Journal of Machine Learning Research 6 (2005) 1297–1321 Submitted 9/04; Revised 4/05; Published 9/05

Universal Algorithms for Learning Theory
Part I : Piecewise Constant Functions

Peter Binev BINEV@MATH.SC.EDU
Industrial Mathematics Institute
Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA

Albert Cohen COHEN@ANN.JUSSIEU.FR
Laboratoire Jacques-Louis Lions
Université Pierre et Marie Curie
175, rue du Chevaleret
75013 Paris, France

Wolfgang Dahmen DAHMEN@IGPM.RWTH-AACHEN.DE
Institut für Geometrie und Praktische Mathematik
RWTH Aachen
Templergraben 55
D-52056 Aachen, Germany

Ronald DeVore DEVORE@MATH.SC.EDU
Vladimir Temlyakov TEMLYAK@MATH.SC.EDU
Industrial Mathematics Institute
Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA

Editor: Peter Bartlett

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.

Keywords: distribution-free learning theory, nonparametric regression, universal algorithms, adaptive approximation, on-line algorithms

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 domain of $\mathbb{R}^{\mathrm{d}}$ 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 ρ, we are interested in estimating the *regression function* $f_\rho(x)$ defined as the conditional expectation of the random variable y at x:

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

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 $[-M, M]$:

$$|y| \le M,$$

almost surely. It follows in particular that $|f_\rho| \le M$ almost everywhere with respect to ρ_X.

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

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

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

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

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

$$\mathcal{E}(f) := \int_Z (y - f(x))^2 d\rho, \tag{1}$$

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

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

where

$$\|\cdot\| := \|\cdot\|_{L_2(X, \rho_X)}. \tag{2}$$

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), as the minimizer of the empirical risk

$$f_{\mathbf{z}} = f_{\mathbf{z},\mathcal{H}} := \operatorname*{argmin}_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f), \quad \text{with} \quad \mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (y_i - f(x_i))^2. \tag{3}$$

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

$$\text{Prob}\{\|f_\rho - f_{\mathbf{z}}\| \geq \eta\}, \quad \eta > 0 \quad \text{or} \quad E(\|f_\rho - f_{\mathbf{z}}\|^2) \tag{4}$$

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 (4) 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 (2) 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 by Györfy et al. (2002), 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 (4), 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 Θ 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 $n \longrightarrow \infty$. 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 entropy numbers satisfy

$$\varepsilon_n(\Theta, \mathcal{B}) \leq Cn^{-r}, \;\; n = 1, 2, \cdots \tag{5}$$

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. Birgé and Massart, 2001). Several recent works (Cucker and Smale, 2001; DeVore et al., 2004b; Konyagin and Temlyakov, 2004b) 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 (Birgé, 2004) that for any class Θ satisfying (5) with $\mathcal{B} = L_2(X, \rho_X)$, there is an estimator $f_{\mathbf{z}}$ satisfying

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \leq Cm^{-\frac{2r}{2r+1}}, \quad m = 1, 2, \ldots \tag{6}$$

whenever $f_\rho \in \Theta$. Lower bounds which match (6) have been given by DeVore et al. (2004b) using a slightly different type of entropy.

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 the paper by Cucker and Smale (2001) and is carried further in several other papers (see DeVore et al., 2004b; Konyagin and Temlyakov, 2004a,b).

One way to circumvent the problem of not knowing the marginal ρ_X is to use coverings in the space $C(X)$ of continuous functions equipped with the uniform norm $\|\cdot\|_{L_\infty}$ rather than in $L_2(X,\rho_X)$, since a good covering for Θ in $C(X)$ gives bounds for the covering in $L_2(X,\rho_X)$ independently of ρ_X. In this approach one would assume that Θ satisfies (5) for $\mathcal{B} = C(X)$ and then build estimators which satisfy (6) 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 $L_2(X,\rho_X)$.

