Open access • Journal Article • DOI:10.1007/S00365-006-0660-5

Thresholding in Learning Theory — Source link

Gerard Kerkyacharian, Dominique Picard

Institutions: University of Paris

Published on: 20 Apr 2007 - Constructive Approximation (Springer-Verlag)

Topics: Reproducing kernel Hilbert space, Besov space, Thresholding, Bounded function and Smoothness (probability theory)

Thresholding in Learning Theory

Gerard Kerkyacharian, Dominique Picard

THRESHOLDING IN LEARNING THEORY

GÉRARD KERKYACHARIAN AND DOMINIQUE PICARD

RÉSUMÉ. In this paper we investigate the problem of learning an unknown bounded function. We be emphasize special cases where it is possible to provide very simple (in terms of computation) estimates enjoying in addition the property of being universal : their construction does not depend on a priori knowledge on regularity conditions on the unknown object and still they have almost optimal properties for a whole bunch of functions spaces. These estimates are constructed using a thresholding schema, which has proven in the last decade in statistics to have very good properties for recovering signals with inhomogeneous smoothness but has not been extensively developed in Learning Theory.

We will basically consider two particular situations. In the first case, we consider the RKHS situation. In this case, we produce a new algorithm and investigate its performances in $L_2(\hat\rho_X)$. The exponential rates of convergences are proved to be almost optimal, and the regularity assumptions are expressed in simple terms.

The second case considers a more specified situation where the X_i's are one dimensional and the estimator is a wavelet thresholding estimate. The results are comparable in this setting to those obtained in the RKHS situation as concern the critical value and the exponential rates. The advantage here is that we are able to state the results in the $L_2(\rho_X)$ norm and the regularity conditions are expressed in terms of standard Hölder spaces.

1. Introduction

In this paper, we are interested in the problem of learning an unknown function defined on a set $\mathbb{X}$ which takes values in a set Y. We assume that $\mathbb{X}$ is a compact domain in $\mathbb{R}^d$ and $Y = [-M/2, M/2]$ is a finite interval in $\mathbb{R}$. This problem, also called regression problem, has a long history in Statistics (many references can be found, for example, in the following books [Ibragimov and Has'minskiĭ, 1981], [Van de Geer, 2001] and [Györfi et al., 2002]). It has recently drawn much attention in the work of [Cucker and Smale, 2002] and amplified upon in [Poggio and Smale, 2003].

We will assume to observe an n sample $Z_1, \dots, Z_n$ of $Z = (X, Y)$. The distribution of Z in denoted by ρ. Our aim is to recover the function f_ρ :

$$f_\rho(x) = \mathbb{E}_\rho[Y|X = x].$$

We shall have as our goal to obtain estimations to f_ρ with the error measured in the $L_2(\mathbb{X}, \rho_X)$ norm, or $L_2(\mathbb{X}, \hat\rho_X)$ where $\hat\rho$ is the empirical measure calculated on the X_i's.

$$\|g\|_\rho^2 = \int_{\mathbb{X}} g(x)^2 d\rho(x), \ \|g\|_{\hat\rho_X}^2 = \frac{1}{n}\sum_{i=1}^n g(X_i)^2$$

Given any $\eta > 0$, if $\hat f$ is an estimator of f_ρ (i.e. a measurable function of $Z_1, \dots, Z_n$, taking its values in the set, say, of bounded functions),

$$\rho^{\otimes n}\{\mathbf{z} : \|\hat f - f_\rho\| > \eta\} \tag{1}$$

measures the confidence we have that the estimator $\hat{f}$ is accurate to tolerance η.

Contrary to Statistics, where people are mainly concerned with evaluation of moments of $\|\hat{f} - f_\rho\|$ (except rare examples, see [Korostelev, 2003] or [Korostelev and Spokoiny, 1996]...) Learning Theory focuses on investigating the decay of (1) as $n \to \infty$ and η increases.

Another difference with the Statistics point of view is that one mail goal in Learning Theory is to obtain results with almost no assumptions on the distribution ρ. However, it is known that it is not possible to have fast rates of convergence without assumptions and a large portion of Statistics and Learning Theory proceeds under the condition that f_ρ is in a known set Θ. Typical choices of Θ are compact sets determined by some smoothness condition or by some prescribed rate of decay for a specific approximation process. Given our prior Θ and the associated class $\mathbb{M}(\Theta)$ of measures ρ, it has been defined in [DeVore et al., 2004], for each $\eta > 0$ the *accuracy confidence function*

$$\mathbf{AC}_n(\Theta, \hat{f}, \eta) := \sup_{\rho \in \mathbb{M}(\Theta)} \rho^{\otimes n}\{\mathbf{z} : \|f_\rho - \hat{f}\| > \eta\}. \tag{2}$$

This quantity measures a uniform confidence (over the space $\mathbb{M}(\Theta)$) we have that the estimator $f_{\mathbf{z}}$ is accurate to tolerance η.

Upper and lower bounds for $\mathbf{AC}$ have been proved in [DeVore et al., 2004]. In most examples, there is a critical $\eta = \eta(n, \Theta)$ after which (2) decreases exponentially. This critical value $\eta(\Theta, n)$ is essential since it yields, as a consequence bounds of type $e_m(\Theta, \hat{f}) \leq C\eta(\Theta, n)^q$ which have been extensively studied in statistics, for

$$e_m(\Theta, \hat{f}) = \sup_{\rho \in \mathbb{M}(\Theta)} \mathbb{E}_{\rho^{\otimes n}} \|f_z - f_\rho\|^q \tag{3}$$

To evaluate lower bounds for the function $\mathbf{AC}_m(\Theta, f_{\mathbf{z}}, \eta)$, [DeVore et al., 2004] considered :

$$\mathbf{AC}_n(\Theta, \eta) := \inf_{\hat{f}} \sup_{\rho \in \mathbb{M}(\Theta)} \rho^{\otimes n}\{\|f_\rho - \hat{f}\| > \eta\}$$

and the following result has been established :

$$\mathbf{AC}_n(\Theta, \eta) \geq C\{ \begin{array}{ll} e^{-cn\eta^2}, & \eta \geq \eta_n, \\ 1, & \eta \leq \eta_n, \end{array}$$

where η_n is defined by the relation : $\ln \bar{N}(\Theta, \eta_n) \sim c^2 n(\eta_n)^2$. $N(\Theta, \eta_n)$ is the 'tight entropy' defined by :

$$\bar{N}(\Theta, \eta) := \sup\{N : \exists\ f_0, f_1, ... f_N \in \Theta, \ \text{ with } c_0\eta \leq \|f_i - f_j\|_{L_2(\rho_X)} \leq c_1\eta, \ \forall i \neq j\}.$$

For instance, $\eta_n = n^{-\frac{s}{2s+d}}$ for the Besov space $B^s_q(L_\infty(\mathbb{R}^d))$ which corresponds to similar results proved in statistics (actually with more restricted assumptions on the set of probabilities ρ) :

$$\inf_{\hat{f}} \sup_{\rho \in \mathbb{M}'(B^s_q(L_\infty(\mathbb{R}^d)))} \mathbb{E}\|f_\rho - \hat{f}\|_{dx} \geq cn^{-\frac{s}{2s+d}}.$$

See, for instance [Ibragimov and Has'minskiĭ, 1981], [Stone, 1982], [Nemirovskiy, 1985] for a slightly more restricted context than Besov spaces, and [Donoho et al., 1995]...

Concerning upper bounds for $\mathbf{AC}_n(\Theta, \eta)$, many reverse properties have been established :see for instance [Yang and Barron, 1999] in statistical context, [Cucker and Smale, 2002], [DeVore et al., 2004], [Konyagyn and Temlyakov, 2004], in learning theory. These upper bounds are generally proved using particular estimation methods more often based on empirical mean square minimization.

$$\hat{f} = \mathsf{Argmin}\{\sum_{i=1}^{n} (Y_i - f(X_i))^2,\ f \in \mathbb{H}_n\}$$

These very nice estimation rules raise nevertheless two important problems : First, they generally require heavy computation times. The second serious problem lies in the fact that their construction (the choice of $\mathbb{H}_n$) is most of the time highly depending on Θ : There also exist universal estimates (see [Temlyakov, 2005]), however these rules are up to now prohibitive in terms of computation time.

Our aim in this paper will be to emphasize special constructions and cases where it is possible to provide very simple (in terms of computation) estimates enjoying in addition the property of being universal : their construction does not depend on a particular Θ and still they have almost optimal properties for a whole bunch of spaces Θ. These estimates are constructed using a thresholding schema, which has proven in the last decade in statistics to have very good properties for recovering signals with inhomogeneous smoothness.

In this paper, we will basically consider two particular situations. In the first case, we consider the RKHS situation. In this case, we produce a new algorithm and investigate its performances in $L_2(\hat{\rho}_X)$. The exponential rates of convergences are good : the critical value η_n is the one predicted by [DeVore et al., 2004], and the exponential rates are comparable to those recently obtained by [Smale and Zhou, 2005], although the loss is not the same ($L_2(\rho_X)$ in SZ), and the regularity assumptions are somewhat different : in SZ, regularity assumptions are expressed in terms of RKHS spaces. These assumptions may seem more intrinsic. However it is difficult to figure out exactly what they mean since they are depending on the unknown measure ρ_X. Our conditions are also depending on the kernel, but easy to figure out.

The second case considers a more specified situation where the X_i's are one dimensional and the estimator is a wavelet thresholding estimate. The results are comparable in this setting to those obtained in the RKHS situation as concern the critical value and the exponential rates. The advantage here is that we are able to state the results in the $L_2(\rho_X)$ norm and the regularity conditions are expressed in terms of standard Hölder spaces.

2. Least squares and thresholding procedures

In this short section, we will consider the construction of our thresholding estimates. To make easier their understanding and motivate their consideration, we give here a connection with general least square estimates. However this construction will not be used in the sequel and can be skipped by a hurried reader which can go directly to the next section.

Empirical mean square minimization consists in considering

$$\hat{f} = \mathsf{Argmin}\{\sum_{i=1}^{n} (Y_i - f(X_i))^2,\ f \in \mathbb{H}_n\}$$

for a specified set $\mathbb{H}_n$. Let us look at particular cases of $\mathbb{H}_n$ leading to especially computable forms of $\hat{f}$. Let us suppose that we have a collection of functions $(e_k)_k$ verifying the following property :

$$(P) : (e_k) : \frac{1}{n} \sum_{i=1}^{n} e_k(X_i) e_l(X_i) = \delta_{kl}$$

(i.e. (e_k) is an orthonormal system for the empirical measure $\hat{\rho}$ on the $X_i's$.)

$$\hat{\rho}_X = \frac{1}{n}\sum_{i=1}^{n} \delta(X_i)$$

if $\delta(x)$ is the Dirac measure at the point x.

