ESAIM: Probability and Statistics March 2003, Vol. 7, 239–250
DOI: 10.1051/ps:2003011

REPLICANT COMPRESSION CODING IN BESOV SPACES

Gérard Kerkyacharian[1, 2] and Dominique Picard[1]

Abstract. We present here a new proof of the theorem of Birman and Solomyak on the metric entropy of the unit ball of a Besov space $B^s_{\pi,q}$ on a regular domain of $\mathbb{R}^d$. The result is: if $s - d(1/\pi - 1/p)_+ > 0$, then the Kolmogorov metric entropy satisfies $H(\epsilon) \sim \epsilon^{-d/s}$. This proof takes advantage of the representation of such spaces on wavelet type bases and extends the result to more general spaces. The lower bound is a consequence of very simple probabilistic exponential inequalities. To prove the upper bound, we provide a new universal coding based on a thresholding-quantizing procedure using replication.

Mathematics Subject Classification. 41A25, 41A46, 65F99, 65N12, 65N55.

Received April 9, 2001. Revised January 20, 2003.

Introduction

The evaluation of the entropy of the balls in Besov spaces is a very important point in modern nonparametric statistics.

Entropy is a measure of the complexity of a parameter space which early proved to be especially appropriate for approximation –see Lorentz [22]– but also for statistical estimation. LeCam was the first to obtain general bounds depending on the metric structure of the parameter space for the risk of estimators –see LeCam [20] and [21]–. After this seminal work, Assouad [1] and Birgé [2] pointed out the link between the metric structure and the minimax rates. Nowadays, entropy or construction of ϵ-nets is a very common approach to obtain lower bounds in minimax evaluations.

Moreover, entropy is also a very powerful tool to derive exponential bounds for the supremum of processes, in particular empirical processes. This has especially important consequences: for instance in many situations, the rate of convergence of the classical MLE or LSE directly follows from entropy calculations. An important tool in modern nonparametric estimation is penalized methods. Here, again exponential bounds on the supremum of the empirical process on a class of functions is a key argument to calibrate the penalties –see for instance, van de Geer [25], for a review–.

On the other hand, Besov spaces among other spaces of regularity appear to be especially suited for approximation and statistical applications: for instance, balls of Besov spaces appear to be maximal sets for linear approximation methods (see for instance Nikolskii [23]). They also appear to be maximal sets –*i.e.* the set of all functions such that a given procedure has a risk bounded by a specified rate of convergence– for general

Keywords and phrases. Entropy, coding, Besov spaces, wavelet bases, replication.

[1] Laboratoire de Probabilités et Modèles Aléatoires, UMR 7599 du CNRS, Université Paris VI et Université Paris VII, 16 rue de Clisson, 75013 Paris, France.

[2] Université Paris X – Nanterre, 200 avenue de la République, 92001 Nanterre Cedex, France; e-mail: picard@math.jussieu.fr

linear smoothing methods under fairly large conditions (see Kerkyacharian and Picard [18], also Härdle *et al.*, Chap. 10 [14]). Moreover weak modifications of the Besov spaces also appear to be maximal sets for a large class of nonlinear methods including thresholding algorithms (see Kerkyacharian and Picard [16]). Hence these spaces, more than Sobolev or Hölder spaces –which are often used as a grid for regularity– appear to be more authentically linked with statistical problems, giving a "natural" way of measuring the sparsity.

If Besov spaces are very important, Besov bodies are also essential. They very well reflect the properties of thresholding algorithms with respect to more general bases than wavelet bases. For this reason, we set our results in the enlarged context of the multiscale type Besov bodies. This enables us to have very general examples such as multidimensional functions with anisotropic regularity –as images– or oscillatory functions –as chirps–.

The evaluation of the entropy of Besov balls goes back to 1967. It follows using interpolation theory from the result of Birman and Solomyak [4]. However, the proof presented there is rather long and difficult.

Moreover, taking advantage of the nice properties of the representation of functions of Besov spaces in wavelet expansion, it was legitimate to hope that entropy evaluation could be recovered using thresholding or m-term approximation methods. More than that these types of methods would hopefully provide in addition an optimal universal compression coding.

These nice ideas are developed in Donoho [12] using a coding deriving from a plain thresholding algorithm. However, the result is not completely optimal because of a logarithmic term appearing in the upper bound.

Birgé and Massart [3] suggested that the default of the method was in the plain thresholding and provided a new coding using a level-dependent thresholding taking advantage of an idea developed in Delyon and Juditski [8] to remove additional logarithmic terms in statistical applications.

Cohen *et al.* [6] provide a beautiful universal coding using tree structures. In particular, they use this specific method to encode the smallest tree containing the m-largest wavelet coefficients. They also recover the right upper bound without additional log-term.

In this short paper we prove that there is no need to modify the thresholding algorithm nor is it necessary to use a tree algorithm. To avoid the difficulty of the logarithmic term we use a replicant code which allows to send the addresses as well as the coefficients without losing length. This code has also the advantage of being easily protected. This ability is important since it is generally a weakness of wavelet codings to be very sensitive to errors in the first bits code.

To achieve this goal we put the problem in a setting which allows to treat the case of $\mathbb{L}_p$-norms as well as H_p-norms for $0 < p < 1$. This setting also enables us, using exponential type inequalities, to obtain the lower bounds in a very elementary way and in less than one page. It also notably enlarges the class of spaces for which the theorem is valid: if the classical Besov spaces are the prime example one can also consider spaces of "chirps" or multidimensional anisotropic regularity spaces.

The paper is organized in the following way: Section 2 quickly recalls the definitions of entropy and coding. Section 3 presents the analytical setting where we settle the problem. Section 4 contains the entropy result and the proof of the lower bound. Section 5 contains the replicant compression coding and the proof of the upper bound.

