

Electronic Journal of Statistics

(2008)

ISSN: 1935-7524

Warped Wavelet and Vertical Thresholding

Pierpaolo Brutti

*Dipartimento di Scienze Economiche e Aziendali**LUISS, Guido Carli**Viale Romania 32, 00197 Roma, Italy.**e-mail: pbrutti@luiss.it**url: <http://docenti.luiss.it/brutti>*

Abstract: Let $\{(X_i, Y_i)\}_{i \in \{1, \dots, n\}}$ be an i.i.d. sample from the random design regression model $Y = f(X) + \varepsilon$ with $(X, Y) \in [0, 1] \times [-M, M]$. In dealing with such a model, adaptation is naturally to be intended in terms of $L^2([0, 1], G_X)$ norm where $G_X(\cdot)$ denotes the (known) marginal distribution of the design variable X . Recently much work has been devoted to the construction of estimators that adapts in this setting (see, for example, [5, 24, 25, 32]), but only a few of them come along with a easy-to-implement computational scheme. Here we propose a family of estimators based on the *warped wavelet* basis recently introduced by Picard and Kerkycharian [36] and a tree-like thresholding rule that takes into account the hierarchical (across-scale) structure of the wavelet coefficients. We show that, if the regression function belongs to a certain class of approximation spaces defined in terms of $G_X(\cdot)$, then our procedure is adaptive and converge to the true regression function with an optimal rate. The results are stated in terms of excess probabilities as in [19].

AMS 2000 subject classifications: Primary 62G07, 60K35; secondary 62G20.

Keywords and phrases: Regression with random design, Wavelets, Block thresholding, Warped Wavelets, Adaptive Approximation, Universal Algorithms, Muckenhoupt weights.

1. Introduction

Wavelet bases are ubiquitous in modern nonparametric statistics starting from the 1994 seminal paper by Donoho and Johnstone [27]. What makes them so appealing to statisticians is their ability to capture the relevant features of smooth signals in a few “big” coefficients at high scales (low frequencies) so that zero thresholding the small ones, results in an effective denoising scheme (see [47]).

Although these well known results about thresholding techniques were usually obtained assuming a fixed (and possibly equispaced) design [27, 28], it was quite reassuring to see how they carry over almost unchanged to the irregular design case. As a matter of fact, in the case of irregular design, various attempts to solve this problem has been made: see, for instance, the interpolation methods of Hall and Turlach [34] and Kovac and Silverman [38]; the binning method of Antoniadis et al. [3]; the transformation method of Cai and Brown [14],

or its recent refinements by Maxim [40] for a random design; the weighted wavelet transform of Foster [29]; the isometric method of Sardy et al. [44]; the penalization method of Antoniadis and Fan [2]; and the specific construction of wavelets adapted to the design of Delouille et al. [21, 22] and Jansen et al. [46]. See also Pensky and Vidakovic [41], and the monograph [32].

The main drawback common to most of the methods just mentioned can be found, with no surprise, on the computational side: compared, for instance, with the usual thresholding technique, the calculations are, in general, less direct. To fix this problem, Kerkycharian and Picard [36] propose *warped* wavelet basis. The idea is as follow. For a signal observed at some design points, $Y(t_i)$, $i \in \{1, \dots, 2^J\}$, if the design is regular ($t_k = k/2^J$), the standard wavelet decomposition algorithm starts with $s_{J,k} = 2^{J/2}Y(k/2^J)$ which approximates the scaling coefficient $\int Y(x)\phi_{J,k}(x)dx$, with $\phi_{J,k}(x) = 2^{J/2}\phi(2^Jx - k)$ and $\phi(\cdot)$ the so-called scaling function or father wavelet (see [39] for further information). Then the cascade algorithm is employed to obtain the wavelet coefficients $d_{j,k}$ for $j \leq J$, which in turn are thresholded. If the design is not regular, and we still employ the *same* algorithm, then for a function $H(\cdot)$ such that $H(k/2^J) = t_k$, we have $s_{J,k} = 2^{J/2}Y(H(k/2^J))$. Essentially what we are doing is to decompose, with respect to a standard wavelet basis, the function $Y(H(x))$ or, if $G \circ H(x) \equiv x$, the original function $Y(x)$ itself but with respect to a new *warped* basis $\{\psi_{j,k}(G(\cdot))\}_{(j,k)}$. In the regression setting, this means replacing the standard wavelet expansion of the function $f(\cdot)$ by its expansion on the new basis $\{\psi_{j,k}(G(\cdot))\}_{(j,k)}$, where $G(\cdot)$ is adapting to the design: it may be the distribution function of the design, or its estimation, when it is unknown (not our case). An appealing feature of this method is that it does not need a new algorithm to be implemented: just standard and widespread tools. Of course the properties of this basis depend on the warping factor $G(\cdot)$. In [36] the authors provide the conditions under which this new basis behaves, at least for statistical purposes, as well as ordinary wavelet bases with respect to $L^p([0, 1], dx)$ norms with $p \in (0, +\infty)$. This condition properly quantifies the departure from the uniform distribution and happens to be associated with the notion of Muckenhoupt weights (see [31, 45]).

Now the problem is that we do not need good estimators in $L^p([0, 1], dx)$. What we need are (easy to compute) estimators that adapt in $L^2([0, 1], G_X)$. As a matter of fact it is possible to prove that the main results contained in [36] can be extended to this new setting once we assume $G_X(\cdot)$ to be known as in [15], the case of an unknown $G_X(\cdot)$ being beyond the scope of this work (see [35]).

Here we propose a particular variation on the basic thresholding procedure advanced in [36], that can be motivated as follow. In a variety of real-life signals, significant wavelet coefficients often occur in clusters at adjacent scales and locations. Irregularities, like a discontinuity for example, in general tend to affect the whole block of coefficients corresponding to wavelet functions whose “support” contains them. For this reason it is reasonable to expect that the risk of “blocked” thresholding rules might compare quite favorably with other classical estimators based on level-wise or global thresholds. The literature is

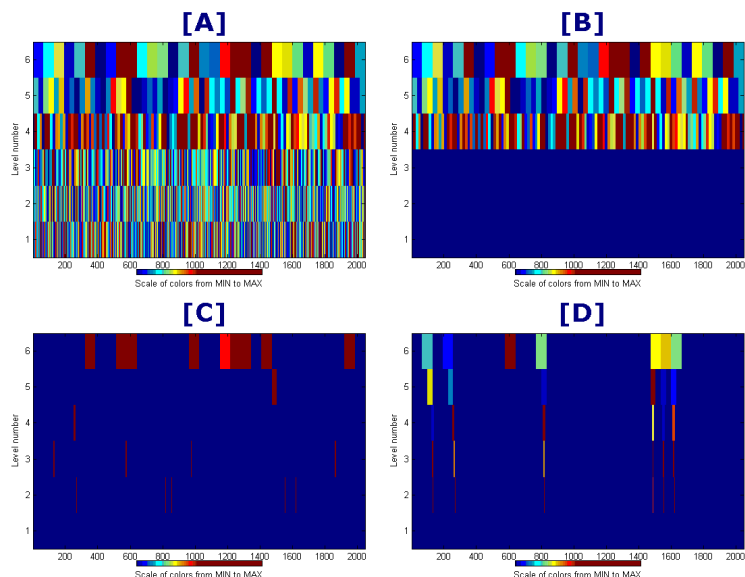


FIG 1. Examples of thresholding rules: [A] - Original wavelet coefficients; [B] - Linear thresholding; [C] - Nonlinear (hard) thresholding; [D] - Vertical (hard) thresholding.

filled with successful examples of “horizontally” (within scales) blocked rules derived from both, purely frequentist arguments [11, 13, 15, 33], or Bayesian reasonings of some flavor [1, 16, 48]. Recently, an increasing amount of work has been devoted to study a new class of “vertically” (across scales, see Figure 1) blocked or *treed* rules [4, 10, 17, 30, 43], that have proved to be of invaluable help in at least two settings of great importance: the construction of *adaptive* pointwise confidence intervals [42] and the derivation of pointwise estimators of a regression function that adapt rate optimally under what we could call a *focused* performance measure [12].

For this reason, adapting some techniques developed in [5] to the current (simplified) setting, in Section 2 we show how vertically zero-thresholding the *warped* wavelet coefficients actually results in an universal smoother with good properties in $L^2([0, 1], G_X)$ over reasonably large approximation spaces.

2. Tree-Structured Warped Approximations

We shall now discuss in greater details nonlinear approximation processes based on warped wavelet bases where a tree structure is pre-imposed on the preserved coefficients. We will start following closely [5] by reviewing some basic facts about partitions and how they are related to adaptive approximation. Then we present the universal algorithm based on adaptive partitions coming from a warped wavelet decomposition and its theoretical properties.