Now, associated to this collection of functions, let us consider the following particular spaces :

$$\mathbb{H}_n^{(1)} = \{f = \sum_{i=1}^{N} \alpha_i e_i\}, \quad \mathbb{H}_n^{(2)} = \{f = \sum_{i=1}^{N} \alpha_i e_i, \ \sum |\alpha_i| \leq \kappa\}$$

$$\mathbb{H}_n^{(3)} = \{f = \sum_{i=1}^{N} \alpha_i e_i, \ \#\{|\alpha_i| \neq 0\} \leq \kappa\}$$

If we now introduce the 3 following estimations of these coefficients :

$$\hat{\alpha}_k = \frac{1}{n}\sum_{i=1}^{n} e_k(X_i)Y_i, \quad \hat{\alpha}_k^{(1)} = sign(\hat{\alpha}_k)|\hat{\alpha}_k - \lambda|_+$$

$$\hat{\alpha}_k^{(2)} = \hat{\alpha}_k I\{|\hat{\alpha}_k| \geq \lambda\}$$

It is easy to prove that there exists $\lambda^i(\kappa)$ such that the following rules are empirical minimizers for the respective spaces $\mathbb{H}_n^{(i)}, \ i \in \{1, 2, 3\}$:

$$\hat{f}^1 = \sum_{k=1}^{N} \hat{\alpha}_k e_k, \quad \hat{f}^2 = \sum_{k=1}^{N} \hat{\alpha}_k^{(1)} e_k$$

$$\hat{f}^3 = \sum_{k=1}^{N} \hat{\alpha}_k^{(2)} e_k$$

These three rules are common in the statistical litterature. $\hat{f}^1$ is generally refered to as linear estimate, whereas, $\hat{f}^2$ and $\hat{f}^3$ are known as (respectively soft and hard) thresholding estimates.

Our aim in this paper is to study the behavior of these estimators, principally $\hat{f}^3$, in different situations. The main difficulty of this paradigm obviously lies in the question : How to choose the functions (e_k) such that condition (P) is verified and suitably chosen tuning constants N, λ ?

This first problem is difficult to solve, if not impossible, and in the sequel, we will not assume that property (P) is verified, but we are going to consider situations where this property can be considered as 'almost true'.

3. RKHS situation

3.1. **Assumptions, estimation rules and regularity conditions.**

3.1.1. *Assumptions on the kernel.* Let us take the case of a symmetric kernel $K(\cdot,\cdot)$ (we do not explicitely need the fact that K is a Mercer kernel). We assume that the kernel K is uniformly bounded by an absolute constant κ. Our fundamental assumption will be the following :

(A) : There exists a set of p determinist points in $\mathbb{R}^d$

$$\{x_1, \dots x_p\}$$

(p will tend to infinity with n) such that the following $p \times p$ matrix M_{np} whose entries are, $(\frac{1}{n}\sum_{i=1}^n K(x_l, X_i)K(X_i, x_k))_{kl}$ is almost diagonal, in the sense that : There exists $0 \le \delta < 1$ such that :

$$\forall x \in \mathbb{R}^p, \quad \|x\|_{l_2}^2 (1-\delta)^2 \le x^t M_{np} x \le \|x\|_{l_2}^2 (1+\delta)^2 \tag{4}$$

$$\|x\|_{l_\infty} (1-\delta) \le \|M_{np} x\|_{l_\infty} \tag{5}$$

We do not assume anything about δ but this quantity will enter into the performances results of the procedure. δ will be desired to be as small as possible. Notice that in general, such an assumption reflects the concentration properties of the kernel, and is quite easy to verify in practical situations where δ can be computed empirically. In particular, we allow in the sequel δ to be a random quantity depending on the observations.

3.1.2. *Estimation rule.* Let us consider the following estimation rule : We will denote by Y the vector with coordinates Y_i, $\varepsilon_i = Y_i - f_\rho(X_i)$, and ε will be the vector with coordinates ε_i. Let us denote by f_X the n dimensional vector which entries are $f_\rho(X_i)$, and K the $p \times n$ matrix which entries $K(x_l, X_i)$ (so $\frac{1}{n}KK^t = M_{np}$), and introduce :

$$\begin{aligned} t_n &= \frac{\log n}{n}, \quad \lambda_n = T\sqrt{t_n}, & (6)\\ z &= (z_1, \dots, z_p)^t = (KK^t)^{-1}KY, & (7)\\ \tilde{z} &= (\tilde{z}_1, \dots, \tilde{z}_p)^t, \quad \tilde{z}_l = z_l I\{|z_l| \ge \lambda_n\} & (8) \end{aligned}$$

T will be chosen so that $T > \sqrt{M^2 + \frac{1}{2}} \vee 4$, and finally, our estimate will be :

$$\hat{f} = \sum_{l=1}^p \tilde{z}_l K(x_l, \cdot). \tag{9}$$

As is easily seen, $\hat{f}$ takes its inspiration into $\hat{f}_3$ and it is worthwhile to notice that its construction do not depend on any regularity parameter.

3.1.3. *Regularity conditions.* We will assume the following sparsity conditions on the function f_ρ :

Let us take

$$p = \lfloor \left(\frac{n}{\log n}\right)^{\frac{1}{2}} \rfloor$$

For any n, there exists $\alpha_1, \dots, \alpha_p$, such that

$$\|f_\rho - \sum_{l=1}^{p} \alpha_l K(x_l, \dot{)}\|_\infty \quad \leq \quad cp^{-3/2} \tag{10}$$

$$\forall \lambda > 0, \; \mathrm{card}\{|\alpha_l| \geq \lambda\} \quad \leq \quad c\lambda^{-\frac{2}{1+2s}} \tag{11}$$

These conditions reflect approximation properties for the function f_ρ by linear combinations of vectors in the RKHS (when K is a Mercer kernel). These properties are quantified by conditions on the coefficients α_i's, which are standard in various situations (Fourier, wavelet coefficients...). As discussed in [Kerkyacharian and Picard, 2000] condition (10) reflects a 'minimal compacity condition' which do not interfere in the entropy calculations (for instance) neither in the minimax rates of convergence. Condition (11) does drive the rates. It is given here with a Lorentz type constraint on the α_i's. These conditions are obviously implied by l_r conditions (for appropriate r) which then looks very much like Besov conditions.

We will measure the error by the following norm (empirical norm) :

$$\|g\|^2_{\hat{\rho}_X} = \frac{1}{n} \sum_{i=1}^{n} g(X_i)^2 \tag{12}$$

$\lfloor x \rfloor$ denotes the integer part of x. Our result is the following :

Theorem 1. *Let us take*

$$p = \lfloor \left(\frac{n}{\log n} \right)^{\frac{1}{2}} \rfloor$$

For any $s > 1/2$, we define,

$$\eta_n = [\frac{n}{\log n}]^{\frac{-s}{1+2s}}.$$

Under the conditions above, there exists a constant D, such that

$$\sup_{\rho \in \mathbb{M}(\Theta)} \rho^{\otimes n}\{\|f_\rho - \hat{f}\|_{\hat{\rho}} > (1-\delta)^{-1}\eta\} \leq T\{ \begin{array}{ll} e^{-\gamma[np^{-1}\eta^2 \vee \log n]}, & \eta \geq D\eta_n, \\ 1, & \eta \leq D\eta_n, \end{array} \tag{13}$$

Remark 1. *As mentioned in the introduction these results prove that the behavior of this estimator is optimal in terms of the critical value η_n as predicted in* [DeVore et al., 2004]*. In terms of exponential rates, they are suboptimal because of the term p^{-1}. However it is worthwhile to notice that these rates still are good : they are comparable to those obtained by* [Smale and Zhou, 2005]*, although the loss is not the same and the regularity assumptions are somewhat different. In addition, we observe that if not entirely opimal, these rates are always better than n^{-c}.*

Finally, it is important to notice the following technical facts which will be crucial in the sequel : because $s > 1/2$, $\eta_n \geq p^{-1}$. Condition (11) can obviously always be replaced by :

$$\forall \lambda > 0, \; \mathrm{card}\{|\alpha_l| \geq \lambda\} \leq c\lambda^{-\frac{2}{1+2s}} \wedge p) \tag{14}$$

3.2. **Proof of the theorem.** First, let us remark that :

$$\begin{aligned}
\|f_\rho - \hat{f}\|_{\hat{\rho}} &\leq \|f - \sum_{l=1}^{p} \alpha_l K(x_l, \dot{})\|_\infty + \|\sum_{l=1}^{p} \alpha_l K(x_l, \dot{}) - \hat{f}\|_{\hat{\rho}} \\
&\leq cp^{-3/2} + \|\sum_{l=1}^{p} (\alpha_l - \tilde{z}_l) K(x_l, \dot{})\|_{\hat{\rho}} \\
&\leq cp^{-3/2} + [(\alpha - \tilde{z}) M_{np} (\alpha - \tilde{z})]^{\frac{1}{2}} \\
&\leq c\eta_n + (1+\delta)[\sum_{l=1}^{p} (\alpha_l - \tilde{z}_l)^2]^{\frac{1}{2}}
\end{aligned}$$

Notice that the first line used hypothesis (10), and the last one (4).

$$\begin{aligned}
\sum_{l=1}^{p} (\alpha_l - \tilde{z}_l)^2 &\leq \sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|z_l| \geq \lambda_n\}[\mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} + \mathbb{I}\{|\alpha_l| < \lambda_n/2\}] \\
&+ \sum_{l=1}^{p} [\alpha_l]^2 \mathbb{I}\{|z_l| < \lambda_n\}[\mathbb{I}\{|\alpha_l| \geq 2\lambda_n\} + \mathbb{I}\{|\alpha_l| < 2\lambda_n\}] \\
&:= \mathsf{BB} + \mathsf{BS} + \mathsf{SB} + \mathsf{SS}
\end{aligned}$$

Let us study the term SS. First we remark that because of condition (11) on f_ρ, we know that

$$\mathrm{card}\{|\alpha_l| \geq \lambda_n\} \leq c\lambda_n^{\frac{-2}{1+2s}}$$

and it is not difficult to prove that (11) is equivalent to the following characterization (the result is standard in Lorenz spaces and in any case can be found in [Cohen et al., 2001]

$$\forall\, \lambda > 0, \sum_{l} \alpha_l^2 \mathbb{I}\{|\alpha_l| < \lambda\} \leq c\lambda^{\frac{4s}{1+2s}} \tag{15}$$

Hence, using (15) :

$$\mathsf{SS} \leq c\lambda_n^{\frac{4s}{1+2s}} = c(\sqrt{t_n}T)^{\frac{4s}{1+2s}} = cT^{\frac{4s}{1+2s}}\eta_n^2$$

