Electronic Journal of Statistics
Vol. 5 (2011) 829–855
ISSN: 1935-7524
DOI: 10.1214/11-EJS628

Ideal denoising within a family of tree-structured wavelet estimators

Florent Autin*

Université d'Aix-Marseille 1, CNRS UMR 6632,
C.M.I., 39 rue F. Joliot Curie, 13453 Marseille Cedex 13
e-mail: autin@cmi.univ-mrs.fr

Jean-Marc Freyermuth† and Rainer von Sachs†

Université Catholique de Louvain,
Institut de Statistique, Biostatistique et Sciences Actuarielles,
Louvain la Neuve, Belgique
e-mail: Jean-Marc.Freyermuth@uclouvain.be; rvs@uclouvain.be

Abstract: We focus on the performances of tree-structured wavelet estimators belonging to a large family of keep-or-kill rules, namely the *Vertical Block Thresholding family*. For each estimator, we provide the maximal functional space (maxiset) for which the quadratic risk reaches a given rate of convergence. Following a discussion on the maxiset embeddings, we identify the ideal estimator of this family, that is the one associated with the largest maxiset. We emphasize the importance of such a result since the ideal estimator is different from the usual (plug-in) estimator used to mimic the performances of the Oracle. Finally, we confirm the good performances of the ideal estimator compared to the other elements of that family through extensive numerical experiments.

AMS 2000 subject classifications: 62G05, 62G20, 41A25, 42C40, 65T60.
Keywords and phrases: Besov spaces, curve estimation, CART, maxiset and oracle approaches, rate of convergence, thresholding methods, tree structure, wavelet estimators.

Received January 2011.

1. Introduction

Wavelet methods are known to be powerful in nonparametric estimation of functions. Indeed, the information of a function is localized in a few large wavelet coefficients for a wide range of function classes. This is the key-point to understanding why Hard and Soft thresholding methods perform well. These methods introduced by Donoho and Johnstone [13] consist in estimating the function by using the empirical wavelet coefficients which are larger than a chosen threshold value. In particular, these estimators were shown to be near optimal over

*Research partly supported by A.N.R. project *Oracle*.
†Financial support from the contract "Projet d'Actions de Recherche Concertées" nr. 07/12/002 of the "Communauté française de Belgique", granted by the "Académie universitaire Louvain", is gratefully acknowledged.

Besov spaces while they are adaptive for the regularity parameter (see Donoho and Johnstone [13, 14]). As mentioned by Autin [3] such thresholding rules are *elitist* in the sense that small empirical wavelet coefficients are not used in the reconstruction of the function.

Recent developments in wavelet thresholding have shown that elitist procedures can be outperformed in both theoretical and practical way by methods which refine the choice of the wavelet coefficients to be used in the reconstruction. This refined choice makes use of information from neighbored coefficients, e.g., block thresholding methods (see among others Cai [9], Autin [3, 5]) or impose that the empirical coefficients used for the reconstruction of the signal are arranged over a rooted connected tree (see Baraniuk [8], Autin [4]). We denote the latter as Tree Structured Wavelets (TSW) estimators. Interest in TSW already appeared in the works of Donoho [14] and Engel [15, 16]. In particular they pointed out the connection between TSW and CART. TSW have been proved useful in curve denoising (see among others Jansen [19], Lee [20], Autin [4]) but their interest goes beyond as they furnish specific abilities to be used in signal processing (Shapiro [23], Cohen et al. [10]), edge detection (Sun et al. [24]), construction of statistical models in the coefficient domain (Freyermuth et al. [17])... This paper is not in the line of comparing TSW to other well established methods in curve denoising. Its aim is to suggest a formal treatment of an algorithm that is closely related to the dyadic CART and to emphasize an important aspect about the selection of the ideal procedure among a 'natural' family of TSW estimators. The family of estimators that we will consider includes as special cases two popular TSW estimators, the *CART-like estimator* obtained by model selection (see Donoho [14] and Engel [15]) and the *Hard Tree estimator* (see Autin[4])).

The Figures 1-4 show an example of a reconstruction of the *Blip function* using these methods (defined in Section 3) and the associated wavelet coefficient magnitudes (the darker, the larger the coefficient magnitude).

Looking at the positions of the large wavelet coefficients in the Figure 1, we notice a hierarchical structure between them. In particular, there are large wavelet coefficients that persist across scales at the location of the singularity. The two methods of reconstruction give estimators in the Figures 3 and 4 which appear to be close to the target function. Note that the sets of empirical wavelet coefficients used by the two methods are embedded (see Proposition 3.1). In particular, the cardinality of the set of empirical coefficients used in the reconstruction of the CART-like estimator (Figure 3) is smaller than the one of the Hard Tree (Figure 4), quantitative results of section 6 support this remark. These facts will be discussed and interpreted throughout the paper.

Donoho [14] proves that estimation under tree constraints can be solved by a CART-like algorithm. A Tree-Oracle estimator is obtained after a recursive-per-level method based on the comparison of the l_2-mean of vertical blocks of the true wavelet coefficients with the standard deviation. This is the best possible tree-structured estimator minimizing the L_2-risk which is unknown in practice but its performances can be mimicked by plugging-in observed values of the wavelet coefficients and adjusting the threshold value upwards to account for

the noise. This estimator is proven to be near-minimax and to perform well in practice. However, in this paper, adopting the maxiset approach, we show that we should not compare local ℓ_2-norms of empirical wavelet coefficients with the threshold but rather local ℓ_∞-norms.

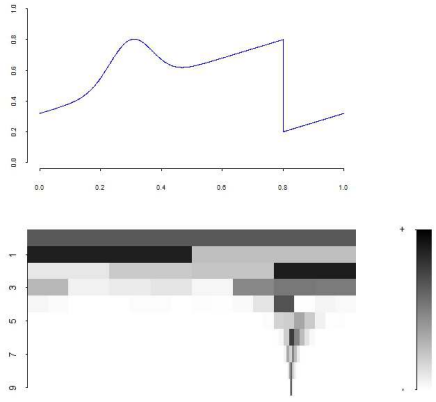

FIG 1. *True function.*

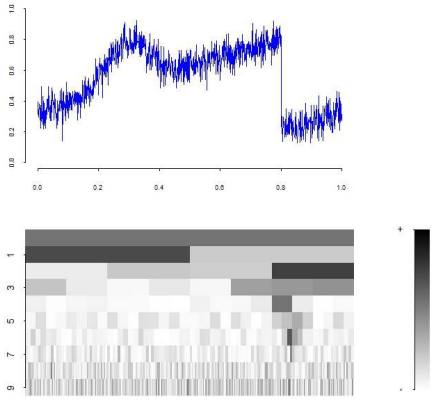

FIG 2. *Noisy data.*

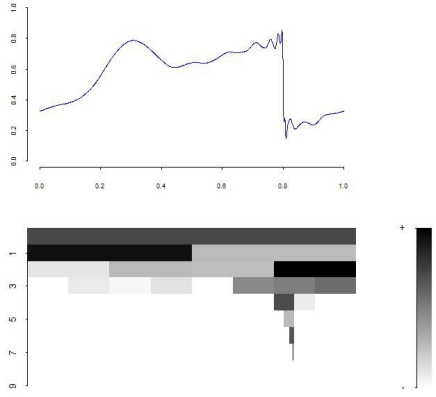

FIG 3. *CART-like estimator.*

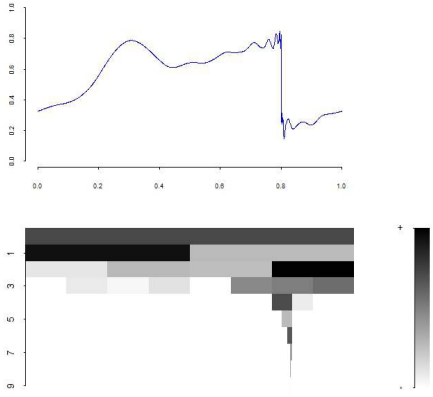

FIG 4. *Hard Tree estimator.*

To reach this goal, that is the main result of this paper, we first introduce in Section 3 a general family of TSW estimators so-called *Vertical Block Thresholding* (VBT) which includes the two previous estimators as special cases. Then, we compute the set of all the functions well estimated by each estimator in that family. Namely, we consider the maxiset approach introduced by Cohen et al

[11]. Its basics are presented in Section 4. This theory is applied in Section 5 to find the *ideal estimator* of the VBT family, that is the one for which the set of well-estimated functions is the largest functional space. The main result of our paper is expressed in Theorem 5.1 and its Corollary 5.1. Section 6 proposes numerical experiments to confirm the superiority of the ideal estimator using as a benchmark the informative results obtained by the Tree-Oracle estimator. Finally after brief conclusive remarks in Section 7, Section 8 presents the proofs of our main results.

2. Model and background

2.1. Wavelet setting and model

Let us consider a compactly supported wavelet basis of $L_2([0,1])$ with V vanishing moments ($V \in \mathbb{N}^*$) which has been previously periodized $\{\phi, \psi_{jk}, j \in \mathbb{N}, k \in \{0,\dots,2^j-1\}\}$. Examples of such bases are given in [12]. Any function $f \in L_2([0,1])$ can be written as follows:

$$f = \alpha\phi + \sum_{j=0}^{\infty}\sum_{k=0}^{2^j-1}\theta_{jk}\psi_{jk}. \tag{1}$$

The coefficient α and the components of $\theta = (\theta_{jk})_{jk}$ are respectively the scaling/wavelet coefficients of f. They correspond to the L_2-scalar products between f and the scaling/wavelet functions ϕ and ψ_{jk}.

We consider the sequential version of the Gaussian white noise model: we dispose of observations of these coefficients which are assumed to be realizations of independent random variables:

$$\begin{aligned}\hat{\alpha} &= \alpha + \epsilon\xi,\\ \hat{\theta}_{jk} &= \theta_{jk} + \epsilon\xi_{jk},\end{aligned} \tag{2}$$

where ξ, ξ_{jk} are i.i.d. $\mathcal{N}(0,1)$, $0 < \epsilon < 1$ is supposed to be the noise level, and where the sequence $(\theta_{jk})_{j,k}$ is sparse, meaning that only a small number of *large* coefficients contain nearly all the information about the signal. That motivates the use of keep-or-kill estimators, for which we recall the Hard thresholding estimator:

$$\hat{f}_{\mathcal{S}} = \hat{\alpha}\phi + \sum_{(j,k)\in\mathcal{S}}\hat{\theta}_{jk}\psi_{jk}, \tag{3}$$

where $\mathcal{S} = \left\{(j,k)\,; j \in \mathbb{N}, j < j_{\lambda_\epsilon}; 0 \le k < 2^j; |\hat{\theta}_{jk}| > \lambda_\epsilon\right\}$. If $\mathcal{S}$ is non empty, it forms an *unstructured* set of indices of 'large' wavelet coefficients (in the sequel, by 'large' coefficients, we understand those which belong to $\mathcal{S}$). Here,

- $\lambda_\epsilon = m\,\epsilon\sqrt{\log(\epsilon^{-1})}, \quad 0 < m < \infty,$
- j_λ is the integer such that $2^{-j_\lambda} \le \lambda^2 < 2^{1-j_\lambda}$ $(0 < \lambda < 1)$. For $\lambda_\epsilon < 1$, $j_{\lambda_\epsilon} - 1$ is the finest level up to which we consider the empirical wavelet coefficients to reconstruct the signal f.

