

# Universal algorithms for learning theory

## Part I : piecewise constant functions \*

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

September, 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 typically 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.

## 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  $\rho$  on a product space  $Z := X \times Y$ . It will be assumed here that  $X$  is a bounded domain of  $\mathbb{R}^d$  and  $Y = \mathbb{R}$ . Given the data  $\mathbf{z} = \{z_1, \dots, z_m\} \subset Z$  of  $m$  independent random observations  $z_i = (x_i, y_i)$ ,  $i = 1, \dots, m$ , identically distributed according to  $\rho$ ,

---

\*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".

we are interested in estimating the *regression function*  $f_\rho(x)$  defined as the conditional expectation of the random variable  $y$  at  $x$ :

$$f_\rho(x) := \int_Y y d\rho(y|x) \quad (1.1)$$

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  $[-M, M]$  :

$$|y| \leq M, \quad (1.2)$$

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

We denote by  $\rho_X$  the marginal probability measure on  $X$  defined by

$$\rho_X(S) := \rho(S \times Y). \quad (1.3)$$

We shall assume that  $\rho_X$  is a Borel measure on  $X$ . We have

$$d\rho(x, y) = d\rho(y|x)d\rho_X(x). \quad (1.4)$$

It is easy to check that  $f_\rho$  is the minimizer of the risk functional

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

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  $\rho_X$ . In fact one has

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

where

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

Our objective is therefore to find an *estimator*  $f_{\mathbf{z}}$  for  $f_\rho$  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

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

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  $f_\rho$ . 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

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

as the sample size  $m$  increases. Here both the expectation and the probability are taken with respect to the product measure  $\rho^m$  defined on  $Z^m$ . An estimation of the above probability will automatically give an estimate in expectation by integrating with respect to  $\eta$ . It is important to note that the measure  $\rho_X$  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 *distribution-free*. A recent survey on distribution free regression theory is provided in the book [18], which includes most existing approaches as well as the analysis of their rate of convergence in the expectation sense. Significantly less appears to be known about estimates in probability.

Estimates for the decay of the quantities in (1.9) are usually obtained under certain assumptions (called *priors*) on  $f_\rho$ . Priors on  $f_\rho$  are typically expressed by a condition of the type  $f_\rho \in \Theta$  where  $\Theta$  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  $\Theta$  is a compact subset of  $L_2(X, \rho_X)$ .

There are three common ways to measure the compactness of a set  $\Theta$ : (i) minimal coverings, (ii) smoothness conditions on the elements of  $\Theta$ , (iii) the rate of approximation of the elements of  $\Theta$  by a specific approximation process. In the learning problem, each of these approaches has to deal with the fact that  $\rho_X$  is unknown.

To describe approach (i), for a given Banach space  $\mathcal{B}$  which contains  $\Theta$ , we define the entropy number  $\varepsilon_n(\Theta, \mathcal{B})$ ,  $n = 1, 2, \dots$ , as the minimal  $\varepsilon$  such that  $\Theta$  can be covered by at most  $2^n$  balls of radius  $\varepsilon$  in  $\mathcal{B}$ . The set  $\Theta$  is compact in  $L_2(X, \rho_X)$  if and only if  $\varepsilon_n(\Theta, L_2(X, \rho_X))$  tends to zero as  $n \rightarrow \infty$ . One can therefore quantify the level of compactness of  $\Theta$  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 entropy numbers satisfy

$$\varepsilon_n(\Theta, \mathcal{B}) \leq Cn^{-r}, \quad n = 1, 2, \dots \quad (1.10)$$

for some  $r > 0$ .

Coverings and entropy numbers have a long history in statistics for deriving optimal bounds for the rate of decay in statistical estimation (see e.g. [5]). Several recent works [8, 12, 20] 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 [4] that for any class  $\Theta$  satisfying (1.10) with  $\mathcal{B} = L_2(X, \rho_X)$ , there is an estimator  $f_{\mathbf{z}}$  satisfying

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

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

The estimators constructed using this approach are made through  $\varepsilon$  nets and are more of theoretical interest (in giving the best possible bounds) but are not practical since  $\rho_X$  is unknown and therefore these  $\varepsilon$  nets are also unknown. One can also use  $\varepsilon$  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, 19, 20].

One way to circumvent the problem of not knowing the marginal  $\rho_X$  is to use coverings in the space  $C(X)$  of continuous functions equipped with the uniform norm  $\|\cdot\|_{C(X)}$  rather

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

Concerning (ii), it is well known that when  $\rho_X$  is the Lebesgue measure, the unit ball of the Sobolev space  $W^s(L_p)$  is a compact set of  $L_2$  if and only if  $\frac{s}{d} > \frac{1}{p} - \frac{1}{2}$ . We recall that when  $s$  is an integer,  $W^s(L_p)$  consists of all  $L_p$  functions whose distributional derivatives of order  $|\alpha| \leq s$  are also in  $L_p$ . It is a Banach space when equipped with the norm

$$\|f\|_{W^s(L_p)} := \sup_{|\alpha| \leq s} \|D^\alpha f\|_{L_p}. \quad (1.12)$$

Similar remarks hold for Sobolev spaces with non-integer  $s$ , as well as for the Besov spaces  $B_q^s(L_p)$  which offer a more refined description of the notion of  $r$ -differentiability in  $L_p$ . We refer to [10] for the precise definition of such spaces.

