THRESHOLDING IN LEARNING THEORY

GÉRARD KERKYACHARIAN AND DOMINIQUE PICARD

RÉSUMÉ. In this paper we investigate the problem of learning an unknown bounded function. We will emphasize special cases where it is possible to provide very simple (in terms of computation) estimates enjoying in addition the property of being universal i.e. their construction does not depend on the a priori knowledge of 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 technique, 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, where 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 concerned 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 : Ibraguimov and Hasminski[10], Van de Geer[25] and Gyorfi et al. [9]). It has recently drawn much attention in the work of Cucker and Smale [4] and amplified upon in Poggio and Smale [19].

We will assume to observe an n sample $Z_1, \dots, Z_n$ of $Z = (X, Y)$. The distribution of Z is 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 ρ_X is the distribution of X and $\hat\rho_X$ is the empirical measure calculated 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.

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

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.

1991 *Mathematics Subject Classification.* 62G08.

Key words and phrases. Learning Theory, Non-linear Methods, Thresholding, Wavelets, Reproducing Kernel Hilbert Spaces.

Contrary to statistics, where people are mainly concerned with evaluation of moments of $\|\hat{f} - f_\rho\|$ (except rare examples, see Korostelev [15] or Korostelev and Spokoiny [14]), learning theory focuses on investigating the decay of (1) as $n \to \infty$ and η increases.

Another difference with the statistical point of view is that one main 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 probability measures ρ such that f_ρ belongs to Θ, it has been defined in DeVore et al.[5], for each $\eta > 0$ the *accuracy confidence function of the procedure* $\hat{f}$:

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

This quantity measures a uniform confidence (over the space $\mathbb{M}(\Theta)$) we have that the estimator $\hat{f}$ is accurate to tolerance η. Upper and lower bounds for $\mathbf{AC}$ have been proved in [5]. In most examples, there is a critical value $\eta_n = \eta(n, \Theta)$ such that for $\eta > \eta_n$, (2) decreases exponentially. This critical value η_n is essential since it yields, as a consequence, bounds of type $e_m(\Theta, \hat{f}) \le C\eta_n^q$ which have been extensively studied in statistics, for

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

The expectation here is taken with respect to $(Z_1, \dots, Z_n)$, with distribution $\rho^{\otimes n}$ (i.e. i.i.d. with common measure ρ).

To evaluate lower bounds for the function $\mathbf{AC}_m(\Theta, \hat{f}, \eta)$, [5] 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 : there exist constants C and D, such that,

$$\mathbf{AC}_n(\Theta, \eta) \ge C\{ \begin{array}{ll} e^{-cn\eta^2}, & \forall\, \eta > D\eta_n, \\ 1, & \forall\, \eta \le D\eta_n, \end{array}$$

where expressions for η_n are given via a tight entropy (defined there) of the set Θ. For instance, $\eta_n = n^{-\frac{s}{2s+d}}$ for the Besov space $B^s_q(L_\infty(\mathbb{R}^d))$ which corresponds to results of minimax type proved in statistics, with more stringent conditions on the measure ρ :

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

Here also an important difference is that the loss function is given by $\|g\|^2_{dx} = \int_{\mathbb{X}} g(x)^2 dx$ replacing the measure ρ_X by the Lebesgue measure. For more details see, for instance, Ibraguimov and Hasminski [10], Stone [23], Nemirovski [17] for a slightly more restricted context than Besov spaces, and Donoho et al. [8]...

Concerning upper bounds for $\mathbf{AC}_n(\Theta, \eta)$, many reverse properties have been established : see for instance Yang and Barron [26] in statistical context, [4], [5], Konyagyn and Temlyakov [13], in learning theory. These upper bounds are generally proved using estimation methods based on empirical mean square minimization :

$$\hat{f} = Argmin\{\sum_{i=1}^{n} (Y_i - f(X_i))^2,\ f \in \mathbb{H}_n\}. \tag{4}$$

$\mathbb{H}_n$ is a functional set associated to the method. These estimation rules raise 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 the knowledge of Θ. There also exist universal estimates (see Temlyakov [24]), however these rules are up to now prohibitive in terms of computation time.

Our aim in this paper will be to emphasize special constructions of procedures not relying on a minimization as in (4) (although, we prove that, in spirit, our construction is not so far from a least square estimate) and conditions under which 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 technique, which has proven in the last decade in statistics to have very good properties for recovering signals with inhomogeneous smoothness.

Our results, although universal will rely on conditions on the kernels or on the underlying classes of functions, which may or may not be verified. Another advantage that we claim is that these conditions are easy to check. Some of them are even directly ascertainable on the data.

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 $L_2(\hat{\rho}_X)$. Our results provide exponential rates of convergences which are good in the following sense. The critical value η_n is the one predicted by [5], and the exponential rates are comparable to those recently obtained by Smale and Zhou [21], although the loss is not the same ($L_2(\rho_X)$ in [21]), and the regularity assumptions are somewhat different. In [21], regularity assumptions are expressed in terms of RKHS spaces. These assumptions may seem more intrinsic. However it is often quite delicate to figure out their meaning since they generally are depending on the unknown measure ρ_X. Our conditions also are depending on the kernel, but they are expressed in such a way that it is always possible given an arbitrary function to check whether they are verified or not. In this sense it is much easier to explain to a potential user the classes of objects that we are effectively able to handle.

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 concerned 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.

It is also interesting to notice that the methods of proofs here mix the technics of deviation inequalities which are popular in learning theory with the standard arguments of thresholding. We use mostly Bernstein inequality in the wavelet case, since we can, most of the time, treat each coefficient separately, and Dvoresky Kiefer and Wolfovitz inequality to bound the deviation of the empirical process. Mac Diarmid inequality is also used. In the RKHS setting, we mostly use Pinelis inequality.

2. Least squares and thresholding procedures

In this short section, we will give a motivation for the construction of our thresholding estimates. To make easier their understanding and motivate their consideration, we give here a connection to 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

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

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 :

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

where $\delta_{kl} = 0$ for $k \neq l$, $\delta_{ll} = 1$. Otherwise, $(e_k)_k$ is an orthonormal system for the empirical measure $\hat{\rho}_X$ on the $X_i's$.

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

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

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

and introduce the three following estimations of these coefficients :

$$\hat{\alpha}_k^{(1)} = \frac{1}{n}\sum_{i=1}^n e_k(X_i)Y_i, \quad \hat{\alpha}_k^{(2)} = sign(\hat{\alpha}_k^{(1)})[|\hat{\alpha}_k^{(1)}| - \lambda]_+$$

$$\hat{\alpha}_k^{(3)} = \hat{\alpha}_k^{(1)}\mathbb{I}\{|\hat{\alpha}_k^{(1)}| \geq \lambda\}.$$

Here and in the sequel $\mathbb{I}\{A\}$ denotes the indicator function of the set A, and $[x]_+ = x\mathbb{I}\{x \geq 0\}$. It is easy to prove that there exists $\lambda^i(\kappa)$ such that the following rules are empirical minimizers for the respective spaces $\mathbb{H}_n^{(i)}$, $i \in \{1, 2, 3\}$:

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

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

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 following questions : How to choose the functions $(e_k)_k$ verifying condition (P) ? How to choose the 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 deterministic points in $\mathbb{R}^d$

$$\{x_1, \ldots x_p\}$$

(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 following sense : there exist $0 \leq \delta < 1$ and $\tau > 0$, such that :

$$\forall u \in \mathbb{R}^p, \quad \|u\|_{l_2}^2(1-\delta)^2 \leq \|M_{np}u\|_{l_2}^2; \quad \|M_{np}{}^{\frac{1}{2}}u\|_{l_2}^2 \leq \|u\|_{l_2}^2\tau^2, \tag{5}$$

$$\|u\|_{l_\infty}(1-\delta) \leq \|M_{np}u\|_{l_\infty}. \tag{6}$$

The second part of (5) only reflects the fact that K is a bounded kernel, which is a relatively harmless assumption. The first part of (5) is closely related to a situation which has been investigated in Smale and Zhou [20] and [22]. In these papers, the richness $(\lambda_{\bar{x}})^2$ of the data $\bar{x} = (X_1, \ldots, X_n)$ is defined as the smallest eigenvalue of the matrix nM_{np} which in our case is clearly related to δ. In particular, lower bounds for $\lambda_{\bar{x}}$ are given for instance in Proposition 7 of [20], which allow direct evaluations of the constant δ in most situations.