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

$$\|f\|_{W^r(L_p)} := \sup_{|\alpha|\le r} \|D^\alpha f\|_{L_p}.$$

Similar remarks hold for Sobolev spaces with non-integer r, as well as for the Besov spaces $B^r_q(L_p)$ which offer a more refined description of the notion of r-differentiability in L_p. We refer to DeVore (1998) for the precise definition of such spaces.

However, there is no general approach to defining smoothness spaces with respect to general Borel measures ρ_X which 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 $W^r(L_\infty)$, 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 (S_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

$$\inf_{g\in S_n} \|f-g\| \le C\alpha_n$$

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 S_n (see the discussion in DeVore et al., 2004b). 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 by DeVore et al. (2004b) and the follow up papers (Konyagin and Temlyakov, 2004a,b).

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 is provided by the following result (see Györfy et al., 2002, Theorem 11.3): if $\mathcal{H}$ is taken as a linear space of dimension N and if the least-square estimator (3) is post-processed by application of the truncation operator $y \mapsto T_M(y) = \text{sign}(y)\min\{|y|, M\}$, then

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

Using this, one can derive specific rates of convergence in expectation by balancing both terms. For example, if Θ 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

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \leq C(\frac{m}{\log m})^{-\frac{2r}{d+2r}}. \tag{7}$$

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 $L_2(X, \rho_X)$.

The first deficiency motivates the notion of *adaptive* or *universal* estimators: 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 by adding a complexity regularization term in the empirical risk minimization process (see Barron, 1991; Baraud, 2002; Birgé and Massart, 2001; DeVore et al., 2004b; Györfy et al., 2002, Chapter 12). In particular, one can construct one estimator which simultaneously obtains the optimal rate (7) for all finite balls in each of the class $W^r(L_\infty)$, $0 < r \leq k$ where k is arbitrary but fixed, by the selection of an appropriate uniform partition.

Fixing the second deficiency means that in the case where the marginal ρ_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 s in L_p whenever $s > \frac{d}{p} - \frac{d}{2}$ (see DeVore, 1998). 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 (Breiman et al., 1984), which limits the search within 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 (Donoho and Johnstone, 1998, 1995; Donoho et al., 1996a,b). 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 by DeVore et al. (2004a) 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 adaptive partitions Λ. Our partitions have the same tree structure as those used in the CART algorithm (Breiman et al., 1984), 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 (Donoho, 1997), 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_ρ.

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 Györfy et al., 2002, Theorem 12.1).

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. Finally, in Section 3 we give results about the consistency of our estimator.

2. The Basic Strategy and the Main Results

In this section we start in §2.1 with some basic facts about adaptive approximation. Then in we continue in §2.2 with some results about least-squares fitting on fixed partition. The universal algorithm is described in §2.3 where the main results of this paper are formulated. In §2.4 we discuss the on-line implementation of our algorithm.

2.1 Partitions and Adaptive Approximation

A typical way of generating partitions Λ 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 Λ_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 Λ_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

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

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 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

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

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

$$P_\Lambda f := \sum_{I \in \Lambda} c_I \chi_I \tag{1}$$

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

$$c_I := \frac{\alpha_I}{\rho_I}, \text{ with } \alpha_I := \int_I f d\rho_X \text{ and } \rho_I := \rho_X(I). \tag{2}$$

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

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

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

$$E_n(f) \le M_0 a^{-ns}, \quad n = 0, 1, \ldots. \tag{3}$$

Notice that $\#(\Lambda_n) = a^n$ so that the decay in (3) is like N^{-s} with N the number of elements in the partition. The smallest M_0 for which (3) 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 $\mathcal{A}^{s/d} = B^s_\infty(L_2)$, $0 < s \le 1$, with equivalent norms. Here $B^s_\infty(L_2)$ is the Besov space which can be described in terms of differences as

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

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 by Cohen et al. (2001) 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

$$\varepsilon_I(f)^2 := \sum_{J \in C(I)} \frac{\left(\int_J f d\rho_X\right)^2}{\rho_J} - \frac{\left(\int_I f d\rho_X\right)^2}{\rho_I}, \tag{4}$$

