Maximal Spaces with given rate of convergence for thresholding algorithms *

Albert Cohen, Ronald DeVore, Gerard Kerkyacharian, Dominique Picard

June 16, 1999

1 Introduction

In recent years, various nonlinear methods have been proposed and deeply investigated in the context of nonparametric estimation: shrinkage methods [21], locally adaptive bandwidth selection [16] and wavelet thresholding [7].

One way of comparing the performances of two different method is to fix a class of functions to be estimated and to measure the estimation rate achieved by each method over this class. In this context, most of these methods have been proved to achieve minimax rate for a given loss function, over various classes modelled by the unit balls of function spaces: Hölder, Sobolev and more generally Besov and Triebel-Lizorkin spaces.

It should be noted that the choice of such a class is quite subjective. Moreover it happens very often that the minimax properties can be extended (without deteriorating the rate of convergence) to larger spaces (see e.g. [11]).

It is thus natural to address the following question: given an estimation method and a prescribed estimation rate for a given loss function, what is the *maximal space*, over which this rate is achieved ? If it exists, such a space will appear as naturally linked with the method under consideration. The goal of this paper is to discuss the existence and the nature of maximal spaces in the context of nonlinear methods based on thresholding (or shrinkage) procedures.

Before going further, some remarks should be made:

- The maximal space will be of particular interest if in addition the rate provided by the method is minimax in this space. Note that if the method is proved to achieve a minimax rate $n^{-\alpha}$ on some space E, then the same necessarily holds for the maximal space F associated to this rate since it contains E.

- Such an approach has been investigated for linear methods. We have in mind the classical Parzen kernel estimates or linear spline and wavelet methods [14]. The results take the following form. For an L_p-loss function ($p \geq 2$) the maximal set where a rate $n^{-\alpha p}$ is attained is a ball of the Besov space $B^s_\infty(L_p)$ where s is defined

*This work has been supported in part by the Office of Naval Research Contract N0014-91-J1076 and the Army Research Office Contract DAAG55-98-1-0002

by the equation $\alpha = s/(1+2s)$. An especially interesting remark is that this result is rather robust with respect to the method of estimation provided that the method is linear. This suggests the following questions: (i) are the maximal spaces associated to non linear methods also very robust and not much varying from one method to the other? (ii) are they bigger than those associated to linear methods?

- While it is not difficult to see that the set of functions corresponding to a certain rate of estimation for a linear method is indeed a linear space, this point is not clear (and probably not true in general) for a nonlinear method: f and g could be estimated at rate $n^{-\alpha}$ while $f+g$ is estimated at the rate $n^{-\beta}$ with $\beta < \alpha$. In the context of the methods which are considered in the present paper, the maximal space will be proved to be linear.

As we have already pointed out, we are interested in estimation methods based on thresholding procedures. Typically, the function f to be estimated is assumed to have an expansion in some basis, i.e. $f = \sum_{k \geq 0} c_k e_k$, and the method consists in three steps:

- A linear step corresponding to the estimation of the c_k's by some estimators $\hat{c}_k$.
- A nonlinear step consisting in a thresholding procedure $\hat{c}_k \mapsto \hat{c}_k I\{|\hat{c}_k| \geq t_k\}$.
- A reconstruction step to derive the global estimator $\hat{f} = \sum_{k \geq 0} \hat{c}_k I\{|\hat{c}_k| \geq t_k\} e_k$.

Although the basis $(e_k)_{k \geq 0}$ can be of any type, a natural setting to derive these results is provided by *wavelet bases*. Indeed, such bases provide characterizations of various classical smoothness classes from the approximation rate of (deterministic) thresholding procedures. While this fact is well-known in approximation theory, we shall see here that similar statements hold for statistical estimation. It turns out in this setting that for the nonlinear estimation methods under consideration, the maximal spaces will coincide with known smoothness classes.

Before proceeding further with the description of our results in a simple case, we shall shall introduce the notation we shall utilize for wavelet bases.

1.1 Wavelets

Wavelet bases have been documented in numerous textbooks and survey papers (see e.g. [4] and [17] for a general treatment). With a little effort, they can be adapted to a bounded interval [1] and to more general domains $\Omega \subset \mathbb{R}^d$ (see [3] for a survey of these adaptations as well as a discussion of the characterizations of function spaces on Ω by wavelet coefficients).

A wavelet basis consists of two types of functions: scaling functions φ_λ and wavelet functions ψ_λ. The index λ concatenates the usual scale and space parameters j,k. Thus for standard wavelet bases on $I\!R$, we simply have $\psi_\lambda = \psi_{j,k} = 2^{j/2}\psi(2^j \cdot - k)$ (and similarily for φ_λ). However, the notation ψ_λ takes into account the possible adaptations of wavelets to multivariate bounded domains in which case the functions ψ_λ and φ_λ usually change form near the boundary.

With this notation, the wavelet decomposition takes the form

$$f = \sum_{\lambda \in \Gamma_{j_0}} \alpha_\lambda \varphi_\lambda + \sum_{j \geq j_0} \sum_{\lambda \in \Delta'_j} \beta_\lambda \psi_\lambda, \tag{1}$$

where $(\varphi_\lambda)_{\lambda \in \Gamma_j}$ is the scaling function basis spanning the approximation at level j, $(\psi_\lambda)_{\lambda \in \Delta'_j}$ is the wavelet basis spanning the details at level j, and $\alpha = \alpha(f)$ and $\beta_\lambda = \beta_\lambda(f)$ are the scaling function and wavelet coefficients of f respectively. In what follows, we shall (merely for notational convenience) always take $j_0 := 0$.

The approximation and detail coefficients of f are linear functionals of f which can be evaluated according to

$$\alpha_\lambda = \langle f, \tilde{\varphi}_\lambda \rangle \;\text{ and }\; \beta_\lambda = \langle f, \tilde{\psi}_\lambda \rangle, \tag{2}$$

where $\tilde{\varphi}_\lambda$ and $\tilde{\psi}_\lambda$ are the corresponding dual scaling functions and wavelets. In the orthonormal case, these are the same as the primal scaling functions and wavelets φ_λ and ψ_λ.

We also use the notation $|\lambda| = j$ if $\lambda \in \Gamma_j$ or $\lambda \in \Delta'_j$. Finally, to simplify notation even more, we define $\Lambda_0 := \Gamma_0 \cup \Delta'_0$ and $\Delta_j := \Delta'_j$, $j > 1$. Then, with $\Delta = \cup_{j \geq 0} \Delta_j$, we have an even simplier notation

$$f = \sum_{\lambda \in \Delta} \langle f, \tilde{\psi}_\lambda \rangle \psi_\lambda = \sum_{j=0}^{\infty} \sum_{\lambda \in \Delta_j} \langle f, \tilde{\psi}_\lambda \rangle \psi_\lambda. \tag{3}$$

It is well known (see e.g. [3]) that wavelet bases provide characterizations of smoothness spaces such as the Hölder spaces C^s, Sobolev spaces $W^s(L_p)$ and Besov spaces $B^s_q(L_p)$ for a range of indices s that depend both on the smoothness properties of ψ and $\tilde{\psi}$. In the scale of Besov spaces (which includes $C^s = B^s_\infty(L_\infty)$ and $W^s(L_p) = B^s_p(L_p)$ if $s \notin \mathbb{N}$ as particular cases), the characterization result has the form

$$\|f\|_{B^s_q(L_p)} \sim \|(2^{s|\lambda|} 2^{d|\lambda|(1/2-1/p)} \|(\beta_\lambda)_{\lambda \in \Delta_j}\|_{\ell_p})_{j \geq 0}\|_{\ell_q}, \tag{4}$$

where d is the space dimension ($\Omega \subset \mathbb{R}^d$).

1.2 A simple example

Let us consider the white noise model on the unit interval

$$dY(t) = f(t)dt + \frac{1}{\sqrt{n}} dW(t), \quad t \in [0, 1]. \tag{5}$$

We want to reconstruct f from observations of Y, which is a polution of f by the addition ofthe white noise W. The local thresholding estimators proposed in [7] have the following general form

$$\hat{f}^n = \sum_{|\lambda| \leq j_0(n)} \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq t(n)\} \psi_\lambda, \tag{6}$$

where $\hat{\beta}_\lambda = \int \tilde{\psi}_\lambda(t) dY(t) dt$ is the linearly estimated coefficient.

We choose here the L_2 norm as the loss function, i.e. we are interested in the mean square error $\mathbb{E}(\|\hat{f}^n - f\|_{L_2}^2)$. This error is majorized by

$$\mathbb{E}(\|\hat{f}^n - f\|_{L_2}^2) \leq 2[B(n) + V(n)], \tag{7}$$

where

$$\begin{aligned} B(n) &= \|f - \textstyle\sum_{|\lambda| \leq j_0(n)} \beta_\lambda I\{|\beta_\lambda| \geq t(n)\}\psi_\lambda\|_{L_2}^2 \\ &\leq C \textstyle\sum_{|\lambda| > j_0(n) \text{ or } |\beta_\lambda| < t(n)} |\beta_\lambda|^2, \end{aligned}$$

denotes the bias term, and

$$\begin{aligned} V(n) &= \mathbb{E}(\| \textstyle\sum_{|\lambda| \leq j_0(n)} (\beta_\lambda I\{|\beta_\lambda| \geq t(n)\} - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq t(n)\})\psi_\lambda\|_{L_2}^2) \\ &\leq C \textstyle\sum_{|\lambda| \leq j_0(n)} \mathbb{E}(|\beta_\lambda I\{|\beta_\lambda| \geq t(n)\} - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq t(n)\}|^2), \end{aligned}$$

the variance term. In both case, the constant C is 1 for an orthonormal wavelet basis, and fixed in the case of biorthogonal wavelets.

The choice of an appropriate threshold $t(n)$ and maximal scale $j_0(n)$ have been widely investigated in the context where the properties of f are modelled by Besov smoothness: one typically assumes that f sits in a class

$$V(r, s) = \{\|f\|_{B_\infty^r(L_2)} \leq C_r\} \cap \{\|f\|_{B_p^s(L_p)} \leq C_s\}, \tag{8}$$

where $0 < r < s$ and p is given by $1/p = 1/2 + s$. The parameters s and r should be viewed as two different measure of smoothness:

- ¿From (4), $\|f\|_{B_p^s(L_p)}$ is equivalent to the discrete ℓ_p norm of the wavelet coefficients of f. This space provides a measure the L_2 error resulting from a thresholding procedure by the estimate

  $$\|f - \sum_{|\beta_\lambda| \geq t} \beta_\lambda \psi_\lambda\|_{L_2}^2 \leq C \sum_{|\beta_\lambda| < t} |\beta_\lambda|^2 \leq C t^{2-p} \|f\|_{B_p^s(L_p)}^p. \tag{9}$$

  Thus, in the context of thresholding, the parameter s is thus a natural measure of the sparsity of a function in a wavelet basis. Note that the space $B_p^s(L_p)$ is embedded in L_2 but not compactly.