Our results will not assume anything about δ but this quantity will play a key role to measure the performances of the procedure. In particular -and this is essential in practice- we allow in the sequel δ to be a random quantity depending on the observations (as it is the case for evaluations following [20]). Of course, δ will be desired to be as small as possible.

3.1.2. *Estimation rule.* Let us consider the following estimation rule : We will denote by Y the n dimensional vector with coordinates Y_i, ε will be the n dimensional vector with coordinates $\varepsilon_i = Y_i - f_\rho(X_i)$. 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 $\frac{1}{n}KK^t = M_{np}$, if K^t denotes the matrix transposed of K), and introduce :

$$t_n = \frac{\log n}{n}, \quad \lambda_n = T\sqrt{t_n}, \tag{7}$$

$$z = (z_1, \dots, z_p)^t = (KK^t)^{-1}KY, \tag{8}$$

$$\widetilde{z} = (\widetilde{z}_1, \dots, \widetilde{z}_p)^t, \quad \widetilde{z}_l = z_l \mathbb{I}\{|z_l| \geq \lambda_n\} \tag{9}$$

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

$$\hat{f} = \sum_{l=1}^{p} \widetilde{z}_l K(x_l, \cdot). \tag{10}$$

As is easily seen, $\hat{f}$ is expressed in a rather similar way as $\hat{f}_3$ and it is worthwhile to notice that its construction does not depend on any regularity parameter. Another analogy which may help to interprete this construction is the linear model in statistics of the following form $Y = K^t\alpha + noise$, where the mean square estimate of α is $\hat{\alpha} = [KK^t]^{-1}KY$, in the case where KK^t is close to the identity matrix, and where we seek for sparsity by thresholding the coefficients.

3.1.3. *Regularity conditions.* We will assume the following sparsity conditions on the function f_ρ :
Let us take, if $\lfloor x \rfloor$ denotes the integer part of x,

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

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

$$\|f_\rho - \sum_{l=1}^{p} \alpha_l K(x_l, \cdot)\|_{\hat{\rho}_X} \leq cp^{-[\frac{2s}{1+2s}+\frac{1}{2}]}, \tag{11}$$

$$\forall \lambda > 0, \ card\{|\alpha_l| \geq \lambda\} \leq c\lambda^{-\frac{2}{1+2s}} \wedge p. \tag{12}$$

(We recall that as usual, $x \vee y$ denotes the maximum of x and y, whereas $x \wedge y$ denotes the minimum.)

The conditions (11) and (12) 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 [11], (11) reflects a 'minimal compactness condition' which generally does not interfere in the entropy calculations (for instance) neither in the minimax rates of convergence. Condition (12) does drive the rates. It is given here with a Lorentz type constraint on the α_i's. Condition (12) is obviously implied by l_r conditions (for $r = \frac{2}{1+2s}$) which are related to Besov conditions.

Let us now discuss briefly an example of such a situation which has been investigated in Smale and Zhou [20] and [22]. Let us consider the space of functions $f = \sum_{t \in \bar{t}} \alpha_t K(t, \cdot)$, where $\bar{t}$ is the sequence $\{x_l, \ l \in \mathbb{N}\}$, K is a predefined kernel. For instance, we can take $K(t, \cdot) = \psi(t - \cdot)$. Interesting choices for ψ in such situations are for instance the gaussian function or the cardinal sine function. In the above papers, lower bounds are given for δ. In this setting, if (5) is verified, simple calculations prove that if the sequence $(\alpha_l)_{l \in \mathbb{N}}$ is such that, $\sup_n n^{\frac{2s}{1+2s}+\frac{1}{2}}[\sum_{l \geq n} \alpha_l^2]^{\frac{1}{2}} \leq c$, (11) is verified. In this case, it is also obvious that f belongs to the reproducing kernel Hilbert space $\mathcal{H}_K$ induced by the kernel K.

It is an interesting fact to notice that the conditions (11) and (12) are readable directly on the coefficients α_l. This seems to be a big advantage compared to the usual conditions that f_ρ belongs to the range of the operator L_K^r (see [21]), since this operator is defined via ρ_X which is unknown.

If we look at the usual conditions which are assumed in similar situations in statistics, it seems that in (11) the right exponent in p should be $\frac{-2s}{1+2s}$ instead of $-[\frac{2s}{1+2s} + \frac{1}{2}]$. Actually, in most situation, and especially in the example above, it can be proved that $\frac{-2s}{1+2s}$ is enough. However, in full generality, we have not been able to reduce this rate.

Theorem 1. *Let us take*

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

and for any $s > 1/2$, we define,

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

Under the conditions above, there exist positive constants D, γ, and L such that if $T \geq T_0$,

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

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 [5]. 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 [21], 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 $Ln^{-\gamma}$.*

The constants D, γ, T_0, and L follow from the proof. However we have not sought for optimality of the constants, so they probably can be improved.

The loss function here is the norm associated with the empirical measure. In statistic this approach is rather classical whereas in learning theory, the favorite loss, is, as considered in the second part of this paper, $L_2(\rho_X)$. Of course using an ergodic argument, the two norms are close. However, the difference is of exponential order, and since this is the range of our results, thorough checking is needed. This is beyond the scope of this paper.

Finally, it is important to notice the following technical fact which will be crucial in the sequel : because $s > 1/2$, $\eta_n \geq p^{-1}$.

3.2. **Proof of the theorem.** First, we will bound the risk by the sum of a bias term (coming from the fact that we stop the expansion at the level p) and a stochastic term. The bias term will be bounded using assumption (11). The second term will be reduced to a quadratic expression using (5). We will denote by α the p dimensional vector with coordinates α_l.

$$\begin{aligned} \|f_\rho - \hat{f}\|_{\hat{\rho}_X} &\leq \|f_\rho - \sum_{l=1}^p \alpha_l K(x_l,\cdot)\|_{\hat{\rho}_X} + \|\sum_{l=1}^p \alpha_l K(x_l,\cdot) - \hat{f}\|_{\hat{\rho}_X} \\ &\leq c\eta_n + \|\sum_{l=1}^p (\alpha_l - \widetilde{z}_l) K(x_l,\cdot)\|_{\hat{\rho}_X} \\ &\leq c\eta_n + [(\alpha - \widetilde{z})^t M_{np} (\alpha - \widetilde{z})]^{\frac{1}{2}} \\ &\leq c\eta_n + \tau [\sum_{l=1}^p (\alpha_l - \widetilde{z}_l)^2]^{\frac{1}{2}}. \end{aligned} \tag{14}$$

The stochatic term will now be decomposed according to the fact that the α_l's as well as their estimates z_l's agree or not.

$$\begin{aligned} \sum_{l=1}^p (\alpha_l - \widetilde{z}_l)^2 &\leq \sum_{l=1}^p (\alpha_l - z_l)^2 \mathbb{I}\{|z_l| \geq \lambda_n\} \Big\{ \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} + \mathbb{I}\{|\alpha_l| < \lambda_n/2\} \Big\} \\ &+ \sum_{l=1}^p [\alpha_l]^2 \mathbb{I}\{|z_l| < \lambda_n\} \Big\{ \mathbb{I}\{|\alpha_l| \geq 2\lambda_n\} + \mathbb{I}\{|\alpha_l| < 2\lambda_n\} \Big\} \\ &:= BB + BS + SB + SS \end{aligned} \tag{15}$$