Let us now investigate the term SB : We observe that $\mathbb{I}\{|z_l| < \lambda_n\}\mathbb{I}\{|\alpha_l| \geq 2\lambda_n\} \leq \mathbb{I}\{|\alpha_l - z_l| \geq |\alpha_l|/2\}\mathbb{I}\{|\alpha_l| \geq 2\lambda_n\}$, hence :

$$\begin{aligned}
\mathsf{SB} &\leq \sum_{l=1}^{p} [\alpha_l]^2 \mathbb{I}\{|z_l - \alpha_l| \geq |\alpha_l|/2\}\mathbb{I}\{|\alpha_l| \geq 2\lambda_n\} \\
&\leq 4\sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq 2\lambda_n\}
\end{aligned}$$

In the same way :

$$\begin{aligned} \mathsf{BB} &= \sum_{l=1}^{p} [\alpha_l - z_l]^2 \mathbb{I}\{|z_l| \geq \lambda_n;\ |\alpha_l| \geq \lambda_n/2\} \\ &\leq \sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \end{aligned}$$

So BB and SB can be treated in the same way, since

$$\sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq 2\lambda_n\} \leq \sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\}$$

$$\mathsf{BB} + \mathsf{SB} \leq 5 \sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\}$$

Let

$$p^* = \mathrm{card}\{|\alpha_l| \geq \lambda_n/2\} \leq c(\lambda_n/2)^{\frac{-2}{1+2s}} \tag{16}$$

3.2.1. *Study of* $\sum_{l=1}^{p}(\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\}$. Let us denote by $\bar{f}_X$ the vector with coordinates $[\bar{f}_X]_i = \bar{f}(X_i) = \sum_{l=1}^{p} \alpha_l K(x_l, X_i)$:

$$\bar{f}_X = K^t \alpha$$

Let us recall that f_X is the n dimensional vector which entries are $f(X_i)$. and by hypothesis (10), $|f(X_i) - \bar{f}(X_i)| \leq cp^{-3/2}$ So that,

$$\begin{aligned} \alpha &= (KK^t)^{-1} K \bar{f}_X, \\ z &= (KK^t)^{-1} KY = (KK^t)^{-1} K[f_X + \varepsilon], \\ \alpha - z &= (KK^t)^{-1} K\varepsilon + (KK^t)^{-1} K[\bar{f}_X - f_X] \end{aligned}$$

¿From this we deduce,

$$\|\alpha - z\|_{l_2} \leq \|(KK^t)^{-1} K\varepsilon\|_{l_2(p)} + \|(KK^t)^{-1} K[\bar{f}_X - f_X]\|_{l_2(p)}$$

But, since $(KK^t)^{-1} = \frac{1}{n} {M_{np}}^{-1}$, and using (4),

$$\begin{aligned} \|(KK^t)^{-1} K[\bar{f}_X - f_X]\|_{l_2(p)} &= \frac{1}{n} \|M_{np}^{-1} K[\bar{f}_X - f_X]\|_{l_2(p)} \\ &\leq (1-\delta)^{-1} \|\frac{1}{n} K[\bar{f}_X - f_X]\|_{l_2(p)} \\ &\leq (1-\delta)^{-1} \kappa \|f_X - \bar{f}_X\|_\infty \sqrt{p} \\ &\leq (1-\delta)^{-1} c \frac{1}{p} \kappa \leq c(1-\delta)^{-1} \kappa \eta_n \end{aligned} \tag{17}$$

¿From the calculations above and (4), we deduce,

$$\sum_{l=1}^{p} (\alpha_l \quad - \quad z_l)^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \leq \sum_{l=1}^{p} ((KK^t)^{-1} K\varepsilon)_l^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} + c(1-\delta)^{-1} [\kappa\eta_n]^2 \tag{18}$$

Let us now recall the following inequality due to Pinelis [Pinelis, 1994], assuming that the ξ_i's are Hilbert space valued, independent random variables, such that $\|\xi_i - \mathbb{E}(\xi_i)\| \leq \widetilde{M}$ and $\mathbb{E}\|\xi_i - \mathbb{E}(\xi_i)\|^2 \leq \sigma^2(\xi)$,

$$\mathsf{Prob}\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\| \geq \lambda\big) \leq 2\exp\{\frac{-n\lambda^2}{2(\lambda\widetilde{M}/3 + \sigma^2(\xi))}\} \tag{19}$$

Now as $\sigma^2(\xi) \leq \widetilde{M}^2$, replacing $\sigma^2(\xi)$ in the RHS, we get :

$$\mathsf{Prob}\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\| \geq \lambda\big) \leq 2\exp\{\frac{-n\lambda^2}{2(\lambda\widetilde{M}/3 + \widetilde{M}^2)}\} \tag{20}$$

As only $\lambda \leq \widetilde{M}$ is significant, since $\mathsf{Prob}\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\| \geq \lambda\big) = 0$, for $\lambda > \widetilde{M}$,

$$\mathsf{Prob}\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\| \geq \lambda\big) \leq 2\exp\ \frac{-3n\lambda^2}{8\widetilde{M}^2} \tag{21}$$

Let us now take $\xi_i \in \mathbb{R}^p$:

$$(\xi_i)_l = (K(x_l, X_i)\varepsilon_i)_l$$

in such a way that,

$$\sum_i \xi_i = K\varepsilon$$

and the ξ_i are independent. It is easy to verify that $\mathbb{E}(\xi_i) = 0$.
Let us for all $U \in \mathbb{R}^p$ define the following Hilbertian norm :

$$\|U\|_A^2 = \sum_{l=1}^{p}(n(KK^t)^{-1}U)_l^2\mathbb{I}\ \{|\alpha_l| \geq \lambda_n/2\} = \sum_{l=1}^{p}((M_{np})^{-1}U)_l^2\mathbb{I}\ \{|\alpha_l| \geq \lambda_n/2\}$$

Then,

$$\sum_{l=1}^{p}((KK^t)^{-1}K\varepsilon)_l^2\mathbb{I}\ \{|\alpha_l| \geq \lambda_n/2\} = \|\frac{1}{n}\sum_i \xi_i\|_A^2$$

Now, we have using (5)

$$\begin{aligned}
\|\xi_i\|_A^2 &= \sum_{l=1}^{p}(M_{np}^{-1}\xi_i)_l^2\mathbb{I}\ \{|\alpha_l| \geq \lambda_n/2\} \\
&\leq p^*(\sup_l(M_{np}^{-1}\xi_i)_l)^2 \\
&\leq p^*(\sup_l(K(x_l, X_i)\varepsilon_i)^2\frac{1}{(1-\delta)^2} \\
&\leq p^*(\kappa\varepsilon_i)^2\frac{1}{(1-\delta)^2} \leq p^*\frac{(M\kappa)^2}{(1-\delta)^2}
\end{aligned}$$

Now, using (21),

$$\mathsf{Prob}\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\|^2 \geq \frac{(a\eta)^2}{(1-\delta)^2}\big) \leq 2\exp -\{\frac{3}{8}n\eta^2\frac{a^2}{p^*M^2\kappa^2}\}$$

So for $a > 0$ suitably chosen, and taking account that $\eta > \eta_n$

$$\begin{aligned}
\rho^{\otimes n} & \Big(\quad \sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \geq \frac{(2a\eta)^2}{(1-\delta)^2}\Big) \\
& \leq \rho^{\otimes n}\Big(\sum_{l=1}^{p} ((KK^t)^{-1}K\varepsilon)_l^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} + [c\kappa(1-\delta)^{-1}\eta_n]^2 \geq \frac{(2a\eta)^2}{(1-\delta)^2}\Big) \\
& \leq \rho^{\otimes n}\Big(\sum_{l=1}^{p} ((KK^t)^{-1}K\varepsilon)_l^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \geq \frac{(a\eta)^2}{(1-\delta)^2}\Big) \leq 2\exp -\{\frac{3}{8} n\eta^2 \frac{a^2}{p^* M^2 \kappa^2}\}
\end{aligned}$$

Now, if we recall that $\eta_n = (\frac{\log n}{n})^{s/(1+2s)}$; $p^{-1} = \sqrt{t_n}$; $\lambda_n = T\sqrt{t_n}$ and $p^* \leq 4c(Tt_n)^{\frac{-1}{1+2s}} \wedge p$, evaluation at the point $\eta = \eta_n$ gives :

$$2\exp -\{\frac{3}{8} n\eta_n^2 \frac{c^2}{p^* M^2\kappa^2}\} = 2\exp -\{\frac{3}{8} \log n \frac{a^2}{c(\frac{1}{T})^{2/1+2s} M^2 \kappa^2}\}.$$

Hence

$$\rho^{\otimes n}(\sum_l [\frac{1}{n} \sum_i \varepsilon_i K(x_l, X_i)]^2 \geq \eta^2/2) \leq \exp -C[n\eta^2 p^{-1} \vee \log n]$$

3.2.2. *Study of* $\sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|z_l| \geq \lambda_n\}\mathbb{I}\{|\alpha_l| < \lambda_n/2\}$. It remains now to study the term :

$$BS = \sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|z_l| \geq \lambda_n\}\mathbb{I}\{|\alpha_l| < \lambda_n/2\} \leq \sum_{l=1}^{p} (\alpha_l - z_l)^2 \mathbb{I}\{|z_l - \alpha_l| \geq \lambda_n/2\}$$

Using the previous result with p instead of p^*, we get,

$$\begin{aligned}
\rho^{\otimes n} & \Big(\quad BS \geq \frac{(2a\eta)^2}{(1-\delta)^2}\Big) \\
& \leq \rho^{\otimes n}\Big(\sum_{l=1}^{p} (\alpha_l - z_l)^2 \geq \frac{(2a\eta)^2}{(1-\delta)^2}\Big) \\
& \leq \rho^{\otimes n}\Big(\sum_{l=1}^{p} ((KK^t)^{-1}K\varepsilon)_l^2 \geq \frac{(a\eta)^2}{(1-\delta)^2}\Big) \\
& \leq 2\exp -\{\frac{3}{8} n\eta^2 \frac{a^2}{p M^2\kappa^2}\}
\end{aligned}$$

We proceed as in the previous section, and obtain using (5) :

$$\begin{aligned}
|\alpha_l - z_l| & \leq |[(KK^t)^{-1}K\varepsilon]_l| + \|(KK^t)^{-1}K[\bar{f}_X - f_X]\|_{l_\infty} \\
& \leq |\frac{1}{n}({M_{np}}^{-1}K\varepsilon)_l| + (1-\delta)^{-1}\kappa p^{-3/2} \\
& \leq (1-\delta)^{-1}\|\frac{1}{n}K\varepsilon\|_{l_\infty} + (1-\delta)^{-1}\sqrt{t_n} M\kappa
\end{aligned}$$