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 also accounts for the decreased projection error when I is refined. In fact, one easily verifies that

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

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

$$|f|^p_{\mathcal{B}^s} := \sup_{\eta>0} \eta^p \#(\mathcal{T}(f,\eta)), \quad \text{where } p := (s+1/2)^{-1} \tag{5}$$

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

$$\|f - P_{\Lambda(f,\eta)}\| \leq C_s |f|^{\frac{1}{2s+1}}_{\mathcal{B}^s} \eta^{\frac{2s}{2s+1}} \leq C_s |f|_{\mathcal{B}^s} N^{-s}, \quad N := \#(\mathcal{T}(f,\eta)), \tag{6}$$

where the constant C_s depends only on s. For the proof of this fact we refer the reader to the paper by Cohen et al. (2001) 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 $0 < q \leq \infty$ arbitrary, is contained in $\mathcal{B}^{s/d}$ (see Cohen et al., 2001). 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 (6), it provides optimal convergence rates for the class $\mathcal{B}^s$. There is actually a slightly better strategy described in the paper by Binev and DeVore (2004) 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_ρ from the given data. We shall use the functions in $\mathcal{S}_\Lambda$ for this purpose. Let us first observe that

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

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

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

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 (3) 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

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

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

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

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

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

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

$$c_I(\mathbf{z}) = \frac{\alpha_I(\mathbf{z})}{\rho_I(\mathbf{z})}, \text{ where } \alpha_I(\mathbf{z}) := \frac{1}{m}\sum_{i=1}^m y_i \chi_I(x_i), \quad \rho_I(\mathbf{z}) := \frac{1}{m}\sum_{i=1}^m \chi_I(x_i).x_i \in I\}. \tag{7}$$

Thus, we can rewrite the estimator as

$$f_{\mathbf{z},\Lambda} = \sum_{I\in\Lambda} c_I(\mathbf{z})\chi_I. \tag{8}$$

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,

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

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

Theorem 1 *For any partition Λ and any $\eta > 0$,*

$$\text{Prob}\left\{\|P_\Lambda f_\rho - f_{\mathbf{z},\Lambda}\| > \eta\right\} \leq 4Ne^{-c\frac{m\eta^2}{N}}, \tag{9}$$

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 (9) over $\eta > 0$.

Corollary 1 *If Λ is any partition, the mean square error is bounded by*

$$E\left(\|P_\Lambda f_\rho - f_{\mathbf{z},\Lambda}\|^2\right) \leq C\frac{N\log N}{m}, \tag{10}$$

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 1 and Corollary 1 and obtain the following result.

Theorem 2 *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*

$$\text{Prob}\left\{\|f_\rho - f_{\mathbf{z}}\| > (\tilde{c} + |f_\rho|_{\mathcal{A}^s})\left(\frac{\log m}{m}\right)^{\frac{s}{2s+1}}\right\} \leq Cm^{-\beta}, \tag{11}$$

and

$$E\left(\|f_\rho - f_{\mathbf{z}}\|^2\right) \leq (C + |f_\rho|^2_{\mathcal{A}^s})\left(\frac{\log m}{m}\right)^{\frac{2s}{2s+1}}. \tag{12}$$

where C depends only on a and M.

Remark 1 *It is also possible to prove Corollary 1 using the result by of Cucker and Smale (2001, Theorem C*). The expectation estimate (12) in Theorem 2 can also be obtained as a consequence of Theorem 11.3 by Györfy et al. (2002) quoted in our introduction. In order to prepare for the subsequent developments direct proofs of these results are given later in §3.*

Theorem 2 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 (Donoho and Johnstone, 1998, 1995; Donoho et al., 1996a,b) 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 (4),

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

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

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

We define the threshold

$$\tau_m := \kappa\sqrt{\frac{\log m}{m}}, \tag{13}$$

where the constant κ is absolute and will be fixed later in the proof of Theorem 3 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 $a^j \le \tau_m^{-1/\gamma}$. We next consider the smallest proper tree $\mathcal{T}(\mathbf{z},m)$ which contains the set

