ESAIM: PS 17 (2013) 472–484
DOI: 10.1051/ps/2012002

ESAIM: Probability and Statistics
www.esaim-ps.org

WHY MINIMAX IS NOT THAT PESSIMISTIC*

Aurelia Fraysse[1]

Abstract. In nonparametric statistics a classical optimality criterion for estimation procedures is provided by the minimax rate of convergence. However this point of view can be subject to controversy as it requires to look for the worst behavior of an estimation procedure in a given space. The purpose of this paper is to introduce a new criterion based on generic behavior of estimators. We are here interested in the rate of convergence obtained with some classical estimators on almost every, in the sense of prevalence, function in a Besov space. We also show that generic results coincide with minimax ones in these cases.

Mathematics Subject Classification. 62C20, 28C20, 46E35.

Received April 21, 2011.

1. Introduction

Since its introduction in the seventies, nonparametric estimation has taken a large place in the work of mathematical or signal processing communities. Often a signal is too wide to be closely approximated by parametric estimators. Therefore new estimation procedures, based on approximation of functions have been introduced. But which kind of estimator is the most appropriate in these cases?

This question raised a lot of definitions and discussions in the statistical community. How can two estimators be compared when they point out infinite dimensional objects and what kind of optimal behavior can be expected. One of the most common ways to test the performance of a procedure is to compare its convergence rate with an optimal one given by minimax theory. Nonetheless, this technique comes from a particular definition which can be subject to controversy. Indeed, in the minimax theory we are looking to the estimation procedure which yields the minimum of a maximum risk, in a sense to be defined, over a function space. The main drawback is the pessimistic point of view of this theory, which looks for the worst rate of estimation obtained in a given space. But the worst case could be a misleading one and a method can be rejected although it is a good one for a lot of functions. The purpose of this paper is to introduce a new test of the risk, obtained thanks to genericity results. Thanks to this new kind of test we show that in fact minimax risk corresponds to a generic one.

Let us first introduce what is meant by a generic set of function in this paper. In a finite dimensional space, we say that a property holds almost everywhere if the set of points where it is not true is of vanishing Lebesgue measure. The Lebesgue measure has here a preponderant role, as it is the only σ-finite and translation invariant measure. Unfortunately, no measure share those properties in infinite dimensional Banach spaces. A way to

Keywords and phrases. Minimax theory, maxiset theory, Besov spaces, prevalence, wavelet bases.

* *This work was performed when the author was at LTCI, Telecom ParisTech.*

[1] L2S, SUPELEC, CNRS, University Paris-Sud, 3 rue Joliot-Curie, 91190 Gif-Sur-Yvette, France. fraysse@lss.supelec.fr

Article published by EDP Sciences

© EDP Sciences, SMAI 2013

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

recover a natural "almost every" notion in infinite vector spaces is thus defined as follows by Christensen in 1972 see [2,4,13].

Definition 1.1. Let V be a complete metric vector space. A Borel set $A \subset V$ is Haar-null (or shy) if there exists a compactly supported probability measure μ such that

$$\forall x \in V, \quad \mu(x + A) = 0. \tag{1.1}$$

If this property holds, the measure μ is said to be transverse to A.

A subset of V is called Haar-null if it is contained in a Haar-null Borel set. The complement of a Haar-null set is called a prevalent set.

As it can be seen in the definition of prevalence, the main issue in proofs is to construct transverse measures to a Borel Haar-null set. We remind here two classical ways to construct such a measure.

Remark 1.2.

(1) A finite dimensional subspace of V, P, is called a probe for a prevalent set $T \subset V$ if the Lebesgue measure on P is transverse to the complement of T.
This measure is not a compactly supported probability measure. However one immediately checks that this notion can be defined in the same way but stated with the Lebesgue measure defined on the unit ball of P. Note that in this case, the support of the measure is included in the unit ball of a finite dimensional subspace. The compactness assumption is therefore fulfilled.

(2) If V is a function space, a probability measure on V can be defined by a random process X_t whose sample paths are almost surely in V. The condition $\mu(f + A) = 0$ means that the event $X_t - f \in A$ has probability zero. Therefore, a way to check that a property $\mathcal{P}$ holds only on a Haar-null set is to exhibit a random process X_t whose sample paths are in V and such that

$$\forall f \in V, \text{ a.s.} \quad X_t + f \ \text{ does not satisfy } \ \mathcal{P}.$$

The fact that a set is Haar-null is independent of the chosen transverse measure, as soon as the translation invariance condition is satisfied. Furthermore, as proved in [7], if a set is Haar null the set of its transverse measures is generic in the Baire's category sense. However this property cannot provide the exact characterization of null sets.