Let us study the term SS. We will use condition (12) on f_ρ. It is not difficult to prove that (12) is equivalent to the following characterization (the result is standard in Lorenz spaces and in any case can be found in Cohen et al.[3]) :

$$\forall\, \lambda > 0, \sum_l \alpha_l^2 \mathbb{I}\{|\alpha_l| < \lambda\} \leq c\lambda^{\frac{4s}{1+2s}}. \tag{16}$$

Hence, using (16),

$$SS \leq c\lambda_n^{\frac{4s}{1+2s}} = cT^{\frac{4s}{1+2s}}\eta_n^2. \quad (17)$$

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,

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

In the same way,

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

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

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

$$BB + SB \leq 5\sum_{l=1}^{p}(\alpha_l - z_l)^2\mathbb{I}\{|\alpha_l| \geq \lambda_n/2\}. \quad (18)$$

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 $\bar{f}$ the function $\sum_{l=1}^{p}\alpha_l K(x_l, \cdot)$ and $\bar{f}_X$ the vector n dimensional with coordinates $[\bar{f}_X]_i = \bar{f}(X_i) = \sum_{l=1}^{p}\alpha_l K(x_l, X_i)$, so that,

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

Let us recall that f_X is the n dimensional vector which entries are $f_\rho(X_i)$, and by hypothesis (11), $|f_\rho(X_i) - \bar{f}(X_i)| \leq cp^{-[\frac{2s}{1+2s}+\frac{1}{2}]}$. 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,

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

But, since $(KK^t)^{-1} = \frac{1}{n}{M_{np}}^{-1}$, and using (5),

$$\begin{aligned} \|(KK^t)^{-1}K[\bar{f}_X - f_X]\|_{l_2} &= \frac{1}{n}\|M_{np}^{-1}K[\bar{f}_X - f_X]\|_{l_2} \\ &\leq (1-\delta)^{-1}\tau\|f_X - \bar{f}_X\|_{\hat{\rho}}\sqrt{p} \\ &\leq (1-\delta)^{-1}\tau\eta_n. \end{aligned} \quad (19)$$

From the calculations above and (5), we deduce,

$$\sum_{l=1}^{p}(\alpha_l - z_l)^2\mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \leq 2\sum_{l=1}^{p}((KK^t)^{-1}K\varepsilon)_l^2\mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} + 2(1-\delta)^{-2}[\tau\eta_n]^2. \quad (20)$$

Let us now recall the following inequality due to Pinelis [18], 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 $\mathbb{E}\|\xi_i - \mathbb{E}(\xi_i)\|^2 \leq \sigma^2(\xi)$, then if *Prob* denotes the distribution of the vector $(\xi_1, \ldots, \xi_n)$,

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

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

$$Prob\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\| \geq \lambda\big) \leq 2\exp\{\frac{-n\lambda^2}{2(\lambda\widetilde{M}/3 + \widetilde{M}^2)}\}. \tag{22}$$

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

$$Prob\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\| \geq \lambda\big) \leq 2\exp\ \frac{-3n\lambda^2}{8\widetilde{M}^2}. \tag{23}$$

Let us consider the X_i's as fixed and use for a while, the probability ρ, conditioned to $(X_1, \ldots, X_n)$, $\rho(\cdot\,|X_1, \ldots, X_n)$, and consider the p dimensional vector ξ_i such that :

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

In such a way that,

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

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

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

Let us now define :

$$p^* = card\{|\alpha_l| \geq \lambda_n/2\}.$$

Because of (12), we know that :

$$p^* \leq c(\lambda_n/2)^{\frac{-2}{1+2s}}. \tag{24}$$

Now, we have using (6)

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

Now, using (23), for $a > 0$,

$$\rho^{\otimes n}\big(\|\frac{1}{n}\sum_{i=1}^{n}[\xi_i - \mathbb{E}\xi_i]\|^2 \geq \frac{(a\eta)^2}{(1-\delta)^2}\big|X_1, \ldots, X_n\big) \leq 2\exp -\{\frac{3}{8}n\eta^2 \frac{a^2}{p^* M^2\kappa^2}\}.$$

As the right hand side does not depend on $(X_1, \ldots, X_n)$, the bound is also true, without conditioning. So for $a > 0$, $D > \frac{\kappa}{a}$ and taking account that $\eta > D\eta_n$, using (20), we have

$$\rho^{\otimes n} \quad \Big(\quad \sum_{l=1}^p (\alpha_l - z_l)^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \geq \frac{4(a\eta)^2}{(1-\delta)^2}\Big)$$

$$\leq \quad \rho^{\otimes n}\Big(\sum_{l=1}^p ((KK^t)^{-1}K\varepsilon)_l^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} + [\kappa(1-\delta)^{-1}\eta_n]^2 \geq \frac{2(a\eta)^2}{(1-\delta)^2}\Big)$$

$$\leq \quad \rho^{\otimes n}\Big(\sum_{l=1}^p ((KK^t)^{-1}K\varepsilon)_l^2 \mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \geq \frac{(a\eta)^2}{(1-\delta)^2}\Big) \leq 2\exp\Big(-\{\frac{3}{8}n\eta^2\frac{a^2}{p^* M^2\kappa^2}\}\Big).$$

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 :

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

Hence

$$\rho^{\otimes n}\Big(\sum_{l=1}^p (\alpha_l - z_l)^2\mathbb{I}\{|\alpha_l| \geq \lambda_n/2\} \geq \frac{2(2a\eta)^2}{(1-\delta)^2}\Big) \leq \exp\Big(-C[n\eta^2 p^{-1} \vee \log n]\Big). \quad (25)$$

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 following term :

$$BS = \sum_{l=1}^p (\alpha_l - z_l)^2\mathbb{I}\{|z_l| \geq \lambda_n\}\mathbb{I}\{|\alpha_l| < \lambda_n/2\} \leq \sum_{l=1}^p (\alpha_l - z_l)^2\mathbb{I}\{|z_l - \alpha_l| \geq \lambda_n/2\} \quad (26)$$

This term will be treated in the same way as the previous one except that we will not restrict to the big α_l's (so the indicator function $\mathbb{I}\{|\alpha_l| \geq \lambda_n/2\}$ will be omitted and as a consequence p will be replacing p^*).

$$\rho^{\otimes n} \quad \Big(\quad BS \geq \frac{4(a\eta)^2}{(1-\delta)^2}\Big)$$

$$\leq \quad \rho^{\otimes n}\Big(\sum_{l=1}^p (\alpha_l - z_l)^2 \geq \frac{4(a\eta)^2}{(1-\delta)^2}\Big)$$

$$\leq \quad \rho^{\otimes n}\Big(\sum_{l=1}^p ((KK^t)^{-1}K\varepsilon)_l^2 \geq \frac{(a\eta)^2}{(1-\delta)^2}\Big)$$

$$\leq \quad 2\exp\Big(-\{\frac{3}{8}n\eta^2\frac{a^2}{pM^2\kappa^2}\}\Big)$$

Notice that in this bound we did not take into account the indicator function in (26). If we do so we will be able to get the '$\vee Logn$' in the exponential bound of the theorem : We proceed as in the previous subsection, and obtain using (6) and (11) and $s > \frac{1}{2}$:

$$\begin{aligned} |\alpha_l - z_l| &\leq |[(KK^t)^{-1}K\varepsilon]_l| + \|(KK^t)^{-1}K[\bar{f}_X - f_X]\|_{l_\infty} \\ &\leq |\frac{1}{n}(M_{np}{}^{-1}K\varepsilon)_l| + (1-\delta)^{-1}\tau\|\bar{f}_X - f_X\|_{\hat{\rho}} \\ &\leq (1-\delta)^{-1}\|\frac{1}{n}K\varepsilon\|_{l_\infty} + (1-\delta)^{-1}(t_n)^{\frac{3}{4}}\tau \end{aligned}$$

