Universal algorithms for learning theory Part II : piecewise polynomial functions

Peter Binev, Albert Cohen, Wolfgang Dahmen, and Ronald DeVore *

December 6, 2005

Abstract

This paper is concerned with estimating the regression function f_ρ in supervised learning by utilizing piecewise polynomial approximations on adaptively generated partitions. The main point of interest is algorithms that with high probability are optimal in terms of the least square error achieved for a given number m of observed data. In a previous paper [1], we have developed for each $\beta > 0$ an algorithm for piecewise constant approximation which is proven to provide such optimal order estimates with probability larger than $1 - m^{-\beta}$. In this paper, we consider the case of higher degree polynomials. We show that for general probability measures ρ empirical least squares minimization will not provide optimal error estimates with high probability. We go further in identifying certain conditions on the probability measure ρ which will allow optimal estimates with high probability.

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

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

1 Introduction

This paper is concerned with providing estimates in probability for the approximation of the regression function in supervised learning when using piecewise polynomials on adaptively generated partitions. We shall work in the following setting. We suppose that ρ is an unknown measure on a product space $Z := X \times Y$, where X is a bounded domain of $\mathbb{R}^d$ and $Y = \mathbb{R}$. Given m independent random observations $z_i = (x_i, y_i)$, $i = 1, \dots, m$,

*This research was supported by the Office of Naval Resarch Contracts ONR-N00014-03-1-0051, ONR/DEPSCoR N00014-03-1-0675 and ONR/DEPSCoR N00014-00-1-0470; the Army Research Office Contract DAAD 19-02-1-0028; the AFOSR Contract UF/USAF F49620-03-1-0381; and NSF contracts DMS-0221642 and DMS-0200187 and by the European Community's Human Potential Programme under contract HPRN-CT-202-00286, (BREAKING COMPLEXITY) and by the National Science Foundation Grant DMS-0200665.

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

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

with $\rho(y|x)$ the conditional probability measure on Y with respect to x. We shall use $\mathbf{z} = \{z_1, \ldots, z_m\} \subset Z^m$ to denote the set of observations.

One of the goals of learning is to provide estimates under minimal restrictions on the measure ρ since this measure is unknown to us. In this paper, we shall always work under the assumption that

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

almost surely. It follows in particular that $|f_\rho| \leq M$. This property of ρ can often be inferred in practical applications.

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

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

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

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

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

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

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

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|^2, \quad f \in L_2(X, \rho_X), \tag{1.6}$$

where

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

Our objective will be 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 with high probability. This type of regression problem is referred to as *distribution-free*. A recent survey on distribution free regression theory is provided in the book [8], which includes most existing approaches as well as the analysis of their rate of convergence in the expectation sense.

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

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

In other words, $f_{\mathbf{z}}$ is the best approximation to $(y_j)_{j=1}^m$ from $\mathcal{H}$ in the the empirical norm

$$\|g\|_m^2 := \frac{1}{m}\sum_{j=1}^m |g(x_j)|^2. \tag{1.9}$$

Typically, $\mathcal{H} = \mathcal{H}_m$ depends on a finite number $n = n(m)$ of parameters. In some algorithms, the number n is chosen using an a priori assumption on f_ρ. We want to avoid such prior assumptions. 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 this 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

$$\mathbb{P}\{\|f_\rho - f_{\mathbf{z}}\| \geq \eta\}, \quad \eta > 0 \quad \text{or} \quad \mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2) \tag{1.10}$$

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. Estimations in probability are to be preferred since they give more information about the success of a particular algorithm and they automatically yield an estimate in expectation by integrating with respect to η. Much more is known about the performance of algorithms in expectation than in probability as we shall explain below. The present paper will be mainly concerned about estimates in probability and we shall show that this problem has some interesting twists.

Estimates for the decay of the quantities in (1.10) are usually obtained under certain assumptions (called *priors*) on f_ρ. We emphasize that the algorithms should not depend on prior assumptions on f_ρ. Only in the analysis of the algorithms do we impose such prior assumptions in order to see how well the algorithm performs.

Priors on f_ρ are typically expressed by a condition of the type $f_\rho \in \Theta$ where Θ is a class of functions that necessarily must be contained in $L_2(X, \rho_X)$. If we wish the error, as measured in (1.10), 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. We have discussed the advantages and limitations of each of these approaches in [1] (see also [6]). In the present paper we shall take the view of (iii) and seek estimators which are optimal for a certain collection of priors of this type. Describing compactness in this way provides a bench mark for what could be achieved at best by a concrete estimator.

Our previous work [1] has considered the special case of approximation using *piecewise constants* on adaptively generated partitions. In that case, we have introduced algorithms that we prove to be optimal in the sense that their rate of convergence is best possible, for a large collection of prior classes, among all methods that utilize piecewise constant approximation on adaptively generated partitions based on isotropic refinements. Moreover, the methods we proposed in [1] had certain aesthetic and numerical advantages. For example, they are implemented by a simple thresholding procedure that can be done on line. This means that in the case of streaming data only a small number of updates are necessary as new data appear. Also, the analysis of our methods provided not only estimates in expectation but also the estimates in probability which we seek.

On the other hand, from an approximation theoretic point of view, using just piecewise constants severely limits the range of error decay rates that can be obtained even for very regular approximands. Much better rates can be expected when employing *higher order* piecewise polynomials which would result in equally local and, due to higher accuracy, overall more economical procedures.

However, in this previous work, we have purposefully not considered the general case of piecewise polynomial approximation because we had already identified certain notable - perhaps at first sight surprising - distinctions with the piecewise constant case. The purpose of the present paper is to analyze the case of general piecewise polynomial approximation, and in particular, to draw out these distinctions. We mention a few of these in this introduction.

In the piecewise constant case, we have shown that estimators built on empirical risk minimization (1.8) are guaranteed with high probability to approximate the regression function with optimal accuracy (in terms of rates of convergence). Here, by high probability we mean that the probability of not providing an optimal order of approximation tends to zero faster than $m^{-\beta}$ for any prescribed $\beta > 0$. We shall show in §3 that in general such probability estimates do not hold when using empirical risk minimization with piecewise polynomial approximation. This means that if we seek estimators which perform well in probability then either we must assume something more about the underlying probability measure ρ or we must find an alternative to empirical risk minimization.

In §4 we put some additional restrictions on the measure ρ_X and show that under these restrictions, we can again design algorithms based on empirical risk minimization which perform optimally with high probability. While, as we have already mentioned, these assumption on ρ_X are undesirable, we believe that in view of the counter example of §3 they represent roughly what can be done if one proceeds only with empirical risk minimization.

2 Approximating the Regression Function: General Strategies

In studying the estimation of the regression function, the question arises at the outset as to what are the best approximation methods to use in deriving algorithms for approximating f_ρ and therefore indirectly in defining prior classes? With no additional knowledge of ρ (and thereby f_ρ) there is no general answer to this question. However, we can draw some distinctions between certain strategies.

Suppose that we seek to approximate f_ρ by the elements from a hypothesis class $\mathcal{H} = \Sigma_n$. Here the parameter n measures the complexity associated to the process. In the case of approximation by elements from linear spaces we will take the space Σ_n to be of dimension n. For nonlinear methods, the space Σ_n is not linear and now n represents the number of parameters used in the approximation. For example, if we choose to approximate by piecewise polynomials on partitions with the degree r of the polynomials fixed then n could be chosen as the number of cells in the partition. The potential effectiveness of the approximation process for our regression problem would be measured by the error of approximation in the $L_2(X, \rho_X)$ norm. We define this error for a function

$g \in L_2(X, \rho_X)$ by

$$E_n(g) := E(g, \Sigma_n) := \inf_{S \in \Sigma_n} \|g - S\|, \quad n = 1, 2, \ldots. \tag{2.1}$$

If we have two approximation methods corresponding to sequences of approximation spaces (Σ_n) and (Σ'_n), then the second process would be superior to the first in terms of rates of approximation if $E'_n(g) \le CE_n(g)$ for all g and an absolute constant $C > 0$. For example, approximation using piecewise linear functions would in this sense be superior to using approximation by piecewise constants. In our learning context however, there are other considerations since: (i) the rate of approximation need not translate directly into results about estimating f_ρ because of the uncertainty in our observations, (ii) it may be that the superior approximation method is in fact much more difficult (or impossible) to implement in practice. For example, a typical nonlinear method may consist of finding an approximation to g from a family of linear spaces each of dimension N. The larger the family the more powerful the approximation method. However, too large of a family will generally make the numerical implementation of this method of approximation impossible.