In the spirit of the recent paper by Cucker and Smale [20], we will measure the performances of our estimator by studying its convergence both in probability and expectation. More specifically, let $P\{\cdot\}$ be a – generally unknown or partially unknown– Borel measure defined on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y} \subset \mathbb{R}^d \times \mathbb{R}$, and consider again a nonparametric regression problem where we want to estimate the conditional mean $f(\mathbf{x}) = \mathbb{E}(Y|\mathbf{X} = \mathbf{x})$ from an i.i.d. sample of size n , $\mathbf{z} = \mathbf{z}_n = \{(\mathbf{x}_i, y_i)\}_{i \in \{1, \dots, n\}}$, drawn from $P\{\cdot\}$. Assume further that, chosen an hypothesis space \mathcal{H} from which our candidate estimators $f_{\mathbf{z}}(\cdot)$ comes from, we shall measure the approximation error of $f_{\mathbf{z}}(\cdot)$ in the $L^2(\mathcal{X}, G_{\mathbf{X}})$ norm, where $G_{\mathbf{X}}(\cdot)$ is the (marginal) distribution of the design variable \mathbf{X} . Here, as in the previous section, we will assume $G_{\mathbf{X}}(\cdot)$ to be known. So, given $f_{\mathbf{z}} \in \mathcal{H}$, the quality of its performance is measured by

$$\|f - f_{\mathbf{z}}\| = \|f - f_{\mathbf{z}}\|_{L^2(\mathcal{X}, G_{\mathbf{X}})}.$$

Clearly this quantity is stochastic in nature and, consequently, it is generally not possible to say anything about it for a fixed \mathbf{z} . Instead we look at the behavior in *probability* as measured by

$$P^{\otimes n} \{\mathbf{z} : \|f - f_{\mathbf{z}}\| > \eta\}, \quad \eta > 0$$

or the expected error

$$\mathbb{E}^{\otimes n}(\|f - f_{\mathbf{z}}\|) = \int \|f - f_{\mathbf{z}}\| dP^{\otimes n},$$

where $P^{\otimes n}\{\cdot\}$ denotes the n -fold tensor product of $P\{\cdot\}$. Clearly, given a bound for $P^{\otimes n} \{\mathbf{z} : \|f - f_{\mathbf{z}}\| > \eta\}$, we can immediately obtain another bound for the expected error since

$$\mathbb{E}^{\otimes n}(\|f - f_{\mathbf{z}}\|) = \int_0^{+\infty} P^{\otimes n} \{\mathbf{z} : \|f - f_{\mathbf{z}}\| > \eta\} d\eta. \quad (1)$$

As we will see in Section 4, bounding probabilities like $P^{\otimes n}\{\cdot\}$ usually requires some kind of concentration of measure inequalities (see [8]).

Now, suppose that we have chosen a reasonable hypothesis space \mathcal{H} . We still need to address the problem of how to find an estimator $f_{\mathbf{z}}(\cdot)$ for the regression function $f(\cdot)$. One of the most widespread criteria (see [20, 24, 32], and references therein) is the so called *empirical risk minimization* (least-square data fitting).

Empirical risk minimization is motivated by the fact that the regression function $f(\cdot)$ is the minimizer of

$$\mathcal{E}(w) = \int [w(\mathbf{x}) - y]^2 dP.$$

That is

$$\mathcal{E}(f) = \inf_{w \in L^2(\mathcal{X}, G_{\mathbf{X}})} \mathcal{E}(w).$$

This suggests to consider the problem of minimizing the empirical loss

$$\mathcal{E}_{\mathbf{z}}(w) = \frac{1}{n} \sum_{i=1}^n [w(\mathbf{x}_i) - y_i]^2,$$

over all $w \in \mathcal{H}$. So, in the end, we found an implementable form for our candidate estimator

$$f_{\mathbf{z}} = f_{\mathbf{z}, \mathcal{H}} = \arg \min_{w \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(w),$$

the so-called *empirical minimizer*. Notice that given a finite ball in a linear or nonlinear finite dimensional space, the problem of finding $f_{\mathbf{z}}(\cdot)$ is numerically solvable.

In the following we will see how to build the hypothesis space \mathcal{H} from refinable partitions of the design space \mathcal{X} and then, how this is related to (warped) wavelet basis. Typically $\mathcal{H} = \mathcal{H}_n$ depends on a finite number $J(n)$ of parameters as, for example, the dimension of a linear space or, equivalently, the number of basis functions we use to generate it. In many cases, this number J is chosen using some a priori assumption on the regression function. In other procedures, the number J avoids any a priori assumptions by adapting to the data. We shall be interested in estimators of the latter type.

2.1. Partitions, Adaptive Approximation and Least-Squares Fitting

We will now review some basic facts about partitions and how they are related to adaptive approximation. The treatment follows closely [5]. A partition Λ of $\mathcal{X} \subset [0, 1]^d$ is usually built through a refinement strategy. We first describe the prototypical example of dyadic partitions and then, in the following section, we will make the link with orthonormal expansions through a wavelet basis. So let $\mathcal{X} = [0, 1]^d$, and denote by $\mathcal{D}_j = \mathcal{D}_j(\mathcal{X})$ the collection of dyadic subcubes of \mathcal{X} of sidelength 2^{-j} and $\mathcal{D} = \bigcup_{j=0}^{\infty} \mathcal{D}_j$. These cubes are naturally aligned on a tree $\mathcal{T} = \mathcal{T}(\mathcal{D})$. Each node of the tree \mathcal{T} is a cube $l \in \mathcal{D}$. If $l \in \mathcal{D}_j$, then its children are the 2^d dyadic cubes of \mathcal{D}_{j+1} with $J \subset l$. We denote the set of children of l by $\mathcal{C}(l)$. We call l the parent of each such child J and write $l = \mathcal{P}(J)$. The cubes in $\mathcal{D}_j(\mathcal{X})$ form a uniform partition in which every cube has the same measure $2^{-j d}$.

More in general, we say that a collection of nodes $\tilde{\mathcal{T}}$ is a *proper* subtree of \mathcal{T} if:

- the root node $l \equiv \mathcal{X}$ is in $\tilde{\mathcal{T}}$,
- if $l \neq \mathcal{X}$ is in $\tilde{\mathcal{T}}$ then its parent $\mathcal{P}(l)$ is also in $\tilde{\mathcal{T}}$.

Any finite proper subtree $\tilde{\mathcal{T}}$ is associated to a unique partition $\Lambda = \Lambda(\tilde{\mathcal{T}})$ which consists of its outer leaves, by which we mean those $J \in \mathcal{T}$ such that $J \notin \tilde{\mathcal{T}}$ but $\mathcal{P}(J)$ is in $\tilde{\mathcal{T}}$. One way of generating adaptive partitions is through some refinement strategy. One begins at the root \mathcal{X} and decides whether to refine \mathcal{X} (i.e. subdivide \mathcal{X}) based on some refinement criteria. If \mathcal{X} is subdivided, then

one examines each child and decides whether or not to refine such a child based on the refinement strategy.

We could also consider more general refinements. Assume, for instance, that $a \geq 2$ is a fixed integer. We assume that if \mathcal{X} is to be refined, then its children consist of a subsets of \mathcal{X} which are a partition of \mathcal{X} . Similarly, for each such child there is a rule which spells out how this child is refined. We assume that the child is also refined into a sets which form a partition of the child. Such a refinement strategy also results in a tree \mathcal{T} (called the *master tree*) and children, parents, proper trees and partitions are defined as above for the special case of dyadic partitions. The refinement level j of a node is the smallest number of refinements (starting at root) to create this node. Note that to describe these more general refinements in terms of basis functions, we need to introduce the concept of *warped* multi-wavelets and wavelet packets, but this is beyond the scope of the present work.

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

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

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

Given a partition Λ , we can easily use it to approximate functions supported on \mathcal{X} . More specifically, let us denote by \mathcal{S}_Λ the space of piecewise constant functions – normalized in $L^2(\mathcal{X}, G_X)$ – subordinate to Λ . Each $f \in \mathcal{S}_\Lambda$ can then be written as

$$f(\cdot) = \sum_{l \in \Lambda} c_l \frac{1}{\sqrt{G_X(l)}} \mathbf{1}_l(\cdot),$$

where $\mathbf{1}_l(\cdot)$ denotes the indicator function of any set $l \subset \mathcal{X}$. The best approximation of a given function $f \in L^2(\mathcal{X}, G_X)$ by the elements of \mathcal{S}_Λ is given by

$$\Pi_\Lambda(f)(\cdot) = \sum_{l \in \Lambda} s_l \frac{1}{\sqrt{G_X(l)}} \mathbf{1}_l(\cdot),$$

where

$$s_l = \left\langle f, \frac{1}{\sqrt{G_X(l)}} \mathbf{1}_l \right\rangle_{L^2(G_X)}, \quad (2)$$

and $s_l \equiv 0$ in case $G_X(l) \equiv 0$.

In practice, we can consider two types of approximations corresponding to *uniform refinement* and *adaptive refinement*. We first discuss uniform refinement. Let

$$\mathcal{E}_J(f) = \|f - \Pi_{\Lambda_J}(f)\|_{L^2(G_X)}, \quad J \in \mathbb{N}_0,$$

which is the error for uniform refinement. The decay of this error to zero is connected with the smoothness of $f(\cdot)$ as measured in $L^2(\mathcal{X}, G_{\mathbf{X}})$. We shall denote by \mathcal{A}^s the approximation space (see the review in [23]), consisting of all functions $f \in L^2(\mathcal{X}, G_{\mathbf{X}})$ such that

$$\mathcal{E}_J(f) \leq M_0 a^{-J s}, \quad J \in \mathbb{N}_0. \quad (3)$$

Notice that $\text{card}(\Lambda_J) = a^J$, so that the decay in Equation (3) is like N^{-s} with N the number of elements in the partition. The smallest M_0 for which Equation (3) holds serves to define the semi-norm $|f|_{\mathcal{A}^s}$ on \mathcal{A}^s . The space \mathcal{A}^s can be viewed as a smoothness space of order $s > 0$ with smoothness measured with respect to $G_{\mathbf{X}}(\cdot)$. For example, if $G_{\mathbf{X}}(\cdot)$ is the Lebesgue measure and we use dyadic partitioning then $\mathcal{A}^{s/d} = \mathcal{B}_{\infty}^{2,s}$, $s \in (0, 1]$, with equivalent norms. Here $\mathcal{B}_{\infty}^{2,s}$ is the Besov space which can be described in terms of the differences as

$$\|w(\cdot + h) - w(\cdot)\|_{L^2(dx)} \leq M_0 |h|^s, \quad x, h \in \mathcal{X}.$$

Instead of working with a-priori fixed partitions there is a second kind of approximation where the partition is generated adaptively and will vary with $f(\cdot)$. Adaptive partitions are typically generated by using some refinement criterion that determines whether or not to subdivide a given cell. We shall consider a refinement criteria that was introduced to build adaptive wavelet constructions such as those given by Cohen *et al.* in [17] for image compression. This criteria is analogous to thresholding wavelet coefficients. Indeed, it would be exactly this criteria if we were to construct a wavelet (*Haar* like) bases for $L^2(\mathcal{X}, G_{\mathbf{X}})$. For each cell l in the master tree \mathcal{T} and any $w \in L^2(\mathcal{X}, G_{\mathbf{X}})$ we define

$$\nu_l = \nu_l(w) = \sqrt{\sum_{j \in \mathcal{C}(l)} s_j^2 - s_l^2}, \quad (4)$$

which describes the amount of $L^2(\mathcal{X}, G_{\mathbf{X}})$ energy which is increased in the projection of $w(\cdot)$ onto \mathcal{S}_{Λ} when the element l is refined. It also accounts for the decreased projection error when l is refined. If we were in a classical situation of Lebesgue measure and dyadic refinement, then $\nu_l^2(w)$ would be exactly the sum of squares of the (scaling) Haar coefficients of $w(\cdot)$ corresponding to l .