However, there is no general approach to defining smoothness spaces with respect to general Borel measures  $\rho_X$  which therefore precludes the direct use of classification according to (ii). One way to circumvent this is to define smoothness in  $C(X)$ , that is systematically use the spaces  $W^s(L_p)$  which compactly embed into  $C(X)$ . but then this suffers from the same deficiency of omitting many smoothness spaces that are compact subsets of  $L_2(X, \rho_X)$ .

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

$$\inf_{g \in S_n} \|f - g\| \leq C\alpha_n \quad (1.13)$$

where  $C$  is a fixed constant and  $(\alpha_n)$  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  $\Sigma_n$  in place of the linear spaces  $S_n$  (see the discussion 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 specified. We discuss this in more detail later in this paper. 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 [19] and [20].

We should mention that in classical settings, for example when  $\rho_X$  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  $\Theta$ .

A very useful method for bounding the performance of such estimators is provided by the following result in [18] (see Theorem 11.3): if  $\mathcal{H}$  is taken as 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

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \leq C \frac{N \log(m)}{m} + \inf_{g \in \mathcal{H}} \|f_\rho - g\|^2. \quad (1.14)$$

From this, one can derive specific rates of convergence in expectation by balancing both terms. For example, if  $\Theta$  is a ball of the Sobolev space  $W^r(L_\infty)$  and  $\mathcal{H}$  is taken as a space of piecewise polynomial functions of degree no larger than  $r - 1$  on uniform partitions of  $X$ , one derives

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \leq C \left( \frac{m}{\log m} \right)^{-\frac{2r}{d+2r}}. \quad (1.15)$$

This estimate is optimal for this class  $\Theta$ , up to the logarithmic factor.

The above approach can be expanded to include methods of nonlinear approximation as well. A simple example is  $n$ -term approximation by functions from a fixed basis  $\Phi = \{\phi_k\}_{k=1}^\infty$  defined on  $X$ . We fix a positive number  $a > 1$  and consider the space  $\Sigma_n$  consisting of all functions  $S$  which can be written as a linear combination of at most  $n$  of the basis function  $\phi_k$  with  $1 \leq k \leq n^a$ . Another way to describe the elements of  $\Sigma_n$  is as elements of the  $\binom{n^a}{n}$  subspaces  $L_n$  obtained as the span of  $n$  elements chosen from  $\{\phi_k\}_{k=1}^{n^a}$ . This type of nonlinear approximation is discussed in more detail in [12]. In this setting one easily generalizes (1.15) to approximation from  $\Sigma_n$ .

The deficiency of the above approach is that it usually chooses the hypothesis classes in advance and typically assumes knowledge of the prior for this choice. This deficiency motivates the notion of *adaptive* or *universal* estimators. These are estimators which exhibit the optimal rate for a wide class of priors  $\Phi$ . A classical way to reach this goal is to perform model selection by adding a complexity regularization term in the empirical risk minimization process, see [2, 1, 5], Chapter 12 in [18], and [12].

We give a simple example of complexity regularization which will serve to illustrate this technique. Let  $(L_n)$  be a sequence of linear spaces with  $L_n$  of dimension  $n$  which are to be used as hypothesis classes. Given a data set  $\mathbf{z}$ , we would like to choose one of these spaces from which to approximate  $f_\rho$ , i.e. we wish to choose the dimension  $n(\mathbf{z})$ . One such choice is

$$n := n(\mathbf{z}) := \arg \min_{1 \leq j \leq n} (\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z}, L_j}) + \frac{A j \ln m}{m}) \quad (1.16)$$

where  $A > 1$  is a fixed constant. Here,  $f_{\mathbf{z}, L_j}$  is the function from  $L_j$  which minimizes the empirical error. With this choice of  $n(\mathbf{z})$ , the function

$$f_{\mathbf{z}} := T_M(f_{\mathbf{z}, L_{n(\mathbf{z})}}) \quad (1.17)$$

is defined to be the estimator to  $f_\rho$  based on the data  $\mathbf{z}$ . Here,  $T_M$  is the truncation operator  $T_M(g)(x) := \min(|g(x)|, M) \text{sign} g(x)$  which is used in treating the variance.

It is proven in [18] (see Theorem 12.1) that the estimator (1.17) gives the optimal decay in expectation (see (1.15)), save for a logarithmic factor, universally for all the

classes defined by approximation as in (1.13) with  $\alpha_n = n^{-r}$ ,  $r > 0$ . In particular, taking for each  $L_n$ , a space of piecewise polynomials of degree  $k$  on a uniform partition into cubes gives an estimator which simultaneously obtains the optimal rate (1.15) for all finite balls in each of the class  $W^s(L_\infty)$ ,  $0 < s \leq k$  by the selection of an appropriate uniform partition.

The same approach can be used to define estimators based on nonlinear approximation which again give optimal (save for logarithms) rates of convergence on more general classes of functions.

Fixing this deficiency means that in the special case where the marginal  $\rho_X$  is Lebesgue measure, the estimator would necessarily have to be optimal for all Sobolev and Besov classes which compactly embed into  $L_2(X, \rho_X)$ . These spaces correspond to smoothness spaces of order  $r$  in  $L_p$  whenever  $r > \frac{d}{p} - \frac{d}{2}$  (see [10]). This can be achieved by introducing spatially adaptive partitions. The selection of an appropriate adaptive partition in the complexity regularization framework can be implemented by the CART algorithm [6], which limits the search to a set of admissible partitions based on a tree structured splitting rule.