Suppose that we have chosen the space Σ_n to be used as our hypothesis class $\mathcal{H}$ in the approximation of f_ρ from our given data $\mathbf{z}$. How should we define our approximation? As we have noted in the introduction, the most common approach is empirical risk minimization which gives the function $\hat{f}_{\mathbf{z}} := \hat{f}_{\mathbf{z},\Sigma_n}$ defined by (1.8). However, since we know $|f_\rho| \le M$, the approximation will be improved if we post-truncate $\hat{f}_{\mathbf{z}}$ by M. For this, we define the truncation operator

$$T_M(x) := \min(|x|, M) \operatorname{sign}(x) \tag{2.2}$$

for any real number x and define

$$f_{\mathbf{z}} := f_{\mathbf{z},\mathcal{H}} := T_M(\hat{f}_{\mathbf{z},\mathcal{H}}). \tag{2.3}$$

There are general results that provide estimates for how well $f_{\mathbf{z}}$ approximates f_ρ. One such estimate given in [8] (see Theorem 11.3) applies when $\mathcal{H}$ is a linear space of dimension n and gives [1]

$$\mathbb{E}(\|f_\rho - f_{\mathcal{Z}}\|^2) \lesssim \frac{n \log(m)}{m} + \inf_{g \in \mathcal{H}} \|f_\rho - g\|^2. \tag{2.4}$$

The second term is the bias and equals our approximation error $E_n(f_\rho)$ for approximation using the elements of $\mathcal{H}$. The first term is the variance which bounds the error due to uncertainty. One can derive rates of convergence in expectation by balancing both terms (see [8] and [6]) for specific applications.

The deficiency of this approach is that one needs to know the behavior of $E_n(f_\rho)$ in order to choose the best value of n and this requires a priori knowledge of f_ρ. There is a general procedure known as *model selection* which circumvents this difficulty and tries to automatically choose a good value of n (depending on f_ρ) by introducing a penalty term. Suppose that $(\Sigma_n)_{n=1}^m$ is a family of linear spaces each of dimension n. For each

[1] Here and later in this paper we use the notation $A \lesssim B$ to mean $A \le CB$ for some absolute constant C

$n = 1, 2, \ldots, m$, we have the corresponding estimator $f_{\mathbf{z},\Sigma_n}$ defined by (2.3) and the empirical error

$$E_{n,\mathbf{z}} := \frac{1}{m}\sum_{j=1}^{m}(y_j - f_{\mathbf{z},\Sigma_n}(x_j))^2. \tag{2.5}$$

Notice that $E_{n,\mathbf{z}}$ is a computable quantity which we can view as an estimate for $E_n(f_\rho)$. In complexity regularization, one chooses a value of n by

$$n^* := n^*(\mathbf{z}) := \operatorname{argmin}\left\{E_{n,\mathbf{z}} + \frac{n\log m}{m}\right\}. \tag{2.6}$$

We now define

$$f_{\mathbf{z}} := f_{\mathbf{z},\Sigma_{n^*}} \tag{2.7}$$

as our estimator to f_ρ. One can then prove (see Chapter 12 of [8]) that whenever f_ρ can be approximated to accuracy $E_n(f_\rho) \le Mn^{-s}$ for some $s > 0$, then[2]

$$\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2 \le C\Big[\frac{\log m}{m}\Big]^{\frac{2s}{2s+1}} \tag{2.8}$$

which save for the logarithm is an optimal rate estimation in expectation.

For a certain range of s, one can also prove similar estimates in probability (see [6]). Notice that the estimator did not need to have knowledge of s and nevertheless obtains the optimal performance.

Model selection can also be applied in the setting of nonlinear approximation, i.e. when the spaces Σ_n are nonlinear but in this case, one needs to invoke conditions on the compatibility of the penalty with the complexity of the approximation process as measured by an entropy restriction. We refer the reader to Chapter 12 of [8] for a more detailed discussion of this topic and will briefly take up this point again in §2.3.

Let us also note that the penalty approach is not always compatible with the practical requirement of *on-line* computations. By on-line computation, we mean that the estimator for the sample size m can be derived by a simple update of the estimator for the sample size $m-1$. In penalty methods, the optimization problem needs to be globally re-solved when adding a new sample. However, when there is additional structure in the approximation process such as the adaptive partitioning that we discuss in the next section, then there are algorithms that circumvent this difficulty (see the discussion of CART algorithms given in the following section).

2.1 Adaptive Partitioning

In this paper, we shall be interested in approximation by piecewise polynomials on partitions generated by adaptive partitioning. We shall restrict our discussion to the case $X = [0,1]^d$ and the case of dyadic partitions. However, all results would follow in the more general setting described in [1].

[2]We use the following conventions concerning constants throughout this paper. Constants like $C, c, \tilde{c}$ depend on the specified parameters but they may vary at each occurrence, even in the same line. We shall indicate the dependence of the constant on other parameters whenever this is important.

Let $\mathcal{D}_j = \mathcal{D}_j(X)$ be 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 $\mathcal{T}$ corresponds to a cube $I \in \mathcal{D}$. If $I \in \mathcal{D}_j$, then its children are the 2^d dyadic cubes $J \subset \mathcal{D}_{j+1}$ with $J \subset I$. We denote the set of children of I by $\mathcal{C}(I)$. We call I the parent of each such child J and write $I = P(J)$. A *proper* subtree $\mathcal{T}_0$ of $\mathcal{T}$ is a collection of nodes of $\mathcal{T}$ with the properties: (i) the root node $I = X$ is in $\mathcal{T}_0$, (ii) if $I \neq X$ is in $\mathcal{T}_0$ then its parent is also in $\mathcal{T}_0$. We obtain (dyadic) partitions Λ of X from finite proper subtrees $\mathcal{T}_0$ of $\mathcal{T}$. Given any such $\mathcal{T}_0$ the *outer leaves* of $\mathcal{T}_0$ consist of all $J \in \mathcal{T}$ such that $J \notin \mathcal{T}_0$ but $P(J)$ is in $\mathcal{T}_0$. The collection $\Lambda = \Lambda(\mathcal{T}_0)$ of outer leaves of $\mathcal{T}_0$ is a partition of X into dyadic cubes. It is easily checked that

$$\#(\mathcal{T}_0) \leq \#(\Lambda) \leq 2^d \#(\mathcal{T}_0). \tag{2.9}$$

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

We let Π_K denote the space of multivariate polynomials of total degree K with $K \geq 0$ a fixed integer. In the analysis we present, the space Π_K could be replaced by any space of functions of fixed finite dimension without affecting our general discussion. Given a dyadic cube $I \in \mathcal{D}$, and a function $f \in L_2(X, \rho_X)$, we denote by $p_I(f)$ the best approximation to f on I:

$$p_I(f) := \operatorname{argmin}_{p \in \Pi_K} \|f - p\|_{L_2(I,\rho_X)}. \tag{2.10}$$

Given $K > 0$ and a partition Λ, let us denote by $\mathcal{S}_\Lambda^K$ the space of piecewise polynomial functions of degree K subordinate to Λ. Each $S \in \mathcal{S}_\Lambda^K$ can be written

$$S = \sum_{I \in \Lambda} p_I \chi_I, \quad p_I \in \Pi_K, \tag{2.11}$$

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 the approximation of a given function $f \in L_2(X, \rho_X)$ by the elements of S_Λ^K. The best approximation to f in this space is given by

$$P_\Lambda f := \sum_{I \in \Lambda} p_I(f) \chi_I. \tag{2.12}$$

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

$$E_n(f) := \|f - P_{\Lambda_n} f\|, \quad n = 0, 1, \ldots \tag{2.13}$$

which is the error for uniform refinement. We shall denote by $\mathcal{A}^s$ the *approximation class* consisting of all functions $f \in L_2(X, \rho_X)$ such that

$$E_n(f) \leq M_0 2^{-nds}, \quad n = 0, 1, \ldots. \tag{2.14}$$

Notice that $\#(\Lambda_n) = 2^{nd}$ so that the decay in (2.14) is like N^{-s} with N the number of elements in the partition. The smallest M_0 for which (2.14) 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 $ds > 0$ with smoothness measured with respect to ρ_X. For example, if ρ_X is the Lebesgue measure then $\mathcal{A}^{s/d} = B^s_\infty(L_2)$, $0 < s \leq 1$, with equivalent norms. Here $B^s_\infty(L_2)$ is a Besov space. For $s < K$ one can take $\sup_t t^{-s}\omega_K(f,t)_{L_2}$ as a norm for this space, where $\omega_K(f,t)_{L_2}$ is the Kth order modulus of smoothness in L_2 (see [2] for the definition and properties of Besov spaces).

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 criterion that is motivated by adaptive wavelet constructions such as those given in [4] for image compression. Given a function $f \in L_2(X, \rho_X)$, we define the local atoms

$$\psi_I(f) := \sum_{J \in \mathcal{C}(I)} p_J(f)\chi_J - p_I(f)\chi_I, \quad I \neq X, \quad \psi_X(f) := p_X(f), \tag{2.15}$$

and

$$\epsilon_I(f) := \|\psi_I(f)\|. \tag{2.16}$$

Clearly, we have

$$f = \sum_{I \in \mathcal{D}} \psi_I(f), \tag{2.17}$$

and since the ψ_I are mutually orthogonal, we also have

$$\|f\|^2_{L_2(X,\rho_X)} = \sum_{I \in \mathcal{D}} \epsilon_I(f)^2. \tag{2.18}$$

The number $\epsilon_I(f)$ gives the improvement in the $L_2(X, \rho_X)$ error squared when the cell I is refined.