So

$$\rho^{\otimes n}(\exists l \in \{1, \dots, p\},\ |\alpha_l - z_l| \geq \lambda_n) \leq \rho^{\otimes n}((1-\delta)^{-1} \sup_l |\frac{1}{n}(K\varepsilon)_l| + (1-\delta)^{-1}\kappa p^{-3/2} \geq \lambda_n)$$

$$\leq \sum_{l=1}^{p} \rho^{\otimes n}((1-\delta)^{-1}|\frac{1}{n}\sum_{i} K(x_l, X_i)\varepsilon_i| + (1-\delta)^{-1}\kappa p^{-3/2} \geq T\sqrt{t_n})$$

$$\leq \sum_{l=1}^{p} \rho^{\otimes n}(|\frac{1}{n}\sum_{i} K(x_l, X_i)\varepsilon_i| \geq \sqrt{\log n/n}(\frac{T}{1-\delta} - \kappa(\frac{\log n}{n})^{1/4})$$

But for n large enough

$$(\frac{T}{1-\delta} - \kappa(\frac{\log n}{n})^{1/4}) \geq \frac{T}{2(1-\delta)}$$

and using Hoeffding inequality

$$\rho^{\otimes n}(|\frac{1}{n}\sum_{i} K(x_l, X_i)\varepsilon_i| \geq \frac{T}{2(1-\delta)\sqrt{\log n/n}}) \leq 2\exp - \frac{T^2 \log n}{8(1-\delta)^2\kappa^2 M^2}$$

So

$$\rho^{\otimes n}(\exists l \in \ \{1, \dots, p\},\ |\alpha_l - z_l| \geq \lambda_n) \leq 2p\exp - \frac{T^2 \log n}{8(1-\delta)^2\kappa^2 M^2} \leq Cn^{-\alpha}$$

with $\alpha > 0$ if T is large enough.

So :

$$\rho^{\otimes n}(BS \geq \frac{a\eta}{1-\delta}) \leq 2\exp - \{\frac{3}{8}n\eta^2 \frac{a^2}{pM^2\kappa^2}\} \wedge n^{-\alpha}$$

This yields the results.

4. Wavelet results

4.1. Assumptions and estimation rules.

4.1.1. *Assumptions on the model.* In this section, we will concentrate on the case of dimension 1 : the random variables X_i's are now taking their values in $\mathbb{X}$ = compact domain of $\mathbb{R}$. This case can easily be generalized to the case where the measure ρ_X is a tensor product of measures ρ_{X_i}, $i = 1, \dots, d$. However the full generalization to dimension d is more involved and will not be discussed in this paper. In the case $d = 1$, we define the distribution function G such that

$$\forall t \in \mathbb{R}, \quad G(t) = \rho(X \leq t) \in [0, 1]$$

and assume that it is a derivable function. We also define,

$$\forall x \in [0, 1], \quad G^{-1}(x) = \inf\{t \in \mathbb{R},\ G(t) \geq x\}.$$

Again, we will assume that f_ρ has sparsity conditions which can be in this case directly expressed in terms of regularity conditions. More precisely, we will denote by $\mathcal{M}(\Theta_s)$, the set of measures ρ verifying all the assumptions above with in addition the fact that $f_\rho(G^{-1}) \in B^s_\infty(L_\infty([0, 1]))(M)$ (the ball of radius M of the Besov space). Notice that as we will only consider the case where $s > 0$ (in fact $s > 1/2$) f_ρ will always be bounded by M.

Let us consider $\{\psi_{j,k}, j \geq \underline{j} + 1,\ 0 \leq k < 2^j\}$ a wavelet basis on $[0, 1]$ (at least continuously differentiable, with enough moment conditions ; the length of of the support of $\psi_{j,k}$ the will be supposed to be less than $N2^{-j}$). We recall that : $\psi_{\underline{j},k} = \varphi_{\underline{j}k}$ denotes the scaling function. These assumptions are standard (see [Cohen et al., 1993]).

Let us expand f in the wavelet basis :

$$f(G^{-1}) = \sum_{j=\underline{j}}^{\infty} \sum_{k\in} \beta_{j,k}\psi_{j,k}.$$

and it is well known that for $0 \leq \gamma < \infty$, $f(G^{-1})$ belongs to $B^{\gamma}_{\infty}(L_{\infty}([0,1]))$ iff (and we will take this as the $B^{\gamma}_{\infty}(L_{\infty}([0,1]))$—norm) :

$$\sup_{j\geq\underline{j}} 2^{j(\gamma+\frac{1}{2})} \sup_{0\leq k<2^j} |\beta_{j,k}| =: \|f\|_{B^{\gamma}_{\infty}} < \infty.$$

In this section our loss will be measured in term of $\mathbb{L}_2$, with respect to the measure $d\rho_X$:

$$\|f\|_{\rho_X} = [\int f(x)^2 d\rho_X(x)]^{\frac{1}{2}}.$$

4.1.2. *Estimation Algorithm :* Again, we put

$$t_n := \frac{\log n}{n}, \quad \lambda_n = \kappa\sqrt{t_n},$$

define :

$$\hat{G}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I\{X_i \leq x\},$$

and let us introduce the ordered statistic : $X_{(1)} \leq \ldots \leq X_{(n)}$. Doing this, we introduce a new ordering on the indices $\{1,\ldots,n\}$. Keeping this ordering, we denote $Y_{(1)},\ldots,Y_{(n)}$. Note that $Y_{(1)},\ldots Y_{(n)}$ is generally not the ordered statistic of $Y_1,\ldots Y_n$.
The estimator is constructed in the following way :

- Step 1 : Estimation of the wavelet coefficients :

$$\hat{\beta}_{jk} = \frac{1}{n}\sum_{i=1}^{n} Y_i\psi_{jk}(\hat{G}_n(\frac{i}{n})) = \frac{1}{n}\sum_{i=1}^{n} Y_{(i)}\psi_{jk}(X_{(i)})$$

- Step 2 : Thresholding

$$\tilde{\beta}_{jk} = \hat{\beta}_{jk}\mathbb{I}\{|\hat{\beta}_{jk}| \geq \lambda_n\}$$

- Step 3 : Reconstruction

$$\hat{f} = \sum_{j=\underline{j}}^{J} \sum_{k} \tilde{\beta}_{jk}\psi_{jk}(\hat{G}_n)$$

Note that this algorithm is an adaptation of the standard wavelet algorithm introduced in [Donoho and Johnstone, 1994] in the case of an equispaced design. It has been investigated in [Kerkyacharian and Picard, 2004], where the expectation properties of the $L_p(dx)$ losses have investigated (instead of here the deviation properties of the $\mathbb{L}_2 d\rho_X$). It proves to have very powerful properties. One of them is its remarkable simplicity in terms of computation. To illustrate this, we give here the main steps of the computation algorithm :
Algorithm :

(1) *Sort the X_i's,*

(2) *Change the numbering in such a way that X_i has rank i,*

(3) Calculate the highest level alpha*-coefficients using the formula :*

$$\hat{\alpha}_{J'k} = \frac{1}{n}\sum_{i=1}^{n} \varphi_{J'k}(i/n)Y_i, \quad (2^{J'} = n)$$

(4) Calculate the wavelet coefficients using the classical pyramidal algorithm

(5) Perform a thresholding algorithm giving rise to $\widetilde{\beta}_{jk}$ *coefficients,*

(6) Reconstruct the estimator, using again the standard backward pyramidal algorithm, obtaining

$$\hat{f} = \sum_{j=\underline{j}}^{J} \sum_{0\le k<2^j} \widetilde{\beta}_{jk}\psi_{jk}(\hat{G}_n(x))$$

which is a function especially easy to draw.

Our aim in this section is to prove the following theorem.

Theorem 2. *With the conditions above,* $\forall s > \frac{1}{2}$*,*

$$\eta_n = [\frac{n}{\log n}]^{\frac{-s}{1+2s}},$$

there exist positive constants γ, T, D *such that,*

$$\sup_{\rho\in\mathcal{M}(\Theta_s)} \rho^{\otimes n}\{\|f_\rho - \hat{f}\| > \eta\} \le T\{ \begin{matrix} e^{-\gamma[n2^{-J}J^{-1}\eta^2 \vee \log n]}, & \eta \ge D\eta_n, \\ 1, & \eta \le D\eta_n, \end{matrix}$$

as long as

$$[\frac{n}{\log n}]^{\frac{1}{1+2s}} \le 2^J \le [\frac{n}{\log n}]^{\frac{1}{2}}$$

Remark 2. *As mentioned in the introduction these results are comparable to those obtained in the RKHS situation as concern the critical value and the exponential rates. The advantage here is that we are able to state the results in the* $L_2(\rho_X)$ *norm and the regularity conditions are expressed in terms of standard Hölder spaces. We expressed the results in a slightly different way, leaving the choice of* J*, as an option. If we optimize oour results in* J*, we take* $2^J = [\frac{n}{\log n}]^{\frac{1}{1+2s}}$ *which gives better rate results but fails in being adaptive. If we want our estimate to be universal (work for any* $s > 1/2$*) we need to take* $2^J \le [\frac{n}{\log n}]^{\frac{1}{2}}$*.*

4.2. **Proof of the theorem.** Throughout the proof, the constant c will denote a constant which may vary from one line to the other, but may be explicitely calculated. For a sake of simplicity we will not make explicite the constants obtained in the proof (although it could be done easily) since we do not think that they are optimal in any sense.

