Universal Piecewise Polynomial Estimators for Machine Learning

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

December 13, 2006

Abstract

We review and expand somewhat on some recent developments concerning the construction and analysis of piecewise polynomial estimators for the regression problem in Mathematical Learning Theory. The discussion will center on two issues. The first of these is computational efficiency including possible online capability. The second is universality by which we mean the capability of the estimator to give rise to optimal convergence rates for a possibly wide class of prior classes without using any a-priori knowledge on the memebership of the regression function to any of these classes. More precisely, the main point of interest are estimators for which the probability of exceeding an optimal rate tends to zero as the number m of observed data increases. We focus on nonlinear methods built on piecewise polynomial approximation on adaptively refined partitions. We describe a class of schemes that are inspired by thresholding concepts for wavelet expansions. We point out obstacles to treating piecewise polynomials of degree higher than one as compared with piecewise constant estimators and discuss several possible remedies.

Key Words: Regression, universal piecewise polynomial estimators, complexity regularization, optimal convergence rates in probability, adaptive partitioning, thresholding

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

1 Introduction

Increasingly complex measuring devices along with growing computing and data storage capacities lead to the aquisition of enormous data sites typically hiding the essential information one is looking for. The quantifiable extraction of information embedded in large data sets that are typically polluted by noise is therefore a central task which is of

*This research was supported by the Office of Naval Research Contracts ONR-N00014-03-1-0051, ONR/DEPSCoR N00014-03-1-0675, and ONR/DEPSCoR N00014-05-1-0715; AFOSR Contract UF/USAF F49620-03-1-0381; DARPA/NGA Contract HM1582-05-2-0001; ARO/DoD Contract W911NF-05-1-0227; National Science Foundation Grant DMS-354707; the French-German PROCOPE contract 11418YB; and the Leibniz Programme of the DFG.

growing importance in many application areas ranging from science and technology over finance to social sciences. This is reflected by the rapid developments in Mathematical Learning Theory that address such issues and provides a theoretical foundation for tasks like pattern recognition, classification and regression. Mathematical Learning Theory draws on concepts from nonparametric statistics, functional analysis, numerical analysis and last but not least from approximation theory. It is fair to say that the potential synergies offered by the interplay of these disciplines have not been exhausted yet.

We do not attempt to give an even nearly representative overview of all of the most recent exciting developments but rather focus on a segment that emphasizes the roles of computational complexity and nonlinear approximation. Specifically, we will be 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 $z_i = (x_i, y_i)$, $i = 1, \dots, m$, 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 $\mathbf{z} = \{z_1, \dots, z_m\} \subset Z^m$ to denote the set of observations. We denote by ρ_X the marginal probability measure on X defined by

$$\rho_X(S) := \rho(S \times Y), \tag{1.2}$$

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

$$d\rho(x, y) = d\rho(y|x) d\rho_X(x). \tag{1.3}$$

The interest in f_ρ lies among other things in the following fact. Defining the risk functional

$$\mathcal{E}(f) := \int_Z (y - f(x))^2 d\rho, \tag{1.4}$$

it is easy to check that

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|^2, \tag{1.5}$$

where

$$\|\cdot\| := \|\cdot\|_{L_2(X,\rho_X)}, \tag{1.6}$$

and $L_2(X, \rho_X)$ consists of all functions from X to Y which are square integrable with respect to ρ_X. Thus, f_ρ is the minimizer of $\mathcal{E}(f)$ over $f \in L_2(X, \rho_X)$.

This type of regression problem is referred to as *distribution-free* since we make no assumptions of the distribution ρ. A recent survey on distribution free regression theory is provided in the monograph [13], which includes most existing approaches to analyzing

their rate of convergence in the *expectation sense*, i.e. to provide estimates for $\mathbb{E}(\|f_{\mathbf{z}}-f_\rho\|^2)$ where the expectation is taken with respect to the product measure ρ^m.

It should be emphasized that a central issue in Learning Theory is to provide estimates for f_ρ 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 for each x,

$$|y| \leq M, \tag{1.7}$$

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

It is desirable to obtain stronger estimates than just for the expectation. Therefore, 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 in probability. Specifically, we would like to bound $\mathbb{P}\{\|f_{\mathbf{z}}-f_\rho\| > \eta\}$ for thresholds η that are allowed to decay with increasing sample size m. Generally speaking bounds in probability are much stronger than those in expectation since from probability bounds we can infer good estimates in expectation while in the other direction estimates in expectation imply only rather weak probability bounds.

Our guiding criteria for the construction of estimators are discussed in § 2 centering, in particular, on the notion of *universality*. In § 3 we review briefly some known concepts such as complexity regularization as a means to realize universality. In particular, we apply this to piecewise polynomial estimators on isotropic and anisotropic partitions. We indicate how to increase in this case the efficiency of complexity regularization in the spirit of CART algorithms by exploiting the special additive structure of the objective functionals. § 4 is devoted to universal piecewise polynomial estimators based on adaptive partitioning by thresholding. On one hand, this complies better with online demands. On the other hand, in the piecewise constant case this gives rise to the desired optimal rates in probability. We indicate why this result does not carry over in full generality to the case of higher polynomial degrees and discuss circumstances under which optimal rates are retrieved. We conclude in § 5 with further possible ways of improving the results for piecewise polynomial estimators either by modifying the estimator or by estimating the performance for individual regression functions rather than for classes of such functions.

2 Guiding Criteria

A common approach to approximating the regression function is to choose an hypothesis (or *model*) class $\mathcal{H} = \mathcal{H}_N$ that offers N degrees of freedom, where $N = N(m)$ will typically depend on the sample size m and then look for elements $f_{\mathbf{z}} \in \mathcal{H}_N$ that approximate f_ρ as well as possible based on the knowledge of $\mathbf{z}$. The construction of $f_{\mathbf{z}}$ should take two aspects into account that typically work against each other, namely (I) *computational efficiency* - one has to handle possibly very large data sets - and (II) quality. According to the preceding discussion, quality can be expressed in terms of estimates for the error $\|f_\rho - f_{\mathbf{z}}\|$ which itself is a random variable. So one can ask for decay rates of the quantities

$$\mathbb{P}\{\|f_\rho - f_{\mathbf{z}}\| \geq \eta\}, \quad \eta > 0 \quad \text{or} \quad \mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2) \tag{2.1}$$

as the sample size m increases.

Of course, concrete rates for either quantity can only be expected under some assumptions on f_ρ. A typical assumption is that f_ρ belongs to some *compact* subset Θ of $L_2(X,\rho_X)$ often referred to as a *prior* on f_ρ.

Compactness can be described in various ways, e.g. by the asymptotic behavior of *entropy* or *covering* numbers which is a measure of the metric thickness of a set. Another way is to impose smoothness on f_ρ relative to $L_2(X,\rho_X)$. However, since ρ_X is unknown it is not clear what this means. A third way, which is the one adopted here, is to characterize compactness through *approximability*. To explain this, we think of any given hypothesis class $\mathcal{H}$ as a collection of functions on X that can be described by $d(\mathcal{H})$ parameters – degrees of freedom. When $\mathcal{H}$ is a linear space one would have $d(\mathcal{H}) = \dim \mathcal{H}$. Now given a family of such sets $\{\mathcal{H}\}$ we consider the corresponding approximation classes

$$\begin{aligned}\mathcal{A}^s &= \mathcal{A}^s(\{\mathcal{H}\}) := \{f \in L_2(X,\rho_X) : \inf_{d(\mathcal{H})\le N}\inf_{g\in\mathcal{H}} \|f-g\| =: \sigma_N(f,\{\mathcal{H}\}) \le CN^{-s} \\ &\qquad \text{for some } \ C<\infty\}. \end{aligned} \tag{2.2}$$

For a given f the infimum of all constants C for which the above bound holds is $|f|_{\mathcal{A}^s} := \sup_{N\in\mathbb{N}} N^s\sigma_N(f,\{\mathcal{H}\})$ which is a (quasi-) seminorm and $\|\cdot\|_{\mathcal{A}^s} = \|\cdot\| + |\cdot|_{\mathcal{A}^s}$ defines a quasi-norm for the space $\mathcal{A}^s$. Clearly for each $s>0$ any bounded subset of $\mathcal{A}^s$ is compact in $L_2(X,\rho_X)$.

Of course, the space $\mathcal{A}^s(\{\mathcal{H}\})$ depends on the collection $\{\mathcal{H}\}$ of hypothesis spaces and, as we shall point out later, in some cases $\mathcal{A}^s$ can also be described by regularity properties. One expects that the richer $\{\mathcal{H}\}$ is the larger is the class $\mathcal{A}^s(\{\mathcal{H}\})$, i.e. the more functions can be recovered at a given rate N^{-s} using at most N degrees of freedom within that framework. An important distinction is when for each N there exists at most one class $\mathcal{H} = \mathcal{H}_N$ in $\{\mathcal{H}\}$ with $d(\mathcal{H}) = N$ and $\mathcal{H}_N$ is a *linear space*. The corresponding approximation method (of taking the $L_2(X,\rho_X)$-orthogonal projection from $\mathcal{H}_N$) is then called *a linear approximation method*. This is to be contrasted with the case where several competing choices of $\mathcal{H}$ each being determined by N parameters. The selection among all the competing equally complex candidates that minimizes the projection error is then *nonlinear* since it depends on the particular approximant. Nonlinear methods will play an important role in what follows.