We let $\mathcal{T}(f, \eta)$ be the smallest proper tree that contains all $I \in \mathcal{D}$ such that $\epsilon_I(f) > \eta$. Corresponding to this tree we have the partition $\Lambda(f, \eta)$ consisting of the outer leaves of $\mathcal{T}(f, \eta)$. We shall define some new smoothness spaces $\mathcal{B}^s$ which measure the regularity of a given function f by the size of the tree $\mathcal{T}(f, \eta)$. Given $s > 0$, we let $\mathcal{B}^s$ be the collection of all $f \in L_2(X, \rho_X)$ such that for $p = (s + 1/2)^{-1/2}$, the following is finite

$$|f|^p_{\mathcal{B}^s} := \sup_{\eta > 0} \eta^p \#(\mathcal{T}(f, \eta)). \tag{2.19}$$

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

$$\|f - P_{\Lambda(f,\eta)} f\| \leq C_s |f|^{\frac{1}{2s+1}}_{\mathcal{B}^s} \eta^{\frac{2s}{2s+1}} \leq C_s |f|_{\mathcal{B}^s} N^{-s}, \quad N := \#(\mathcal{T}(f, \eta)), \tag{2.20}$$

where the constant C_s depends only on s. The proof of this estimate can be based on the same strategy as used in [4] where a similar result is proven in the case of the Lebesgue

measure: one introduces the trees $\mathcal{T}_j := \mathcal{T}(f, 2^{-j}\eta)$ which have the property $\mathcal{T}_j \subset \mathcal{T}_{j+1}$, and then writes

$$\begin{aligned} \|f - P_{\Lambda(f,\eta)} f\|^2 &= \textstyle\sum_{I \notin \mathcal{T}(f,\eta)} \|\psi_I\|^2 \\ &= \textstyle\sum_{j\geq 0} \sum_{I \in \mathcal{T}_{j+1} \setminus \mathcal{T}_j} \|\psi_I\|^2 \\ &\leq \textstyle\sum_{j\geq 0} \#(\mathcal{T}_{j+1})(2^{-j}\eta)^2 \\ &\leq |f|^p_{\mathcal{B}^s} \textstyle\sum_{j\geq 0} (2^{-j}\eta)^{2-p}, \end{aligned}$$

which gives (2.20) with $C_s := \sum_{j\geq 0} 2^{(p-2)j}$.

Invoking (2.9), 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 $0 < q \leq \infty$, is contained in $\mathcal{B}^{s/d}$ (see [4]). 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.

2.2 An adaptive algorithm for learning

In the learning context, we cannot use the algorithm described in the previous section since the regression function f_ρ and the measure ρ are not known to us. Instead we shall use an empirical version of this adaptive procedure.

Given the data $\mathbf{z}$ and any Borel set $I \subset X$, we define

$$p_{I,\mathbf{z}} := \operatorname{argmin}_{p\in\Pi_K} \frac{1}{m} \sum_{i=1}^m (p(x_i) - y_i)^2 \chi_I(x_i). \tag{2.21}$$

When there are no x_i in I, we set $p_{I,\mathbf{z}} = 0$.

Given a partition Λ of X we define the estimator $f_{\mathbf{z}}$ as

$$f_{\mathbf{z}} = f_{\mathbf{z},\Lambda} := \sum_{I\in\Lambda} T_M(p_{I,\mathbf{z}})\chi_I \tag{2.22}$$

with T_M the truncation operator defined earlier. Note that the empirical minimization (2.21) is not done over the truncated polynomials, since this is not numerically feasible. Instead, truncation is only used as a post processing.

As in the previous section, we have two ways to generate partitions Λ. The first is to use the uniform partition Λ_n consisting of all dyadic cells in $\mathcal{D}_n$. The second is to define an empirical analogue of the ϵ_I. For each cell I in the master tree $\mathcal{T}$, we define

$$\epsilon_I(\mathbf{z}) := \|T_M(\sum_{J\in\mathcal{C}(I)} p_{J,\mathbf{z}}\chi_J - p_{I,\mathbf{z}}\chi_I)\|_m, \tag{2.23}$$

where $\|\cdot\|_m$ is the empirical norm defined in (1.9).

A data based adaptive partitioning requires limiting the depth of corresponding trees. To this end, let $\gamma > 0$ be an arbitrary but fixed constant. We define $j_0 = j_0(m,\gamma)$ as

the smallest integer j such that $2^{jd} \geq \tau_m^{-1/\gamma}$. We then consider the smallest tree $\mathcal{T}(\tau_m, \mathbf{z})$ which contains the set

$$\Sigma(\mathbf{z}, m) := \{I \in \cup_{j \leq j_0} \Lambda_j \quad : \quad \epsilon_I(\mathbf{z}) \geq \tau_m\}, \tag{2.24}$$

where τ_m is a threshold to be set further. We then define the partition $\Lambda = \Lambda(\tau_m, \mathbf{z})$ associated to this tree and the corresponding estimator $f_{\mathbf{z}} := f_{\mathbf{z},\Lambda}$. Obviously, the role of the integer j_0 is to limit the depth search for the coefficient $\epsilon_I(\mathbf{z})$ which are larger than the threshold τ_m. Without this restriction, the tree $\mathcal{T}(\tau_m, \mathbf{z})$ could be infinite preventing a numerical implementation. The essential steps of the adaptive algorithm in the present setting read as follows:

Algorithm: *Given* $\mathbf{z}$*, choose* $\gamma > 0$*,* $\tau_m := \kappa\sqrt{\frac{\log m}{m}}$ *and*

- *for* $j_0(m, \gamma)$ *determine the set* $\Sigma(\mathbf{z}, m)$ *according to (2.24);*
- *form* $\mathcal{T}(\tau_m, \mathbf{z}), \Lambda(\tau_m, \mathbf{z})$ *and compute* $f_{\mathbf{z}}$ *according to (2.22).*

For further comments concerning the treatment of streaming data we refer to an analogous strategy outlined in [1].

In our previous work [1], we have analyzed the above algorithm in the case of *piecewise constant approximation* and we have proved the following result.

Theorem 2.1 *Let* $\beta, \gamma > 0$ *be arbitrary. Then, using piecewise constant approximations in the above scheme, i.e.* $K = 0$*, there exists* $\kappa_0 = \kappa_0(\beta, \gamma, M)$ *such that if* $\kappa \geq \kappa_0$ *in the definition of* τ_m*, then whenever* $f_\rho \in \mathcal{A}^\gamma \cap \mathcal{B}^s$ *for some* $s > 0$*, the following concentration estimate holds*

$$\mathbb{P}\left\{\|f_\rho - f_{\mathbf{z}}\| \geq \tilde{c}\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\right\} \leq Cm^{-\beta}, \tag{2.25}$$

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

Let us make some remarks on this theorem. First note that truncation does not play any role in the case of piecewise constant approximation since in that case the constant of best empirical approximation automatically is $\leq M$ in absolute value. The theorem gives an estimate for the error $\|f_\rho - f_{\mathbf{z}}\|$ in probability which is the type of estimate we are looking for in this paper. From this one obtains a corresponding estimate in expectation. The order of approximation can be shown to be optimal save for the logarithmic term by using the results on lower estimates from [6]. Finally, note that the role of the space $\mathcal{A}^\gamma$ is a minor one since the only assumption on γ is that it be positive. This assumption merely guarantees that a finite depth search will behave close to an infinite depth search.

The goal of the present paper is to determine whether the analogue of Theorem 2.1 holds when piecewise polynomials are used in place of piecewise constants. We shall see that this is not the case by means of a counterexample in §3. We shall then show that such estimates are possible if we place restrictions on the measure ρ_X.

2.3 Estimates in Expectation

Before starting the analysis of results in probability for the higher order piecewise polynomial case, let us point out that it is possible to derive estimates in expectation for adaptive partitions constructed by other strategies, such as model selection by complexity regularization. In particular, the following result can easily be derived from Theorem 12.1 in [8].

Theorem 2.2 *Let $\gamma > 0$ be arbitrary and let $j_0 = j_0(\gamma, m)$ be defined again as the smallest integer j such that $2^{-jd} \leq (\log m/m)^{1/2\gamma}$. Consider the set $\mathcal{M}$ of all partitions Λ induced by proper trees $\mathcal{T} \subset \cup_{j \leq j_0} \Lambda_j$. Then, there exists $\kappa_0 = \kappa_0(d, K)$ such that if*

$$\mathrm{pen}_m(\Lambda) = \frac{\kappa \log m}{m} \#(\Lambda).$$

for some $\kappa \geq \kappa_0$, the estimator defined by $f_{\mathbf{z}} := f_{\mathbf{z},\Lambda^*}$ with

$$\Lambda^* := \mathrm{argmin}_{\Lambda \in \mathcal{M}} \left\{ \|f_{\mathbf{z},\Lambda} - y\|_m^2 + \mathrm{pen}_m(\Lambda) \right\},$$

satisfies

$$\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|) \leq C \Big(\frac{\log m}{m} \Big)^{\frac{s}{2s+1}}, \quad m = 1, 2, \ldots, \tag{2.26}$$

if $f_\rho \in \mathcal{A}^\gamma \cap \mathcal{B}^s$ where the constant C depends on $\kappa, M, |f_\rho|_{\mathcal{B}^s}, |f_\rho|_{\mathcal{A}^s}$, but not on m.

Let us also remark that the search of the optimal partition Λ^* in the above theorem can be performed at a reasonable computational cost using a CART algorithm (see e.g. [3] or [7]). Note that our approach for selecting the appropriate partition differs from the CART algorithms, in the sense that it is based on a thresholding procedure rather than solving an optimization problem.