A practical limitation of the above described complexity regularization approach is that it is not generally 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 minimization problem needs to be globally re-solved when adding a new sample.

In two slightly different contexts, namely density estimation and denoising on a fixed design, estimation procedures based on *wavelet thresholding* have been proposed as a natural alternative to model selection by complexity regularization [14, 15, 16, 17]. These procedures are particularly attractive since they combine optimal convergence rates for the largest possible array of unknown priors together with simple and fast algorithms which are on-line implementable. In the learning theory context, the wavelet thresholding has also been used in [11] for estimation of a modification of the regression function  $f_\rho$ , namely, for estimating  $(d\rho_X/dx)f_\rho$ , where  $\rho_X$  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  $\rho_X$  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  $\rho_X$  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 adaptive partitions  $\Lambda$ . Our partitions have the same tree structure as those used in the CART algorithm [6], yet the selection of the appropriate partition is operated quite differently since it is not based on an optimization problem which would have to be re-solved when a new sample is added: instead 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 connection between CART and thresholding in one or several orthonormal bases is well understood in the fixed design denoising context

[13], this connection is not clear to us in our present context. As it will be demonstrated, our estimation schemes enjoy 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 convergence rates are optimal in probability and expectation (up to logarithmic factors) 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  $f_\rho$ .

The present choice of piecewise constant functions limits the optimal convergence rate to classes of low or no pointwise regularity. While the practical 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. This is so far a weakness of our approach from the theoretical perspective, compared to the complexity regularization approach for which optimal convergence results could be obtained in the piecewise polynomial context, using for instance Theorem 12.1 in [18].

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

A typical way of generating partitions  $\Lambda$  of  $X$  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  $I = \mathcal{P}(J)$ . The cubes in  $\mathcal{D}_j(X)$  form a uniform partition in which every cube has the same measure  $2^{-jd}$ .

More general adaptive partitions are defined as follow. A *proper* subtree  $\tilde{\mathcal{T}}$  of  $\mathcal{T}$  is a collection of nodes of  $\mathcal{T}$  with the properties: (i) the root node  $I = X$  is in  $\tilde{\mathcal{T}}$ , (ii) if  $I \neq X$  is in  $\tilde{\mathcal{T}}$  then its parent  $\mathcal{P}(I)$  is also in  $\tilde{\mathcal{T}}$ . Any finite proper subtree  $\tilde{\mathcal{T}}$  is associated to a unique partition  $\Lambda = \Lambda(\tilde{\mathcal{T}})$  which consists of its *outer leaves*, by which we mean those  $J \in \mathcal{T}$  such that  $J \notin \tilde{\mathcal{T}}$  but  $\mathcal{P}(J)$  is in  $\tilde{\mathcal{T}}$ . One way of generating adaptive 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.

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. Such a refinement strategy also results in a tree  $\mathcal{T}$  (called the *master tree*) and children, parents, proper trees 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  $< j$  and we denote by  $\Lambda_j$  the partition associated to  $\mathcal{T}_j$ , which coincides with  $\mathcal{D}_j(X)$  in the above described dyadic partition case. Note that in contrast to this case, the  $a$  children may not be similar in which case the partitions  $\Lambda_j$  are not spatially uniform (we could also work with even more generality and allow the number of children to depend on the cell to be refined, while remaining globally bounded by some fixed  $a$ ). It is important to note that the cardinalities of a proper tree  $\tilde{\mathcal{T}}$  and of its associated partition  $\Lambda(\tilde{\mathcal{T}})$  are equivalent. In fact one easily checks that

$$\#(\Lambda(\tilde{\mathcal{T}})) = (a - 1)\#\tilde{\mathcal{T}} + 1, \quad (2.1)$$

by remarking that each time a new node gets refined in the process of building an adaptive partition,  $\#\tilde{\mathcal{T}}$  is incremented by 1 and  $\#\Lambda$  by  $a - 1$ .

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

$$S = \sum_{I \in \Lambda} a_I \chi_I, \quad (2.2)$$

where for  $G \subset X$  we denote by  $\chi_G$  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  $\mathcal{S}_\Lambda$ . The best approximation to  $f$  in this space is given by

$$P_\Lambda f := \sum_{I \in \Lambda} c_I \chi_I \quad (2.3)$$

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

$$c_I := \frac{\alpha_I}{\rho_I}, \quad \text{with } \alpha_I := \int_I f d\rho_X \quad \text{and } \rho_I := \rho_X(I). \quad (2.4)$$

In the case where  $\rho_I = 0$ , both  $f_\rho$  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

$$E_n(f) := \|f - P_{\Lambda_n} f\|, \quad n = 0, 1, \dots \quad (2.5)$$



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

$$E_n(f) \leq M_0 a^{-ns}, \quad n = 0, 1, \dots \quad (2.6)$$

Notice that  $\#(\Lambda_n) = a^n$  so that the decay in (2.6) is like  $N^{-s}$  with  $N$  the number of elements in the partition. The smallest  $M_0$  for which (2.6) 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  $\rho_X$ .

