Bernoulli **2**(3), 1996, 229–247

L^p adaptive density estimation

GÉRARD KERKYACHARIAN[1]*, DOMINIQUE PICARD[2] and KARINE TRIBOULEY[3]

[1]*URA 1321 et Mathématiques et Informatiques, Université de Picardie, 80039 Amiens, France*
[2]*Mathématiques, URA 1321, Université de Paris VII, 2 place Jussieu, 75251 Paris Cedex 05, France*
[3]*Université Paris Sud, URA 743, Bâtiment 425, 91405 Orsay Cedex, France*

We provide global adaptive wavelet-type density estimates. Our procedures illustrate the refinement which can be obtained by replacing the Fourier basis by the wavelet basis in estimation methods. The main argument consists in observing that the estimated total energy of the details of a specified level j will be smaller or greater than some known threshold if precisely j is above or below the theoretical optimal level calculated with the a priori knowledge of the regularity of the density. This balancing effect leads directly to an adaptation procedure, and some natural extensions. We investigate the minimax properties of these procedures and explain their evolution for different global error measures.

Keywords: adaptive estimation; Besov spaces; density estimation; minimax estimation; U-estimate; wavelet orthonormal bases

1. Introduction

This paper investigates the problem of global adaptation using particular threshold wavelet-type estimates in the context of probability density estimation, that is, the problem of estimating a density function on the basis of $X_1, \ldots, X_n$ independent and identically distributed drawn from f.

Various methods can be used in nonparametric estimation, such as kernel estimation, orthogonal projection estimation, smoothing splines, wavelets. An overview of traditional methods and of a part of the vast literature on density estimation is given in Devroye (1985), Silverman (1986) and Scott (1992). The performances of all these procedures depend strongly on the choice of a smoothing parameter or bandwidth. This choice is in fact by no means an easy task. Different approaches have been considered, generally corresponding to some optimal solution of some well-posed problem (see, for example, Bretagnolle and Carol-Huber 1979; Pinsker 1980; Efromovich and Pinsker 1982; Ibragimov and Has'minskii 1982; Stone 1982; Birgé 1983; Nussbaum 1985). As an example, if the regularity class of the estimated function is assumed to be known, then it is possible to choose the bandwidth so that the estimate attains the minimax rate. Of course, from a practical point of view, this is not entirely satisfactory since it requires some extra knowledge. Various attempts have also been investigated to reduce this knowledge.

* To whom correspondence should be addressed.

1350-7265

Among these, the recent appearance of explicit orthonormal bases based on multiresolution analysis has given different opportunities to solve this problem. Indeed, unlike traditional Fourier bases, wavelet bases, since they have localization properties in space as well as in frequency, enable expansions of a function into coefficients which are reliable indicators of its regularity.

If we now focus on the problem where we do not know the regularity of the function, one possible approach is to start from the evaluation of the risk of a procedure. In almost every case, this risk can be decomposed by means of the well-known formula $\mathrm{E}\|\hat{f}-f\|_2^2 \sim C_1(nh)^{-1}+C_2h^{2s}$ into a sum of a stochastic term whose behaviour is not affected by the regularity and a bias term which depends strongly on this parameter s. The optimal choice for the bandwidth consists in balancing these two contributions. See for instance Kerkyacharian and Picard (1992), where it can be found as well as an introduction to Besov spaces in this framework.

However, some nice phenomena appear in the wavelet framework. Let us suppose that the wavelet basis is derived from $\phi_{jk}(x)=2^{j/2}\phi(2^jx-k)$, $k\in\mathbb{Z}$, and $\psi_{jk}(x)=2^{j/2}\psi(2^jx-k)$, $k\in\mathbb{Z}$, $j\in\mathbb{Z}$, where ϕ and ψ are the scaling function and the mother wavelet, respectively. The probability density has formal expansion

$$f(x)=\sum_k \alpha_k\phi_{0k}(x)+\sum_{j\geq 0}\sum_k \beta_{jk}\psi_{jk}(x). \tag{1}$$

In this context, the bandwidth selection corresponds to choosing the level parameter j_0 at which to stop the sum in (1). Of course, when one wants to stop at some level j_0-1, a natural investigation consists in looking at the next layer of 'details'

$$\sum \beta_{j_0k}\psi_{j_0k}(x). \tag{2}$$

However, the striking fact is that the 'energy' of (2) (the p-power of the L_P-norm) is of order $2^{-j(s+1/2-1/p)p}$, whereas this quantity can be estimated (roughly) with an error less than $2^jn^{-p/2}$. A consequence is that the level at which the error becomes more important than the estimated quantity is of the same order as the optimal 'bandwidth' $2^{j_0}\sim n^{1/(1+2s)}$. This balancing effect leads directly to a strategy of adaptation by thresholding, and to some natural variations around this strategy. The aim of this paper is to investigate the properties of these procedures.

Our results are the following. We take as a global error measure for estimating the whole density the L_p error

$$R_n(\hat{f},f)=\mathrm{E}\|\hat{f}_n-f\|_p^p=\mathrm{E}\int|\hat{f}_n-f|^p\,\mathrm{d}x.$$

We consider the case $\infty>p\geq 2$. We look at the worst performance over a variety of functional spaces:

$$R_n(\hat{f};\mathcal{F})=\sup_{f\in\mathcal{F}}\mathrm{E}\|\hat{f}_n-f\|_p^p,$$

where $\mathcal{F}$ will be a subset of density functions, compactly supported with fixed (but unknown) support bounded in the norm of the Besov space B_{spq}. Let

$$\mathcal{F}_{spq}(M,B)=\{f \text{ density, supp }(f)\subset[-B,B],\ \|f\|_{spq}\leq M\}.$$

A density estimate $f^\star$ will be called *adaptive* for a class $\{\mathcal{C}(\alpha), \alpha \in \mathcal{A}\}$ if there exists some constant C such that

$$\forall \alpha \in \mathcal{A},\ R_n(f^\star; \mathcal{C}(\alpha)) \leq C \inf_{\hat{f}} R_n(\hat{f}; \mathcal{C}(\alpha)).$$

We will show that our procedures are adaptive for the class

$$\mathcal{F}_{spq}(M, B), \qquad 1/p < s < r+1,\ 1 \leq q \leq \infty,\ 0 < M < \infty,\ 0 < B < +\infty.$$

The problem of adaptivity has been widely investigated in the recent statistical literature: among other papers, the following have especially inspired the spirit of our work: the problem of adaptive estimation in the L_2-norm for the class $\{\mathcal{F}_{s22}(M, B), s, M\}$ was stated in Stone (1982) and solved by Härdle and Marron (1985) in the nonparametric regression scheme, and by Efromovich (1985) in the density problem; in the L_p-norm for the class $\{\mathcal{F}_{s\infty\infty}(M, B), s, M\}$ for the white noise model by Lepskii (1990; 1991); in the L_π-norm for the class $\{\mathcal{F}_{spq}(M, B), s, p, q, B\}$ by Donoho and Johnstone (1993; 1995) and Donoho *et al.* (1995a; 1996b).

The comparison between our procedures and those investigated in Efromovich (1985) provides an explicit illustration of the refinement that can be obtained by replacing the Fourier basis by the wavelet basis. Indeed, if the estimates are close enough, the wavelet tools give at the same time a better understanding of the packets T_j of Fourier coefficients, and the opportunity of solving the problem for norms different from the L_2-norm.

As will be explained later on, our first estimate is quite close to that obtained by Lepskii's procedure; a main advantage is its extreme computational simplicity.

The main difference between our method and other adaptive wavelet procedures is essentially its global aspect: instead of thresholding each coefficient, we consider the global level j. This different point of view has advantages as well as drawbacks: the classes of adaptation in both cases are essentially different. The local adaptation allows us to solve the difficult problem of finding one single procedure achieving nearly optimal performance over a variety of global error measures and over a variety of function spaces, but provides an extra logarithmic factor, and requires knowledge of the radius M of the balls; the global procedure, on the other hand, can be performed without knowledge of M and enjoys exact convergence rates. A practical aspect of this comparison seems also to be that, like cross-validation procedures, this one does a good job for a reasonable amount of data.