3 A counterample

We begin by showing that in general we cannot obtain optimal estimates with high probability when using empirical risk minimization with piecewise polynomials of degree larger than zero. We shall first consider the case of approximation by linear functions on the interval $X = [-1, 1]$ for the bound $M = 1$. For each $m = 1, 2, \ldots$, we will construct a measure $\rho = \rho_m$ on $[-1, 1] \times [-1, 1]$ for which empirical risk minimization does not perform well in probability.

Let p be the polynomial

$$p = \mathrm{argmin}_{g \in \Pi_1} \mathbb{E}(|g - y|^2) = \mathrm{argmin}_{g \in \Pi_1} \|f_\rho - g\|^2, \tag{3.1}$$

with f_ρ the regression function and $\|\cdot\|$ the $L_2(X, \rho_X)$ norm. Consider also the empirical least square minimizer

$$\hat{p} = \mathrm{argmin}_{g \in \Pi_1} \sum_{i=1}^m |g(x_i) - y_i|^2. \tag{3.2}$$

We are interested in the concentration properties between $T(p)$ and $T(\hat{p})$ in the $\|\cdot\|$ metric, where $T(u) = \text{sign}(u)\max\{1, |u|\}$ is the truncation operator. We shall prove the following result, which expresses the fact that we cannot hope for a distribution free concentration inequality with a fast rate of decay of the probability.

Lemma 3.1 *Given any $\beta > 2$, there exist absolute constants $c, \tilde{c} > 0$ such that for each $m = 1, 2, \ldots$, there is a distribution $\rho = \rho_m$ such that the following inequalities hold*

$$\mathbb{P}\{\|T(p) - T(\hat{p})\| \geq c\} \geq \tilde{c}m^{-\beta+1} \tag{3.3}$$

and

$$\mathbb{P}\{\|f_\rho - T(\hat{p})\| \geq c\} \geq \tilde{c}m^{-\beta+1}. \tag{3.4}$$

On the other hand, we have

$$\|f_\rho - T(p)\|^2 \leq C_0[m^{-2\beta+4} + m^{-\beta}] \tag{3.5}$$

for an absolute constant C_0.

Remark 3.2 *Note that this clearly implies the same results as in (3.3) and (3.4) with the truncation operator removed. By contrast, for least square approximation by constants q, we have (see [1]) for all ρ, m and η*

$$\mathbb{P}\{\|q - \hat{q}\| \geq \eta\} \lesssim e^{-cm\eta^2}. \tag{3.6}$$

Proof of Lemma 3.1: In order to prove this result, we consider the probability measure

$$\rho_X := (1/2 - \kappa)(\delta_{-\gamma} + \delta_\gamma) + \kappa(\delta_{-1} + \delta_1), \tag{3.7}$$

where $\gamma := \gamma_m = \frac{1}{3m}$ and $\kappa := \kappa_m := m^{-\beta}$. We then define $\rho = \rho_m$ completely by

$$y(\gamma) = 1,\ y(-\gamma) = -1,\ y(\pm 1) = 0, \quad \text{with probability 1.} \tag{3.8}$$

Therefore, there is no randomness in the y direction. It follows that

$$f_\rho(\gamma) = 1,\ f_\rho(-\gamma) = -1,\ f_\rho(\pm 1) = 0. \tag{3.9}$$

We next proceed in three steps.

1. Properties of p: By symmetry, the linear function that best approximates f_ρ in $L_2(\rho_X)$ is of the form $p(x) = ax$, where a minimizes

$$F(a) = (1/2 - \kappa)(a\gamma - 1)^2 + \kappa a^2. \tag{3.10}$$

We therefore find that

$$p(\pm\gamma) = \pm a\gamma = \pm\frac{(1/2 - \kappa)\gamma^2}{(1/2 - \kappa)\gamma^2 + \kappa} = \pm 1 + O(m^{-\beta+2}). \tag{3.11}$$

This shows that

$$\|f_\rho - Tp\|^2 \leq C_0 m^{-2\beta+4} + 2\kappa \leq C_0[m^{-2\beta+4} + m^{-\beta}] \tag{3.12}$$

with C_0 an absolute constant.

2. Properties of $\hat{p}$: We can write the empirical least square polynomial $\hat{p}$ as

$$\hat{p}(x) = \hat{b} + \hat{a}(x - \hat{\xi}), \tag{3.13}$$

where $\hat{b} = \frac{1}{m}\sum_{i=1}^m y_i$ and $\hat{\xi} := \frac{1}{m}\sum_{i=1}^m x_i$. (Notice that 1 and $x - \hat{\xi}$ are orthogonal with respect to the empirical measure $\frac{1}{m}\sum_{i=1}^m \delta_{x_i}$.) From

$$\frac{\hat{p}(\gamma) - \hat{p}(-\gamma)}{2\gamma} = \hat{a} = \frac{\hat{p}(\hat{\xi}) - \hat{p}(\gamma)}{(\hat{\xi} - \gamma)}, \tag{3.14}$$

it follows that

$$\hat{p}(\gamma) = \hat{p}(-\gamma)(1 + \frac{2\gamma}{\hat{\xi} - \gamma})^{-1} + \hat{p}(\hat{\xi})(1 + \frac{\hat{\xi} - \gamma}{2\gamma})^{-1}. \tag{3.15}$$

Since $\hat{p}(\hat{\xi}) = \hat{b} \in [-1, 1]$, (3.15) shows that whenever $\hat{\xi} \geq 2\gamma$, then either $\hat{p}(\gamma) \leq 1/2$ or $\hat{p}(-\gamma) \geq -1/2$. It follows that whenever $\hat{\xi} \geq 2\gamma$, we have

$$\|f_\rho - T(\hat{p})\|^2 \geq (1/2 - \kappa)(1/2)^2 \geq c^2 \tag{3.16}$$

with c an absolute constant. Using (3.12), we see that (3.3) is also satisfied provided that $\mathbb{P}\{\hat{\xi} > 2\gamma\} \geq \tilde{c} m^{-\beta+1}$.

3. Study of $\hat{\xi}$: We consider the event where $x_i = 1$ for one $i \in \{1, \cdots, m\}$ and $x_j = \gamma$ or $-\gamma$ for the other $j \neq i$. In such an event, we have

$$\hat{\xi} \geq \frac{1}{m}(1 - (m-1)\gamma) > \frac{1}{m}(1 - 1/3) = \frac{2}{3m} = 2\gamma. \tag{3.17}$$

The probability of this event is

$$P = m\kappa(1 - 2\kappa)^{m-1} \geq \tilde{c} m^{-\beta+1}. \tag{3.18}$$

This concludes the proof of (3.3). □

Let us now adjust the example of the lemma to give information about piecewise linear approximation on adaptively generated dyadic partitions. We first note that we can rescale the measure of the lemma to any given interval I. We choose a dyadic interval I of length 2^{-m} and scale the measure of the lemma to that interval. We denote this measure again by ρ_m. The regression function f_ρ will be denoted by f_m and it will take values ± 1 at the rescaled points corresponding to γ and $-\gamma$ and will take the value 0 everywhere else. We know from Lemma 3.1 that we can approximate f_m by a single polynomial to accuracy

$$\|f_\rho - p\|^2 \leq C_0[m^{-2\beta+4} + m^{-\beta}]. \tag{3.19}$$

Also, if we allow dyadic partitions with more than m elements, we can approximate f_ρ exactly. In other words, f_ρ is in $\mathcal{B}^s$ with $s = \min(\beta - 2, \beta/2)$.

On the other hand, any adaptively generated partition with at most m elements will have an interval J containing I. The empirical data sets will be rescaled to I and the empirical $\hat{p}_J$'s will all be the empirical linear functions of Lemma 3.1 scaled to I. Hence, for any of the bad draws $\mathbf{z}$ of Lemma 3.1, we will have

$$\|f_\rho - \hat{f}_{\mathbf{z}}\| \geq c \tag{3.20}$$

on a set of $\mathbf{z}$ with probability larger than $\tilde{c}m^{-\beta+1}$.

This shows that empirical least squares using piecewise linear functions on adaptively generated partitions will not provide optimal bounds with high probability. Note that the above results are not in contradiction with optimal estimates in expectation. The counter example also indicates, however, that the arguments leading to optimal rates in expectation based on complexity regularization cannot be expected to be refined towards estimates in probability.

4 Optimal results in probability under regularity assumptions on ρ_X

In view of the results of the previous section, it is not possible to prove optimal convergence rates with high probability for adaptive methods based on piecewise polynomials for the general setting of the regression problem in learning. In this section, we want to show that if we impose some restrictions on the marginal measure ρ_X, then high probability results are possible.