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

$$\|f(\cdot + h) - f(\cdot)\|_{L_2} \leq M_0 |h|^s, \quad x, h \in X. \quad (2.7)$$

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 [7] 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

$$\varepsilon_I(f)^2 := \sum_{J \in \mathcal{C}(I)} \frac{\left(\int_J f d\rho_X\right)^2}{\rho_J} - \frac{\left(\int_I f d\rho_X\right)^2}{\rho_I}, \quad (2.8)$$

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

$$\varepsilon_I(f)^2 = \|f - c_I\|_{L_2(I, \rho_X)}^2 - \sum_{J \in \mathcal{C}(I)} \|f - c_J\|_{L_2(J, \rho_X)}^2. \quad (2.9)$$

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  $\varepsilon_I(f) \geq \eta$ . This tree can also be described as the set of all  $J \in \mathcal{T}$  such that there exists  $I \subset J$  such that  $\varepsilon_I(f) \geq \eta$ . Note that since  $f \in L^2(X, \rho_X)$  the set of nodes such that  $\varepsilon_I(f) \geq \eta$  is always finite and so is  $\mathcal{T}(f, \eta)$ . 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  $\mathcal{T}(f, \eta)$ .

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

$$|f|_{\mathcal{B}^s}^p := \sup_{\eta > 0} \eta^p \#(\mathcal{T}(f, \eta)), \quad \text{where } p := (s + 1/2)^{-1} \quad (2.10)$$

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

$$\|f - P_{\Lambda(f, \eta)}\| \leq C_s |f|_{\mathcal{B}^s}^{\frac{1}{2s+1}} \eta^{\frac{2s}{2s+1}} \leq C_s |f|_{\mathcal{B}^s} N^{-s}, \quad N := \#(\mathcal{T}(f, \eta)), \quad (2.11)$$

where the constant  $C_s$  depends only on  $s$ . For the proof of this fact we refer the reader to [7] 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  $\Lambda$  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  $B_q^s(L_\tau)$  with  $\tau > (s/d + 1/2)^{-1}$  and  $0 < q \leq \infty$  arbitrary, is contained in  $\mathcal{B}^{s/d}$  (see [7]). This should be compared with the  $\mathcal{A}^s$  where we know that  $\mathcal{A}^{s/d} = B_\infty^s(L_2)$  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 (2.11), it provides optimal convergence rates for the class  $\mathcal{B}^s$ . There is actually a slightly better strategy described in [3] 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 estimating  $f_\rho$  from the given data. We shall use the functions in  $\mathcal{S}_\Lambda$  for this purpose. Let us first observe that

$$P_\Lambda f_\rho = \operatorname{argmin}_{f \in \mathcal{S}_\Lambda} \mathcal{E}(f) = \operatorname{argmin}_{f \in \mathcal{S}_\Lambda} \int_Z (y - f(x))^2 d\rho. \quad (2.12)$$

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

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|^2 \quad (2.13)$$

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  $f_\rho$  on  $\mathcal{S}_\Lambda$  as the empirical counterpart of  $P_\Lambda f_\rho$  obtained as the solution of the least-squares problem

$$f_{\mathbf{z},\Lambda} := \operatorname{argmin}_{f \in \mathcal{S}_\Lambda} \mathcal{E}_{\mathbf{z}}(f) = \operatorname{argmin}_{f \in \mathcal{S}_\Lambda} \frac{1}{m} \sum_{i=1}^m (y_i - f(x_i))^2. \quad (2.14)$$

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

$$\|y\|_{L_2(X, \delta_x)}^2 := \frac{1}{m} \sum_{i=1}^m |y(x_i)|^2, \quad (2.15)$$

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

$$\min_{c \in \mathbb{R}} \frac{1}{m} \sum_{i=1}^m (y_i - c)^2 \chi_I(x_i), \quad (2.16)$$

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

$$c_I(\mathbf{z}) = \frac{\alpha_I(\mathbf{z})}{\rho_I(\mathbf{z})}, \quad \text{where } \alpha_I(\mathbf{z}) := \frac{1}{m} \sum_{i=1}^m y_i \chi_I(x_i), \quad \rho_I(\mathbf{z}) := \frac{1}{m} \sum_{i=1}^m \chi_I(x_i). \quad (2.17)$$

Thus, we can rewrite the estimator as

$$f_{\mathbf{z},\Lambda} = \sum_{I \in \Lambda} c_I(\mathbf{z}) \chi_I. \quad (2.18)$$

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  $\mathcal{S}_\Lambda$ ,

$$\|f_\rho - f_{\mathbf{z},\Lambda}\|^2 = \|f_\rho - P_\Lambda f_\rho\|^2 + \|P_\Lambda f_\rho - f_{\mathbf{z},\Lambda}\|^2 =: e_1 + e_2. \quad (2.19)$$

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

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

$$\operatorname{Prob} \{ \|P_\Lambda f_\rho - f_{\mathbf{z},\Lambda}\| > \eta \} \leq 4N e^{-c \frac{m\eta^2}{N}}, \quad (2.20)$$

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  $\Lambda$  is any partition, the mean square error is bounded by*

$$E\left(\|P_\Lambda f_\rho - f_{\mathbf{z},\Lambda}\|^2\right) \leq C \frac{N \log N}{m}, \quad (2.21)$$