1. Metric entropy and coding

1.1. Metric entropy

Let us recall the following definitions: let (K, d) be a metric space. For every $\epsilon > 0$, we define $N(\epsilon, K, d)$ as the minimum number of balls of radius ϵ, covering K. We define the *metric entropy* of K as $H(\epsilon, K, d) = \log_2(N(\epsilon, K, d))$.

For (X, d) a metric space and K included in X, for every $\epsilon > 0$, we define $N(\epsilon, K, X, d)$ as the minimum number of balls of radius ϵ, centered in X, covering K. We define the *metric entropy relative to* X as

$H(\epsilon, K, X, d) = \log_2(N(\epsilon, K, X, d))$. If K is considered with the induced metric, we obviously have:

$$H(\epsilon, K, d) \geq H(\epsilon, K, X, d) \geq H(2\epsilon, K, d).$$

Because of the inequalities above, in the sequel we shall generally not distinguish between the two entropies.

1.2. **Coding**

Let (X, d) be a metric space and K be a subset of X. An *ϵ-coding of K of length l* is given by two functions:
$C : K \longrightarrow \{0,1\}^l$ (the "encoding" function) and
$D : \{0,1\}^l \longrightarrow X$ (the "decoding" function), such that

$$d(DC(x), x) \leq \epsilon.$$

Let us define $L(\epsilon, K, X, d)$ as the minimum length l of an ϵ-coding of K.

It is obvious that:

$$H(\epsilon, K, X, d) \leq L(\epsilon, K, X, d) \leq H(\epsilon, K, X, d) + 1.$$

2. Multiscale type Besov bodies

Let us now describe the type of function spaces that we are going to consider. As will become obvious, our framework will take classical Besov spaces as a model but leads to a much wider setting.

2.1. **Multiscale setting**

The framework is the following. Let X be a Banach space, or a $\tau-$Banach space, with $0 < \tau \leq 1$. (This means that instead of the usual triangular inequality we have $\|f + g\|^\tau \leq \|f\|^\tau + \|g\|^\tau$.) Typical examples are $X = \mathbb{L}^p$ for $1 \leq p \leq \infty$ and for $0 < p < 1$, $X = H_p$ (the Hardy space, and then $\tau = p$).

Let $\mathcal{E} = \{\psi_{j,k},\ j \in \mathbb{N},\ k \in \Lambda_j\}$ be a family in X with the following properties:

- for each $j \in \mathbb{N}$, Λ_j is a set of cardinality of order 2^{jd}. *i.e.* there exist two constants c_1, c_2 –not depending on j– such that $c_1 2^{jd} \leq \mathrm{Card}(\Lambda_j) \leq c_2 2^{jd}$ (d will be a dimension index);
- there exist a real number $0 < p \leq \infty$ and a constant $0 < C < \infty$, such that

$$\forall j \in \mathbb{N},\ \ \frac{1}{C}\left(\sum_{k\in\Lambda_j} |\beta_{j,k}|^p\right)^{1/p} \leq \left\|\sum_{k\in\Lambda_j} \beta_{j,k}\psi_{j,k}\right\|_X \leq C\left(\sum_{k\in\Lambda_j} |\beta_{j,k}|^p\right)^{1/p}. \qquad (1)$$

  We substitute the usual modification for $p = \infty$. The real number p will be of major importance in some cases. When necessary we shall precise that the setting is a p-multiscale setting.

Remarks. These properties obviously are verified for instance when $\mathcal{E}$ is a multiscale analysis associated to a compactly supported wavelet basis and $X = \mathbb{L}_p$ for $1 \leq p \leq \infty$, normalized in such a way that for all (j,k) we have $\|\psi_{j,k}\|_X = 1$.

However this condition is not necessary. Particularly we *do not need that* $\mathcal{E}$ is an unconditional basis of X not even a topological basis.

For instance the following family on $[0,1]$:

$$\psi_{jk}(x) = 2^{\frac{2j+1}{p}} I\left\{\left[\frac{k-1}{2^{2j+1}}, \frac{k}{2^{2j+1}}\right]\right\}(x),\ j \in \mathbb{N},\ k \in \mathbb{N},\ 1 + 2^{2j+1} - 2^{j+1} \leq k \leq 1 + 2^{2j+1} - 2^j$$

can be used to describe a phenomena on the unit interval which is more and more oscillating, like a "chirp". It satisfies the properties of a p-multiscale setting when the space X is $\mathbb{L}^p$ for $1 \leq p \leq \infty$. ($I\{B\}$ is the indicator function of the set B).

In the multidimensional framework, we can also proceed to build families of functions taking into account local anisotropy: let ψ be a wavelet function supported on $[0,1]$. Let $\psi_{j,k}(x) = 2^{\frac{j}{p}}\psi(2^j x - k)$ be the p-normalized family at the level j. Let us form the product

$$\psi_{j_1,k_1}(x_1)\dots\psi_{j_d,k_d}(x_d) = \Psi_I(x_1,\dots,x_d)$$

and let us index this product by its support: $I = [\frac{k_1}{2^{j_1}}, \frac{k_1+1}{2^{j_1}}] \times \dots \times [\frac{k_d}{2^{j_d}}, \frac{k_d+1}{2^{j_d}}]$. For each $j \in \mathbb{N}$, let us select Λ_j as a choice of hyperectangles $I_1,\dots,I_N$, such that $j_1+\dots+j_d$ is always equal to j (all rectangles have the same surface), $I_l \cap I_i = \emptyset$ unless $i = l$, and $\cup_{i=1}^N I_i = [0,1]^d$ (the hyperectangles are forming a partition of $[0,1]^d$). Of course, $N = 2^{jd}$ and it is not difficult to prove that such a family $\{\psi_I,\ I \in \Lambda_j,\ j \in \mathbb{N}\}$ will again satisfy the conditions above. Of course the uniform choice $j_1 = \dots = j_d = \frac{j}{d}$ corresponds to the isotropic case, but a choice introducing long and thin hyperectangles will be more suitable to handle anisotropic situations.