We fix the value K of the polynomial space Π_K in this section. For each dyadic cube $I \in \mathbb{R}^d$ and any function $f \in L_2(X, \rho_X)$, we define $p_I(f)$ as in (2.10). This means that given any dyadic partition Λ, the best approximation to f from $\mathcal{S}_\Lambda$ is given by the projector

$$P_\Lambda f = \sum_{I \in \Lambda} p_I(f)\chi_I. \tag{4.1}$$

We shall work under the following assumption in this section:

Assumption A : *There exists a constant $C_A > 0$ such that for each dyadic cube I, there exists an $L_2(I, \rho_X)$-orthonormal basis $(L_{I,k})_{k=1,\dots,\lambda}$ of Π_K (with λ the algebraic dimension of Π_K) such that*

$$\|L_{I,k}\|_{L_\infty(I)} \leq C_A(\rho_X(I))^{-1/2}, \quad k = 1, \dots, \lambda. \tag{4.2}$$

Note that this assumption implies that the least squares projection is uniquely defined on each I by

$$p_I(f) = \sum_{k=1}^{\lambda} \langle f, L_{I,k}\rangle_{L_2(I,\rho_X)} L_{I,k}, \tag{4.3}$$

and in particular $\rho_X(I) \neq 0$. It also implies that for all partitions Λ and for all $f \in L_\infty(X)$,

$$\|P_\Lambda f\|_{L_\infty} \leq \lambda C_A \|f\|_{L_\infty}, \tag{4.4}$$

i.e. the projectors P_Λ are bounded in L_∞ independently of Λ. Indeed, from Cauchy-Schwartz inequality, we have

$$\|L_{I,k}\|_{L_1(I,\rho_X)} \leq (\rho_X(I))^{1/2} \|L_{I,k}\|_{L_2(I,\rho_X)} = (\rho_X(I))^{1/2} \tag{4.5}$$

and therefore for all $x \in I \in \Lambda$

$$|(P_\Lambda f)(x)| \leq \sum_{k=1}^{\lambda} |\langle f, L_{I,k}\chi_I\rangle L_{I,k}(x)| \leq \lambda C_A \|f\|_{L_\infty(I)}. \tag{4.6}$$

It is readily seen that Assumption A holds when ρ_X is the Lebesgue measure dx. On the other hand, it may not hold when ρ_X is a very irregular measure. A particular simple case where Assumption A always holds is the following:

Assumption B : *We have $d\rho_X = \omega(x)dx$ and there exists a constant $C_B > 0$ such that for all $I \in \mathcal{D}$, there exists a convex subset $D \subset I$ with $|I| \leq B|D|$ and such that $0 < \rho_X(I) \leq C_B|I| \inf_{x\in D} \omega(x)$.*

This assumption is obviously fulfilled by weight functions such that $0 < c \leq \omega(x) \leq C$, but it is also fulfilled by functions which may vanish (or go to $+\infty$) with a power-like behaviour at isolated points or lower dimensional manifolds. Let us explain why Assumption B implies Assumption A. Any convex domain D can be framed by $E \subset D \subset \tilde{E}$ where E is an ellipsoid and $\tilde{E}$ a dilated version of E by a fixed factor depending only on the dimension d. Therefore, if P is a polynomial with $\|P\|_{L_2(I,\rho_X)} = 1$, we have $\|P\|^2_{L_\infty(I)} \leq C\|P\|^2_{L_\infty(D)}$ holds for a constant C depending only on the degree K, the spatial dimension d and the constant B in Assumption B. Hence, using Assumption B, we further conclude

$$\begin{aligned}
\|P\|^2_{L_\infty(I)} &\lesssim \|P\|^2_{L_\infty(D)} \lesssim |D|^{-1} \int_D |P(x)|^2 dx \\
&\lesssim |D|^{-1} \int_D |P(x)|^2 \rho_X(I)^{-1} |I| \omega(x) dx \\
&\lesssim \rho_X(I)^{-1} \int_I |P(x)|^2 d\rho \;=\; \rho_X(I)^{-1}.
\end{aligned}$$

Here, all the constants in the inequalities depend on d, K, B, C_B but are independent of $I \in \mathcal{T}$.

4.1 Probability estimates for the empirical estimator

We shall now show that under Assumption A, we can estimate the discrepancy between the truncated least squares polynomial approximation to f_ρ and the truncated least squares

polynomial fit to the empirical data. This should be compared with the counterexample of the last section which showed that for general ρ we do not have this property.

Lemma 4.1 *There exists a constant $c > 0$ which depends on the constant C_A in Assumption A, on the polynomial space dimension $\lambda = \lambda(K)$ of Π_K and on the bound M, such that for all $I \in \mathcal{D}$*

$$\mathbb{P}\{\|T_M(p_I)\chi_I - T_M(p_{I,\mathbf{z}})\chi_I\| > \eta\} \leq \tilde{c}e^{-cm\eta^2}, \tag{4.7}$$

where $\tilde{c} = 2(\lambda + \lambda^2)$,

Proof : We clearly have

$$\|T_M(p_I)\chi_I - T_M(p_{I,\mathbf{z}})\chi_I\|^2 \leq 4M^2\rho_X(I), \tag{4.8}$$

and therefore it suffices to prove (4.7) for those I such that $\eta^2 < 4M^2\rho_X(I)$. We use the basis $(L_{I,k})_{k=1,\cdots,L}$ of Assumption A to express p_I and $p_{I,\mathbf{z}}$ according to

$$p_I = \sum_{k=1}^{\lambda} c_{I,k}L_{I,k} \text{ and } p_{I,\mathbf{z}} = \sum_{k=1}^{\lambda} c_{I,k}(\mathbf{z})L_{I,k}, \tag{4.9}$$

and we denote by c_I and $c_I(\mathbf{z})$ the corresponding coordinate vectors. We next remark that since T_M is a contraction

$$\|T_M(p_I)\chi_I - T_M(p_{I,\mathbf{z}})\chi_I\| \leq \|p_I\chi_I - p_{I,\mathbf{z}}\chi_I\|, \tag{4.10}$$

it follows that

$$\|T_M(p_I)\chi_I - T_M(p_{I,\mathbf{z}})\chi_I\|^2 \leq \sum_{k=1}^{\lambda} |c_{I,k} - c_{I,k}(\mathbf{z})|^2 = \|c_I - c_I(\mathbf{z})\|^2_{\ell_2(\mathbb{R}^\lambda)}, \tag{4.11}$$

were $\|\cdot\|_{\ell_2(\mathbb{R}^\lambda)}$ is the λ-dimensional Euclidean norm. Introducing

$$\alpha_{I,k} = \int_Z yL_{I,k}(x)\chi_I(x)d\rho \text{ and } \alpha_{I,k}(\mathbf{z}) := \frac{1}{m}\sum_{i=1}^{m} y_iL_{I,k}(x_i)\chi_I(x_i), \tag{4.12}$$

with α_I and $\alpha_I(\mathbf{z})$ the corresponding coordinate vectors, we clearly have

$$c_I = \alpha_I. \tag{4.13}$$

On the other hand, we have

$$G_I(\mathbf{z})c_I(\mathbf{z}) = \alpha_I(\mathbf{z}),$$

where

$$(G_I(\mathbf{z}))_{k,l} := \frac{1}{m}\sum_{i=1}^{m} L_{I,k}(x_i)L_{I,l}(x_i) = \langle L_{I,k}, L_{I,l}\rangle_m,$$

and $c_I(\mathbf{z}) := 0$ when there are no x_i's in I. Therefore, when $G_I(\mathbf{z})$ is invertible, we can write

$$c_I(\mathbf{z}) - c_I = G_I(\mathbf{z})^{-1}\alpha_I(\mathbf{z}) - \alpha_I = G_I(\mathbf{z})^{-1}[\alpha_I(\mathbf{z}) - G_I(\mathbf{z})\alpha_I],$$

and therefore

$$\|c_I(\mathbf{z}) - c_I\|_{\ell_2(\mathbb{R}^\lambda)} \leq \|G_I(\mathbf{z})^{-1}\|_{\ell_2(\mathbb{R}^\lambda)}\Big(\|\alpha_I(\mathbf{z}) - \alpha_I\|_{\ell_2(\mathbb{R}^\lambda)} + \|G_I(\mathbf{z}) - I\|_{\ell_2(\mathbb{R}^\lambda)}\|\alpha_I\|_{\ell_2(\mathbb{R}^\lambda)}\Big), \tag{4.14}$$

where we also denote by $\|G\|_{\ell_2(\mathbb{R}^\lambda)}$ the spectral norm for an $\lambda \times \lambda$ matrix G. Since $\|G_I(\mathbf{z}) - I\|_{\ell_2(\mathbb{R}^\lambda)} \leq 1/2$ implies $\|G_I(\mathbf{z})^{-1}\|_{\ell_2(\mathbb{R}^\lambda)} \leq 2$ it follows that $\|c_I(\mathbf{z}) - c_I\|_{\ell_2(\mathbb{R}^\lambda)} \leq \eta$ provided that we have jointly

$$\|\alpha_I(\mathbf{z}) - \alpha_I\|_{\ell_2(\mathbb{R}^\lambda)} \leq \frac{\eta}{4}, \tag{4.15}$$

and

$$\|G_I(\mathbf{z}) - I\|_{\ell_2(\mathbb{R}^\lambda)} \leq \min\Big\{\frac{1}{2}, \frac{\eta}{4}\|\alpha_I\|^{-1}_{\ell_2(\mathbb{R}^\lambda)}\Big\}. \tag{4.16}$$