This term by term thresholding does not take into account the information that give us the clusters of wavelet coefficients that we observed in the Figure 1. But this knowledge has the practical application that, on the one hand, we would not use in the reconstruction a large isolated wavelet coefficient because it is not likely to be part of the signal; on the other hand, a small coefficient in the neighborhood of large coefficients would be kept. This motivates the use of refined thresholding methods such as the *tree-structured* wavelets (Autin [4] and Baraniuk [8]) which we describe in the next section.

2.2. Tree-structured wavelet estimators

Tree-structured wavelet (TSW) estimators are based on the hierarchical interpretation of the wavelet expansion (1). The periodized wavelets $\{\psi_{jk}\}_{jk}$ are arranged over a nested multiscale structure such that the support of each ψ_{jk} contains the supports of $\psi_{j+1,2k}$ and $\psi_{j+1,2k+1}$. This induces a hierarchy among the wavelet coefficients which can be represented over a binary tree rooted in $(0,0)$ (see Figure 5). Hence, at the location of a singularity in the signal, we observe the persistence of large wavelet coefficients over all scales (see Figure 1).

Therefore, considering the wavelet coefficients as a multiresolution sequence provides additional information which we aim to benefit from by imposing a tree/hereditary constraint. The hereditary constraint requires that the set of non zero wavelet coefficients after thresholding forms a *connected rooted subtree*. In other words, it cannot include an empirical wavelet coefficient unless all its ancestors (defined in equation (4) below) are large.

We denote as $\mathcal{T}_J$ the binary tree of depth J for which the nodes are the couples of indices (j,k), $\left(0 \le j < J, k \in \{0,\dots,2^j-1\}\right)$ (see the Figure 5). For any couple of indices (j,k), following Engel [16], we define the set which contains:

- its ancestors

$$\mathcal{P}(j,k) = \{(j-m, \lceil k/2^m \rceil)\,;\ m = 0,\dots,j\}\,, \tag{4}$$

  where $\lceil x \rceil$ denotes the smallest integer smaller than or equal to x;
- its descendants

$$\begin{aligned} \mathcal{C}(j,k) = \{&(j,k)\,,(j+1,2k)\,,(j+1,2k+1)\,,\\ &(j+\mu,2^\mu k)\,,\dots,(j+\mu,2^\mu(k+1)-1)\,,\dots\,;\ \mu = 2,3,\dots\}\,. \end{aligned} \tag{5}$$

Note that to each node of indices (j,k) correspond $2^{j'-j}$ **descendants** at levels j' $(j \le j' < J)$ and $j+1$ **ancestors**.

Remark 2.1. When using smooth wavelets, the presence of an edge 'generates' several large wavelet coefficients at each scale, due to the overlapping supports of the wavelets. This idea is the leitmotiv of block thresholding methods (intra-scales), but could also be applied to TSW. In such a case, the heredity constraint would mean that a node at the scale j in the tree have more than two descendants at the scale $j+1$ (see Baraniuk [8], Averkamp and Houdré [7]). In this paper, we

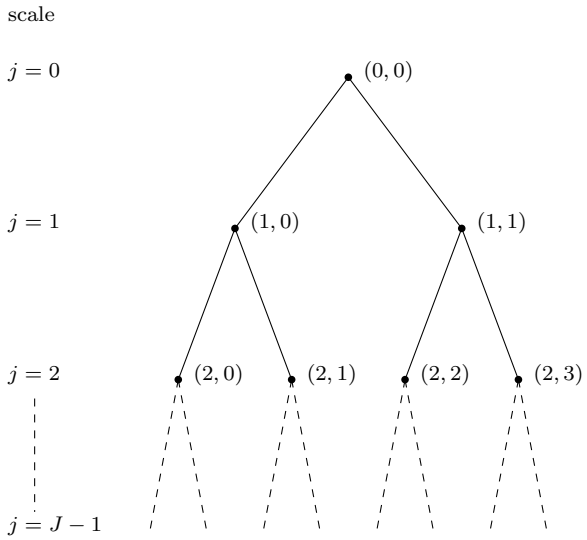

FIG 5. *Binary tree of depth J.*

consider the situation where the ancestors have two descendants and therefore, we naturally associate binary trees to wavelet coefficient sequences.

Let us now introduce the definition of a *tree-structured estimator* in our setting.

Definition 2.1. We call *tree-structured* estimator of a signal f satisfying (1) any keep-or-kill estimator

$$\hat{f}_{\mathcal{T}} = \hat{\alpha}\phi + \sum_{(j,k)\in\mathcal{T}} \hat{\theta}_{jk}\psi_{jk},$$

where the set of the indices $\mathcal{T}$ satisfies the *hereditary constraint* formulated in Engel [15], that is, if (j, k) is in $\mathcal{T}$ then all its ancestors are in $\mathcal{T}$.

In the sequel we denote by $|\mathcal{T}|$ the cardinality of the tree $\mathcal{T}$, i.e., the number of active wavelet coefficients kept in the estimator $\hat{f}_{\mathcal{T}}$. Analogously to the Hard thresholding estimator defined in (3), we only use the empirical wavelet coefficients on levels smaller than j_{λ_ϵ}.

Donoho [14] used the Oracle approach to propose a tree-structured near optimal estimator. His idea was to find a tree-structured estimator which mimics the optimal risk $\mathcal{R}_\epsilon(f)$ only attained by the "Tree-Oracle", that is

$$\mathcal{R}_\epsilon(f) = \min_{\hat{f}_{\mathcal{T}},\ \mathcal{T}\subseteq\mathcal{T}_{j_{\lambda_\epsilon}}} \mathbb{E}\|\hat{f}_{\mathcal{T}} - f\|_2^2 = \min_{\mathcal{T}\subseteq\mathcal{T}_{j_{\lambda_\epsilon}}} \left(\sum_{(j,k)\notin\mathcal{T}} \theta_{jk}^2 + \epsilon^2\ (|\mathcal{T}| + 1) \right),$$

where the minimum is taken over all the tree-structured estimators. Donoho [14] showed that the solution of this optimization problem under a tree constraint has an inheritance property and therefore can be solved by a CART-like algorithm applied to the true wavelet coefficients using ϵ as the threshold value. In the sequel $\mathcal{T}^{\mathcal{O}}$ stands for the set of coefficients selected by the Tree-Oracle. In practice, $\hat{f}_{\mathcal{O}} = \hat{\alpha}\phi + \sum_{(j,k)\in\mathcal{T}^{\mathcal{O}}} \hat{\theta}_{jk}\psi_{jk}$ is not available. Donoho [14] proposed to consider the estimator $\hat{f}_{cart}$ which minimizes the empirical complexity, that is

$$\hat{f}_{cart} = \underset{\hat{f}_{\mathcal{T}},\ \mathcal{T}\subseteq\mathcal{T}_{j_{\lambda_\epsilon}}}{\arg\min} \left(\sum_{(j,k)\notin\mathcal{T}} \hat{\theta}_{jk}^2 + \lambda_\epsilon^2\ (|\mathcal{T}|+1) \right).$$

Furthermore it was shown that the risk of $\hat{f}_{cart}$ is of the same order as the optimal risk up to a logarithmic term. Precisely, for m large enough, there exists a constant $K > 0$ not depending on ϵ such that for any $f \in L_2([0,1])$:

$$\mathbb{E}\|\hat{f}_{cart} - f\|_2^2 \leq K \log(\epsilon^{-1})\mathcal{R}_\epsilon(f).$$

3. Vertical block thresholding estimators

Let us now define a general Vertical Block Thresholding (VBT) estimator $\hat{f}_p$, for any $1 \leq p \leq \infty$, as follows:

Definition 3.1 ((λ,p)-VBT-method)**.** For given $0 < \lambda < 1$, $1 \leq p \leq \infty$ and any set of real numbers $\left(\theta_{jk}, 0 \leq j < j_\lambda, 0 \leq k < 2^j\right)$ we define the sets of indices, $\mathcal{E}_{jk}(\theta,\lambda)$, for any (j,k), iteratively as follows:

- For $j = j_\lambda - 1$ and for any k,

$$\begin{aligned} \mathcal{E}_{jk}(\theta,\lambda) &= \{(j,k)\} \quad \text{if } |\theta_{jk}| > \lambda, \\ \mathcal{E}_{jk}(\theta,\lambda) &= \emptyset \quad \text{otherwise.} \end{aligned}$$

- For any $0 \leq j < j_\lambda - 1$ and any k, we put

$$\mathcal{F}_{jk}(\theta,\lambda) := \{(j,k)\} \cup \{\mathcal{E}_{j+1,k'}(\theta,\lambda); (j,k) \in \mathcal{P}(j+1,k')\}.$$

  Then

$$\begin{aligned} \mathcal{E}_{jk}(\theta,\lambda) &= \mathcal{F}_{jk}(\theta,\lambda) \quad \text{if } \|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda)\|_p > \lambda, \\ \mathcal{E}_{jk}(\theta,\lambda) &= \emptyset \quad \text{otherwise,} \end{aligned}$$

where

$$\begin{aligned} \|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda)\|_p &:= \left(\frac{1}{\#\mathcal{F}_{jk}(\theta,\lambda)} \sum_{(j',k')\in\mathcal{F}_{jk}(\theta,\lambda)} |\theta_{j'k'}|^p \right)^{1/p} \quad \text{for } 1 \leq p < \infty, \\ \|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda)\|_\infty &:= \max_{(j',k')\in\mathcal{F}_{jk}(\theta,\lambda)} |\theta_{j'k'}|. \end{aligned}$$

The (λ, p)-VBT-method is illustrated on an example in the Appendix 8.1. For any real valued $p \in [1, \infty]$, it is associated to the following estimator:

$$\begin{aligned} \hat{f}_p &:= \hat{f}_{\mathcal{T}^p} := \hat{\alpha}\phi + \sum_{(j,k)\in\mathcal{T}^p} \hat{\theta}_{jk}\psi_{jk} \\ &= \hat{\alpha}\phi + \sum_{j\in\mathbb{N}, j<j_{\lambda_\epsilon}} \sum_{k=0}^{2^j-1} \hat{\theta}_{jk}\, \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\hat{\theta} \,/\, \mathcal{F}_{j'k'}(\hat{\theta}, \lambda_\epsilon)\|_p > \lambda_\epsilon \right\} \psi_{jk}, \end{aligned} \tag{6}$$

where $\mathcal{T}^p$ is the set of empirical wavelet coefficients used in the reconstruction following the VBT method based on ℓ_p-norms.