The following results enumerate important properties of prevalence and show that these notions supply a natural generalization of "zero measure" and "almost every" in finite-dimensional spaces, see [2,4,13].

Proposition 1.3. *Let V be a complete metric vector space. Then the following properties holds:*

- *if S is Haar-null, then $\forall x \in V$, $x + S$ is Haar-null;*
- *if $\dim(V) < \infty$, S is Haar-null if and only if $\text{meas}(S) = 0$ (where meas denotes the Lebesgue measure);*
- *prevalent sets are dense;*
- *the intersection of a countable collection of prevalent sets is prevalent;*
- *if V is infinite dimensional, compact subsets of V are Haar-null.*

As we can see from the properties of prevalent sets, this theory provides a natural generalization of the finite dimensional notion of almost everywhere. Since its definition, it has been mainly used in the context of differential geometry [13] and regularity type properties [12]. A classical example is given in [12], where it is proved that the set of nowhere differentiable functions is prevalent in the space of continuous functions. Surprisingly even in the finite dimensional case, genericity approach has no longer been studied in statistics. The only actual result involving genericity in statistics is due to Doob, see [28] in the context of Bayesian estimation in parametric statistics.

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

Using this theory, a natural way to exhibit a test of performance for an estimation procedure would be to look at the risk reached on almost every function of a function space, in the sense of prevalence.

In this paper we study performances in terms of generic approximation of two classical estimation procedures in the white noise model in Besov spaces. With these two techniques, we will see that minimax and generic results coincide.

2. Models and estimation procedures

In the following, we consider the classical Gaussian white noise model. Following the definition of [14], we suppose that we observe Y_t such that

$$\mathrm{d}Y_t = f(t)\mathrm{d}t + \frac{1}{\sqrt{n}}\mathrm{d}W_t, \quad t \in (0,1)^d, \tag{2.1}$$

where dW_t stands for the d-dimensional Wiener measure, n is known and f is the unknown function to be estimated.

The estimation procedures that we deal with are defined thanks to a decomposition of the functions to be estimated. To define them, we first introduce the wavelet bases. In our framework, those bases allow both to define function spaces and estimation procedures. It provides thus a key tool to introduce our results. The wavelet transform is a powerful approximation tool widely used in statistics and signal processing, thanks to its properties of localization in time and frequency domains. Indeed, this property allows to reconstruct a signal with few coefficients. Its use in statistical studies and the development of wavelet based estimators are thus natural, as introduced in [20].

To define wavelets, we refer to [6] where it is proved that for r large enough there exists 2^d-1 functions $\psi^{(i)}$ with compact support and which are r regular. Furthermore each $\psi^{(i)}$ has r vanishing moments and the set of functions $\{\psi^{(i)}_{j,k}(x) = 2^{dj/2}\psi^{(i)}(2^j x - k), \quad j \in \mathbb{N},\ k \in \{0,\dots,2^j-1\}^d, i \in \{1,\dots,2^d-1\}\}$ forms an orthonormal basis of $L^2([0,1]^d)$. It is also noticed in [21] that wavelets provide unconditional bases of $L^p([0,1]^d)$ as far as $1 < p < \infty$.

Thus any function $f \in L^p([0,1]^d)$ can be written as

$$f(x) = \sum_{i,j,k} c^{(i)}_{j,k}\psi^{(i)}_{j,k}(x)$$

where

$$c^{(i)}_{j,k} = 2^{jd/2}\int f(x)\psi^{(i)}(2^j x - k)\mathrm{d}x.$$

In the following we stand in isotropic cases. Thus the direction of the wavelets is not involved and for the sake of simplicity we omit the directional index i.

As the collection of $\{2^{dj/2}\psi(2^j x - k), \quad j \in \mathbb{N},\ k \in \{0,\dots,2^j-1\}^d\}$ form an orthonormal basis of $L^2([0,1]^d)$, observing the whole trajectory of Y_t in (2.1) is equivalent to treat the following problem, in which is observed $(y_{j,k})_{j\in\mathbb{N},k\in\{0,\dots,2^j-1\}^d} \in \ell^2(\mathbb{N}^{d+1})$ such that $\forall j,k$,

$$y_{j,k} = \theta_{j,k} + \frac{1}{\sqrt{n}}v_{j,k}, \tag{2.2}$$

where $y_{j,k} = \int \psi_{j,k}\mathrm{d}Y(x)$, v_i are i.i.d. Gaussian random variables and $(\theta_{j,k})_{j,k}$ is the sequence to be estimated.