We can use $\nu_l(w)$ to generate an adaptive partition. Given any $\lambda > 0$, let $\mathcal{T}(w, \lambda)$ be the smallest proper tree that contains all $l \in \mathcal{T}$ for which $\nu_l(w) \geq \lambda$. This tree can also be described as the set of all $J \in \mathcal{T}$ such that there exists $l \subset J$ which verifies $\nu_l(w) \geq \lambda$. Note that since $w \in L^2(\mathcal{X}, G_{\mathbf{X}})$, the set of nodes such that $\nu_l(w) \geq \lambda$ is always finite and so is $\mathcal{T}(w, \lambda)$. Corresponding to this tree we have the partition $\Lambda(w, \lambda)$ consisting of the outer leaves of $\mathcal{T}(w, \lambda)$. We shall define some new approximation spaces \mathcal{B}^s which measure the regularity of a given function $w(\cdot)$ by the size of the tree $\mathcal{T}(w, \lambda)$.

Given $s > 0$, we let \mathcal{B}^s be the collection of all $w \in L^2(\mathcal{X}, G_{\mathbf{X}})$ such that the following is finite

$$|w|_{\mathcal{B}^s}^p = \sup_{\lambda \geq 0} \{ \lambda^p \text{card}(\mathcal{T}(w, \lambda)) \}, \quad \text{with } p = (s + \frac{1}{2})^{-1}. \quad (5)$$

We obtain the norm for \mathcal{B}^s by adding $\|w\|_{L^2(G_{\mathbf{X}})}$ to $|w|_{\mathcal{B}^s}$. One can show that

$$\|w - \Pi_{\Lambda(w,\lambda)}(w)\|_{L^2(G_{\mathbf{X}})} \leq C(s) |w|_{\mathcal{B}^s}^{\frac{1}{2s+1}} \lambda^{\frac{2s}{2s+1}} \leq C(s) |w|_{\mathcal{B}^s} N^{-s}, \quad (6)$$

where $N = \text{card}(\mathcal{T}(w, \lambda))$ and the constant $C(s)$ depends only on s (see Cohen *et al.* [17]). It follows that every function $w \in \mathcal{B}^s$ can be approximated to order $\mathcal{O}(N^{-s})$ by $\Pi_{\Lambda}(w)(\cdot)$ for some partition Λ with $\text{card}(\Lambda) = N$. 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 $\mathcal{B}_q^{\tau,s}$ with $\tau > (s/d + 1/2)^{-1}$ and $q \in (0, \infty]$ arbitrary, is contained in $\mathcal{B}^{s/d}$ (see [17]). This should be compared with the \mathcal{A}^s where we know that $\mathcal{A}^{s/d} = \mathcal{B}_{\infty}^{2,s}$ as we have noted earlier. In the next section we will see how to “visualize” these approximation spaces when we use *warped* wavelet bases to build our partitions.

Until now, we have only considered the problem of approximating elements of some smoothness class by approximators associated to (adaptive) partitions of their domain \mathcal{X} : no data, no noise; just functions. Here, instead, we assume that $f(\cdot)$ denotes, as before, the regression function and we return to the problem of estimating it from a given data-set. Clearly, we can use the functions in $\mathcal{H} = \mathcal{S}_{\Lambda}$ for this purpose, so that the “incarnation” in this context of what we called the *empirical minimizer*, is given by

$$f_{\mathbf{z},\Lambda} = \arg \min_{w \in \mathcal{S}_{\Lambda}} \frac{1}{n} \sum_{i=1}^n [w(\mathbf{x}_i) - y_i]^2,$$

the orthogonal projection of $y = y(\mathbf{x})$ onto \mathcal{S}_{Λ} with respect to the empirical norm

$$\|y\|_{L^2(\mathcal{X}, \delta_{\mathbf{X}})}^2 = \frac{1}{n} \sum_{i=1}^n |y(\mathbf{x}_i)|^2,$$

with $y(\mathbf{x}_i) = y_i$, and we can compute it by solving $\text{card}(\Lambda)$ independent problems, one for each element $l \in \Lambda$. The resulting estimator can then be written as

$$f_{\mathbf{z},\Lambda}(\cdot) = \sum_{l \in \Lambda} s_l(\mathbf{z}) \frac{1}{\sqrt{\mathbb{G}_{\mathbf{X},n}(l)}} \mathbf{1}_l(\cdot),$$

where, for each $l \in \Lambda$,

$$s_l(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n y_i \frac{1}{\sqrt{\mathbb{G}_{\mathbf{X},n}(l)}} \mathbf{1}_l(\mathbf{x}_i) \quad \text{and} \quad \mathbb{G}_{\mathbf{X},n}(l) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}_l(\mathbf{x}_i),$$

are the empirical counterparts of the theoretical coefficients defined in Equation (2). With the coefficients $\{s_l(\mathbf{z})\}_{l \in \Lambda}$ at hand, we can build linear estimators $f_{\mathbf{z}}(\cdot)$ corresponding to uniform partitions with cardinality suitably chosen to balance the bias and variance of $f_{\mathbf{z}}(\cdot)$ when the true regression function $f(\cdot)$ belongs to

Algorithm: Least-Squares on Adaptive Partitions**Require:** Sample $\mathbf{z} = \{(\mathbf{x}_i, y_i)\}_{i \in \{1, \dots, n\}}$; threshold λ_n , $\gamma > 0$ smoothness index**Output:** An estimator $f_{\mathbf{z}}(\cdot)$ for the regression function $f(\cdot)$ **Setup:**1 : Define $J^* = \min \{j \in \mathbb{N} : 2^j \leq \lambda_n^{-1/\gamma}\}$ **Generator:**2 : Compute $\nu_l(\mathbf{z})$ for the nodes l at a refinement level $j < J^*$ 3 : Threshold $\{\nu_l(\mathbf{z})\}_l$ at level λ_n obtaining the set: $\Sigma(\mathbf{z}, n) = \{l \in \mathcal{T}_{J^*} : \nu_l(\mathbf{z}) \geq \lambda_n\}$ 4 : Complete $\Sigma(\mathbf{z}, n)$ to a tree $\mathcal{T}(\mathbf{z}, n)$ by adding nodes $J \supset l \in \Sigma(\mathbf{z}, n)$ 7 : **Return** The estimator $f_{\mathbf{z}}(\cdot)$ that minimizes the empirical risk on $\Lambda(\mathbf{z}, n)$

TABLE 1

Least-squares on adaptive partitions driven by the empirical residuals $\nu_l(\mathbf{z})$ defined in Equation (7). Adapted from [5].

some specific smoothness class. Alternatively, defining the empirical versions of the residuals introduced in Equation (4) as

$$\nu_l(\mathbf{z}) = \sqrt{\sum_{J \in \mathcal{C}(l)} s_J^2(\mathbf{z}) - s_l^2(\mathbf{z})}, \quad (7)$$

we can mimic the adaptive procedure introduced in the previous section (see Table 1) to get universal¹ estimators based on adaptive partitions. These partitions have the same tree structure as those used in the CART algorithm [9], yet the selection of the right partition is quite different since it is not based on an optimization problem but on a thresholding technique applied to empirical quantities computed at each node of the tree which play a role similar to wavelet coefficients as we will see in the following (see [26] for a connection between CART and thresholding in one or several orthonormal bases).

2.2. A Universal Algorithm Based on Warped Wavelets

The choice we made in the previous Section of adopting piecewise constant functions as approximators, severely limits the optimal convergence rate to approximation spaces corresponding to smoothness classes of low or no pointwise regularity (see [6] for an interesting extension based on piecewise polynomial approximations). A possible way to fix this problem would be to use the complexity regularization approach for which optimal convergence results could be obtained in the piecewise polynomial context (see for instance Theorem 12.1 in [32], and the paper by Kohler [37]).

In the present context where the marginal design distribution $G_X(\cdot)$ is assumed to be known, we have another option based on the *warped* systems introduced in Section 1.

It is worth mentioning that in this section we will concentrate on the $\mathcal{X} \equiv [0, 1]$. The present setting could be generalized to the case where $G_X(\cdot)$ is a d -

¹A synonymous of “adaptive”: the estimator does not require any prior knowledge of the smoothness of the regression function $f(\cdot)$.

dimensional tensor product. However, the full generalization to dimension $d > 1$ is more involved and will not be discussed here.

To “translate” the concepts highlighted in the previous two sections in terms of *warped* systems, consider a compactly supported wavelet basis $\{\psi_{j,k}(\cdot), j \geq -1, k \in \mathbb{Z}\}$, where $\psi_{-1,k}(\cdot) = \phi_{0,k}(\cdot)$ denotes the scaling function, and its *warped* version $\{\psi_{j,k}(G_X(\cdot)), j \geq -1, k \in \mathbb{Z}\}$. Then, for each $f \in L^2([0, 1], G_X)$, consider its expansion in this basis

$$f(x) = \sum_{j,k} d_{j,k} \psi_{j,k}(G_X(x)).$$

In this context, a *tree* is a finite set \mathcal{T} of indexes (j, k) , $j \in \mathbb{N}_0$ and $k \in \{0, \dots, 2^j - 1\}$, such that $(j, k) \in \mathcal{T}$ implies $(j - 1, \lfloor k/2 \rfloor) \in \mathcal{T}$, i.e., all “ancestors” of the point (j, k) in the dyadic grid also belong to the tree.