By Bernstein's inequality, we have

$$\begin{aligned}\mathbb{P}\left\{\|\alpha_I(\mathbf{z}) - \alpha_I\|_{\ell_2(\mathbb{R}^\lambda)} > \frac{\eta}{4}\right\} &\leq \mathbb{P}\{|\alpha_{I,k}(\mathbf{z}) - \alpha_{I,k}| > \frac{\eta}{4\lambda^{1/2}} \text{ for some } k\} \\ &\leq 2\lambda e^{-c\frac{m\eta^2}{\sigma^2 + S\eta}},\end{aligned} \tag{4.17}$$

with c depending on λ where, on account of (4.2),

$$S := \sup_k \sup_{x,y} |yL_{I,k}(x)| \leq MC_A\rho_X(I)^{-1/2},$$

and

$$\sigma^2 := \sup_k \mathbb{E}(y^2|L_{i,k}|^2) \leq M^2.$$

Since $\eta^2 < 2M^2\rho_X(I)$, we thus have, up to a change in the constant c,

$$\mathbb{P}\left\{\|\alpha_I(\mathbf{z}) - \alpha_I\|_{\ell_2(\mathbb{R}^\lambda)} > \frac{\eta}{4}\right\} \leq 2\lambda e^{-c\frac{m\eta^2}{M^2 + M\rho_X(I)^{-1/2}\eta}} \leq 2\lambda e^{-cm\eta^2}, \tag{4.18}$$

with c depending on λ, C_A and M. In a similar way, we obtain by Bernstein's inequality that

$$\begin{aligned}\mathbb{P}\left\{\|G_I(\mathbf{z}) - I\|_{\ell_2(\mathbb{R}^\lambda)} > \xi\right\} &\leq \mathbb{P}\left\{|(G_I(\mathbf{z}))_{k,l} - \delta_{k,l}| > \frac{\xi}{\lambda} \text{ for some } (k,l)\right\} \\ &\leq \lambda^2 e^{-c\frac{m\xi^2}{\sigma^2 + S\xi}},\end{aligned} \tag{4.19}$$

with c depending on λ, where now again by (4.2)

$$S := \sup_{k,l} \sup_x |L_{I,k}(x)L_{I,l}(x)| \leq C_A^2\rho_X(I)^{-1}, \tag{4.20}$$

and

$$\sigma^2 := \sup_{k,l} \mathbb{E}(|L_{I,k}(x)L_{I,l}(x)|^2) \leq C_A^2\rho_X(I)^{-1}. \tag{4.21}$$

In the case $\xi = \frac{1}{2} \leq \frac{\eta}{4}\|\alpha_I\|^{-1}$, we thus obtain from (4.19), using that $\eta^2 \leq 2M^2\rho_X(I)$,

$$\mathbb{P}\left\{\|G_I(\mathbf{z}) - I\|_{\ell_2(\mathbb{R}^\lambda)} > \xi\right\} \leq 2\lambda^2 e^{-c\frac{m}{\rho_X(I)^{-1}}} \leq 2\lambda^2 e^{-cm\eta^2}, \tag{4.22}$$

with c depending on λ, C_A and M. In the case $\xi = \frac{\eta}{4}\|\alpha_I\|^{-1} \leq \frac{1}{2}$, we notice that

$$\|\alpha_I\|_{\ell_2(\mathbb{R}^\lambda)} \leq \lambda^{1/2} M \sup_k \|L_{I,k}\|_{L^1(I,\rho_X)} \leq \lambda^{1/2} M C_A \rho_X(I)^{1/2}, \tag{4.23}$$

because of Assumption A. It follows that $\xi \geq \eta(16\lambda M^2 C_A^2 \rho_X(I))^{-1/2}$ and the absolute value of the exponent in the exponential of (4.19) is bounded from below by

$$\frac{cm\eta^2\rho_X(I)^{-1}}{24\lambda M^2 C_A^4 \rho_X(I)^{-1}} \geq cm\eta^2, \tag{4.24}$$

with c depending on λ, C_A and M. The proof of (4.7) is therefore complete. □

Remark 4.2 *The constant c in (4.7) depends on M and C_A and behaves like $(MC_A^2)^{-2}$.*

4.2 Learning on Fixed and Uniform Partitions

From the basic estimate (4.7), we can immediately derive an estimate for an arbitrary but fixed partition Λ consisting of disjoint dyadic cubes. If $|\Lambda| = N$, we have

$$\begin{aligned}\mathbb{P}\left\{\|f_{\mathbf{z},\Lambda} - T_M(P_\Lambda f_\rho)\| > \eta\right\} \quad \leq \quad & \mathbb{P}\left\{\|T_M(p_I)\chi_I - T_M(p_{I,\mathbf{z}})\chi_I\| > \frac{\eta}{N^{1/2}}\right. \\ & \left.\text{for some } I \in \Lambda\right\},\end{aligned}$$

which yields the following analogue to Theorem 2.1 of [1].

Remark 4.3 *Under Assumption A one has for any fixed integer $K \geq 0$, any partition Λ and $\eta > 0$*

$$\mathbb{P}\left\{\|f_{\mathbf{z},\Lambda} - T_M(P_\Lambda f_\rho)\| > \eta\right\} \leq C_0 N e^{-c\frac{m\eta^2}{N}}, \tag{4.25}$$

where $N := \#(\Lambda)$ and $C_0 = C_0(\lambda)$ and $c = c(\lambda, M, C_A)$.

We can then derive by integration over $\eta > 0$ an estimate in the mean square sense

$$\mathbb{E}\left(\|f_{\mathbf{z},\Lambda} - T_M(P_\Lambda f_\rho)\|^2\right) \leq C\frac{N\log N}{m}, \tag{4.26}$$

similar to Corollary 2.2 of [1], with $C = C(M, C_A, \lambda)$. Combining these estimates with the definition of the approximation classes $\mathcal{A}^s$, we obtain similar convergence rates as in Theorem 2.3 of [1].

Theorem 4.4 *If $f_\rho \in \mathcal{A}^s$ and the estimator is defined by $f_{\mathbf{z}} := f_{\Lambda_j,\mathbf{z}}$ with j chosen as the smallest integer such that $2^{jd(1+2s)} \geq \frac{m}{\log m}$, then given any $\beta > 0$, there exist constants $C = C(M, C_A, \lambda, \beta)$ and $\tilde{c} = \tilde{c}(M, C_A, \lambda, \beta)$ such that*

$$\mathbb{P}\left\{\|f_\rho - f_{\mathbf{z}}\| > (\tilde{c} + |f_\rho|_{\mathcal{A}^s})\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\right\} \leq Cm^{-\beta}. \tag{4.27}$$

There also exists a constant $C = C(M, C_A, \lambda)$ such that

$$\mathbb{E}\left(\|f_\rho - f_{\mathbf{z}}\|^2\right) \leq (C + |f_\rho|_{\mathcal{A}^s}^2)\Big(\frac{\log m}{m}\Big)^{\frac{2s}{2s+1}}. \tag{4.28}$$

As we have noted earlier, the second estimate (4.28) could have been obtained in a different way, namely by using Theorem 11.3 in [8] and without Assumption A. On the other hand, it is not clear that the probability estimate (4.27) could be obtained without this assumption.

4.3 Learning on Adaptive Partitions

We now turn to the adaptive algorithm as defined in the §2.2. Here, we shall extend Theorem 2.5 of [1] in two ways. Recall that the depth of the tree is limited by $j_0 = j_0(m, \gamma)$ the smallest integer j such that $2^{jd} \geq \tau_m^{-1/\gamma}$.

Theorem 4.5 *We fix an arbitrary $\beta \geq 1$ and $\gamma > 0$. If we take the threshold $\tau_m := \kappa\sqrt{\frac{\log m}{m}}$ with $\kappa \geq \kappa_0 = \kappa_0(\gamma, \beta, M, \lambda, d, C_A)$, then for any ρ such that $f_\rho \in \mathcal{A}^\gamma \cap \mathcal{B}^s$, $s > 0$, the adaptive algorithm gives an $f_{\mathbf{z}}$ satisfying the concentration estimate*

$$\mathbb{P}\left\{\|f_\rho - f_{\mathbf{z}}\| \geq c\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\right\} \leq m^{-\beta}, \tag{4.29}$$

with $c = c(s, d, \lambda, \beta, |f_\rho|_{\mathcal{B}^s}, |f_\rho|_{\mathcal{A}^\gamma})$. We also have the following expectation bound

$$\mathbb{E}\left(\|f_\rho - f_{\mathbf{z}}\|^2\right) \leq C\Big(\frac{\log m}{m}\Big)^{\frac{2s}{2s+1}}, \tag{4.30}$$

with $C = C(s, \lambda, M, C_A, d, \beta, |f_\rho|_{\mathcal{B}^s}, |f_\rho|_{\mathcal{A}^\gamma})$.

A defect of this theorem is the dependency of κ_0 on C_A which may be unknown in our regression problem. We shall propose later another convergence theorem where this defect is removed.

The strategy for proving Theorem 4.5 is similar to the one for Theorem 2.5 in [1]. The idea is to show that the set of coefficients chosen by the adaptive empirical algorithm are with high probability similar to the set that would be chosen if the adaptive thresholding took place directly on f_ρ. This will be established by probability estimates which control the discrepancy between ϵ_I and $\epsilon_I(\mathbf{z})$. This is given by the following result, which is a substitute to Lemma 4.1 in [1].