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

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 $f_{\mathbf{z}} := f_{\mathbf{z},\Lambda}$. 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 τ_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 3 *Let $\beta,\gamma > 0$ be arbitrary. Then, there exists $\kappa_0 = \kappa_0(\beta,\gamma,M)$ such that if $\kappa \ge \kappa_0$, then whenever $f_\rho \in \mathcal{A}^\gamma \cap \mathcal{B}^s$ for some $s > 0$, the following concentration estimate holds*

$$\text{Prob}\left\{\|f_\rho - f_{\mathbf{z}}\| \ge \tilde{c}\left(\frac{\log m}{m}\right)^{\frac{s}{2s+1}}\right\} \le Cm^{-\beta}, \tag{14}$$

as well as the following expectation bound

$$E(\|f_\rho - f_{\mathbf{z}}\|^2) \le C\left(\frac{\log m}{m}\right)^{\frac{2s}{2s+1}}, \tag{15}$$

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

Theorem 3 is more satisfactory than Theorem 2 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 by Binev and DeVore (2004). 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

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

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

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

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 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

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

and

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

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 1. Let Λ be any partition. By (1) and (8), we can write

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

According to their definitions (2), (7), both c_I and $c_I(\mathbf{z})$ are bounded in modulus by M. Therefore, given $\eta > 0$, if we define

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

we clearly have

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

We next consider the complement set $\Lambda^+ = \Lambda \setminus \Lambda^-$. In order to prove (9), it now suffices to establish that for all $I \in \Lambda^+$

$$\text{Prob}\left\{ |c_I(\mathbf{z}) - c_I|^2 \geq \frac{\eta^2}{2N\rho_I} \right\} \leq 4e^{-c\frac{m\eta^2}{N}}. \tag{1}$$

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

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

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

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

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

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

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 $\varepsilon > 0$

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

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

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

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 $\frac{1}{2}\rho_I\leq\frac{\eta\rho_I^{3/2}}{4\sqrt{2N}|\alpha_I|}$

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

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

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

with $c=[\frac{448}{6}M^2]^{-1}$ since $|\alpha_I|\leq M\rho_I$. Therefore, we obtain (1) with the smallest of the three values of c, namely $c=[\frac{256}{3}M^2]^{-1}$, which concludes the proof of Theorem 1.

Remark 2 *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 (9) can be obtained by a direct use of the concentration property of the Gaussian.*

The proof of Corollary 1 follows by integration of (9) over η:

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

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 (10).

Finally, to prove the estimates in Theorem 2, 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

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

Hence, using Theorem 1, we see that the probability on the left of (11) is bounded from above by

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

which does not exceed $Cm^{-\beta}$ provided $\tilde{c}^2 c > a^2(1+\beta)$. The proof of (12) follows in a similar way from Corollary 1.

4. Proof of Theorem 3

This section is devoted to a proof of Theorem 3. 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 (13) 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

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

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

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

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 of 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 ε_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 ε_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

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

Therefore we conclude that

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

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

The third term e_3 is treated by the estimate (9) of Theorem 1:

$$\mathrm{Prob}\{e_3>\eta\}\le 4Ne^{-c\frac{m\eta^2}{N}}, \tag{2}$$

with

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

Hence we infer from (5) that

$$N \le b^p\tau_m^{-p}|f_\rho|^p_{\mathcal{B}^s} = b^p\tau_m^{-\frac{2}{2s+1}}|f_\rho|^p_{\mathcal{B}^s} = b^p\kappa^{-\frac{2}{2s+1}}|f_\rho|^p_{\mathcal{B}^s}\left(\frac{m}{\log m}\right)^{\frac{1}{2s+1}}, \tag{3}$$

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

$$\mathrm{Prob}\{e_2>0\}+\mathrm{Prob}\{e_4>0\}\le Cm^{-\beta}, \tag{4}$$

whenever $\kappa\ge\kappa_0$ with κ_0 depending on β, γ, and M and with C depending only on a.