It will be essential in the sequel to notice that with the assumptions above, we have :

$$\|f\|_{L_2(\mathbb{X}\rho_X)} = \|f(G^{-1})\|_{L_2([0,1],dx)}.$$

Since $\|f\|_{\rho_X} = \|f(G^{-1})\|_{dx}$, we have if

$$f_\rho(G^{-1}) = \sum_{j,k} \beta_{jk}\psi_{jk}$$

$$\begin{aligned}\|\hat{f}-f_\rho\|_{\rho X} &= \|\hat{f}(G^{-1})-f_\rho(G^{-1})\|_{dx}\\ &= \|\sum_{j=\underline{j}}^{J}\sum_k \widetilde{\beta}_{jk}\psi_{jk}(\hat{G}_n(G^{-1}))-\sum_{j,k}\beta_{jk}\psi_{jk}\|_{dx}\\ &\leq \|\sum_{j=\underline{j}}^{J}\sum_k \widetilde{\beta}_{jk}[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]\|_{dx}+\|\sum_{j=\underline{j}}^{J}\sum_k[\widetilde{\beta}_{jk}-\beta_{jk}]\psi_{jk}\|_{dx}\\ &+ \|\sum_{j=\geq J+1}\sum_k \beta_{jk}\psi_{jk}\|_{dx}\end{aligned}$$

Hence

$$\begin{aligned}\|\hat{f}-f_\rho\|^2_{\rho X} &\leq 3[\|\sum_{j=\underline{j}}^{J}\sum_k \widetilde{\beta}_{jk}[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]\|^2_{dx}+\sum_{j=\underline{j}}^{J}\sum_k[\widetilde{\beta}_{jk}-\beta_{jk}]^2+\sum_{j\geq J+1}\sum_k\beta^2_{jk}]\\ &\leq (I)+(II)+(III)\end{aligned}$$

If $f_\rho(G^{-1})\in B^s_\infty(L_\infty([0,1]))(M)$, then

$$III=\sum_{j\geq J+1}\sum_k\beta^2_{jk}\leq\sum_{j\geq J+1}2^j\sup_k\beta^2_{jk}\leq M^2\sum_{j\geq J+1}2^j2^{-j(2s+1)}\leq M^22^{-2Js}\leq M^2\eta^2_n$$

if $2^J\geq t_n^{\frac{-1}{1+2s}}=(\sqrt{\frac{\log n}{n}})^{\frac{-1}{1+2s}}$

Let us now study the second term :

$$\begin{aligned}(II) &\leq \sum_{j=\underline{j}}^{J}\sum_k[\hat{\beta}_{jk}-\beta_{jk}]^2\mathbb{I}\{|\hat{\beta}_{jk}|\geq\lambda_n\}\,[\mathbb{I}\{|\beta_{jk}|\geq\lambda_n/2\}+\mathbb{I}\{|\beta_{jk}|<\lambda_n/2\}]\\ &+ \sum_{j=\underline{j}}^{J}\sum_k[\beta_{jk}]^2\mathbb{I}\{|\hat{\beta}_{jk}|<\lambda_n\}\,[\mathbb{I}\{|\beta_{jk}|\geq 2\lambda_n\}+\mathbb{I}\{|\beta_{jk}|<2\lambda_n\}]\\ &:= BB+BS+SB+SS\end{aligned}$$

Let us study the term SS. First we remark that, as $f_\rho(G^{-1})\in B^s_\infty(L_\infty([0,1]))(M)$, then $|\beta_{jk}|\leq M2^{-j(s+\frac{1}{2})}$, hence if we denote :

$$2^{j_s}=t_n^{\frac{-1}{1+2s}}$$

$$\begin{aligned} SS &\leq \sum_{j=\underline{j}}^{j_s} \sum_k [\beta_{jk}]^2 \mathbb{I}\{|\beta_{jk}| < 2\lambda_n\} + \sum_{j=j_s}^{J} \sum_k [\beta_{jk}]^2 \mathbb{I}\{|\beta_{jk}| < 2\lambda_n\} \\ &\leq \sum_{j=\underline{j}}^{j_s} \sum_k [2\lambda_n]^2 + \sum_{j=j_s}^{J} \sum_k [\beta_{jk}]^2 \\ &\leq 22^{j_s}(2\lambda_n)^2 + \sum_{j=j_s}^{J} 2^j M^2 2^{-2j(s+\frac{1}{2})} \\ &\leq (8\kappa^2 + 2M^2)\eta_n^2 \end{aligned}$$

Let us now investigate the term SB : We observe that

$$\mathbb{I}\{|\hat{\beta}_{jk}| < \lambda_n\}\mathbb{I}\{|\beta_{jk}| \geq 2\lambda_n\} \leq \mathbb{I}\{|\hat{\beta}_{jk} - \beta_{jk}| \geq |\beta_{jk}|/2\}\mathbb{I}\{|\beta_{jk}| \geq 2\lambda_n\}$$

, hence :

$$\begin{aligned} SB &\leq \sum_{j=\underline{j}}^{J} \sum_k [\beta_{jk}]^2 \mathbb{I}\{|\hat{\beta}_{jk} - \beta_{jk}| \geq |\beta_{jk}|/2\}\mathbb{I}\{|\beta_{jk}| \geq 2\lambda_n\} \\ &\leq 4\sum_{j=\underline{j}}^{J} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^2 \mathbb{I}\{|\beta_{jk}| \geq 2\lambda_n\} \end{aligned}$$

So

$$BB + SB \leq 5\sum_{j=\underline{j}}^{J} \sum_k |\hat{\beta}_{jk} - \beta_{jk}|^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} = 5BB'$$

Now, we investigate the term BB'.

If we recall that $X_{(1)} \leq \ldots \leq X_{(n)}$. Doing this we introduce a new ordering on the indices $\{1, \ldots, n\}$, and that we keep this ordering, to denote $Y_{(1)}, \ldots, Y_{(n)}$. We also introduce $U_i = G(X_i),\ i = 1, \ldots, n$, as well as the associated $U_{(1)}, \ldots, U_{(n)}$. Notice that the $U_{(i)}$'s are ordered (since G is increasing) and the U_i's are i.i.d. uniformly distributed.

$$\begin{aligned} \hat{\beta}_{jk} - \beta_{jk} &= \frac{1}{n}\sum_{i=1}^{n} Y_{(i)}\psi_{jk}(\frac{i}{n}) - \beta_{jk} \\ &= [\frac{1}{n}\sum_{i=1}^{n} f_\rho(G^{-1}(U_{(i)}))\psi_{jk}(\frac{i}{n}) - \int \psi_{jk} f_\rho(G^{-1})] + [\frac{1}{n}\sum_{i=1}^{n} \varepsilon_{(i)}\psi_{jk}(\frac{i}{n})] \\ &= [\frac{1}{n}\sum_{i=1}^{n} f_\rho(G^{-1}(U_{(i)}))\psi_{jk}(\frac{i}{n}) - \int \psi_{jk} f_\rho(G^{-1})] + [\frac{1}{n}\sum_{i=1}^{n} \varepsilon_i\psi_{jk}(\frac{i}{n})] \\ &:= A_{jk} + B_{jk} \end{aligned} \tag{22}$$

Let us begin by the following lemma which proof is obvious (but which will be useful in the sequel :

Lemma 1. *For any* $r \geq 1$*, we have*

$$\frac{1}{n}\sum_{i=1}^{n} |\psi_{jk}(\frac{i}{n})|^r \leq \tau_r 2^{j(\frac{r}{2}-1)} + \tau_r' \frac{2^{j(1+\frac{r}{2})}}{n} \tag{23}$$

with $\tau_r = N\|\psi\|_\infty$ *and* $\tau_r' = Nr\|\psi'\|_\infty(\|\psi\|_\infty)^{r-1}$

Let us put :

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\{U_i \leq x\}, \quad \Delta_n := \sup_{x\in[0,1]} |\hat{F}_n(x) - x|.$$

and $\bar{s} = s \wedge 1$, using (23) for the third inequality,

$$\begin{aligned}
|A_{jk}| &\leq \frac{1}{n}\sum_{i=1}^{n} |f_\rho(G^{-1}(U_{(i)})) - f_\rho(G^{-1}(\frac{i}{n}))||\psi_{jk}(\frac{i}{n})| \\
&+ \sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} |f_\rho(G^{-1}(x)\psi_{jk}(x) - f_\rho(G^{-1}(\frac{i}{n}))\psi_{jk}(\frac{i}{n})| \\
&\leq \Delta_n^{\bar{s}}\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}\frac{1}{n}\sum_{i=1}^{n}|\psi_{jk}(\frac{i}{n})| \\
&+ \sum_{i=1}^{n}\int_{(i-1)/n}^{i/n}[2^{j/2}\|\psi\|_\infty\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}n^{-\bar{s}} + \|\psi\|_{1\infty\infty}\|f_\rho(G^{-1})\|_\infty\frac{2^{3j/2}}{n}]\mathbb{I}\{x \in [\frac{k}{2^j}, \frac{k+N}{2^j}]\}dx \\
&\leq \Delta_n^{\bar{s}}\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}\{\tau_1 2^{-j/2} + \tau_1'\frac{2^{3j/2}}{n}\} \\
&+ N\|\psi\|_\infty\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}n^{-\bar{s}}2^{-\frac{j}{2}} + N\|\psi\|_{1\infty\infty}\|f_\rho(G^{-1})\|_\infty\frac{2^{j/2}}{n} \\
&\leq C_1\Delta_n^{\bar{s}}2^{-j/2} + C_2\frac{2^{j/2}}{n} + C_3 n^{-\bar{s}}2^{-\frac{j}{2}} \qquad (24)
\end{aligned}$$

where

$$C_1 = \tau_1 + \tau_1', \; C_2 = N\|\psi'\|_\infty, \; C_3 = N\|\psi\|_\infty\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}$$

The last line uses the fact that for $j \leq J$, $2^{2j} \leq n$. We can then state the following lemma :

Lemma 2. *For* J *such that* $t_n^{\frac{-1}{1+2s}} \leq 2^J \leq t_n^{-1/2}$*, we have :*

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_k A_{jk}^2 \geq \eta^2) \leq \exp -Cn2^{-J}\eta^2 \vee \log n, \tag{25}$$

for all $\eta \geq D\eta_n$*, where* $C = 2(2C_1^2N)^{\frac{-1}{\bar{s}}}$

Proof of the lemma :

We observe that

$$\sum_{j=\underline{j}}^{J}\sum_k[\frac{2^{j/2}}{n}]^2 \leq c\frac{2^{2J}}{n^2} \leq c\frac{1}{n} << \eta_n^2.$$

$$\text{and} \quad \sum_{j=\underline{j}}^{J} \sum_k n^{-2\bar{s}} 2^{-j} \leq J n^{-2\bar{s}} << \eta_n^2$$

.

$$\sum_{j=\underline{j}}^{J} \sum_k \Delta_n^{2\bar{s}} 2^{-j} \leq J \Delta_n^{2\bar{s}} \tag{26}$$

Hence, for $\eta \geq D\eta_n$, and n large enough,

$$\begin{aligned} \rho^{\otimes n}(\sum_{j=\underline{j}}^{J} \sum_k A_{jk}^2 \geq \eta^2) &\leq \rho^{\otimes n}(C_1^2 J \Delta_n^{2\bar{s}} \geq \eta^2/2) \\ &\leq K \exp -cn[\eta J^{-1/2}]^{\frac{2}{\bar{s}}} \mathbb{I}\{\eta^2 \leq 2C_1^2 J\} \end{aligned}$$