At any rate, whatever the collection $\{\mathcal{H}\}$ is, one faces two questions: When f_ρ belongs to some class $\mathcal{A}^s(\{\mathcal{H}\})$, i.e. it can be approximated in $\|\cdot\|$ by elements from some $\mathcal{H}$ with $d(\mathcal{H}) \le N$ to accuracy N^{-s},

(i) what is the best decay rate of the quantities in (2.1) when the sample size m grows?

(ii) how can one construct estimators $f_{\mathbf{z}}$ that match this rate?

As we shall see, the difficulties to be faced for (ii) depend very much on how much information one is willing to assume about f_ρ. For instance, is the estimator allowed to use s as an active parameter to find a good compromise between goodness of fit and variance? Of course, in most practical situations one would not know beforehand the prior class Θ, or values of s for which $f_\rho \in \mathcal{A}^s$. So a more refined question would be:

How to construct $f_{\mathbf{z}}$ so that it recovers the best possible rate for a possibly large range of $s > 0$ without using s in the actual algorithm?

An estimation scheme with this latter property is called *universal* and universality is a central issue throughout the subsequent discussion. We shall briefly review next two known concepts that reflect the essence of the problem.

3 Some Known Results on Estimates in Expectation

3.1 Linear Methods

When $\mathcal{H} = \mathcal{H}_N$ is a *linear* space of dimension N, a natural way to build an estimator is to mimic (1.4), i.e. to choose $f_{\mathbf{z}}$ as the minimizer of the empirical risk

$$\tilde{f}_{\mathbf{z}} := \operatorname*{argmin}_{f\in\mathcal{H}} \mathcal{E}_{\mathbf{z}}(f), \quad \text{with} \quad \mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m}\sum_{j=1}^{m}(y_j - f(x_j))^2. \tag{3.1}$$

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

$$\|g\|_m^2 := \frac{1}{m}\sum_{j=1}^{m}|g(x_j)|^2. \tag{3.2}$$

So the computation of $\tilde{f}_{\mathbf{z}}$ is essentially reduced to solving (possibly large) linear systems.

While f_ρ is known to be bounded by M, such a least squares fit might very well give rise to approximations violating this bound significantly, a point that will be taken up later again. Such a violation is a serious obstruction to the analysis of the performance of such estimators which typically involves concentration inequalities requiring L_∞ bounds. Therefore, one applies a truncation step as a postprocessing to obtain

$$f_{\mathbf{z}} := T_M(\tilde{f}_{\mathbf{z}}), \quad \text{with} \quad T_M(g) := \operatorname{sgn}(g)\min\{M, |g|\}. \tag{3.3}$$

For estimators of this type the following general result can be found in [13].

Theorem 3.1 *For an arbitrary linear space $\mathcal{H}_N$ of dimension N and $f_{\mathbf{z}}$ defined by* (3.3)*, one has* [1]

$$\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2) \lesssim \frac{N\log m}{m} + \inf_{g\in\mathcal{H}_N}\|f_\rho - g\|^2. \tag{3.4}$$

The first term in this bound reflects the uncertainty due the variance of the data while the second one describes the approximability of f_ρ. Clearly, the bound is minimized by equilibrating variance and bias. Hence whenever f_ρ belongs to $\mathcal{A}^s(\{\mathcal{H}_N\}_N)$ for some $s > 0$,

[1] Here and later we use the notation $A \lesssim B$ to mean that $A \le CB$ with C a constant that does not depend on the parameters involved unless explicitly stated. Similarly $A \sim B$ means $A \le CB$ and $B \le CA$.

the term $\inf_{g\in\mathcal{H}_N}\|f_\rho - g\|^2$ can be bounded by $|f_\rho|^2_{\mathcal{A}^s}N^{-2s}$. These bounds are balanced when choosing $N \sim (m/\log m)^{1/2s+1}$ which yields

$$\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2) \lesssim \left(\frac{\log m}{m}\right)^{\frac{2s}{2s+1}}. \tag{3.5}$$

The rate shown in (3.5) is actually (up to logarithmic factors) *best possible*, see [11]. However, to realize it, one needs to know s in order to choose the right dimension $N = N(m,s)$. So if this knowledge is not available - as would be usually the case in practice - an improper choice of N would give rise to an unsatisfactory performance of the estimator which is clearly not universal. Note that in the above situation the reference approximation method is linear.

3.2 Nonlinear Methods - Model Selection

The price to be paid for obtaining universality is to employ *nonlinear* estimation schemes. A widely used and very flexible paradigm is *model selection*. Instead of an a-priori prescription of the trial space $\mathcal{H}$ (*independent* of the data) one allows the estimator to select from a class $\mathcal{M}_m = \{\mathcal{H}\}$ of "models" where the class $\mathcal{M}_m$ depends on the data size m. A common way of organizing data dependent selection is through *complexity regularization*. The complexity or richness of each hypothesis space $\mathcal{H}_{p,m}$ is described by a *penalty term* $\mathrm{pen}_m(\mathcal{H})$. For each class one takes again the truncated least squares estimator

$$f_{\mathcal{H},m} := T_M\left(\operatorname*{argmin}_{g\in\mathcal{H}}\left(\frac{1}{m}\sum_{i=1}^m (g(x_i)-y_i)^2\right)\right), \tag{3.6}$$

and then selects one of them by

$$f_{\mathbf{z}} := \operatorname*{argmin}_{\mathcal{H}\in\mathcal{M}_m}\left\{\frac{1}{m}\sum_{i=1}^m (f_{\mathcal{H},m}(x_i)-y_i)^2 + \mathrm{pen}_m(\mathcal{H})\right\}. \tag{3.7}$$

The following theorem ([13, Thm. 12.1]) gives sufficient conditions for this estimator to exhibit optimal performance. In its formulation $\mathcal{H}^+$ denotes the set of subgraphs of the functions in $\mathcal{H}$ and $V_{\mathcal{H}^+}$ denote its VC-dimension. Note that when $\mathcal{H}$ is a linear space one has $V_{\mathcal{H}^+} \le \dim\mathcal{H} + 1$, see e.g. [13].

Theorem 3.2 *Suppose*

$$\mathrm{pen}_m(\mathcal{H}) \gtrsim \frac{(\log m)V_{\mathcal{H}^+} + \frac{c_{\mathcal{H}}}{2}}{m}, \quad \mathcal{H}\in\mathcal{M}_m, \tag{3.8}$$

where $c_{\mathcal{H}} > 0$ satisfies

$$\sum_{\mathcal{H}\in\mathcal{M}_m} e^{-c_{\mathcal{H}}} \le 1. \tag{3.9}$$

Then,

$$\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2) \le 2\inf_{\mathcal{H}\in\mathcal{M}_m}\left\{\inf_{g\in\mathcal{H}_{p,m}}\|f_\rho - g\|^2 + \mathrm{pen}_m(p)\right\} + \frac{c}{m}. \tag{3.10}$$

Note that the penalization is not controlled by some smoothness measure but solely by the complexity of the corresponding model. To illustrate the role of the various ingredients it is instructive to consider the following example.

3.2.1 Piecewise Polynomials on Adaptive Partitions

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 [2].

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

$$\#(\mathcal{T}_0) \leq \#(\Lambda) \leq 2^d \#(\mathcal{T}_0). \tag{3.11}$$

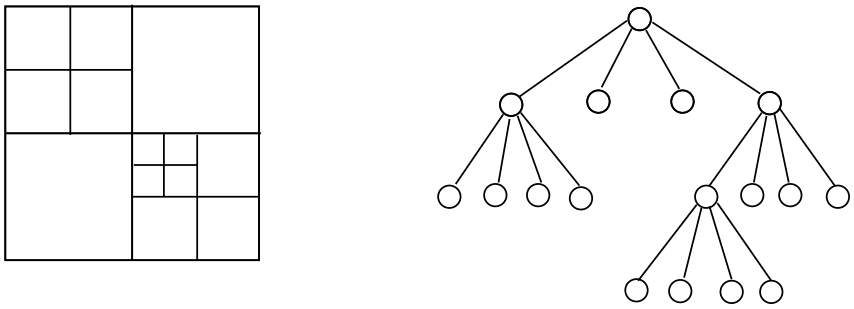

Figure 3.1: Local mesh refinement

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 the corresponding uniform partition Λ_j has the same measure 2^{-jd}. Another way of generating partitions is through some possibly *local* 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*.

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:

$$p_I(f) := \operatorname*{argmin}_{p \in \Pi_K} \|f - p\|, \tag{3.12}$$

where Π_K is the space of polynomials of degree at most K in d variables.

Given $K \in \mathbb{N}$ and a partition Λ of X, 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 in the

form

$$S = \sum_{I \in \Lambda} p_I \chi_I, \quad p_I \in \Pi_K, \tag{3.13}$$

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^K_Λ. The *best approximation* to f in this space is given by

$$P_\Lambda f := \sum_{I \in \Lambda} p_I(f) \chi_I, \quad p_I(f) := \operatorname*{argmin}_{p \in \Pi_K} \|f - p\|. \tag{3.14}$$

This suggests a natural discrete counterpart as an estimator from $\mathcal{H} = \mathcal{S}^K_\Lambda$. Given the data $\mathbf{z}$ and any Borel set $I \subset X$, we define

$$p_{I,\mathbf{z}} := \operatorname*{argmin}_{p \in \Pi_K} \frac{1}{m} \sum_{i=1}^{m} (p(x_i) - y_i)^2 \chi_I(x_i). \tag{3.15}$$

When there are no x_i in I, we set $p_{I,\mathbf{z}} = 0$. Moreover, for any partition Λ of X we define the estimator $f_{\mathbf{z}}$ as

$$f_{\mathbf{z}} := f_{\mathcal{S}^K_\Lambda} := \sum_{I \in \Lambda} T_M(p_{I,\mathbf{z}}) \chi_I \tag{3.16}$$

with T_M the truncation operator defined earlier. Note that this requires solving (in parallel if needed) only small least squares problems of fixed size dim Π_K. Moreover, the empirical minimization (3.15) is not done over the truncated polynomials, since this is numerically much more delicate and expensive. Instead, truncation is only used as a post processing. Note that $V_{(\mathcal{S}^K_\Lambda)^+} \lesssim_{K,d} \#\Lambda$.