We encourage the reader to check that for $p = 2$ (resp. $p = \infty$) the estimator $\hat{f}_p$ is the CART-like estimator (resp. the Hard Tree estimator). These estimators have an interesting interpretation using the terminology of wavelet thresholding. At each node (j, k), we consider the coefficient at (j, k) and those which survive the previous step (i.e., at scale $j+1$). They form a connected **subtree** $\mathcal{F}_{jk}(\theta, \lambda)$ of $\mathcal{C}(j, k)$ rooted to (j, k). The decision to keep-or-kill this block of coefficients depends on its ℓ_p-mean which is compared with the threshold λ_ϵ. We remark that unlike other block thresholding methods there is no need for controlling the size of the blocks by any additional parameter.

From now on, we will study the performance of these VBT estimators to address the following question: is the ℓ_2-norm the best choice to consider among $\hat{f}_p$ estimators $(1 \le p \le \infty)$? In the next sections we use the maxiset approach to prove that the answer is NO.

Define the *Vertical Block Thresholding* family $(\mathcal{VBT}_\epsilon)$ as

$$\mathcal{VBT}_\epsilon = \left\{ \hat{f}_p,\ 1 \le p \le \infty \right\}.$$

At first glance, as $1 \le p \le \infty$ is real-valued, this family of estimators $\mathcal{VBT}_\epsilon$ seems to be uncountable. But it is not since the estimators are clearly tree-structured. More precisely,

Proposition 3.1. *For any $1 \le p \le q$ and for given λ_ϵ,*

1. *$\mathcal{T}^p$ and $\mathcal{T}^q$ constitute trees of indices,*
2. *$\mathcal{T}^p \subseteq \mathcal{T}^q$,*
3. *$\mathcal{T}^\infty$ is the smallest tree (in terms of cardinality) which contains all 'large' empirical wavelet coefficients.*

According to the previous proposition, we deduce that $\mathcal{VBT}_\epsilon$ is a family of tree-structured estimators with embedded trees. The larger p, the bigger the tree.

4. Maxiset approach

In this section we recall the maxiset approach. The maxiset point of view has been proposed by Cohen et al. [11] to measure the performance of estimators.

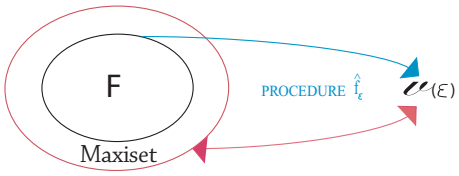

FIG 6. *Maxiset and Minimax*

For a given estimator $\hat{f}$ and a chosen sequence $v = (v_\epsilon)_\epsilon$ tending to 0 when ϵ goes to 0, this approach consists in providing the set of all the functions (maxiset) for which the rate of convergence of the quadratic-risk of $\hat{f}$ is at least as fast as v.

In this setting, the functional space $\mathcal{G}$ will be called maxiset of $\hat{f}$ for the rate of convergence v if and only if the following property holds:

$$\sup_{0<\epsilon<1} v_\epsilon^{-1}\mathbb{E}\|\hat{f} - f\|_2^2 < \infty \iff f \in \mathcal{G}.$$

From now on we shall adopt the following notation: $MS\big(\hat{f}, (v_\epsilon)_\epsilon\big) = \mathcal{G}$.

Note that, if $\hat{f}$ reaches the minimax rate v on a functional space F, then $F \subseteq MS(\hat{f}_\epsilon, (v_\epsilon)_\epsilon)$. Hence, the maxiset approach appears to be more optimistic than the minimax one. The following scheme illustrates this idea.

The maxiset setting allows to compare efficiently different procedures. This approach lies on the fact that the larger the maxiset, the better the procedure. Following Kerkyacharian and Picard [21, 22] and Autin [3], this way to measure the performance of procedures is often successfully applicable to discriminate procedures that are equivalent in the minimax sense, and to give theoretical explanations for some phenomena observed in practice (see Section 6).

5. Main results

5.1. Functional spaces: Definitions and embeddings

In this paragraph, we characterize the functional spaces which shall appear in the maxiset study of our estimators. Recall that, for later use of these functional spaces, we shall consider wavelet bases with V vanishing moments.

Definition 5.1. Let $1 \le \gamma \le \infty$, $0 < u < V$. We say that a function $f \in L_2([0,1])$ belongs to the Besov space $\mathcal{B}^u_{\gamma,\infty}$ if and only if:

$$\sup_{J\in\mathbb{N}} 2^{\gamma(u-\frac{1}{\gamma}+\frac{1}{2})J} \sum_{j\ge J} \sum_{k=0}^{2^j-1} |\theta_{jk}|^\gamma < \infty.$$

Besov spaces naturally appear in estimation problems (see Autin [3] and Cohen et al. [11]). These spaces characterize the functions for which the energy of wavelet coefficients on levels larger than J ($J \in \mathbb{N}$) is decreasing exponentially in J. We recall some properties of embeddings. Let $1 \leq \gamma \leq \gamma' \leq \infty$.

$$\begin{aligned} \mathcal{B}^u_{\gamma,\infty} &\subsetneq \mathcal{B}^{u'}_{\gamma,\infty} \text{ for } u > u', \\ \mathcal{B}^u_{\gamma,\infty} &\subsetneq \mathcal{B}^{u'}_{\gamma',\infty} \text{ for } u' - \frac{1}{\gamma'} < u - \frac{1}{\gamma}. \end{aligned}$$

For an overview of these spaces, see Härdle et al. [18].

Let us now define a new function space which is the key to our results:

Definition 5.2. Let $0 < r < 2$ and $1 \leq p \leq \infty$. We say that a function f belongs to the space $W_{r,p}$ if and only if:

$$\sup_{0<\lambda<1} \lambda^{r-2} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \ / \ \mathcal{F}_{j'k'}(\theta,\lambda)\|_p \leq \lambda \right\} < \infty.$$

First, note that the larger r, the larger the functional space; second, in contrast to weak Besov spaces (see Cohen et al. [11] for an explicit definition) which appear in the maxiset results for Hard and Soft thresholding estimators, the spaces $W_{r,p}$ ($0 < r < 2$) are not invariant under permutations of wavelet coefficients within each scale. This property makes them able to distinguish functions according to the "clustering properties" of their wavelet coefficients. These functional spaces are quite large as suggested by our following Proposition 5.1.

Proposition 5.1. *For any $0 < s < V$, and any $2 \leq p \leq \infty$*

$$\mathcal{B}^s_{2,\infty} \subseteq W_{\frac{2}{1+2s},p}. \tag{7}$$

Our following Proposition 5.2 shows that, for the same parameter r ($0 < r < 2$), the functional spaces $W_{r,p}$ ($p \geq 1$) are embedded. The larger p the larger $W_{r,p}$. Moreover, in Theorem 5.1, the intersections of function spaces appearing in equation (9) below are shown to be directly related to the maxisets of the estimators $\hat{f}_p \in \mathcal{VBT}_\epsilon$.

Proposition 5.2. *For any $1 \leq p < q$ and any $0 < r < 2$, we have the following embeddings of spaces:*

$$W_{r,p} \subseteq W_{r,q}, \tag{8}$$

$$\mathcal{B}^u_{2,\infty} \cap W_{\frac{2}{1+2s},2} \subsetneq \mathcal{B}^u_{2,\infty} \cap W_{\frac{2}{1+2s},\infty}, \quad \text{for any } u < \frac{s}{1+2s}. \tag{9}$$

5.2. Maxiset results

In this paragraph we provide the maximal space (maxiset) of any $\hat{f}_p \in \mathcal{VBT}_\epsilon$ associated with the rate $\lambda_\epsilon^{\frac{4s}{1+2s}}$ ($s > 0$). This corresponds to the optimal minimax

rate over Besov spaces $\mathcal{B}^s_{\gamma,\infty}$, $s > \frac{1}{\gamma} - \frac{1}{2}$ under the L_2-risk with a logarithm term. In the maxiset context, this is a traditional choice which has nothing to do with a price to pay for adaptivity. It gives a maxiset that is simpler to interpret than the one we would get using the exact optimal minimax rate. And, for our purpose, this is unnecessary complications since the choice of the rate will not make any difference in the identification of the maxiset-ideal method among the (λ, p)-VBT family.

Theorem 5.1. *Let $s > 0$, $1 \leq p \leq \infty$ and $\lambda_\epsilon = m\ \epsilon\sqrt{\log(\epsilon^{-1})}$. For any $m \geq 4\sqrt{3}$, we have the following equivalence:*

$$\sup_{0<\epsilon<1} \lambda_\epsilon^{-\frac{4s}{1+2s}}\ \mathbb{E}\|\hat{f}_p - f\|_2^2 < \infty \iff f \in \mathcal{B}^{\frac{s}{1+2s}}_{2,\infty} \cap W_{\frac{2}{1+2s},p},$$

that is to say, using the maxiset notation, $MS(\hat{f}_p, (\lambda_\epsilon^{\frac{4s}{1+2s}})_\epsilon) = \mathcal{B}^{\frac{s}{1+2s}}_{2,\infty} \cap W_{\frac{2}{1+2s},p}$.

Note that these maxisets are large functional spaces since from Proposition 5.1 we deduce that the functional space $\mathcal{B}^{\frac{s}{1+2s}}_{2,\infty} \cap W_{\frac{2}{1+2s},p}$ contains the space $\mathcal{B}^s_{\gamma,\infty}$ for any $\gamma \geq \min(2, s^{-1})$. Hence this maxiset contains many functions that cannot be reconstructed by linear procedures at the rate $\lambda_\epsilon^{-\frac{4s}{1+2s}}$ (for more details see Autin et al. [6]).

We now state the main result of the paper through the following corollary.

Corollary 5.1. *Let $\lambda_\epsilon = m\ \epsilon\sqrt{\log(\epsilon^{-1})}$, with $m \geq 4\sqrt{3}$, then $\hat{f}_\infty$ is the ideal estimator in the maxiset sense among the $\mathcal{VBT}_\epsilon$ family.*

Proof. Theorem 5.1 establishes the maxiset associated with any estimator $\hat{f}_p$ built with the $(\lambda_\epsilon, p)-$VBT method. According to (8) of Proposition 5.2 we deduce that the maxisets of these estimators are embedded and that the largest maxiset is the one associated with $\hat{f}_\infty$ (Hard Tree estimator). □

Although $\hat{f}_2$ was shown to be very powerful by using the Oracle approach (see Donoho [14]), $\hat{f}_\infty$ is better in the maxiset sense. This result is interpretable as the necessity to keep all empirical wavelet coefficients larger than λ_ϵ in the reconstruction. Missing some of them has a huge maxiset-cost which corresponds to the exclusion of many functions estimated at the same rate. Moreover this suggests to include not only all the 'large' empirical wavelet coefficients but also some well chosen small ones. Autin [3] already underlies this important issue through what he calls *cautious rules*. In particular he proved that $\hat{f}_\infty$ outperforms Hard and Soft thresholding estimators in the maxiset sense.