2. Main results

2.1. SUMMARY

We begin in Section 2.2 by describing elements of the basic theory of wavelet methods and Besov spaces. In Section 2.3 we introduce the wavelet-based estimates, present the different estimation procedures and discuss the importance of reducing the bias by using U-estimates. Our main results are given there. Section 2.4 gives a summary of the essential material used in the proofs which are collected in Section 3.

2.2. WAVELETS AND BESOV SPACES

The key ingredients of our analysis are described in much greater detail in Peetre (1976), Bergh and Löfström (1976), Meyer (1990), Daubechies (1992) and Triebel (1992).

We first review the very basic features of the multiresolution analysis of Meyer (1990). One can construct a real function ϕ (the scaling function) such that:

(1) the sequence $\{\phi_{0,k} = \phi(. - k) | k \in \mathbb{Z}\}$ is an orthonormal family of $L_2(R)$. Let us call V_0 the subspace spanned by this sequence

(2) if V_j denotes the subspace spanned by the sequence $\{\phi_{j,k} = 2^{j/2}\phi(2^j. - k) | k \in \mathbb{Z}\}$, then $\{V_j\}_{j\in\mathbb{Z}}$ is an increasing sequence of nested spaces such that $\cap_{j\in\mathbb{Z}} V_j = \{0\}$ and, if $\int \phi = 1$, $\overline{\cup_{j\in\mathbb{Z}} V_j} = L_2$.

It is possible to require in addition that ϕ is of class $\mathcal{C}^r$ with a compact support (Daubechies wavelets). In the sequel, we will work with such a scaling function ϕ. We define the space W_j by the following: $V_{j+1} = V_j \oplus W_j$. There also exists a function ψ (the wavelet) such that:

(1) ψ is of class $\mathcal{C}^r$ with a compact support;
(2) $\{\psi_{0,k} = \psi(. - k) | k \in \mathbb{Z}\}$ is an orthonormal basis of W_0;
(3) $\{\psi_{j,k} = 2^{j/2}\psi(2^j. - k) | k \in \mathbb{Z}, j \in \mathbb{Z}\}$ is an orthonormal basis of L_2.

For $j_0 \in \mathbb{Z}$, the following decomposition is also true:

$$\forall f \in L_2, \qquad f = \sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\phi_{j_0,k} + \sum_{j\geq j_0}\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k},$$

where

$$\alpha_{j,k} = \int f(x)\phi_{j,k}(x)\,\mathrm{d}x, \qquad \beta_{j,k} = \int f(x)\psi_{j,k}(x)\,\mathrm{d}x. \tag{3}$$

The following lemma will be of some importance. It provides explicit expansions of the L_p-norms of the details $\sum_{k\in\mathbb{Z}} \beta_{j,k}\psi_{j,k}$ at level j and of the low-frequency part $\sum_{k\in\mathbb{Z}} \alpha_{j_0,k}\phi_{j_0,k}$ in terms of the wavelet coefficients:

Lemma 1 (Meyer). *Let g be either ϕ or ψ with the conditions above; let $\theta(x) = \theta_g(x) = \sum_{k\in\mathbb{Z}} |g(x-k)|$, and $\|\theta\|_p = (\int_0^1 |\theta(x)|^p\,\mathrm{d}x)^{1/p}$. Let $f(x) = \sum_{k\in\mathbb{Z}} \lambda_k 2^{j/2} g(2^j x - k)$. If $1 \leq p \leq \infty$ and p_1 satisfies $1/p + 1/p_1 = 1$, then*

$$\frac{1}{\|\theta\|_1^{1/p_1}\|\theta\|_\infty^{1/p}} 2^{j(1/2-1/p)}\|\lambda\|_{l_p} \leq \|f\|_{L_p} \leq \|\theta\|_p 2^{j(1/2-1/p)}\|\lambda\|_{l_p}. \tag{4}$$

Let us now define Besov spaces in terms of wavelet coefficients. For the classical definitions, in terms for example of the modulus of continuity, we refer to Peetre (1976), Bergh and Löfström (1976) and Triebel (1992). The following definition is especially convenient for statistical purposes as it gives a description of the space in terms of a sequence of coefficients, see Meyer (1990). Besov spaces depend on three parameters $s > 0$,

$1 \le p \le +\infty$ and $1 \le q \le +\infty$ and are denoted B_{spq}. Let s be smaller than r (see (2)), let ϕ and ψ be subject to the conditions above, and let α_{jk}, β_{jk} be defined as in (3). We say that $f \in B_{spq}$ if and only if

$$\| f \|_{spq} = \|\alpha_{0.}\|_p + \left(\sum_{j \ge 0} (2^{j(s+1/2-1/p)} \|\beta_{j.}\|_p)^q \right)^{1/q} < +\infty \tag{5}$$

(we have set $\|u_{j.}\|_p = (\sum_k |u_{jk}|^p)^{1/p}$), the necessary condition is true up to $s < r+1$.

Because of classical results on $\|u_.\|_p$ the following inequalities are true and will be essential later in this paper:

$$\forall \ell \le p \qquad \sum_{k \in E} |\beta_{jk}|^\ell \le \left(\sum_{k \in E} |\beta_{jk}|^p \right)^{\ell/p} (\text{card } E)^{(1-\ell/p)}, \tag{6}$$

$$\forall \ell \ge p \qquad \sum_{k \in E} |\beta_{jk}|^\ell \le \left(\sum_{k \in E} |\beta_{jk}|^p \right)^{\ell/p}. \tag{7}$$

Let us now denote

$$\Theta_j = \sum_k |\beta_{jk}|^p. \tag{8}$$

In this sum, only a finite number of β_{jk} are non-zero as soon as f is compacted supported. This number is less than $2^j AB^{-1}$ where $2B$, $2A$ are the respective length of the supports of f and ψ.

2.3. ESTIMATES AND RESULTS

Now let $(X_1, \dots, X_n)$ be n independent and identically distributed variables according to a distribution P. We assume that P is absolutely continuous with respect to Lebesgue measure: let $f : \mathbb{R} \to \mathbb{R}^+$ be the unknown density of P. Furthermore, we suppose that

$$f \in F_{spq}(M, B) = \left\{ f \in B_{spq}, \int f = 1,\, f \ge 0,\, \text{supp}\,(f) \subset [-B, B],\ \|f\|_{spq} \le M \right\},$$

where $s - 1/p > 0$. Let us consider the following weighted linear estimate:

$$\hat{f} = \sum_k \hat{\alpha}_{j_0 k} \phi_{j_0 k} + \sum_{j=j_0}^{j_1} \hat{\eta}_j \sum_k \hat{\beta}_{jk} \psi_{jk},$$

where

$$\hat{\alpha}_{j_0 k} = \frac{1}{n} \sum_{i=1}^n \phi_{j_0 k}(X_i),$$

$$\hat{\beta}_{jk} = \frac{1}{n} \sum_{i=1}^n \psi_{jk}(X_i).$$

The first sum is an estimate of the low-frequency part of f. The level $j_0 = 0$ can be chosen arbitrarily here so that the 'variance term' will not significantly contribute to the error. In the same spirit, the level $j_1(n) = \log_2(n)$ can also be chosen so that the bias term will never contribute for $1/p < s < r + 1$. The important task now is to determine $\hat{\eta}_j$.