Before proving this result, let us show that the combination (1), (2), (3) and (4) imply the validity of the estimates (14) and (15) in Theorem 3. We fix the value of β and we fix any constant κ for which (4) holds. Let $\eta_1 := \tilde{c}(\frac{\log m}{m})^{\frac{s}{2s+1}}$ with $\tilde{c}$ from (14) 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 (1) it follows that for $\tilde{c}>c_0$ we have $\mathrm{Prob}\{\|f_\rho - f_{\mathbf{z},m}\|>\eta_1\}\le\mathrm{Prob}\{e_2+e_3+e_4>\eta_1-\eta_2\}$. Hence, defining $\eta = (\tilde{c}-c_0)(\frac{\log m}{m})^{\frac{s}{2s+1}}$, the probability on the left side of (14) does not exceed

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

Moreover, on account of (2) and (3), we can estimate $\mathrm{Prob}\{e_3>\eta\}$ by

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

where $D^2 := \frac{(\tilde{c}-c_0)^2}{\kappa^{\frac{2}{2s+1}} b^p |f|^p_{\mathcal{B}^s}}$. The concentration estimate (14) follows now by taking $\tilde{c}$ large enough so that $1 - cD^2 + \beta \le 0$.

For the expectation estimate (15), we recall that according to Corollary 1, we have

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

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

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

by choosing β larger than $2s/(2s+1)$, for example $\beta = 1$. The same holds for e_4 and therefore we obtain (15).

It remains to prove (4). 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 *For any $\eta > 0$ and any element $I \in \mathcal{T}$, one has*

$$\text{Prob}\{\varepsilon_I(\mathbf{z}) \le \eta \ \text{ and } \ \varepsilon_I \ge b\eta\} \le Ce^{-cm\eta^2} \tag{5}$$

and

$$\text{Prob}\{\varepsilon_I \le \eta \ \text{ and } \ \varepsilon_I(\mathbf{z}) \ge b\eta\} \le Ce^{-cm\eta^2} \tag{6}$$

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

Before proving Lemma 4, let us show how this result implies (4). 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 $\varepsilon_I(\mathbf{z}) < \tau_m$ and $\varepsilon_I \ge b\tau_m$ or there exists an I such that both $\varepsilon_I(\mathbf{z}) \ge \tau_m$ and $\varepsilon_I < \tau_m/b$. It follows that

$$\begin{array}{rcl} \text{Prob}\{e_2 > 0\} & \le & \displaystyle\sum_{I\in\mathcal{T}_{j_0}} \text{Prob}\{\varepsilon_I(\mathbf{z}) \le \tau_m \ \text{ and } \ \varepsilon_I \ge b\tau_m\} \\ \varepsilon_I \le b\tau_m\}. & & \displaystyle + \sum_{I\in\mathcal{T}_{j_0}} \text{Prob}\{\varepsilon_I(\mathbf{z}) \ge \tau_m \ \text{ and } \ \varepsilon_I \le \tau_m/b\}. \end{array} \tag{7}$$

Using (5) with $\eta = \tau_m$ yields

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

We can treat the second sum in (7) the same way and obtain the same bound as the one for e_4 below. By similar considerations, we obtain

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

and we use (6) with $\eta = \tau_m/b$ which yields $\text{Prob}\{e_4 > 0\} \le Cm^{1/\gamma - c\kappa^2/b^2}$. We therefore obtain (4) by choosing $\kappa \ge \kappa_0$ with $c\kappa_0^2 = b^2(\beta + 1/\gamma)$.