6. Numerical experiments

We first introduce the notations of the nonparametric model we are dealing with:

$$Y_i = f\left(\frac{i}{N}\right) + \sigma\zeta_i,\ 1 \leq i \leq N,\ \zeta_i \text{ are i.i.d. } \mathcal{N}(0,1)\,. \tag{10}$$

We refer the reader to the classical literature (e.g., Tsybakov [25]) for details about the equivalence between this nonparametric regression model and the sequence model given by equation (2). We only recall that the noise level ϵ is such that $\epsilon = \frac{\sigma}{\sqrt{N}}$.

This section proposes numerical experiments designed to check whether the choice of the ℓ_∞ norm should be preferred as claimed by the corollary 5.1. The previous theory does not model all the complexity encountered in practice with the choice of the wavelet function, of the primary resolution scale, etc. Therefore, we choose a classical setting for numerical experiments, using Daubechies 8 Least Asymmetric. In addition, our theoretical model do not consider neither method-dependent threshold nor data-driven threshold. Hence, for these experiments, we naturally decide to use the universal threshold value for all methods, i.e., $\hat{\lambda} = \hat{\sigma}\sqrt{2N^{-1}\log N}$. We follow a standard approach to estimate σ by the Median Absolute Deviation (MAD) divided by 0.6745 over the wavelet coefficients at the finest wavelet scale $J-1$ (see e.g., Vidakovic [26]).

We generate the data sets from a large panel of functions often used in wavelet estimation studies (Antoniadis et al. [2]) with various Signal to Noise Ratios $SNR = \{5, 10, 15, 20\}$ and sample sizes $N = \{512, 1024, 2048\}$. We define the SNR as the logarithm decibel scale of the ratio of the standard deviation of the function values to the standard deviation of the noise. We compute the Integrated Squared Error of the estimators $\hat{f}_p,\ p \in \{1, 2, 5, 10, \infty\}$ at the m-th Monte Carlo replication ($ISE^{(m)}(\hat{f}_p),\ 1 \le m \le M$) as follows:

$$ISE^{(m)}\left(\hat{f}_p\right) = \frac{1}{N}\sum_{i=1}^{N}\left(\hat{f}_p^{(m)}\left(\frac{i}{N}\right) - f\left(\frac{i}{N}\right)\right)^2. \tag{11}$$

The Mean ISE is $MISE(\hat{f}_p) = M^{-1}\sum_{m=1}^{M} ISE^{(m)}(\hat{f}_p)$ and its standard error is $SE(\hat{f}_p) = M^{-\frac{1}{2}}\ \hat{\sigma}_{ISE(\hat{f}_p)}$.

In this context we are particularly interested in comparing the results of the estimators for $p = 2$ with $p = \infty$. In addition to that, we propose to test the null hypothesis: $H_0 : MISE(\hat{f}_2) = MISE(\hat{f}_\infty)$ against the alternative: $H_A : MISE(\hat{f}_2) \neq MISE(\hat{f}_\infty)$ using the Wilcoxon Signed-Rank test for paired samples. Therefore, we can choose the number of Monte Carlo replications M in order to ensure that the power of the test at level I-error of 5% is about 80% to detect a difference in means of about 1% of $MISE(\hat{f}_\infty)$.

There are numerous connections between keep-or-kill estimation and hypothesis testing (see Abramovich et al. [1]). We will get an interesting insight into these methods by computing the number of false positives/negatives (i.e., type I/II errors). To do so, we compare the set of indices of wavelet coefficients kept by each estimators ($\mathcal{T}^p$) and by the Tree-Oracle ($\mathcal{T}^{\mathcal{O}}$) with the one of the keep-or-kill Oracle estimator $\mathcal{S}^{\mathcal{O}} = \left\{(j,k)\,; j \in \mathbb{N}, j < j_{\lambda_{\frac{\sigma}{\sqrt{N}}}}\,; 0 \le k < 2^j; |\theta_{jk}| > \frac{\sigma}{\sqrt{N}}\right\}$. In addition, we give in Tables 1 and 2 the size (number of nodes) of the trees.

The results suggest similar behavior for different values of N and SNR. To keep clear the presentation of the results, we only report those for $N = 1024$ and $SNR = 10$ in Tables 1-2 and summarize the MISE results in Figure 7.

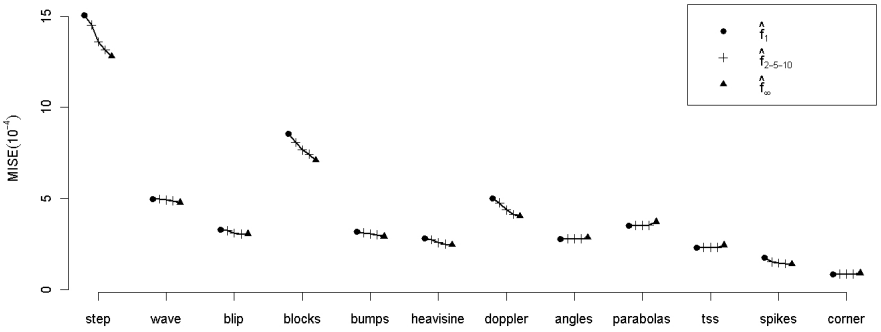

FIG 7. *MISE of* (λ, p)*-VBT estimator for five different values of* p *for estimating various functions with a SNR equal to 10.*

TABLE 1

MISE (10^{-4})*, number of false positives/negatives and average size of the tree (number of non zero empirical wavelet coefficients in the estimator).*

	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_5$	$\hat{f}_{10}$	$\hat{f}_\infty$	Tree-Oracle
	Function: Step					
MISE	15.06	14.52	13.60	13.15	12.80	4.45
False positives	0.01	0.01	0.02	0.07	0.60	1
False negatives	22.27	21.62	20.51	19.93	19.20	2
Size	19.73	20.39	21.51	22.14	23.41	41
	Function: Wave					
MISE	4.99	4.98	4.92	4.87	4.79	1.37
False positives	8.01	8.01	8.01	8.05	8.38	8
False negatives	29.34	29.25	28.87	28.52	27.70	0
Size	26.67	26.76	27.14	27.52	28.68	56
	Function: Blip					
MISE	3.31	3.25	3.12	3.05	3.06	1.39
False positives	0.00	0.00	0.00	0.04	0.73	0
False negatives	14.77	14.58	14.19	13.94	13.52	0
Size	19.24	19.42	19.82	20.10	21.21	34
	Function: Blocks					
MISE	8.56	8.08	7.66	7.44	7.10	2.30
False positives	3.32	3.90	4.61	5.04	5.62	9
False negatives	68.65	67.33	65.64	64.49	62.63	1
Size	50.67	52.57	54.97	56.55	58.99	124
	Function: Bumps					
MISE	3.20	3.12	3.06	3.01	2.93	0.97
False positives	1.02	1.08	1.42	1.69	2.09	5
False negatives	66.95	66.40	65.48	64.64	63.10	2
Size	74.07	74.69	75.94	77.04	78.99	143
	Function: Heavisine					
MISE	2.82	2.74	2.59	2.50	2.47	1.26
False positives	0.01	0.01	0.01	0.06	0.95	0
False negatives	13.12	12.90	12.45	12.16	11.61	3
Size	9.88	10.11	10.56	10.90	12.34	20

TABLE 2
MISE (10^{-4}), *number of false positives/negatives and average size of the tree (number of non zero empirical wavelet coefficients in the estimator).*

	$\hat{f}_1$	$\hat{f}_2$	$\hat{f}_5$	$\hat{f}_{10}$	$\hat{f}_\infty$	Tree-Oracle
	Function: Doppler					
MISE	5.02	4.77	4.38	4.14	4.03	2.23
False positives	2.93	3.08	4.75	6.00	7.45	11
False negatives	22.88	22.60	21.72	21.11	20.57	5
Size	32.05	32.47	35.04	36.89	38.88	58
	Function: Angles					
MISE	2.80	2.80	2.80	2.80	2.87	1.32
False positives	0.01	0.02	0.07	0.14	0.82	1
False negatives	9.93	9.93	9.88	9.84	9.72	0
Size	19.08	19.09	19.18	19.30	20.10	30
	Function: Parabolas					
MISE	3.52	3.52	3.53	3.54	3.72	1.54
False positives	1.01	1.01	1.02	1.08	1.99	2
False negatives	7.66	7.66	7.64	7.62	7.51	0
Size	14.35	14.35	14.38	14.46	15.48	23
	Function: time.shift.sine					
MISE	2.32	2.32	2.32	2.33	2.45	1.09
False positives	0.01	0.01	0.01	0.06	0.86	0
False negatives	5.56	5.56	5.56	5.55	5.51	0
Size	17.45	17.45	17.46	17.51	18.36	23
	Function: Spikes					
MISE	1.77	1.54	1.47	1.44	1.41	0.62
False positives	0.84	1.00	1.02	1.05	1.31	1
False negatives	20.37	19.27	18.55	18.11	17.55	1
Size	37.47	38.73	39.47	39.94	40.76	57
	Function: Corner					
MISE	0.85	0.85	0.85	0.86	0.91	0.44
False positives	0.00	0.00	0.01	0.06	0.92	0
False negatives	6.44	6.44	6.44	6.43	6.35	1
Size	13.56	13.56	13.56	13.63	14.57	19

Comparing the MISE of $\hat{f}_2$ with $\hat{f}_\infty$ we observe the optimality of the latter for most of the test functions with sometimes important improvements, up to 16% for the function 'doppler'. In the other cases, the loss of $\hat{f}_\infty$ against $\hat{f}_2$ remains under 7%. More than that, for many of these functions we have a monotone decrease in the MISE as the value of p increases, reflecting the embeddings of the maxisets of the $\mathcal{VBT}_\epsilon$ estimators (see Section 5).

Looking at the number of false positives/negatives, we can check that $\hat{f}_\infty$ allows to reduce the number of false negatives with a comparatively small increase in the number of false positives yielding its good performances in terms of MISE. Comparing the results to those of the Tree-Oracle we observe that there are potentially huge improvements achievable by reducing the number of false negatives. Indeed, the number of active coefficients of the Tree-Oracle estimators (see Section 2.2), $\left|\mathcal{T}^{\mathcal{O}}\right|$ is about 25% to 110% larger than $|\mathcal{T}^{\infty}|$.