2.2. Multiscale Besov bodies

Our next step is to formulate the definition of Besov bodies associated to the previous multiscale setting:

Definition 1. For $0 < s < \infty,\ 0 < \pi \le \infty,\ 0 < r \le \infty,\ d \ge 1$ we define the following "multiscale Besov body" associated to the p-multiscale setting introduced above:

$$B^s_{\pi,r} = \left\{ f = \sum_{j=0}^{\infty}\sum_{k\in\Lambda_j}\beta_{j,k}\psi_{j,k};\ \left[\sum_{j=0}^{\infty}\left[2^{j(s+d(1/p-1/\pi))}\left(\sum_{k\in\Lambda_j}|\beta_{j,k}|^{\pi}\right)^{1/\pi}\right]^r\right]^{1/r} := \|f\|_{B^s_{\pi,r}} \right\} < \infty.$$

We substitute the obvious modifications for $r = \infty,\ \pi = \infty, p = \infty$.

Remarks.

(1) The following definition exactly corresponds to the usual characterization of standard Besov spaces $B_{s\pi r}$ when the multiscale setting is such that $X = \mathbb{L}_p([0,1]^d)$ and the family $\mathcal{E}$ is a multiscale analysis associated to a compactly supported wavelet basis normalized in $\mathbb{L}_p$ –see DeVore [10], DeVore *et al.* [9], Cohen *et al.* [5]–.

(2) As usual we have the standard embeddings:

$$\begin{aligned}
&\text{for}\quad 0 < s' \le s, \quad 0 < \pi' \le \pi \le \infty,\ \ 0 < r \le r', \quad && \|f\|_{B^{s'}_{\pi',r'}} \le \|f\|_{B^s_{\pi,r}};\\
&\text{for}\quad 0 < s' \le s, \quad 0 < \pi \le \pi' \le \infty,\ \ s - d/\pi = s' - d/\pi' \quad && \|f\|_{B^{s'}_{\pi',r}} \le \|f\|_{B^s_{\pi,r}}.
\end{aligned}$$

Let us now observe that there is generally no reasons for $B^s_{\pi,r}$ defined as above to be included in X. However, the following proposition proves that this occurs under the following condition:

Proposition 1. *If* $\delta := s - d(1/\pi - 1/p)_+ > 0$, *then* $B^s_{\pi,r} \subset X$.

Proof of Proposition. It is enough to prove that $\sum_j \|\sum_{k\in\Lambda_j}\beta_{j,k}\psi_{j,k}\|_X < \infty$ (or $\sum_j \|\sum_{k\in\Lambda_j}\beta_{j,k}\psi_{j,k}\|_X^{\tau} < \infty$ in the case of a $\tau-$Banach).

Our assumption implies $\|\beta_{j,.}\|_{l_\pi} = \epsilon_j 2^{-j(s+d(1/p-1/\pi))}$ with $\epsilon_. \in l_q$.

(1) $0 < \pi \le p \le \infty$.

$$\left\|\sum_{k\in\Lambda_j}\beta_{j,k}\psi_{j,k}\right\|_X \le C\|\beta_{j,.}\|_{l_p} \le C\|\beta_{j,.}\|_{l_\pi} \le C\epsilon_j 2^{-j(s+d(1/p-1/\pi))} = C\epsilon_j 2^{-j\delta}.$$

(2) $0 < p \le \pi \le \infty$. Using Hölder inequality, as $\mathrm{card}(\Lambda_j)$ is of order 2^{jd}:

$$\left\| \sum_{k\in\Lambda_j} \beta_{j,k}\psi_{j,k} \right\|_X \le C\|\beta_{j,.}\|_{l_p} \le C'\|\beta_{j,.}\|_{l_\pi} 2^{jd(1/p-1/\pi)} \le C'\epsilon_j 2^{-js} = C'\epsilon_j 2^{-j\delta}.$$

3. Main entropy result

The following theorem is the main result of this paper. It concerns the evaluations of the entropies for balls of the Besov bodies introduced above.

Theorem 1 (Birman and Solomyak)**.** *For $0 < p \le \infty$, $0 < s < \infty$, $0 < \pi \le \infty$, $0 < r \le \infty$, if $\delta = s - d(1/\pi - 1/p)_+ > 0$, then the unit ball $U^s_{\pi,r}$ of $B^s_{\pi,r}$ is such that there exist two constants $c(s,\pi,p) > 0$ and $C(s,\pi,p)$ such that*

$$c(s,\pi,p)\epsilon^{-d/s} \le H(\epsilon, U^s_{\pi,r}, X) \le C(s,\pi,p)\epsilon^{-d/s}.$$

Remark. We deduce from our proof the following values of the constants:

$$\begin{aligned} c(s,\pi,p) &= \frac{1}{8\log 2} 2^{d(1-2s/p)} C^{-d/s} 2^{-d} \\ C(s,\pi,p) &= 10\left[\frac{2^d}{2^d-1} + \frac{2^{\delta q}}{1-2^{-d\pi}}\right] + \frac{2(\delta\pi + d)2^{-\delta\pi}}{(1-2^{-\delta\pi})^2} + \frac{4}{2^d-1} + \frac{2}{1-2^{-d}}\left(\frac{d2^{-d}}{q(1-2^{-d})} + \frac{1}{\pi}\right) \\ q &= \frac{d}{s+d/p}\cdot \end{aligned}$$

However, they are in no way to be considered as optimal so it is difficult to draw conclusions upon them.