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  $f_{\mathbf{z}} := f_{\mathbf{z}, \Lambda_j}$  with  $j$  chosen as the smallest integer such that  $a^{j(1+2s)} \geq \frac{m}{\log m}$ . Then, given any  $\beta > 0$ , there is a constant  $\tilde{c} = \tilde{c}(M, \beta, a)$  such that*

$$\text{Prob} \left\{ \|f_\rho - f_{\mathbf{z}}\| > (\tilde{c} + |f_\rho|_{\mathcal{A}^s}) \left( \frac{\log m}{m} \right)^{\frac{s}{2s+1}} \right\} \leq C m^{-\beta}, \quad (2.22)$$

and

$$E \left( \|f_\rho - f_{\mathbf{z}}\|^2 \right) \leq (C + |f_\rho|_{\mathcal{A}^s}^2) \left( \frac{\log m}{m} \right)^{\frac{2s}{2s+1}}. \quad (2.23)$$

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 (2.23) in Theorem 2.3 can also be obtained as a consequence of Theorem 11.3 in [18] quoted in our introduction. In order to prepare for the subsequent developments direct proofs of these results are given later in §3.*

Theorem 2.3 is satisfactory in the sense that it is obtained under no assumption on the measure  $\rho_X$  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  $L_\infty$ . 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 [14, 15, 16, 17] 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  $\rho_X$  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  $f_\rho$  but rather will learn the smoothness from the data. Thus, it will automatically choose the right size for the partition  $\Lambda$ .

Our starting point is the adaptive procedure introduced in §2.1 applied to the function  $f_\rho$ . We use the notation  $\varepsilon_I := \varepsilon_I(f_\rho)$  in this case. Then, by (2.8),

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

The selection of the partition  $\Lambda$  in our learning scheme will be based on the empirical coefficients

$$\varepsilon_I^2(\mathbf{z}) := \sum_{J \in \mathcal{C}(I)} \frac{\alpha_J^2(\mathbf{z})}{\rho_J(\mathbf{z})} - \frac{\alpha_I^2(\mathbf{z})}{\rho_I(\mathbf{z})}. \quad (2.25)$$

We define the threshold

$$\tau_m := \kappa \sqrt{\frac{\log m}{m}}, \quad (2.26)$$

where the constant  $\kappa$  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 proper tree  $\mathcal{T}(\mathbf{z}, m)$  which contains the set

$$\Sigma(\mathbf{z}, m) := \{I \in \mathcal{T}_{j_0} ; \varepsilon_I(\mathbf{z}) \geq \tau_m\}. \quad (2.27)$$

This tree can also be described as the set of all  $J \in \mathcal{T}_{j_0}$  such that there exists  $I \subset J$  such that  $I \in \Sigma(\mathbf{z}, m)$ . 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 the nodes  $I$  of generation  $j < j_0$ .
- (ii) Threshold these quantities at level  $\tau_m$  to obtain the set  $\Sigma(\mathbf{z}, m)$ .
- (iii) Complete  $\Sigma(\mathbf{z}, m)$  to  $\mathcal{T}(\mathbf{z}, m)$  by adding the nodes  $J$  which contain an  $I \in \Sigma(\mathbf{z}, m)$ .
- (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  $\kappa_0 = \kappa_0(\beta, \gamma, M)$  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*

$$\text{Prob} \left\{ \|f_\rho - f_{\mathbf{z}}\| \geq \tilde{c} \left( \frac{\log m}{m} \right)^{\frac{s}{2s+1}} \right\} \leq C m^{-\beta}, \quad (2.28)$$

as well as the following expectation bound

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \leq C \left( \frac{\log m}{m} \right)^{\frac{2s}{2s+1}}, \quad (2.29)$$

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  $\left(\frac{\log m}{m}\right)^{\frac{s}{2s+1}}$  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  $\gamma$  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 [3]. 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

$$\varepsilon_I(\mathbf{z})^2 \leq \mathcal{E}_I(\mathbf{z}) \quad (2.30)$$

with  $\mathcal{E}_I(\mathbf{z}) := \|y - c_I(\mathbf{z})\|_{L_2(\delta_{\mathcal{X}, I})}^2$  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:

$$J \subset I \Rightarrow \mathcal{E}_J(\mathbf{z}) \leq \mathcal{E}_I(\mathbf{z}). \quad (2.31)$$

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 descendants of an element  $I$  for which  $\mathcal{E}_I(\mathbf{z})$  is less than  $\tau_m$ .

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  $\rho_I$  and  $\alpha_I$  are

$$\rho_I(\mathbf{z}^+) = \frac{m}{m+1}(\rho_I(\mathbf{z}) + \chi_I(x_{m+1})) \quad (2.32)$$

and

$$\alpha_I(\mathbf{z}^+) = \frac{m}{m+1}(\alpha_I(\mathbf{z}) + y_{m+1}\chi_I(x_{m+1})). \quad (2.33)$$

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  $\tau_m$  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  $\Lambda$  be any partition. By (2.3) and (2.18), we can write

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

According to their definitions (2.4), (2.17), both  $c_I$  and  $c_I(\mathbf{z})$  are bounded in modulus by  $M$ . Therefore, given  $\eta > 0$ , if we define

$$\Lambda^- := \{I \in \Lambda : \rho_I \leq \frac{\eta^2}{8NM^2}\}, \quad (3.2)$$

we clearly have

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

We next consider the complement set  $\Lambda^+ = \Lambda \setminus \Lambda^-$ . In order to prove (2.20), it now suffices to establish that for all  $I \in \Lambda^+$

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

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

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

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

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

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

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

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  $\zeta_i$  of a random variable  $\zeta$  such that  $|\zeta(z) - E(\zeta)| \leq M_0$  and  $\text{Var}(\zeta) = \sigma^2$ , one has for any  $\varepsilon > 0$

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

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

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

with  $c = [\frac{256}{3}M^2]^{-1}$ , where we have used in the second inequality 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

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

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  $c = [\frac{256}{3}M^2]^{-1}$ , 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  $M \rightarrow +\infty$ . 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 (2.20) can be obtained by a direct use of the concentration property of the gaussian.*

The proof of Corollary 2.2 follows by integration of (2.20) over  $\eta$ :

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

where  $\eta_0$  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 (2.21).