7. Conclusions

In this paper we introduced the family of the Vertical Block Thresholding estimators. We studied their performances under L_2-risk using the maxiset approach, and we identified the ideal procedure, that is the one obtained from the $(\lambda_\epsilon, \infty)$-VBT-method. The main message of this paper is that the ideal estimator is different from the classical one obtained by plugging-in empirical quantities in the Tree-Oracle which corresponds to the estimator built from the $(\lambda_\epsilon, 2)$-VBT-method. Indeed, compared to the latter one, the ideal estimator is able to reconstruct more functions at the chosen rate.

It is important to emphasize that we compared both theoretically and numerically all these estimators for a fixed threshold value. We have chosen to use the universal threshold value for the numerical experiments although it is known to be too conservative in practice, simply in order to use the most standard choice for our comparisons.

Our theoretical and numerical results emphasize the importance of reducing the number of false negatives while maintaining the number of false positives. In addition, the numerical experiments which implement the Tree-Oracle estimator show us the important potential in reducing the amount of false negatives. To do so, using these methods, we should either consider more complex hereditary constraints or allow lower threshold values. Indeed, large threshold values lead to suboptimal estimation of the localized structure in the underlying curve. It would be more convenient to use a minimum risk threshold rather than the universal threshold (cf. Jansen [19]) but, when used with Hard thresholding, the estimate often shows unappealing visual artifacts (spurious bumps) due to large wavelet coefficients at fine resolution scales generated from the random noise ("false positives"). In this context, and as part of future research, we expect the vertical block thresholding algorithms also for $p < \infty$ to be powerful as they adaptively keep-or-kill blocks of coefficients even if they contain coefficients larger than the threshold value. Hence, the control of false positives is not only achieved by the threshold value but by the algorithm too. The conclusive words for the present results is that practical application would require to optimize simultaneously over the parameter p and over the threshold value.

8. Appendix

8.1. Illustration of the Definition 3.1

In order to illustrate the definition 3.1 and the proposition 3.1 let us consider the example of a sequence $|\theta|$ of wavelet coefficients magnitudes given by the tree A in the Figure 8. We apply to this tree the (λ, p)-VBT-method for $p = \{2, \infty\}$ with $\lambda = 0.9$ that yields the trees B and C.

In what follows we give the detailed steps of the iterative algorithm described in the definition 3.1 for the $(\lambda = 0.9, p = 2)$-VBT method:

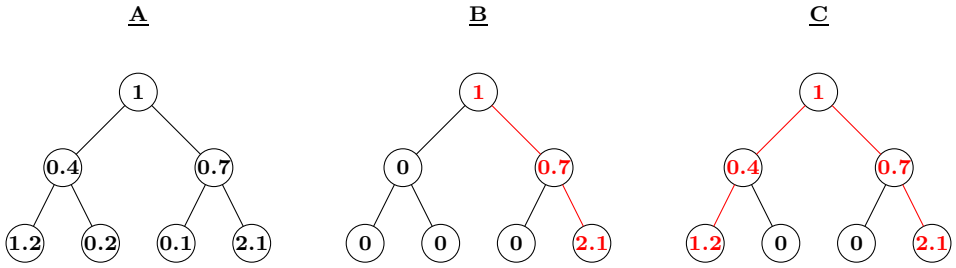

FIG 8. *A: an example of a tree of wavelet coefficient magnitudes* $|\theta_{j,k}|$*; B: the result of* $(\lambda = 0.9, p = 2)$*-VBT method applied to this example; C: the result of* $(\lambda = 0.9, p = \infty)$*-VBT method applied to this example.*

1. step 1: $j = j_\lambda - 1 = 2,\ k = \{0,1,2,3\}$, we easily see that $\mathcal{E}_{2,1}(\theta,\lambda) = \mathcal{E}_{2,2}(\theta,\lambda) = \emptyset$, $\mathcal{E}_{2,0}(\theta,\lambda) = \{(2,0)\}$, $\mathcal{E}_{2,3}(\theta,\lambda) = \{(2,3)\}$.
2. step 2: $j = 1, k = \{0,1\}$, note that:
   $\{\mathcal{E}_{j+1,k'}(\theta,\lambda)\} = \{\mathcal{E}_{2,0}(\theta,\lambda), \mathcal{E}_{2,1}(\theta,\lambda), \mathcal{E}_{2,2}(\theta,\lambda), \mathcal{E}_{2,3}(\theta,\lambda)\}$
   and that $\{\mathcal{E}_{j+1,k'}(\theta,\lambda), (j,k) \in \mathcal{P}(j+1,k')\}$ corresponds to the direct descendants of (j,k), hence:
   (a) for $k = 0$,
   $\mathcal{F}_{1,0}(\theta,\lambda) = \{(1,0), \mathcal{E}_{2,0}(\theta,\lambda), \mathcal{E}_{2,1}(\theta,\lambda)\} = \{(1,0),(2,0)\}$.
   We compute the $\|\theta/\mathcal{F}_{1,0}(\theta,\lambda)\|_2 \approx 0.89 < \lambda$
   Then, we set $\mathcal{E}_{1,0} = \emptyset$.
   (b) For $k = 1$,
   $\mathcal{F}_{1,1}(\theta,\lambda) = \{(1,1),(2,3)\}$, $\|\theta/\mathcal{F}_{1,1}(\theta,\lambda)\|_2 \approx 1.57 > \lambda$. Then, we set $\mathcal{E}_{1,1}(\theta,\lambda) = \{(1,1),(2,3)\}$.
3. step 3: $j = 0,\ k = 0$,
   $\mathcal{F}_{0,0}(\theta,\lambda) = \{(0,0), \mathcal{E}_{1,0}(\theta,\lambda), \mathcal{E}_{1,1}(\theta,\lambda)\} = \{(0,0),(1,1),(2,3)\}$,
   $\|\theta/\mathcal{F}_{0,0}(\theta,\lambda)\|_2 \approx 1.40 > \lambda$.

8.2. *Proof of Proposition 3.1*

The proofs of 1. and 3. are obvious. To prove 2., we first notice that from Definition 3.1 the sets $\mathcal{E}_{jk}(\theta,\lambda)$ and $\mathcal{F}_{jk}(\theta,\lambda)$ depend on p. For notational convenience, we suppress the dependence on this parameter in the paper except for this proof as it is a crucial aspect to consider. Then we need the following Lemma 8.1.

Lemma 8.1. *Let* $1 \le p < q \le \infty$, $0 < \lambda < 1$ *and let consider a sequence of real numbers* $\theta = (\theta_{jk}, 0 \le j < j_\lambda, 0 \le k < 2^j)$. *Then the following embedding property holds for any couple of indices* (j,k)*:*

$$\mathcal{E}_{jk}(\theta,\lambda,p) \subseteq \mathcal{E}_{jk}(\theta,\lambda,q). \tag{12}$$

Proof. We prove property (12) by level-recurrence arguments on $\mathcal{E}_{jk}(\theta,\lambda,\bullet)$. For $j=j_\lambda-1$ and any k, we have that $\mathcal{E}_{jk}(\theta,\lambda,p)=\mathcal{E}_{jk}(\theta,\lambda,q)$. If $j=j_\lambda-2$ then, for any $0\leq k<2^j$,

$$\mathcal{F}_{jk}(\theta,\lambda,p)=\mathcal{F}_{jk}(\theta,\lambda,q).$$

When comparing the norms $\|.\|_p$ and $\|.\|_q$, one gets

$$\begin{aligned}\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,p)\|_p &= \|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\|_p\\ &\leq \|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\|_q.\end{aligned}$$

Hence $\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,p)\|_p>\lambda \Longrightarrow \|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\|_q>\lambda$. It implies that $\mathcal{E}_{jk}(\theta,\lambda,p)\subseteq\mathcal{E}_{jk}(\theta,\lambda,q)$.

Suppose now that property (12) holds at a level $j+1$ such that $0\leq j<j_\lambda-1$ and for any $0\leq k'<2^{j+1}$. Then, for any $0\leq k<2^j$

$$\mathcal{F}_{jk}(\theta,\lambda,p)\subseteq\mathcal{F}_{jk}(\theta,\lambda,q).$$

Since

$$\mathcal{E}_{jk}(\theta,\lambda,\bullet)\in\{\emptyset,\mathcal{F}_{jk}(\theta,\lambda,\bullet)\},$$

property (12) clearly holds if $\mathcal{E}_{jk}(\theta,\lambda,q)=\mathcal{F}_{jk}(\theta,\lambda,q)$. We only have to prove the property for the case $\mathcal{E}_{jk}(\theta,\lambda,q)=\emptyset$, i.e. when $\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\|_q\leq\lambda$. First, we note that $\mathcal{F}_{jk}(\theta,\lambda,q)=\mathcal{F}_{jk}(\theta,\lambda,p)\cup(\mathcal{F}_{jk}(\theta,\lambda,q)\setminus\mathcal{F}_{jk}(\theta,\lambda,p))$ and secondly, we note that, $\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\setminus\mathcal{F}_{jk}(\theta,\lambda,p)\|_q>\lambda$. The latter statement comes from the fact that if the set of indices $\mathcal{F}_{jk}(\theta,\lambda,q)\setminus\mathcal{F}_{jk}(\theta,\lambda,p)$ is pruned by the (λ,q)-VBT-method, that means that its ℓ_q mean is lower than the threshold λ.

Let us set $\mathcal{F}_{jk}(\theta,\lambda,q)=\mathcal{F}_{jk}(\theta,\lambda,p)\cup\left(\mathcal{F}_{jk}(\theta,\lambda,q)\setminus\mathcal{F}_{jk}(\theta,\lambda,p)\right)$. Therefore,

$$\begin{aligned}\lambda^q &\geq \|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\|_q^q\\ &= \frac{\#\mathcal{F}_{jk}(\theta,\lambda,p)}{\#\mathcal{F}_{jk}(\theta,\lambda,q)}\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,p)\|_q^q\\ &\quad+\frac{\#\mathcal{F}_{jk}(\theta,\lambda,q)-\#\mathcal{F}_{jk}(\theta,\lambda,p)}{\#\mathcal{F}_{jk}(\theta,\lambda,q)}\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\setminus\mathcal{F}_{jk}(\theta,\lambda,p)\|_q^q.\end{aligned}$$

So

$$\lambda^q\geq\frac{\#\mathcal{F}_{jk}(\theta,\lambda,p)}{\#\mathcal{F}_{jk}(\theta,\lambda,q)}\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,p)\|_q^q+\frac{\#\mathcal{F}_{jk}(\theta,\lambda,q)-\#\mathcal{F}_{jk}(\theta,\lambda,p)}{\#\mathcal{F}_{jk}(\theta,\lambda,q)}\lambda^q,$$

and $\lambda\geq\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,p)\|_q$. When comparing norms $\|.\|_p$ and $\|.\|_q$ one gets

$$\lambda\geq\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,p)\|_p.$$

So $\mathcal{E}_{jk}(\theta,\lambda,p)=\emptyset$ that is to say $\mathcal{E}_{jk}(\theta,\lambda,p)\subseteq\mathcal{E}_{jk}(\theta,\lambda,q)$.