3.1. **Lower bound**

Proposition 2. *There exists a constant $c(s,\pi,r) > 0$ such that: $\forall\epsilon > 0$, $H(\epsilon, U^s_{\pi,r}, X) \ge c(s,\pi,r)\epsilon^{-d/s}$.*

Proof of Proposition 2. As $H(\epsilon, U^s_{\pi,r}, X)$ is a non decreasing function of ϵ, it is enough to find a non increasing sequence of non negative numbers $(\epsilon_j)_{j\in\mathbb{N}}$, such that $\lim \epsilon_j = 0$, $\frac{\epsilon_j}{\epsilon_{j+1}} \le A < \infty$ and $H(\epsilon_j, U^s_{\pi,r}, X) \ge K\epsilon_j^{-d/s}$. Let us consider the following set:

$$\mathcal{A}_j = \left\{ 2^{-j(s+d/p)} \sum_{k\in\Lambda_j} \delta_k \psi_{j,k}, \quad \delta_k \in \{0,1\} \right\}\cdot$$

Obviously, for any j in $\mathbb{N}$, $\mathcal{A}_j \subset U^s_{\pi,q}$, so $H(\epsilon, U^s_{\pi,r}, X) \ge H(\epsilon, \mathcal{A}_j, X)$.

Proposition 3. *The set $\Omega_n = \{0,1\}^n$, with the l_1 distance: $\|\omega - \omega'\| = \sum_{i=1}^n |\omega_i - \omega'_i|$ satisfies:*

$$H(n/4, \Omega_n, l_1) \ge \frac{n}{8\log(2)}\cdot$$

Proof of Proposition 3. This proposition is well known (Varshamov–Gilbert lemma, see [7]). We give here a sketch of proof for the reader's convenience.

Let P be the uniform probability on Ω_n. The coordinate functions $X_i(\omega) = \omega_i$ are then independent Bernoulli random variables. Let us consider a covering of Ω_n, by N balls B_j of radius $n/4$. We have

$$1 = P(\Omega_n) \le \sum_{j=1}^N P(B_j) = N\ P(B(0, n/4)),$$

as obviously all the balls B_j have the same probability. But,

$$P(B(0,n/4)) = P\left(\sum_{i=1}^{n} X_i \le n/4\right) = P\left(\sum_{i=1}^{n}(1/2 - X_i) \ge n/4\right) \le \exp(-n/8),$$

using Hoeffding inequality (see [13] or [24]). This ends the proof of Proposition 3.

Proof of Proposition 2 (continuing). Let us prove that: for every $\epsilon > 0$, we have

$$H(\epsilon, \mathcal{A}_j, X) \ge H((C\epsilon)^p 2^{j(sp+d)}, \Omega_{2^{jd}}, l_1). \tag{2}$$

Let us consider a covering of $\mathcal{A}_j$ by N balls of radius ϵ centered on $\mathcal{A}_j$. As we have

$$\left\| 2^{-j(s+\frac{d}{p})} \sum_{k\in\Lambda_j} \delta_k \psi_{j,k} - 2^{-j(s+\frac{d}{p})} \sum_{k\in\Lambda_j} \delta'_k \psi_{j,k} \right\|_X \ge \frac{1}{C} 2^{-j(s+\frac{d}{p})} \left(\sum_{k\in\Lambda_j} |\delta_k - \delta'_k|^p\right)^{1/p},$$

it is clear that this covering corresponds to a covering of $\{0,1\}^{\Lambda_j}$ by N balls of radius less then $(C\epsilon)^p 2^{j(sp+d)}$, with the l_1 distance. This implies (2).

Let us now choose ϵ_j such that $(C\epsilon_j)^p 2^{j(sp+d)} = \frac{2^{jd}}{4}$. Using the previous proposition we get:

$$H(\epsilon_j, \mathcal{A}_j, X) \ge \frac{2^{jd}}{8\log(2)} \ge K\epsilon_j^{-d/s}.$$

(Implicitly we took here $0 < p < \infty$. The case $p = \infty$ is simpler and can be handled directly.)

We deduce from our proof the following values of the constants: $c(s,\pi,p) = \frac{1}{8\log 2} 2^{d(1-2s/p)} C^{-d/s} 2^{-d}$ if we notice that we can take the constant A defined above as equal to 2^s.

4. Replicant coding and upper bound

As was mentioned above, the upper bound will follow from the construction of a coding procedure. Moreover, as $U^s_{\pi,r}$ is included in $U^s_{\pi,\infty}$, it is enough to consider the problem for $U^s_{\pi,\infty}$.

4.1. Quantization algorithm

Let us introduce the following quantization procedures:

Definition 2. For $0 < \lambda < \infty$, $\beta \in \mathbb{R}$, we define $Q_\lambda(\beta) = \operatorname{sign}(\beta)\left[\frac{|\beta|}{\lambda}\right]\lambda$, where $[x]$ denotes the integer part of $x \in \mathbb{R}^+$.

For $f = \sum_{j=0}^{\infty} \sum_{k\in\Lambda_j} \beta_{j,k}\psi_{j,k}$, we define:

$$Q_\lambda(f) = \sum_{j=0}^{\infty} \sum_{k\in\Lambda_j} Q_\lambda(\beta_{j,k})\psi_{j,k}$$

$$Q^J_\lambda(f) = \sum_{j=0}^{J} \sum_{k\in\Lambda_j} Q_\lambda(\beta_{j,k})\psi_{j,k}.$$

The following theorem describes the rates of approximation of the procedures described above when the object has a $B^s_{\pi,\infty}$ regularity:

Theorem 2. *For* $0 < p \leq \infty,\ 0 < s < \infty,\ 0 < \pi \leq \infty,\ 0 < r \leq \infty,\ d \geq 1,$ *and* $\delta = s - d(1/\pi - 1/p)_+ > 0$, *let us define* q *by:*

$$s + \frac{d}{p} = \frac{d}{q}\cdot$$

There exists a constant D *depending only on* $p, \pi,$ *and* s, *such that if* $\|f\|_{B^s_{\pi,\infty}} \leq 1$ *then, for any* $\lambda > 0$,

$$\mathrm{card}\{(j,k), Q_\lambda(\beta_{j,k}) \neq 0\} = \mathrm{card}\{(j,k), |\beta_{j,k}| \geq \lambda\} \leq D_1 \lambda^{-q}. \tag{3}$$