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

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



Hence, using Theorem 2.1, we see that the probability on the left of (2.22) is bounded from above by

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

which does not exceed  $Cm^{-\beta}$  provided  $\tilde{c}^2 c > a^2(1 + \beta)$ . The proof of (2.23) 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  $\eta$ .  $\Lambda(f_\rho, \eta)$  is the partition induced by the outer leaves of  $\mathcal{T}(f_\rho, \eta)$ . We use  $\tau_m$  as defined in (2.26) and  $j_0 = j_0(m)$  is the largest integer such that  $a^{j_0} \leq \tau_m^{-1/\gamma}$ . For any partition  $\Lambda$  we write  $f_{\mathbf{z}, \Lambda} = \sum_{I \in \Lambda} c_I(\mathbf{z}) \chi_I$ .

If  $\Lambda_0$  and  $\Lambda_1$  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

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

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

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

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  $b := 2\sqrt{a-1} > 1$ . This type of splitting is classically used in the analysis wavelet thresholding procedures, in order to deal with the fact that the partition built from those  $I$  such that  $\varepsilon_I(\mathbf{z}) \geq \tau_m$  does not exactly coincides with the partition which would be chosen by an oracle based on those  $I$  such that  $\varepsilon_I \geq \tau_m$ . This is accounted by the terms  $e_2$  and  $e_4$  which correspond to those  $I$  such that  $\varepsilon_I(\mathbf{z})$  is significantly larger or smaller than  $\varepsilon_I$  respectively, and which will be proved to be small in probability. The remaining terms  $e_1$  and  $e_3$  respectively correspond to the bias and variance of oracle estimators based on partitions obtained by thresholding the unknown coefficients  $\varepsilon_I$ .

The first term  $e_1$  is therefore 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

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

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

The third term  $e_3$  is treated by the estimate (2.20) of Theorem 2.1:

$$\text{Prob}\{e_3 > \eta\} \leq 4Ne^{-c\frac{m\eta^2}{N}}, \quad (4.4)$$

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 (2.10) that

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

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

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

whenever  $\kappa \geq \kappa_0$  with  $\kappa_0$  depending on  $\beta$ ,  $\gamma$ , and  $M$  and with  $C$  depending only on  $a$ .

Before proving this result, let us show that the combination (4.3), (4.4), (4.5) and (4.6) imply the validity of the estimates (2.28) and (2.29) in Theorem 2.5. We fix the value of  $\beta$  and we fix any constant  $\kappa$  for which (4.6) holds. Let  $\eta_1 := \tilde{c}(\frac{\log m}{m})^{\frac{s}{2s+1}}$  with  $\tilde{c}$  from (2.28) 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 (2.28) 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

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

$$\begin{aligned}
&= C \left( \frac{m}{\log m} \right)^{\frac{1}{2s+1}} e^{-cD^2 m \left( \frac{\log m}{m} \right)} \\
&= C \left( \frac{m}{\log m} \right)^{\frac{1}{2s+1}} m^{-cD^2} \\
&\leq Cm^{1-cD^2}
\end{aligned}$$

where  $D^2 := \frac{(\tilde{c}-c_0)^2}{\kappa^{2s+1} b^p |f|_{\mathcal{B}^s}^p}$ . The concentration estimate (2.28) follows now by taking  $\tilde{c}$  large enough so that  $1 - cD^2 + \beta \leq 0$ .

For the expectation estimate (2.29), we recall that according to Corollary 2.2, we have

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

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

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

by choosing  $\beta$  larger than  $2s/(2s+1)$ , for example  $\beta = 1$ . The same holds for  $e_4$  and therefore we obtain (2.29).

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  $\varepsilon_I$  with respect to the threshold. This is expressed by the following lemma.

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

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

and

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

where the constant  $c$  depends only on  $M$  and the constant  $C$  depends only on  $a$ .

Before proving Lemma 4.1, let us show how this result 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

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

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

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

We can treat the second sum in (4.11) the same way and obtain the same bound as the one for  $e_4$  below. By similar considerations, we obtain

$$\text{Prob}\{e_4 > 0\} \leq \sum_{I \in \mathcal{T}_{j_0}} \text{Prob}\{\varepsilon_I(\mathbf{z}) \geq \tau_m \text{ and } \varepsilon_I \leq \tau_m/b\}, \quad (4.12)$$

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  $c\kappa_0^2 = b^2(\beta + 1/\gamma)$ .