We conclude that property (12) holds at level j. This ends the proof. □

Corollary 8.1. *Let $1\leq p<q\leq\infty$, $0<\lambda<1$ and let consider a sequence of real numbers $\theta:=\left(\theta_{jk},0\leq j<j_\lambda,0\leq k<2^j\right)$. Then, for any couple of indices (j,k), the following property holds:*

$$\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,q)\|_q\leq\lambda\Longrightarrow\|\theta\ /\ \mathcal{F}_{jk}(\theta,\lambda,p)\|_p\leq\lambda. \tag{13}$$

Proof. Because of the $(\lambda, \bullet)$-VBT-method, property (13) holds if and only if

$$\mathcal{E}_{jk}(\theta, \lambda, q) = \emptyset \Longrightarrow \mathcal{E}_{jk}(\theta, \lambda, p) = \emptyset.$$

This statement is a consequence of Lemma 8.1.

The proof of Proposition 3.1 is then deduced from the corollary above. □

8.3. Proof of Proposition 5.1

Proof. According to (8) of Proposition 5.2, it suffices to state the embedding for the case $p = 2$.

Let $f \in \mathcal{B}^s_{2,\infty}$. There exists $C > 0$ such that, for any $j \in \mathbb{N}$, the wavelet coefficients of f satisfy:

$$\sum_{k=0}^{2^j-1} \theta_{jk}^2 \leq C\ 2^{-2js}.$$

Fix $0 < \lambda < 1$. Let $j_{\lambda,s}$ be the integer such that $2^{-j_{\lambda,s}} \leq \lambda^{\frac{2}{1+2s}} < 2^{1-j_{\lambda,s}}$. Notice that $j_{\lambda,s} \leq j_\lambda$.

$$\begin{aligned}
&\sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \ / \ \mathcal{F}_{j'k'}(\theta,\lambda)\|_2 \leq \lambda \right\} \\
\leq\ &\sum_{j<j_{\lambda,s}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \ / \ \mathcal{F}_{j'k'}(\theta,\lambda)\|_2 \leq \lambda \right\} + \sum_{j\geq j_{\lambda,s}} \sum_{k} \theta_{jk}^2 \\
\leq\ &2^{j_{\lambda,s}}\lambda^2 + C\ 2^{-2sj_{\lambda,s}} \\
\leq\ &(2+C)\ \lambda^{\frac{4s}{1+2s}}.
\end{aligned}$$

Hence

$$\sup_{0<\lambda<1} \lambda^{-\frac{4s}{1+2s}} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \ / \ \mathcal{F}_{j'k'}(\theta,\lambda)\|_2 \leq \lambda \right\} < \infty,$$

that is to say, $f \in W_{\frac{2}{1+2s},2}$. □

8.4. Proof of the maxiset results

In this section, we first provide technical lemmas which shall be used to prove the maxiset result established in Theorem 5.1. Then we prove Proposition 5.2 and Theorem 5.1.

8.4.1. Technical lemmas and their proof

Lemma 8.2. *Let $0 < r < 2$ and let f belong to the space $\mathcal{B}_{2,\infty}^{\frac{2-r}{4}} \cap W_{r,p}$. Then:*

$$\sup_{0<\lambda<1} \lambda^r \left[\log\left(\frac{1}{\lambda}\right)\right]^{-1} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p > \lambda\right\} < \infty.$$

Proof. Let $f \in \mathcal{B}_{2,\infty}^{\frac{2-r}{4}} \cap W_{r,p}$. Then its wavelet coefficients satisfy:

$$\sup_{j\in\mathbb{N}} 2^{\frac{2-r}{2}j} \sum_k \theta_{jk}^2 < \infty,$$

$$\sup_{0<\lambda<1} \lambda^{r-2} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p \le \lambda\right\} \quad < \quad \infty.$$

For any $n \in \mathbb{N}$, we denote by $j_{\lambda,n}$ the smallest integer such that

$$2^{-j_{\lambda,n}} \le (\lambda 2^{1+n})^2.$$

$$\begin{aligned}
&\sup_{0<\lambda<1} \lambda^r \left[\log\left(\frac{1}{\lambda}\right)\right]^{-1} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p > \lambda\right\} \\
&= \sup_{0<\lambda<1} \lambda^r \left[\log\left(\frac{1}{\lambda}\right)\right]^{-1} \\
&\quad \times \sum_{n\in\mathbb{N}} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\lambda 2^n < \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p \le \lambda 2^{1+n}\right\}.
\end{aligned}$$

For any $n \in \mathbb{N}$, the number of wavelet coefficients under interest can be upper bounded by counting $j+1$ ancestors for each leave at level j in the tree (a leave is a coefficient satisfying $\lambda 2^n < \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p \le \lambda 2^{1+n}$ and $|\theta_{jk}| > \lambda 2^n$). So,

$$\begin{aligned}
&\sup_{0<\lambda<1} \lambda^r \left[\log\left(\frac{1}{\lambda}\right)\right]^{-1} \sum_{n\in\mathbb{N}} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \mathbf{1}\left\{\lambda 2^n < \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p \le \lambda 2^{1+n}\right\} \\
&\le \sup_{0<\lambda<1} \lambda^r \left[\log\left(\frac{1}{\lambda}\right)\right]^{-1} \\
&\quad \times \sum_{n\in\mathbb{N}} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} (j+1)\, \mathbf{1}\left\{|\theta_{jk}| > \lambda 2^n;\ \lambda 2^n < \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p \le \lambda 2^{1+n}\right\}.
\end{aligned}$$

For any $n \in \mathbb{N}$, the leaves (j,k) with level $j < j_{\lambda,n}$ are the same as the ones got from the $(\lambda 2^n, p)$-VBT-method satisfying $\min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda 2^n)\|_p \le \lambda 2^{1+n}$. Moreover the number of such leaves is smaller than or equal to the

number of wavelet coefficients θ_{jk} with absolute value strictly larger than $\lambda 2^n$ and such that $\min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda 2^{1+n})\|_p \le \lambda 2^{1+n}$. So

$$\begin{aligned}
&\sup_{0<\lambda<1} \lambda^r \left[\log\left(\frac{1}{\lambda}\right)\right]^{-1} \\
&\quad\times \sum_{n\in\mathbb{N}} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} (j+1)\, \mathbf{1}\Big\{|\theta_{jk}| > \lambda 2^n;\ \lambda 2^n < \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_p \le \lambda 2^{1+n}\Big\} \\
&\le \sup_{0<\lambda<1} \lambda^r \left[\log\left(\frac{1}{\lambda}\right)\right]^{-1} \\
&\quad\times \sum_{n\in\mathbb{N}} \sum_{j<j_{\lambda,n}} \sum_{k=0}^{2^j-1} j_\lambda\, \mathbf{1}\Big\{|\theta_{jk}| > \lambda 2^n; \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda 2^{1+n})\|_p \le \lambda 2^{1+n}\Big\} \\
&\quad+ \sup_{0<\lambda<1} \lambda^{r-2} \sum_{n\in\mathbb{N}} 4^{1-n} \sum_{j\ge j_{\lambda,n}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \\
&= A + B.
\end{aligned}$$

Since $f \in W_{r,p}$,

$$\begin{aligned}
A \;&\le\; \sup_{0<\lambda<1} \lambda^{r-2} \sum_{n\in\mathbb{N}} 4^{1-n} \\
&\qquad\qquad \times \sum_{j<j_{\lambda,n}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\Big\{\min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda 2^{1+n})\|_p \le \lambda 2^{1+n}\Big\} \\
&< \infty.
\end{aligned}$$

Since $f \in \mathcal{B}_{2,\infty}^{\frac{2-r}{4}}$, one has $B < \infty$. This ends the proof. □

Lemma 8.3. *Let $0 < \lambda < 1$, $1 \le p \le \infty$, (j,k) be a couple of indices and θ be a sequence of wavelet coefficients. The two following properties are equivalent:*

i) $\|\theta \,/\, \mathcal{F}_{jk}(\theta,\lambda)\|_p > \lambda$.

ii) There exists a tree $\mathcal{T}$ rooted at (j,k) such that:

$$\begin{aligned}
\frac{1}{\#\mathcal{T}} \sum_{(u,v)\in\mathcal{T}} |\theta_{uv}|^p \;&>\; \lambda^p \quad \text{if } 1 \le p < \infty, \\
\max_{(u,v)\in\mathcal{T}} |\theta_{uv}| \;&>\; \lambda \quad \text{if } p = \infty.
\end{aligned}$$

Proof. We only prove the equivalence property for any $1 \le p < \infty$ since the proof for the case $p = \infty$ is analogous.

$i) \Longrightarrow ii)$

Choose $\mathcal{T} = \mathcal{F}_{jk}(\theta,\lambda)$. From $i)$ one gets

$$\frac{1}{\#\mathcal{T}} \sum_{(u,v)\in\mathcal{T}} |\theta_{uv}|^p = \|\theta \,/\, \mathcal{F}_{jk}(\theta,\lambda)\|_p^p \;>\; \lambda^p.$$

Hence $ii)$ is satisfied.