In this framework a simple class $\mathcal{M}_m$ of models would comprise the spaces $\mathcal{S}^K_{\Lambda_j}$ $j \le \bar{j}$, which is a hierarchy of spaces of piecewise polynomials on uniform partitions up to some data dependent level $\bar{j}$ of resolution. In this case, one can choose the penalty weights $c_{\mathcal{S}^K_{\Lambda_j}} \sim \#\Lambda_j \sim 2^{jd}$ (with a constant depending on K and d) so that (3.9) is satisfied and Theorem 3.6 applies, see [13, Theorem 12.1].

A richer model class is to consider any partition of X into dyadic cubes generated by a tree of limited depth, namely $\mathcal{M}_m := \{\mathcal{S}^K_\Lambda : \Lambda(\mathcal{T}), \mathcal{T}$ a tree of bounded depth $\bar{j}\}$. Counting the number of possible partitions of a given size, one can continue to use the penalty terms as above. Namely, one can choose $\mathrm{pen}_m(\mathcal{S}^K_\Lambda) \sim \frac{\#\Lambda \ \log m}{m}$ to obtain, upon balancing the approximation error bound and the penalty, the following immediate consequence of Theorem 3.2.

Corollary 3.3 *Let $\gamma > 0$ be arbitrary and let $j_0 = j_0(\gamma, m)$ be defined as the smallest integer j such that $2^{-jd} \le (\log m/m)^{1/2\gamma}$. Consider the set $\mathcal{M}_m := \{\mathcal{S}^K_\Lambda\}$ corresponding to partitions Λ induced by proper trees $\mathcal{T} \subset \cup_{j \le j_0} \Lambda_j$. Then, there exists $\kappa_0 = \kappa_0(d, K)$ such that if*

$$\mathrm{pen}_m(\mathcal{S}^K_\Lambda) = \frac{\kappa \log m}{m} \#(\Lambda)$$

for some $\kappa \geq \kappa_0$, the estimator $f_{\mathbf{z}}$: defined by 3.7 *satisfies*

$$\mathbb{E}(\|f_\rho - f_{\mathbf{z}}\|^2) \leq C\Big(\frac{\log m}{m}\Big)^{\frac{2s}{2s+1}}, \quad m = 1, 2, \ldots, \tag{3.17}$$

whenever $f_\rho \in \mathcal{A}^\gamma(\{\mathcal{S}^K_{\Lambda_j}\}_{j\in\mathbb{N}}) \cap \mathcal{A}^s(\{\mathcal{S}^K_\Lambda\}_{\#\Lambda\leq N, N\in\mathbb{N}})$. Here, the constant C depends on

$$\kappa, M, |f_\rho|_{\mathcal{A}^s(\{\mathcal{S}^K_\Lambda\}_{\#\Lambda\leq N,N\in\mathbb{N}})}, |f_\rho|_{\mathcal{A}^\gamma(\{\mathcal{S}^K_{\Lambda_j}\}_{j\in\mathbb{N}})},$$

but not on m.

The assumption on f_ρ that guarantees the above rate in expectation consists of two parts. First $f_\rho \in \mathcal{A}^\gamma(\{\mathcal{S}^K_{\Lambda_j}\}_{j\in\mathbb{N}})$ means that for the above γ, which could be taken arbitrarily small, f_ρ should be approximable at that rate by a *linear method* based on a hierarchy of *uniformly* refined piecewise polynomials. The algorithm depends on that γ through the choice of the largest tree depth j_0. The smaller γ, the larger j_0 and the larger the computational effort. Nevertheless, no precise knowledge of γ is needed to achieve an optimal rate. Its choice only affects the *range* of those functions for which an optimal rate is attained by the scheme. The approximability of f_ρ by the nonlinear model class expressed by $f_\rho \in \mathcal{A}^s(\{\mathcal{S}^K_\Lambda\}_{\#\Lambda\leq N,N\in\mathbb{N}})$ does *not* enter the scheme but determines the accomplished rate. In this sense the scheme is universal for the range of rates attainable by piecewise polynomials on adaptive partitions.

To illustrate the meaning of the assumption $f_\rho \in \mathcal{A}^s(\{\mathcal{S}^K_\Lambda\})$ let us consider the special case where ρ_X is equivalent to the Lebesgue measure. In this case it is known that

$$B^{sd}_p(L_p) \;\hookrightarrow\; \mathcal{A}^s \;\text{ as long as }\; \frac{1}{p} < s + \frac{1}{2}, \quad s \leq k.$$

Thus the smoothness that is needed to guarantee a certain adaptive approximation rate in L_2 is significantly weaker than the corresponding smoothness measured also in L_2 that would guarantee the same rate if linear methods were used.

This also hints at the problems that will be faced when dealing with large spatial dimensions d, often referred to as *Curse of Dimension*. Realizing a rate s requires larger and larger smoothness ds, or in other words, the computational effort required by schemes based on such refinement strategies grows exponentially in d. Methods based on dyadic partitions are therefore not suitable for large d.

Note that the above approach yields estimates in expectation. We shall see later that it fails in general to provide the sharper estimates in probability mentioned in (2.1). A further major drawback of this very versatile principle is that it is often extremely (and sometimes even prohibitively) expensive from a computational point of view, and therefore far from being compatible with online demands. However, for the above particular example things are slightly different as will be pointed out next.

3.2.2 Adaptive Splits Using CART

One can turn to even richer model classes than those obtained as above by (isotropic) dyadic splits, while still maintaining essentially the same conclusions. Instead of subdividing cells in all directions one can consider anisotropic splits e.g. by halving cells with the aid of d hyperplanes that are orthogonal to one of the d coordinate axes.

Any such partition of X will be denoted by $\Lambda(X)$ or briefly Λ. Accordingly, for any subset $I \subset X$ we sometimes write $\Lambda(I)$ to indicate that we are dealing with a partition of I. If we decide to refine a cell I we have d different choices correpsonding to partitions $I = I_j^0 \cup I_j^1$, $j = 1, \dots, d$. We shall say a partition Λ is *admissible* if it is obtained by refinement which at each applications replaces a cell I by one of the pairs $\{I_j^0, I_j^1\}$. We denote by $\hat{\mathcal{P}}(I)$ the set of all admissible partitions $\Lambda = \Lambda(I)$ of I and write briefly $\hat{\mathcal{P}} = \hat{\mathcal{P}}(X)$.

Any admissible partition Λ can be identified with a *labelled binary tree* $\mathcal{T}(\Lambda)$ (with root I when $\Lambda \in \hat{\mathcal{P}}(I)$), where the label j at each node indicates that the children are obtained by a j-split. Thus we could label all elements in Λ as $I_{\mathbf{j}}^{\mathbf{e}}$ where, for $\mathbf{e} = (e_1, \dots, e_n)$, $e_i \in \{0, 1\}$, $\mathbf{j} = (j_1, \dots, j_n) \in \{1, \dots, d\}^n$ we have

$$I_{j_1,\dots,j_{n-1}}^{e_1,\dots,e_{n-1}} = I_{j_1,\dots,j_{n-1},j_n}^{e_1,\dots,e_{n-1},0} \cup I_{j_1,\dots,j_{n-1},j_n}^{e_1,\dots,e_{n-1},1} \tag{3.18}$$

i.e. I has resulted from n successive splits of X. The vector $\mathbf{j}$ encodes the type of splits used along the way and $\mathbf{e}$ records which of the two children have been used at each prior stage.

To obtain a numerically feasible scheme we shall always restrict $\hat{\mathcal{P}}$ (and hence the $\hat{\mathcal{P}}(I)$) to a finite set obtained by requiring that in the split history $\mathbf{j}$ of any of its cells each split type j may appear at most j_0 times, i.e. the highest spatial resolution is again bounded.

We now take $\mathcal{M}_m$ as the set of all $\mathcal{H} = \mathcal{S}_\Lambda^K$ with the above restrictions on the depth of Λ. Note that $\dim \mathcal{S}_\Lambda^K = (\#\Lambda)\binom{k+d}{d}$ and hence is proportional to $\#\Lambda$. We shall define the penalty function $\mathrm{pen}(\mathcal{H})$ and the constants c_Λ for $\mathcal{H} = \mathcal{S}_\Lambda^K \in \mathcal{M}_m$ exactly as before. Namely, $c_\Lambda = c^* \#(\Lambda)$ (where the constant c^* is yet to be fixed) and

$$\mathrm{pen}_m(\mathcal{H}) \geq \frac{c(\log m)\#\Lambda + c_\Lambda}{m}. \tag{3.19}$$

To verify (3.9) we note that the binary tree corresponding to a partition Λ with $N + 1 = \#\Lambda$ cells has N interior nodes (nodes that are not leaves) each of which can be labelled in d ways. Let $t(N)$ be the number of binary trees with $2N + 1$ nodes (hence N interior nodes). Then the number of possible partitions with N cells is given by $t(N)d^N$. Moreover, it is known that $t(N) = (N+1)^{-1}\binom{2N}{N} \lesssim 4^n/n^{3/2}$, see e.g. [6]. Hence, one can still ensure that

$$\sum_{\mathcal{H} \in \mathcal{M}_m} e^{-c_\mathcal{H}} \leq C' \sum_{N=1}^{\infty} d^N 4^N N^{-3/2} e^{-\gamma N} \leq 1 \tag{3.20}$$

for c^* sufficiently large. We can therefore apply Theorem 3.2 with $f_{\mathcal{H},m}$ and $f_{\mathbf{z}}$ defined as in (3.6) and (3.7) respectively and arrive at the analogue of Corollary 3.3.