The last line uses the following Dvoreski, Kiefer and Wolfovitz bound (see for instance the review on the subject in Devroye Lugosi section 12.) : For any $\lambda > 0$, there exists a universal constant K, such that :

$$\mathbb{P}(\Delta_n \geq \lambda) \leq K \exp -2n\lambda^2 \tag{27}$$

(and noticing that $\Delta_n \leq 1$)
Now, for $s \geq 1$, $n[\eta]^{\frac{2}{\bar{s}}} J^{\frac{-1}{2\bar{s}}} = n\eta^2 J^{-1/2} \geq n\eta^2 2^{-J} \vee \log n$.
Identically, for $1/2 < s < 1$ and $\eta \geq D\eta_n$,

$$\begin{aligned} n[\eta]^{\frac{2}{s}} J^{\frac{-1}{2s}} &\geq n\eta^2 2^{-J} \eta_n^{2(\frac{1}{s}-1)} 2^J J^{\frac{-1}{2s}} \\ &\geq n\eta^2 2^{-J} 2^{-2sj_s(\frac{1}{s}-1)} 2^{j_s} J^{\frac{-1}{2s}} \\ &\geq n\eta^2 2^{-J} 2^{j_s(2s-1)} J^{\frac{-1}{2s}} \geq n\eta^2 2^{-J} \vee \log n \end{aligned}$$

This ends up the proof of the lemma. □

Let us now investigate the term corresponding to the B_{jk}'s. We have the following lemma :

Lemma 3. *For* J *such that* $t_n^{\frac{-1}{1+2s}} \leq 2^J \leq t_n^{-1/2}$, *there exists a constant* c *such that :*

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J} \sum_k B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2) \leq c \exp -cn2^{-J}\eta^2 \vee \log n, \tag{28}$$

for all $1 \geq \eta \geq D\eta_n$

Proof of the lemma :

Let us first remark that since $f_\rho(G^{-1}) \in B_\infty^s(L_\infty([0,1]))(M)$, then $|\beta_{jk}| \leq M2^{-j(s+\frac{1}{2})}$ and then, if $\kappa \geq 2M$, $|\beta_{jk}| \geq \lambda_n/2$ implies $j \leq j_s$, hence :

$$\begin{aligned} \sum_{j=\underline{j}}^{J} \sum_k B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} &\leq \sum_{j=\underline{j}}^{j_s} \sum_k B_{jk}^2 \\ &\leq \sum_{j=\underline{j}}^{j_s} 2^j \sup_k B_{jk}^2 \leq 2 2^{j_s} \sup_{jk} B_{jk}^2 \end{aligned}$$

We will investigate separately the cases $\eta \leq 1$, and $\eta \geq 1$. Let us begin with the fist case :

$$\begin{aligned}
\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_{k} B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2) &\leq \rho^{\otimes n}(2^{j_s+1}\sup_{jk} B_{jk}^2 \geq \eta^2) \\
&\leq \sum_{j=\underline{j}}^{J}\sum_{k} \rho^{\otimes n}(|\sum_{i=1}^{n}\psi_{jk}(\frac{i}{n})\varepsilon_i| \geq n\eta 2^{-j_s/2}/\sqrt{2}) \\
&\leq \sum_{j=\underline{j}}^{J}\sum_{k} \exp\{\frac{-n^2\eta^2 2^{-j_s}/2}{2(nC_3 + n\eta M\|\psi\|_\infty 2^{(j-j_s)/2}/3)}\} \\
&\leq 2^{j_s+1}\exp\{\frac{-n\eta^2 2^{-j_s}}{4(C_3+\eta M\|\psi\|_\infty/3)}\} \qquad (29)
\end{aligned}$$

In the last line we used Bernstein inequality (cf Bernstein [Bernstein, 1946]), since the variables $\psi_{jk}(\frac{i}{n})\varepsilon_i$ are a sequence of independent bounded random variables (by $M\|\psi\|_\infty 2^{\frac{j}{2}}$), with zero mean and

$$\mathbb{E}[\sum_i \psi_{jk}(\frac{i}{n})\varepsilon_i]^2 \leq C_3 n$$

$(M^2(\tau_2 + \tau_2') := C_3$ using (23).)
Hence we obtain :

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_{k} B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2) \leq 2\exp\{-cn\eta^2 2^{-j_s} + \frac{1}{2}\log n\} \qquad (30)$$

with $c = 4(C_3 + M\|\psi\|_\infty/3)^{-1}$ since $\eta \leq 1$. As $\eta \geq D\eta_n$, it is easy to see that for D large enough, $cn\eta^2 2^{-j_s} \geq 2\log n$. Hence in this case, we get the bound : $\exp -cn2^{-J}\eta^2 \vee \log n$

Let us now study the case where $\eta \geq 1$, we'll use Mac Diarmid's inequality (see [Diarmid, 1989] we have the following lemma :

Lemma 4. *For* J *such that* $t_n^{\frac{-1}{1+2s}} \leq 2^J \leq t_n^{-1/2}$, *we have :*

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_{k} B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2) \leq \exp -Cn2^{-J}\eta^2 \vee \log n, \qquad (31)$$

for all $\eta \geq 1$, *and* $C = \frac{1}{2B^2}$, $B^2 = 2M^2N^2\|\psi\|_\infty$.

Proof of the lemma : We have :

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_{k} B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2) \leq \rho^{\otimes n}(F(\varepsilon_1,\ldots,\varepsilon_n) \geq \eta^2)$$

with :

$$F(\varepsilon_1,\ldots,\varepsilon_l,\ldots,\varepsilon_n) = \sum_{j=\underline{j}}^{j_s}\sum_{k}\frac{1}{n^2}[\sum_{i=1}^{n}\psi_{jk}(\frac{i}{n})\varepsilon_i]^2$$

$$
\begin{aligned}
|\Delta F_l| &= |F(\varepsilon_1,\dots,\varepsilon_l,\dots,\varepsilon_n) - F(\varepsilon_1,\dots,\varepsilon'_l,\dots,\varepsilon_n)| \\
&= \sum_{j=\underline{j}}^{j_s} \sum_k \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \frac{1}{n^2} \left([\sum_{i=1}^n \psi_{jk}(\frac{i}{n})\varepsilon_i]^2 - [\sum_{i=1}^n \psi_{jk}(\frac{i}{n})\varepsilon_i + \psi_{jk}(\frac{l}{n})(\varepsilon'_l - \varepsilon_l)]^2 \right) \\
&\leq 2M^2 \sum_{j=\underline{j}}^{j_s} \sum_{k,\, |\frac{l}{n}-\frac{k}{2^j}| \leq \frac{N}{2^j}} \frac{1}{n^2} \sum_{i=1}^n |\psi_{jk}(\frac{i}{n})||\psi_{jk}(\frac{l}{n})| \\
&\leq 2M^2N^2 \frac{1}{n^2} \|\psi\|_\infty^2 \sum_{j=\underline{j}}^{j_s} 2^j \frac{n}{2^j} \\
&\leq 2M^2N^2 \|\psi\|_\infty^2 \frac{J}{n} =: B^2 \frac{j_s}{n}
\end{aligned}
$$

On the other hand,

$$
\begin{aligned}
\mathbb{E}_{\rho^{\otimes n}} F(\varepsilon_1,\dots,\varepsilon_n) &\leq \sum_{j=\underline{j}}^{j_s} \sum_k \frac{1}{n^2} [\sum_{i=1}^n \psi_{jk}(\frac{i}{n})\varepsilon_i]^2 \\
&\leq \sum_{j=\underline{j}}^{j_s} \sum_k \frac{1}{n^2} \sum_{i=1}^n \psi_{jk}(\frac{i}{n})^2 M^2 \\
&\leq M^2 C_3 \sum_{j=\underline{j}}^{j_s} \sum_k \frac{1}{n} \\
&\leq 2M^2 C_3 \frac{2^{j_s}}{n} \leq c\eta_n^2
\end{aligned}
$$

Hence, for $\eta \geq D\eta_n$,

$$
\begin{aligned}
\rho^{\otimes n}(\sum_{j=\underline{j}}^{J} \sum_k B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2) &\leq \rho^{\otimes n}(|F(\varepsilon_1,\dots,\varepsilon_n) - \mathbb{E}_{\rho^{\otimes n}} F(\varepsilon_1,\dots,\varepsilon_n)| \geq \eta^2/2) \\
&\leq \exp \frac{-2\eta^4}{4n(\frac{BJ}{n})^2} \leq \exp -nC\frac{\eta^4}{J^2}
\end{aligned}
$$

Now, for $\eta \geq 1$, we obviously have $Cn\frac{\eta^4}{J^2} \geq Cn2^{-J}\eta^2 \vee \log n$, which proves the result of the lemma. □

Notice also that, using exactly the same proof, we have also the following result, which will be used later :

Lemma 5. *For* J *such that* $t_n^{\frac{-1}{1+2s}} \leq 2^J \leq t_n^{-1/2}$, *we have :*

$$
\rho^{\otimes n}(\sum_{j=\underline{j}}^{J} \sum_k B_{jk}^2 \geq \lambda^2) \leq \exp -Cn\lambda^4/J^2 \vee \log n, \tag{32}
$$

for all $\lambda^2 \geq 2M^2C_1t_n^{1/2}$, $C = \frac{1}{2B^2}$, $B^2 = 2M^2N^2\|\psi\|_\infty$.

This achieves bounding the term (BB). We now proceed to bound the term (BS) :

$$\begin{aligned}\sum_{j=\underline{j}}^{J}\sum_{k}(\hat{\beta}_{jk}-\beta_{jk})^2\mathbb{I}\{|\beta_{jk}|<\lambda_n/2\}\mathbb{I}\{|\hat{\beta}_{jk}-\beta_{jk}|\geq\lambda_n/2\} &\leq \sum_{j=\underline{j}}^{J}\sum_{k}(\hat{\beta}_{jk}-\beta_{jk})^2\mathbb{I}\{|\hat{\beta}_{jk}-\beta_{jk}|\geq\lambda_n/2\}\\ &\leq 2^{J+1}\sup_{jk}\{(\hat{\beta}_{jk}-\beta_{jk})^2;\ |\hat{\beta}_{jk}-\beta_{jk}|\geq\lambda_n/2\}\end{aligned}$$

Hence

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_{k}(\hat{\beta}_{jk}-\beta_{jk})^2\mathbb{I}\{|\hat{\beta}_{jk}-\beta_{jk}|\geq\lambda_n/2\}\geq\eta^2\}) \quad\leq\quad 2^{J+1}\rho^{\otimes n}(|\hat{\beta}_{jk}-\beta_{jk}|\geq\eta 2^{-J/2}/2\vee\lambda_n/2)$$