- The parameter r measures the smoothness of f in L_2, or equivalently the approximation error commited by truncating the function at some scale. One can indeed easily derive from (4) that

  $$\|f\|_{B_\infty^s(L_2)} \sim \|f\|_{L_2} + \sup_{j \geq 0} 2^{js} \|f - \sum_{|\lambda| \leq j} \beta_\lambda \psi_\lambda\|_{L_2}. \tag{10}$$

  An assumption of minimal smoothness in this classical sense is unavoidable in order to limit the thresholding procedure in (6) to a finite number of coefficients below a maximal scale $j_0(n)$ and to discard all other coefficients at higher scales.

With such assumptions, it is known that the minimax estimation rate

$$\mathbb{E}(\|f - \hat{f}^n\|_{L_2}^2) \leq C[\frac{n}{\log(n)}]^{-\frac{2s}{1+2s}}, \tag{11}$$

can be achieved with the choices $2^{j_0(n)} \sim n^{\frac{s}{r+2rs}}$ and $t(n) = \kappa\sqrt{\log(n)/n}$ for κ large enough (depending on r). Such choices correspond to balancing the upper bounds for the bias and variance terms given above. The constant C in (11) depends on the constants C_r and C_s in the definition of the class $V(r,s)$.

It turns out that the above rate can be extended to a slightly larger class $\widetilde{V}(r,s)$, where the assumption $\|f\|_{B^s_p(L_p)} \leq C_s$ is replaced by the assumption that the following thresholding estimate holds:

$$\|f - \sum_{|\beta_\lambda| \geq t} \beta_\lambda \psi_\lambda\|_{L_2} \leq C_s t^{1-p/2}. \tag{12}$$

This estimate can be shown to be equivalent to requiring that the wavelet coefficients of f belong to the weak space ℓ^w_p, i.e.

$$\#\{\lambda \in \Delta \ ; \ |\beta_\lambda| \geq t\} \leq C t^{-p}, \quad t > 0. \tag{13}$$

Unlike $B^s_p(L_p)$, the corresponding function space $B^{s,w}_p(L_p)$ does not correspond to a classical smoothness class, although it can be viewed as an interpolation space: if $s' > s$, $1/p' = 1/2 + s'$ and $(1-\theta)/2 + \theta/p' = 1/p$, elementary interpolation results on ℓ_p spaces give

$$B^{s,w}_p(L_p) = [L_2, B^{s'}_{p'}(L_{p'})]_{\theta,\infty}. \tag{14}$$

Our general results will reveal as a particular case that $B^r_\infty(L_2) \cap B^{s,w}_p(L_p)$ is precisely the maximal space associated to the above described thresholding estimator: if, for a given function f, the rate (11) is achieved, then one necessarily has

$$\|f\|_{B^{s,w}_p(L_p)} + \|f\|_{B^r_\infty(L_2)} \leq \widetilde{C}, \tag{15}$$

where $\widetilde{C}$ depends on the constant C in (11).

1.3 Contents of the paper

In this paper, we shall consider various thresholding methods:

- local (coefficients are thresholded individually),
- global (the sets of coefficients $(\beta_\lambda)_{\lambda \in \Delta_j}$ are globally thresholded),
- block (subsets of $(\beta_\lambda)_{\lambda \in \Delta_j}$ are globally thresholded).

We shall also consider two different types of loss functions:

- Norms which can be expressed as weighted ℓ_p norm of the corresponding wavelet coefficients, e.g. L_2 and $B^s_p(L_p)$. We refer to such norms as "p-sequential".
- L_p norms (which are not p-sequential for $p \neq 2$).

In the second case, which is technically more difficult, the results are only obtained for the local thresholding, whereas in the first case comparison between the different methods are obtained.

The main conclusions of the present paper are the following: the spaces associated to nonlinear methods vary from one method to another. We do not obtain a precise classification of the different types thresholding, unless we leave aside the distinction between rates of convergence that differ by a logarithmic factor. It is interesting to note that the maximal spaces obtained here correspond to the widespread idea that the thresholding methods behave well on spaces whose functions have a regular behaviour except on a lower dimensional set of singularities. The advantage of these spaces is to carefully measure what is the amount of singular behaviour that can be tolerated in the function to be estimated.

Another interesting observation is that the gain of nonlinear methods over linear methods decreases as the loss function tends the L_∞ norm.

This paper is organized as follows. A general setting for nonlinear estimation is introduced in §2, which allows different thresholding methods (local, global or block) as special examples. In §3, we introduce a class of weak spaces that occur in nonlinear approximation. These weak spaces will provide our description of maximal spaces. In that section, we recall various results of nonlinear approximation theory which are central to our analysis, both for proving the estimation rate for a given weak space and the maximality of a weak space for a given rate. The p-sequential case is addressed in §4. A general theorem is given which shows that the saturation space of a general thresholding procedure is the intersection of a certain weak space with a compact subest of the space induced by the loss. This compact subset can be thought of as a minimal basic regularity below which the method seriously degrades. We discuss cases for which the intersection with this compact set is not necessary. The special case $p = \infty$ is treated separately. We devote §5 to the application of the general theorem to different examples of wavelet thresholding. The case of measuring loss in L_p is adressed in §6, in the context of local thresholding.

2 Models and thresholding estimators

In this section, we shall describe the general setting and the notation we shall employ throughout this paper. We shall also give several examples which fall into our general setting.

Our aim is to estimate a function f belonging to some Banach space V. We are given a sequence of models indexed by $n > 0$ from which we derive a sequence of estimators $\hat{f}^n$, which aims to converge to f in some average sense. Generally speaking, our loss will be of the form $\|\hat{f}^n - f\|^p$ where $\|\cdot\| = \|\cdot\|_V$ is the norm of V and $p > 0$. We are thus interested in the behaviour of $\mathbb{E}(\|\hat{f}^n - f\|^p)$ as n goes to $+\infty$.

For instance, in the white noise model (5), the estimators $\hat{f}^n$ are obtained from the observation modelized as f deteriorated by an additive noise of variance $1/n$. Other closely related model are regression (we observe $Y_i = f(i/n) + \varepsilon_i$ for $i = 1, \cdots, n$ where ε_i are i.i.d. Gaussian variables) and density estimation (we observe n independent realizations X_i of a random variable from which we want to estimate the density $f(x)dx$).

Our results will deal with the application of thresholding estimators to such models. We now describe typical examples of these estimators.

2.1 General thresholding estimators

Throughout this paper, we shall assume that the functions to be estimated have the following *atomic decomposition* :

$$f = \sum_{i \in \mathbb{N}} f_i, \tag{16}$$

where the series converge in V and each function f_i is in a fixed closed subspace V_i of V.

It can happen that the norm $\|.\|$ for V in which we measure the loss has a special behavior with respect to this atomic decompostion, in the sense that there exists a sequence of positive numbers p_i and two constants $C_1, C_2 > 0$ such that for all f

$$C_1 \sum_{i \in \mathbb{N}} p_i \|f_i\|_i^p \le \|f\|^p \le C_2 \sum_{i \in \mathbb{N}} p_i \|f_i\|_i^p, \tag{17}$$

where for each $i \in \mathbb{N}$, $\|.\|_i$ is a norm for V_i. In this case, we say that the norm is p-sequential. A trivial case of p-sequential norm is the discrete ℓ_p norm, in which case the f_i are simply the i-th component of the sequence f and $\|f_i\|_i = |f_i|$.

To the atomic decomposition of f corresponds a thresholding estimator:

$$\hat{f}^n = \sum_{i \in \lambda_n} \hat{f}_i^n + \sum_{i \in \Lambda_n \setminus \lambda_n} \hat{f}_i^n I\{\|\hat{f}_i^n\|_i \ge \kappa c(n)\}, \tag{18}$$

where $\hat{f}_i^n \in V_i$ is an estimator of f_i coming from our observation.

The two subsets $\lambda_n \subset \Lambda_n$ reflect the idea that at the "low level of resolution" $(i \in \lambda_n)$ one does not threshold, while at the "very high level" $(i \notin \Lambda_n)$ one does not estimate. In between these two sets, we operate a thresholding according to the rate $\kappa c(n)$. Our estimation procedure is thus characterized by λ_n, Λ_n and $\kappa c(n)$, which all depend on n. The estimators $\hat{f}_i^n$ are coming from our observation.

The assumptions on the sequence of models will only be made through the behaviour of the estimators $\hat{f}_i^n$. For sake of simplicity the index n will be omitted in $\hat{f}$ as well as in $\hat{f}_i$ when no confusion is possible. We will also investigate later the case where the norm can be represented as previously but with a formula where the sum in i is replaced by a supremum.

2.2 Examples of thresholding estimators

2.2.1 Local wavelet thresholding

Let us consider the white noise model (5) and let $(\psi_\lambda)_{\lambda \in \Delta}$ be a wavelet basis adapted to the interval $[0,1]$ (we use here the notation of §1.1).

In [9], [7] and [8], the following estimator, based on wavelet thresholding, was proposed and studied:

$$\hat{f}^n = \sum_{|\lambda| \le j_0(n)} \hat{\beta}_\lambda \psi_\lambda I\{|\hat{\beta}_\lambda| \ge \kappa t_n\}, \tag{19}$$

where

$$\hat\beta_\lambda = \int \tilde\psi_\lambda(t)dY(t), \ \ 2^{j_0(n)} \simeq n/\log n \ \text{ and } \ t_n = (\frac{\log n}{n})^{1/2}. \tag{20}$$

For models other than (5) - e.g. density estimation, regression spectral density estimation, drift or volatility of a diffusion model - the estimator has the same general form (19), and the modifications occur only in the estimation formula for the coefficients $\hat\beta_\lambda$.

This enters the general framework of §2.1 with the $f_i = f_\lambda$ identified as the components $f_\lambda = \beta_\lambda\psi_\lambda$ of the function f in the wavelet basis, $\|f_i\|_i = |\beta_\lambda|$, $\Lambda_n := \{\lambda \in \Delta \ ; \ |\lambda| \leq j_0(n)\}$, $c(n) = t_n$ and λ_n the empty set.

If we measure the loss in the Besov norm $\|\cdot\|_{B^\sigma_p(L_p)}$ for some $\sigma \geq 0$ and $p < \infty$, then according to (4), we are in the p-sequential case, with the coefficients p_i given by $p_i = p_\lambda = 2^{|\lambda|p(\sigma+(1/2-1/p))}$. The case $B^\sigma_\infty(L_\infty)$ corresponds to replacing the summation in (17) by a supremum with exponent $p = 1$ and $p_i = p_\lambda = 2^{|\lambda|(\sigma+1/2)}$.

2.2.2 Global thresholding

In [15], a global thresholding strategy was proposed for the density estimation model. For simplicity, we describe the estimator in the white noise setting (5). We estimate the function f by

$$\hat f^n = \sum_{j<j_1(n)} \sum_{|\lambda|=j} \hat\beta_\lambda\psi_\lambda + \sum_{j=j_1(n)}^{j_0(n)} I\{\sum_{|\lambda|=j} |\hat\beta_\lambda|^2 \geq \kappa 2^j/n\} \sum_{|\lambda|=j} \hat\beta_\lambda\psi_\lambda, \tag{21}$$

where $2^{j_0(n)} \simeq n$ and $2^{j_1(n)} \simeq n^\varepsilon$ for some fixed $\varepsilon \in]0,1[$.