Let us briefly point out next that, due to the particular structure of the cost functional, complexity regularization can in this case of adaptive piecewise estimators (in the previous setting of isotropic refinements as well as in the more general framework of anisotropic refinements) be realized in a relatively efficient way, see [12]. It suffices to explain this for the deterministic setting and for anisotropic splits.

According to (3.20) we can take $c_\Lambda = c^* \# \Lambda$ and

$$\mathrm{pen}_m(\Lambda) = \mu \# \Lambda, \tag{3.21}$$

where $\mu := \mu(m) := \kappa \log m/m$. For any admissible cell I and any admissible partition Λ of I, consider the *local objective functional*

$$\phi(\Lambda|I) := \|f_\rho - P_\Lambda(f_\rho)\|^2_{L_2(\rho_X, I)} + \mathrm{pen}_\mu(\Lambda).$$

The minimization of such functionals will greatly benefit from the above form of the penalty term which entails the following *additive* structure. Suppose that $I = I' \cup I''$ and Λ is a partition for I whose restrictions to I', I'' are denoted by Λ', Λ'', respectively. Then,

$$\phi(\Lambda|I) = \phi(\Lambda'|I') + \phi(\Lambda''|I''). \tag{3.22}$$

In order to find a partition Λ^* that minimizes $\phi(\Lambda|X)$ one can proceed as follows. Consider for any admissible cell I the local square error

$$e_I := \int_I (f_\rho - p_I(f_\rho))^2 d\rho_X = \inf_{p \in \Pi_k} \int_I (f_\rho - p)^2 d\rho_X.$$

The key observation is how to build from locally optimal partitions on pairs of cells an optimal partition on the parent cell. Let

$$w(I) := \phi(\Lambda^*(I,\mu)|I) \quad \text{with} \quad \Lambda^*(I,\mu) = \operatorname*{argmin}_{\Lambda \in \hat{\mathcal{P}}(I)} \phi(\Lambda|I). \tag{3.23}$$

Remark 3.4 *For any I, we have*

$$w(I) = e_I + \mu, \qquad \Lambda^*(I,\mu) = \{I\} \tag{3.24}$$

if and only if

$$e_I + \mu \le w(I^0_{i^*}) + w(I^1_{i^*}) := \min_{i=1,\dots,d} w(I^0_i) + w(I^1_i). \tag{3.25}$$

Moreover, if $w(I) \ge w(I^0_{i^*}) + w(I^1_{i^*})$, then

$$w(I) = w(I^0_{i^*}) + w(I^1_{i^*}), \qquad \Lambda^*(I,\mu) = \Lambda^*(I^0_{i^*},\mu) \cup \Lambda^*(I^1_{i^*},\mu). \tag{3.26}$$

Also

$$e_I \le \mu \quad \Longrightarrow \quad \Lambda^*(I,\mu) = \{I\}. \tag{3.27}$$

Proof: The equivalence of (3.24) and (3.25) as well as (3.26) follow immediately from the additivity property (3.22). Since $\min_{i=1,\dots,d} w(I^0_i) + w(I^1_i) \ge 2\mu$, (3.27) implies the validity of (3.25) □

In principle, this shows how to minimize $\phi(\Lambda|X)$ over $\Lambda \in \hat{\mathcal{P}}$. Finding the true optimum seems to require knowing all the e_I, for any I that may appear in an element of $\hat{\mathcal{P}}$. The computation of these quantities (or later their empirical counterparts) can be organized from coarse to fine, picturing all possible partitions in a d-dimensional array

of copies of X where in direction $i \leq d$ only a refinement of the ith coordinate takes place. The advantage is that whenever some I is encountered for which $e_I \leq \mu$ this cell will, according to (3.27), never be refined. This may reduce the amount of computations needed in the whole process.

Once the e_I have been computed one can start pruning from the highest level downward to coarser levels. Let $\hat{\Lambda}_{j_0}$ denote the uniform dyadic partition of level j_0 of X. Clearly, any $J \in \hat{\Lambda}_{j_0}$ has a sibling J' such that $J \cup J' = I$ is a cell in some $\Lambda' \in \hat{\mathcal{P}}(X)$. $w(J), w(J')$ being known we know from Remark 3.4 how to form $\Lambda^*(I, \mu)$. Successively merging lower level siblings from optimal higher level partitions eventually yields $\Lambda^*(X, \mu)$.

In the case of isotropic refinements one has to deal, of course, with significantly fewer comparisons to determine local optimality. Let us also remark that in this case the search of the optimal partition Λ^* is known to be performed at a reasonable computational cost using a CART algorithm (see e.g. [8] or [12]).

Notice that the estimator did not need to have knowledge of s and nevertheless obtains the optimal performance. For a certain restricted range of s, one can actually prove similar estimates also in probability (see [11]).

All the above strategies involve postprocessing a least squares estimator by a truncation so that the estimator is in general no longer an element of the approximation classes $\mathcal{H}$ under consideration. This can be avoided by another approach developed in [14] and also discussed in [13]. The least squares procedure there is confined to the intersection of the approximation class $\mathcal{H}$ with some L_∞-ball. On one hand, it is then possible to establish optimal rates in expectation and probability for an estimator that remains in the chosen approximation class. On the other hand, one has to perform now a quadratic minimization under an L_∞ state constraint which is a numerically much more demanding task than solving the linear problems yielding the unconstrained least squares estimator. Moreover, it is not clear which L_∞ bound should be chosen so as to avoid an unnecessary degrading of the approximation properties of the estimator. Therefore we confine the subsequent discussion to unconstrained least squares estimators.

4 Adaptive Partitioning by Thresholding

Despite the possible computational speedup offered by CART in connection with complexity regularization, a principal limitation of this approach seems to be that, in general, it does not lead to optimal estimates in probability, see (2.1) - a fact that will become clearer later.

In the case of isotropic partitioning, an alternative is offered by adaptive partitioning based on thresholding techniques as proposed recently in [2, 3]. Let us briefly recall the main ingredients of this approach adhering to the above notation for isotropic refinements. In particular, we denote again by Λ_j the uniform partition of level j, giving rise to the approximation spaces $\mathcal{A}^s(\{\mathcal{S}^K_{\Lambda_j}\}_{j\in\mathbb{N}})$.

We shall consider adaptive partitions that are obtained from a refinement criterion that is motivated by adaptive wavelet constructions such as those given in [9] for image

compression. Given a function $f \in L_2(X, \rho_X)$, we define the local atoms

$$\psi_I(f) := \sum_{J \in \mathcal{C}(I)} p_J(f)\chi_J - p_I(f)\chi_I, \quad I \neq X, \quad \psi_X(f) := p_X(f), \tag{4.1}$$

and

$$\epsilon_I(f) := \|\psi_I(f)\|. \tag{4.2}$$

Clearly, we have

$$f = \sum_{I \in \mathcal{D}} \psi_I(f), \tag{4.3}$$

and since the ψ_I are mutually orthogonal, we also have

$$\|f\|^2_{L_2(X,\rho_X)} = \sum_{I \in \mathcal{D}} \epsilon_I(f)^2. \tag{4.4}$$

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

$$|f|^p_{\mathcal{B}^s} := \sup_{\eta > 0} \eta^p \#(\mathcal{T}(f, \eta)). \tag{4.5}$$

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

$$\|f - P_{\Lambda(f,\eta)} f\| \leq C_s |f|^{\frac{1}{2s+1}}_{\mathcal{B}^s} \eta^{\frac{2s}{2s+1}} \leq C_s |f|_{\mathcal{B}^s} N^{-s}, \quad N := \#(\mathcal{T}(f, \eta)), \tag{4.6}$$

where the constant C_s depends only on s. The proof of this estimate can be based on the same strategy as used in [9] where a similar result is proven in the case of the Lebesgue measure.

Invoking (3.11), 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 $\#(\Lambda) = N$, i.e. $\mathcal{B}^s \subseteq \mathcal{A}^s(\{\mathcal{S}^K_\Lambda\}_{\#\Lambda \leq N})$. This should be contrasted with $\mathcal{A}^s = \mathcal{A}^s(\{\mathcal{S}^K_{\Lambda_j}\}_{j \in \mathbb{N}})$ 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 [9]). 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.

4.1 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 one can

use an empirical version of this adaptive procedure based on the estimator given by (3.15) and (3.16).

Our adaptive partitions are based now on an empirical analogue of the ϵ_I. For each cell I in the master tree $\mathcal{T}$, we define

$$\epsilon_I(\mathbf{z}) := \|T_M(\sum_{J\in\mathcal{C}(I)} p_{J,\mathbf{z}}\chi_J - p_{I,\mathbf{z}}\chi_I)\|_m, \tag{4.7}$$

where $\|\cdot\|_m$ is the empirical norm defined in (3.2).