Efromovich (1985) has demonstrated the virtue, in the context of Fourier series and L_2 error, of choosing $\hat{\eta}_j$ as a 'soft thresholding' based on the two following principles:

(1) Let us denote by j_s, the optimal 'bandwidth' selection: $2^{j_s} = n^{1/(1+2s)}$ if the regularity s was known. We have the following inequalities:

$$f \in B_{spq} \Rightarrow \Theta_j \leq C 2^{-j(s+1/2-1/p)p},$$

$$j \leq j_s \Leftrightarrow 2^{-j(s+1/2-1/p)p} \geq \frac{2^j}{n^{p/2}}.$$

Thus, if $\{\Theta_j \geq 2^j/n^{p/2}\}$ this means that $j \leq j_s$. Conversely, if $j \leq j_s$ of course it can occur that $\{\Theta_j \leq 2^j/n^{p/2}\}$, but in this case it is actually interesting to threshold the level j since it can only give a better performance than the linear estimate – which would give a global error of order $2^j/n^{p/2}$ (see (4)). Then it turns out to be reasonable that a level j should be kept if and only if

$$\Theta_j \geq \frac{2^j}{n^{p/2}}.$$

(2) Θ_j has to be estimated carefully: a natural candidate could be $\sum_k |\hat{\beta}_{jk}|^p$, as in Efromovich (1985). Unfortunately, it happens that this estimate has too large a bias especially for large values of j and $p > 2$. The effect is an undersmoothing (i.e. choosing too small a bandwidth). The solution to this drawback is provided by choosing the associated U-estimate in the case where p is an even integer and to interpolate in the other cases. For $p \in 2\mathbb{N}$

$$\hat{\Theta}_j(p) = (C_n^p)^{-1} \sum_{(i_1, \ldots, i_p) \in S_p} \sum_k \psi_{jk}(X_{i_1}) \ldots \psi_{jk}(X_{i_p}), \tag{9}$$

where S_p is the set of p-dimensional vectors of $\{1, \ldots, n\}^p$ such that all the coordinates are different. For $p = \alpha p_1 + (1-\alpha)p_2$, where $\alpha \in]\hat{\Theta}, 1[$, $p_1, p_2 \in 2\mathbb{N}$:

$$\hat{\Theta}_j = (\hat{\Theta}_j(p_1))^{\alpha}(\hat{\Theta}_j(p_2))^{1-\alpha}. \tag{10}$$

Many kinds of thresholding are available. We present the results in the two following different settings:

$$\hat{\eta}_j^{\mathrm{H}} = 1_{\{\hat{\Theta})_j \geq 2^j/n^{p/2}\}} \qquad \text{(hard thresholding)} \tag{11}$$

$$\hat{\eta}_j^{\mathrm{S}} = \frac{\hat{\Theta}_j - 2^j/n^{p/2}}{\hat{\Theta}_j} 1_{\{\hat{\Theta})_j \geq 2^j/n^{p/2}\}} \qquad \text{(soft thresholding).} \tag{12}$$

Soft thresholding provides a generalization in L_p of Efromovich's (1985) procedure, whereas hard thresholding is very close to Lepskii (1990). Indeed, because of Lemma 1, $2^{j(p/2-1)}\Theta_j$ is, up to some constant, $\|\mathrm{Proj}_{W_{j+1}} f - \mathrm{Proj}_{W_j} f\|_p^p$.

Theorem 1. *Let $p \geq 2$. Let*

$$\hat{f} = \sum_k \hat{\alpha}_{j_0k}\phi_{j_0k} + \sum_{j=j_0}^{j_1} \hat{\eta}_j \sum_k \hat{\beta}_{jk}\psi_{jk},$$

where $j_0 = 0$, $j_1 = \log_2(n)$ and $\hat{\eta}_j$ is either $\hat{\eta}_j^{\mathrm{H}}$ or $\hat{\eta}_j^{\mathrm{S}}$. Then, for $s \in]1/p, r+1[$, $q \in [1, +\infty]$, there exists a constant C such that

$$\sup_{f \in F_{spq}(M,B)} \mathrm{E}_f \, \| \hat{f}_{j_0,j_1} - f \|_p^p \leq Cn^{-sp/(1+2s)},$$

i.e. $\hat{f}$ is adaptive in the class $\{\mathcal{F}_{spq}(M,B), s, q, M, B\}$.

Another way to understand this result is as follows: if, for s known, we denote by $\hat{f}_L = \sum_k \hat{\alpha}_{j_0k}\phi_{j_0k} + \sum_{j=j_0}^{j_s} \sum_k \hat{\beta}_{jk}\psi_{jk}$ the best linear estimator, we have:

$$\sup_{f \in F_{spq}(M,B)} \mathrm{E}_f \, \| \hat{f}_{j_0,j_1} - \hat{f}_L \|_p^p \leq Cn^{-sp/(1+2s)}.$$

2.4. BASIC INGREDIENTS OF THE PROOFS

We shall first give two lemmas describing the behaviour of the moments of the estimate $\hat{\Theta}_j$. As can be seen later on in the proof, the first one will essentially be used in proving that when the statistic $\hat{\Theta}_j$ is large, the associated level j has a small enough 'energy' to be omitted with high probability. The second one concerns the centred moment and will be useful in the opposite situation when the problem is to prove that if the statistic is too small to threshold the level j, then it is significant with high probability. Lemma 4 establishes the behaviour of the linear estimates of $\mathrm{Proj}_{W_j} f$ or $\mathrm{Proj}_{V_j} f$. Its proof uses the Rosenthal inequality. It will be given in the Appendix.

Let us begin with some notation:

$$\mathrm{E}(\hat{\Theta}_j)^m = (C_n^p)^{-m} \sum_{k^1,\ldots,k^m} \sum_{(i_1^1,\ldots,i_p^1) \in S_p} \cdots \sum_{(i_1^m,\ldots,i_p^m) \in S_p} \mathrm{E} \prod_{l=1}^{p} \prod_{h=1}^{m} \psi_{jk^h}(X_{i_l^h}).$$

Let us investigate, for the sake of simplicity, the case where the wavelet ψ yields to the Haar basis and has support $[0, 1]$ (of course, it is not the case if the wavelet has regularity greater than 1, but the reader may be convinced that the argument is not very different when the support is of finite size). Let us consider one term in the sum above. A useful remark should be made immediately: as soon as two indices i_r^s, $i_{r'}^{s'}$ are equal (for two different indices s and s'), because of the property of the support of ψ, the two indices k_s and $k_{s'}$ have to be equal (otherwise the term is zero). (Here is the main difference with a ψ of arbitrary compact support, where there is more than one non-zero term when $i_r^s = i_{r'}^{s'}$.) Let λ_i denote the number of times a product $\mathrm{E}(\psi_{jk}(X_h))^i$ of size i appears in the term of the sum under consideration. Of course $mp = \sum i\lambda_i$. Associated with $(i_1^j, \dots, i_p^j) \in S_p$, let $\alpha_j = \{i_1^j, \dots, i_p^j\}$ and let us introduce the following equivalence relationship: $\alpha_i \sim \alpha_j$ if $1_{\alpha_i} 1_{\alpha_j} \neq 0$. The equivalence classes are $R_1, \dots, R_r$. For each class R_s, we define $\lambda_i^s = \operatorname{card}\{(\sum_{\alpha_j \in R_s} 1_{\alpha_j})^{-1}(\{i\})\}$. So we have $\lambda_i = \sum_{s=1}^r \lambda_i^s$, $p(\operatorname{card} R_s) = \sum_{i=1}^m i\lambda_i^s$. Now let κ denote the number of classes R_s such that $\lambda_1^s \leq p-1$. We have the following lemma:

Lemma 2. *For all $m \in 2\mathbb{N}$ and with $\lambda_1, \dots, \lambda_m$, κ as defined previously, we have:*

$$\mathrm{E}(\hat{\Theta}_j)^m \leq C \sum_{\lambda_1, \dots, \lambda_m} \sum_{\kappa} \frac{2^{j(mp/2 - \sum \lambda_i + \lambda_1/2)}}{n^{mp - \sum \lambda_i}} 2^{-j(s+1/2-1/p)\lambda_1} 2^{j\kappa}. \tag{13}$$

Moreover,

$$\frac{mp}{2} - \sum \lambda_i + \frac{\lambda_1}{2} \geq 0, \tag{14}$$

$$\frac{\lambda_1}{p} + \kappa - m \leq \frac{-\kappa}{p}. \tag{15}$$

$$\text{If} \quad \frac{mp}{2} - \sum \lambda_i \geq -\frac{mp}{4} \quad \text{then} \quad m - \kappa \geq \frac{m}{8}. \tag{16}$$