This enters the general framework of §2.1, with $f_j := \sum_{|\lambda|=j} \beta_\lambda\psi_\lambda$ the component of f at scale level j, $\|f_j\|_j := 2^{-j/2}[\sum_{|\lambda|=j} |\beta_\lambda|^2]^{1/2} \sim 2^{-j/2}\|f_j\|_{L_2}$, $\lambda_n := \{j < j_1(n)\}$, $\Lambda_n := \{j \leq j_0(n)\}$ and $c(n) := n^{-1/2}$. Here, V_j is the space spanned by the wavelet function ψ_λ, $|\lambda| = j$.

If we measure the loss in the Besov norm $\|\cdot\|_{B^\sigma_2(L_2)}$ for some $\sigma \geq 0$, then according to (4), we are in the p-sequential case, with $p = 2$ and with the coefficients p_j given by $p_j = 2^{j(2\sigma+1)}$.

More general global thresholding estimators can be constructed using integral operators which approximate the identity: in this setting the function f is approximated at scale 2^{-j} by E_jf where $E_jf(x) = \int E_j(x,y)\,f(y)\,dy$. Here E_j can either be a convolution, i.e. $E_j(x,y) = 2^jE(2^j(x-y))$, or a projection (or more generally quasi-interpolation), i.e. $E_j(x,y) = \sum_{|\lambda|=j} \varphi_\lambda(x)\tilde\varphi_\lambda(y)$. Introducing the "innovation" kernel, $D_j = E_{j+1} - E_j$, and setting $D_0 := E_1$, we have the formal decomposition

$$f := \sum_{j\geq 0} D_jf. \tag{22}$$

In the white noise model, the D_jf are naturally estimated according to

$$\hat D_j := \hat D^n_j := \int_0^1 D_j(t,s)dY(s). \tag{23}$$

Given such an estimation procedure for the individual D_jf, we can thus derive the corresponding global thresholding estimator by setting $f_j = D_jf$, $\|f_j\|_j \sim 2^{-j/2}\|f_j\|_{L_2}$,

$\lambda_n := \{j < j_1(n)\}$, $\Lambda_n := \{j \leq j_0(n)\}$ and $c(n) := n^{-1/2}$. Here V_j is the range of the operator D_j.

If we measure the loss in the Besov norm $\|\cdot\|_{B_2^\sigma(L_2)}$ for some $\sigma \geq 0$, then it is well known that under suitable assumptions on the kernel E_j we have the norm equivalence

$$\|f\|^2_{B_2^\sigma(L_2)} \sim \sum_{j\geq 0} 2^{2j\sigma}\|D_j f\|^2_{L_2}. \tag{24}$$

Therefore, we are again in the p-sequential case, with $p = 2$ and with the coefficients p_i given by $p_j = 2^{j(2\sigma+1)}$.

2.2.3 Block thresholding using a kernel

In [11], block thresholding has been studied in the density estimation model, based on a decomposition of the type (22).

At each level j, one considers the partition of $\mathbb{R}$ into intervals $I_{jk} := [k2^{-j}l_j, (k+1)2^{-j}l_j]$, $k \in \mathbb{Z}$, where $l_j \geq 1$ is a sequence of positive numbers such that $l_j \to +\infty$ and $2^{-j}l_j \to 0$ as j goes to $+\infty$. We then define the block $B_{j,k} = \chi_{I_{j,k}} D_j f$ and its estimator $\hat{B}_{j,k} = \chi_{I_{j,k}} \hat{D}_j f$. The choice $l_j := j^2$, which is studied in [11], corresponds to a "logarithmic growth" of the relative size of the blocks with respect to the resolution.

The corresponding block-thresholded estimator has the form

$$\hat{f}^n = \sum_{j<j_1(n)} \hat{D}_j + \sum_{j_1(n)\leq j\leq j_0(n)} \sum_k \hat{B}_{j,k} I\{l_j^{-1}\int_{I_{j,k}} |\hat{B}_{j,k}|^2 > \kappa n^{-1}\}, \tag{25}$$

where $2^{j_0(n)} \simeq n$, $2^{j_1(n)} \simeq n^\epsilon$ for some fixed $\varepsilon \in]0,1[$.

This enters the general framework of §2.1. with f_i identified to the blocks $B_{j,k}$, $\|f_i\|_i = l_j^{-1/2}\|B_{j,k}\|_{L^2}$, $\lambda_n := \{(j,k)\ ;\ j < j_1(n)\}$, $\Lambda_n := \{(j,k)\ ;\ j \leq j_0(n)\}$ and $c(n) := n^{-1/2}$.

If we measure the loss in the $L_2 = B^0_{2,2}$ norm, then from (24) with $\sigma = 0$, we are in the p-sequential case, with $p = 2$ and with the coefficients $p_i = p_{j,k}$ given by $p_{j,k} = l_j$. For $\sigma > 0$, the expression of the loss in $B^\sigma_{2,2}$ as in (17) with $p_{j,k} = l_j 2^{2j\sigma}$ requires smoother cut-off than $\chi_{I_{j,k}}$ in the definition of the blocks $B_{j,k}$.

3 $V_q^*(\mu)$-spaces and approximation results

We shall introduce certain spaces related to V which occur naturally in analyzing the approximation performance of thresholding. The classical setting is when we use wavelet decompositions. Then, the spaces we introduce are related to Besov spaces.

Let V be a space in which we measure error. Let μ be any positive measure defined on $\mathbb{N}$ (in the case of wavelet bases, μ is defined on Δ the set of wavelet indices). For any $0 < q < \infty$, we define the space $V_q^*(\mu)$ to be the collection of all functions in V such that for all $\epsilon > 0$,

$$\text{(I)} \quad \mu(\{i\ :\ \|f_i\|_i > \epsilon\}) \leq A\epsilon^{-q},$$

for some constant A. The smallest constant $A(f)$ for which (I) is valid is the "norm "on $V_q^*(\mu)$. The condition (I) is the same as saying that the sequence $(\|f_i\|_i)_{i\in\mathbb{N}}$ is in the sequence space weak $\ell_q(\mu)$. This is a slightly weaker condition than requiring that this sequence is in $\ell_q(\mu)$.

In §4, we shall only deal with the case when the norm on V is p-sequential (17) with $0 < p < \infty$. We will utilize the measure

$$\mu^*\{i\} = p_i,$$

where p_i be the numbers appearing in (17). In this case, we shall simply write $V_q^* := V_q^*(\mu^*)$. In §6, when we treat the non sequential case $V = L_p$, we shall utilize other measures μ.

In the case of μ^*, it is easy to show that an equivalent statement to (I) is that for some $r > q$, we have

$$\text{(II)} \qquad \sum_{\|f_i\|_i \le \epsilon} p_i \|f_i\|_i^r = \int \|f_i\|_i^r I\{\|f_i\|_i \le \epsilon\} d\mu \le B\epsilon^{r-q},$$

for all $\epsilon > 0$ with B a constant. Moreover, the smallest constant B for which (II) is valid is equivalent to $\|f\|_{V_q^*}$. To see this, let

$$\Lambda_j(f,\epsilon) := \{i : \; 2^{-j-1}\epsilon < \|f_i\|_i \le 2^{-j}\epsilon\}.$$

If (I) holds then, for all $r > q$, we have

$$\begin{aligned} \textstyle\sum_{\|f_i\|_i \le \epsilon} p_i \|f_i\|_i^r &= \textstyle\sum_{j \ge 0} \sum_{i \in \Lambda_j(f,\epsilon)} p_i \|f_i\|_i^r \\ &\le \textstyle\sum_{j \ge 0} [2^{-j}\epsilon]^r \sum_{i \in \Lambda_j(f,\epsilon)} p_i \\ &\le A \textstyle\sum_{j \ge 0} [2^{-j}\epsilon]^{r-q} \le CA\epsilon^{r-q}. \end{aligned}$$

On the other hand, if (II) holds for some $r > q$, then we have

$$\begin{aligned} \textstyle\sum_{\|f_i\|_i > \epsilon} p_i &\le \textstyle\sum_{j<0} \sum_{i \in \Lambda_j(f,\epsilon)} p_i \\ &\le 2^r \textstyle\sum_{j<0} \sum_{i \in \Lambda_j(f,\epsilon)} p_i \|f_i\|_i^r [2^{-j}\epsilon]^{-r} \\ &\le 2^r B \textstyle\sum_{j<0} [2^{-j}\epsilon]^{r-q} [2^{-j}\epsilon]^{-r} \\ &\le CB\epsilon^{-q}. \end{aligned}$$

Consider now the setting of the wavelet decompositions. For each $s \ge 0$, $0 < p < \infty$, the space $V = B_p^s(L_q)$ is a p-sequential space. Here $f_\lambda := \langle f, \tilde{\psi}_\lambda \rangle \psi_\lambda$, V_λ is the one dimensional space spanned by ψ_λ and its norm is $|\langle f, \tilde{\psi}_\lambda \rangle|$. The norm equivalence (17) holds with $p_\lambda = 2^{\beta p |\lambda|}$, $\beta = s + 1/2 - 1/p$. Thus, $f \in B_p^s(L_p)$ is equivalent to saying that the sequence $(2^{\beta|\lambda|}\|f_\lambda\|_{V_\lambda})_{\lambda \in \Delta}$ is in ℓ_p.

The above holds in particular when $V = B_p^0(L_p)$, in which case $p_\lambda = 2^{|\lambda|(p/2-1)}$. In this case, for $q < p$, property (II) says that $f \in V_q^*$ if and only if

$$\|f - \sum_{|\beta_\lambda| > \varepsilon} \beta_\lambda \psi_\lambda\|_{B_p^0(L_p)} \le C\epsilon^{p-q}.$$

That is, V_q^* is characterized by the approximation performance of wavelet thresholding in $B_p^0(L_p)$. If $q < p$, the space V_q^* is very close to a Besov space. Recall that by (I), $f \in V_q^*$ if and only if $(\beta_\lambda(f))_{\lambda \in \Delta}$ is in weak $\ell_q(\mu)$. It is easy to see that for $q < p$, the space of functions f with $(\beta_\lambda(f))_{\lambda\in\Delta}$ in $\ell_q(\mu)$ is the Besov space $B^s_q(L_q)$ with s and q related by $s = (p/q-1)/2$. Thus, V^*_q can be viewed as a weak Besov space ($V^*_q = weak B^s_q(L_q)$. For further discussion of thresholding and the spaces V^*_q, we refer the reader to [2]

Surprisingly, when $1 < p < \infty$, similar results can be obtained with $B^0_p(L_p)$ replaced by the space L_p, which is not p-sequential. We close this section by mentioning some results on thresholding in L_p which will be useful in §6. These results are straightforward with $B^0_p(L_p)$ in place of L_p. For $p \leq 1$, similar results hold with the Hardy H_p space in place of L_p.

The next result was given in Lemma 5.1 of [2] and generalizes a result of Temlyakov [24].

