen,de

Ronald DeVore, Industrial Mathematics Institute, University of South Carolina, Columbia, SC 29208, devore@math.sc.edu

Vladimir Temlyakov, Industrial Mathematics Institute, University of South Carolina, Columbia, SC 29208, temlyak@math.sc.edu