So

$$\rho^{\otimes n}\Big(BS \geq \frac{4(a\eta)^2}{(1-\delta)^2}\Big) \leq \rho^{\otimes n}\Big(\exists l \in \{1,\ldots,p\},\ |\alpha_l - z_l| \geq \lambda_n\Big) \leq \rho^{\otimes n}\Big((1-\delta)^{-1}\sup_l|\frac{1}{n}(K\varepsilon)_l| + (1-\delta)^{-1}(t_n)^{\frac{3}{4}}\tau \geq \lambda_n\Big)$$

$$\leq \sum_{l=1}^p \rho^{\otimes n}\Big((1-\delta)^{-1}|\frac{1}{n}\sum_i K(x_l, X_i)\varepsilon_i| + (1-\delta)^{-1}\sqrt{t_n}\tau \geq T\sqrt{t_n}\Big)$$

$$\leq \sum_{l=1}^{p} \rho^{\otimes n}\Big(|\frac{1}{n}\sum_i K(x_l, X_i)\varepsilon_i| \geq \sqrt{\log n/n}\big[\frac{T}{1-\delta} - \tau(\frac{\log n}{n})^{1/4}\big]\Big).$$

But for $n \geq 2$ and $T > 2(1-\delta)\kappa(\frac{\log 2}{2})^{1/4}$, and using Hoeffding inequality,

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

So

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

with $\gamma > 0$ if $T \geq 4(1-\delta)\kappa M$.

So, putting together the two bounds obtained above, we get,

$$\rho^{\otimes n}(BS \geq \frac{4(a\eta)^2}{(1-\delta)^2}) \leq 2\Big\{\exp -\{\frac{3}{8}n\eta^2\frac{a^2}{pM^2\kappa^2}\}\Big\} \wedge n^{-\gamma}. \tag{27}$$

Putting together (14), (15), (17), (18), (25) and (27) yields the results. Notice that, if needed, we can also choose T (then depending on γ) so that γ is arbitrarily large.

4. Wavelet results

4.1. **Assumptions and estimation rules.**

4.1.1. *Assumptions on the model.* In this section, we will concentrate on the one dimensional case : the random variables X_i's are now taking their values in a compact domain of $\mathbb{R}$, $\mathbb{X} = [0,1]$ (for simplicity). 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 are able to 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 this function is differentiable. 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 a sparse representation. We will denote by $\mathcal{M}(\Theta_s)$, the set of measures ρ verifying all the assumptions above with in addition the fact that $f_\rho(G^{-1})$ belongs to the standard Besov ball $B^s_\infty(L_\infty([0,1]))(M)$. (We recall that $B^s_q(L_p([0,1]))$ denotes the Besov space of parameters s, p, q, and $B^s_q(L_p([0,1]))(M)$ the ball of radius M in this 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 $\{\psi_{jk}, j \geq \underline{j},\ 0 \leq k < N2^j\}$ a compactly supported wavelet basis on $[0,1]$. As usual, for $j > \underline{j}$, $\psi_{jk}(x) = 2^{\frac{j}{2}}\psi(2^j x - k)$ whereas, $\psi_{\underline{j},k} = \varphi_{\underline{j}k}$ denotes the scaling function. We suppose that φ and ψ are sufficiently differentiable and ψ has enough moment conditions. In addition, we suppose that the length of the support of ψ_{jk} is less than $N2^{-j}$). All these assumptions are standard (see for instance, Cohen et al.[2]).

Let us expand f_ρ in the wavelet basis in the following sense :

$$f_\rho(G^{-1}) = \sum_{j=\underline{j}}^{\infty}\sum_{0\leq k<N2^j}\beta_{jk}\psi_{jk}.$$

Notice that it may seem strange to expand $f_\rho(G^{-1})$ instead of f_ρ, however the following change of variable formula $\int \psi_{jk}(G(x))f_\rho(x)d\rho_X(x) = \int_a^b \psi_{jk}(y)f_\rho(G^{-1}(y))dy$ will shed some light on this, since expanding $f_\rho(G^{-1})$ in the basis $\{\psi_{jk}, j \geq \underline{j},\ 0 \leq k < N2^j\}$ appears then as expanding f_ρ in the 'warped basis' $\{\psi_{jk}(G), j \geq \underline{j},\ 0 \leq k < N2^{\underline{j}}\}$, but using a scalar product associated to the measure $d\rho_X(x)$ which is genuine in the problem (see Kerkyacharian and Picard[12]).

It is well known that for $0 \leq \gamma < \infty$, $f_\rho(G^{-1})$ belongs to $B^\gamma_\infty(L_\infty([0,1]))$ if and only if (and we will take this as the $B^\gamma_\infty(L_\infty([0,1]))-$norm) :

$$\sup_{j\geq\underline{j}} 2^{j(\gamma+\frac{1}{2})}\sup_{0\leq k<N2^j}|\beta_{jk}| =: \|f\|_{\gamma\infty\infty} < \infty.$$

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

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

and define :

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

Let us introduce the ordered statistic : $X_{(1)} \leq \ldots \leq X_{(n)}$ constructed by ordering $X_1, \ldots, X_n$. 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 :

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

– Step 2 : Thresholding

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

– Step 3 : Reconstruction

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

Note that this algorithm is an adaptation of the standard wavelet algorithm introduced in Donoho and Johnstone[7] in the case of an equispaced design. In Kerkyacharian and Picard[12], the expectation properties of the $L_p(dx)$ losses have been investigated (instead of here the deviation properties of the $\mathbb{L}_2(d\rho_X)$). It proves there 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*

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

which is a function especially easy to draw.

Theorem 2. *Under the conditions above, for all $s > \frac{1}{2}$, if we denote again,*

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

there exist positive constants γ, L, D such that,

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

as long as

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

Remark 2. *As mentioned in the introduction these results are comparable to those obtained in the RKHS situation as concerned 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 our 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 $2^J \leq [\frac{n}{\log n}]^{\frac{1}{2}}$.*

We leave in appendix the proof of this theorem. The proof uses similar arguments as the proof of the previous theorem in the RKHS situation. The concentration inequalities are different. The use of wavelets allows to express conditions which are close to conditions (5) and (6), in term of the regularity of the function f_ρ. Of course, the fact that our conditions are expressed in terms of $f_\rho(G^{-1})$ instead of f_ρ itself is more restrictive. If we take the example where the measure ρ_X has the density

$$g(x) = (\alpha+1)x^\alpha I\{[0,1]\}(x)$$

$f_\rho(G^{-1})(x) = f_\rho(x^{\frac{1}{\alpha+1}})$. Obviously, if for instance $s \leq k$ and $\alpha \leq -1 + 1/k$, $f_\rho \in B^s_\infty(L_\infty([0,1]))$ implies $f_\rho(G^{-1}) \in B^s_\infty(L_\infty([0,1]))$. But in general it may be untrue.

5. Appendix : Proof of Theorem 2

This proof is of the same essence as the proof of Theorem 1. However some technical details will be different. Before entering into the main stream of the proof, let us begin with some lemmas which will be essential in the sequel.

5.1. Preliminary lemmas.

The first lemma bounds the following Riemann sums. It will be applied in the sequel when h equals either φ or ψ, and we will not later on make the difference between the two functions, because the properties that will be used will not necessitate this. The proof of the lemma is obvious, obtained using standard arguments comparing the Riemann sum and the integral.