Lemma 1 *Let $1 < p < \infty$ and let $\mu(\lambda) := 2^{(p/2-1)|\lambda|}$, $\lambda \in \Delta$. Then, there exists constants $C_1(p)$, and $C_2(p)$ such that, for every $E \subset \Delta$ we have*

$$C_1 \inf_{\lambda\in E} |c_\lambda|^p \mu(E) \leq \| \sum_{\lambda\in E} c_\lambda \psi_\lambda \|^p_{L_p} \leq C_2 \sup_{\lambda\in E} |c_\lambda|^p \mu(E).$$

.

Lemma 2 *Let $2 < p < \infty$ and let $\mu(\lambda) := 2^{(p/2-1)|\lambda|}$, $\lambda \in \Delta$. If $0 < q < p$, then $V^*_q(\mu) \subset L_p$.*

For a proof see [2].

Lemma 3 *Let $1 < p < \infty$, $\mu(\lambda) := 2^{(p/2-1)|\lambda|}$, $\lambda \in \Delta$, and $0 < q < p$. The following properties are equivalent:*

(i) $f \in V^*_q(\mu)$.

(ii) for all $\epsilon > 0$, $\|f - \sum_{|\beta_\lambda|>\varepsilon} \beta_\lambda\psi_\lambda\|^p_{L_p} \leq B\epsilon^{p-q}$,

for some constant $B = B(f) > 0$. Moreover, the smallest constant satisfying (ii) is equivalent to the norm of f in $V^*_q(\mu)$.

For a proof, see [2].

We finally recall the characterization of L_p by the square function (see e.g. [17])

Lemma 4 *For $1 < p < \infty$, there exists constants $C_1(p)$, and $C_2(p)$ such that we have*

$$C_1(p)\|f\|^p_{L^p} \leq \|(\sum_{\lambda\in\Delta} |\beta_\lambda(f)\psi_\lambda|^2)^{1/2}\|_{L_p} \leq C_2(p)\|f\|_{L^p}.$$

4 Results in the case of p-sequential norms

4.1 Result in the general setting

In this section, we shall prove a theorem (Theorem 1) which analyzes the performance of general thresholding. The ingredients of this theorem are to assume certain properties of the thresholding procedure (these are give in assumptions (a), (b), (c) of Theorem 1) and then to characterize the functions which are approximated with a specified rate of decrease of error in terms of the weak spaces V^*_q. Our point is to isolate conditions on the thresholding which are sufficient for such a characterization. The apropriateness of our assumptions are justified in §5 where we give several examples where these assumptions apply.

We place ourselves under the assumptions of §2.1. Thus, we assume that V is a space which has a p-sequential case (17) with $0 < p < \infty$ and weights p_i, $i \in \mathbb{N}$. We recall the weak spaces $V_q^*(\mu)$ defined in §2.3, with $\mu(i) = p_i$. We shall analyze the performance of the general estimator $\hat{f}^n$ defined by (18).

For $\alpha \in]0,1[$, we also define by $BS = BS_\alpha$ a space of "basic smoothness" associated with our procedure: $f \in BS$ if and only if

$$\sum_{i \notin \Lambda_n} p_i \|f_i\|_i^p \leq Ac(n)^{\alpha p}, \quad n = 1, 2, \ldots.$$

We then have the following result.

Theorem 1 *Let V be a space with a p-sequential norm (17. Let $\alpha \in]0,1[$ and define $q := (1-\alpha)p$. Further, let $c(n)$ be a decreasing sequence tending to 0 such that* $\limsup c(n)/c(n+1) < +\infty$*. For the general thresholding estimator* (18)*, we assume that for each $n \geq 1$, the estimator $\hat{f}_i^n$ satisfies*

(a) $\mathbb{E}\|\hat{f}_i^n - f_i\|_i^{2p} \leq Cc(n)^{2p}$, $i \in \Lambda_n$,

(b) $P(\|\hat{f}_i^n - f_i\|_i \geq \kappa c(n)/2) \leq Kc(n)^\gamma$, $i \in \Lambda_n \setminus \lambda_n$, *for some fixed* $\gamma \geq \alpha p$,

(c)
$$\sum_{i \in \Lambda_n} p_i \leq Cc(n)^{-q-\gamma/2} \text{ and } \sum_{i \in \lambda_n} p_i \leq c(n)^{-q}. \tag{26}$$

Then for $f \in V$, the following conditions are equivalent

$$\begin{array}{l} (i)\ \mathbb{E}\|\hat{f}^n - f\|^p \leq Cc(n)^{\alpha p}, \quad n = 1, 2, \ldots, \\ (ii)\ f \in V_q^* \cap BS. \end{array}$$

Remarks:
(R1) We see that the sequence $c(n)$ appears in three strongly connected points: the conditions (a) and (b) which describe the rate of convergence of the "individual" estimators $\hat{f}_i^n$ to f_i, the rate of thresholding in (18), and finally the resulting rate of convergence in (i).
(R2) Condition (a) is a usual moment condition on the sequence of estimators. Note that, by the Cauchy-Schwartz inequality, it implies $\mathbb{E}\|\hat{f}_i^n - f_i\|_i^p \leq Cc(n)^p$. Condition (b) should be viewed as a concentration property. As will be shown in the examples, it is often the consequence of an exponential inequality.

4.2 Proof of Theorem 1

Here and after, C denotes a constant which may change from one line to the next. The p-sequential assumption (17) implies that

$$\mathbb{E}\|\hat{f}^n - f\|^p \sim \sum_{i \in \lambda_n} p_i \mathbb{E}\|\hat{f}_i^n - f_i\|_i^p + \sum_{i \notin \Lambda_n} p_i \|f_i\|_i^p + \sum_{i \in \Lambda_n \setminus \lambda_n} p_i \mathbb{E}\|\hat{f}_i^n I\{\|\hat{f}_i^n\|_i \geq \kappa c(n)\} - f_i\|_i^p. \tag{27}$$

We shall make use of this equivalence in both directions of the proof.

4.2.1 (i)⇒(ii)

Assuming that (i) holds, we first obtain

$$C_1 \sum_{i \notin \Lambda_n} p_i \|f_i\|_i^p \le C_1 \mathbb{E}\|\hat{f}^n - f\|^p \le Ac(n)^{\alpha p}, \tag{28}$$

for all n, so that $f \in BS$.

Next, we remark that $\|f_i\|_i \le \kappa c(n)/2$ implies $\|f_i\|_i \le \|f_i - \hat{f}_i I\{\|\hat{f}_i\|_i \ge \kappa c(n)\}\|_i$. It follows that

$$\begin{aligned} \textstyle\int \|f_i\|_i^p \quad & I\{i \notin \lambda_n, \|f_i\|_i \le \kappa c(n)/2\} d\mu = \textstyle\sum_{i \notin \lambda_n, \|f_i\|_i \le \kappa c(n)/2} p_i \|f_i\|_i^p \\ & \le \textstyle\sum_{i \notin \Lambda_n} p_i \|f_i\|_i^p + \sum_{i \in \Lambda_n \setminus \lambda_n} p_i \mathbb{E}\|\hat{f}_i^n I\{\|\hat{f}_i^n\|_i \ge \kappa c(n)\} - f_i\|_i^p \\ & \le C\mathbb{E}\|\hat{f}_n - f\|^p \le Cc(n)^{\alpha p} = Cc(n)^{p-q}. \end{aligned}$$

For the indices $i \in \lambda_n$, we note that

$$\int \|f_i\|_i^p I\{i \in \lambda_n, \|f_i\|_i \le \kappa c(n)/2\} d\mu \le (\kappa c(n)/2)^p \sum_{i \in \lambda_n} p_i \le Cc(n)^{p-q}, \tag{29}$$

where we have used the second assumption in (26).

We thus obtain that $\int \|f_i\|_i^p I\{\|f_i\|_i \le \kappa c(n)/2\} d\mu \le Cc(n)^{p-q}$ for all $n > 0$. Using the condition $\limsup c(n)/c(n+1) < +\infty$, it is not difficult to show that this extends to $\int \|f_i\|_i^p I\{\|f_i\|_i \le \epsilon\} d\mu \le C\epsilon^{p-q}$, for all $\epsilon \le \kappa c(1)/2$. For $\epsilon > \kappa c(1)/2$, the same property immediately follows from the fact that $\sum_i p_i \|f_i\|_i^p$ is bounded and $q < p$. We thus conclude that $f \in V_q^*$.

4.2.2 (ii)⇒ (i)

Assuming now that (ii) is true, we first notice that since $f \in BS$, the second term $\sum_{i \notin \Lambda_n} p_i \|f_i\|_i^p$ on the right hand side of (27) is bounded by $Cc(n)^{\alpha p}$.

The first term $\sum_{i \in \lambda_n} p_i \mathbb{E}\|\hat{f}_i^n - f_i\|_i^p$ is also bounded by $Cc(n)^{\alpha p}$ using the moment assumption (a) together with the second assumption in (26).

It remains to estimate the last term $\sum_{i \in \Lambda_n \setminus \lambda_n} p_i \mathbb{E}\|\hat{f}_i^n I\{\|\hat{f}_i^n\|_i \ge \kappa c(n)\} - f_i\|_i^p$. This term can be split into

$$\sum_{i \in \Lambda_n \setminus \lambda_n} p_i \|f_i\|_i^p P\{\|\hat{f}_i^n\|_i < \kappa c(n)\} + \sum_{i \in \Lambda_n \setminus \lambda_n} p_i \mathbb{E}(\|\hat{f}_i^n - f_i\|_i^p I\{\|\hat{f}_i^n\|_i \ge \kappa c(n)\})$$

We further more split this sum $I + II$ into $I_A + I_B + II_A + II_B$ where these terms are defined and estimated as follows:

$$\begin{aligned} I_A \quad & := \textstyle\sum_{i \in \Lambda_n \setminus \lambda_n} I\{\|f_i\|_i \ge 2\kappa c(n)\} p_i \|f_i\|_i^p P\{\|\hat{f}_i^n\|_i < \kappa c(n)\} \\ & \le \textstyle\sum_{i \in \Lambda_n \setminus \lambda_n} p_i \|f_i\|_i^p P\{\|\hat{f}_i^n - f_i\|_i \ge \kappa c(n)\} \\ & \le Cc(n)^\gamma \textstyle\sum_{i \in \Lambda_n \setminus \lambda_n} p_i \|f_i\|_i^p \\ & \le Cc(n)^\gamma \|f\|^p \le Cc(n)^{\alpha p}, \end{aligned}$$

where we have used (b) and the assumption $\gamma \ge \alpha$.

$$\begin{aligned} I_B \quad &:= \textstyle\sum_{i\in\Lambda_n\setminus\lambda_n} I\{\|f_i\|_i < 2\kappa c(n)\} p_i \|f_i\|_i^p P\{\|\hat{f}_i^n\|_i < \kappa c(n)\} \\ &\le \textstyle\sum_{i\in\Lambda_n\setminus\lambda_n} I\{\|f_i\|_i < 2\kappa c(n)\} p_i \|f_i\|_i^p \\ &\le C(2\kappa c(n))^{p-q} = Cc(n)^{\alpha p}, \end{aligned}$$

where we have used the V_q^* assumption in the form (II).