One can then consider the best *tree-structured* approximation to $f(\cdot)$, by trying to minimize

$$\left\| f - \sum_{(j,k) \in \mathcal{T}} d_{j,k} \psi_{j,k}(G) \right\|_{L^2(G_X)}^2,$$

over all tree \mathcal{T} having the same cardinality N , and all choices of $d_{j,k}$. However the procedure of selecting the optimal tree is costly in computational time, in comparison to the simple reordering that characterize the classical thresholding procedure described in the previous section. A more reasonable approach is to use suboptimal tree selection algorithms inspired by the adaptive procedure introduced before. In detail, we start from an initial tree $\mathcal{T}_0 = \{(0, 0)\}$ and let it “grow” as follow:

1. Given a tree \mathcal{T}_N , define its “leaves” $\mathcal{L}(\mathcal{T}_N)$ as the indexes $(j, k) \notin \mathcal{T}_N$ such that $(j - 1, \lfloor k/2 \rfloor) \in \mathcal{T}_N$.
2. For $(j, k) \in \mathcal{L}(\mathcal{T}_N)$ define the residual

$$\nu_{j,k} = \sqrt{\sum_{l,m \in I_{j,k}} |d_{l,m}|^2},$$

with $I_{j,k} = [2^{-j}k, 2^{-j}(k + 1)]$.

3. Choose $(j_0, k_0) \in \mathcal{L}(\mathcal{T}_N)$ such that

$$\nu_{j_0, k_0} = \max_{(j,k) \in \mathcal{L}(\mathcal{T}_N)} \nu_{j,k},$$

4. Define $\mathcal{T}_{N+1} = \mathcal{T}_N \cup \{(j_0, k_0)\}$.

Note that this algorithm can either be controlled by the cardinality N of the tree, or by the size of the residuals as in Table 1.

Now, let Λ be the dyadic partition associated to any such tree, and define

$$\Pi_\Lambda(f)(x) = \sum_{l \in \Lambda} d_l \psi_l(G(x)), \quad \text{with } d_l = \langle f, \psi_l(G) \rangle_{L^2(G_X)},$$

and their empirical counterparts

$$f_{\Lambda, \mathbf{z}}(x) = \sum_{l \in \Lambda} d_l(\mathbf{z}) \psi_l(G(x)), \quad \text{with} \quad d_l(\mathbf{z}) = \frac{1}{n} \sum_{i=1}^n Y_i \psi_l(G(X_i)).$$

Then, by adapting the techniques used in [5], in Section 4.1 we prove the following result for uniform partitions:

Theorem 2.1. (Optimality for uniform partitions) *Assume that $f \in \mathcal{A}^s$ and define the estimator $f_{\mathbf{z}} = f_{\mathbf{z}, \Lambda_{j^*}}$, with*

$$J^* = J^*(n) = \min \left\{ j \in \mathbb{N} : 2^{j(1+2s)} \geq \frac{n}{\log(n)} \right\}.$$

Then, given any $\beta > 0$, there is a constant \tilde{c} such that

$$\mathbb{P}^{\otimes n} \left\{ \|f - f_{\mathbf{z}}\|_{L^2(G_X)}^2 > (\tilde{c} + |f|_{\mathcal{A}^s}) \left(\frac{\log n}{n} \right)^{\frac{s}{2s+1}} \right\} \leq C n^{-\beta}$$

and

$$\mathbb{E}^{\otimes n} \left\{ \|f - f_{\mathbf{z}}\|_{L^2(G_X)}^2 \right\} \leq (C + |f|_{\mathcal{A}^s}^2) \left(\frac{\log n}{n} \right)^{\frac{2s}{2s+1}},$$

where C depends only on M .

Theorem 2.1 is satisfactory in the sense that the rate $\left[\frac{\log(n)}{n} \right]^{-s/(2s+1)}$ is known to be optimal (or minimax) over the class \mathcal{A}^s save for the logarithmic factor. However, it is unsatisfactory in the sense that the estimation procedure requires a-priori knowledge of the smoothness parameter s which appears in the choice of the resolution level j . Moreover, as noted before, the smoothness assumption $f \in \mathcal{A}^s$ is too severe. Consequently, our next task, will consist in deriving a method capable of treating both defects. To this end, mimicking Equation (7), we define the empirical residuals as

$$\nu_{j,k}(\mathbf{z}) = \sqrt{\sum_{l_{\ell,m} \subset l_{j,k}} |d_{\ell,m}(\mathbf{z})|^2}.$$

Then, for some² $\kappa > 0$, let

$$\lambda_n = \kappa \sqrt{\frac{\log(n)}{n}},$$

be a given threshold. Now, adapting the algorithm given in Table 1, assume that the estimator $f_{\mathbf{z}}(\cdot)$ is generated as detailed in Table 2. Then, in Section 4.2, we prove the following

² κ is essentially a smoothing parameter to be selected by cross-validation, for instance). Notice that in our theoretical developments we will only assume that κ is “large enough” to ensure the desired concentration inequalities.

Algorithm: “Treed” Approximations from Warped Wavelets

Require: Sample $\mathbf{z} = \{(\mathbf{x}_i, y_i)\}_{i \in \{1, \dots, n\}}$; threshold λ_n , $\gamma \geq \frac{1}{2}$ smoothness index

Output: An estimator $f_{\mathbf{z}}(\cdot)$ for the regression function $f(\cdot)$

Setup:

- 1 : Define $J^* = \min \{j \in \mathbb{N} : 2^j \leq \lambda_n^{-1/\gamma}\}$

Generator:

- 2 : Compute $\nu_{j,k}(\mathbf{z})$ for any node (j, k) at a refinement level $j < J^*$
- 3 : Threshold $\{\nu_{j,k}(\mathbf{z})\}_{j,k}$ at level λ_n obtaining: $\Sigma(\mathbf{z}, n) = \{(j, k) \in \mathcal{T}_{J^*} : \nu_{j,k}(\mathbf{z}) \geq \lambda_n\}$
- 4 : Complete $\Sigma(\mathbf{z}, n)$ to a tree $\mathcal{T}(\mathbf{z}, n)$ by adding nodes $(\ell, m) \in \mathcal{P}(\{(j, k)\})$ for all $(j, k) \in \Sigma(\mathbf{z}, n)$
- 7 : **Return** The estimator $f_{\mathbf{z}}(\cdot) = \sum_{(j,k) \in \Lambda(\mathbf{z}, n)} d_{j,k}(\mathbf{z}) \psi_{j,k}(G_X(\cdot))$

TABLE 2

Tree-structured approximations from warped wavelet decompositions.

Theorem 2.2. (Optimality for “growing” adaptive partitions) *Let β and $\gamma \geq \frac{1}{2}$ be arbitrary. Then, there exists $\kappa > 0$, such that, whenever $f \in \mathcal{A}^\gamma \cap \mathcal{B}^s$ for some $s > 0$, the following inequalities hold*

$$P^{\otimes n} \left\{ \|f - f_{\mathbf{z}}\|_{L^2(G_X)} > \tilde{c} \left(\frac{\log n}{n} \right)^{\frac{s}{2s+1}} \right\} \leq C n^{-\beta},$$

and

$$\mathbb{E}^{\otimes n} \left\{ \|f - f_{\mathbf{z}}\|_{L^2(G_X)}^2 \right\} \leq C \left(\frac{\log n}{n} \right)^{\frac{2s}{2s+1}},$$

where the constants \tilde{c} and C do not depend on the sample size n .

Theorem 2.2 is definitively more satisfactory than Theorem 2.1 in two respects:

- The optimal rate $\left[\frac{\log(n)}{n} \right]^{-s/(2s+1)}$ is now obtained under weaker smoothness assumptions on the regression function, namely, $f \in \mathcal{B}^s$ in place of $f \in \mathcal{A}^s$, with the extra assumption $f \in \mathcal{A}^\gamma$ with $\gamma \geq \frac{1}{2}$ arbitrary.
- The estimator we obtain is adaptive (universal), in the sense that the value of s does not enter the definition of the algorithm. The procedure automatically extract information about the regularity of the regression function from the data at hand.

It is interesting to notice that in standard thresholding (standard denoising or density estimation, for instance) one usually sets the highest level J^* so that $2^{J^*} \sim n/\log(n)$; here we have to stop much sooner, namely, $2^{J^*} \sim \sqrt{n/\log(n)}$, as in [36]. This is especially necessary to obtain the exponential inequalities in Section 4.1 and 4.2.

A final remark on the approximation spaces \mathcal{A}^s and \mathcal{B}^s is in order. In a previous section, we mentioned that, when $G_X(\cdot)$ is the Lebesgue measure, then the spaces \mathcal{A}^s and \mathcal{B}^s are well understood. In particular, each Besov space $\mathcal{B}_q^{\tau,s}$ with $\tau > (s + 1/2)^{-1}$ and $q \in (0, +\infty]$, is contained in \mathcal{B}^s (see Cohen *et al.* [17, 18]), whereas $\mathcal{A}^s = \mathcal{B}_\infty^{2,s}$. For general partitions it is not totally clear how

to express the content of these approximation spaces in terms of reasonably simple variations of common smoothness classes. Things get slightly simpler when we employ a *warped* wavelet basis to generate the partition. As a result, we can map approximation properties imposed on $f(\cdot)$ to regularity properties over its warped version $f \circ G_X^{-1}(\cdot)$. So, assuming $f \in \mathcal{A}^s$ is equivalent to impose $f \circ G_X^{-1} \in \mathcal{B}_\infty^{2,s}$ as soon as \mathcal{A}^s is defined in terms of warped wavelets.