Lemma 1. *For any $r \geq 1$, we have, for $h_{jk}(x) = 2^{\frac{j}{2}}h(2^j x - k)$, if h is a compactly supported function (with support of length N) with bounded derivative,*

$$\frac{1}{n}\sum_{i=1}^n |h_{jk}(\frac{i}{n})|^r \leq \tau_r 2^{j(\frac{r}{2}-1)} + \tau_r' \frac{2^{j(1+\frac{r}{2})}}{n} \tag{28}$$

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

If we recall that $X_{(1)} \leq \ldots \leq X_{(n)}$ denotes the ordered statistics built on $X_1, \ldots, X_n$ and keeping this ordering, we 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), while the $Y_{(i)}$ are not, and the U_i's are i.i.d. uniformly distributed. Now let us denote

$$A_{jk} := \frac{1}{n}\sum_{i=1}^n f_\rho(G^{-1}(U_{(i)}))\psi_{jk}(\frac{i}{n}) - \int \psi_{jk}(x) f_\rho(G^{-1})(x)dx.$$

The following lemma describes the probability of large deviations for the A_{jk}'s.

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

$$\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k A_{jk}^2 \geq \eta^2\Big) \leq \exp -C\left\{n2^{-J}\eta^2 \vee \log n\right\}, \tag{29}$$

for all $\eta \geq D\eta_n$, where $C = 2(2C_1^2 N)^{\frac{-1}{s}}$, $D \geq C_2 + C_3'$

Proof of the lemma : 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 $\bar{s} = s \wedge 1$, we recall that by definition of the Besov spaces, we have :

$$|f(x) - f(y)| \leq \|f\|_{\bar{s}\infty\infty}|x-y|^{\bar{s}} \tag{30}$$

Using (30) for the second inequality, (28) for the third one, we have,

$$\begin{aligned}
|A_{jk}| &\leq \frac{1}{n}\sum_{i=1}^{n} |f_\rho(G^{-1}(U_{(i)})) - f_\rho(G^{-1}(\frac{i}{n}))|\ |\psi_{jk}(\frac{i}{n})| \\
&+ \sum_{i=1}^{n}\int_{(i-1)/n}^{i/n} \left|f_\rho(G^{-1}(x))\psi_{jk}(x) - f_\rho(G^{-1}(\frac{i}{n}))\psi_{jk}(\frac{i}{n})\right| \\
&\leq \Delta_n^{\bar{s}}\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}\frac{1}{n}\sum_{i=1}^{n}|\psi_{jk}(\frac{i}{n})| \\
&+ \sum_{i=1}^{n}\int_{(i-1)/n}^{i/n}\left[2^{j/2}\|\psi\|_\infty\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}n^{-\bar{s}} + \|\psi\|_{1\infty\infty}\|f_\rho(G^{-1})\|_\infty\frac{2^{3j/2}}{n}\right]\mathbb{I}\{x\in[\frac{k}{2^j}, \frac{k+N}{2^j}]\}dx \\
&\leq \Delta_n^{\bar{s}}\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}\{\tau_1 2^{-j/2} + \tau_1'\frac{2^{3j/2}}{n}\} \\
&+ N\|\psi\|_\infty\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}n^{-\bar{s}}2^{-\frac{j}{2}} + N\|\psi\|_{1\infty\infty}\|f_\rho(G^{-1})\|_\infty\frac{2^{j/2}}{n} \\
&\leq C_1\Delta_n^{\bar{s}}2^{-j/2} + C_2\frac{2^{j/2}}{n} + C_3'n^{-\bar{s}}2^{-\frac{j}{2}},
\end{aligned} \tag{31}$$

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 $j \leq J, \quad 2^{2j} \leq n$.

We observe that

$$\sum_{j=\underline{j}}^{J}\sum_k [\frac{2^{j/2}}{n}]^2 \leq c\frac{2^{2J}}{n^2} \leq c\frac{1}{n} \text{ and } \quad \sum_{j=\underline{j}}^{J}\sum_k n^{-2\bar{s}}2^{-j} \leq Jn^{-2\bar{s}}.$$

These two terms are obviously going to zero at a faster rate that η_n^2. They will not play a role in the final bound.

$$\sum_{j=\underline{j}}^{J}\sum_k \Delta_n^{2\bar{s}}2^{-j} \leq NJ\Delta_n^{2\bar{s}} \tag{32}$$

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

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

The last line uses the following Dvoreski, Kiefer and Wolfovitz bound (see, for instance the review on the subject in Devroye et al. [6] section 12.) : for any $\lambda > 0$, there exists a universal constant K, such that :

$$\mathbb{P}(\Delta_n \geq \lambda) \leq K\exp -2n\lambda^2 \tag{33}$$

The indicator function comes from the fact that $\Delta_n \le 1$. Now, for $s \ge 1$, $n[\eta]^{\frac{2}{s}} J^{\frac{-1}{2s}} = n\eta^2 J^{-1/2} \ge n\eta^2 2^{-J} \vee \log n$. Identically, for $1/2 < s < 1$ and $\eta \ge D\eta_n$, if we put as above $2^{j_s} = [\frac{n}{\log n}]^{\frac{1}{1+2s}}$,

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

This ends up the proof of the lemma. □

Let us now denote, as above,

$$\varepsilon_i = Y_i - f_\rho(X_i),$$

and

$$B_{jk} := \frac{1}{n} \sum_{i=1}^{n} \varepsilon_i \psi_{jk}(\frac{i}{n}).$$

We have the following bound.

Lemma 3. *For $\underline{j} \le j_0 \le J$,*

$$\rho^{\otimes n}\{2^{j_0} \sup_{j \le j_0} \sup_k B_{jk}^2 > \lambda^2\} \le 2N 2^{j_0} \exp - \frac{n\lambda^2 2^{-j_0}}{2(C_3 + \lambda M \|\psi\|_\infty/3)}$$

The proof of the lemma is simple :

$$\rho^{\otimes n}\{2^{j_0} \sup_{j \le j_0} \sup_k B_{jk}^2 > \lambda^2\} \le \sum_{j \le j_0} \sum_k \exp - \frac{n\lambda^2 2^{-j_0}}{2(C_3 + \lambda M \|\psi\|_\infty 2^{(j-j_0)/2}/3)} \le 2N 2^{j_0} \exp - \frac{n\lambda^2 2^{-j_0}}{2(C_3 + \lambda M \|\psi\|_\infty/3)}.$$

We used Bernstein inequality (cf Bernstein [1]), 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 variance,

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

(We have used (28) and noticed $C_3 := M^2(\tau_2 + \tau_2')$.) □

To finish this subsection, we will state the following lemma which also concerns bounds for large deviations of the B_{jk}'s, but is covering a different domain and using a different type of probabilist inequality.

Lemma 4. *For j_0 such that $t_n^{\frac{-1}{1+2s}} \le 2^{j_0} \le t_n^{-1/2}$, there exists a constant C :*

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{j_0} \sum_k B_{jk}^2 \ge \lambda^2) \le \exp -C\{n\lambda^4/J^2 \vee \log n\}, \tag{34}$$

for all $\lambda^2 \ge 4NM^2 C_1 \frac{2^{j_0}}{n}$.