To begin the description of the thresholding algorithm, we fix a parameter $\kappa > 0$ which will be described in more detail later. With κ in hand, we define the threshold

$$\tau_m := \kappa\sqrt{\frac{\log m}{m}}. \tag{4.8}$$

As before, 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

$$\Sigma(\mathbf{z}, m) := \{I \in \cup_{j\leq j_0}\Lambda_j \quad : \quad \epsilon_I(\mathbf{z}) \geq \tau_m\}. \tag{4.9}$$

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. The essential steps of the adaptive algorithm in the present setting read as follows:

Algorithm: *Given* $\mathbf{z}$*, choose* $\gamma > 0$*, and*

- *for* $j_0(m,\gamma)$ *determine the set* $\Sigma(\mathbf{z}, m)$ *according to (4.9);*
- *form* $\mathcal{T}(\tau_m, \mathbf{z}), \Lambda(\tau_m, \mathbf{z})$ *and compute* $f_{\mathbf{z}}$ *according to (3.16) for this partition.*

For further comments concerning the treatment of streaming data we refer to an analogous strategy outlined in [2].

The above algorithm has been analyzed in [2] in the case of *piecewise constant approximation* for which the following result could be established.

Theorem 4.1 *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*

$$\mathbb{P}\left\{\|f_\rho - f_{\mathbf{z}}\| \geq \tilde{c}\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\right\} \leq Cm^{-\beta}, \tag{4.10}$$

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

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 the desired estimate for the error $\|f_\rho - f_{\mathbf{z}}\|$ in probability which is somewhat stronger than the estimates obtained in the previous section. As mentioned earlier, from this one obtains a corresponding estimate in expectation, see e.g. [2]. The order of approximation can be shown to be optimal save for the logarithmic term by using the results on lower estimates from [11]. Finally, note that the role of the space $\mathcal{A}^\gamma$, as in Corollary 3.3, 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.

A natural question would be to have an analogous result for piecewise polynomial estimators of higher degree. In fact, the previous estimates in expectation made no distinction concerning the degree of the polynomials and allowed one to fully exploit the superior approximation power offered by higher degrees.

4.2 A Principal Obstruction

In this regard an important observation is that the analogue of Theorem 4.1 does not hold in full generality when piecewise polynomials of degree higher than zero are used in place of piecewise constants. This can be shown with the aid of a counterexample for which empirical risk minimization does not perform well in probability and whose essence is conveyed by the following figures 4.2 and 4.3 below.

Referring to [3] for the technical details, one first considers approximation by linear functions on the interval $X = [-1, 1]$ for the bound $M = 1$ where the data y_i are given exactly as samples of the piecewise contant function indicated in Figure 4.2. For each $m = 1, 2, \ldots$, we define a measure $\rho_X = \rho_{X,m}$ on $[-1, 1]$ is concentrated at the points $\pm 1, \pm\gamma$, namely

$$\rho_X := (1/2 - \kappa)(\delta_{-\gamma} + \delta_\gamma) + \kappa(\delta_{-1} + \delta_1), \tag{4.11}$$

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

$$y(\gamma) = 1,\ y(-\gamma) = -1,\ y(\pm 1) = 0, \text{ with probability 1.} \tag{4.12}$$

Therefore, there is no randomness in the y direction.

One can then show that the empirical least square minimizer

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

assumes with high probability a position indicated in Figure 4.3. In fact, the intersection of the estimator with the x-axis is shown to be at least 2γ away from the origin which implies an error of order one. More precisely, given any $\beta > 2$, there exist absolute constants $c, \tilde{c} > 0$ such that for each $m = 1, 2, \ldots$, the above distribution $\rho = \rho_m$ satisfies

$$\mathbb{P}\{\|f_\rho - T(\hat{p})\| \geq c\} \geq \tilde{c} m^{-\beta+1}. \tag{4.13}$$

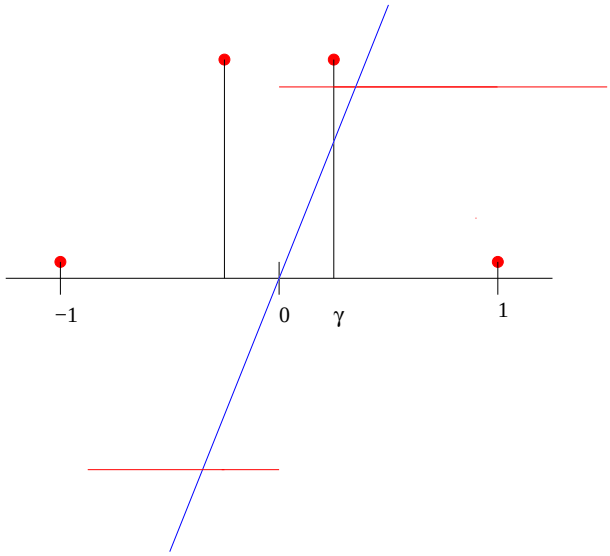

Figure 4.2: Best linear fit

One can now adjust the above situation to give information about piecewise linear approximation on adaptively generated dyadic partitions by rescaling. We let I be an interval at the finest scale allowed under our restrictions of depth search. If we allow dyadic partitions with more than m elements, we can approximate f_ρ exactly so that 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. For any of the corresponding bad draws $\mathbf{z}$ we will have

$$\|f_\rho - \hat{f}_\mathbf{z}\| \geq c \tag{4.14}$$

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 counterexample also indicates, however, that the arguments leading to optimal rates in expectation based on complexity regularization cannot be expected to be refined in general towards estimates in probability.

In view of these observations, the following two options suggest themselves. First, inspired by the above counterexample one can look for (hopefully weak) conditions on the measure ρ under which one might still get optimal rates for piecewise polynomial estimators of higher degree. Second, one can try to modify the estimators to cope with the type of obstructions suggested by the example. In fact, regarding the first option, one can show that, when imposing some restrictions on the marginal measure ρ_X, then high probability results turn out to be possible as we shall next describe.

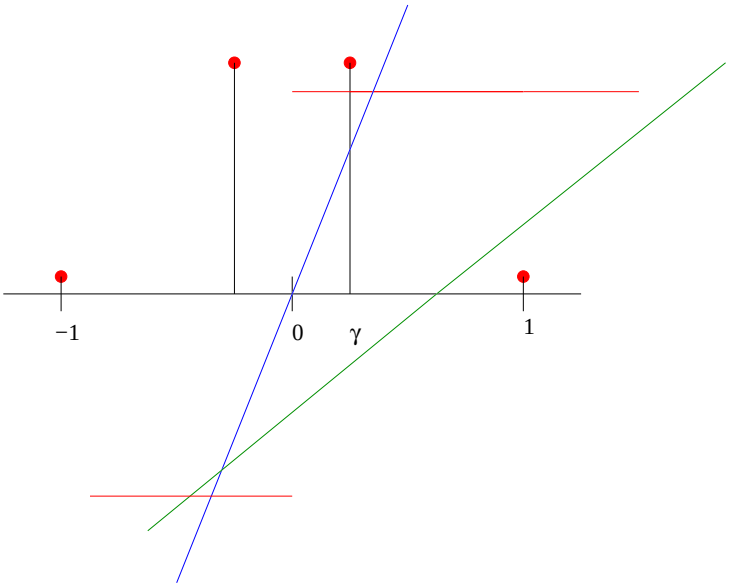

Figure 4.3: Estimator

4.3 The Case of Bounded Projections

For simplicity let us assume now that $M = 1$. The difficulty of bounding the projection $p_I(f_\rho)$ seems to prevent one from showing that

$$p_I := p_I(f_\rho) := \mathrm{Argmin}_{p\in\Pi_K}\|f_\rho - p\| \tag{4.15}$$

and its empirical counterpart

$$\hat{p}_I := T_1\Big(\mathrm{Argmin}_{p\in\Pi_K}\frac{1}{m}\sum_{i=1}^{m}\chi_I(x_i)(y_i - p(x_i))^2\Big), \tag{4.16}$$

are close with high probability. The closeness of these two quantities, however, already played an important role in the analysis of the piecewise constant case [2]. We shall show next that a favorable comparison of these two quantities indeed becomes possible when the projection $p_I(f_\rho)$ remains below some fixed bound

$$\|p_I\|_{L_\infty(I)} \le M'. \tag{4.17}$$

We say I is *good* if (4.17) holds.

Theorem 4.2 *Suppose I is good, i.e. (4.17) holds. Then there exist uniform constants c, c_1, c_2, depending only on M', K, d, such that $f_{\mathbf{z}}$ given by (4.16) satisfies*

$$\mathbb{P}\{\|f_\rho - \hat{p}_I\| > \delta\} \le c_1\delta^{-c_2}e^{-cm\delta^2}, \tag{4.18}$$

provided that

$$\delta \ge 32\|f_\rho - p_I\|. \tag{4.19}$$

Moreover, for any δ satisfying (4.19) one also has

$$\mathbb{P}\{\|p_I - \hat{p}_I\| > \delta\} \le c_1'\delta^{-c_2}e^{-cm\delta^2}. \tag{4.20}$$

The proof of this theorem is based on the following concentration result from [13, Theorem 11.2] that will play also an important role in the subsequent discussion.