Lemma 3. *For all $m \in 2\mathbb{N}$, we have, if $j \geq 0$, $2^j \leq n$,*

$$\mathrm{E}(\hat{\Theta}_j - \Theta_j)^m \leq C \sum_{l=1}^{p} (\Theta_j)^{m - lm/p} \left(\frac{2^j}{n^{p/2}}\right)^{lm/p}.$$

Lemma 4. *Let $g : \mathbb{R} \to \mathbb{R}$ have compact support (supp $g \subset [-A, +A]$) such that $\|g\|_\infty < +\infty$. Let*

$$\gamma_{jk} = \int f(x) 2^{j/2} g(2^j x - k)\, \mathrm{d}x,$$

$$\hat{\gamma}_{jk} = \frac{1}{n} \sum_{i=1}^{n} 2^{j/2} g(2^j X_i - k).$$

If $m \geq 2$, f has compact support (supp $f \subset [-B, +B]$) and is such that $\|f\|_\infty < +\infty$; furthermore, if $2^j \leq n$, then

$$\sum_{k \in \mathbb{Z}} (\mathrm{E}|\hat{\gamma}_{jk} - \gamma_{jk}|^{mr})^{1/r} \leq C \frac{2^j}{n^{m/2}},$$

where C is a constant depending on $A, B, \|g\|_\infty, \|f\|_\infty$.

Lemma 5. *For $p \geq 1$*

$$\left\| \sum_{j_0}^{j_1} \sum_k \beta_{jk} \psi_{jk} \right\|_p^p \leq \begin{cases} 2^{j_1 \epsilon p/2} \sum_{j_0}^{j_1} 2^{-j\epsilon p/2} \| \sum_k \beta_{jk} \psi_{jk} \|_p^p & \text{if } \epsilon > 0, \\ 2^{j_0 \epsilon p/2} \sum_{j_0}^{j_1} 2^{-j\epsilon p/2} \| \sum_k \beta_{jk} \psi_{jk} \|_p^p & \text{if } \epsilon < 0. \end{cases}$$

This last result is a simple consequence of the inclusion $B_{0,p,p\wedge 2} \subset L^p$ for $p \geq 1$ (see Triebel 1992) and of the Hölder inequality.

3. Proofs

3.1. PROOF OF LEMMA 2

Inequality (14) is easily obtained by the following remark: $mp = \sum i\lambda_i \geq \lambda_1 + 2\sum_{i>1}\lambda_i$.

Turning to inequality (15), if κ represents the number of classes R_s such that $\lambda_1^s \leq p-1$, then for such a class R_s we have $\mathrm{card}(R_s) \geq 2$. Moreover,

$$p+1 \leq (\mathrm{card}(R_s)-1)p+1 = \mathrm{card}(R_s)p-(p-1) \leq \mathrm{card}(R_s)p-\lambda_1^s.$$

Hence

$$\kappa(p+1) \leq \sum_{s/\lambda_1^s \leq p-1} \sum_{i=2}^{m} i\lambda_i^s \leq mp-\lambda_1.$$

As for inequality (16), it is enough to prove that if $\sum \lambda_i \leq 3mp/4$ then $m-m_0 \geq m/8$, where m_0 is the number of equivalence classes. Put $m_0 = \nu m$. If $\nu \leq \frac{1}{2}$, then the result is obvious. If $\nu > \frac{1}{2}$, then it is clear that $(2\nu-1)m$ equivalence classes are reduced to one element, and for those classes, $\lambda_1^s = p$. Then $\lambda_1 \geq (2\nu-1)mp$. But from $\lambda_1 \leq 3mp/4$, we obtain $(2\nu-1)mp \leq 3mp/4$.

Finally, inequality (13) is obtained just by counting the number of times that a fixed configuration $\lambda_1, \ldots, \lambda_{mp}$ occurs and using the definition of Besov spaces, (6) and (7):

$$\text{for } l>1 \qquad \sum_k \mathrm{E}(|\psi_{jk}(X_i)|^l) \leq C2^{j(l/2-1)},$$

and

$$\sum_k (\mathrm{E}(\psi_{jk}(X_i)))^l \leq C2^{j[-(s+1/2-1/p)l+(1-(\min(l,p))/p)]},$$

which concludes the proof of lemma 2.

3.2. PROOF OF LEMMA 3

Let us denote $\Delta x_i = x_i - \beta$. Then:

$$\prod_1^p x_i - \beta^p = \sum_{l=1}^{p} \beta^{p-l} \sum_{1\leq j_1<\ldots<j_l\leq p} \prod_{i=1}^{l} \Delta x_{j_i}.$$

If α is a subset of length $p = |\alpha|$ of $\{1,\ldots,n\}$, let us denote $\psi_{jk}^{\otimes p}(X_\alpha) = \prod_{i\in\alpha}\psi_{jk}(X_i)$. Moreover, $\hat{\Theta}_j - \Theta_j = \sum_k (C_n^p)^{-1} \sum_{\alpha,|\alpha|=p}(\psi_{jk}^{\otimes p}(X_\alpha) - \beta_{jk}^p)$. Hence, using the previous formula, we obtain:

$$\hat{\Theta}_j - \Theta_j = \sum_k \frac{1}{C_n^p} \sum_{\alpha, |\alpha| = p} \sum_{l=1}^{p} \beta_{jk}^{p-l} \sum_{\gamma \subset \alpha, |\gamma| = l} \Delta\psi_{jk}^{\otimes l}(X_\gamma),$$

where $\Delta\psi_{jk}() = \psi_{jk}() - \beta_{jk}$. Now, reversing the order of integration, we obtain:

$$\hat{\Theta}_j - \Theta_j = \sum_k \sum_{l=1}^{p} \frac{C_{n-l}^{p-l}}{C_n^p} \beta_{jk}^{p-l} \sum_{\gamma, |\gamma| = l} \Delta\psi_{jk}^{\otimes l}(X_\gamma).$$

Now let m be an ever integer. Then

$$\mathrm{E}(\hat{\Theta}_j - \Theta_j)^m \le p^{m-1} \sum_{l=1}^{p} \left(\frac{C_{n-l}^{p-l}}{C_n^p} \right)^m \sum_{k_1, \ldots, k_m} (|\beta_{jk_1} \ldots \beta_{jk_m}|)^{p-l} \sum_{\gamma_1, \ldots, \gamma_m, |\gamma_i| = l} \mathrm{E} \left| \prod_i \Delta\psi_{jk}^{\otimes l}(X_{\gamma_i}) \right|.$$

If we denote

$$Q_{m,l} = \left| \sum_{\gamma_1, \ldots, \gamma_m, |\gamma_i| = l} \mathrm{E}\{\Delta\psi_{jk}^{\otimes l}(X_{\gamma_1}) \ldots \Delta\psi_{jk}^{\otimes l}(X_{\gamma_l})\} \right|, \tag{17}$$

we shall prove that

$$Q_{m,l} \le Cn^{ml/2}. \tag{18}$$

Indeed, let us look at the set of subsets of integers $\{\gamma_1, \ldots, \gamma_m\}$, if it is not the case that $\sum 1_{\gamma_i} \ge 2 1_{\cup \gamma_i}$, then $\mathrm{E}\{\Delta\psi_{jk}^{\otimes l}(X_{\gamma_1}) \ldots \Delta\psi_{jk}^{\otimes l}(X_{\gamma_l})\} = 0$. Hence, only the family of subsets $\{\gamma_1, \ldots, \gamma_m\}$ verifying $\sum 1_{\gamma_i} = \sum_{j=2}^{m} j 1_{A_j}$, where A_j are disjoint sets of integers of size (say) λ_j, has to be taken into account. We then have $ml = \sum_{j=2}^{m} j\lambda_j$. For such a configuration, we have, as in the proof of Lemma 2:

$$|\mathrm{E}\{\Delta\psi_{jk}^{\otimes l}(X_{\gamma_1}) \ldots \Delta\psi_{jk}^{\otimes l}(X_{\gamma_l})\}| \le C 2^{j(\sum_{i=2}^{m} (i-2)\lambda_i)/2}.$$

The number of such terms may be bounded by $Cn^{(\Sigma \lambda_i)}$, but, since $\lambda_1 = 0$, certainly $ml/2 - \sum \lambda_i \ge 0$. So due to the fact that $2^j \le n$, we get $n^{(\Sigma \lambda_i)} 2^{j(ml/2 - \Sigma \lambda_i)} \le n^{ml/2}$. The result follows using inequality (6).