$ii) \Longrightarrow i)$

Assume that there exists a subtree $\mathcal{T}$ rooted at (j,k) such that

$$\frac{1}{\#\mathcal{T}} \sum_{(u,v)\in\mathcal{T}} |\theta_{uv}|^p \quad > \quad \lambda^p.$$

and consider the tree $\mathcal{F}_{jk}(\theta,\lambda)$ obtained with the (λ,p)-VBT-method. The proof is trivial if $\mathcal{T} = \mathcal{F}_{jk}(\theta,\lambda)$. Otherwise, when looking at the possibly different nodes (j',k') between $\mathcal{T}$ and $\mathcal{F}_{jk}(\theta,\lambda)$, one has:

- if $\mathcal{F}_0 := \mathcal{F}_{jk}(\theta,\lambda) \setminus \mathcal{T} \neq \emptyset$ then

$$\frac{1}{\#\mathcal{F}_0} \sum_{(u,v)\in\mathcal{F}_0} |\theta_{uv}|^p > \lambda^p,$$

  Indeed, for each set of indices $(j',k') \in \mathcal{F}_{jk}(\theta,\lambda)$ the (λ,p)-VBT-method method has verified that $\|\theta/\mathcal{F}_{j'k'}(\theta,\lambda)\|_p > \lambda$.
- if $\mathcal{T}_0 := \mathcal{T} \setminus \mathcal{F}_{jk}(\theta,\lambda) \neq \emptyset$, conversely to the argument of the previous assertion, we have that

$$\frac{1}{\#\mathcal{T}_0} \sum_{(u,v)\in\mathcal{T}_0} |\theta_{uv}|^p \leq \lambda^p.$$

Therefore, since $\mathcal{F}_{jk}(\theta,\lambda) = (\mathcal{F}_0 \cup \mathcal{T}) \setminus \mathcal{T}_0$, we deduce that

$$\frac{1}{\#\mathcal{F}_{jk}(\theta,\lambda)} \sum_{(u,v)\in\mathcal{F}_{jk}(\theta,\lambda)} |\theta_{uv}|^p = \|\theta \ / \ \mathcal{F}_{jk}(\theta,\lambda)\|_p^p > \lambda^p.$$

So $i)$ is satisfied. This ends the proof. □

Lemma 8.4. *Let $0 < \lambda < \frac{1}{2}$ and let $(\theta_{jk}^{(1)}, 0 \leq j < j_\lambda, 0 \leq k < 2^j)$ and $(\theta_{jk}^{(2)}, 0 \leq j < j_\lambda, 0 \leq k < 2^j)$ be two sequences of real numbers. Suppose that the following property holds:*

$$\exists (j',k') \text{ such that } \|\theta^{(2)} \ / \ \mathcal{F}_{j'k'}(\theta^{(2)}, 2\lambda)\|_p > 2\lambda$$
$$\text{and } \|\theta^{(1)} \ / \ \mathcal{F}_{j'k'}(\theta^{(1)}, \lambda)\|_p \leq \lambda.$$

Then there exists $(j",k")$ such that $0 \leq j" < j_\lambda$, $0 \leq k" < 2^j$ and

$$|\theta_{j"k"}^{(2)} - \theta_{j"k"}^{(1)}| > \lambda.$$

Proof. Suppose that for any $(j",k")$ such that $0 \leq j" < j_\lambda$, $0 \leq k" < 2^j$ one has

$$|\theta_{j"k"}^{(2)} - \theta_{j"k"}^{(1)}| \leq \lambda. \tag{14}$$

Then

$$\left(\frac{1}{\#\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}\sum_{(u,v)\in\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}|\theta^{(2)}_{uv}|^p\right)^{1/p} - \left(\frac{1}{\#\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}\sum_{(u,v)\in\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}|\theta^{(1)}_{uv}|^p\right)^{1/p}$$
$$\leq \left(\frac{1}{\#\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}\sum_{(u,v)\in\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}\left|\theta^{(1)}_{uv}-\theta^{(2)}_{uv}\right|^p\right)^{1/p}$$
$$\leq \lambda.$$

So, due to the assumption on the sequence $(\theta^{(2)}_{jk}, 0\leq j<j_\lambda, 0\leq k<2^j)$,

$$\frac{1}{\#\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}\sum_{(u,v)\in\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)}|\theta^{(1)}_{uv}|^p>\lambda^p.$$

Since $\mathcal{T}:=\mathcal{F}_{j'k'}(\theta^{(2)},2\lambda)$ is a tree rooted at (j',k'), when using Lemma 8.3 one gets

$$\|\theta^{(1)}\ /\ \mathcal{F}_{j'k'}(\theta^{(1)},\lambda)\|_p^p=\frac{1}{\#\mathcal{F}_{j'k'}(\theta^{(1)},\lambda)}\sum_{(u,v)\in\mathcal{F}_{j'k'}(\theta^{(1)},\lambda)}|\theta^{(1)}_{uv}|^p>\lambda^p.$$

Thus, this ends the proof by contradiction with the assumption on the sequence $(\theta^{(1)}_{jk}, 0\leq j<j_\lambda, 0\leq k<2^j)$. □

8.4.2. Proof of Proposition 5.2

Let $0<s<V$ and $0<u<\frac{s}{1+2s}$.

(8) The embedding property is a direct consequence of 2. of Proposition 3.1.

(9) The large inclusions are due to (8). To prove the strict embedding we construct a function which belongs to $\mathcal{B}^u_{2,\infty}\cap W_{\frac{2}{1+2s},\infty}$ but not to $W_{\frac{2}{1+2s},2}$. The main idea to construct such a function is to ensure that $\hat{f}_\infty$ uses all the coefficients up to the finest scale and that $\hat{f}_2$ thresholds the finest scale. To do so, we put non zero coefficients at each odd scales j and, within each scales, only one non zero coefficient over two. Hence, the ℓ_2 norm of the first block of coefficients, i.e., $\mathcal{F}_{j_\lambda-2,k}$ is lower than the threshold. In other words $\hat{f}_2$ sets a whole scale of coefficients to zero whereas $\hat{f}_\infty$ keeps them. Now formally, let us consider the function h with wavelet coefficients $(\theta_{jk})_{jk}$ satisfying:

$$\begin{aligned}\theta_{jk} &= 2^{-\frac{j}{2}} \text{ if } j \text{ and } k \text{ are odd and } 0\leq\ k<2(j+1)\ 2^{\frac{j}{1+2s}},\\ \theta_{jk} &= 0 \text{ otherwise.}\end{aligned}$$

This function h belongs to the space $\mathcal{B}^u_{2,\infty} \cap W_{\frac{2}{1+2s},\infty}$ but does not belong to the space $W_{\frac{2}{1+2s},2}$, which we show now.

Proof. Put $r = 2(1+2s)^{-1}$. For any level j large enough

$$\sum_{k=0}^{2^j-1} \theta_{jk}^2 \leq \left[(j+1)2^{\frac{j}{1+2s}} + 1\right] 2^{-j} = (j+2)2^{-\frac{2js}{1+2s}} \leq 2^{-2ju}.$$

Hence $h \in \mathcal{B}^u_{2,\infty}$. Moreover $h \in W_{\frac{2}{1+2s},\infty}$ since it is clear that

$$\sup_{0<\lambda<1} \lambda^{r-2} \sum_{j<j_\lambda} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_\infty \leq \lambda \right\} = 0.$$

But

$$\begin{aligned} & \sup_{0<\lambda<1} \lambda^{r-2} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta,\lambda)\|_2 \leq \lambda \right\} \\ \geq & \sup_{0<\lambda<1} \left[j_\lambda \, 2^{\frac{j_\lambda - 1}{1+2s}} - 1 \right] \, \lambda^{r-2} \, 2^{-j_\lambda+1} \\ > & \sup_{0<\lambda<1} j_\lambda - 1 \\ = & +\infty. \end{aligned}$$

Hence $h \notin W_{\frac{2}{1+2s},2}$. □

8.4.3. Proof of Theorem 5.1

For this proof, we use the following concentration inequality: $P\left[|Z| > \lambda\right] \leq 2\exp\left(\frac{-\lambda^2}{2}\right)$ where Z denotes a standard Gaussian random variable. Here and later, we shall denote by C a constant which may be different from line to line.

Proof. Notice that the result can be proven by replacing the supremum over ϵ in $[0,1[$ by the supremum over ϵ in $]0,\epsilon_m[$, where ϵ_m is such that $0 < \epsilon_m < \frac{1}{2}$ and such that $0 < \lambda_{\epsilon_m} = m\epsilon_m\sqrt{\log(\epsilon_m^{-1})} < \frac{1}{2}$.

$\Longrightarrow$

Suppose that, for any $0 < \epsilon < \epsilon_m$, $\mathbb{E}\|\hat{f}_p - f\|_2^2 \leq C\ \lambda_\epsilon^{\frac{4s}{1+2s}}$. Then,

$$\begin{aligned} \sum_{j\geq j_{\lambda_\epsilon}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 & \leq \mathbb{E}\|\hat{f}_p - f\|_2^2 \\ & \leq C\, \lambda_\epsilon^{\frac{4s}{1+2s}} \\ & \leq C\, 2^{-\frac{2s}{1+2s} j_{\lambda_\epsilon}}. \end{aligned}$$

So, using argument of continuity, $f \in \mathcal{B}_{2,\infty}^{\frac{s}{1+2s}}$. Moreover

$$\left(\frac{\lambda_\epsilon}{2}\right)^{-\frac{4s}{1+2s}} \sum_{j<j_{\lambda_\epsilon}+2} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\theta \,/\, \mathcal{F}_{j'k'}(\theta, \frac{\lambda_\epsilon}{2})\|_p \le \frac{\lambda_\epsilon}{2} \right\}$$
$$\le A_1 + A_2 + A_3,$$

with

$$A_1 = \left(\frac{\lambda_\epsilon}{2}\right)^{-\frac{4s}{1+2s}} \sum_{j<j_{\lambda_\epsilon}} \sum_{k=0}^{2^j-1} \mathbb{E}\left[\theta_{jk}^2 \mathbf{1}\left\{ \min_{(j',k')\in\mathcal{P}(j,k)} \|\hat{\theta} \,/\, \mathcal{F}_{j'k'}(\hat{\theta}, \lambda_\epsilon)\|_p \le \lambda_\epsilon \right\}\right]$$
$$\le \left(\frac{\lambda_\epsilon}{2}\right)^{-\frac{4s}{1+2s}} \mathbb{E}\|\hat{f}_p - f\|_2^2$$
$$\le C,$$

$$A_2 = \left(\frac{\lambda_\epsilon}{2}\right)^{-\frac{4s}{1+2s}}$$
$$\times \mathbb{E}\left[\sum_{j<j_{\lambda_\epsilon}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbf{1}\left\{ \exists (j',k') \in \mathcal{P}(j,k) : \|\theta / \mathcal{F}_{j'k'}(\theta, \frac{\lambda_\epsilon}{2})\|_p \le \frac{\lambda_\epsilon}{2}, \|\hat{\theta} \,/\, \mathcal{F}_{j'k'}(\hat{\theta}, \lambda_\epsilon)\|_p > \lambda_\epsilon \right\}\right]$$
$$\le C\, 2^{j_{\lambda_\epsilon}}\, \lambda_\epsilon^{-\frac{4s}{1+2s}} \sum_{j<j_{\lambda_\epsilon}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \mathbb{P}\left(|\hat{\theta}_{jk} - \theta_{jk}| > \frac{\lambda_\epsilon}{2}\right)$$
$$\le C\, 2^{j_{\lambda_\epsilon}}\, \lambda_\epsilon^{-\frac{4s}{1+2s}} \epsilon^{\frac{m^2}{8}}$$
$$\le C\, \lambda_\epsilon^{\frac{m^2}{8}-4}$$
$$\le C.$$

The last inequalities require Lemma 8.4 and $m \ge 4\sqrt{2}$ to hold.