3. Discussion

The dependence on the design marginal $G_X(\cdot)$ is so far a clear weakness of our approach from both a theoretical and a practical perspective. Nevertheless, an obvious option to extend our tree-structured procedure to the case of an unknown $G_X(\cdot)$, would probably end up combining the arguments introduced in [5] with those considered by Kerkycharian and Picard in [36] and [35]. Another practical option might be to adopt a *split sample* approach and measure smoothness in terms of the *discrete* norm induced by the data. Here we also mention the fact that in Theorem 2.2 we require the knowledge of the parameter γ which can be arbitrary close to $1/2$. As in [5], it is probably possible to remove the dependency on γ at the price of using the much more complicated construction proposed by Binev and DeVore in [7].

4. Proofs for Section 2

4.1. Proof of Theorem 2.1

For any given partition Λ , a natural way to control $\|f - f_{\mathbf{z},\Lambda}\|_{L^2(G_X)}^2$ is by splitting it into a bias and variance term denoted respectively with e_1 and e_2 in the following equation

$$\|f - f_{\mathbf{z},\Lambda}\|_{L^2(G_X)}^2 = \|f - \Pi_\Lambda(f)\|_{L^2(G_X)}^2 + \|\Pi_\Lambda(f) - f_{\mathbf{z},\Lambda}\|_{L^2(G_X)}^2 = e_1 + e_2. \quad (8)$$

e_1 will be controlled by using the smoothness assumptions we made in the statement of the theorem, whereas the variance term e_2 will be controlled by Bernstein's inequality.

Lets start with this second step observing that, by denoting $[d_l - d_l(\mathbf{z})]$ with $\Delta_l(\mathbf{z})$, then by orthonormality of the warped system we have

$$\|\Pi_\Lambda(f) - f_{\mathbf{z},\Lambda}\|_{L^2(G_X)}^2 = \left\| \sum_{l \in \Lambda} [d_l - d_l(\mathbf{z})] \psi_l(G_X(\cdot)) \right\|_{L^2(G_X)}^2 = \sum_{l \in \Lambda} \Delta_l^2(\mathbf{z}).$$

Hence, for any $\eta > 0$,

$$\begin{aligned} \mathbb{P}^{\otimes n} \{ \|\Pi_\Lambda(f) - f_{\mathbf{z},\Lambda}\|_{L^2(G_X)} > \eta \} &= \mathbb{P}^{\otimes n} \left\{ \sum_{l \in \Lambda} \Delta_l^2(\mathbf{z}) > \eta^2 \right\} \leq \quad (9) \\ &\leq \text{card}(\Lambda) \cdot \mathbb{P}^{\otimes n} \left\{ \Delta_l^2(\mathbf{z}) > \frac{\eta^2}{\text{card}(\Lambda)} \right\} = \\ &= \text{card}(\Lambda) \cdot \mathbb{P}^{\otimes n} \left\{ |\Delta_l(\mathbf{z})| > \frac{\eta}{\sqrt{\text{card}(\Lambda)}} \right\}, \end{aligned}$$

Consequently to control e_2 we just need to control $|\Delta_1(\mathbf{z})|$ and the cardinality of Λ . Now, if we define $U = Y\psi_{j,k}(X)$, then

- $M = \|U - \mathbb{E}\{U\}\|_\infty \leq 2 \cdot 2^{j/2} \|\psi\|_\infty \|f\|_\infty$,
- $\sigma^2 = \mathbb{E}\{|U - \mathbb{E}(U)|^2\} \leq \mathbb{E}\{|U|^2\} \leq \|f\|_\infty^2$,

as

$$\begin{aligned} \mathbb{E}\{|\psi_{j,k}(G(X))|^2\} &= \int |\psi_{j,k}(G(x))|^2 dG_X(x) = \int_0^1 |\psi_{j,k}(G(G^{-1}(y)))|^2 dy = \\ &= \int_0^1 |\psi_{j,k}(y)|^2 dy = 1. \end{aligned}$$

Hence, for any $\eta > 0$, by Bernstein's inequality we get

$$\begin{aligned} \mathbb{P}^{\otimes n} \left\{ |\Delta_1(\mathbf{z})| \geq \frac{\eta}{\sqrt{\text{card}(\Lambda)}} \right\} &\leq 2 \exp \left\{ - \frac{n \eta^2}{2 \text{card}(\Lambda) \left[\|f\|_\infty^2 + \frac{2\|\psi\|_\infty \|f\|_\infty}{3} \frac{2^{j/2}}{\sqrt{\text{card}(\Lambda)}} \eta \right]} \right\} \leq \\ &\leq 2 \exp \left\{ - \frac{3n \eta^2}{C' \text{card}(\Lambda)(3 + \eta)} \right\}, \end{aligned} \quad (10)$$

where $C' = 2 \max\{\|f\|_\infty^2, 2\|\psi\|_\infty \|f\|_\infty\}$, and the last inequality comes from the fact that for any $l \in \Lambda$ we have $2^j = |l|^{-1} \leq \text{card}(\Lambda) = 2^J$ for some $J \in \mathbb{N}$, being Λ a dyadic partition.

Now, back to our specific case. First of all remember that, by definition,

$$J^* = J^*(n) = \min \left\{ j \in \mathbb{N} : 2^{j(1+2s)} \geq \frac{n}{\log(n)} \right\},$$

so

$$\text{card}(\Lambda_{J^*}) \leq 2^{J^*+1} \leq 2^{2J^*-1} \leq 2^2 \left[\frac{\log(n)}{n} \right]^{-\frac{1}{1+2s}}. \quad (11)$$

Hence, by definition of \mathcal{A}^s , we get the following bound for e_1 :

$$\|f - \Pi_{\Lambda_{J^*}}(f)\|_{L^2(G_X)} \leq |f|_{\mathcal{A}^s} 2^{-J^*s} \leq |f|_{\mathcal{A}^s} \left[\frac{\log(n)}{n} \right]^{\frac{s}{1+2s}}. \quad (12)$$

From Equation (8) we then get

$$\|f - f_{\mathbf{z}, \Lambda_{J^*}}\|_{L^2(G_X)}^2 \leq |f|_{\mathcal{A}^s}^2 \left[\frac{\log(n)}{n} \right]^{\frac{2s}{1+2s}} + e_2^2,$$

therefore, for all $\delta > 0$

$$\mathbb{P}^{\otimes n} \left\{ \|f - f_{\mathbf{z}, \Lambda_{J^*}}\|_{L^2(G_X)} \geq \delta \right\} \leq \mathbb{P}^{\otimes n} \left\{ e_2 > \delta - |f|_{\mathcal{A}^s} \left[\frac{\log(n)}{n} \right]^{\frac{s}{1+2s}} \right\}.$$

Taking $\delta = (\tilde{c} + |f|_{\mathcal{A}^s}) \left[\frac{\log(n)}{n} \right]^{\frac{s}{2s+1}}$ as in the statement of Theorem 2.1, and applying Equations (9) and (10) noticing that $\left[\frac{\log(n)}{n} \right]^{\frac{s}{2s+1}} < 1$ for every $s > 0$, we obtain

$$\mathbb{P}^{\otimes n} \left\{ e_2 > \tilde{c} \left[\frac{\log(n)}{n} \right]^{\frac{s}{1+2s}} \right\} \leq 2 \text{card}(\Lambda_{J^*}) \exp \left\{ - \left(\frac{n}{\text{card}(\Lambda_{J^*})} \left[\frac{\log(n)}{n} \right]^{\frac{2s}{1+2s}} \right) \frac{3\tilde{c}^2}{C'(3+\tilde{c})} \right\}.$$

But from Equation (11) we know how to bound the cardinality of our partition, therefore

$$\begin{aligned} \left(\frac{n}{\text{card}(\Lambda_{J^*})} \left[\frac{\log(n)}{n} \right]^{\frac{2s}{1+2s}} \right) \frac{3\tilde{c}^2}{C'(3+\tilde{c})} &> \left(\frac{n}{2^2} \left[\frac{\log(n)}{n} \right]^{\frac{1}{1+2s}} \left[\frac{\log(n)}{n} \right]^{\frac{2s}{1+2s}} \right) \frac{3\tilde{c}^2}{C'(3+\tilde{c})} = \\ &= g(\tilde{c}) \cdot \log(n), \end{aligned}$$

with

$$g(\tilde{c}) = \frac{3\tilde{c}^2}{4C'(3+\tilde{c})},$$

so that

$$\begin{aligned} \mathbb{P}^{\otimes n} \left\{ \|f - f_{\mathbf{z}, \Lambda_{J^*}}\|_{L^2(G_X)} \geq \delta \right\} &= \mathbb{P}^{\otimes n} \left\{ e_2 > \tilde{c} \left[\frac{\log(n)}{n} \right]^{\frac{s}{1+2s}} \right\} \leq \\ &\leq 2 \cdot 2^2 \left[\frac{\log(n)}{n} \right]^{-\frac{1}{1+2s}} \exp \left\{ \log \left[n^{-g(\tilde{c})} \right] \right\} \leq \\ &\leq 8 n^{-[g(\tilde{c})-1]} \leq 8 n^{-\beta}, \end{aligned}$$

where the last inequality holds as soon as $g(\tilde{c}) - 1 > \beta$. And this complete the proof since from here we can easily derive a bound for the risk by using Equation (1).

4.2. Proof of Theorem 2.2

Lets start with a bit of notation. First of all, for each $\lambda > 0$, we will denote by

- $\mathcal{T}(f, \lambda)$: smallest tree which contains all dyadic intervals I such that $\nu_I > \lambda$.
- $\Lambda(f, \lambda)$: partition induced by the outer leaves of $\mathcal{T}(f, \lambda)$.
- $\mathcal{T}(f, \lambda, \mathbf{z})$: smallest tree which contains all dyadic intervals I such that $\nu_I(\mathbf{z}) > \lambda$.
- $\Lambda(f, \lambda, \mathbf{z})$: partition induced by the outer leaves of $\mathcal{T}(f, \lambda, \mathbf{z})$.

If Λ_0 and Λ_1 are partitions associated to the tree \mathcal{T}_0 and \mathcal{T}_1 , then we denote by

- $\Lambda_0 \vee \Lambda_1$ the partition associated to the tree $\mathcal{T}_0 \cup \mathcal{T}_1$,
- $\Lambda_0 \wedge \Lambda_1$ the partition associated to the tree $\mathcal{T}_0 \cap \mathcal{T}_1$.