3.3. PROOF OF THEOREM 1 IN THE CASE OF HARD THRESHOLDING

In what follows C will denote a positive constant which may change from place to place. Let us first investigate the basic case of hard thresholding when $p \in 2\mathbb{N}$. Let $f \in F_{spq}(M, B)$ and let j_s be such that $2^{j_s} = n^{1/(1+2s)}$. As

$$\mathrm{E}\| \hat{f} - f \|_p^p \le 3^{p-1} \left[\mathrm{E} \left\| \sum_k (\hat{\alpha}_{j_0k} - \alpha_{j_0k}) \phi_{j_0k} \right\|_p^p + \mathrm{E} \left\| \sum_{j_0 \le j \le j_1} \sum_k (\hat{\eta}_j \hat{\beta}_{jk} - \beta_{jk}) \psi_{jk} \right\|_p^p + \left\| \sum_{j_1 \le j} \sum_k \beta_{jk} \psi_{jk} \right\|_p^p \right],$$

we are going to prove that each of the three terms on the right-hand side is bounded by $Cn^{-sp/(1+2s)}$. In the following inequalities, we have emboldened certain terms which give the required rate of convergence.

For the linear stochastic term, using Lemmas 1 and 4, we have:

$$\mathrm{E}\left\|\sum_k(\hat{\alpha}_{j_0k}-\alpha_{j_0k})\phi_{j_0k}\right\|_p^p \leq C_p\,\mathrm{E}\left[2^{j_0(p/2-1)}\sum_k|\hat{\alpha}_{j_0k}-\alpha_{j_0k}|^p\right]$$

$$\leq C_p\,\mathbf{2^{j_0(p/2-1)}}\,\frac{2^{j_0}}{\boldsymbol{n^{p/2}}} \leq C_p n^{-sp/(1+2s)}.$$

For the bias term we apply the definition of Besov spaces in terms of wavelet coefficients:

$$\left\|\sum_{j_1\leq j}\sum_k\beta_{jk}\psi_{jk}\right\|_p^p \leq \sum_{j_1\leq j}\left\|\sum_k\beta_{jk}\psi_{jk}\right\|_p$$

$$\leq \sum_{j_1\leq j}\mathbf{2^{-js}}\varepsilon_j$$

$$\leq 2^{-j_1 s}\left(\sum_{j_1\leq j}\varepsilon_j^q\right)^{1/q} \leq Mn^{-s/(1+2s)}.$$

Finally, we decompose the nonlinear stochastic term into four terms:

$$\mathrm{E}\left\|\sum_{j_0\leq j\leq j_1}\sum_k(\hat{\eta}_j\beta_{jk}-\beta_{jk})\psi_{jk}\right\|_p^p$$

$$\leq 4^{p-1}\left[\mathrm{E}\left\|\sum_{j_0\leq j\leq j_s}\sum_k(\hat{\eta}_j\hat{\beta}_{jk}-\beta_{jk})\psi_{jk}1_{\{\hat{\Theta}_j\leq 2^j/n^{p/2}\}}\right\|_p^p\right.$$

$$+\mathrm{E}\left\|\sum_{j_0\leq j\leq j_1}\sum_k(\hat{\eta}_j\hat{\beta}_{jk}-\beta_{jk})\psi_{jk}1_{\{\hat{\Theta}_j\leq 2^j/n^{p/2}\}}\right\|_p^p$$

$$+\mathrm{E}\left\|\sum_{j_s\leq j\leq j_s}\sum_k(\hat{\eta}_j\hat{\beta}_{jk}-\beta_{jk})\psi_{jk}1_{\{\hat{\Theta}_j\leq 2^j/n^{p/2}\}}\right\|_p^p$$

$$\left.+\mathrm{E}\left\|\sum_{j_s\leq j\leq j_1}\sum_k(\hat{\eta}_j\hat{\beta}_{jk}-\beta_{jk})\psi_{jk}1_{\{\hat{\Theta}_j\leq 2^j/n^{p/2}\}}\right\|_p^p\right]$$

$$\leq T_1+T_2+T_3+T_4.$$

Using Lemma 5, for some $\epsilon>0$, and Lemma 4, we bound the first term:

$$T_1 \leq C_p 2^{j_s\epsilon p/2}\sum_{j_0\leq j\leq j_s}2^{j(p/2-1)}2^{-j\epsilon p/2}\sum_k \mathrm{E}|\hat{\beta}_{jk}-\beta_{jk}|^p 1_{\{\hat{\Theta}_j\geq 2^j/n^{p/2}\}} \tag{19}$$

$$\leq C_p 2^{j_s \epsilon p/2} \sum_{j_0 \leq j \leq j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \frac{2^j}{n^{p/2}} \leq C_p n^{-sp/(1+2s)}. \tag{20}$$

In the same way as for the bias term, we can derive the same upper bound for T_4:

$$T_4 \leq \left(\sum_{j_s \leq j \leq j_1} \left\| \sum_k \beta_{jk} \psi_{jk} \right\|_p \right)^p \leq M^p 2^{-j_s sp} \leq C n^{-sp/(1+2s)}.$$

Now the thresholding comes into play. To study T_3, we will use Lemma 2; as for T_4, we will use Lemma 3. In both cases, the main point will be to observe that $\hat{\Theta}_j$ is not far from Θ_j, but in the first case there is no need to consider the centred moment since there Θ_j is small.

Let us now study T_3. Using Lemma 5, for some $\epsilon < 0$, the Chebyshev inequality for some $m \in 2\mathbb{N}$, Lemma 4 and the Hölder inequality for $1/m' + 1/m'' = 1$, we obtain the following chain of inequalities:

$$\begin{aligned} T_3 &\leq C 2^{j_s \epsilon p/2} \sum_{j_s \leq j \leq j_1} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \sum_k \mathrm{E}[|\hat{\beta}_{jk} - \beta_{jk}|^p 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}}] \\ &\leq C 2^{j_s \epsilon p/2} \sum_{j_s \leq j \leq j_1} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \sum_k \mathrm{E}|\hat{\beta}_{jk} - \beta_{jk}|^{pm'})^{1/m'} P\left(\hat{\Theta}_j \geq \frac{2^j}{n^{p/2}}\right)^{1/m''} \\ &\leq C 2^{j_s \epsilon p/2} \sum_{j_s \leq j \leq j_1} \mathbf{2}^{\boldsymbol{j(p/2-1)}} 2^{-j\epsilon p/2} \frac{2^j}{\boldsymbol{n^{p/2}}} \left[\mathrm{E}(\hat{\Theta}_j)^m \left(\frac{2^j}{n^{p/2}}\right)^{-m} \right]^{1/m''}. \end{aligned} \tag{21}$$

We will now apply Lemma 2. Let us denote:

$$\Gamma = \left\{ \lambda = (\lambda_1, \ldots, \lambda_{mp}, \kappa) \Big/ \sum_{1 \leq i \leq mp} i\lambda_i = mp \right\} = \Gamma_1 \cup \Gamma_2 \cup \Gamma_3$$

where

$$\begin{aligned} \Gamma_1 &= \left\{ \lambda \in \Gamma \Big/ \frac{mp}{2} - \sum \lambda_i \geq 0 \right\}, \\ \Gamma_2 &= \left\{ \lambda \in \Gamma \Big/ \frac{mp}{2} - \sum \lambda_i < -\frac{mp}{4} \right\}, \\ \Gamma_3 &= \left\{ \lambda \in \Gamma \Big/ -\frac{mp}{4} \leq \frac{mp}{2} - \sum \lambda_i < 0 \right\}. \end{aligned}$$

Let us remark that the number of elements of these sets is independent of n. For all three sets, we give a bound for κ.