Furthermore wavelets are useful as they provide a simple characterization of Besov spaces. Homogeneous Besov spaces are characterized, for $p,q>0$ and $s \in \mathbb{R}$, by:

$$f \in B^{s,q}_p([0,1]^d) \iff \exists c>0,\ \ \|f\|_{B^{s,q}_p} := \sum_{j\geq 0}\left(\sum_{k\in\{0,\dots,2^j-1\}^d}|c_{j,k}|^p 2^{(sp-d+\frac{pd}{2})j}\right)^{q/p} \leq c. \tag{2.3}$$

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

This characterization is independent from the chosen wavelet as soon as ψ has r vanishing moments, with $r \geq s$. We also denote by $B^{s,q}_{p,c}([0,1]^d)$, the closed ball in $B^{s,q}_p([0,1]^d)$ of radius $c > 0$.

In our framework, the minimax paradigm induces that one supposes that a function f belongs to $B^{s,q}_p([0,1]^d)$. Then one defines a risk or loss function thanks to a pseudo-distance on $B^{s,q}_p([0,1]^d)$, denoted $R(.,.)$. Given a radius $c > 0$ and an estimator $\widehat{f_n}$ of f which is a measurable function of the observations, the maximal risk of $\hat{f}_n$ on $B^{s,q}_{p,c}([0,1]^d)$ is defined by:

$$R^n(\hat{f}_n) = \sup_{f \in B^{s,q}_{p,c}([0,1]^d)} \mathbb{E}(R(\hat{f}_n, f)). \tag{2.4}$$

If $\mathcal{T}_n$ denotes the set of all measurable estimation procedures defined thanks to a given model, the minimax risk on $B^{s,q}_{p,c}([0,1]^d)$ is then given by:

$$R^n(B^{s,q}_{p,c}([0,1]^d)) = \inf_{\hat{f}_n \in \mathcal{T}_n} \sup_{f \in B^{s,q}_{p,c}([0,1]^d)} \mathbb{E}(R(\hat{f}_n, f)),$$

see for instance [27] for more details on the minimax theory.

This minimax risk gives an optimal bound over the function class $B^{s,q}_{p,c}([0,1]^d)$. It is thus natural for estimation procedures to attempt to reach this risk, at least asymptotically when n tends to infinity.

In terms of wavelets approximation, or in any base, a classical way to define estimators is given by linear estimators.

Definition 2.1. Suppose that we stand in the model (2.2). Linear estimators $\hat{f}^L_n$ are defined by

$$\hat{f}^L_n(x) = \sum_{j \geq 0} \sum_{k \in \{0,\dots,2^j-1\}^d} \hat{\theta}^{(n)}_{j,k} \psi_{j,k}(x), \tag{2.5}$$

where

$$\hat{\theta}^{(n)}_{j,k} = \lambda^{(n)}_{j,k} y_{j,k}.$$

Parameters $(\lambda^{(n)}_{j,k})_{j,k}$ can be seen as smoothing weights lying in $[0,1]$. Those weights can be of different natures. Classical ones are:

- Projection weights: $\lambda^{(n)}_{j,k} = \mathbb{1}_{\{j < T_n\}}$;
- Pinsker weights: $\lambda^{(n)}_{j,k} = (1 - (\frac{j}{T_n})^\alpha)_+$,

where $(T_n)_n$ is an increasing sequence depending on n.

The localization property of wavelet expansions is such that a given signal may have a sparse representation in those bases. Thus a natural estimation procedure in the white noise model, defined in [8] and widely used in the signal community is to take away small wavelet coefficients. This is the principle of wavelet thresholding.