$$\begin{aligned} II_A \quad &:= \textstyle\sum_{i\in\Lambda_n\setminus\lambda_n} I\{\|f_i\|_i \ge \kappa c(n)/2\} p_i \mathbb{E}(\|\hat{f}_i^n - f_i\|_i^p I\{\|\hat{f}_i^n\|_i \ge \kappa c(n)\}) \\ &\le Cc(n)^p \mu\{i: \ \|f_i\|_i \ge \kappa c(n)/2\} \le Cc(n)^{p-q} = Cc(n)^{\alpha p}, \end{aligned}$$

where we have used (a) and the V_q^* assumption in the form (I).

$$\begin{aligned} II_B \quad &:= \textstyle\sum_{i\in\Lambda_n\setminus\lambda_n} I\{\|f_i\|_i < \kappa c(n)/2\} p_i \mathbb{E}(\|\hat{f}_i^n - f_i\|_i^p I\{\|\hat{f}_i^n\|_i \ge \kappa c(n)\}) \\ &\le \textstyle\sum_{i\in\Lambda_n\setminus\lambda_n} p_i \mathbb{E}(\|\hat{f}_i^n - f_i\|_i^p I\{\|\hat{f}_i^n - f_i\|_i \ge \kappa c(n)/2\}) \\ &\le \textstyle\sum_{i\in\Lambda_n\setminus\lambda_n} p_i (\mathbb{E}\|\hat{f}_i^n - f_i\|_i^{2p})^{1/2} (P\{\|\hat{f}_i^n - f_i\|_i \ge \kappa c(n)/2\})^{1/2} \\ &\le Cc(n)^p c(n)^{\gamma/2} \textstyle\sum_{i\in\Lambda_n\setminus\lambda_n} p_i \le Cc(n)^{\alpha p}, \end{aligned}$$

where we have used Schwarz inequality, (a), (b) and the first assumption in (26).

We thus concludes that the estimation rate (i) is statisfied.

4.3 When the BS condition is not necessary

Under certain conditions, we can avoid the extra BS condition in the statement (ii) of the Theorem 1. This essentially occurs when the set Λ_n is related to the ordering with the weights p_i. More precisely, we suppose that there exists $\tau > -1$ such that for all $r > 0$,

$$\#\{i \,;\, p_i \le r\} \le K r^\tau. \tag{30}$$

We define $\tilde{\Lambda}_n = \{i \in \mathbb{N}: \ p_i \le c(n)^{-\delta}\}$ for some fixed $\delta > 0$, and consider the following thresholding estimator

$$\hat{f}_n = \sum_{i\in\lambda_n} \hat{f}_i^n + \sum_{i\in\tilde{\Lambda}_n\setminus\lambda_n} \hat{f}_i^n I(\|\hat{f}_i^n\|_i \ge \kappa c(n)) \tag{31}$$

Theorem 2 *Let $\alpha \in]0,1[$. Assume that* (30) *holds and consider the modified thresholding estimator* (31)*. We suppose that the sequence of estimators $\hat{f}_i^n$ satisfies the assumptions* (a) *and* (b) *of Theorem 1 with $\gamma \ge \alpha p$ and that $c(n)$ is a decreasing sequence tending to* 0 *such that* $\limsup c(n)/c(n+1) < +\infty$*. Setting $q := (1-\alpha)p$, we assume that $\delta \ge q$, that $\gamma/2 - (\tau+1)\delta + q \ge 0$, and that $\sum_{i\in\lambda_n} p_i \le c(n)^{-q}$. Then, for $f \in V$, the following conditions are equivalent*

$$\begin{aligned} &(i)\ \mathbb{E}\|\hat{f}^n - f\|^p \le Cc(n)^{\alpha p}, \quad n = 1, 2, \ldots, \\ &(ii)\ f \in V_q^*. \end{aligned}$$

Proof

The proof of (i) ⇒ (ii) is exactly the same as the proof of the L_q^* property in the first part (i) ⇒ (ii) of Theorem 1.

For proving (ii) ⇒ (i), we again consider the equivalent expression of the loss (27) where we replace Λ_n by $\tilde{\Lambda}_n$. We can bound the first term $\sum_{i\in\lambda_n} p_i \mathbb{E}\|\hat{f}_i - f_i\|_i^p$ by $Cc(n)^{\alpha p}$ in the same way as in the proof of Theorem 1.

For the second term, we remark that since $f \in V_q^*$ and $\delta \geq q$, there exists a constant C such that we have

$$\sum_{\|f_i\|_i \geq Cc(n)} p_i \leq c(n)^{-\delta}.$$

Therefore, if $i \notin \tilde{\Lambda}_n$, i.e. $p_i > c(n)^{-\delta}$, we necessarily have $\|f_i\|_i \leq Cc(n)$. In turn, we obtain

$$\sum_{i \notin \tilde{\Lambda}_n} p_i \|f_i\|_i^p \leq \sum_{\|f_i\|_i \leq Cc(n)} p_i \|f_i\|_i^p \leq Cc(n)^{\alpha p}, \tag{32}$$

where in the last inequality we used the fact that $f \in V_q^*$ in the form of (II) of §3 with $r = p$.

For the third term, the estimation of I_A, I_B and II_A is left unchanged. For the estimation of II_B, starting as in the proof of Theorem 1, we obtain

$$\begin{aligned} II_B &\leq Cc(n)^{p+\gamma/2} \textstyle\sum_{i \in \tilde{\Lambda}_n \setminus \lambda_n} p_i \\ &\leq Cc(n)^{p+\gamma/2} \textstyle\sum_{p_i \leq c(n)^{-\delta}} p_i \\ &\leq Cc(n)^{p+\gamma/2} \textstyle\sum_{j \geq 0} \sum_{2^{-j-1} c(n)^{-\delta} < p_i \leq 2^{-j} c(n)^{-\delta}} p_i \\ &\leq Cc(n)^{p+\gamma/2} \textstyle\sum_{j \geq 0} [2^{-j} c(n)^{-\delta}]^{\tau+1} \\ &\leq Cc(n)^{p+\gamma/2-\delta(\tau+1)} \leq Cc(n)^{p-q} = Cc(n)^{\alpha p}, \end{aligned}$$

where we have used the assumption $\gamma/2 - (\tau + 1)\delta + q \geq 0$. □

4.4 The supremum case

We now consider the case where the loss function is expressed by $\|f\|_V = \sup_{i \geq 0} p_i \|f_i\|_i$. For $\alpha \in]0, 1[$, we define the space V_∞^α which consists of all functions $f \in V$ such that for all $\epsilon > 0$,

$$\text{(I)} \quad \sup_{\|f_i\|_i| \leq \lambda} \|f_i\|_i p_i \leq C\lambda^\alpha.$$

One easily checks that (I) is equivalent to

$$\text{(II)} \quad \sup_i \|f_i\|_i p_i^{1+\rho} < \infty,$$

with $1/\rho = 1/\alpha - 1$.

Let us now suppose that condition (30) is satisfied, i.e. there exists $\tau > -1$ such that $\#\{i \ ; \ p_i \leq r\} \leq Kr^\tau$. As in §4.3, we let $\tilde{\Lambda}_n = \{i \in \mathbb{N} : \ p_i \leq c(n)^{-\delta}\}$ for some fixed $\delta > 0$, and we consider the modified thresholding estimator (31).

Theorem 3 *Let $\alpha \in]0, 1[$. Assume that* (30) *holds and consider the modified thresholding estimator* (31). *We suppose that the sequence of estimators $\hat{f}_i^n$ satisfies the assumptions*

(a′) $\mathbb{E}(\sup_{i \in \tilde{\Lambda}_n} \|\hat{f}_i^n - f_i\|_i^2) \leq Cc(n)^2, \quad n = 1, 2, \ldots,$

(b′) $P(\sup_{i \in \tilde{\Lambda}_n} \|\hat{f}_i^n - f_i\|_i \geq \kappa c(n)/2) \leq Kc(n)^\gamma, \quad n = 1, 2, \ldots,$

(c)

$$\sum_{i \in \Lambda_n} p_i \leq Cc(n)^{-q-\gamma/2} \text{ and } \sum_{i \in \lambda_n} p_i \leq c(n)^{-q}. \tag{33}$$

with γ such that $\alpha \le \min\{1+\gamma/2-\delta, \gamma\}$, and with $(c(n))$ a decreasing sequence tending to 0 *such that* $\limsup c(n)/c(n+1) < +\infty$*. Moreover, we assume that $\delta \ge 1-\alpha$ and $\#(\lambda_n) \le Cc(n)^{\delta+\alpha-1}$. Then for $f \in V$, the following conditions are equivalent*

$$\begin{array}{l} (i)\ \mathbb{E}\|\hat{f}^n - f\| \le Cc(n)^{\alpha}, \quad n = 1, 2, \ldots, \\ (ii)\ f \in V_{\infty}^{\alpha}. \end{array}$$

Proof:

4.4.1 (i)⇒ (ii)

If $\mathbb{E}\|\hat{f}^n - f\| \le Cc(n)^{\alpha}$ then $\sup_{i \notin \tilde{\Lambda}_n} p_i \|f_i\|_i \le Kc(n)^{\alpha}$ by defintion fo the estimator. Next, we remark that if $\|f_i\|_i \le \kappa c(n)/2$ then $\|f_i\|_i \le \|f_i - \hat{f}_i^n I\{\|\hat{f}_i^n\|_i \ge \kappa c(n)\}\|_i$. It follows that

$$\begin{array}{ll} \sup_{i \in \tilde{\Lambda}_n \ \|f_i\|_i \le \kappa c(n)/2} p_i \|f_i\|_i & \le \mathbb{E}(\sup_{i \in \tilde{\Lambda}_n \ \|f_i\|_i \le \kappa c(n)/2} p_i \|f_i - \hat{f}_i I\{\|\hat{f}_i\|_i \ge \kappa c(n)\}\|_i) \\ & \le Cc(n)^{\alpha}. \end{array}$$

Combining these estimates, we obtain that $\sup_{\|f_i\|_i \le \kappa c(n)/2} p_i \|f_i\|_i \le Cc(n)^{\alpha}$, and thus

$$\sup_{\|f_i\|_i \le \lambda} p_i \|f_i\|_i \le C\lambda^{\alpha},$$

for all $\lambda \le \kappa c(1)/2$. For the large values of λ, the above is true since $\|f\|_V = \sup_i p_i \|f_i\|_i < \infty$. Therefore $f \in V_{\infty}^{\alpha}$.

4.4.2 (ii)⇒ (i)

Suppose now that $f \in V_{\infty}^{\alpha}$. We bound the estimation error by

$$\mathbb{E}\|\hat{f}^n - f\| \le C[\mathbb{E}(\sup_{i \in \tilde{\Lambda}_n} p_i \|\hat{f}_i^n I\{\|\hat{f}_i^n\| \ge \kappa c(n)\} - f_i\|_i) + \sup_{i \notin \tilde{\Lambda}_n} p_i \|f_i\|_i + \mathbb{E}(\sup_{i \in \lambda_n} p_i \|\hat{f}_i^n - f_i\|_i)].$$

¿From the equivalent form (II) of the definition of V_{∞}^{α}, we have for the second term the estimate

$$\sup_{i \notin \tilde{\Lambda}_n} p_i \|f_i\|_i \le \sup_{i \notin \tilde{\Lambda}_n} p_i^{-\rho} \le Cc(n)^{\delta\rho} \le Cc(n)^{\alpha},$$

where we have used that $\delta\rho \ge (1-\alpha)\rho = \alpha$.