Finally, let $\lambda_n = \kappa \sqrt{\frac{\log(n)}{n}}$ for some $\kappa > 0$, and

$$J^* = \min \{j \in \mathbb{N} : 2^j \leq \lambda_n^{-1/\gamma}\}.$$

Then for each $\lambda > 0$, define the partitions

- $\Lambda(\lambda) = \Lambda(f, \lambda) \wedge \Lambda_{J^*}$,
- $\Lambda(\lambda, \mathbf{z}) = \Lambda(f, \lambda, \mathbf{z}) \wedge \Lambda_{J^*}$.

Therefore, in this section, we consider the adaptive estimator

$$f_{\mathbf{z},n}(x) = f_{\mathbf{z},\Lambda(\lambda_n,\mathbf{z})}(x) = \sum_{l \in \Lambda(\tau_n,\mathbf{z})} d_l(\mathbf{z})\psi_l(G_X(x)).$$

Lets now start the proof observing that, using the triangle inequality, we can decompose the loss as follow

$$\|f - f_{\mathbf{z},n}\|_{L^2(G_X)} = e_1 + e_2 + e_3 + e_4,$$

where

- $e_1 = \|f - \Pi_{\Lambda(\lambda_n,\mathbf{z}) \vee \Lambda(2\lambda_n)}(f)\|_{L^2(G_X)}$,
- $e_2 = \|\Pi_{\Lambda(\lambda_n,\mathbf{z}) \vee \Lambda(2\lambda_n)}(f) - \Pi_{\Lambda(\lambda_n,\mathbf{z}) \wedge \Lambda(2^{-1}\lambda_n)}(f)\|_{L^2(G_X)}$,
- $e_3 = \|\Pi_{\Lambda(\lambda_n,\mathbf{z}) \wedge \Lambda(2^{-1}\lambda_n)}(f) - f_{\mathbf{z},\Lambda(\lambda_n,\mathbf{z}) \wedge \Lambda(2^{-1}\lambda_n)}\|_{L^2(G_X)}$,
- $e_4 = \|f_{\mathbf{z},\Lambda(\lambda_n,\mathbf{z}) \wedge \Lambda(2^{-1}\lambda_n)} - f_{\mathbf{z},\Lambda(\tau_n,\mathbf{z})}\|_{L^2(G_X)}$.

This type of splitting is frequently used in the analysis of wavelet thresholding procedures to deal with the fact that the partition built from those l such that $\nu_l(\mathbf{z}) \geq \lambda_n$, does not exactly coincides with the partition which would be chosen by an oracle based on those l such that $\nu_l \geq \lambda_n$. This is accounted by the terms e_2 and e_4 which correspond to those dyadic interval l such that $\nu_l(\mathbf{z})$ is significantly larger or smaller than ν_l respectively, and which will proved to be small in probability. The remaining terms e_1 and e_3 correspond respectively to the bias and variance of oracle estimators based on partitions obtained by zero-thresholding based on the unknown quantities $\{\nu_l\}_l$.

The first term e_1 , being a bias, is treated by a deterministic estimate as in [5]. More specifically, since $\Lambda(\lambda_n, \mathbf{z}) \vee \Lambda(2\lambda_n)$ is a refinement of $\Lambda(2\lambda_n) = \Lambda(f, 2\lambda_n) \wedge \Lambda_{J^*}$, we have (almost surely):

$$\begin{aligned} e_1 &\leq \|f - \Pi_{\Lambda(2\lambda_n)}(f)\|_{L^2(G_X)} \leq \\ &\leq \|f - \Pi_{\Lambda(f,2\lambda_n)}(f)\|_{L^2(G_X)} + \|\Pi_{\Lambda(f,2\lambda_n)}(f) - \Pi_{\Lambda(2\lambda_n)}(f)\|_{L^2(G_X)} \leq \\ &\leq \|f - \Pi_{\Lambda(f,2\lambda_n)}(f)\|_{L^2(G_X)} + \|f - \Pi_{\Lambda_{J^*}}(f)\|_{L^2(G_X)} \leq \\ &\leq C(s)[2\lambda_n]^{\frac{2s}{2s+1}} |f|_{B^s} + 2^{-\gamma J^*} |f|_{A^\gamma} \leq \\ &\leq C(s)[2\lambda_n]^{\frac{2s}{2s+1}} |f|_{B^s} + 2^{-\gamma} \lambda_n |f|_{A^\gamma}. \end{aligned}$$

Therefore

$$e_1 \leq C(s) \left\{ (2\kappa)^{\frac{2s}{2s+1}} + 2^\gamma \kappa \right\} \max \{ |f|_{A^\gamma}, |f|_{B^s} \} \left[\frac{\log(n)}{n} \right]^{\frac{s}{2s+1}} = c_1 \left[\frac{\log(n)}{n} \right]^{\frac{s}{2s+1}},$$

as soon as $f \in \mathcal{B}^s \cap \mathcal{A}^\gamma$, with $c_1 = C(s) \left\{ (2\kappa)^{\frac{2s}{2s+1}} + 2^\gamma \kappa \right\} \max \{ |f|_{\mathcal{A}^\gamma}, |f|_{\mathcal{B}^s} \}$.

The third term e_3 is treated by the estimate provided by combining Equations (9) and (10)

$$\mathbb{P}^{\otimes n} \{e_3 > \eta\} \leq 2 \text{card}(\Lambda_3) \exp \left\{ -\frac{3n \eta^2}{C' \text{card}(\Lambda_3)(3 + \eta)} \right\}, \quad (13)$$

where $\Lambda_3 = \Lambda(\lambda_n, \mathbf{z}) \wedge \Lambda(2^{-1}\lambda_n)$. So

$$\begin{aligned} \text{card}(\Lambda_3) &\leq \text{card}(\Lambda(2^{-1}\lambda_n)) = \text{card}(\Lambda(f, 2^{-1}\lambda_n) \wedge \Lambda_{J^*}) \leq \\ &\leq \text{card}(\Lambda(f, 2^{-1}\lambda_n)) \leq (2^{-1}\lambda_n)^{-p} |f|_{\mathcal{B}^s}^p = 2^p \lambda_n^{-\frac{2}{1+2s}} |f|_{\mathcal{B}^s}^p = \\ &= 2^p \kappa^{-\frac{2}{1+2s}} |f|_{\mathcal{B}^s}^p \left[\frac{\log(n)}{n} \right]^{-\frac{2}{2(1+2s)}} = c_3 \left[\frac{\log(n)}{n} \right]^{-\frac{1}{1+2s}}, \end{aligned} \quad (14)$$

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

For the remaining two terms, e_2 and e_4 we will show that $\forall \beta > 0$ we fix, there exists a constant $C' > 0$ such that:

$$\mathbb{P}^{\otimes n} \{e_2 > 0\} + \mathbb{P}^{\otimes n} \{e_4 > 0\} \leq C' n^{-\beta}. \quad (15)$$

Before we prove this, let's show why it is sufficient. Let $0 < \delta = \tilde{c} \left[\frac{\log(n)}{n} \right]^{\frac{1}{1+2s}}$ as in the statement of Theorem 2.2. Then we have

$$\begin{aligned} \mathbb{P}^{\otimes n} \{ \|f - f_{\mathbf{z},n}\|_{L^2(G_X)} \geq \delta \} &\leq \mathbb{P}^{\otimes n} \{e_1 + e_2 + e_3 + e_4 \geq \delta\} \leq \\ &\leq \mathbb{P}^{\otimes n} \left\{ e_2 + e_3 + e_4 \geq (\tilde{c} - c_1) \left[\frac{\log(n)}{n} \right]^{\frac{s}{2s+1}} \right\} \leq \\ &\leq \mathbb{P}^{\otimes n} \{e_2 > 0\} + \mathbb{P}^{\otimes n} \{e_4 > 0\} + \mathbb{P}^{\otimes n} \{e_3 \geq \tilde{\delta}\} \leq \\ &\stackrel{\text{by Eq. (15)}}{\leq} C' n^{-\beta} + \mathbb{P}^{\otimes n} \{e_3 \geq \tilde{\delta}\}, \end{aligned}$$

where $\tilde{\delta} = (\tilde{c} - c_1) \left[\frac{\log(n)}{n} \right]^{\frac{s}{2s+1}}$. Repeating the steps used to derive the bound we needed in Section 4.1, from Equations (13) and (14), we obtain

$$\begin{aligned} &\left(\frac{n}{\text{card}(\Lambda_3)} \left[\frac{\log(n)}{n} \right]^{\frac{2s}{1+2s}} \right) \frac{3(\tilde{c} - c_1)^2}{C'[3 + (\tilde{c} - c_1)]} > \\ &> \left(\frac{n}{c_3} \left[\frac{\log(n)}{n} \right]^{\frac{1}{1+2s}} \left[\frac{\log(n)}{n} \right]^{\frac{2s}{1+2s}} \right) \frac{3(\tilde{c} - c_1)^2}{C'[3 + (\tilde{c} - c_1)]} = \\ &= g(\tilde{c}) \cdot \log(n), \end{aligned}$$

where

$$g(\tilde{c}) = \frac{3(\tilde{c} - c_1)^2}{c_3 C'[3 + (\tilde{c} - c_1)]}.$$

Therefore

$$\mathbf{P}^{\otimes n} \left\{ e_3 \geq \tilde{\delta} \right\} \leq 2c_3 \left[\frac{\log(n)}{n} \right]^{-\frac{1}{1+2s}} n^{-g(\tilde{c})} \leq c' n^{-[g(\tilde{c})-1]} \leq c' n^{-\beta},$$

as soon as $g(\tilde{c}) - 1 \geq \beta$. And this would conclude the proof of Theorem 2.2.