Definition 2.2. The wavelet thresholding procedure is defined by

$$\hat{f}^T_n(x) = \sum_{j=0}^{j(n)} \sum_{k \in \{0,\dots,2^j-1\}^d} \theta^T_{j,k} \psi_{j,k}(x). \tag{2.6}$$

Here the weights are given by:

$$\theta^T_{j,k} = y_{j,k} \mathbb{1}_{\{|y_{j,k}| \geq \kappa t_n\}}, \tag{2.7}$$

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

in the case of hard thresholding, or

$$\theta_{j,k}^T = \text{sign}(y_{j,k})(|y_{j,k}| - \kappa t_n)_+, \tag{2.8}$$

for the soft thresholding. Furthermore,

$$t_n = \sqrt{\frac{\log n}{n}},$$

stands for the universal threshold and $j(n)$ is such that

$$2^{-j(n)} \leq \frac{\log n}{n} < 2^{-j(n)+1},$$

κ being a constant large enough.

Concerning the minimax theory, a large class of results exist in different function spaces and with different risk functions. Historically, the first one is the result of Pinsker [23] which shows that suitable linear estimators reach the optimal L^2 risk rate on L^2 Sobolev classes. If the risk function is given by an L^p norm, [3, 14] show that, under certain conditions, kernel estimators are optimal in the sense of minimax theory in the same function spaces. More recent results, such as those of [22], stated that linear estimators cannot reach the optimal bound in nonlinear regression, as soon as we take the L^p risk and Sobolev classes.

In this paper we focus on Besov spaces and take the general L^p norm as a risk function. The interest of studying Besov spaces is motivated by its practical use in approximation theory and its theoretical simplicity in terms of wavelet expansions. Furthermore, in the theoretical point of view, they also generalize some classical function spaces, such as Hölder and L^2 Sobolev spaces.

3. Statement of the main result

Let us recall some classical minimax results in Besov spaces. Taking the L^p norm, where $1 \leq p < \infty$, as loss function, we know from [10] that the minimax lower bound in closed balls in Besov spaces is given by the following proposition.

Proposition 3.1. *Let $1 \leq r \leq \infty$, $1 \leq p < \infty$ and $s > \frac{d}{r}$. Then, there exists $C > 0$ such that*

$$R^n(B_{r,c}^{s,\infty}) = \inf_{\widehat{f}_n} \sup_{f \in B_{r,c}^{s,\infty}} \mathbb{E}\left\|\widehat{f}_n - f\right\|_{L^p}^p \geq C r_n(s,r,p)$$

where

$$r_n(s,r,p) = \begin{cases} n^{-\frac{ps}{2s+d}} & \text{if } r > \frac{dp}{2s+d}, \\ \left(\frac{n}{\log n}\right)^{-\frac{p(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d}} & \text{else.} \end{cases}$$

Although it is proved in [10] that thresholding procedures reach asymptotically the optimal rate up to a logarithmic correction, it is not always the case for linear procedures. As it can be seen in [9], with L^2 risk, linear estimators do not attain the minimax rate when studied functions have a sparse representation in a given base. This result is generalized by the following proposition from [9] which gives the optimal rate that can be reached in this case.

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

Proposition 3.2. *Let $1 \leq r \leq \infty$, $1 \leq p < \infty$ and $s > \frac{d}{r}$. There exist $C > 0$ such that*

$$R^n_{lin}(B^{s,\infty}_{r,c}) = \inf_{\widehat{f}_n \, linear} \sup_{f \in B^{s,\infty}_{r,c}} \mathbb{E}\|\widehat{f}_n - f\|^p_{L^p} \geq C\tilde{r}_n(s,r,p)$$

where

$$\tilde{r}_n(s,r,p) = \begin{cases} n^{-\frac{ps}{2s+d}} & \text{if } r > p \\ \left(\frac{n}{\log n}\right)^{-\frac{ps'}{2s'+d}} & \text{else,} \end{cases}$$

and $s' = s - \frac{d}{r} + \frac{d}{p}$.

We see in the following theorem that results of Proposition 3.2 remain valid if we replace the risk maximum by the risk reached on almost every function. We also prove that in the same context thresholding algorithms attain the minimax risk given in Proposition 3.1 up to a logarithmic term.