For the third term, we have the estimate

$$\mathbb{E}(\sup_{i \in \lambda_n} p_i \|\hat{f}_i^n - f_i\|_i) \le \mathbb{E}(\sum_{i \in \lambda_n} p_i \|\hat{f}_i^n - f_i\|_i) \le Cc(n) \sum_{i \in \lambda_n} p_i \le Cc(n)^{1-\delta} \#(\lambda_n) \le Cc(n)^{\alpha},$$

where we have used the assumtion on the cardinality of λ_n and the fact that if $i \in \lambda_n \subset \tilde{\Lambda}_n$ we have $p_i \le c(n)^{-\delta}$.

It remains to estimate the first term. As in the proof of Theorem 1, we write

$$\mathbb{E}(\sup_{i \in \tilde{\Lambda}_n} p_i \|\hat{f}_i^n I\{\|\hat{f}_i^n\| \ge \kappa c(n)\} - f_i\|_i \le I_A + I_B + II_A + II_B,$$

where these four terms are defined and estimated as follows.

$$\begin{aligned} I_A &:= \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n, \|f_i\|_i \geq 2\kappa c(n)} p_i\|f_i\|_i I\{\|\hat{f}_i^n\|_i < \kappa c(n)\}) \\ &\leq \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n, \|f_i\|_i \geq 2\kappa c(n)} p_i\|f_i\|_i I\{\|\hat{f}_i^n - f_i\|_i \geq \kappa c(n)\}) \\ &\leq \|f\|_V P\{\sup_{\infty\in\tilde{\Lambda}_n} \|\hat{f}_i^n - f_I\|_i \geq \kappa c)n)\} \\ &\leq Cc(n)^\gamma \leq Cc(n)^\alpha, \end{aligned}$$

$$\begin{aligned} I_B &:= \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n, \|f_i\|_i < 2\kappa c(n)} p_i\|f_i\|_i I\{\|\hat{f}_i^n\|_i < \kappa c(n)\}) \\ &\leq \sup_{\|f_i\|_i < 2\kappa c(n)} p_i\|f_i\|_i \\ &\leq Cc(n)^\alpha, \end{aligned}$$

directly from the V_∞^α assumption.

$$\begin{aligned} II_A &:= \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n, \|f_i\|_i \geq \kappa c(n)/2} p_i\|\hat{f}_i^n - f_i\|_i I\{\|\hat{f}_i^n\|_i \geq \kappa c(n)\}) \\ &\leq \sup_{\|f_i\|_i \geq \kappa c(n)/2} p_i \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n} \|\hat{f}_i^n - f_i\|_i) \\ &\leq Cc(n)^{-1/(1+\rho)} \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n} \|\hat{f}_i^n - f_i\|_i) \\ &\leq Cc(n)^{-1/(1+\rho)} \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n} \|\hat{f}_i^n - f_i\|_i) \\ &\leq Cc(n)^{\rho/(1+\rho)} = Cc(n)^\alpha, \end{aligned}$$

where we have used that $f \in V_\infty^\alpha$ in the form of (II) and then (a').

$$\begin{aligned} II_B &:= \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n, \|f_i\|_i < \kappa c(n)/2} p_i\|\hat{f}_i^n - f_i\|_i I\{\|\hat{f}_i^n\|_i \geq \kappa c(n)\}) \\ &\leq \mathbb{E}(\sup_{i\in\tilde{\Lambda}_n, \|f_i\|_i < \kappa c(n)/2} p_i\|\hat{f}_i^n - f_i\|_i I\{\|\hat{f}_i^n - f_i\|_i \geq \kappa c(n)/2\}) \\ &\leq (\sup_{i\in\tilde{\Lambda}_n} p_i)(\mathbb{E}(\|\hat{f}_i^n - f_i\|_i^2)^{1/2}(P(I\{\|\hat{f}_i^n\|_i \geq \kappa c(n)\})^{1/2} \\ &\leq Cc(n)^{-\delta} c(n) c(n)^{\gamma/2}, \end{aligned}$$

where we used in the last inequality the definition of $\tilde{\Lambda}_n$, and (a') and (b').

5 Examples

We shall now show how Theorems 1-3 can be applied to specific thresholding estimators based on wavelet decompositions. We shall restrict our examples to the settings put forward in §2.

5.1 Local Thresholding

We consider the setting of §2.2.1. The thresholding estimator $\hat{f}^n$ is given by (19) with the thresholding parameter $c(n) = (\frac{\log(n)}{n})^{1/2}$ which was introduced in [7]. Thus $\lambda_n = \emptyset$ and $\Lambda_n = \tilde{\Lambda}_n = \{\lambda : |\lambda| \leq j_0(n)\}$ with $j_0(n)$ defined by the relation $2^{j_0(n)} \approx \frac{n}{\log n}$. The space V can be taken as any of the Besov spaces $B_p^\sigma(L_p)$ (with $p = \infty$ in Theorem 3), $\sigma \geq 0$. The Besov norm $\|\cdot\|_{B_p^\sigma(L_p)}$ is a p-sequential measurement of the loss with the weights $p_\lambda = 2^{(\sigma p + p/2 - 1)|\lambda|}$, $\lambda \in \Delta$.

We shall discuss the conditions of Theorems 1-3 for the white noise model (5). In this case, the $\hat{f}_i^n - f_i$ form an orthonormal sequence of iid $N(0, 1/n)$ variables. Consider first the conditions of Theorem 1. The properties (a) and (b) are very classical. Let us only remark that in fact for any arbitrary $\gamma > 0$, there exists $\kappa(\gamma)$ such that (b) is fulfilled, since using the concentration properties of the Gaussian measure, we have $P(|X| > \lambda) \leq 2\exp(-\lambda^2/2)$. Concerning condition (c) of Theorem 1, one has that $\sum_{\lambda\in\Lambda_n} p_\lambda = \sum_{j=0}^{j_0(n)} 2^j 2^{j(\sigma p+p/2-1)} \approx (c(n))^{-(2\sigma+1)p}$. Hence, the first requirement in (c) is satisfied provided that $q+\gamma/2 \geq (2\sigma+1)p$. By our remarks that γ can be chosen arbitrarily large we see that the first condition of (c) is always satisfied. The second condition in (c) is satisfied automatically since $\lambda_n = \emptyset$. Hence, all conditions are satisfied with no restrictions on the parameters.

We can also apply Theorem 2 with the same thresholding estimator (19). We need to check condition (30). From the definition of the p_λ, we see that this condition is satisfied provided $\sigma p+p-1>0$ and $\tau \geq (\sigma p+p-1)^{-1}$ (note that we need only need to check (30) for $r \geq 1$ since for $r<1$ the set in (30) is empty). We shall take $\tau := (\sigma p+p-1)^{-1}$. If we take $\delta := 2(\sigma p+p/2-1)$ then the set $\tilde{\Lambda}_n = \Lambda_n$ and the estimator of Theorem 2 coincides with (19). We also have the requirement in Theorem 2 that $\delta \geq q = (1-\alpha)p$. This will be satisfied with the above choice of δ provided $\sigma - 1/p > -\alpha/2$. The final condition $\gamma/2-(\tau+2)\delta+q \geq 0$ will be satisfied if γ is sufficiently large. Since, as observed earlier, we can choose γ as large as we wish, Theorem 2 is applicable whenever $\sigma-1/p > -\alpha/2$ provided γ is sufficiently large.

Regarding the application of Theorem 3 in this setting, all conditions of that theorem are satisfied if γ is sufficiently large.

With these calculations behind us, we see that Theorem 1-3 give the following theorem.

Theorem 4 *Let $0<p<\infty$, $\sigma \geq 0$, $\alpha \in]0,1[$ and $q=(1-\alpha)p$. Let $\kappa=\kappa(\gamma)$ be associated to any sufficiently large γ (e.g. $\gamma > (\tau+2\delta)+q$ will suffice). For the thresholding estimator (19) we have that the maximal space G consisting of those functions $f \in V = B^\sigma_{p,p}([0,1])$ such that*

$$\mathbb{E}\|\hat{f}_n - f\|_V^p \leq C(\frac{\log n}{n})^{\alpha p/2},$$

coincides with:

1. *The weak space V_q^*, i.e. $WB^s_q(L_q)$, with $s=\sigma p/q+(p/q-1)/2$, in the case $\sigma-1/p > -\alpha/2$ (by application of Theorem 2),*

2. *The intersection $V_q^* \cap B^{\alpha/2+\sigma}_\infty(L_p)$ in the case $\sigma - 1/p \leq -\alpha/2$ (by application of Theorem 1).*

3. *In the case where $p=\infty$, with $\kappa=\kappa(\gamma)$ associated to any sufficiently large γ, the maximal space G consisting of those functions $f\in V = B^\sigma_{\infty,\infty}([0,1])$ such that*

   $$\mathbb{E}\|\hat{f}_n - f\|_V \leq C(\frac{\log n}{n})^{\alpha/2},$$

   coincides with the space $V^\alpha_\infty = B^s_{\infty,\infty}$ with $\alpha = 2(s-\sigma)/(1+2s)$ (by application of Theorem 3).

Remarks

(R1) Note that the rate of convergence obtained here is minimax since it is known to be minimax for $B^s_{q,q}$ which is a subspace of $WB^s_{q,q}$.

(R2) We discussed the case of a white noise model. Of course, the same result holds for the regression case at least with Gaussian errors. In the density estimation, the result is still true if the maximal space includes L^∞ since in this case it seems necessary to assume f bounded in order to obtain the concentration property (b) (see [7]).

-This result can easily be extended to the estimation of multivariate functions, using isotropic tensor product type wavelets which characterize Besov spaces in several dimensions.

5.2 Global thresholding

This technique was described in §2.2.2, with the Besov norm $\|\cdot\|_{B^\sigma_2(L_2)}$ as a p-sequential measurement of the loss (with $p = 2$), and weights $p_j := 2^{(2\sigma+1)j}$. The threshold estimator f^n is given by (21) with the thresholding parameter $c(n) = n^{-1/2}$ and with $j_0(n)$ and $j_1(n)$ chosen so that $2^{j_0(n)} \sim n$ and $2^{j_1(n)} \sim n^\epsilon$.