In the cases Γ_1 and Γ_3, we have $m - \kappa \geq m/8$ (see Lemma 2). Otherwise, we will only use $m - \kappa \geq 0$.

Coming back to the proof, we have:

$$T_3 \leq C2^{j_s\epsilon p/2}\sum_{j_s}^{j_1}\left(\frac{\mathbf{2^j}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}}2^{-j\epsilon p/2}\sum_{\Gamma_1}+\sum_{\Gamma_2}+\sum_{\Gamma_3}\left(\frac{2^{j(mp/2-\Sigma\lambda_i)}}{n^{mp/2-\Sigma\lambda_i}}2^{-j(s-1/p)\lambda_1}2^{-j(m-\kappa)}\right)^{1/m''}$$

$$\leq Q_1+Q_2+Q_3.$$

First, let us study Q_1. Let us choose ϵ, m, m'' such that $m/8 \geq (1-\epsilon)m''p/2$. As $2^j/n \leq 1$ we have

$$\begin{aligned} Q_1 &\leq C2^{j_s\epsilon p/2}\sum_{j_s}^{j_1}\left(\frac{\mathbf{2^j}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}}2^{-j\epsilon p/2}\sum_{\Gamma_1}(2^{-j(s-1/p)\lambda_1}2^{-j(m-\kappa)})^{1/m''} \\ &\leq C\left(\frac{\mathbf{2^{j_s}}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}}\sum_{\Gamma_1}(2^{-j_s(s-1/p)\lambda_1}2^{-j_s(m-\kappa)})^{1/m''} \\ &\leq C\left(\frac{\mathbf{2^{j_s}}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}} \leq Cn^{-sp/(1+2s)}. \end{aligned}$$

To evaluate Q_2, we just need to observe that $mp/4 \geq m/8 \geq (1-\epsilon)m''p/2$ so 2^j has a negative power. Then, using inequalities (15) and (14) of Lemma 2:

$$\begin{aligned} Q_2 &\leq C2^{j_s\epsilon p/2}\sum_{\Gamma_2}2^{-j_s(s-1/p)\lambda_1/m''}\sum_{j_s}^{j_1}\left(\frac{\mathbf{2^j}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}}\left(\frac{2^j}{n}\right)^{(mp/2-\Sigma\lambda_i)/m''}2^{j((-m+\kappa)/m''-\epsilon p/2)} \\ &\leq C\left(\frac{\mathbf{2^{j_s}}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}}\sum_{\Gamma_2}2^{j_s(mp/2-m-\Sigma\lambda_i+\lambda_1/2)/m''}2^{-j_s(s+1/2-1/p)\lambda_1/m''}2^{j_s\kappa/m''}n^{-(mp/2+\Sigma\lambda_i)/m''} \\ &\leq C\left(\frac{\mathbf{2^{j_s}}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}}\sum_{\Gamma_2}n^{(-2s/(1+2s)(mp/2-\Sigma\lambda_i+\lambda_1/2)/m'')}n^{-1/(1+2s)(m-\kappa-\lambda_1/p)/m''} \\ &\leq C\left(\frac{\mathbf{2^{j_s}}}{\boldsymbol{n}}\right)^{\boldsymbol{p/2}} \leq Cn^{-sp/(1+2s)} \end{aligned}$$

For the last term Q_3, we just observe that 2^j has a negative power and we then use the same argument as in the previous case.

Finally, we have to look at the last term T_2. Using Lemma 5, for some $\epsilon > 0$, we write:

$$\begin{aligned} T_2 &\leq C_p2^{j_s\epsilon p/2}\sum_{j_0\leq j\leq j_s}2^{j(p/2-1)}2^{-j\epsilon p/2}\mathrm{E}(\Theta_j 1_{\{\hat{\Theta}_j\leq 2^j/n^{p/2}\}}1_{\{\Theta_j\leq 2^j/n^{p/2}\}} \\ &\quad + \Theta_j 1_{\{\hat{\Theta}_j\leq 2^j/n^{p/2}\}}1_{\{\Theta_j\geq 2^j/n^{p/2}\}} \\ &\leq C_p2^{j_s\epsilon p/2}\sum_{j_0\leq j\leq j_s}2^{j(p/2-1)}2^{-j\epsilon p/2}\left[\frac{\mathbf{2^j}}{\boldsymbol{n^{p/2}}}+\Theta_j P(|\hat{\Theta}_j-\Theta_j|>\tfrac{1}{2}\Theta_j)_{\{\Theta_j\geq 2(2^j/n^{p/2})\}}\right]. \end{aligned}$$

Using now the Chebyshev inequality for $m = p$ and Lemma 3 we get:

$$T_2 \le Cn^{-sp/(1+2s)} + C_p 2^{j_s\epsilon p/2} \sum_{j_0 \le j \le j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \Theta_j \Theta_j^{-m} \mathrm{E}(\hat{\Theta}_j - \Theta_j)^m 1_{\{\Theta_j \ge 2(2^j/n^{p/2})\}}$$

$$\le Cn^{-sp/(1+2s)} + C_p 2^{j_s\epsilon p/2} \sum_{j_0 \le j \le j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \Theta_j \Theta_j^{-m}$$

$$\times \sum_{l=1}^{p} (\Theta_j)^{m-lm/p} \left(\frac{2^j}{n^{p/2}}\right)^{lm/p} 1_{\{\Theta_j \ge 2(2^j/n^{p/2})\}}$$

$$\le Cn^{-sp/(1+2s)} + C_p 2^{j_s\epsilon p/2} \sum_{j_0 \le j \le j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \Theta_j \sum_{l=1}^{p} \left(\frac{2^j}{n^{p/2}} \Theta_j^{-1}\right)^{lm/p} 1_{\{\Theta_j \ge 2(2^j/n^{p/2})\}}$$

$$\le Cn^{-sp/(1+2s)} + C_p 2^{j_s\epsilon p/2} \sum_{j_0 \le j \le j_s} \mathbf{2}^{\boldsymbol{j(p/2-1)}} 2^{-j\epsilon p/2} \frac{2^j}{\boldsymbol{n^{p/2}}} \le Cn^{-sp/(1+2s)}.$$

Let us now investigate the case where p is not an even integer: let $p = \alpha p_1 + (1-\alpha)p_2$, where $\alpha \in]0,1[$ and $p_1, p_2 \in 2\mathbb{N}$. In this case, we replace the U-estimate by the interpolated estimate (see (10)).

Only the factors T_2 and T_3 have to be investigated separately. For T_3, the task is not difficult since it is enough to bound in (21) $\mathrm{E}(\hat{\Theta}_j)^m$ by $(\mathrm{E}\hat{\Theta}_j(p_1)^m)^{\alpha}(\mathrm{E}\hat{\Theta}_j(p_2)^m)^{1-\alpha}$ and then to use the same arguments as those following inequality (21):

$$T_3 \le C2^{j_s\epsilon p/2} \sum_{j_s \le j \le j_1} \left[\mathbf{2}^{\boldsymbol{j(p_1/2-1)}} 2^{-j\epsilon p/2} \frac{2^j}{\boldsymbol{n^{p_1/2}}} \left[\mathrm{E}(\hat{\Theta}_j(p_1))^m \left(\frac{2^j}{n^{p_1/2}}\right)^{-m} \right]^{1/m''} \right]^{\alpha}$$

$$\times \left[\mathbf{2}^{\boldsymbol{j(p_2/2-1)}} 2^{-j\epsilon p_2/2} \frac{2^j}{\boldsymbol{n^{p_2/2}}} \left[\mathrm{E}(\hat{\Theta}_j(p_2))^m \left(\frac{2^j}{n^{p_2/2}}\right)^{-m} \right]^{1/m''} \right]^{1-\alpha}.$$

Now, we have $B_{spq} \subset B_{sp_1q}$ as the functions have common impact support and $B_{spq} \subset B_{s_2p_2q}$ with $s - 1/p = s_2 - 1/p_2$. This implies that $\Theta_j(p_1) \le C2^{-j(s+1/2-1/p_1)p_1}$ and $\Theta_j(p_2) \le C2^{-j(s_2+1/2-1/p_2)p_2}$. It remains then just to extend the later proof for T_3 with these bounds. This does not present any difficulty.