Theorem 4.3 *Let $\mathcal{F}$ be any set of bounded functions f and consider the discrete norm*

$$\|f\|_{\mathbf{t}} := \frac{1}{2m}\sum_{j=1}^{2m}|f(t_j)|^2. \tag{4.21}$$

Let $\mathcal{N}(\mathcal{F},\eta,\mathbf{t})$ denote the covering number which is the smallest number of balls of radius η which cover $\mathcal{F}$ with respect to the norm $\|\cdot\|_{\mathbf{t}}$. Then one has

$$\mathbb{P}\left\{\|f\| - 2\|f\|_m > \eta \text{ for some } f\in\mathcal{F}\right\} \le 3e^{-\frac{m\eta^2}{288M^2}}\mathbb{E}\left(\mathcal{N}(\mathcal{F},\eta,\mathbf{t})\right). \tag{4.22}$$

Here the probability is with respect to $\mathbf{z}$ (note that $\|\cdot\|_m$ (see (3.2)) is a random variable since it depends on $\mathbf{z}$) and the expectation is with respect to $\mathbf{t}$.

It is well known that if V is a linear space of dimension q and $\mathcal{G} := \{T_M g : g\in V\}$ then

$$\mathcal{N}(\mathcal{G},\eta,\mathbf{t}) \le (C\eta)^{-(2q+1)}, \quad 0<\eta\le 1, \tag{4.23}$$

with $C = C(M)$ (see e.g. Theorems 9.4 and 9.5 in [13]).

Proof of Theorem 4.2: Given any sample set $\mathbf{z}$ as above let $\mathbf{x} := \{x_1,\dots,x_m\}$ be the x component of $\mathbf{z}$. We employ the empirical norm

$$\|f\|^2_{\mathbf{x},m} := \frac{1}{m}\sum_{i=1}^{m}|f(x_i)\chi_I(x_i)|^2,$$

imitating $\|\cdot\| = \|\cdot\|_{L_2(\rho_X,I)}$. In order to use results involving this empirical norm we write

$$\mathbb{P}\left\{\|p_{I,\mathbf{z}} - f_\rho\| > \epsilon\right\} \le \underbrace{\mathbb{P}\left\{\|p_{I,\mathbf{z}} - f_\rho\| - 2\|p_{I,\mathbf{z}} - f_\rho\|_{\mathbf{x},m} > \epsilon/2\right\}}_{=:P_1} + \underbrace{\mathbb{P}\left\{\|p_{I,\mathbf{z}} - f_\rho\|_{\mathbf{x},m} > \epsilon/4\right\}}_{=:P_2} \tag{4.24}$$

Denoting by $\lambda = \lambda(K,d)$ the dimension of Π_K, we can invoke Theorem 4.3 applied to the set $\mathcal{F}$ of functions $f_\rho - p_{I,\mathbf{z}}$. This gives

$$P_1 \;\lesssim\; \epsilon^{-\hat{c}\lambda}e^{-cm\epsilon^2}, \tag{4.25}$$

taking care of the first term on the right hand side of (4.24).

As for P_2, we write

$$P_2 = \int\left\{\mathbb{P}\{\|p_{I,\mathbf{z}} - f_\rho\|_{\mathbf{x},m} > \epsilon/4 \mid \mathbf{x}\}\right\}d\rho_X^m \tag{4.26}$$

and we bound the probability inside the integral as follows. For fixed $\mathbf{x} = \{x_1,\cdots,x_m\}$ we can write

$$y_i = f_\rho(x_i) + B_i, \tag{4.27}$$

where the B_i are independent random variables and such that $|B_i| \leq 1$ and $\mathbb{E}(B_i) = 0$. We denote by $\mathbf{y}$ and $\mathbf{B}$ the corresponding vectors comprised of those y_i, B_i for which $x_i \in I$. Now let $P_{\mathbf{x}} : \mathbb{R}^m \to \mathcal{H}(\mathbf{x})$, where $\mathcal{H}(\mathbf{x})$ is the space of traces of the elements in Π_K on $\mathbf{x}|_I$, be the $\|\cdot\|_{\mathbf{x},m}$-orthogonal projector onto $\mathcal{H}(\mathbf{x})$. In a slight abuse of notation we denote now by $f_\rho, p_{I,\mathbf{z}}$ also their traces on $\mathbf{x}|_I$ consisting of those x_i that belong to I. In these terms we can reexpress

$$p_{I,\mathbf{z}} := T_1 P_{\mathbf{x}}(f_\rho + \mathbf{B}) = T_1(P_{\mathbf{x}} f_\rho + P_{\mathbf{x}}\mathbf{B}). \tag{4.28}$$

Clearly the dimension of $\mathcal{H}(\mathbf{x})$ is at most λ. Next we make use of the following elementary observations

Remark 4.4 *Suppose that $|a| \leq 1$. Then one has*

$$|a - T_1(c+d)| \leq |a - T_3(c)| + |d|, \quad |a - T_3(b)| \leq |a - b|. \tag{4.29}$$

As a consequence of the first inequality in (4.29) we obtain with $a = f_\rho(x_i)$, $c = P_{\mathbf{x}} f_\rho$, $d = P_{\mathbf{x}}\mathbf{B}$, on account of (4.28),

$$\|f_\rho - p_{I,\mathbf{z}}\|_{\mathbf{x},m} \leq \|f_\rho - T_3 P_{\mathbf{x}} f_\rho\|_{\mathbf{x},m} + \|P_{\mathbf{x}}\mathbf{B}\|_{\mathbf{x},m}. \tag{4.30}$$

Moreover, by the second inequality in (4.29) we have

$$\|f_\rho - T_3 P_{\mathbf{x}} f_\rho\|_{\mathbf{x},m} \leq \|f_\rho - P_{\mathbf{x}} f_\rho\|_{\mathbf{x},m} \leq \|f_\rho - p_I(f_\rho)\|_{\mathbf{x},m},$$

where we have used the optimality of $P_{\mathbf{x}} f_\rho$ with respect to $\|\cdot\|_{\mathbf{x},m}$. Therefore we conclude from (4.30) that

$$\|f_\rho - p_{I,\mathbf{z}}\|_{\mathbf{x},m} \leq \|f_\rho - p_I(f_\rho)\|_{\mathbf{x},m} + \|P_{\mathbf{x}}\mathbf{B}\|_{\mathbf{x},m}. \tag{4.31}$$

Here we do not need any further truncation of $P_{\mathcal{H}} f_\rho$ because of (4.17) Setting

$$P_3 := \mathbb{P}\left\{\|f_\rho - p_I(f_\rho)\|_{\mathbf{x},m} > \epsilon/8\right\}, \tag{4.32}$$

and

$$P_4 := \int \mathbb{P}\left\{\|P_{\mathbf{x}}\mathbf{B}\|_{\mathbf{x},m} > \epsilon/8 \mid \mathbf{x}\right\} d\rho_X^m, \tag{4.33}$$

it follows that

$$P_2 = \mathbb{P}\left\{\|p_{I,\mathbf{z}} - f_\rho\|_{\mathbf{x},m} > \epsilon/4\right\} \leq P_3 + P_4. \tag{4.34}$$

As for P_3, we remark that since $\|f_\rho - p_I(f_\rho)\| \leq \epsilon/32$ (see (4.19)), it follows that

$$P_3 \leq \mathbb{P}\{\|f_\rho - p_I(f_\rho)\|_{\mathbf{x},m} - 2\|f_\rho - p_I(f_\rho)\| > \epsilon/16\}. \tag{4.35}$$

The function $F := f_\rho - p_I(f_\rho)$ is, by (4.17), bounded $|F(x)| \leq 1 + M'$. Invoking a symmetric version of Theorem 4.3 to conclude that

$$P_3 \leq C\epsilon^{-\bar{c}\lambda} e^{-cm\epsilon^2}, \tag{4.36}$$

which gives an exponential bound similar to P_1.

For estimating P_4, we fix $\mathbf{x} = \{x_1, \cdots, x_m\}$ and define $A^1, \cdots A^q$ an $\|\cdot\|_{\mathbf{x},m}$-orthonormal basis of $\mathcal{H}(\mathbf{x})$. Note that $q \leq \lambda$. We now have

$$\|P_{\mathbf{x}}\mathbf{B}\|_{\mathbf{x},m}^2 = \sum_{j=1}^{q} |\langle \mathbf{B}, A^j \rangle|^2, \tag{4.37}$$

and therefore

$$\mathbb{P}\{\|P_{\mathbf{x}}\mathbf{B}\|_m > \epsilon/8\} \leq \sum_{j=1}^{q} \mathbb{P}\{|\langle \mathbf{B}, A^j \rangle| \geq \frac{\epsilon}{8\sqrt{\lambda}}\}. \tag{4.38}$$

Now, we have $\langle \mathbf{B}, A^j \rangle = \frac{1}{m} \sum_{i=1}^{m} B_i A_i^j \chi_I(x_i)$. We apply the following version of Hoeffding's inequality : if $\zeta_1, \cdots, \zeta_m$ are independent variables such that $|\zeta_i| \leq M_i$ and $E(\zeta_i) = 0$ then

$$\mathbb{P}\{|\frac{1}{m} \sum_{i=1}^{m} \zeta_i| \geq \epsilon\} \leq 2e^{-2\frac{m\epsilon^2}{\frac{1}{m}\sum_{i=1}^{m} M_i^2}}. \tag{4.39}$$

Here $M = 1$, $\zeta_i = \chi_I(x_i) B_i A_i^j$ and $M_i = |A_i^j|$ so that $\frac{1}{m} \sum_{i=1}^{m} M_i^2 = 1$. Therefore

$$\mathbb{P}\{\|P_{\mathbf{x}}\mathbf{B}\|_m > \epsilon/8\} \leq 2\lambda e^{-\frac{m\epsilon^2}{32\lambda}}. \tag{4.40}$$

It follows that

$$P_4 \leq 2\lambda e^{-\frac{m\epsilon^2}{32\lambda}}. \tag{4.41}$$

Combining all these results, the assertion follows. □

4.4 Optimal Results in Probability under Regularity Assumptions on ρ_X

We shall exhibit next some conditions on the measure ρ that guarantee the validity of (4.17) for any cell I. As before, we fix the polynomial degree K and define the projector P_Λ for any dyadic partition Λ by (3.14). The example in Section 4.2 indicates that a strong concentration of ρ_X may cause steep slopes in the estimator and hence leads to large deviations. The following assumption prohibits such a strong concentration and ensures that the least squares projection is uniquely defined.