Moreover, we have:

$$\|f - Q_\lambda(f)\|_X \leq D_2 \lambda^{\frac{sq}{d}} \tag{4}$$

$$\|f - Q^J_\lambda(f)\|_X \leq D_2 \left(\lambda^{\frac{sq}{d}} + 2^{-J\delta}\right). \tag{5}$$

Proof of Theorem 2. It is enough to give the proof in the case $0 \leq \pi \leq p \leq \infty$. Since, for $\pi > p$ we have $\|f\|_{B^s_{p,\infty}} \leq \|f\|_{B^s_{\pi,\infty}}$ and we can take $\pi = p$. The case $\pi = p = \infty$ is easy to verify directly. Hence, we shall assume in the sequel that $0 \leq \pi \leq p \leq \infty,\ \pi < \infty$. So $\delta = s - d(1/\pi - 1/p) > 0$ and using the hypothesis $(\sum_{k \in \Lambda_j} |\beta_{j,k}|^\pi)^{1/\pi} \leq 2^{-j(s+d(1/p-1/\pi))} = 2^{-j\delta}$, we have:

(1)

$$\mathrm{card}\{k \in \Lambda_j, |\beta_{j,k}| \geq \lambda\} \leq 2^{jd} \wedge \frac{2^{-j\delta\pi}}{\lambda^\pi}$$

$$\text{and, } \mathrm{card}\{(j,k), |\beta_{j,k}| \geq \lambda\} \leq \sum_{j \geq 0} 2^{jd} \wedge \frac{2^{-j\delta\pi}}{\lambda^\pi}\cdot$$

Let J_0 such that $2^{J_0 d} \sim \frac{2^{-J_0\delta\pi}}{\lambda^\pi}$ $(i.e.\ 2^{J_0 d} \leq \frac{2^{-J_0\delta\pi}}{\lambda^\pi} \leq \lambda^{\frac{-\pi}{d+\delta\pi}} 2^{(J_0+1)d})$

$$\mathrm{card}\{(j,k), |\beta_{j,k}| \geq \lambda\} \leq \sum_{0 \leq j \leq J_0} 2^{jd} + \sum_{J_0 < j} \frac{2^{-j\delta\pi}}{\lambda^\pi} \leq D_1 \lambda^{\frac{-d\pi}{d+\delta\pi}}$$

with $D_1 = \frac{2^d}{2^d - 1} + \frac{2^{\delta q}}{1 - 2^{-d\pi}}$ since one easily verifies that $\frac{d\pi}{d+\delta\pi} = q$.

(2) Let us suppose that X is a Banach space. (The τ-Banach case does not lead to additional difficulty.) Let us also suppose that p is finite: the case $p=\infty$ is let to the reader.

$$\begin{aligned}
\|f-Q_\lambda(f)\|_X &\le \sum_{j\ge 0}\left\|\sum_{k\in\Lambda_j}(Q_\lambda(\beta_{j,k})-\beta_{j,k})\psi_{j,k}\right\|_X\\
&\le C\sum_{j\ge 0}\left[\sum_{k\in\Lambda_j}|Q_\lambda(\beta_{j,k})-\beta_{j,k}|^p\right]^{1/p}\\
&\le C\sum_{j\ge 0}\left[\sum_{k\in\Lambda_j,|\beta_{j,k}|<\lambda}|\beta_{j,k}|^p+\lambda^p\mathrm{card}\{k\in\Lambda_j,|\beta_{j,k}|\ge\lambda\}\right]^{1/p}\\
&\le C\sum_{j\ge 0}\left[\sum_{k\in\Lambda_j,|\beta_{j,k}|<\lambda}|\beta_{j,k}|^p\right]^{1/p}+C\lambda\sum_{j\ge 0}[\mathrm{card}\{k\in\Lambda_j,|\beta_{j,k}|\ge\lambda\}]^{1/p}.
\end{aligned}$$

As $\pi\le p$, we have:

$$\left[\sum_{k\in\Lambda_j,|\beta_{j,k}|<\lambda}|\beta_{j,k}|^p\right]^{1/p}=\left[\sum_{k\in\Lambda_j,|\beta_{j,k}|<\lambda}|\beta_{j,k}|^{p-\pi}|\beta_{j,k}|^\pi\right]^{1/p}\le\lambda 2^{jd/p}\wedge\lambda^{1-p/\pi}2^{-j\delta\pi/p}.$$

Using (3), we get:

$$\begin{aligned}
\|f-Q_\lambda(f)\|_X &\le C\sum_{j\ge 0}\lambda 2^{jd/p}\wedge\lambda^{1-p/\pi}2^{-j\delta\pi/p}+C\lambda\sum_{j\ge 0}\left[2^{jd}\wedge\frac{2^{-j\delta\pi}}{\lambda^\pi}\right]^{1/p}\\
&=2C\lambda\sum_{j\ge 0}\left[2^{jd}\wedge\frac{2^{-j\delta\pi}}{\lambda^\pi}\right]^{1/p}\\
&\le 2C\lambda\left[\sum_{0\le j\le J_0}2^{jd/p}+\frac{1}{\lambda^{\pi p}}\sum_{J_0<j}2^{-j\delta\pi/p}\right]
\end{aligned}$$

where as previously, we have chosen $2^{J_0d}\sim\frac{2^{-J_0\delta\pi}}{\lambda^\pi}$. So

$$\|f-Q_\lambda(f)\|_X\le D_2\lambda^{\frac{p-q}{p}}=D_2\lambda^{sq}$$

with $D_2=[2C(\frac{2^{d/p}}{2^{d/p}-1}+\frac{2^{\delta q/p}}{1-2^{\delta\pi/p}})]$.