For T_2, let us formulate the proof, for the sake of simplicity, when $p_2 - p_1 = 2$. Let $p_0 \le p \le p_0'$, $p_0' = p_0 + 2$. We have to replace Lemma 3 by the following result.

Lemma 6. *For all $m \in 2\mathbb{N}$, we have, if $j \ge 0, 2^j \le n$,*

$$\mathrm{E}[n^{-1}\hat{\Theta}_j(p_0) - \Theta_j(p_0')]^m \le C \sum_{l=1}^{p_0'} [\Theta_j(p_0' - l)]^m n^{-lm/2}.$$

Proof. For proving this Lemma 6, we follow the same scheme as we used for proving

Lemma 3. Keeping the same notation, we have

$$n^{-1}\hat{\Theta}_j(p_0) - \Theta_j(p_0') = \sum_k \frac{1}{C_n^{p_0}} \sum_{\alpha,\,|\alpha|=p_0} \sum_{l=1}^{p_0'} \beta_{jk}^{p_0'-l} \times \Bigg[\sum_{\gamma\subset\alpha,\,|\gamma|=l} \Delta\psi_{jk}^{\otimes l}(X_\gamma) + \frac{1}{n^{1/2}} \sum_{\gamma\subset\alpha,\,|\gamma|=l-1} \Delta\psi_{jk}^{\otimes(l-1)}(X_\gamma) + \frac{1}{n} \sum_{\gamma\subset\alpha,\,|\gamma|=l-2} \Delta\psi_{jk}^{\otimes(l-2)}(X_\gamma)\Bigg],$$

with the convention that

$$\sum_{\gamma\subset\alpha,\,|\gamma|=0} \Delta\psi_{jk}^{\otimes 0}(X_\gamma) = 2 \qquad \text{and} \qquad \sum_{\gamma\subset\alpha,\,|\gamma|=l<0} \Delta\psi_{jk}^{\otimes l}(X_\gamma) = 0.$$

Now, again reversing the order of summation, we obtain:

$$\mathrm{E}[n^{-1}\hat{\Theta}_j(p_0) - \Theta_j(p_0')]^m \le C \sum_{l=1}^{p_0'} \sum_{k_1,\dots,k_m} \prod_{t=1}^{m} |\beta_{jk_t}|^{p_0'-l} \left(\frac{Q_{m,l}}{n^{ml}} + \frac{Q_{m,l-1}}{n^{m(l-1/2)}} + \frac{Q_{m,l-2}}{n^{m(l-1)}} \right).$$

Using inequality (18), we obtain the result. □

Returning to the behaviour of T_2, we only have to look at the factor:

$$T_2 \le C_p 2^{j_s\epsilon p/2} \sum_{j_0\le j\le j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \mathrm{E}(\Theta_j 1_{\{\Theta_j \le 2(2^j/n^{p/2})\}} + \Theta_j 1_{\{\hat{\Theta}_j \le 2^j/n^{p/2}\}} 1_{\{\Theta_j \ge 2(2^j/n^{p/2})\}})$$

$$\le C_p 2^{j_s\epsilon p/2} \sum_{j_0\le j\le j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} [\Theta_j 1_{\{\Theta_j \le 2(2^j/n^{p/2})\}} + \Theta_j (1_{\{\hat{\Theta}_j(p_0) \le 2^j/n^{p_0/2}\}} 1_{\{\Theta_j \ge 2(2^j/n^{p/2})\}} + 1_{\{\hat{\Theta}_j(p_0') \le 2^j/n^{p_0'/2}\}} 1_{\{\Theta_j \le 2(2^j/n^{p/2})\}})].$$

But we have

$$\left\{\Theta_j \ge 2\frac{2^j}{n^{p/2}}\right\} \subset \left\{\Theta_j(p_0') \ge 2\frac{2^j}{n^{p_0'/2}}\right\} \tag{22}$$

Therefore,

$$T_2 \le C_p 2^{j_s\epsilon p/2} \sum_{j_0\le j\le j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \Bigg[\frac{2^j}{n^{p/2}} + \Theta_j 1_{\{\Theta_j(p_0') \ge 2(2^j/n^{p_0'/2})\}} \times \left(P(|\hat{\Theta}_j(p_0') - \Theta_j(p_0')| \ge \Theta_j(p_0')/2) + P\left(\left|\frac{1}{n}\hat{\Theta}_j(p_0) - \Theta_j(p_0')\right| \ge \Theta_j(p_0')/2) \right) \right) \Bigg].$$

Using Chebyshev inequality, Lemmas 3 and 6 and inequality (6), we get for $m \in 2\mathbb{N}$, $m > p$:

$$\begin{aligned}
T_2 &\leq C_p 2^{j_s \epsilon p/2} 2 \sum_{j_0 \leq j \leq j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \\
&\quad \times \left[\frac{2^j}{n^{p/2}} + \sum_{l=1}^{p_0'} \Theta_j(p_0' - l)^m n^{-ml/2} \Theta_j(p_0')^{1-m} 1_{\{\Theta_j(p_0') \geq 2(2^j/n^{p_0'/2})\}} \right] \\
&\leq C_p 2^{j_s \epsilon p/2} 2 \sum_{j_0 \leq j \leq j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \\
&\quad \times \left[\frac{2^j}{n^{p/2}} + \sum_{l=1}^{p_0'} n^{-ml/2} \Theta_j(p_0')^{(p-ml)/p_0'} 2^{j(1-p/p_0'+ml/p_0')} 1_{\{\Theta_j(p_0') \geq 2(2^j/n^{p_0'/2})\}} \right] \\
&\leq C_p 2^{j_s \epsilon p/2} 2 \sum_{j_0 \leq j \leq j_s} 2^{j(p/2-1)} 2^{-j\epsilon p/2} \left[\frac{\mathbf{2^j}}{\mathbf{n^{p/2}}} + \sum_{l=1}^{p_0'} n^{-ml/2} \left(\frac{2^j}{\mathbf{n^{p_0'/2}}} \right)^{(\mathbf{p}-ml)/\mathbf{p_0'}} \mathbf{2}^{\mathbf{j}(\mathbf{1}-p/p_0'+ml/p_0')} \right] \\
&\leq C_p n^{-sp/(1+2s)}.
\end{aligned}$$

3.4. PROOF OF THEOREM 1 IN THE CASE OF SOFT THRESHOLDING

The proof for the case of soft thresholding is exactly similar to the proof for the hard thresholding case. The bias term and the stochastic term are the same. For the nonlinear stochastic term, let us denote by R_1, R_2, R_3, R_4 the quantities associated with T_1, T_2, T_3, T_4. R_2 and R_4 are identical to T_2 and T_4 (for $\hat{\Theta}_j \leq 2^j/n^{p/2}$); R_1 and R_3 are bounded as T_1 (see (19)) and T_3 (see (21)) (for $\hat{\Theta}_j \geq 2^j/n^{p/2}$). Indeed, using Lemma 5 once again for some $\epsilon_1 > 0$ and some $\epsilon_2 < 0$:

$$\begin{aligned}
R_1 &= \left\| \sum_{j_0 \leq j \leq j_s} \sum_k (\hat{\eta}_j \hat{\beta}_{jk} - \beta_{jk}) \psi_{jk} 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} \right\|_p^p \leq C 2^{j_s \epsilon_1 p/2} \sum_{j_0}^{j_s} 2^{j(p/2-1)} 2^{j\epsilon_1 p/2} \\
&\quad \times \left[\mathrm{E} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^p 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} + \mathrm{E} \left(\frac{2^j}{n^{p/2}} \right)^p \hat{\Theta}_j^{1-p} 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} \right] \\
&\leq C 2^{j_s \epsilon_1 p/2} \sum_{j_0}^{j_s} 2^{j(p/2-1)} 2^{j\epsilon_1 p/2} \left[\mathrm{E} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^p + \frac{2^j}{n^{p/2}} \right] \leq C n^{-sp/(1+2s)}.
\end{aligned}$$