Theorem 3.3. *Let $1 \leq r \leq \infty$, $1 \leq p < \infty$ and $s > \frac{d}{r}$. Then, in the context of (2.1):*

- *for almost every function f in $B^{s,\infty}_r([0,1]^d)$,*

$$\inf_{\hat{f}^L_n \, linear} \mathbb{E}\|\hat{f}^L_n - f\|^p_{L^p} \approx n^{-\alpha p}, \tag{3.1}$$

where

$$\alpha = \begin{cases} \frac{s}{2s+d} & \text{if } r \geq p, \\ \frac{s-\frac{d}{r}+\frac{d}{p}}{2(s-\frac{d}{r}+\frac{d}{p})+d} & \text{else;} \end{cases} \tag{3.2}$$

- *for almost every function in $B^{s,\infty}_r([0,1]^d)$, and for thresholding estimator $\hat{f}^T_n$*

$$\mathbb{E}\|\hat{f}^T_n - f\|^p_{L^p} \approx n^{-\alpha p} \tag{3.3}$$

where

$$\alpha = \begin{cases} \frac{s}{2s+d} & \text{if } r > \frac{pd}{2s+d} \\ \frac{s-\frac{d}{r}+\frac{d}{p}}{2s-\frac{d}{r}+d} & \text{else,} \end{cases} \tag{3.4}$$

where $a_n \approx b_n$ means that $\frac{\log a_n}{\log b_n} \to 1$.

As mentioned earlier, this generic result, such as Doob's theorem does not provide the exact behavior of a given function. However, it introduces a new vision of a generic behavior in Besov spaces and of the minimax theory as in these particular cases, generic and minimax results coincide. In this paper we have focused on the polynomial behavior of our estimators respectively to n. Concerning the thresholding procedure, we will prove that this result is true up to a logarithm term.

As we will see in the following section, the proofs of these results are quite simple. They are mainly based on the maxisets theory. Once known the maxiset associated to an estimation procedure, one study the genericity of such a set in the involved function space. The advantage is that our theorem can be easily extended to another kind of estimation procedures, thanks for instance to the results of [1, 25] or to other function spaces, such as Sobolev spaces.

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

4. Proof of Theorem 3.3

For the sake of completeness, let us recall some basic facts upon the maxisets theory.

4.1. Maxiset theory

The maxiset theory introduced recently in [5, 18, 19] is an alternative way to compare different estimation procedures. In our case, it provides a crucial key to prove Theorem 3.3. The main idea is to look for the maximal space on which an estimator reach a given rate instead of searching an optimal rate for a given space.

Definition 4.1. Let ρ be a risk function and $(v_n)_{n\in\mathbb{N}}$ a sequence such that $v_n \to 0$. For $\hat{f}_n$ an estimator, the maximal space associated to ρ, v_n and a constant T is given by

$$MS(\hat{f}_n, \rho, v_n, T) = \left\{ f;\ \sup_n v_n^{-1} \mathbb{E}(\rho(\hat{f}_n, f)) < T \right\}.$$

Several improvements were made in nonparametric theory thanks to this theory. For instance, it is shown in [5] that, for the density estimation model the thresholding procedure is more efficient than the linear procedure, whose maxiset is given in [17]. And in the heteroscedastic white noise model, [24, 25] shown that thresholding procedures are better than linear estimators and as good as Bayesian procedures. In the case of white noise model, we recall the following result which is a particular case of [24].

Proposition 4.2. *Let $1 \le p < \infty$ and $0 < s < \infty$ be two reals numbers. Let $\hat{f}_n^L$ be the linear estimator given in Definition 2.1. For any $T_n > 0$ suppose we are given $(\lambda_{j,k}^{(n)}(T_n))_{j,k}$ weights in $[0,1]$ such that:*

- *there exists $c < 1$ such that for all $T_n > 0$ and $j \ge T_n$, $\lambda_{j,k}^{(n)}(T_n) \le c$;*
- *there exists $c_s \in \mathbb{R}$ such that for any $T_n > 1$*

$$\sum_{1\le j\le T_n} \sum_{k\in\{0,\dots,2^j-1\}^d} \left(\lambda_{j-1,k}^{(n)}(T_n) - \lambda_{j,k}^{(n)}(T_n)\right) \left(1 - \lambda_{j,k}^{(n)}(T_n)\right)^{p-1} \left(\frac{j}{T_n}\right)^{-ps} \le c_s;$$

- *$T_n \to \infty$ as $n \to \infty$.*

We suppose that there exists a positive constant T such that for any $n \in \mathbb{N}$,

$$\frac{T_n^{ps}}{n^{p/2}} \sum_{j,k} \int |\psi_{j,k}|^p \le T. \tag{4.1}$$

Then for every f, there exists a positive constant C such that for any $n \in \mathbb{N}$,

$$\mathbb{E}\|\hat{f}_n^L - f\|_p^p \le C T_n^{-sp}$$

if and only if $f \in B_p^{s,\infty}([0,1]^d)$.