We need to prove Equation (15). The main tool is the following lemma

Lemma 4.1. *For each $I \in \Lambda_{J^*}$, we have*

- $\mathbf{P}^{\otimes n} (\{\nu_I(\mathbf{z}) \leq \lambda_n\} \cap \{\nu_I \geq 2\lambda_n\}) \leq 4 n^{-g(\kappa)}$,
- $\mathbf{P}^{\otimes n} (\{\nu_I(\mathbf{z}) \geq \lambda_n\} \cap \{\nu_I \leq 2^{-1}\lambda_n\}) \leq 4 n^{-g(\kappa)}$,

where

$$g(\kappa) = \frac{3\kappa^2}{8C' \left(3 + \kappa^{1-\frac{1}{2\gamma}} \right)}.$$

Before we prove Lemma 4.1, let's show why this is sufficient. Remember that

$$e_2 = \|\Pi_{\Lambda(\lambda_n, \mathbf{z}) \vee \Lambda(2\lambda_n)}(f) - \Pi_{\Lambda(\lambda_n, \mathbf{z}) \wedge \Lambda(2^{-1}\lambda_n)}(f)\|_{L^2(G_X)}.$$

Consequently

- $e_2 \equiv 0$ if $\mathcal{T}(\lambda_n, \mathbf{z}) \cup \mathcal{T}(2\lambda_n) \equiv \mathcal{T}(\lambda_n, \mathbf{z}) \cap \mathcal{T}(2^{-1}\lambda_n)$,
- $e_2 > 0$ if

$$\begin{aligned} \mathcal{T}(\lambda_n, \mathbf{z}) \cup \mathcal{T}(2\lambda_n) \supset \mathcal{T}(\lambda_n, \mathbf{z}) \cap \mathcal{T}(2^{-1}\lambda_n) &\Leftarrow \begin{cases} \mathcal{T}(\lambda_n, \mathbf{z}) \not\subset \mathcal{T}(2^{-1}\lambda_n) \\ \text{or} \\ \mathcal{T}(2\lambda_n) \not\subset \mathcal{T}(\lambda_n, \mathbf{z}) \end{cases} \\ &\Leftarrow \exists \text{ s.t. } \begin{cases} \{\nu(\mathbf{z}) \leq \lambda_n\} \cap \{\nu \geq 2\lambda_n\} \\ \text{or} \\ \{\nu(\mathbf{z}) \geq \lambda_n\} \cap \{\nu \leq 2^{-1}\lambda_n\} \end{cases}. \end{aligned}$$

Therefore

$$\begin{aligned} \mathbf{P}^{\otimes n} \{e_2 > 0\} &\leq \sum_{I \in \Lambda_{J^*}} \mathbf{P}^{\otimes n} (\{\nu_I(\mathbf{z}) \leq \lambda_n\} \cap \{\nu_I \geq 2\lambda_n\}) + \\ &\quad + \sum_{I \in \Lambda_{J^*}} \mathbf{P}^{\otimes n} (\{\nu_I(\mathbf{z}) \geq \lambda_n\} \cap \{\nu_I \leq 2^{-1}\lambda_n\}) = R_1 + R_2. \end{aligned} \tag{16}$$

Then, by applying the first part of Lemma 4.1, we get

$$\begin{aligned} R_1 &\leq \text{card}(\Lambda_{J^*}) 4 n^{-g(\kappa)} \leq \text{card}(\Lambda_0) 2^{J^*} 4 n^{-g(\kappa)} \leq \\ &\leq \text{card}(\Lambda_0) \lambda_n^{-1/\gamma} 4 n^{-g(\kappa)} \leq \text{card}(\Lambda_0) \kappa^{-1/\gamma} \left[\frac{n}{\log(n)} \right]^{\frac{1}{2\gamma}} 4 n^{-g(\kappa)} \leq \\ &\leq \text{card}(\Lambda_0) \kappa^{-1/\gamma} n^{\frac{1}{\gamma}} 4 n^{-g(\kappa)} = C' n^{-\left[g(\kappa) - \frac{1}{\gamma} \right]}, \end{aligned} \tag{17}$$

and analogously, by the second part of Lemma 4.1, we obtain

$$R_2 \leq C' n^{-\left[g(\kappa) - \frac{1}{\gamma} \right]}. \tag{18}$$

Applying again the second part of Lemma 4.1, we are also able to bound e_4 as follow

$$\mathbf{P}^{\otimes n} \{e_4 > 0\} \leq \sum_{l \in \Lambda_{J^*}} \mathbf{P}^{\otimes n} (\{\nu_l(\mathbf{z}) \geq \lambda_n\} \cap \{\nu_l \leq 2^{-1} \lambda_n\}) \leq C' n^{-\left[\frac{g(\kappa)-1}{\gamma}\right]}. \quad (19)$$

Combining Equations (16), (17), (18), and (19), we complete the proof of Equation (15). In fact, given β and $\gamma \geq \frac{1}{2}$, we can find κ such that the theorem holds.

4.2.1. Proof of Lemma 4.1

Lets starting noticing that, for each $\eta > 0$,

$$\{\nu_l(\mathbf{z}) \leq \eta\} \cap \{\nu_l \geq 2\eta\} \subseteq \{|\nu_l(\mathbf{z}) - \nu_l| \geq \eta\},$$

hence

$$\mathbf{P}^{\otimes n} (\{\nu_l(\mathbf{z}) \leq \eta\} \cap \{\nu_l \geq 2\eta\}) \leq \mathbf{P}^{\otimes n} \{|\nu_l(\mathbf{z}) - \nu_l| \geq \eta\}.$$

In addition

$$\begin{aligned} |\nu_l(\mathbf{z}) - \nu_l| &= \left| \sqrt{\sum_{J \in \mathcal{C}(l)} d_l^2(\mathbf{z})} - \sqrt{\sum_{J \in \mathcal{C}(l)} d_l^2} \right| = \left| \|\mathbf{d}(\mathbf{z})\|_2 - \|\mathbf{d}\|_2 \right| \leq \\ &\leq \|\mathbf{d}(\mathbf{z}) - \mathbf{d}\|_2 \stackrel{\text{dyadic}}{=} \sqrt{[d_{l^+}(\mathbf{z}) - d_{l^+}]^2 + [d_{l^-}(\mathbf{z}) - d_{l^-}]^2}, \end{aligned}$$

where l^+ and l^- denote respectively the left and right child of l . So

$$\begin{aligned} \{|\nu_l(\mathbf{z}) - \nu_l| \geq \eta\} &\Leftrightarrow \{|\nu_l(\mathbf{z}) - \nu_l|^2 \geq \eta^2\} \Leftrightarrow \begin{cases} [d_{l^+}(\mathbf{z}) - d_{l^+}]^2 \geq \frac{\eta^2}{2} \\ [d_{l^-}(\mathbf{z}) - d_{l^-}]^2 \geq \frac{\eta^2}{2} \end{cases} \\ &\Leftrightarrow \begin{cases} |d_{l^+}(\mathbf{z}) - d_{l^+}| \geq \frac{\eta}{\sqrt{2}} \\ |d_{l^-}(\mathbf{z}) - d_{l^-}| \geq \frac{\eta}{\sqrt{2}} \end{cases}. \end{aligned}$$

Therefore

$$\mathbf{P}^{\otimes n} (\{\nu_l(\mathbf{z}) \leq \lambda_n\} \cap \{\nu_l \geq 2\lambda_n\}) \leq \mathbf{P}^{\otimes n} (|\Delta_{l^+}(\mathbf{z})| \geq \frac{\eta}{\sqrt{2}}) + \mathbf{P}^{\otimes n} (|\Delta_{l^-}(\mathbf{z})| \geq \frac{\eta}{\sqrt{2}}).$$

If we now take $\eta = \kappa \sqrt{\frac{\log(n)}{n}}$, by applying the Bernstein's inequality as in Section 4.1, for $J \in \{l^+, l^-\}$ we obtain³

$$\begin{aligned} \mathbf{P}^{\otimes n} \left(|\Delta_J(\mathbf{z})| \geq \frac{\kappa}{\sqrt{2}} \sqrt{\frac{\log(n)}{n}} \right) &\leq 2 \exp \left\{ -\frac{3\kappa^2 \log(n)}{2C' [3 + 2^{(j+1)/2} 2^{-1/2} \kappa \sqrt{\frac{\log(n)}{n}}]} \right\} \leq \\ &\leq 2 \exp \left\{ -\frac{3\kappa^2 \log(n)}{8C' [3 + 2^{j/2} \kappa \sqrt{\frac{\log(n)}{n}}]} \right\}. \end{aligned}$$

³Compare with the proof of Proposition 3 in [36].

Now, by hypothesis, we know that

$$2^j \leq \lambda_n^{-1/\gamma} = \left[\frac{1}{\kappa} \sqrt{\frac{n}{\log(n)}} \right]^{\frac{1}{\gamma}} \Rightarrow 2^{j/2} \leq \left[\frac{1}{\kappa} \sqrt{\frac{n}{\log(n)}} \right]^{\frac{1}{2\gamma}},$$

therefore

$$\frac{2^{j/2}}{\frac{1}{\kappa} \sqrt{\frac{n}{\log(n)}}} \leq \left[\frac{1}{\kappa} \sqrt{\frac{n}{\log(n)}} \right]^{-\left(1 - \frac{1}{2\gamma}\right)} = \left[\kappa \sqrt{\frac{\log(n)}{n}} \right]^{1 - \frac{1}{2\gamma}} \stackrel{\gamma \geq \frac{1}{2}}{\leq} \kappa^{1 - \frac{1}{2\gamma}},$$

hence

$$\begin{aligned} \mathbf{P}^{\otimes n} \left(|\Delta_{\mathbf{J}}(\mathbf{z})| \geq \frac{\kappa}{\sqrt{2}} \sqrt{\frac{\log(n)}{n}} \right) &\leq 2 \exp \left\{ -\frac{3\kappa^2 \log(n)}{8C' \left(3 + \kappa^{1 - \frac{1}{2\gamma}}\right)} \right\} = \\ &= 2 \exp \{-g(\mathbf{z}) \log(n)\} = 2 n^{-g(\mathbf{z})}, \end{aligned}$$

with

$$g(\kappa) = \frac{3\kappa^2}{8C' \left(3 + \kappa^{1 - \frac{1}{2\gamma}}\right)}.$$