(3)

$$\|f-Q^J_\lambda(f)\|_X\le\|f-Q_\lambda(f)\|_X+\|Q_\lambda(f)-Q^J_\lambda(f)\|_X$$

$$\begin{aligned}
\|Q_\lambda(f)-Q^J_\lambda(f)\|_X &\le\sum_{j>J}\left\|\sum_{k\in\Lambda_j}Q_\lambda(\beta_{j,k})\psi_{j,k}\right\|_X\\
&\le C\sum_{j>J}\left[\sum_{k\in\Lambda_j}|Q_\lambda(\beta_{j,k})|^p\right]^{1/p}\le C\sum_{j>J}\left[\sum_{k\in\Lambda_j}|\beta_{j,k}|^p\right]^{1/p}.
\end{aligned}$$

But as $s - d/\pi = \delta - d/p$ and $\pi \le p$, we have $\|f\|_{B^\delta_{p,\infty}} \le \|f\|_{B^s_{\pi,\infty}} \le 1$, so

$$\sum_{j>J} \left[\sum_{k \in \Lambda_j} |\beta_{j,k}|^p \right]^{1/p} \le \sum_{j>J} 2^{-j\delta} \le \frac{1}{1-2^{-\delta}} 2^{-J\delta}.$$

4.2. **Replicant universal coding**

We use a procedure inspired by [12] and which has been improved in [17]. Let us consider:

$$Q^J_\lambda(f) = \sum_{j=0}^{J} \sum_{k \in \Lambda_j} Q_\lambda(\beta_{j,k}) \psi_{j,k}$$

with the tuning constants J and λ defined by: $\frac{\epsilon}{2D} = 2^{-J\delta} = \lambda^{\frac{sq}{d}}$. This choice ensures that:

$$\|f - Q^J_\lambda(f)\|_X \le \epsilon.$$

Now we need to encode $Q^J_\lambda(f)$ and compute the length of this ϵ-coding.

Let us first explain why, it is needed to improve the procedure in [12]: if we use a binary representation of $[\frac{|\beta_{j,k}|}{\lambda}]$, this will cost for each (j,k) such that $|\beta_{j,k}| \ge \lambda$, $1 + \log_2([\frac{|\beta_{j,k}|}{\lambda}])$ bits. But for each $0 \le j$:

$$\sup_{k \in \Lambda_j} |\beta_{j,k}| \le \left(\sum_{k \in \Lambda_j} |\beta_{j,k}|^\pi \right)^{1/\pi} \le 2^{-j\delta}.$$

So we obtain a bound of order $\log_2([\frac{1}{\lambda}]) \sim \log_2([\frac{1}{\epsilon}])$. Using Theorem 2, we know that we have of order $\lambda^{-q} \sim \epsilon^{-d/s}$ of such terms ($|\beta_{j,k}| \ge \lambda$). We have in addition to encode the signs and the addresses of such β's. This will have a cost of the order of $Jd \sim \log_2([\frac{1}{\epsilon}])$ bits. So if we keep in advance a fixed number of bits to encode the addresses and the $Q_\lambda(\beta_{j,k})$'s we shall use up to constants, $\epsilon^{-d/s} \log_2([\frac{1}{\epsilon}])$ bits. This obviously gives an undesirable extra logarithmic term.

So instead of keeping in advance each time a fixed allocation for the addresses, and the $Q_\lambda(\beta_{j,k})$'s, we shall encode them dynamically: first the sign, then the binary representation of $[\frac{|\beta_{j,k}|}{\lambda}]$ and then the difference between two successive adresses. To do this, we first suppose that (j,k) is the number $2^{jd} + k$. Then, we obviously need a separator between each of the triples (sign, $[\frac{|\beta_{j,k}|}{\lambda}]$, address). For this purpose, we use 01 as separator, and we replicate each bit in the binary expansion of the triple. (For instance, the expansion 0110100 becomes 00111100110000). This obviously gives us an injective coding which also has the advantage of being easily protected. For instance, using 0101 as separator instead of 01 even increases the protection.

Let us now be more precise about the way of encoding the addresses: let

$$\Lambda'_j = \{k \in \Lambda_j,\ |\beta_{j,k}| \ge \lambda\} = \{k_{1,j}, \dots, k_{n_j,j}\};\quad 0 \le k_{i,j} < 2^{jd};\quad n_j = \mathrm{Card}(\Lambda'_j) \le 2^{jd} \wedge \frac{2^{-j\delta\pi}}{\lambda^\pi}.$$

Let us define the representation $\alpha[(j, k_{l,j})]$ for $k_{l,j} \in \Lambda'_j$ –assuming then that $\Lambda'_j \ne \emptyset$–. Because we encode the difference between two successive addresses, there will be a difference between the cases $l = 1$ and $l > 1$.

For $l = 1$, $j \ne 0$, let us introduce the previous –non void– level to be encoded: $j' = \sup\{i < j \ /\Lambda'_i \ne \emptyset\}$. Then,

$$\alpha[(j, k_{1,j})] = 2^j + k_{1,j} - (2^{j'} + k_{n_{j'},j'}).$$

For $l = 1,\ j = 0$,

$$\alpha[(0, k_{1,0})] = k_{1,0}.$$

For $l > 1$, we stay on the same level and put $\alpha[(j, k_{l,j})] = 2^j + k_{l,j} - (2^j + k_{l-1,j}) = k_{l,j} - k_{l-1,j}$.