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  $\{J_1, \dots, J_a\}$  can be decomposed into  $a - 1$  steps consisting of splitting an element into a pair of elements: defining  $I_n := I \setminus (J_1 \cup \dots \cup J_n)$  we start from  $I = I_0$  and refine iteratively  $I_{n-1}$  into the two elements  $I_n$  and  $J_n$ , for  $n = 1, \dots, a - 1$ . By orthogonality, we can write

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

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

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

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

$$\text{Prob}\{\varepsilon_I \leq \eta \text{ and } \varepsilon_I(\mathbf{z}) \geq b\eta\} \leq \sum_{n=0}^{a-2} \text{Prob}\{\varepsilon_{I_n} \leq \eta \text{ and } \varepsilon_{I_n}(\mathbf{z}) \geq 2\eta\}, \quad (4.15)$$

and similarly

$$\text{Prob}\{\varepsilon_I(\mathbf{z}) \leq \eta \text{ and } \varepsilon_I \geq b\eta\} \leq \sum_{n=0}^{a-2} \text{Prob}\{\varepsilon_{I_n}(\mathbf{z}) \leq \eta \text{ and } \varepsilon_{I_n} \geq 2\eta\}, \quad (4.16)$$

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  $\varepsilon_I$  as follows

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

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

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

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

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

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  $\beta_I$  and  $\beta_I(\mathbf{z})$  as

$$\beta_I = a_{I^+}\alpha_{I^+} - a_{I^-}\alpha_{I^-} \quad (4.19)$$

and

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

It follows that

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

We next introduce the numbers  $\delta_J$  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

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

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

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

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

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

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

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

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

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

with

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

and

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

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

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

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

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

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

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

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

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

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

$$q_{K,J} \leq e^{-cm\eta^2} \quad (4.29)$$

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  $\varepsilon_I \leq \eta$  and  $\varepsilon_I(\mathbf{z}) \geq 2\eta$ . In this case, we remark that we have

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

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

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

where now

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

and

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

and the additional probability is given by

$$\tilde{p}_J := \text{Prob}\{\eta^2 \leq M^2\rho_J(\mathbf{z}) \text{ given } \eta^2 \geq 2M^2\rho_J\}. \quad (4.34)$$

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

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

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

## 5 Universal consistency of the estimator

In this last section, we discuss the consistency of our estimator when no smoothness assumption is made on the regression function  $f_\rho \in L2(X, \rho_X)$ . Of course it is still assumed that  $|y| \leq M$  almost surely, so that  $|f_\rho| \leq M$ .

We are therefore interested in checking the convergence property

$$\lim_{m \rightarrow +\infty} E(\|f_\rho - f_{\mathbf{z},m}\|^2) = 0, \quad (5.1)$$

which in turns implies the convergence in probability: for all  $\varepsilon > 0$ ,

$$\lim_{m \rightarrow +\infty} \text{Prob}\{\|f_\rho - f_{\mathbf{z},m}\| > \varepsilon\} = 0. \quad (5.2)$$

For this purpose, we use the same estimation of the error by  $e_1 + e_2 + e_3 + e_4$  as in the proof of Theorem 2.5.

We first remark that the proof of the estimate

$$E(e_2) + E(e_4) \leq Cm^{-\beta}, \quad (5.3)$$

remains unchanged under no smoothness assumption made on  $f_\rho$ .

Concerning the approximation term  $e_1$ , we have seen that

$$e_1 \leq \|f_\rho - P_{\Lambda(f_\rho, b\tau_m)} f_\rho\| + \|f_\rho - P_{\Lambda_{j_0}} f_\rho\|. \quad (5.4)$$

Under no smoothness assumptions, the convergence to 0 of these two terms still occurs when  $j_0 \rightarrow +\infty$  and  $\tau_m \rightarrow 0$ , that is as  $m \rightarrow +\infty$ . This requires however that the union of the spaces  $(S_{\Lambda_j})_{j \geq 0}$  is dense in  $L2(X, \rho_X)$ . This is ensured by imposing natural restrictions on the splitting procedure generating the partitions which should be such that

$$\lim_{j \rightarrow +\infty} \sup_{I \in \Lambda_j} |I| = 0, \quad (5.5)$$

where  $|I|$  is the Lebesgue measure of  $I$ . This is obviously true for dyadic partitions, and more generally when the splitting rule is such that

$$\sup_{J \in \mathcal{C}(I)} |J| \leq \nu |I|, \quad (5.6)$$

with  $\nu < 1$  independent of  $I \in \mathcal{T}$ . Under this restriction, classical results of measure theory indicates that  $P_{\Lambda_j} f$  converges to  $f$  in  $L2(X, \rho_X)$  as  $j \rightarrow +\infty$  for all  $f \in L2(\rho_X)$ .

We are therefore ensured that  $\|f_\rho - P_{\Lambda_{j_0}} f_\rho\|$  tends to 0 as  $m \rightarrow +\infty$ . For the first term  $\|f_\rho - P_{\Lambda(f_\rho, b\tau_m)} f_\rho\|$ , we remark that the convergence of  $P_{\Lambda_j} f$  to  $f$  also implies that  $f$  can be written as the sum of an  $L2(X, \rho_X)$ -orthogonal series

$$f = p_0 \chi_X + \sum_{I \in \mathcal{T}} \psi_I, \quad \text{with } \psi_I := \sup_{J \in \mathcal{C}(I)} p_J \chi_J - p_I \chi_I, \quad (5.7)$$

It follows that for  $\eta > 0$

$$\|f - P_{\Lambda(f, \eta)} f\|^2 = \sum_{I \notin \mathcal{T}(f, \eta)} \varepsilon_I(f)^2 \leq \sum_{\varepsilon_I(f) \leq \eta} \varepsilon_I(f)^2. \quad (5.8)$$

Since by Parseval inequality,

$$\sum_{I \in \mathcal{T}} \varepsilon_I(f)^2 = \|f\|^2 - \|p_0 \chi_X\|^2 < +\infty, \quad (5.9)$$

it follows that  $\|f - P_{\Lambda(f, \eta)} f\|$  tends to 0 as  $\eta \rightarrow 0$ . Therefore  $\|f_\rho - P_{\Lambda(f_\rho, b\tau_m)} f_\rho\|$  tends to 0 as  $m \rightarrow +\infty$ .