Proof of the lemma : We will use here Mac Diarmid inequality (see [16]) . We have :

$$\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{j_0} \sum_k B_{jk}^2 \ge \lambda^2\Big) = \rho^{\otimes n}(F(\varepsilon_1, \dots, \varepsilon_n) \ge \lambda^2)$$

with :

$$F(\varepsilon_1, \dots, \varepsilon_l, \dots, \varepsilon_n) = \sum_{j=\underline{j}}^{j_0} \sum_k \frac{1}{n^2} \Big(\sum_{i=1}^{n} \psi_{jk}(\frac{i}{n})\varepsilon_i\Big)^2$$

$$\begin{aligned}
|\Delta F_l| &= |F(\varepsilon_1,\dots,\varepsilon_l,\dots,\varepsilon_n) - F(\varepsilon_1,\dots,\varepsilon_l',\dots,\varepsilon_n)| \\
&= \sum_{j=\underline{j}}^{j_0}\sum_k \frac{1}{n^2}\left([\sum_{i=1}^n \psi_{jk}(\frac{i}{n})\varepsilon_i]^2 - [\sum_{i=1}^n \psi_{jk}(\frac{i}{n})\varepsilon_i + \psi_{jk}(\frac{l}{n})(\varepsilon_l' - \varepsilon_l)]^2\right) \\
&\le 8M^2 \sum_{j=\underline{j}}^{j_0} \sum_{k,\ |\frac{l}{n}-\frac{k}{2^j}|\le \frac{N}{2^j}} \frac{1}{n^2}\sum_{i=1}^n |\psi_{jk}(\frac{i}{n})||\psi_{jk}(\frac{l}{n})| \\
&\le 8M^2N^2\frac{1}{n^2}\|\psi\|_\infty^2 \sum_{j=\underline{j}}^{j_0} 2^j\frac{n}{2^j} \\
&\le 8M^2N^2\|\psi\|_\infty^2\frac{j_0}{n} =: B^2\frac{j_0}{n}
\end{aligned}$$

On the other hand, using again lemma 28,

$$\begin{aligned}
\mathbb{E}_{\rho^{\otimes n}} F(\varepsilon_1,\dots,\varepsilon_n) &\le \sum_{j=\underline{j}}^{j_0}\sum_k \frac{1}{n^2}[\sum_{i=1}^n \psi_{jk}(\frac{i}{n})\varepsilon_i]^2 \\
&\le \sum_{j=\underline{j}}^{j_0}\sum_k \frac{1}{n^2}\sum_{i=1}^n \psi_{jk}(\frac{i}{n})^2 M^2 \\
&\le M^2C_3\sum_{j=\underline{j}}^{j_0}\sum_k \frac{1}{n} \\
&\le 2NM^2C_3\frac{2^{j_0}}{n}
\end{aligned}$$

Hence, for $\lambda^2 \ge 4NM^2C_1\frac{2^{j_0}}{n}$, using Mac Diarmid inequality (see [16]),

$$\begin{aligned}
\rho^{\otimes n}(\sum_{j=\underline{j}}^{j_0}\sum_k B_{jk}^2 \ge \lambda^2) &\le \rho^{\otimes n}(|F(\varepsilon_1,\dots,\varepsilon_n) - \mathbb{E}_{\rho^{\otimes n}}F(\varepsilon_1,\dots,\varepsilon_n)| \ge \lambda^2/2) \\
&\le \exp\frac{-2\lambda^4}{4n(\frac{BJ}{n})^2} \le \exp -nC'\frac{\lambda^4}{J^2}.
\end{aligned}$$

Now, for $\lambda^2 \ge 4NM^2C_1\frac{2^{j_0}}{n}$, we have $C'n\frac{\lambda^4}{J^2} \ge C"n^{-1}2^{2j_0}\eta^2 \ge C"n^{\frac{1-2s}{1+2s}}(\log n)^{\frac{2}{1+2s}} \ge C"\log n$, which proves the result of the lemma, with $C = C' \wedge C"$. □

5.2. **Main part of the proof.** 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 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, if we recall

$$\|h\|_{dx} := \Big[\int_{[0,1]} h^2(x)dx\Big]^{\frac{1}{2}},$$

$$\|f\|_{\rho_X} = \|f(G^{-1})\|_{dx}.$$

Using the above formula, we have if

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

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

Hence

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

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

$$
(III) = \sum_{j\ge J+1}\sum_k \beta^2_{jk} \le N \sum_{j\ge J+1} 2^j \sup_k \beta^2_{jk} \le NM^2 \sum_{j\ge J+1} 2^j 2^{-j(2s+1)} \le NM^2 2^{-2Js} \le M^2\eta_n^2
$$

if $2^J \ge t_n^{\frac{-1}{1+2s}}$. This bound is parallel to (14) in the proof of Theorem 1, and now we have the following parallel to (15) :

Let us now study the second term :

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

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 $|\beta_{jk}| \le M2^{-j(s+\frac{1}{2})}$. Hence if we put :

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

$$
\begin{aligned}
SS &\le \sum_{j=\underline{j}}^{j_s}\sum_k [\beta_{jk}]^2\mathbb{I}\{|\beta_{jk}|<2\lambda_n\} + \sum_{j=j_s}^{J}\sum_k [\beta_{jk}]^2\mathbb{I}\{|\beta_{jk}|<2\lambda_n\} \\
&\le \sum_{j=\underline{j}}^{j_s}\sum_k [2\lambda_n]^2 + \sum_{j=j_s}^{J}\sum_k [\beta_{jk}]^2 \\
&\le N2^{j_s+1}(2\lambda_n)^2 + N\sum_{j=j_s}^{J} 2^j M^2 2^{-2j(s+\frac{1}{2})} \\
&\le N(8T^2 + 2M^2)\eta_n^2.
\end{aligned}
$$

This is parallel to (17) in the proof of Theorem 1.

Let us now investigate the term SB. We observe that

$$
\mathbb{I}\{|\hat\beta_{jk}|<\lambda_n\}\mathbb{I}\{|\beta_{jk}|\ge 2\lambda_n\} \le \mathbb{I}\{|\hat\beta_{jk}-\beta_{jk}|\ge|\beta_{jk}|/2\}\mathbb{I}\{|\beta_{jk}|\ge 2\lambda_n\}.
$$

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

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

This is parallel to (18) in the proof of Theorem 1. Let us now investigate the term BB'.

If we recall that $X_{(1)} \leq \ldots \leq X_{(n)}$ and that keeping this ordering, we have denoted $Y_{(1)}, \ldots, Y_{(n)}$ and introduced $U_i = G(X_i)$, $i = 1, \ldots, n$, as well as the associated $U_{(1)}, \ldots, U_{(n)}$. We have with the notations introduced in the previous subsections of preliminary lemmas.

$$\begin{aligned} \hat\beta_{jk} - \beta_{jk} &= \frac{1}{n}\sum_{i=1}^{n} Y_{(i)}\psi_{jk}(\frac{i}{n}) - \beta_{jk} \\ &= [\frac{1}{n}\sum_{i=1}^{n} f_\rho(G^{-1}(U_{(i)}))\psi_{jk}(\frac{i}{n}) - \int \psi_{jk} f_\rho(G^{-1})] + [\frac{1}{n}\sum_{i=1}^{n}\varepsilon_{(i)}\psi_{jk}(\frac{i}{n})] \\ &:= A_{jk} + B_{jk} \end{aligned} \tag{35}$$

Lemma 2 exactly bounds $\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}\sum_k A_{jk}^2 > \eta^2)$. To bound the deviations associated with the B_{jk}'s, let us first remark that since $f_\rho(G^{-1}) \in B^s_\infty(L_\infty([0,1]))(M)$, then $|\beta_{jk}| \leq M2^{-j(s+\frac{1}{2})}$ and then, if $T \geq 2M$, $|\beta_{jk}| \geq \lambda_n/2$ implies $j \leq j_s$ (remember that $2^{j_s} = t_n^{\frac{-1}{1+2s}}$). Hence,

$$\begin{aligned} \sum_{j=\underline{j}}^{J}\sum_k B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} &\leq \sum_{j=\underline{j}}^{j_s}\sum_k B_{jk}^2 && (36) \\ &\leq \sum_{j=\underline{j}}^{j_s} 2^j N \sup_k B_{jk}^2 \leq 2^{j_s+1} N \sup_{j\leq j_s, k} B_{jk}^2. && (37) \end{aligned}$$

We will investigate separately the cases $\eta \leq 1$, and $\eta \geq 1$. Let us begin with the first case. We have, using lemma 3 :

$$\begin{aligned} \rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2\Big) &\leq \rho^{\otimes n}(2^{j_s+1} N \sup_{j\leq j_s,k} B_{jk}^2 \geq \eta^2) \\ &\leq 2^{j_s+1}\exp\{\frac{-n\eta^2 N^2 2^{-j_s}}{(C_3 + \eta 4N^2 M\|\psi\|_\infty/3)}\}. && (38) \end{aligned}$$

Hence we obtain :

$$\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2\Big) \leq 2\exp\{-cn\eta^2 2^{-j_s} + \frac{1}{2}\log n\}, \tag{39}$$

with $c = N^2(C_3 + M\|\psi\|_\infty/3)^{-1}$ since $\eta \leq 1$. As $\eta \geq D\eta_n$, it is easy to check that for $D > \frac{2}{c}$, $cn\eta^2 2^{-j_s} \geq 2\log n$. Hence in this case, we get the bound : $\exp\Big(-c\{n2^{-J}\eta^2 \vee \log n\}\Big)$.