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 $(L_{I,k})_{k=1,\cdots,\lambda}$ of Π_K (with λ the algebraic dimension of Π_K) such that*

$$\|L_{I,k}\|_{L_\infty(I)} \leq C_A(\rho_X(I))^{-1/2}, \quad k = 1, \ldots, \lambda. \tag{4.42}$$

Hence on each I one has

$$p_I(f) = \sum_{k=1}^{\lambda} \langle f, L_{I,k} \rangle_{L_2(I,\rho_X)} L_{I,k}, \tag{4.43}$$

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

$$\|P_\Lambda f\|_{L_\infty} \leq \lambda C_A \|f\|_{L_\infty}, \tag{4.44}$$

i.e. the projectors P_Λ are bounded in L_∞ independently of Λ, [3].

It is readily seen that Assumption A holds when ρ_X is the Lebesgue measure dx or, more generally for $d\rho_X = \omega(x)dx$ where $0 < c \leq \omega(x) \leq C$. For further examples see [3].

Under Assumption A, one 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. The following result was established in [3] directly under Assumption A. Note that a slightly weaker estimate is given by (4.20) under a weaker assumption.

Theorem 4.5 ([3]) *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 $I \in \mathcal{D}$*

$$\mathbb{P}\{\|T_M(p_I)\chi_I - T_M(p_{I,\mathbf{z}})\chi_I\| > \eta\} \leq \tilde{c}e^{-cm\eta^2}, \tag{4.45}$$

where $\tilde{c} = 2(\lambda + \lambda^2)$, and the constant c in (4.45) depends on M and C_A and behaves like $(MC_A^2)^{-2}$.

From the basic estimate (4.45), 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 [2].

Remark 4.6 *Under Assumption A one has for any fixed integer $K \geq 0$, any partition Λ and $\eta > 0$*

$$\mathbb{P}\left\{\|f_{\mathbf{z},\Lambda} - T_M(P_\Lambda f_\rho)\| > \eta\right\} \leq C_0 N e^{-c\frac{m\eta^2}{N}}, \tag{4.46}$$

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

$$\mathbb{E}\left(\|f_{\mathbf{z},\Lambda} - T_M(P_\Lambda f_\rho)\|^2\right) \leq C\frac{N \log N}{m}, \tag{4.47}$$

similar to Corollary 2.2 of [2], with $C = C(M, C_A, \lambda)$.

Based on these findings one can derive now optimal approximation rates for post-truncated least squares estimators on uniform partitions Λ_j of the form

$$\mathbb{P}\left\{\|f_\rho - f_{\mathbf{z}}\| > (\tilde{c} + |f_\rho|_{\mathcal{A}^s})\Big(\frac{\log m}{m}\Big)^{\frac{s}{2s+1}}\right\} \leq Cm^{-\beta}, \tag{4.48}$$

where, however, the choice of the right dyadic refinement level hinges on the knowledge of s, so that these estimators are not universal. Therefore we focus in what follows on the adaptive counterpart given by the Algorithm in Section 4.1.

Learning on Adaptive Partitions

We now turn to an analysis of the adaptive algorithm defined in the §4.1. This entails extensions of Theorem 4.1 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 $2^{jd} \geq \tau_m^{-1/\gamma}$.

We continue to assume that the measure ρ satisfies Assumption A. One roadblock to having a self contained algorithm is the fact that the constant C_A is unknown to us. This has a simple remedy which is to enlarge the threshold somewhat. To illustrate this, let us take $\tau_m := \frac{\log m}{\sqrt{m}}$. Using this threshold, the same analysis as in §5 of [2] shows that this algorithm is universally consistent. Moreover, we have the following theorem (see [3]) for the performance of this algorithm.

Theorem 4.7 ([3]) *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*

$$\mathbb{P}\left\{\|f_\rho - f_{\mathbf{z}}\| \geq c\Big(\frac{\log m}{\sqrt{m}}\Big)^{\frac{2s}{2s+1}}\right\} \leq m^{-\beta}, \tag{4.49}$$

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

$$\mathbb{E}\left(\|f_\rho - f_{\mathbf{z}}\|^2\right) \leq C\Big(\frac{\log m}{\sqrt{m}}\Big)^{\frac{4s}{2s+1}} \tag{4.50}$$

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

$$\lim_{m\to+\infty} E(\|f_\rho - f_{\mathbf{z}}\|^2) = 0, \tag{4.51}$$

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

$$\lim_{m\to+\infty} \mathbb{P}\{\|f_\rho - f_{\mathbf{z}}\| > \epsilon\} = 0. \tag{4.52}$$

The same conclusion could be obtained for any threshold of the form

$$\tau_m := \kappa(m)\sqrt{\frac{\log m}{m}}, \tag{4.53}$$

where $\kappa(m)$ is a sequence which grows very slowly to $+\infty$. This would result in a slightly different logarithmic factor in the excess rates. In fact, one could actually take just $\tau_m := \kappa_0\sqrt{\frac{\log m}{m}}$ if C_A was known to us (see [3] for details).

The strategy for proving Theorem 4.7 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.

Lemma 4.8 *For any $\eta > 0$ and any element $I \in \Lambda_{j_0}$, one has*

$$\mathbb{P}\left\{\epsilon_I(\mathbf{z}) \leq \eta \text{ and } \epsilon_I \geq 8\lambda C_A \eta\right\} \leq \tilde{c}(1+\eta^{-C})e^{-cm\eta^2} \tag{4.54}$$

and

$$\mathbb{P}\left\{\epsilon_I \leq \eta \text{ and } \epsilon_I(\mathbf{z}) \geq 4\eta\right\} \leq \tilde{c}(1+\eta^{-C})e^{-cm\eta^2}, \tag{4.55}$$

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

The proof of Lemma 4.8 is rather different from the counterpart in [2] for the case of piecewise constants. It is based among other things on the concentration result given in Theorem 4.3, see also [13, Theorem 11.2].

5 Some Perspectives and Open Problems

We mention next two directions for improving the above results.

5.1 Proving probability results for piecewise polynomial approximation

We have seen that when an interval I is good in the sense of (4.17), we are able to meet our goal of directly estimating the performance of the empirical algorithm on I in probability. We want to show that in the case that I is not good, it is possible to find in I a good interval J and use this to construct an estimator which performs well in probability. We shall confine the discussion in what follows to the case $X = [0,1]$, $K = 1$ and $M = 1$. As a possible modification of estimators considered so far we shall consider the following two-step procedure: (i) Given I, find $J \subseteq I$ that is good in the sense of (4.17) for some fixed bound M'. (ii) Then construct an estimator based on samples contained only in J. More precisely, let

$$f_{I,\mathbf{z}} := T_M(\tilde{f}_{J,\mathbf{z}}), \qquad \tilde{f}_{J,\mathbf{z}} := \operatorname*{argmin}_{g \in \Pi_K} \frac{1}{m}\sum_{i=1}^{m}(g(x_i) - y_i)^2 \chi_J. \tag{5.1}$$

Let us first point out how it is possible to find a good interval J inside I, at least on a theoretical level. For any interval J, let

$$\rho_J := \int_X \chi_J(x) d\rho_X, \quad x_J := \int_X x\chi_J d\rho_X, \quad \xi_J := \frac{x_J}{\rho_J}. \tag{5.2}$$

Given I, we are going to create now a nested sequence of interval $J_0 = I \supset J_1 \ldots \supset J_k \ldots$. In the case I is good, this sequence consists of only the one interval $J_0 = I$. Given that J_k is already defined, if J_k is good we terminate the sequence. If J_k is not good, we let $J_{k+1} := \bar{J}_{k+1} \cap J_k$ where $\bar{J}_{k+1} = J_k \cap [\xi_{J_k} - |J_k|/2, \xi_{J_k} + |J_k|/2]$ is the interval centered at ξ_{J_k} with length $|J_k|/2$. In going further, we define $J_k' := J_k \setminus J_{k+1}$. Now either this sequence terminates in a good interval J_k or else $I = \{x_0\} \cup J_1' \cup J_2' \cdots$ where x_0 is some point from I. In the latter case, we define $k := \infty$.

We define $f_{I,\mathbf{z}}$ by (5.1) for the good interval we have extracted. To analyze the performance of $f_{I,\mathbf{z}}$, we first establish a bound for the measure of $\cup_{j=1}^{k} J_j'$, i.e. of the complement of the good interval. To do this we introduce some notation. For any two intervals $L \subset K$, we shall use the notation

$$E_K(L) := \|f_\rho - p_K\|_{L_2(L,\rho_X)} \tag{5.3}$$

which is the error in approximating f_ρ by p_K on the interval L.