Let us now calculate the length of the coding. It is obviously less than:

$$6\text{card}\{(j,k), |\beta_{j,k}| \geq \lambda\} + 2\sum_{j=0}^{J}\sum_{k\in\Lambda'_j}\left\{\log_2\left[\frac{|\beta_{j,k}|}{\lambda}\right] + 1\right\} + 2\sum_{j=0}^{J}\sum_{l=1}^{n_j}\{\log_2 \alpha[(j, k_{l,j})] + 1\}$$

$$= 10\text{card}\{(j,k), |\beta_{j,k}| \geq \lambda\} + 2\sum_{j=0}^{J}\sum_{k\in\Lambda'_j}\log_2\left(\left[\frac{|\beta_{j,k}|}{\lambda}\right]\right) + 2\sum_{j=0}^{J}\sum_{l=1}^{n_j}\log_2(\alpha[(j,k_l)]).$$

As using Theorem 2, we have $\text{card}\{(j,k), |\beta_{j,k}| \geq \lambda\} \leq D_1\lambda^{-q} = D_1\epsilon^{-d/s}$, we now need to prove the two following bounds:

$$\sum_{j=0}^{J}\sum_{k\in\Lambda'_j}\log_2\left[\frac{|\beta_{j,k}|}{\lambda}\right] \leq D_4\epsilon^{-d/s} \tag{6}$$

$$\sum_{j=0}^{J}\sum_{l=1}^{n_j}\log_2\alpha[(j,k_{l,j})] \leq D_7\epsilon^{-d/s}. \tag{7}$$

Inequality (6):

$$\sum_{k\in\Lambda'_j}\log_2\left[\frac{|\beta_{j,k}|}{\lambda}\right] \leq \sum_{k\in\Lambda'_j}\log_2\frac{|\beta_{j,k}|}{\lambda} \leq \frac{1}{\pi}\text{Card}(\Lambda'_j)\ \frac{1}{\text{card}(\Lambda'_j)}\sum_{k\in\Lambda'_j}\log_2\frac{|\beta_{j,k}|^{\pi}}{\lambda^{\pi}}\cdot$$

Using Jensen inequality, and the fact that $\log_2$ is concave, we can bound the last quantity by:

$$\leq \frac{1}{\pi}\text{Card}(\Lambda'_j)\log_2\left(\frac{1}{\text{Card}(\Lambda'_j)}\sum_{k\in A'_j}\left[\frac{|\beta_{j,k}|^{\pi}}{\lambda^{\pi}}\right]\right) \leq \frac{1}{\pi}\text{Card}(\Lambda'_j)\log_2\left(\frac{2^{-j\delta\pi}}{\text{Card}(\Lambda'_j)\ \lambda^{\pi}}\right)\cdot$$

Let us recall that

$$\text{Card}(\Lambda'_j) \leq 2^{jd} \wedge \frac{2^{-j\delta\pi}}{\lambda^{\pi}} \tag{8}$$

$$\text{and}\quad \sup_{0\leq x\leq a\wedge K} x\log_2\frac{K}{x} = 1_{ae\leq K}\ a\log_2\frac{K}{a} + 1_{ae\geq K}\frac{\log_2 e}{e}K \leq \frac{\log_2 e}{e}K. \tag{9}$$

Let us choose J_0 as before ($2^{J_0 d} \sim \frac{2^{-J_0\delta\pi}}{\lambda^{\pi}}$). We have:

$$\sum_{j=0}^{J}\sum_{k\in\Lambda'_j}\log_2\left(\left[\frac{|\beta_{j,k}|}{\lambda}\right]\right) \leq \sum_{0\leq j<J_0}\frac{1}{\pi}2^{jd}\log_2\left(\frac{2^{-j\delta\pi}}{2^{jd}\ \lambda^{\pi}}\right) + \sum_{j=J0}^{J}\frac{1}{\pi}\frac{\log_2 e}{e}\ \frac{2^{-j\delta\pi}}{\lambda^{\pi}}$$

$$\leq \frac{1}{\pi}\left[\sum_{0\leq j<J_0}2^{jd}(J_0 - j)(\delta\pi + d) + 2^{J_0 d}\sum_{j=J0}^{J}2^{-(j-J_0)\delta\pi}\right]$$

$$\leq \frac{1}{\pi}\left[2^{J_0 d}\left((\delta\pi + d)\sum_{0}^{\infty}j2^{-jd} + \sum_{0}^{\infty}2^{-j\delta\pi}\right)\right] \leq D_4\epsilon^{-d/s}.$$

If we recall that we have $s+\frac{d}{p}=\delta+\frac{d}{\pi}$, and denote:

$$D_4=\frac{1}{1-2^{-d}}\left(\frac{d2^{-d}}{q(1-2^{-d})}+\frac{1}{\pi}\right)\cdot$$

Inequality (7): Again, we shall see that it is necessary to separate the first term ($l=1$) and the other ones. More precisely:

$$\sum_{j=0}^{J}\log_2\{\alpha[(j,k_{1,j})]\}\leq KJ^2=\mathcal{O}(\log(\epsilon)^2).$$

This term is obviously not significant. Whereas,

$$\begin{aligned}\sum_{l=2}^{n_j}\log_2\{\alpha[(j,k_{l,j})]\} &= n_j\frac{1}{n_j}\sum_{l=2}^{n_j}\log_2\{\alpha[(j,k_{l,j})]\}\\ &\leq n_j\log_2\left\{\frac{1}{n_j}\sum_{l=2}^{n_j}\alpha[(j,k_{l,j})]\right\}\leq n_j\log_2\frac{2^{jd}}{n_j}\cdot\end{aligned}$$