Lemma 4.6 *For any $\eta > 0$ and any element $I \in \Lambda_{j_0}$, one has*

$$\mathbb{P}\left\{\epsilon_I(\mathbf{z}) \leq \eta \text{ and } \epsilon_I \geq 8\lambda C_A \eta\right\} \leq \tilde{c}(1 + \eta^{-C})e^{-cm\eta^2} \tag{4.31}$$

and

$$\mathbb{P}\left\{\epsilon_I \leq \eta \text{ and } \epsilon_I(\mathbf{z}) \geq 4\eta\right\} \leq \tilde{c}(1 + \eta^{-C})e^{-cm\eta^2}, \tag{4.32}$$

where $\tilde{c} = \tilde{c}(\lambda, M, d)$, $c = c(\lambda, M, C_A, d)$ and $C = C(\lambda, d)$.

The proof of Lemma 4.6 is rather different from the counterpart in [1] and is postponed to the end of this section. Its usage in deriving the claimed bounds in Theorem 4.5, however, is very similar to the proof in [1]. Because of a few changes due to the truncation

operator we briefly recall first its main steps suppressing at times details that could be recovered from [1].

For any two partitions Λ_0, Λ_1 we denote by $\Lambda_0 \cap \Lambda_1$, $\Lambda_0 \cup \Lambda_1$ the partitions induced by the corresponding trees $\mathcal{T}_0 \cap \mathcal{T}_1$, $\mathcal{T}_0 \cup \mathcal{T}_1$, respectively. Recall that the data dependent partitions $\Lambda(\eta, \mathbf{z})$, defined in (2.24), do not contain cubes from levels higher than j_0. Likewise, for any threshold $\eta > 0$, we need the deterministic analogue $\Lambda(\eta) := \Lambda(f_\rho, \eta)$, where $\Lambda(f_\rho, \eta)$ is the partition associated to the smallest tree $\mathcal{T}(f_\rho, \eta)$ that contains all I for which $|\epsilon_I(f_\rho)| > \eta$.

As in the proof of Theorem 2.5 in [1], we estimate the error by

$$\|f_\rho - f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})}\| \leq e_1 + e_2 + e_3 + e_4 \tag{4.33}$$

with each term now defined by

$$\begin{aligned} e_1 &:= \|f_\rho - T_M(P_{\Lambda(\tau_m,\mathbf{z})\cup\Lambda(8C_A\lambda\tau_m)}f_\rho)\|, \\ e_2 &:= \|T_M(P_{\Lambda(\tau_m,\mathbf{z})\cup\Lambda(8C_A\lambda\tau_m)}f_\rho) - T_M(P_{\Lambda(\tau_m,\mathbf{z})\cap\Lambda(\tau_m/4)}f_\rho)\|, \\ e_3 &:= \|T_M(P_{\Lambda(\tau_m,\mathbf{z})\cap\Lambda(\tau_m/4)}f_\rho) - f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})\cap\Lambda(\tau_m/4)}\|, \\ e_4 &:= \|f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})\cap\Lambda(\tau_m/4)} - f_{\mathbf{z},\Lambda(\tau_m,\mathbf{z})}\|. \end{aligned}$$

The first term e_1 is treated by approximation

$$e_1 \leq C_s(8C_A\lambda\tau_m)^{\frac{2s}{2s+1}}|f_\rho|_{\mathcal{B}^s} + \tau_m|f_\rho|_{\mathcal{A}^\gamma} \leq c(|f_\rho|_{\mathcal{B}^s} + |f_\rho|_{\mathcal{A}^\gamma})\Big(\frac{m}{\log m}\Big)^{-\frac{s}{2s+1}}, \tag{4.34}$$

where $c = c(s, C_A, \lambda, \kappa)$. The second summand in the first estimate of (4.34) stems from the fact that optimal trees might be clipped by the restriction to Λ_{j_0} and this missing part exploits the assumption $f_\rho \in \mathcal{A}^\gamma$.

The third term e_3 can be estimated by an inequality similar to (4.25) although $\Lambda(\tau_m, \mathbf{z}) \cap \Lambda(\tau_m/4)$ is a data-dependent partition. We use the fact that all the cubes in this partition are always contained in the tree consisting of $\mathcal{T}(f_\rho, \tau_m/4)$ completed by its outer leaves, i.e. by the cubes of $\Lambda(\tau_m/4)$. We denote by $\mathcal{T}^*(f_\rho, \tau_m/4)$ this tree and by N its cardinality. Using the same argument that led us to (4.25) for a fixed partition, we obtain

$$\mathbb{P}\{e_3 > \eta\} \leq C_0 N e^{-c\frac{m\eta^2}{N}} \tag{4.35}$$

with c and C_0 the same constants as in (4.25). From the definition of the space $\mathcal{B}^s$, we have the estimate

$$N \leq (2^d+1)\#(\mathcal{T}(f_\rho, \tau_m/4)) \leq (2^d+1)(4\kappa^{-1}|f_\rho|_{\mathcal{B}^s})^{\frac{2}{2s+1}}\Big(\frac{m}{\log m}\Big)^{\frac{1}{2s+1}}. \tag{4.36}$$

Thus, for κ larger than $\kappa_0(\beta)$ we obtain that

$$\mathbb{P}\Big\{e_3 > c\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\Big\} \leq m^{-\beta}, \tag{4.37}$$

with $c = c(d, \lambda, \beta, C_A, |f_\rho|_{\mathcal{B}^s}, s, \kappa, d)$.

The terms e_2 and e_4 are treated using the probabilistic estimates of Lemma 4.6. From these estimates it follows that

$$\mathbb{P}\{e_2 > 0\} + \Pr\{e_4 > 0\} \leq 2\#(\Lambda_{j_0})\tilde{c}(1+\eta^{-C})e^{-cm\eta^2} \tag{4.38}$$

with $\eta = \tau_m/4$. It follows that for $\kappa \geq \kappa_0(\beta, \gamma, \lambda, M, C_A, d)$, we have

$$\mathbb{P}\{e_2 > 0\} + \mathbb{P}\{e_4 > 0\} \leq m^{-\beta}. \tag{4.39}$$

Combining the estimates for e_1, e_2, e_3 and e_4, we therefore obtain the probabilistic estimate (2.25).

For the expectation estimate, we clearly infer from (4.34)

$$\mathbb{E}(e_1^2) \leq c^2(|f_\rho|^2_{\mathcal{B}^s} + |f_\rho|^2_{\mathcal{A}^\gamma})\Big(\frac{\log m}{m}\Big)^{\frac{2s}{2s+1}} \tag{4.40}$$

with $c = c(s, C_A, \lambda, \kappa)$. Also, since $e_2^2, e_4^2 \leq 4M^2$, we obtain

$$\mathbb{E}(e_2^2) + \mathbb{E}(e_4^2) \leq *M^2 m^{-\beta}, \tag{4.41}$$

with $C = C(\kappa, \lambda, C_A, M, d)$. We can estimate $\mathbb{E}(e_3^2)$ by integration of (4.35) which gives

$$\mathbb{E}(e_3^2) \leq 4M^2 m^{-\beta} + c^2\Big(\frac{\log m}{m}\Big)^{\frac{2s}{2s+1}}, \tag{4.42}$$

with c as in (4.37). This completes the proof of the Theorem 4.5. □

We now turn to the proof of Lemma 4.6. We shall give a different approach than that used in [1] for piecewise constant approximation. Recall that

$$\epsilon_I(\mathbf{z}) := \|T_M(\sum_{J\in\mathcal{C}(I)} p_{J,\mathbf{z}}\chi_J - p_{I,\mathbf{z}}\chi_I)\|_m, \tag{4.43}$$

and

$$\epsilon_I := \epsilon_I(f_\rho) := \|\sum_{J\in\mathcal{C}(I)} p_J\chi_J - p_I\chi_I\|. \tag{4.44}$$

We introduce the auxiliary quantities

$$\tilde{\epsilon}_I(\mathbf{z}) := \|T_M(\sum_{J\in\mathcal{C}(I)} p_{J,\mathbf{z}}\chi_J - p_{I,\mathbf{z}}\chi_I)\|, \tag{4.45}$$

and

$$\epsilon_I^* := \|T_M(\sum_{J\in\mathcal{C}(I)} p_J\chi_J - p_I\chi_I)\|. \tag{4.46}$$

For the proof of (4.31), we first remark that from the L_∞ bound (4.4) on the least square projection, we know that

$$\|\sum_{J\in\mathcal{C}(I)} p_J\chi_J - p_I\chi_I\|_{L_\infty} \leq \sup_{J\in\mathcal{C}(I)} \|p_J\chi_J\|_{L_\infty} + \|p_I\chi_I\|_{L_\infty} \leq 2\lambda C_A M. \tag{4.47}$$

It follows that the inequality

$$|\sum_{J\in\mathcal{C}(I)} p_J\chi_J - p_I\chi_I| \le 2\lambda C_A |T_M(\sum_{J\in\mathcal{C}(I)} p_J\chi_J - p_I\chi_I|), \tag{4.48}$$

holds at all points, and therefore $\epsilon_I \le 2\lambda C_A \epsilon_I^*$ so that

$$\mathbb{P}\{\epsilon_I(\mathbf{z}) \le \eta \text{ and } \epsilon_I \ge 8\lambda C_A\eta\} \le \mathbb{P}\{\epsilon_I(\mathbf{z}) \le \eta \text{ and } \epsilon_I^* \ge 4\eta\}. \tag{4.49}$$