Lemma 5.1 *For each $1 \le j < k$, we have*

$$\rho_X(J_j') \le (M'-1)^{-2}(E_{J_j}(J_j)^2 - E_{J_{j+1}}(J_{j+1})^2) \tag{5.4}$$

Proof: Since J_j is not good, we have $|p_{J_j}(x)| \ge M'$, $x \in J_j'$. Hence, $|f_\rho(x) - p_{J_j}(x)| \ge M'-1$ on this interval. If we square this and integrate, we get that

$$E_{J_j}(J_j)^2 - E_{J_{j+1}}(J_{j+1})^2 \ge E_{J_j}(J_j)^2 - E_{J_j}(J_{j+1})^2 = E_{J_j}(J_j')^2 \ge (M'-1)^2\rho_X(J_j') \tag{5.5}$$

which gives (5.4). □

Lemma 5.2 *If $J = J_k \subseteq I$ is good, then*

$$\rho_X(I \setminus J) \le \Big(M'-1\Big)^{-2}\|f_\rho - p_I\|^2_{L_2(I,\rho_X)}.$$

Proof: We have by definition

$$I = J_k \cup J_{k-1}' \cup \cdots \cup J_0', \qquad \rho_X(I \setminus J_k) = \sum_{j=0}^{k-1} \rho_X(J_j').$$

We infer from Lemma 5.1 that for $J = J_k$

$$\begin{aligned} \rho_X(I \setminus J) &\le (M'-1)^{-2}\sum_{j=0}^{k-1}(E_{J_j}(J_j)^2 - E_{J_{j+1}}(J_{j+1})^2) \\ &\le (M'-1)^{-2}E_{J_0}(J_0)^2 = (M'-1)^{-2}\|f_\rho - p_I\|^2. \end{aligned} \tag{5.6}$$

This completes the proof. □

With these estimates in hand, we have

$$\begin{aligned} \|f_\rho - f_{I,\mathbf{z}}\|^2_{L_2(I,\rho_X)} &= \|f_\rho - T_1(\tilde f_{J,\mathbf{z}})\|^2_{L_2(J,\rho_X)} + \|f_\rho - T_1(\tilde f_{J,\mathbf{z}})\|^2_{L_2(I\setminus J,\rho_X)} \\ &\le \|f_\rho - T_1(\tilde f_{J,\mathbf{z}})\|^2_{L_2(J,\rho_X)} + 4\rho_X(I \setminus J) \\ &\le \|f_\rho - T_1(\tilde f_{J,\mathbf{z}})\|^2_{L_2(J,\rho_X)} + 4(M'-1)^{-2}\|f_\rho - f_I\|^2_{L_2(I,\rho_X)} \\ &\le (\|f_\rho - T_1(\tilde f_{J,\mathbf{z}})\|_{L_2(J,\rho_X)} + 2(M'-1)^{-1}\|f_\rho - f_I\|_{L_2(I,\rho_X)})^2 \end{aligned}$$

Hence, whenever

$$\eta \ge \max\{4(M'-1)^{-1}\|f_\rho - p_I\|_{L_2(I,\rho_X)}, 64\|f_\rho - p_J(f_\rho)\|_{L_2(J,\rho_X)}\},$$

we can invoke Theorem 4.2, (4.18), to conclude that

$$\mathbb{P}\{\|f_\rho - f_{I,\mathbf{z}}\|_{L_2(I,\rho_X)} \geq \eta\} \leq \mathbb{P}\{\|f_\rho - T_1(\tilde{f}_{J,\mathbf{z}})\|_{L_2(J,\rho_X)} \geq \eta/2\} \leq c_1 2^{c_2} \eta^{-c_2} e^{-cm\eta^2/4}, \quad (5.7)$$

which indeed provides a concentration estimate of the desired kind for the modified estimator.

The above procedure, while interesting, is not an empirical algorithm. To bring this approach to completion, we would need a numerical procedure to identify the good interval J in I. The natural approach of replacing the above arguments with empirical quantities will fail due to the difficulty of estimating the quantity ξ_I with sufficiently high probability, regardless of the interval being good or not. Therefore it is not clear whether this line will ultimately be successful.

5.2 Improved Probability Results for Adaptive Piecewise Constant Approximation

We have shown that it is possible to give near optimal estimates in probability for the performance of piecewise constant adaptive algorithms on certain approximation classes. We want now to show that it is possible to improve these results and obtain results which are near optimal on individual regression functions rather than only classes.

The tool for obtaining these improved results is the idea of near-best adaptive tree approximation as studied in [5] for the deterministic case. This type of approximation studies all possible partitions that can be described by trees of the form we have been discussing. Given the data $\mathbf{z}$, we use the empirical local error estimators $e_I(\mathbf{z})$ defined in (4.7). In the case of piecewise constant approximation (as we are now discussing), the truncation operator T_M is not needed. For a partition Λ associated to such a tree we denote by $E_\Lambda := \sum_{I\in\Lambda} e_I^2$ the error of piecewise approximation by piecewise constants for this partition and by $E_{\Lambda,\mathbf{z}}$ the corresponding empirical error. The local error $e_I(\mathbf{z})$ satisfy

$$\sum_{I'\in\mathcal{C}(I)} e_{I'}(\mathbf{z})^2 \leq e_I^2(\mathbf{a}) \quad (5.8)$$

which is the subadditivity property needed to apply the results of [5].

Applying the algorithm of [5] to the empirical data yields a partition $\Lambda^* := \Lambda^\star(\mathbf{z})$ which satisfies

$$E_{\Lambda^\star,\mathbf{z}} \leq C_1 \inf_{\#\Lambda\leq C_2\#\Lambda^\star} E_{\Lambda,\mathbf{z}} \ , \quad (5.9)$$

where the constants C_1 and C_2 are absolute. One can now prove that the piecewise constant function built on the partition Λ^* approximates f_ρ well in probability. Indeed, suppose that at a certain stage of this refinement we receive a partition $\Lambda^\star$. Then for any partition Λ with $\#\Lambda \leq C_2 N$ we have

$$E_{\Lambda^\star,\mathbf{z}} \leq C_1 E_{\Lambda,\mathbf{z}} \ . \quad (5.10)$$

We denote by $\eta^2 := E_{\Lambda^\star,\mathbf{z}}$ and $N = \#\Lambda^\star$ and consider the random variables $r_I := (y - q_I)^2\chi_I(x)$ and their empirical realizations

$$r_{I,\mathbf{z}} := \frac{1}{m}\sum_{i=1}^m (y_i - q_I)^2\chi_I(x_i) = \frac{1}{m}\sum_{i=1}^m \left[(y_i - q_{I,\mathbf{z}})^2 + (q_I - q_{I,\mathbf{z}})^2\right]\chi_I(x_i)$$

$$= \quad e_{I,\mathbf{z}} + (q_I - q_{I,\mathbf{z}})^2 \rho_{I,\mathbf{z}}, \tag{5.11}$$

where as usual $\rho_{I,\mathbf{z}} := \frac{1}{m}\sum_{i=1}^{m} \chi_I(x_i)$. We use the above relation and concentration of measure inequalities in a similar way as in [2] to establish the following estimate for any partition Λ with $\#(\Lambda) \leq C_2 N$.

$$\mathbb{P}\left\{|E_\Lambda - E_{\Lambda,\mathbf{z}}| > \frac{\eta^2}{2C_1}\right\} \leq 4Ne^{-\frac{cm\eta^2}{NM^2}} \tag{5.12}$$

Thus, from the computable quantity $E_{\Lambda,\mathbf{z}}$ we get an estimate for the true error (namely, $E_\Lambda \leq E_{\Lambda,\mathbf{z}} + \frac{\eta^2}{2C_1}$ which holds with high probability and we have a computable bound for this probability (the right side of (5.12)). The estimate (5.12) for $\Lambda^\star$ has a slightly different flavor than our previous results. As we run the algorithm thereby enlarging the tree, the estimate we have for the error will decrease but the bound for the probability of failure of this estimate will increase. The user can decide when to terminate the algorithm and accept the given bounds.

References

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

[2] 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.

[3] Binev, P., A. Cohen, W. Dahmen, and R. DeVore (2005), *Universal algorithms in learning theory - Part II : piecewise polynomial functions*, IGPM Report # 254, RWTH-Aachen, Dec. 2005, to appear in Constructive Approximation (2007).

[4] Binev, P., W. Dahmen, and R. DeVore (2004), *Adaptive Finite Element Methods with Convergence Rates*, Numer. Math. **97**, 219-268.

[5] Binev, P., and R. DeVore (2004), *Fast Computation in Adaptive Tree Approximation*, Numer. Math. **97**, 193-217.

[6] G. Blanchard, C. Schäfer, Y. Rozenholc, Oracle bounds and exact algorithm for dyadic classification trees, Preprint

[7] de Boor, C. (1990), *Quasiinterpolants and approximation power of multivariate splines*, in Computations of curves and surfaces, Dahmen, Gasca, Micchelli (eds.), Kluwer (Dordrecht, Netherlands), 313–345;

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

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

[10] 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.

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

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

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

[14] Temlyakov, V. (2005), *Approximation in learning theory*, IMI preprints **05**, University of South Carolina, 1-42.

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