Now

$$A_3 = \left(\frac{\lambda_\epsilon}{2}\right)^{-\frac{4s}{1+2s}} \left[\sum_{k=0}^{2^j-1} \theta_{j_{\lambda_\epsilon}k}^2 + \sum_{k=0}^{2^j-1} \theta_{j_{\lambda_\epsilon}+1k}^2\right]$$
$$\le C \left(\frac{\lambda_\epsilon}{2}\right)^{-\frac{4s}{1+2s}} 2^{-\frac{2s}{1+2s} j_{\lambda_\epsilon}}$$
$$\le C.$$

The last inequality holds since we have already proved that $f \in \mathcal{B}_{2,\infty}^{\frac{s}{1+2s}}$. When combining the bounds of A_1, A_2 and A_3 and using argument of continuity, one deduces that $f \in W_{\frac{2}{1+2s},p}$.

$\Longleftarrow$

Suppose that $f \in \mathcal{B}_{2,\infty}^{\frac{s}{1+2s}} \cap W_{\frac{2}{1+2s},p}$. For any $0 < \epsilon < \epsilon_m$, the quadratic risk of the estimator $\hat{f}_p$ can be decomposed as follows:

$$
\begin{aligned}
& \mathbb{E}\|\hat{f}_p - f\|_2^2 \\
= \;& \mathbb{E}\left[\sum_{j<j_{\lambda_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)}\|\hat{\theta}\;/\;\mathcal{F}_{j'k'}(\hat{\theta},\lambda_\epsilon)\|_p \le \lambda_\epsilon\right\}\right] \\
& + \sum_{j<j_{\lambda_\epsilon}}\sum_{k=0}^{2^j-1}\mathbb{E}\left[(\hat{\theta}_{jk}-\theta_{jk})^2\mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)}\|\hat{\theta}\;/\;\mathcal{F}_{j'k'}(\hat{\theta},\lambda_\epsilon)\|_p > \lambda_\epsilon\right\}\right] \\
& + \sum_{j\ge j_{\lambda_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2 + \epsilon^2 \\
= \;& B_1 + B_2 + B_3.
\end{aligned}
$$

Since $f \in \mathcal{B}_{2,\infty}^{\frac{s}{1+2s}} \cap W_{\frac{2}{1+2s},p}$ and due to Lemma 8.4

$$
\begin{aligned}
B_1 &= \mathbb{E}\left[\sum_{j<j_{\lambda_\epsilon}}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)}\|\hat{\theta}\;/\;\mathcal{F}_{j'k'}(\hat{\theta},\lambda_\epsilon)\|_p \le \lambda_\epsilon\right\}\right] \\
&\le \sum_{j<j_{\lambda_\epsilon}-2}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)}\|\theta\;/\;\mathcal{F}_{j'k'}(\theta,2\lambda_\epsilon)\|_p \le 2\lambda_\epsilon\right\} \\
&\quad + C\;2^{-\frac{2s}{1+2s}j_{\lambda_\epsilon}} + 2^{j_{\lambda_\epsilon}}\sum_{j<j_{\lambda_\epsilon}-2}\sum_{k=0}^{2^j-1}\theta_{jk}^2\mathbb{P}\left(|\hat{\theta}_{jk}-\theta_{jk}| > \lambda_\epsilon\right) \\
&\le C\left(\lambda_\epsilon^{\frac{4s}{1+2s}} + 2^{j_{\lambda_\epsilon}}\;\epsilon^{\frac{m^2}{2}}\right) \\
&\le C\,\lambda_\epsilon^{\frac{4s}{1+2s}},
\end{aligned}
$$

as soon as $m \ge 2\sqrt{2}$.

By using Lemmas 8.2 and 8.4, and the Cauchy-Schwarz inequality

$$
\begin{aligned}
B_2 &= \sum_{j<j_{\lambda_\epsilon}}\sum_{k=0}^{2^j-1}\mathbb{E}\left[(\hat{\theta}_{jk}-\theta_{jk})^2\mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)}\|\hat{\theta}\;/\;\mathcal{F}_{j'k'}(\hat{\theta},\lambda_\epsilon)\|_p > \lambda_\epsilon\right\}\right] \\
&\le \epsilon^2\sum_{j<j_{\lambda_\epsilon}+2}\sum_{k=0}^{2^j-1}\mathbf{1}\left\{\min_{(j',k')\in\mathcal{P}(j,k)}\|\theta\;/\;\mathcal{F}_{j'k'}(\theta,\frac{\lambda_\epsilon}{2})\|_p > \frac{\lambda_\epsilon}{2}\right\} \\
&\quad + C\;2^{\frac{j_{\lambda_\epsilon}}{2}}\epsilon^2\sum_{j<j_{\lambda_\epsilon}}\sum_{k=0}^{2^j-1}\mathbb{P}^{1/2}\left(|\hat{\theta}_{jk}-\theta_{jk}| > \frac{\lambda_\epsilon}{2}\right) \\
&\le C\left(\lambda_\epsilon^{\frac{4s}{1+2s}} + 2^{\frac{j_{\lambda_\epsilon}}{2}}\;\epsilon^{\frac{m^2}{16}}\right) \\
&\le C\left(\lambda_\epsilon^{\frac{4s}{1+2s}} + \lambda_\epsilon^{\frac{m^2}{16}-1}\right) \\
&\le C\,\lambda_\epsilon^{\frac{4s}{1+2s}},
\end{aligned}
$$

as soon as $m \geq 4\sqrt{3}$. Since $f \in \mathcal{B}_{2,\infty}^{\frac{s}{1+2s}}$,

$$\begin{aligned} B_3 &= \epsilon^2 + \sum_{j \geq j_{\lambda_\epsilon}} \sum_{k=0}^{2^j-1} \theta_{jk}^2 \\ &\leq \epsilon^2 + C\, 2^{-\frac{2s}{1+2s} j_{\lambda_\epsilon}} \\ &\leq C\, \lambda_\epsilon^{\frac{4s}{1+2s}}. \end{aligned}$$

When combining the bounds of B_1, B_2 and B_3 one deduces that

$$\sup_{0<\epsilon<1} \lambda_\epsilon^{-\frac{4s}{1+2s}} \mathbb{E}\|\hat{f}_p - f\|_2^2 < \infty.$$

This ends the proof. □

Acknowledgments

The authors thank the associate editor and two anonymous referees for the helpful comments and suggestions which led to the considerable improvement of our article.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D., Johnstone, I. (2006). Adapting to Unknown Sparsity by Controlling the False Discovery Rate. *Annals of Statistics*, **34**(2), 584-653. MR2281879

[2] Antoniadis, A., Bigot, J., Sapatinas, T. (2001). Wavelet Estimators in Nonparametric Regression: a Comparative Simulation Study. *Journal of Statistical Software*, **6**(6), 1-83.

[3] Autin, F. (2004). Maxiset Point of View in Nonparametric Estimation. *Ph.D.* at university of Paris 7 - France.

[4] Autin, F. (2008). On the Performances of a New Thresholding Procedure using Tree Structure. *Electronic Journal of Statistics*, **2**, 412-431. MR2411441

[5] Autin, F. (2008). Maxisets for μ-thresholding Rules. *Test*, **17**(2), 332-349. MR2434331

[6] Autin, F., Picard, D., and Rivoirard, V. (2006). Large variance gaussian priors in Bayesian nonparametric estimation: a maxiset approach. *Mathematical Methods of Statistics*, **15**(4), 349-373. MR2301657

[7] Averkamp, R., Houdré (2005). Wavelet Thresholding for Non Necessarily Gaussian Noise: Functionality. *Annals of Statistics*, **33**(5), 2164-2193. MR2211083

[8] Baraniuk, R. (1999). Optimal Tree Approximation Using Wavelets. *Proceedings of SPIE Conference on Wavelet Applications in Signal and Image Processing VII*, Eds A. J. Aldroubi and M. Unser, Bellingham, WA:SPIE, 196-207.

[9] Cai, T. (1999). Adaptive Wavelet Estimation: a Block Thresholding and Oracle Inequality Approach. *Annals of Statistics*, **27**(3), 898-924. MR1724035

[10] Cohen, A., Dahmen W., Daubechies I., and DeVore, R. (2001). Tree Approximation and Optimal Encoding. *Applied and Computational Harmonic Analysis*, **11**(2), 192-226. MR1848303

[11] Cohen, A., De Vore, R., Kerkyacharian, G., and Picard, D. (2001). Maximal Spaces with Given Rate of Convergence for Thresholding Algorithms. *Applied and Computational Harmonic Analysis*, **11**, 167-191. MR1848302

[12] Daubechies, I. (1992). *Ten Lectures on Wavelets*. SIAM, Philadelphia. MR1162107

[13] Donoho, D.L., and Johnstone, I.M. (1994). Ideal Spatial Adaptation by Wavelet Shrinkage. *Biometrika*, **81**(3), 425-455. MR1311089

[14] Donoho, D.L. (1997). CART and Best-ortho-basis. *Annals of Statistics*, **25**(5), 1870-1911. MR1474073

[15] Engel, J. (1994). A simple Wavelet Approach to Nonparametric Regression from Recursive Partitioning Schemes. *Journal of Multivariate Analysis*, **49**(2), 242-254. MR1276437

[16] Engel, J. (1999). Tree Structured Estimation with Haar Wavelets. Verlag, 159 pp.

[17] Freyermuth, J.-M., Ombao, H., and von Sachs R. (2010). Tree-Structured Wavelet Estimation in a Mixed Effects Model for Spectra of Replicated Time Series. *Journal of the American Statistical Association*, **105**(490), 634-646. MR2724848

[18] Härdle, W. and Kerkycharian, G. and Picard D. and Tsybakov, A. (1998). Wavelets, approximation, and statistical applications. Springer Verlag, Lectures Notes in Statistics, vol. 129. MR1618204

[19] Jansen, M. (2001). Noise Reduction by Wavelet Thresholding. Springer Verlag, Lecture Notes in Statistics, vol. 161, 224 pp. MR1848545

[20] Lee, T. (2002). Tree based wavelet regression for correlated data using the minimum description length principle. *Australian and New Zealand Journal of Statistics*, **44**(1), 23-39. MR1894978

[21] Kerkyacharian, G., and Picard, D. (2000). Thresholding Algorithms, Maxisets and Well Concentrated Bases. *Test*, **9**(2), 283-344. MR1821645

[22] Kerkyacharian, G., and Picard, D. (2002). Minimax or maxisets? *Bernoulli*, **8**(2), 219-253. MR1895892

[23] Shapiro, J. (1993). Embedded image coding using zero trees of wavelet coefficients. *IEEE Transactions on Signal Processing*, **41**(12), 3445-3462.

[24] Sun, J., Gu, D., Chen, Y., and Zhang, S. (2004). A multiscale edge detection algorithm based on wavelet domain vector hidden markov tree model. *Pattern Recognition*, **37**, 1315-1324.

[25] Tsybakov, A. (2008). Introduction to Nonparametric Estimation. Springer Series in Statistics, 214 pp. MR2724359

[26] Vidakovic, B. (1999). Statistical Modelling by Wavelets, John Wiley & Sons, Inc., New York, 384 pp. MR1681904