Using (8) and (9), we have: for $0\leq j\leq J_0,\quad n_j\log_2(\frac{2^{jd}}{n_j})\leq\frac{\log_2 e}{e}2^{jd}$

$$\text{so}\quad\sum_{0\leq j\leq J_0}n_j\log_2\left(\frac{2^{jd}}{n_j}\right)\leq\sum_{0\leq j\leq J_0}\frac{\log_2 e}{e}2^{jd}\leq D_5\epsilon^{-d/s},$$

with

$$D_5\leq\frac{2}{2^d-1}$$

and for $J_0<j\leq J$,

$$\begin{aligned}n_j\log_2\left(\frac{2^{jd}}{n_j}\right) &\leq\frac{2^{-j\delta\pi}}{\lambda^\pi}\log_2\left(2^{jd}\frac{\lambda^\pi}{2^{-j\delta\pi}}\right)\\ &\leq 2^{J_0d}2^{-(j-J_0)\delta\pi}\ (j-J_0)(\delta\pi+d)\\ &\leq\epsilon^{-d/s}2^{-(j-J_0)\delta\pi}\ (j-J_0)(\delta\pi+d)\end{aligned}$$

$$\text{so}\quad\sum_{J_0<j\leq J,}n_j\log_2\left(\frac{2^{jd}}{n_j}\right)\leq\sum_{J_0<j\leq J,}\epsilon^{-d/s}2^{-(j-J_0)\delta\pi}\ (j-J_0)(\delta\pi+d)\leq D_6\epsilon^{-d/s},$$

with,

$$D_6=(\delta\pi+d)\frac{2^{-\delta\pi}}{(1-2^{-\delta\pi})^2}\cdot$$

So:

$$\sum_{j=0}^{J}\sum_{k\in\Lambda'_j}\log_2\{\alpha[(j,k)]\}\leq\sum_{0\leq j\leq J}n_j\log_2\left(\frac{2^{jd}}{n_j}\right)+\sum_{j=0}^{J}\log_2\{\alpha[(j,k_1)]\}\leq D_7\epsilon^{-d/s},$$

with $D_7=D_5+D_6$.

Finally, we get,

$$H(\epsilon,X,d)\leq L(\epsilon,X,d)\leq C(s,\pi,p)\epsilon^{-d/s},$$

with,

$$\begin{aligned} C(s,\pi,p) &\le 10D_1 + 2D_7 + 2D_4 \\ &= 10\left[\frac{2^d}{2^d-1} + \frac{2^{\delta q}}{1-2^{-d\pi}}\right] + \frac{2(\delta\pi+d)2^{-\delta\pi}}{(1-2^{-\delta\pi})^2} + \frac{4}{2^d-1} + \frac{2}{1-2^{-d}}\left(\frac{d2^{-d}}{q(1-2^{-d})} + \frac{1}{\pi}\right)\cdot \end{aligned}$$

□

References

[1] P. Assouad, Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sr. I Math.* **296** (1983) 1021-1024.
[2] L. Birgé, Sur un thorme de minimax et son application aux tests. *Probab. Math. Statist.* **3** (1984) 259-282.
[3] L. Birgé and P. Massart, An adaptative compression algorithm in Besov spaces. *Constr. Approx.* **16** (2000) 1-36.
[4] M.S. Birman and M.Z. Solomiak, Piecewise-polynomial approximation of functions of the classes W_p. *Mat. Sbornik* **73** (1967) 295-317.
[5] A. Cohen, R. DeVore and W. Dahmen, Multiscale methods on bounded domains. *Trans. AMS* **352** (2000) 3651-3685.
[6] A. Cohen, W. Dahmen, I. Daubechies and R. DeVore, Tree approximation and optimal encoding. *Appl. Comput. Harmon. Anal.* **11** (2001) 192-226.
[7] T.A. Cover and J.A. Thomas, *Element of Information Theory.* Wiley Interscience (1991).
[8] B. Delyon and A. Juditski, On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** (1996) 215-228.
[9] R. DeVore, R. Kyriazis and P. Wang, Multiscale characterization of Besov spaces on bounded domains. *J. Approx. Theory* **93** (1998) 273-292.
[10] R. DeVore, *Nonlinear approximation.* Cambridge University Press, *Acta Numer.* **7** (1998) 51-150.
[11] R. DeVore and G. Lorentz, *Constructive Approximation.* Springer-Verlag, New York (1993).
[12] D.L. Donoho, Unconditional bases and bit-level compression. *Appl. Comput. Harmon. Anal.* **3** (1996) 388-392.
[13] W. Hoeffding, Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58** (1963) 13-30.
[14] W. Härdle, G. Kerkyacharian, D. Picard and A. Tsybakov, *Wavelet, Approximation and Statistical Applications.* Springer Verlag, New York, *Lecture Notes in Statist.* **129** (1998).
[15] G. Kerkyacharian and D. Picard, Thresholding algorithms, maxisets and well-concentrated bases, with discussion. *Test* **9** (2000) 283-345.
[16] G. Kerkyacharian and D. Picard, Minimax or maxisets? *Bernoulli* **8** (2002) 219-253.
[17] G. Kerkyacharian and D. Picard, *Entropy, Universal coding, Approximation and bases properties.* Technical Report (2001).
[18] G. Kerkyacharian and D. Picard, Density Estimation by Kernel and Wavelets methods – Optimality of Besov spaces. *Statist. Probab. Lett.* **18** (1993) 327-336.
[19] A.N. Kolmogorov and V.M. Tikhomirov, ϵ-entropy and ϵ-capacity. *Uspekhi Mat. Nauk* **14** (1959) 3-86. (Engl. Translation: *Amer. Math. Soc. Transl. Ser. 2* **17**, 277-364.)
[20] L. Le Cam, Convergence of estimator under dimensionality restrictions. *Ann. Statist.* **1** (1973) 38-53.
[21] L. Le Cam, Metric dimension and statistical estimation, in *Advances in mathematical sciences: CRM's 25 years.* Montreal, PQ (1994) 303-311.
[22] G.G. Lorentz, Metric entropy and approximation. *Bull. Amer. Math. Soc.* **72** (1966) 903-937.
[23] S.M. Nikolskii, *Approximation of functions of several variables and imbedding theorems* (Russian), Second Ed. Moskva, Nauka (1977). English translation of the first Ed., Berlin (1975).
[24] V.V. Petrov, *Limit Theorems of Probability Theory: Sequences of independent Random Variables.* Oxford University Press (1995).
[25] S.A. van de Geer, *Empirical processes in M-estimation.* Cambridge University Press (2000).