So finally

$$\mathbf{P}^{\otimes n} (\{\nu_1(\mathbf{z}) \leq \lambda_n\} \cap \{\nu_1 \geq 2\lambda_n\}) \leq 4 n^{-g(\mathbf{z})}.$$

Now, let's evaluate the other term in a similar manner, starting from

$$\begin{aligned} \mathbf{P}^{\otimes n} (\{\nu_1(\mathbf{z}) \geq \lambda_n\} \cap \{\nu_1 \leq 2^{-1}\lambda_n\}) &\leq \mathbf{P}^{\otimes n} \{|\nu_1(\mathbf{z}) - \nu_1| \geq 2^{-1}\lambda_n\} \leq \\ &\leq \sum_{\mathbf{J} \in \{I^+, I^-\}} \mathbf{P}^{\otimes n} \left(|\Delta_{\mathbf{J}}(\mathbf{z})| \geq \frac{\lambda_n}{2\sqrt{2}} \right). \end{aligned}$$

By the same arguments adopted before, we see that, for each $\mathbf{J} \in \{I^+, I^-\}$,

$$\mathbf{P}^{\otimes n} \left(|\Delta_{\mathbf{J}}(\mathbf{z})| \geq \frac{\lambda_n}{2\sqrt{2}} \right) \leq 2 n^{-g(\mathbf{z})},$$

and consequently

$$\mathbf{P}^{\otimes n} (\{\nu_1(\mathbf{z}) \geq \lambda_n\} \cap \{\nu_1 \leq 2^{-1}\lambda_n\}) \leq 4 n^{-g(\mathbf{z})}.$$

References

- [1] F. Abramovich, P. Besbeas, and T. Sapatinas. Empirical Bayes approach to block wavelet function estimation. *Computational Statistics and Data Analysis*, 39:435–451, 2002.
- [2] A. Antoniadis and J. Fan. Regularization of wavelet approximations. *Journal of the American Statistical Association*, 96:939–967, 2001.

- [3] A. Antoniadis, G. Grégoire, and P. Vial. Random design wavelet curve smoothing. *Statistics and Probability Letters*, 35:225–232, 1997.
- [4] F. Autin, D. Picard, and V. Rivoirard. Maxiset approach for Bayesian nonparametric estimation *Mathematical Methods of Statistics*, 15(4):349–373, 2006.
- [5] P. Binev, A. Cohen, W. Dahmen, R. DeVore, and V. N. Temlyakov. Universal algorithms for learning theory part I: piecewise constant functions. *Journal of Machine Learning Research*, 6:1297–1321, 2005.
- [6] P. Binev, A. Cohen, W. Dahmen, and R. DeVore. Universal algorithms for learning theory part II: piecewise polynomial functions. *Constructive Approximation*, 26(2):127–152, August 2007.
- [7] P. Binev and R. DeVore. Fast computation in adaptive tree approximation. *Numerische Math.*, 97(02:11):193–217, 2004.
- [8] S. Boucheron, O. Bousquet, and G. Lugosi. Concentration inequalities. In O. Bousquet, U. v. Luxburg, and Rätsch G., editors, *Advanced Lectures in Machine Learning*, pages 208–240. Springer, 2004.
- [9] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and regression trees*. Wadsworth International, 1984.
- [10] P. Brutti. Variable bandwidth schemes for local polynomial smoothers via vertical wavelet thresholding. In S. Barber R.G. Aykroyd and K.V. Mardia, editors, *Bioinformatics, Images, and Wavelets*, pages 119–121. Department of Statistics, University of Leeds, 2004.
- [11] T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *The Annals of Statistics*, 27:898–924, 1999.
- [12] T. Cai and M. G. Low. Nonparametric estimation over shrinking neighborhoods: superefficiency and adaptation. *The Annals of Statistics*, 33(1), 2005.
- [13] T. Cai and W. Silverman. Incorporating information on the neighboring coefficients into wavelet estimation. *Sankhyā, Series B*, 63:127–148, 2001. Special issue on wavelets.
- [14] T. T. Cai and L. D. Brown. Wavelet shrinkage for nonequispaced samples. *The Annals of Statistics*, 26(5):1783–1799, October 1998.
- [15] C. Chesneau. Wavelet block thresholding for samples with random design: a minimax approach under the L^p risk. *Electronic Journal of Statistics*, 1:331–346, 2007.
- [16] D. De Canditiis and B. Vidakovic. Wavelet Bayesian block shrinkage via mixtures of normal–inverse gamma priors. Technical Report RT 234/01, Istituto per le Applicazioni del Calcolo , Sezione di Napoli, 2001.
- [17] A. Cohen, W. Dahmen, I. Daubechies, and R. DeVore. Tree approximation and optimal encoding. *Applied Computational and Harmonic Analysis*, 11(2):167–191, 1999.
- [18] A. Cohen, I. Daubechies, O. G. Guleryuz, and M. T. Orchard. On the importance of combining wavelet-based nonlinear approximation with coding strategies. *IEEE Transactions on Information Theory*, 48(7):1895–1921, 2002.
- [19] F. Cucker and S. Smale. On the mathematical foundations of learning

- theory. *Bulletin. Amer. Math. Soc.*, 39:1–49, 2002.
- [20] Felipe Cucker and Steve Smale. On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1):1–49 (electronic), 2002.
- [21] V. Delouille, J. Franke, and R. von Sachs. Nonparametric stochastic regression with design-adapted wavelets. *Sankhyā, Series A*, 63(3):328–366, 2001.
- [22] V. Delouille, J. Simoens, and R. von Sachs. Smooth design-adapted wavelets for nonparametric stochastic regression. *Journal of the American Statistical Association*, 99(467):643–658, 2004.
- [23] R. DeVore. Nonlinear approximation. *Acta Numerica*, pages 1–99, 1998.
- [24] R. DeVore, G. Kerkyacharian, D. Picard, and V. N. Temlyakov. Mathematical methods for supervised learning. Research Report 04:22, Industrial Mathematics Institute, 2004.
- [25] R. DeVore, G. Kerkyacharian, D. Picard, and V. N. Temlyakov. On mathematical methods of learning. Research Report 04:10, Industrial Mathematics Institute, 2004.
- [26] D. L. Donoho. CART and best-ortho-basis: A connection. *The Annals of Statistics*, 25(5):1870–1911, October 1997.
- [27] D. L. Donoho and I. M. Johnstone. Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(425–455):425–455, 1994.
- [28] D. L. Donoho, I. M. Johnstone, G. Kerkyacharian, and D. Picard. Wavelet shrinkage: asymptotia. *Journal of the Royal Statistical Society, Series B*, 57(2):301–370, 1995. with discussion.
- [29] G. Foster. Wavelet for period analysis of unequally sampled time series. *Astronomy Journal*, 112:1709–1729, 1996.
- [30] P. Fryzlewicz. Bivariate hard thresholding in wavelet function estimation. Technical Report TR-04-03, Department of Mathematics, Imperial College London, UK, 2004.
- [31] J. García-Cuerva and J. M. Martell. Wavelet characterization of weighted spaces. *Journal of Geometrical Analysis*, 11:241–264, 2001.
- [32] L. Györfi, M. Kohler, A. Krzyżak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, 2002.
- [33] P. Hall, G. Kerkyacharian, and D. Picard. On the minimax optimality of block thresholded wavelet estimators. *Statistica Sinica*, 9:33–50, 1999.
- [34] P. Hall and B. A. Turlach. Interpolation methods for nonlinear wavelet regression with irregularly spaced design. *The Annals of Statistics*, 25:1912–1925, 1997.
- [35] G. Kerkyacharian and D. Picard. Thresholding in learning theory. *Constructive Approximation*, 26(2):173–203, August 2007.
- [36] Gérard Kerkyacharian and Dominique Picard. Regression in random design and warped wavelets. *Bernoulli*, 10(6):1053–1105, 2004.
- [37] M. Kohler. Nonlinear orthogonal series estimates for random design regression. *Journal of statistical Planning and Inference*, 115:491–520, 2003.
- [38] A. Kovac and B. W. Silverman. Extending the scope of wavelet regression methods by coefficient-dependent thresholding. *Journal of the American Statistical Association*, 95:172–183, 2000.

- [39] S. Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, second edition, 1998.
- [40] V. Maxim. Denoising signals observed on a random design. In *Fifth AFA-SMAI Conference on Curves and Surfaces*, 2002.
- [41] M. Pensky and B. Vidakovic. On non-equally spaced wavelet regression. *Ann. Inst. Statist. Math. Soc.*, 53:681–690, 2001.
- [42] D. Picard and K. Tribouley. Adaptive confidence interval for pointwise curve estimation. *The Annals of Statistics*, 28(1):298–335, 2000.
- [43] J. Romberg, H. Choi, and R. Baraniuk. Bayesian tree-structured image modeling using wavelet domain hidden Markov models. *IEEE Transactions on Image Processing*, 10(7):1056–1068, 2001.
- [44] S. Sardy, D. B. Percival, A. G. Bruce, H-Y Gao, and W. Stuetzle. Wavelet de-noising for unequally spaced data. *Statistics and Computing*, 9:65–75, 1999.
- [45] M. L. Stein. Spline smoothing with an estimated order parameter. *The Annals of Statistic*, 21(3):1522–1544, 1993.
- [46] E. Vanraes, M. Jansen, and A. Bultheel. Stabilized wavelet transforms for non-equispaced data smoothing. *Signal Processing*, 82(12):1979–1990, 2002.
- [47] B. Vidakovic. *Statistical Modeling by Wavelets*. Wiley-Interscience, New York, 1999.
- [48] X. W. Wang and A. T. A. Wood. Empirical Bayes block shrinkage of wavelet coefficients via the non-central χ^2 distribution. Technical Report 03–01, University of Nottingham, Division of Statistics, 2003.