We wish to apply Theorem 2, and therefore, we check that the conditions of this theorem are satisfied. Note that $\delta = 2$. ¿From the definition of the p_j, we have that

$$\#\{j\ ;\ p_j \le r\} \le \frac{\log r}{2\sigma}.$$

Therefore, condition (30) is satisfied for any $\tau > 0$ (note that we need only check this condition for $r \ge 1$ since the set in (30) is empty if $r \le 1$). Condition (a) is immediate to prove. We also easily obtain that for any arbitrary γ, there exists $\kappa(\gamma)$ such that (b) is true, since $P\{2^{-j}n\sum_{|\lambda|=j}|\hat{\beta}_\lambda - \beta_\lambda|^2 \ge C\} = P\{\sum_i Y_i^2 \ge 2^jC\}$ where the Y_i are $C2^j$ independent $N(0,1)$ random variables. Hence $P\{\sum_i Y_i^2 \ge 2^jC\} \le \exp\{-2^jh(C)\} \le \exp\{-n^\varepsilon h(C)\}$ where h is the Cramer transform of Y_i^2 and ε the parameter such that $2^{j_1(n)} \sim n^\varepsilon$. Since we are working on a finite interval, e.g. $[0,1]$, we can also write $\sum_{j\in\lambda_n} p_i = 1 + \sum_{j=0}^{j_1(n)} 2^{j(2\sigma+1)} \le Cc(n)^{-2(2\sigma+1)\varepsilon}$. So the condition on λ_n in Theorem 2 is satisfied if $2(2\sigma+1)\varepsilon \ge q$. The condition $\delta \ge q$ is automatically satisfied since $\delta = 2$ and $q = (1-\alpha)2$.

Application of Theorem 2 then gives the following

Theorem 5 *Let $\alpha \in]0,1[$, $q = 2(1-\alpha) \le 2(2\sigma+1)\varepsilon$ and assume that $\kappa = \kappa(\gamma)$ is associated to any sufficiently large γ (e.g. $\gamma > 4\alpha$ will do). Then the maximal space G consisting of those functions $f \in V = B^\sigma_2(L_2[0,1])$ such that*

$$\mathbb{E}\|\hat{f}^n - f\|_V^2 \le Cn^{-\alpha},$$

coincides with V_q^*.

Remarks

(R1) The space V_q^* is turns out to be a weak space for $B^s_2(L_q)$, $s = \sigma + t$ where $t > 0$ is such that $\alpha = 4t/(1+2t)$. Again the rate of convergence is minimax since it is for $B^s_2(L_q)$.

(R2) If we want to compare global thresholding with local thresholding, we find that the rate of convergence is better (since $c(n)$ does not contain any additional logarithmic term) but the space $B^s_2(L_q)$ is smaller than $B^s_q(L_q)$ (and the the same is true for their associated the weak spaces).

5.3 Block thresholding using kernel

In this case, described in §2.2.3, we give without details the application of Theorem 1. Here we only consider the L^2-loss. It is not surprizing that we obtain intermediate results between local and global thresholdings in terms of the magnitude of the maximal space, for projection kernels associated to a multiresolution analysis. Let us mention however that the maximal spaces obtained with other kernels are much more difficult to identify.

We define $\bar{B}^{\alpha/2}_{2,\infty}$ the space of functions f such that $\|D_j f\|_2^2 \leq C(j2^j)^{-\alpha}$. We then have the following.

Theorem 6 *The maximal space G consisting of those functions $f \in V = L^2([0,1])$ such that*

$$\mathbb{E}\|\hat{f}_n - f\|_V^2 \leq Cn^{-\alpha},$$

coincides with the intersection of $L^*_q(\mu)$ and $\bar{B}^{\alpha/2}_{2,\infty}$.

6 L_p-loss in the local thresholding framework

The key-point of t§4 was the sequential form of the norm and all the calculations highly rely on this fact. A natural question arises: what happens if one uses the L_p-loss which is *not p-sequential* when $p \neq 2$? We shall answer this question in the case of local thresholding and show that, when $1 < p < \infty$, the saturation results are the same for L_p as for $B^0_p(L_p)$.

6.1 Statement of the result

For simplicity, we consider the setting of wavelet decompositions on the interval $[0,1]$ and we shall take the thresholding parameter $c(n) = (\frac{\log(n)}{n})^{1/2}$ which was introduced in [7]. We define

$$\Lambda_n = \{\lambda \in \Delta \ ; \ 2^{-|\lambda|} \leq c(n)^2\},$$

and the corresponding thresholding estimator

$$\hat{f}_n = \sum_{\lambda \in \Lambda_n} \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\}\psi_\lambda.$$

For p fixed in $]1,\infty[$, we set $\mu(\lambda) = 2^{(p/2-1)|\lambda|}$ and define $L^*_q(\mu)$ the corresponding weak space.

Let us make some observations which show that the estimators $\hat{f}^n$ satisfy the conditions of §4.1 for the space $B^0_p(L_p)$. Note that $\lambda_n = \emptyset$. Let $N := N(n)$ be defined by

$2^N = \frac{n}{\log n} = c(n)^{-2}$. This means that $\Lambda_n = \{\lambda : \; |\lambda| \leq N\}$ and therefore

$$\mu(\Lambda_n) = \sum_{|\lambda| \leq N} 2^{(p/2-1)|\lambda|} = \sum_{j=0}^{N} 2^{jp/2} \leq C 2^{Np/2} = Cc(n)^{-p}. \tag{34}$$

Also, for the p-sequential space $B^0_p(L_p)$, we have $p_\lambda = \mu(\lambda)$ and therefore

$$\sum_{\lambda \in \Lambda_n} p_\lambda = \mu(\Lambda_n) \leq Cc(n)^{-p} \tag{35}$$

This means that condition (c) of Theorem 1 applies for the space $V = B^0_p(L_p)$ provided $\gamma \geq \alpha p$. We now show that we can replace $B^0_p(L_p)$ by L_p in that theorem.

Theorem 7 *Let $0 < q < p$ and $\alpha > 0$ such that $q = (1-\alpha)p$. We assume that the estimators $\hat{\beta}^n_\lambda$ satisfy*

(a) $\mathbb{E}|\hat{\beta}^n_\lambda - \beta_\lambda|^{2p} \leq Cc(n)^{2p}, \; \lambda \in \Lambda_n$

(b) $P(|\hat{\beta}^n_\lambda - \beta_\lambda| \geq \kappa c(n)/2) \leq Kc(n)^\gamma, \; \lambda \in \Lambda_n$

with $\gamma \geq 2\alpha p$. Then the following are equivalent

$$\begin{array}{l} \text{(i) } \mathbb{E}\|\hat{f}_n - f\|^p_{L_p} \leq Cc(n)^{\alpha p} \\ \text{(ii) } f \in V^*_q(\mu) \cap B^{\alpha/2}_\infty(L_p). \end{array}$$

6.2 Proof of Theorem 7

6.2.1 (i) ⇒ (ii)

We assume $\mathbb{E}\|\hat{f}_n - f\|^p_{L_p} \leq Cc(n)^{\alpha p}$, $n = 1, 2, \ldots$. Since $(\psi_\lambda)_{\lambda \in \Delta}$ is an unconditional basis of L_p, we first obtain as an immediate consequence

$$\|\sum_{\lambda \notin \Lambda_n} \beta_\lambda \psi_\lambda\|^p_{L_p} \leq C\mathbb{E}\|\hat{f^n} - f\|^p_{L_p} \leq Cc(n)^{\alpha p}, \tag{36}$$

and

$$\mathbb{E}\|\sum_{\lambda \in \Lambda_n} (\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\})\psi_\lambda\|^p_{L_p} \leq C\mathbb{E}\|\hat{f^n} - f\|^p_{L_p} \leq Cc(n)^{\alpha p}. \tag{37}$$

The estimate (36) is equivalent to

$$\|\sum_{|\lambda| \geq j} \beta_\lambda \psi_\lambda\|^p_{L_p} \leq C 2^{j\alpha p/2},$$

for all $j \geq 0$, i.e. $f \in B^{\alpha/2}_\infty(L_p)$.

Next, observe that $|\beta_\lambda| \leq \kappa c(n)/2$ implies $|\beta_\lambda| \leq |\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\}|$. Since $(\psi_\lambda)_{\lambda \in \Delta}$ is an unconditional basis of L_p, we have

$$\begin{aligned} \|f - \textstyle\sum_{|\beta_\lambda| \geq \kappa c(n)/2} \beta_\lambda \psi_\lambda\|^p_{L_p} &= \|\textstyle\sum_{|\beta_\lambda| \leq \kappa c(n)/2} \beta_\lambda \psi_\lambda\|^p_{L_p} \\ &\leq C[\|\textstyle\sum_{\lambda \notin \Lambda_n} \beta_\lambda \psi_\lambda\|^p_{L_p} + \|\textstyle\sum_{\lambda \in \Lambda_n, |\beta_\lambda|\ leq\kappa c(n)/2} \beta_\lambda \psi_\lambda\|^p_{L_p}] \\ &\leq C[c(n)^{\alpha p} + \|\textstyle\sum_{\lambda \in \Lambda_n, |\beta_\lambda| \leq \kappa c(n)/2} \beta_\lambda \psi_\lambda\|^p_{L_p}] \\ &\leq C[c(n)^{\alpha p} + \mathbb{E}\|\textstyle\sum_{\lambda \in \Lambda_n} (\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\})\psi_\lambda\|^p_{L_p}], \end{aligned}$$

Combining with (37), we obtain the estimate

$$\|f - \sum_{|\beta_\lambda| \geq \kappa c(n)/2} \beta_\lambda \psi_\lambda\|_{L_p}^p \leq Cc(n)^{\alpha p}, \quad n = 1, 2, \ldots.$$

By Lemma 3, we conclude that $f \in V_q^*(\mu)$.

6.2.2 (ii) ⇒ (i)

We first remark that for each $n = 1, 2, \ldots$,

$$\mathbb{E}\|\hat{f}^n - f\|_{L_p}^p \leq C[\mathbb{E}\|\sum_{\lambda \in \Lambda_n} (\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\})\psi_\lambda\|_{L_p}^p + \|\sum_{\lambda \notin \Lambda_n} \beta_\lambda \psi_\lambda\|_{L_p}^p]$$

The second term is bounded by $Cc(n)^{\alpha p}$ since $f \in B_\infty^{\alpha/2}(L_p)$. For the first term, in the case $1 < p \leq 2$, we use our remarks from the begining of this section that Theorem 1 holds for $B_p^0(L_p)$. Since $\|\cdot\|_{L^p} \leq C\|\cdot\|_{B_p^0(L_p)}$, we immediately obtain

$$\begin{aligned}&\mathbb{E}\|\textstyle\sum_{\lambda \in \Lambda_n}(\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\})\psi_\lambda\|_{L_p}^p \\ &\leq C\mathbb{E}\|\textstyle\sum_{\lambda \in \Lambda_n}(\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\})\psi_\lambda\|_{B_{p,p}^0}^p \leq Cc(n)^{\alpha p},\end{aligned}$$

where we have used Theorem 1.