We are left with the proof of Lemma 4. 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

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

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

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

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,

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

and similarly

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

so that the estimates (5) and (6) 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

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

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

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

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

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

and

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

It follows that

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

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

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

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

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

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

$$|\varepsilon_I - \varepsilon_I(\mathbf{z})| \leq \sum_{K=I^+,I^-} 2a_K|\alpha_K - \alpha_K(\mathbf{z})| + 3\delta a_K|\alpha_K|. \tag{8}$$

We first turn to (5), 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

$$\eta^2 \leq \frac{\varepsilon_I^2}{4} = \frac{\rho_{I^+}\rho_{I^-}}{\rho_I}\frac{(c_{I^+} - c_{I^-})^2}{4} \leq M^2\rho_L, \tag{9}$$

for $L = I^+, I^-$ and I. Combining (8) and (9), we estimate the probability by

$$\text{Prob}\{\varepsilon_I(\mathbf{z}) \leq \eta \ \text{ and } \ \varepsilon_I \geq 2\eta\} \leq \sum_{K=I^+,I^-} \Big(p_K + \sum_{J=I^-,I^+,I} q_{K,J} \Big), \tag{10}$$

with

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

and

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

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 $\eta^2 \leq \rho_K M^2$ 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

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

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

$$q_{K,J} \leq e^{-cm\eta^2}$$

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

Using these estimates for p_K and $q_{K,J}$ back in (10), we obtain (5).

We next turn to (6), 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

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

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

$$\text{Prob}\{\varepsilon_I \leq \eta \ \text{ and } \ \varepsilon_I(\mathbf{z}) \geq 2\eta\} \leq \sum_{K=I^+,I^1} \Big(p_K + \tilde{p}_K + \sum_{J=I^-,I^+,I} (q_{K,J} + \tilde{p}_J) \Big), \tag{11}$$

where now

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

and

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

and the additional probability is given by

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

Clearly, p_K and $q_{K,J}$ are estimated as in the proof of (5). 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 (11), we obtain (6), which concludes the proof of the lemma. □

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 $f_\rho \in L^2(X, \rho_X)$. Of course it is still assumed that $|y| \leq M$ almost surely, so that we also have $|f_\rho| \leq M$. For an arbitrary such f_ρ, we are interested in proving the convergence property

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

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

$$\lim_{m \to +\infty} \text{Prob}\{\|f_\rho - f_{\mathbf{z},m}\| > \varepsilon\} = 0.$$

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 3.

We first remark that the proof of the estimate

$$E(e_2^2) + E(e_4^2) \leq Cm^{-\beta},$$

remains unchanged under no smoothness assumption made on f_ρ.

Concerning the approximation term e_1, we have seen that

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

Under no smoothness assumptions, the convergence to 0 of these two terms still occurs when $j_0 \to +\infty$ and $\tau_m \to 0$, and therefore as $m \to +\infty$. This requires however that the union of the spaces $(S_{\Lambda_j})_{j \geq 0}$ is dense in $L^2(X, \rho_X)$. This is ensured by imposing natural restrictions on the splitting procedure generating the partitions which should be such that

$$\lim_{j \to +\infty} \sup_{I \in \Lambda_j} |I| = 0,$$

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

$$\sum_{J\in\mathcal{C}(I)} |J| \le \nu |I|,$$

with $\nu < 1$ independent of $I \in \mathcal{T}$. Under this restriction, classical results of measure theory state that $P_{\Lambda_j} f$ converges to f in $L^2(X,\rho_X)$ as $j \to +\infty$ for all $f \in L^2(\rho_X)$.

We are therefore ensured that $\|f_\rho - P_{\Lambda_{j_0}} f_\rho\|$ tends to 0 as $m \to +\infty$. 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 $L^2(X,\rho_X)$-orthogonal series

$$f = c_X\chi_X + \sum_{I\in\mathcal{T}} \psi_I, \text{ with } \psi_I := \sum_{J\in\mathcal{C}(I)} c_J\chi_J - c_I\chi_I,$$

We remark that $\|\psi_I\| = \varepsilon_I(f)$. It follows that for $\eta > 0$

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

Since by Parseval inequality,

$$\sum_{I\in\mathcal{T}} \varepsilon_I(f)^2 = \|f\|^2 - \|c_X\chi_X\|^2 < +\infty, \tag{1}$$