Let us now study the case where $\eta \geq 1$. Using (36), we apply lemma 4 with $j_0 = j_s$, $\lambda = \eta$, and get the bound :

$$\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k B_{jk}^2 \mathbb{I}\{|\beta_{jk}| \geq \lambda_n/2\} \geq \eta^2\Big) \leq \exp -C\{n2^{-J}\eta^2 \vee \log n\}, \tag{40}$$

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

$$\sum_{j=\underline{j}}^{J}\sum_k(\hat\beta_{jk}-\beta_{jk})^2\mathbb{I}\{|\beta_{jk}|<\lambda_n/2\}\mathbb{I}\{|\hat\beta_{jk}-\beta_{jk}|\geq\lambda_n/2\} \leq \sum_{j=\underline{j}}^{J}\sum_k(\hat\beta_{jk}-\beta_{jk})^2\mathbb{I}\{|\hat\beta_{jk}-\beta_{jk}|\geq\lambda_n/2\}$$
$$\leq 2[\sum_{j=\underline{j}}^{J}\sum_k A_{jk}^2+\sum_{j=\underline{j}}^{J}\sum_k B_{jk}^2\mathbb{I}\{|B_{jk}|\geq\lambda_n/2\}]$$

Hence

$$\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k(\hat\beta_{jk}-\beta_{jk})^2 \qquad \mathbb{I}\{|\hat\beta_{jk}-\beta_{jk}|\geq\lambda_n/2\}\geq\eta^2\}\Big)\leq\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k A_{jk}^2\geq\eta^2/4\Big)$$
$$+\quad\rho^{\otimes n}\Big(2^{J+1}N\sup_{\underline{j}\leq j\leq J}\sup_k B_{jk}^2\geq\eta^2/4\Big)\wedge\rho^{\otimes n}\Big(\sup_{\underline{j}\leq j\leq J}\sup_k B_{jk}^2\geq\lambda_n^2/4\Big)$$

The first term is bounded using Lemma 2, the second one may be bounded, using lemma 3 by : $\exp -c[n2^{-J}\eta^2\vee\log n]$, if $\eta\leq 1$, since $\rho^{\otimes n}\Big(\sup_{\underline{j}\leq j\leq J}\sup_k B_{jk}^2\geq\lambda_n^2/4\Big)\leq\exp\{-c\log n\}$ using again lemma 3.
For $\eta\geq 1$, we have :

$$\sum_{j=\underline{j}}^{J}\sum_k B_{jk}^2\mathbb{I}\{|B_{jk}|\geq\lambda_n/2\}\leq\sum_{j=\underline{j}}^{J}\sum_k B_{jk}^2$$

Then it suffices to use lemma 4.

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

Proposition 1. $\forall s>\frac{1}{2}$, *there exist constants* c, C, D, *such that,*

$$\sup_{\rho\in\mathcal{M}(\Theta_s)}\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k[\widetilde\beta_{jk}-\beta_{jk}]^2>t^2\Big)\leq c\begin{cases} e^{-C\{n2^{-J}t^2\vee\log n\}}, & t\geq D\eta_n,\\ 1, & t\leq D\eta_n,\end{cases}$$

if $[\frac{n}{\log n}]^{\frac{1}{1+2s}}\leq 2^J\leq[\frac{n}{\log n}]^{\frac{1}{2}}$.

5.3. **Investigation of the term (I).** This term is specific to this case (and in some sense was avoided in the case of Theorem 1 because of the difference of the losses). We have :

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

since $|\widetilde\beta_{jk}-\beta_{jk}|\leq|\hat\beta_{jk}-\beta_{jk}|+|\beta_{jk}|$. We will treat the two terms of the RHS separetely, and begin with the second one. If $Z=\sum|\beta_{jk}|\psi_{jk}$ we observe that $\|Z\|_{s\infty\infty}=\|f_\rho(G^{-1})\|_{s\infty\infty}$. So, if again $\bar s=s\wedge 1$, using (30)

$$\begin{aligned}\|\sum_{j=\underline{j}}^{J}\sum_k|\beta_{jk}|\Big|\psi_{jk}(\hat G_n(G^{-1}))-\psi_{jk}\Big|\|_{dx} &= \|Z\big(\hat G_n(G^{-1})\big)\ -Z\|_\infty\\ &\leq \|f_\rho(G^{-1})\|_{\bar s\infty\infty}\Delta_n^{\bar s}.\end{aligned}$$

Hence,

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

with $c = 2\|f_\rho(G^{-1})\|_{\bar{s}\infty\infty}^{\frac{2}{\bar{s}}}$. We have used inequality (33) as above (and argued as in the five lines after (33) in the proof of lemma 2).

Concerning the stochastic term, using (35) we have :

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

Let us investigate separately the two contributions. Concerning the first one, if $Z' = \sum_{j=\underline{j}}^{J}\sum_k |A_{jk}|\psi_{jk}$, using (31), and $s>\frac{1}{2}$,

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

As above, using again $\|Z'((\hat{G}_n(G^{-1}))-Z'\|_\infty\leq\|Z'\|_{1/2\infty\infty}\Delta_n^{1/2}$, we deduce,

$$\rho^{\otimes n}\Big(\| \sum_{j=\underline{j}}^{J}\sum_k |A_{jk}|\Big|\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}\Big|\|_{\infty}\geq \eta\Big) \leq \rho^{\otimes n}\Big(\|Z'\|_{1/2\infty\infty}\Delta_n^{1/2}\geq\eta\Big).$$

Furthermore, using (33),

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

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 $\eta\geq D\eta_n$. On the other hand, for $\widetilde{C}=C_2+C_3$

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

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

$$\|\sum_{j=\underline{j}}^{J}\sum_k |B_{jk}|\Big|\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}\Big|\|_{dx}^2 \leq [\sum_{j=\underline{j}}^{J}\|\sum_k |B_{jk}|\Big|\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}\Big|\|_{dx}]^2$$

$$\|\sum_k |B_{jk}|\Big|\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}\Big|\|_{dx}^2 = \int \Big[\sum_k |B_{jk}|\Big|\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}\Big|\Big]^2$$

$$= \sum_{i=0}^{n}\int_{U_{(i)}}^{U_{(i+1)}}[\sum_k |B_{jk}||\psi_{jk}(\frac{i}{n})-\psi_{jk}(x)|]^2dx$$