We now write

$$\begin{aligned}
\mathbb{P}\{\epsilon_I(\mathbf{z}) \le \eta \text{ and } \epsilon_I^* \ge 4\eta\} &\le \mathbb{P}\{\tilde{\epsilon}_I(\mathbf{z}) \le 3\eta \text{ and } \epsilon_I^* \ge 4\eta\} \\
&\quad + \mathbb{P}\{\epsilon_I(\mathbf{z}) \le \eta \text{ and } \tilde{\epsilon}_I(\mathbf{z}) \ge 3\eta\} \\
&\le \mathbb{P}\{|\tilde{\epsilon}_I(\mathbf{z}) - \epsilon_I^*| \ge \eta\} + \mathbb{P}\{\tilde{\epsilon}_I(\mathbf{z}) - 2\epsilon_I(\mathbf{z}) \ge \eta\} \\
&= p_1 + p_2.
\end{aligned}$$

The first probability p_1 is estimated by

$$p_1 \le \mathbb{P}\{\|T_M(\sum_{J\in\mathcal{C}(I)} p_J\chi_J - p_I\chi_I) - T_M(\sum_{J\in\mathcal{C}(I)} p_{J,\mathbf{z}}\chi_J - p_{I,\mathbf{z}}\chi_I)\| > \eta\} \le \tilde{c}e^{-cm\eta^2}, \tag{4.50}$$

with $\tilde{c} = \tilde{c}(\lambda, d)$ and $c = c(\lambda, M, C_A, d)$, where the last inequality is obtained by the same technique as in the proof of Lemma 4.1 applied to $\sum_{J\in\mathcal{C}(I)} p_J\chi_J - p_I\chi_I$ instead of $p_I\chi_I$ only.

The second probability p_2 is estimated by using results from [8]. Fix I and let $\mathcal{F} = \mathcal{F}(I)$ be the set of all functions f of the form

$$f = T_M(\sum_{J\in\mathcal{C}(I)} q_J\chi_J - q_I\chi_I). \tag{4.51}$$

where q_I and the q_J are arbitrary polynomials from Π_K. Let $\mathbf{t} = (t_1, \ldots, t_{2m})$ with the t_j chosen independently and at random with respect to the measure ρ_X and consider the discrete norm

$$\|f\|_{\mathbf{t}} := \frac{1}{2m}\sum_{j=1}^{2m} |f(t_j)|^2. \tag{4.52}$$

We will need a bound for the covering number $\mathcal{N}(\mathcal{F}, \eta, \mathbf{t})$ which is the smallest number of balls of radius η which cover $\mathcal{F}$ with respect to the norm $\|\cdot\|_{\mathbf{t}}$. It is well known that if V is a linear space of dimension q and $\mathcal{G} := \{T_M g : g \in V\}$ then

$$\mathcal{N}(\mathcal{G}, \eta, \mathbf{t}) \le (C\eta)^{-(2q+1)}, \quad 0 < \eta \le 1, \tag{4.53}$$

with $C = C(M)$ (see e.g. Theorems 9.4 and 9.5 in [8]). In our present situation, this leads to

$$\mathcal{N}(\mathcal{F}, \eta, \mathbf{t}) \le (C\eta)^{-2\lambda[2^d+1]+2}, 0 < \eta < 1, \tag{4.54}$$

We apply this bound on covering numbers in Theorem 11.2 of [8] which states that

$$\mathbb{P}\{\|f\| - 2\|f\|_m > \eta \text{ for some } f \in \mathcal{F}\} \le 3e^{-\frac{m\eta^2}{288M^2}}\mathbb{E}(\mathcal{N}(\mathcal{F}, \eta, \mathbf{t})) \tag{4.55}$$

Here the probability is with respect to $\mathbf{z}$ and the expectation is with respect to $\mathbf{t}$. We can put in the entropy bound (4.54) for $\mathcal{N}$ into (4.55) and obtain

$$p_2 \leq \mathbb{P}\left\{\|f\| - 2\|f\|_m > \eta \text{ for some } f \in \mathcal{F}\right\} \leq \tilde{c}\eta^{-C}e^{-cm\eta^2} \tag{4.56}$$

with $\tilde{c}, C, c$ as stated in the lemma. This is the estimate we wanted for p_2.

For the proof of (4.32), we first remark that obviously $\epsilon_I^* \leq \epsilon_I$ so that

$$\mathbb{P}\left\{\epsilon_I \leq \eta \text{ and } \epsilon_I(\mathbf{z}) \geq 4\eta\right\} \leq \mathbb{P}\left\{\epsilon_I^* \leq \eta \text{ and } \epsilon_I(\mathbf{z}) \geq 4\eta\right\}. \tag{4.57}$$

We proceed similarily to the proof of (4.31) by writing

$$\begin{aligned}
\mathbb{P}\left\{\epsilon_I^* \leq \eta \text{ and } \epsilon_I(\mathbf{z}) \geq 4\eta\right\} &\leq \mathbb{P}\left\{\epsilon_I^* \leq \eta \text{ and } \tilde{\epsilon}_I(\mathbf{z}) \geq \frac{3}{2}\eta\right\} \\
&\qquad + \mathbb{P}\left\{\tilde{\epsilon}_I(\mathbf{z}) \leq \frac{3}{2}\eta \text{ and } \epsilon_I(\mathbf{z}) \geq 4\eta\right\} \\
&\leq \mathbb{P}\left\{|\tilde{\epsilon}_I(\mathbf{z}) - \epsilon_I^*| \geq \eta/2\right\} + \mathbb{P}\left\{\epsilon_I(\mathbf{z}) - 2\tilde{\epsilon}_I(\mathbf{z}) \geq \eta\right\} \\
&= p_1 + p_2.
\end{aligned} \tag{4.58}$$

The first probability p_1 is estimated as previously. For the second probability p_2, we need a symmetric statement to Theorem 11.2 in [8], to derive

$$\mathbb{P}\left\{\|f\|_m - 2\|f\| > \eta \text{ for some } f \in \mathcal{F}\right\} \leq Ce^{-cm\eta^2} \tag{4.59}$$

with c and C as before. It is easily checked from the proof of Theorem 11.2 in [8], that such a statement also holds (one only needs to modify the first page of the proof of this theorem in [8] in order to bound $\mathbb{P}\{\|f\|_m - 2\|f\| > \eta \text{ for some } f \in \mathcal{F}\}$ by $3\mathbb{P}\{\|f\|_m - \|f\|'_m > \eta/4 \text{ for some } f \in \mathcal{F}\}/2$ and the rest of the proof is then identical). The proof of Lemma 4.6 is therefore complete. □

We finally modify our algorithm slightly by choosing a slightly larger threshold which is now independent of the unknown constant C_A, namely $\tau_m := \frac{\log m}{\sqrt{m}}$. We could actually use any threshold of the type

$$\tau_m := \kappa(m)\sqrt{\frac{\log m}{m}}, \tag{4.60}$$

where $\kappa(m)$ is a sequence which grows very slowly to $+\infty$. This results in an additional logarithm factor in our convergence estimates. Moreover, the same analysis as in §5 of [1] shows that this new algorithm is universally consistent. We record this in the following theorem for which we do not give a proof since it is very similar to the proof of Theorem 4.5.

Theorem 4.7 *Given an arbitrary $\beta \geq 1$ and $\gamma > 0$, we take the threshold $\tau_m := \frac{\log m}{\sqrt{m}}$. Then the adaptive algorithm has the property that whenever $f_\rho \in \mathcal{A}^\gamma \cap \mathcal{B}^s$ for some $s > 0$, the following concentration estimate holds*

$$\mathbb{P}\left\{\|f_\rho - f_{\mathbf{z}}\| \geq c\Big(\frac{\log m}{\sqrt{m}}\Big)^{\frac{2s}{2s+1}}\right\} \leq m^{-\beta}, \tag{4.61}$$

with $c = c(s, C_A, \lambda, |f_\rho|_{\mathcal{B}^s}, |f_\rho|_{\mathcal{A}^\gamma})$*, as well as the following expectation bound*

$$\mathbb{E}\left(\|f_\rho - f_{\mathbf{z}}\|^2\right) \leq C\Big(\frac{\log m}{\sqrt{m}}\Big)^{\frac{4s}{2s+1}} \tag{4.62}$$

with $C = C(s, \lambda, M, C_A, d, |f_\rho|_{\mathcal{B}^s}, |f_\rho|_{\mathcal{A}^\gamma})$*. For a general regression function* f_ρ*, we have the universal consistency estimate*

$$\lim_{m\to+\infty} E(\|f_\rho - f_{\mathbf{z}}\|^2) = 0, \tag{4.63}$$

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

$$\lim_{m\to+\infty} \mathbb{P}\{\|f_\rho - f_{\mathbf{z}}\| > \epsilon\} = 0. \tag{4.64}$$

References

[1] Binev, P., A. Cohen, W. Dahmen, R. DeVore, and V. Temlyakov (2004) *Universal algorithms in learning theory - Part I : piecewise constant functions*, Journal of Machine Learning Research (JMLR), 6(2005), 1297–1321.

[2] C. Bennett and R. Sharpley, Interpolation of Operators, Vol. 129 in Pure and Applied Mathematics, Academic Press, N.Y., 1988.

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

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

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

[6] DeVore, R., G. Kerkyacharian, D. Picard and V. Temlyakov (2004) *Mathematical methods for supervised learning*, to appear in J. of FOCM.

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

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

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