$$\begin{aligned}
R_3 &= \left\| \sum_{j_s \leq j \leq j_1} \sum_k (\hat{\eta}_j \hat{\beta}_{jk} - \beta_{jk}) \psi_{jk} 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} \right\|_p^p \leq C 2^{j_s \epsilon_2 p/2} \sum_{j_2}^{j_1} 2^{j(p/2-1)} 2^{j\epsilon_2 p/2} \\
&\quad \times \left[\mathrm{E} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^p 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} + \mathrm{E} \left(\frac{2^j}{n^{p/2}} \right)^p \hat{\Theta}_j^{1-p} 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} \right]
\end{aligned}$$

$$\leq C2^{j_s \epsilon_2 p/2} \sum_{j_2}^{j_1} 2^{j(p/2-1)} 2^{j\epsilon_2 p/2} \left[\mathrm{E} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^p 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} + \mathrm{E}\hat{\Theta}_j 1_{\{\hat{\Theta}_j \geq 2^j/n^{p/2}\}} \right]$$

$$\leq C2^{j_s \epsilon_2 p/2} \sum_{j_2}^{j_1} 2^{j(p/2-1)} 2^{j\epsilon_2 p/2} \left[\sum_k (\mathrm{E}|\hat{\beta}_{jk} - \beta_{jk}|^{pm'})^{1/m'} P\left(\hat{\Theta}_j \geq \frac{2^j}{n^{p/2}}\right)^{1/m''} + \mathrm{E}(\hat{\Theta}_j)^{m'''} \left(\frac{2^j}{n^{p/2}}\right)^{1-m'''} \right]$$

$$\leq C2^{j_s \epsilon_2 p/2} \sum_{j_2}^{j_1} 2^{j(p/2-1)} 2^{j\epsilon_2 p/2} \frac{2^j}{n^{p/2}} \left[\left((\mathrm{E}(\hat{\Theta}_j)^m \left(\frac{2^j}{n^{p/2}}\right)^{-m} \right)^{1/m''} + \mathrm{E}(\hat{\Theta}_j)^{m'''} \left(\frac{2^j}{n^{p/2}}\right)^{-1m'''} \right]$$

$$\leq C n^{sp/(1+2s)}.$$

where m, m', m'' are chosen as for the study of T_3 and $m''' \in 2\mathbb{N}$.

Appendix: Proof of Lemma 4

For $r > 1$, we have, using the Rosenthal inequality, see Rosenthal (1972):

$$\left(\sum_k (\mathrm{E}|\hat{\gamma}_{jk} - \gamma_{jk}|^{mr})^{1/r} \right)$$

$$\leq \sum_k (C_m)^{1/r} \left[\frac{\left(\int |g|^2 (x-k) f(x/2^j)\, \mathrm{d}x \right) (2\|g\|_\infty 2^{j/2})^{mr-2}}{n^{mr-1}} + \frac{\left(\int |g|^2 (x-k) f(x/2^j)\, \mathrm{d}x \right)^{mr/2}}{n^{mr/2}} \right]^{1/r}$$

$$\leq (C_m)^{1/r} \left[\sum_k \left(\int |g|^2 (x-k) f\left(\frac{x}{2^j}\right) \mathrm{d}x \right)^{1/r} \frac{(2\|g\|_\infty)^{m-2/r} 2^{j/2(m-2/r)}}{n^{m-1/r}} + \sum_k \frac{\left(\int |g|^2 (x-k) f(x/2^j)\, \mathrm{d}x \right)^{m/2}}{n^{m/2}} \right]$$

$$\times \frac{1}{n^{m/2}} \sum_k \left(\int |g|^2 (x-k) f\left(\frac{x}{2^j}\right) \mathrm{d}x \right)^{m/2} \leq \frac{1}{n^{m/2}} \sum_k \int |g|^2 (x-k) f^{m/2}\left(\frac{x}{2^j}\right) \mathrm{d}x$$

by the Jensen inequality and $m/2 \geq 1$

$$\text{LHS} \leq \frac{2^j}{n^{m/2}} \left\| \sum_k |g|^2(\bullet - k) \right\|_\infty \int f^{m/2}(x)\,\mathrm{d}x.$$

For the second term, noticing that in the sum all the terms but $-B2^j - A \leq k \leq B2^j - A$ are zero and the rest are bounded, we have:

$$\sum_k \left(\int |g|^2(x-k) f\left(\frac{x}{2^j}\right) \mathrm{d}x \right)^{1/r} \leq \|g\|_2^{2/r} \|f\|_\infty^{1/r} 2^j 2B.$$

So:

$$\sum_k \left(\int |g|^2(x-k) f\left(\frac{x}{2^j}\right) \mathrm{d}x \right)^{1/r} \frac{(2\|g\|_\infty)^{m-2/r} 2^{(j/2)(m-2/r)})}{n^{m-1/r}}$$

$$\leq 2B \frac{2^j}{n^{m/2}} \|g\|_\infty^m 2^{m-2/r} \frac{2^{j(m/2-1/r)}}{n^{m/2-1/r}}.$$

As $2^j \leq n$, the lemma is proved. The proof for $r = 1$ is in the same spirit and in fact easier.

References

Bergh, J. and Löfström, J. (1976) *Interpolation Spaces – An Introduction*. New York: Springer-Verlag.

Birgé, L. (1983) Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **65**, 181–237.

Bretagnolle, J. and Carol-Huber, C. (1979) Estimation des densités: risque minimax. *Z. Wahrscheinlichkeitstheorie Verw. Geb.*, **47**, 119–137.

Daubechies, I. (1992) *Ten Lectures on Wavelets*. Philadelphia: SIAM.

Devroye, L. (1985) *Nonparametric Density Estimation*. New York: Wiley.

Donoho, D.L. and Johnstone, I.M. (1995) Minimax Estimation via Wavelet shrinkage. *Ann. Statist.* To appear.

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

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1995) Wavelet shrinkage: Asymptopia? (with discussion) *J. Roy. Statist. Soc*, Ser. B, **57,** 307–369.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.* To appear.

Efromovich, S.Y. (1985) Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.*, **30**, 557–661.

Efromovich, S.Y. and Pinsker, M.S. (1982) Estimation of square-integrable probability density of a random variable. *Problemy Peredachi Informatsii*, **18**, 19–38 (in Russian). *Problems of Inform. Transmission*, **18** (1983), 175–189 (in English).

Härdle, W. and Marron, J.S. (1985) Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.*, **13**, 1465, 1481.

Ibragimov, I.A. and Has'minskii, R.Z. (1982) Bounds for the risk of nonparametric regression estimates. *Theory Probab. Appl.*, **27**, 84–99.

Kerkyacharian, G. and Picard, D. (1992) Density estimation in Besov spaces. *Statist. Probab. Lett.*, **13**, 15-24.

Lepskii, O.V. (1990) On one problem of adaptive estimation on white Gaussian noise. *Teor. Veroyatnost. i Primenen.*, **35**, 454–466 (in English).

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

Meyer, Y. (1990) *Ondelettes*. Paris: Hermann.

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

Peetre, J. (1976) *New Thoughts on Besov Spaces*. Duke Univ Math. Ser. 1. Durham, NC.

Pinsker, M.S. (1980) Optimal filtering of square integrable signals in Gaussian white noise. *Problems Inform. Transmission*, **16**, 120–133.

Rosenthal, H.P. (1972) On the span in L^p of sequences of independent random variables. In *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 2, pp. 149–167. Berkeley: University of California Press.

Scott, D.W. (1992) *Multivariate Density Estimation*. New York: Wiley.

Silverman, B.W. (1986) *Density Estimation for Statistics and Data Analysis*. London: Chapman & Hall.

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

Triebel, H. (1990) *Theory of Function Spaces 2*. Basel: Birkhäuser Verlag.

Received August 1994 and revised September 1995