When $p > 2$, we first use the square function characterization of Lemma 4 to obtain

$$\begin{aligned}&\mathbb{E}\|\textstyle\sum_{\lambda \in \Lambda_n}(\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\})\psi_\lambda\|_{L_p}^p \\ &\leq \mathbb{E}\textstyle\int(\sum_{\lambda \in \Lambda_n}|(\beta_\lambda - \hat{\beta}_\lambda I\{|\hat{\beta}_\lambda| \geq \kappa c(n)\})\psi_\lambda|^2)^{p/2} \\ &\leq \textstyle\int\Big[\sum_{\lambda \in \Lambda_n,}|\beta_\lambda\psi_\lambda|^2(P_n(|\hat{\beta}_\lambda| \leq \kappa c(n)))^{2/p}\Big]^{p/2} \\ &+\textstyle\int\Big[\sum_{\lambda \in \Lambda_n,}|\psi_\lambda|^2(\mathbb{E}(I\{|\hat{\beta}_\lambda| > \kappa c(n)\}|(\hat{\beta}_\lambda - \beta_\lambda)|^p))^{2/p}\Big]^{p/2} \\ &=: I + II.\end{aligned}$$

where we have used the generalized Minkowski inequality in the second inequality.

Up to a multiplicative constant, we can bound $I + II$ by $I_A + I_B + II_A + II_B$ where these terms (similar to those in the proof of Theorem 1) are defined and estimated as follows:

$$\begin{aligned}I_A &:= \textstyle\int\Big[\sum_{\lambda \in \Lambda_n,\ |\beta_\lambda| > 2\kappa c(n)}|\beta_\lambda\psi_\lambda|^2(P(|\hat{\beta}_\lambda| \leq \kappa c(n)))^{2/p}\Big]^{p/2} \\ &\leq \textstyle\int\Big[\sum_{\lambda \in \Lambda_n,\ |\beta_\lambda| > 2\kappa c(n)}|\beta_\lambda\psi_\lambda|^2(P(|\beta_\lambda - \hat{\beta}_\lambda| \geq \kappa c(n)))^{2/p}\Big]^{p/2} \\ &\leq Cc(n)^\gamma\|\textstyle\sum_{\lambda \in \Lambda_n}\beta_\lambda\psi_\lambda\|_{L_p}^p \leq Cc(n)^\gamma\|f\|_{L_p}^p \leq Cc(n)^{\alpha p}\|f\|_{L_p}^p,\end{aligned}$$

where we have used the concentration property (b) and the fact that $\gamma \geq 2\alpha p \geq \alpha p$.

$$\begin{aligned}I_B &:= \textstyle\int\Big[\sum_{\lambda \in \Lambda_n,\ |\beta_\lambda| \leq 2\kappa c(n)}|\beta_\lambda\psi_\lambda|^2(P(|\hat{\beta}_\lambda| \leq \kappa c(n)))^{2/p}\Big]^{p/2} \\ &\leq \textstyle\int\Big[\sum_{|\beta_\lambda| \leq 2\kappa c(n)}|\beta_\lambda\psi_\lambda|^2\Big]^{p/2} \\ &\leq C\|\textstyle\sum_{|\beta_\lambda| \leq 2\kappa c(n)}\beta_\lambda\psi_\lambda\|_{L_p}^p \\ &\leq Cc(n)^{\alpha p},\end{aligned}$$

where the last inequality stems from Lemma 3.

$$\begin{aligned} II_A \ &:= \int \Big[\textstyle\sum_{\lambda\in\Lambda_n, |\beta_\lambda|>\kappa c(n)/2} |\psi_\lambda|^2 (\mathbb{E}(I\{|\hat\beta_\lambda| > \kappa c(n)\}|(\hat\beta_\lambda - \beta_\lambda)|^p))^{2/p}\Big]^{p/2} \\ &\le Cc(n)^p \int \Big[\textstyle\sum_{\lambda\in\Lambda_n, |\beta_\lambda|>\kappa c(n)/2} |\psi_\lambda|^2 \Big]^{p/2} \\ &\le Cc(n)^p \| \textstyle\sum_{\lambda\in\Lambda_n, |\beta_\lambda|>\kappa c(n)/2} \psi_\lambda \|_{L_p}^p \\ &\le Cc(n)^p \mu(\{\lambda \in \Lambda_n, |\beta_\lambda| > \kappa c(n)/2\}) \\ &\le Cc(n)^{p-q} = Cc(n)^{\alpha p}. \end{aligned}$$

where we have used assumption (a), Lemma 1, and in the last inequality the assumption that $f \in V_q^*(\mu)$.

$$\begin{aligned} II_B \ &:= \int \Big[\textstyle\sum_{\lambda\in\Lambda_n, |\beta_\lambda|\le\kappa c(n)/2} |\psi_\lambda|^2 (\mathbb{E}(I\{|\hat\beta_\lambda| > \kappa c(n)\}|(\hat\beta_\lambda - \beta_\lambda)|^p))^{2/p}\Big]^{p/2} \\ &\le \int \Big[\textstyle\sum_{\lambda\in\Lambda_n, |\beta_\lambda|\le\kappa c(n)/2} |\psi_\lambda|^2 (P(|\beta_\lambda - \hat\beta_\lambda| > \kappa c(n)/2))^{1/p} (\mathbb{E}|(\hat\beta_\lambda - \beta_\lambda)|^{2p})^{1/p}\Big]^{p/2} \\ &\le Cc(n)^{p+\gamma/2} \int \Big[\textstyle\sum_{\lambda\in\Lambda_n, |\beta_\lambda|\le\kappa c(n)/2} |\psi_\lambda|^2 \Big]^{p/2} \\ &\le Cc(n)^{p+\gamma/2} \| \textstyle\sum_{\lambda\in\Lambda_n} \psi_\lambda \|_{L_p}^p \\ &\le Cc(n)^{p+\gamma/2} \mu(\Lambda_n) \le Cc(n)^{\gamma/2} \le Cc(n)^{\alpha p}, \end{aligned}$$

where we have used Schwarz inequality, Lemma 1, assumptions (a) and (b), the property $\mu(\Lambda_n) \le Cc(n)^{-p}$, and the fact that $\gamma \ge \alpha p$.

This concludes the proof of the theorem.

Remark
In Theorem 7, $B_\infty^{\alpha/2}(L_p)$ plays the same role as BS in Theorem 1. When $p \ge q + 2$, it is easy to check that for $f \in L^p$, the property $f \in V_q^*(\mu)$ implies $f \in B_\infty^{\alpha/2}(L_p)$, so that the "basic smoothness" assumption is redundant in (ii).

7 Concluding remarks

Throughout this paper, we have proved in various settings results of the type

$$f \in V_\alpha \ \text{ iff } \ \mathbb{E}(\|\hat f^n - f\|_V^p) \le Cc(n)^{\alpha p}, \tag{38}$$

where V_α is a subspace of V with an intrinsic definition in terms of the atomic decomposition involved in the thresholding procedure.

Although it does not appears explicitly in our computations, there is a more precise dependence between the constant C in the above estimation rate and the norm (or quasi-norm) $\|f\|_{V_\alpha}$: we could actually prove the equivalence

$$\|f\|_{V_\alpha}^p \sim \|f\|_V^p + \sup_{n\ge 0} c(n)^{-\alpha p} \mathbb{E}(\|\hat f^n - f\|_V^p). \tag{39}$$

It is also interesting to note that, by a discrete Hardy inequality, we can obtain an equivalent statement to (38) of the form

$$f \in V_\alpha \ \text{ iff } \ \mathbb{E}(\|\hat f^{2n} - \hat f^n\|_V^p) \le Cc(n)^{-\alpha p}. \tag{40}$$

(provided that the estimator $\hat f_n$ is known to converge to f). In this alternate statement, the rate of decay of $\mathbb{E}(\|\hat f^{2n} - \hat f^n\|_V^p)$ - which can be empirically estimated in contrast to $\mathbb{E}(\|\hat f^n - f\|_V^p)$ - provides a way to estimate the smoothness of the unknown function f.

References

[1] Cohen, A., Daubechies, I. and Vial, P. (1993) *Wavelets and fast wavelet transforms on an interval*, Appl. Comp. Harm. Anal. **1**.

[2] Cohen, A., DeVore, R. and Hochmuth, R. (1997) *Restricted nonlinear approximation*, preprint Laboratoire d'Analyse Numérique, Université Paris VI, to appear in Constructive Apporximation.

[3] Cohen, A. (1998) *Wavelets in numerical analysis*, to appear in the Handbook of Numerical Analysis, vol. VII, P.G. Ciarlet and J.L. Lions, eds., Elsevier, Amsterdam.

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

[5] Donoho, D.L. and Johnstone, I. M (1992) *Minimax Estimation v ia Wavelet shrinkage.* Technical Report, Department of Statistics, Stanford University.

[6] Donoho, D.L. and Johnstone, I. M (1993) *Adapting to unknown smoothness via Wavelet shrinkage.* Technical Report, Department of Statistics, Stanford University.

[7] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) *Wavelet Shrinkage: Asymptopia?* , J. Royal Statistical Soc., **57**, 301–369.

[8] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) *Density estimation by wavelet thresholding.* Annals of stat a completer.

[9] Donoho, D.L., Johnstone. (1996) *Neoclassical minimax problems, thresholding and adaptive function estimation* Bernoulli @(1), 39-62.

[10] DeVore, R., Jawerth, B. and Popov, V. (1991) *Compression of wavelet decompositions*, American J. of Mathematics, **114**, 737–785.

[11] Hall, P., Kerkyacharian, G. and Picard, D. (1996)

[12] Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1992) *Estimation d'une densité de probabilité par méthode d'ondelettes.* Comptes Rendus Acad. Sciences Paris (A) **315** 211-216.

[13] Kerkyacharian, G. and Picard, D. (1992) *Density estimation in Besov Spaces.* Statistics and Probability Letters **13** 15-24

[14] Kerkyacharian, G. and Picard, D. (1992) *Density Estimation by Kernel and Wavelets methods - optimality of Besov spaces.* Submitted. Technical Report, Université de Paris VII.

[15] Kerkyacharian, G., Picard, D and Tribouley, K (1996) *Global thresholding.*

[16] Lepskii, O.V. (1991) *Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates.* Theory Probab. Appl. (36) pp 682-697

[17] Meyer, Y. (1990) *Ondelettes et Opérateurs*, Hermann, Paris.

[18] Nemirovskii, A.S. (1985) *Nonparametric estimation of smooth regression functions.* Izv. Akad. Nauk. SSR Teckhn. Kibernet. **3**, 50-60 (in Russian). J. Comput. Syst. Sci. **23**, 6, 1-11, (1986) (in English).

[19] Nussbaum, M. (1985) *Spline smoothing and asymptotic efficiency in L_2.* Ann. Statist., **13**, 984–997.

[20] Peetre, J. (1975) *New Thoughts on Besov Spaces.* Duke Univ Math. Ser. **1**.

[21] Pinsker, M.S. (1980) *Optimal filtering of square integrable signals in Gaussian white noise*. Problemy Peredatsii Informatsii **16** 52-68 (in Russian); Problems of Information Transmission (1980) 120-133 (in English).

[22] Scott, D. W. (1992) *Multivariate Density Estimation.* Wiley, New York.

[23] Stone, C. (1982) *Optimal global rates of convergence for nonparametric estimates.* Ann. Statist., **10**, 1040-1053.

[24] Temlyakov, V. (1997) *The best m-term approximation and greedy algorithms*, Advances in Comp. Math **8**, 249–265.