It remains to study the variance term  $e_3$  for which we have established

$$E(e_3^2) \leq C \frac{N \log N}{m}, \quad (5.10)$$

with

$$N = \#(\Lambda(\tau_m, \mathbf{z}) \wedge \Lambda(\tau_m/b)) \leq \#(\Lambda(\tau_m/b)). \quad (5.11)$$

Note that since  $(\varepsilon_I)_{I \in \mathcal{T}}$  is a square summable sequence according to (5.9), we have

$$\#\{I \in \mathcal{T} ; \varepsilon_I > \eta\} \leq C\eta^{-2}\varphi(\eta), \quad (5.12)$$

where  $\varphi(\eta) \rightarrow 0$  as  $\eta \rightarrow 0$ . Therefore if  $\#(\Lambda(\tau_m/b))$  was controlled by  $\#\{I \in \mathcal{T} ; \varepsilon_I > \tau_m/b\}$ , we would derive

$$E(e_3^2) \leq C \frac{\tau_m^{-2} \log(\tau_m^{-2})}{m} \varphi(\tau_m) \leq C\varphi(\tau_m), \quad (5.13)$$

and therefore  $E(e_3^2)$  would tend to 0 which terminates the proof of the consistency of the estimator. However,  $\#(\Lambda(\tau_m/b))$  can be significantly larger due to the process of completing the set of thresholded coefficients into a proper tree. We still have the estimate

$$\#(\Lambda(\tau_m/b) \leq j_0 \#\{I \in \mathcal{T} ; \varepsilon_I > \tau_m/b\} \leq C \log m \tau_m^{-2} \varphi(\tau_m), \quad (5.14)$$

where  $C$  depends on  $a$  and  $\gamma$ . It follows that  $E(e_3^2)$  will always go to 0 if the threshold  $\tau_m$  is modified to

$$\tau_m := \frac{\log m}{\sqrt{m}}, \quad (5.15)$$

which amounts to an additional  $\sqrt{\log m}$  factor. It is easily checked that this modification does not affect the other estimates for  $e_1$ ,  $e_2$  and  $e_4$ . However it induces an additional  $\sqrt{\log m}$  factor in the rate of convergence expressed by Theorem 2.5.

An alternate way of ensuring the convergence to zero of  $E(e_3^2)$  is by imposing that  $\gamma > 1/2$ , since we can then write

$$\#(\Lambda(\tau_m/b) \leq \#(\Lambda_{j_0} = a^{j_0} \leq C\tau_m^{-1/\gamma}), \quad (5.16)$$

with  $1/\gamma > 2$ , which readily implies that  $N \log N/m$  tends to 0. However this is a stronger restriction since the optimal convergence rate of the algorithm is maintained only for regression functions which are at least in the uniform approximation space  $\mathcal{A}^{1/2}$ .

## References

- [1] Baraud, Y. (2002) *Model selection for regression on a random design*, ESAIM Prob. et Stats. **6**, 127–146.
- [2] Barron, A. R. (1991) *Complexity regularization with application to artificial neural network*, in *Nonparametric functional estimation and related topics*, G. Roussas (ed.), 561–576, Kluwer Academic Publishers.

- [3] Binev, P. and R. DeVore (2004) *Fast computation in adaptive tree approximation*, Numerische Math. **97**, 193–217.
- [4] Birgé, L. (2004) *Model selection via testing : an alternative to (penalized) maximum likelihood estimators*, preprint, to appear in Ann. IHP.
- [5] Birgé, L. and P. Massart (2001) *Gaussian model selection* J. Eur. Math. Soc. **3**, 203–268.
- [6] Breiman, L., J.H. Friedman, R.A. Olshen and C.J. Stone (1984) *Classification and regression trees*, Wadsworth international, Belmont, CA.
- [7] Cohen, A., W. Dahmen, I. Daubechies and R. DeVore (2001) *Tree-structured approximation and optimal encoding*, App. Comp. Harm. Anal. **11**, 192–226.
- [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] Donoho, D.L. and I. M. Johnstone (1998) *Minimax Estimation via Wavelet shrinkage*, Annals of Statistics **26**, no. 3, 879–921.
- [15] Donoho, D.L. and I.M. Johnstone (1995) *Adapting to unknown smoothness via Wavelet shrinkage*, J. Amer. Statist. Assoc. **90**, no. 432, 1200–1224.
- [16] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) *Wavelet Shrinkage: Asymptopia?* , J. Royal Statistical Soc. **57**, 301–369.
- [17] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) *Density estimation by wavelet thresholding*, Annals of Statistics **24**, 508–539.
- [18] Györfy, L., M. Kohler, A. Krzyzak, A. and H. Walk (2002) *A distribution-free theory of nonparametric regression*, Springer, Berlin.
- [19] Konyagin S.V., Temlyakov V.N. (2004) *Some error estimates in learning theory*, IMI Preprints 05, 1–18.
- [20] 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