Now, using (22) and (26), we get

$$\begin{aligned}2^{J+1}\rho^{\otimes n}(|\hat{\beta}_{jk}-\beta_{jk}|\geq\eta 2^{-J/2}/2 \quad\vee&\quad \lambda_n/2)\leq 2^{J+1}\rho^{\otimes n}(C_1\Delta_n^{\bar{s}}2^{-J/2}\geq\eta 2^{-J/2}/4\vee\lambda_n/8)\\ +&\quad 2^{J+1}\rho^{\otimes n}(|B_{jk}|\geq\eta 4^{-J/2}/2\vee\lambda_n/8)\\ \leq&\quad 2^{J+1}K\exp-cn(\frac{\eta}{4})^{\frac{2}{\bar{s}}}\mathbb{I}\{\eta\leq 4C_1\}\\ +&\quad 2^{J+1}\exp\{\frac{-n(\eta^2 2^{-J}/16\vee\lambda_n^2/64)}{2C_3+(\eta/4\vee\lambda_n/8)2^{J/2})M\|\psi\|_\infty}\}\end{aligned}$$

The first term may be bounded as in Lemma 3, the second one may be bounded by : $\exp-c[n2^J\eta^2\vee\log n]$, with $c=(64C_3)^{-1}$ if $\eta\leq 1$.
For $\eta\geq 1$, we have :

$$\sum_{j=\underline{j}}^{J}\sum_{k}(\hat{\beta}_{jk}-\beta_{jk})^2\mathbb{I}\{|\beta_{jk}|<\lambda_n/2\} \quad\leq\quad 2^{J+1}\sup_{jk}A_{jk}^2+\sum_{j=\underline{j}}^{J}\sum_{k}B_{jk}^2$$

Hence,

$$\begin{aligned}\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_{k}(\hat{\beta}_{jk}-\beta_{jk})^2\mathbb{I}\{|\beta_{jk}|<\lambda_n/2\}\geq\eta^2) &\leq \rho^{\otimes n}(2^{J+1}\sup_{jk}A_{jk}^2\geq\eta^2/2)\\ &+ \rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_{k}B_{jk}^2\geq\eta^2/2\})\end{aligned}$$

The first term, treated as above, gives the same bound since in this case the condition $\eta\leq 1$ was not necessary. For the second term, we use the lemma (5).

This achieves the proof for the term (II), which can be summarised in the following proposition.

Proposition 1. $\forall s>\frac{1}{2}$

$$\sup_{\rho\in\mathcal{M}(\Theta_s)}\rho^{\otimes n}\{\sum_{j=\underline{j}}^{J}\sum_{k}[\tilde{\beta}_{jk}-\beta_{jk}]^2>t^2\}\leq c\{\begin{array}{ll} e^{-Cn2^{-J}t^2\vee\log n}, & t\geq\eta_n,\\ 1, & t\leq\eta_n,\end{array}$$

if $[\frac{n}{\log n}]^{\frac{1}{1+2s}}\leq 2^J\leq[\frac{n}{\log n}]^{\frac{1}{2}}$ *for* $C=(64C_3)^{-1}\wedge(2B^2)^{-1}\wedge 2(\frac{2}{C_1N})^{\frac{1}{\bar{s}}}$

It remains, now to study the term (I).

We have :

$$\begin{aligned}
\|\sum_{j=\underline{j}}^{J}\sum_{k}\widetilde{\beta}_{jk}[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]\|_{dx} &\leq \|\sum_{j=\underline{j}}^{J}\sum_{k}|\widetilde{\beta}_{jk}-\beta_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx}\\
&+ \|\sum_{j=\underline{j}}^{J}\sum_{k}|\beta_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx}\\
&\leq \|\sum_{j=\underline{j}}^{J}\sum_{k}|\hat{\beta}_{jk}-\beta_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx}\\
&+ 2\|\sum_{j=\underline{j}}^{J}\sum_{k}|\beta_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx}
\end{aligned}$$

since $|\widetilde{\beta}_{jk}-\beta_{jk}| \leq |\hat{\beta}_{jk}-\beta_{jk}|+|\beta_{jk}|$. If $Z=\sum|\beta_{jk}|\psi_{jk}$ we observe that $\|Z\|_{s\infty\infty} = \|f_\rho(G^{-1})\|_{s\infty\infty}$, so :

$$\begin{aligned}
\|\sum_{j=\underline{j}}^{J}\sum_{k}|\beta_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx} &= \|Z(\hat{G}_n(G^{-1})-Z\|_{dx}\\
&\leq \|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}\Delta_n^{\bar{s}}
\end{aligned}$$

Hence,

$$\begin{aligned}
\rho^{\otimes n}(\|\sum_{j=\underline{j}}^{J}\sum_{k}|\beta_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx}\geq\eta) &\leq \rho^{\otimes n}(\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}\Delta_n^{\bar{s}}\geq\eta)\\
&\leq K\exp -cn\eta^{\frac{2}{\bar{s}}}\mathbb{I}\{\eta/\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}\leq 1\}\\
&\leq K\exp\{-cn\eta^2 2^{-J}\vee\log n\}
\end{aligned}$$

As above (see the proof of lemma 3), with $c=2\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}^{\frac{2}{\bar{s}}}$, here.

Concerning the stochastic term, using (22) we have :

$$\begin{aligned}
\|\sum_{j=\underline{j}}^{J}\sum_{k}|\hat{\beta}_{jk}-\beta_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx} &\leq \|\sum_{j=\underline{j}}^{J}\sum_{k}|A_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx}\\
&+ \|\sum_{j=\underline{j}}^{J}\sum_{k}|B_{jk}|[[\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}]]\|_{dx}
\end{aligned}$$

Now, if $Z' = \sum_{j=\underline{j}}^{J} \sum_k |A_{jk}|\psi_{jk}$, using (24), and $s > \frac{1}{2}$,

$$\begin{aligned} \|Z'\|_{1/2\infty\infty} &\leq \sup_{j\leq J,k}\{2^j|A_{jk}|\} \\ &\leq \sup_{j\leq J,k}\{C_1\Delta_n^{\bar{s}}2^{j/2} + C_2\frac{2^{3j/2}}{n} + C_3n^{-\bar{s}}\} \\ &\leq C_1\Delta_n^{\bar{s}}2^{J/2} + (C_2+C_3)\frac{2^{3J/2}}{n} \end{aligned}$$

Let us investigate separately the two contributions : As above,

$$\rho^{\otimes n}(\|\sum_{j=\underline{j}}^{J}\sum_k|\hat{\beta}_{jk} - \beta_{jk}|[|\psi_{jk}(\hat{G}_n(G^{-1})) - \psi_{jk}]]\|_{dx} \geq \eta) \quad \leq \quad \rho^{\otimes n}(\|Z'\|_{1/2\infty\infty}\Delta_n^{1/2} \geq \eta)$$

Furthermore,

$$\rho^{\otimes n}(\Delta_n^{1/2+\bar{s}}2^{J/2} \geq \eta/(2C_1)) \leq \exp\{-2n(\frac{\eta}{2C_1}2^{-J/2})^{\frac{2}{\bar{s}+1/2}}\}\mathbb{I}\{\frac{\eta}{2C_1}2^{-J/2} \leq 1\}$$

Now, as $\bar{s} > 1/2$, we have, for $\eta 2^{-J/2} \leq 2C_1$, $n(\eta 2^{-J/2})^{\frac{2}{\bar{s}+1/2}} \geq (2C_1)^{\frac{1-2\bar{s}}{1+\bar{s}/2}}n(\eta 2^{-J/2})^2 \vee \log n$, for $\eta \geq \eta_n$. On the other hand, for $\widetilde{C} = C_2 + C_3$

$$\rho^{\otimes n}(\widetilde{C}\frac{2^{3J/2}}{n}\Delta_n^{1/2} \geq \eta) \leq \exp\{-n2(\widetilde{C})^{-4}(n\eta 2^{-3J/2})^4\}\mathbb{I}\{n\eta 2^{-3J/2} \leq \widetilde{C}\}$$

And obviously, on the range we are considering $n(n\eta 2^{-3J/2})^4 \geq n(\eta 2^{-J/2})^2 \vee \log n$.

Now for the last term, $(\|\sum_{j=\underline{j}}^{J}\sum_k|B_{jk}|[|\psi_{jk}(\hat{G}_n(G^{-1})) - \psi_{jk}]]\|_{dx})$, considering again the $U_{(i)}$'s and putting $U_{(0)} = 0$, $U_{(n+1)} = 1$, we have, on $[U_{(i)}, U_{(i+1)}]$, $\hat{G}(G^{-1}(x)) = \frac{i}{n}$. For any arbitrary $a > 0$, we have

$$\begin{aligned} \|\sum_{j=\underline{j}}^{J}\sum_k \|[[\psi_{jk}(\hat{G}_n(G^{-1})) - \psi_{jk}]]\|_{dx}^2 &\leq [\sum_{j=\underline{j}}^{J}\|\sum_k|B_{jk}|[|\psi_{jk}(\hat{G}_n(G^{-1})) - \psi_{jk}|]\|_{dx}]^2 \\ &\leq 2^{(J+1)a}\sum_{j=\underline{j}}^{J}2^{-ja}\int[\sum_k|B_{jk}||\psi_{jk}(\hat{G}_n(G^{-1})) - \psi_{jk}|]^2 \\ &\leq 2^{(J+1)a}\sum_{j=\underline{j}}^{J}2^{-ja}\sum_{i=0}^{n}\int_{U_{(i)}}^{U_{(i+1)}}[\sum_k|B_{jk}||\psi_{jk}(\frac{i}{n}) - \psi_{jk}(x)|]^2dx \end{aligned}$$

Now, we will distinguish two cases : either $\frac{i}{n} \in [U_{(i)} - \frac{N}{2^j}, U_{(i+1)} + \frac{N}{2^j}]$ (case I) or not (case II, which implies that $\Delta_n 2^j \geq N$).

In case I, if we denote by $\Delta_{n,i} = \sup\{|\frac{i}{n} - U_{(i)}|, |\frac{i}{n} - U_{(i+1)}|\}$, and I_{jk} is the support of ψ_{jk} ,as ψ is continuously differentiable, we get, for $x \in [U_{(i)}, U_{(i+1)}]$:

$[\sum_k|B_{jk}||\psi_{jk}(\frac{i}{n}) - \psi_{jk}(x)|]]^2 \leq [\sum_k|B_{jk}|\|\psi'\|_\infty 2^{3j/2}\Delta_{n,i}\mathbb{I}_{I_{jk}}(x)]^2 \leq N\sum_k|B_{jk}|^2\|\psi'\|_\infty^2 2^{3j}\Delta_{n,i}^2\mathbb{I}_{I_{jk}}(x)$

The last inequality is true because only a finite number of $\mathbb{I}_{I_{jk}}(x)$'s are not zero at the same time.