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 :

$$[\sum_k |B_{jk}|\ |\psi_{jk}(\frac{i}{n})-\psi_{jk}(x)|]^2 \leq 2\left\{[\sum_k|B_{jk}|\ |\psi_{jk}(\frac{i}{n})|]^2+[\sum_k|B_{jk}|\ |\psi_{jk}(x)|]^2\right\}\mathbb{I}\{\Delta_n2^j\geq N\}$$

$$\leq 2\left[N\|\psi\|_\infty^2 2^j\sup_{j\leq J,\,k}B_{jk}^2+[\sum_k|B_{jk}|\ |\psi_{jk}(x)|]^2\right]\mathbb{I}\{\Delta_n2^j\geq N\}$$

Putting the two cases together, we deduce :

$$\|\sum_k|B_{jk}|\quad\Big|\psi_{jk}(\hat{G}_n(G^{-1}))-\psi_{jk}\Big|\|_{dx}^2\leq\sum_{i=0}^{n}\int_{U_{(i)}}^{U_{(i+1)}}[\sum_k|B_{jk}|\ |\psi_{jk}(\frac{i}{n})-\psi_{jk}(x)|]^2dx$$

$$\leq \sum_{i=0}^{n}\int_{U_{(i)}}^{U_{(i+1)}}\Big\{4N^2\sum_k|B_{jk}|^2\|\psi'\|_\infty^2 2^{2j}\Delta_n\mathbb{I}\{x\in I_{jk}\}$$

$$+\left[2N\|\psi\|_\infty^2 2^j\sup_{j\leq J,\,k}B_{jk}^2+[\sum_k|B_{jk}|\ |\psi_{jk}(x)|]^2\right]\mathbb{I}\{\Delta_n2^j\geq N\}\Big\}dx$$

$$\leq c\left\{\sum_k|B_{jk}|^2 2^j\Delta_n+\left[2^j\sup_k B_{jk}^2+\sum_k|B_{jk}|^2N^{-1}\Delta_n2^j\right]\mathbb{I}\{\Delta_n2^j\geq N\}\right\}$$

$$\leq c\left[\sum_k 2^j|B_{jk}|^2\Delta_n+2^j\sup_{j\leq J,\,k}B_{jk}^2\mathbb{I}\{\Delta_n2^j\geq N\}\right] \quad (41)$$

To study the first term coming from (41), again using lemma 4, and (33), we get, for $1>a>0$ arbitrary, $\{\sum_{j=\underline{j}}^{J}[\sum_k 2^j|B_{jk}|^2\Delta_n]^{1/2}\}^2\leq 22^{Ja}\sum_{j=\underline{j}}^{J}\sum_k|B_{jk}|^2\Delta_n2^{j(1-a)}\leq 22^J\sum_k|B_{jk}|^2\Delta_n$, and

$$\rho^{\otimes n}(\{\sum_{j=\underline{j}}^{J}[\sum_k 2^j|B_{jk}|^2\Delta_n]^{1/2}\}^2\geq c\eta^2\Big) \leq \rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k|B_{jk}|^22^J\geq t^2\Big)+\rho^{\otimes n}(\Delta_n\geq c\eta^2/t^2)$$

$$\leq c\exp-c[n\frac{t^4}{J^22^{2J}}\vee\log n]+K\exp-n\frac{c^22\eta^4}{t^4},$$

for any arbitrary t such that $t^2 2^{-J} \geq c t_n^{1/2}$. Optimizing in t, we find, for $t^4 = c\eta^2 2^J J$,

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}[\sum_k 2^j |B_{jk}|^2 \Delta_n]^{1/2}\}^2 \geq \eta^2/6) \leq c \exp -c' n\eta^2 2^{-J} J^{-1}.$$

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

$$\rho^{\otimes n}(\sum_{j=\underline{j}}^{J}[\sum_k 2^j |B_{jk}|^2 \Delta_n]^{1/2}\}^2 \geq \eta^2/6) \leq \exp -[d \log n]$$

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 lemma 3 :

$$\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sum_k |B_{jk}|^2 2^j \geq t^2\Big) \leq \rho^{\otimes n}(\sup_{\underline{j}\leq j\leq J} |B_{jk}|^2 2^{2J} \geq t^2) \leq 2^{J+1} \exp -nct^2 2^{-2J}, \tag{42}$$

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 $t^2 2^{-J/2} = \eta^{4/3} 2^{2J/3} 2^{-J/2} \leq 2^{-J/2}$.

Now, for the second term coming from (41), we have using lemma 3, and inequality (33),

$$\begin{aligned}
\rho^{\otimes n}\Big(\sum_{j=\underline{j}}^{J}\sup_k 2^{j/2}|B_{jk}|\mathbb{I}\{\Delta_n 2^j \geq N\} \geq \frac{\eta^2}{6}\Big) &\leq \sum_{j=\underline{j}}^{J}\rho^{\otimes n}\Big(\sup_k 2^{j/2}|B_{jk}| \geq \frac{\eta^2}{6J}\Big) \wedge \rho^{\otimes n}\Big(\Delta_n 2^j \geq N\Big)\mathbb{I}\{\frac{\eta}{J} \leq c"\} \\
&\leq \sum_{j=\underline{j}}^{J} 2^{j+1} \exp -\frac{n\eta^2 2^{-j}/(36J^2)}{2(C_3 + \frac{\eta M\|\psi\|_\infty}{3J})} \wedge \exp(-2nN^2 2^{-2j})\mathbb{I}\{\frac{\eta}{J} \leq c"\} \\
&\leq C[2^J \exp -(cn\eta^2 \frac{2^{-J}}{J^2}) \wedge \exp(-2nN^2 2^{-2J})]\mathbb{I}\{\frac{\eta}{J} \leq c"\}
\end{aligned}$$

Notice that the indicator function in the inequalities above is coming from $2^j B_{jk}^2 \leq 2^j[\frac{1}{n}\sum_i M 2^{j/2}\mathbb{I}\{\frac{i}{n} \in [\frac{k}{2^j}, \frac{k+n}{2^j}]\}]^2 \leq 2^j[\frac{1}{n} M 2^{j/2}\frac{nN}{2^j}]^2 \leq M^2 N^2 := (c")^2$.

This ends up the proof of the theorem since whenever $n\eta^2 \frac{2^{-J}}{J^2}$ becomes of smaller order than J (i.e. for $\eta^2 \leq cJ^3(n\log n)^{-1/2}$), the rate is given by $\exp(-2nN^2 2^{-2J})]$ (since this one is the smallest as soon as $\eta^2 \leq 2^{-J}J^2$) and precisely gives a rate smaller than $\exp -(2N\log n)$. □

We thanks two referees and Vladimir Temlyakov for allowing us to improve an earlier version and especially pointing out the link of our work with the papers [20] *and* [22] *-which we were aware of but did not pay enough attention-.*

Références

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

[2] Albert Cohen, Ingrid Daubechies, and Pierre Vial. Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.*, 1(1) :54–81, 1993.

[3] Albert Cohen, Ronald DeVore, Gerard Kerkyacharian, and Dominique Picard. Maximal spaces with given rate of convergence for thresholding algorithms. *Appl. Comput. Harmon. Anal.*, 11(2) :167–191, 2001.

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

[5] Ronald DeVore, Gerard Kerkyacharian, Dominique Picard, and Vladimir Temlyakov. Approximation methods for supervised learning. *Found. Comput. Math.*, 6(1) :3–58, 2006.

[6] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New York, 1996.

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

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

[9] László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002.

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

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

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

[13] S. V. Konyagyn and V. N. Temlyakov. 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, 2004.

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

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

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

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

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

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

[20] Steve Smale and Ding-Xuan Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc. (N.S.)*, 41(3) :279–305 (electronic), 2004.

[21] Steve Smale and Ding-Xuan Zhou. Learning theory estimates via operators and their approximations. Technical report, Toyota Technological Institute, 2005.

[22] Steve Smale and Ding-Xuan Zhou. Shannon sampling ii : Connections to learning theory. *Applied and Computational Harmonic Analysis*, 19(3) :285–302 (electronic), 2005.

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

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

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

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

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