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

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

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

with

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

Note that since $(\varepsilon_I)_{I\in\mathcal{T}}$ is a square summable sequence according to (1), we have

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

where $\varphi(\eta) \to 0$ as $\eta \to 0$. Therefore if $\#(\Lambda(\tau_m/b))$ was simply controlled by $\#\{I\in\mathcal{T}\ ;\ \varepsilon_I > \tau_m/b\}$, we would derive that $E(e_3^2)$ would tend to 0 according to

$$E(e_3^2) \le C\frac{\tau_m^{-2}\varphi(\tau_m)\log(\tau_m^{-2}\varphi(\tau_m))}{m} \le \tilde{C}\frac{\tau_m^{-2}\varphi(\tau_m)\log m}{m} = \tilde{C}\varphi(\tau_m).$$

However, $\#(\Lambda(\tau_m/b)$ can be significantly larger due to the process of completing the set of thresholded coefficients into a proper tree. Since this process adds at most $j_0 - 1$ nodes J for each I such that $\varepsilon_I > \tau_m/b$, we have the estimate

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

where C depends on a and γ. It follows that if the threshold τ_m is modified into

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

we find that $E(e_3^2)$ goes to 0 according to

$$E(e_3^2) \leq C\frac{\tau_m^{-2}\varphi(\tau_m)\log m \log(\tau_m^{-2}\varphi(\tau_m)\log m)}{m} \leq \tilde{C}\frac{\tau_m^{-2}\varphi(\tau_m)\log m}{m} = \tilde{C}\varphi(\tau_m).$$

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 which was obtained in Theorem 3.

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

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

so that $N \log N/m$ tends to 0 if $1/\gamma > 2$. 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 $\mathcal{A}^{1/2}$.

Acknowledgments

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”.

References

Y. Baraud. Model selection for regression on a random design. *ESAIM Prob. et Stats.*, 6:127—146, 2002.

A. R. Barron. Complexity regularization with application to artificial neural network. In *Nonparametric functional estimation and related topics*, G. Roussas (ed.), pages 561—576, Kluwer Academic Publishers, 1991.

P. Binev and R. DeVore. Fast computation in adaptive tree approximation. *Numerische Math.*, 97:193—217, 2004.

L. Birgé. Model selection via testing : an alternative to (penalized) maximum likelihood estimators. Preprint, to appear in *Ann. IHP*, 2004.

L. Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203—268, 2001.

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

A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree-structured approximation and optimal encoding. *App. Comp. Harm. Anal.*, 11:192—226, 2001.

S. Cucker and S. Smale. On the mathematical foundations of learning. *Bulletin of AMS*, 39:1—49, 2001.

I. Daubechies. *Ten Lectures on Wavelets*, SIAM, Philadelphia, 1992.

R. A. DeVore. Nonlinear approximation. *Acta Numerica*, 7:51—150, 1998.

R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov. On mathematical methods of learning. *IMI Preprint*, 2004:10, University of South Carolina, 2004a.

R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov. Lower bounds in learning theory. *IMI Preprint*, 2004:22, University of South Carolina, 2004b.

D. L. Donoho. CART and best-ortho-basis : a connection. *Annals of Statistics*, 25:1870—1911, 1997.

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

D. L. Donoho and I. M. Johnstone. Minimax estimation via wavelet shrinkage. *Annals of Statistics*, 26(no. 3):879—921, 1998.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: Asymptopia? *Journal of the Royal Statistical Society*, 57:301—369, 1996a.

D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Annals of Statistics*, 24:508—539, 1996b.

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

S. V. Konyagin and V. N. Temlyakov. Some error estimates in learning theory. *IMI Preprints*, 2004:05, University of South Carolina, 2004a.

S. V. Konyagin and V. N. Temlyakov. The entropy in learning theory: Error estimates. *IMI Preprints*, 2004:09, University of South Carolina, 2004b.