If we now remark that in case I, $\Delta_{n,i} \leq 2N2^{-j} \wedge \Delta_n$ we get, for $x \in [U_{(i)}, U_{(i+1)}]$:

$[\sum_k |B_{jk}||\psi_{jk}(\frac{i}{n}) - \psi_{jk}]]^2(x) \leq 2N^2 \sum_k |B_{jk}|^2 \|\psi'\|_\infty 2^{2j} \Delta_n \mathbb{I}_{I_{jk}}(x).$

In case II, we get, for $x \in [U_{(i)}, U_{(i+1)}]$, using again the fact that only a finite number of ψ_{jk}'s are not zero at the same time :

$$\begin{aligned}
[\sum_k |B_{jk}||\psi_{jk}(\frac{i}{n}) - \psi_{jk}(x)|]^2 &\leq 2\left\{[\sum_k |B_{jk}||\psi_{jk}(\frac{i}{n})|]^2 + [\sum_k |B_{jk}||\psi_{jk}(x)|]^2\right\} \mathbb{I}\{\Delta_n 2^j \geq N\} \\
&\leq 2\left[N\|\psi\|_\infty^2 2^j \sup_{j \leq J,\, k} B_{jk}^2 + [\sum_k |B_{jk}||\psi_{jk}(x)|]^2\right] \mathbb{I}\{\Delta_n 2^j \geq N\}
\end{aligned}$$

Putting the two cases together, we deduce :

$$\begin{aligned}
\| \sum_{j=\underline{j}}^{J} \quad \sum_k \quad |B_{jk}|[|\psi_{jk}(\hat{G}_n(G^{-1})) - \psi_{jk}|]\|_{dx}^2 &\leq c2^{Ja} \sum_{j=\underline{j}}^{J} 2^{-ja} \sum_{i=0}^{n} \int_{U_{(i)}}^{U_{(i+1)}} [\sum_k |B_{jk}||\psi_{jk}(\frac{i}{n}) - \psi_{jk}(x)|]^2 dx \\
\leq \quad & c2^{Ja} \sum_{j=\underline{j}}^{J} 2^{-ja} \sum_{i=0}^{n} \int_{U_{(i)}}^{U_{(i+1)}} \left\{ N^2 \sum_k |B_{jk}|^2 \|\psi'\|_\infty 2^{2j} \Delta_n \right. \\
+ \quad & \left. \left[N\|\psi\|_\infty^2 2^j \sup_{j \leq J,\, k} B_{jk}^2 + [\sum_k |B_{jk}||\psi_{jk}(x)|]^2\right] \mathbb{I}\{\Delta_n 2^j \geq N\} \right\} dx \\
\leq \quad & c2^{Ja} \sum_{j=\underline{j}}^{J} 2^{-ja} \left\{ N^2 \sum_k |B_{jk}|^2 \|\psi'\|_\infty 2^j \Delta_n \right. \\
+ \quad & \left. \left[N\|\psi\|_\infty^2 2^j \sup_{j \leq J,\, k} B_{jk}^2 + \sum_k |B_{jk}|^2 N^{-1} \Delta_n 2^j\right] \mathbb{I}\{\Delta_n 2^j \geq N\} \right\} \\
\leq \quad & c\left[\sum_{j=\underline{j}}^{J} \sum_k |B_{jk}|^2 \Delta_n + 2^j \sup_{j \leq J,\, k} B_{jk}^2 \mathbb{I}\{\Delta_n 2^j \geq N\}\right] := A + B
\end{aligned}$$

To study the first term, again using lemma 5, and (27), we get

$$\begin{aligned}
\rho^{\otimes n}(A \geq \eta^2/3) &\leq \rho^{\otimes n}(\sum_{j=\underline{j}}^{J} \sum_k |B_{jk}|^2 2^J \Delta_n \geq c\eta^2) \\
&\leq \rho^{\otimes n}(\sum_{j=\underline{j}}^{J} \sum_k |B_{jk}|^2 2^J \geq t^2) + \rho^{\otimes n}(\Delta_n \geq c\eta^2/t^2) \\
&\leq c \exp -c[n\frac{t^4}{J^2 2^{2J}} \vee \log n] + K \exp -n\frac{c^2 2\eta^4}{t^4}
\end{aligned}$$

for $t^2 2^{-J} \geq c t_n^{1/2}$: Optimizing in t, we find, for $t^4 = c\eta^2 2^J J$,

$$\rho^{\otimes n}(A \geq \eta^2/3) \leq \exp -n\eta^2 2^{-J} J^{-1}$$

This is valid if $t^2 2^{-J} \geq c t_n^{1/2}$ i.e. $\eta 2^{-J/2} \geq c n^{-1/2}$.

Now taking $t = mJ$, we find

$$\rho^{\otimes n}(A \geq \eta^2/3) \leq \exp -[d \log n]$$

using again the fact that $s > \frac{1}{2}$ and $\eta \geq D\eta_n$.

On the other hand, we have also the following bound using Bernstein inequality (see (30) :

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J} \sum_k |B_{jk}|^2 2^j \geq t^2) \leq 2^J \rho^{\otimes n}(|B_{jk}|^2 2^{2J} \geq t^2) \leq 2^J \exp -nct^2 2^{-2J} \tag{33}$$

For $t2^{-J/2} \leq c'$. If then again, we optimize in t, we find : $t^2 = \eta^{4/3} 2^{2J/3}$ leading to the rate : $\exp -n\eta^{4/3} 2^{-4J/3}$ We have $\eta^{4/3} 2^{-4J/3} \geq \eta^2 2^{-J}$ for $\eta \leq 2^{-J/2}$. In this case, we precisely have $t^2 2^{-J/2} = \eta^{4/3} 2^{2J/3} 2^{-J/2} \leq 2^{-J/2}$.

It is obvious that the second term $(B\ )$ may be bounded (using (27)) by

$$\rho^{\otimes n}(\Delta_n 2^j \geq N) \leq K \exp -2nN^2 2^{-2J} \leq \exp -2N \log n$$

Now, we have, using (30)

$$\rho^{\otimes n}(\sup_{\underline{j} \leq j \leq J,\, k} B_{jk}^2 2^J \geq c' \eta^2) \leq c \exp cn\eta^2 2^{-J}$$

if $\eta \leq c''$. Notice that the constant c'' may be chosen arbitrarily. Of course this choice will change the constant c. Hence, let us take $c'' = MN$, and now, let us remark that,

$2^j B_{jk}^2 \leq 2^j [\frac{1}{n} \sum_i M 2^{J/2} \mathbb{I}\{\frac{i}{n} \in [\frac{k}{2^j}, \frac{k+n}{2^j}]\}]^2 \leq 2^j [\frac{1}{n} M 2^{J/2} \frac{nN}{2^j}]^2 \leq M^2 N^2$. Hence the probability for $\sup_{\underline{j} \leq j \leq J,\, k} B_{jk}^2 2^J$ to exceed η^2 is zero for $\eta^2 > M^2 N^2$.

This achieves bounding the term SS as well as ends up the proof of the theorem.

Références

[Bernstein, 1946] Bernstein, S. (1946). *The theory of Probability.* Gastehizdal Publishing House, Moscow.

[Cohen et al., 1993] Cohen, A., Daubechies, I., and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1) :54–81.

[Cohen et al., 2001] Cohen, A., DeVore, R., Kerkyacharian, G., and Picard, D. (2001). Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2) :167–191.

[Cucker and Smale, 2002] Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *Bull. Amer. Math. Soc. (N.S.)*, 39(1) :1–49 (electronic).

[DeVore et al., 2004] DeVore, R., Kerkyacharian, G., Picard, D., and Temlyakov, V. (2004). Mathematical methods for supervised learning. Technical report, IMI. University of South carolina.

[Diarmid, 1989] Diarmid, M. (1989). On the method of bounded differences. In *Surveys in Combinatorics*, pages 148–188. Cambridge University Press, Cambridge.

[Donoho and Johnstone, 1994] Donoho, D. L. and Johnstone, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika*, 81(3) :425–455.

[Donoho et al., 1995] Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage : Asymptopia ? *Journal of the Royal Statistical Society, Series B*, 57 :301–369. With Discussion.

[Györfi et al., 2002] Györfi, L., Kohler, M., Krzyżak, A., and Walk, H. (2002). *A distribution-free theory of nonparametric regression.* Springer Series in Statistics. Springer-Verlag, New York.

[Ibragimov and Has'minskiĭ, 1981] Ibragimov, I. A. and Has'minskiĭ, R. Z. (1981). *Statistical estimation.* Springer-Verlag, New York. Asymptotic theory, Translated from the Russian by Samuel Kotz.

[Kerkyacharian and Picard, 2000] Kerkyacharian, G. and Picard, D. (2000). Thresholding algorithms and well-concentrated bases. *Test*, 9(2).

[Kerkyacharian and Picard, 2004] Kerkyacharian, G. and Picard, D. (2004). Regression in random design and warped wavelets. *Bernoulli*, 10(6) :1053–1105.

[Konyagyn and Temlyakov, 2004] Konyagyn, S. V. and Temlyakov, V. N. (2004). Some error estimates in learning theory. In *Approximation theory : a volume dedicated to Borislav Bojanov*, pages 126–144. Prof. M. Drinov Acad. Publ. House, Sofia.

[Korostelev, 2003] Korostelev, A. (2003). The Bahadur risk in probability density estimation. *Statist. Decisions*, 21(2) :139–148.

[Korostelev and Spokoiny, 1996] Korostelev, A. P. and Spokoiny, V. G. (1996). Exact asymptotics of minimax Bahadur risk in Lipschitz regression. *Statistics*, 28(1) :13–24.

[Nemirovskiy, 1985] Nemirovskiy, A. S. (1985). Nonparametric estimation of smooth regression functions. *Izv. Akad. Nauk SSSR Tekhn. Kibernet.*, (3) :50–60, 235.

[Pinelis, 1994] Pinelis, I. (1994). Optimum bounds for the distributions of martingales in banach spaces. *Ann. Probab.*, 22 :1679–1706.

[Poggio and Smale, 2003] Poggio, T. and Smale, S. (2003). The mathematics of learning : dealing with data. *Notices Amer. Math. Soc.*, 50(5) :537–544.

[Smale and Zhou, 2005] Smale, S. and Zhou, D.-X. (2005). Learning theory estimates via operators and their approximations. Technical report, Toyota Technological Institute.

[Stone, 1982] Stone, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.*, 10(4) :1040–1053.

[Temlyakov, 2005] Temlyakov, V. (2005). Approximation in learning theory. Technical report, IMI. University of South carolina.

[Van de Geer, 2001] Van de Geer, S. (2001). *Empirical processes in M-estimation.* Cambridge University Press, New York.

[Yang and Barron, 1999] Yang, Y. and Barron, A. (1999). Information-theoretic determination of minimax rates of convergence. *Ann. Statist.*, 27(5) :1564–1599.

CNRS LPMA, 175 rue du Chevaleret, 75013 Paris, France. Université Paris X, 200 Avenue de la République 92001 Nanterre cedex. Université Paris VII, 175 rue du Chevaleret, 75013 Paris, France