Before stating the corresponding result for thresholding algorithms, we define new function spaces closely related to approximation theory. Those spaces, weak Besov spaces, defined in [5] are subsets of Lorentz spaces, and constitute a larger class of functions than Besov spaces.

Definition 4.3. Let $0 < r < p < \infty$. We say that a function $f = \sum_{j,k} c_{j,k}\psi_{j,k}$ belongs to $W(r,p)$ if and only if

$$\sup_{\lambda>0} \lambda^r \sum 2^{j\left(\frac{dp}{2}-d\right)} \sum_k \mathbb{1}_{\{|c_{j,k}|>\lambda\}} < \infty. \tag{4.2}$$

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

A fast calculation shows that the space $W(r,p)$ contains the homogeneous Besov spaces $B_r^{\beta,\infty}$ as soon as $\beta \geq \frac{d}{2}(\frac{p}{r}-1)$.

The maxiset associated with the thresholding estimation procedure is given by a weak Besov space as proved in [5], and developed further in the heteroscedastic regression case in [18].

Proposition 4.4. *Let $1 \leq p < \infty$ and $\tilde{\alpha} \in (0,1)$ be two reals numbers. Let $\hat{f}_n^T$ be the estimator given in Definition 2.2. Then for every f we have the following equivalence:*

$\exists K > 0$ such that $\forall n > 0$,

$$\mathbb{E}\|\hat{f}_n^T - f\|_p^p \leq K\left(\sqrt{n\log(n)^{-1}}\right)^{-\tilde{\alpha}p} \tag{4.3}$$

if and only if $f \in B_p^{\tilde{\alpha}/2,\infty} \cap W((1-\tilde{\alpha})p, p)$.

Furthermore, another important key result involving Besov spaces is the following proposition from [11].

Proposition 4.5. *Let us define the scaling function of a distribution f by*

$$\forall p > 0 \quad s_f(p) = \sup\{s :\ f \in B_p^{s,\infty}\}. \tag{4.4}$$

Let s_0 and p_0 be fixed such that $s_0 - \frac{d}{p_0} > 0$. Outside a Haar-null set in $B_{p_0}^{s_0,\infty}([0,1]^d)$, we have:

$$s_f(p) = \begin{cases} s_0 & \text{if } p \leq p_0 \\ \frac{d}{p} + s_0 - \frac{d}{p_0} & \text{if } p \geq p_0. \end{cases} \tag{4.5}$$

One can check that a lower bound of this scaling function is given by Besov embeddings and interpolation theory, which can be found in [26]. This result states that one cannot have a better regularity than the one given by those embeddings. In our case, we will exploit this result by comparing those critical spaces with the maxiset associated to each procedure.

In the following parts we prove Theorem 3.3. For this purpose we consider the maxisets associated to better rate of convergence than the minimax one. We will prove that, both for linear and thresholding procedures, these maxisets can be written as a countable union of Haar-null sets.

4.2. Generic risk for linear estimators

Let $1 \leq p < \infty$, $1 \leq r < \infty$ and $s > \frac{d}{r}$ be fixed. Denote

$$s' = s - \left(\frac{d}{r} - \frac{d}{p}\right)_+,$$

and

$$\alpha(s') = \frac{s'}{2s'+d}.$$

In this section, we prove the first part of Theorem 3.3. We define the linear estimator as in Definition 2.1. For the sake of simplicity, we assume here that we take projections weights in the definition but the proof remains valid for other types of weights, as soon as assumptions in Proposition 4.2 are satisfied. Let $\theta > 0$ be given. As we are looking for the polynomial behavior of linear estimators, we take $T_n = n^\theta$. Define $j_1(n)$ be such that $2^{j_1(n)} \leq T_n < 2^{j_1(n)+1}$. From [8] we know that there exists $C > 0$ such that for any $n \in \mathbb{N}$, and for any $f \in B_r^{s,\infty}([0,1]^d)$,

$$\mathbb{E}\|\hat{f}_n^L - f\|_p^p \leq C\left(2^{-j_1(n)sp} + \left(\frac{2^{j_1(n)}}{n}\right)^{p/2}\right). \tag{4.6}$$

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

Furthermore, the infimum in (4.6) is obtained when the two terms are balanced, that is for $\theta_0 = \frac{1}{2s'+d}$. And we obtain in this case

$$\inf_{\hat{f}_n^L \ linear} \mathbb{E}\|\hat{f}_n^L - f\|_p^p \leq Cn^{-\alpha(s')p}. \tag{4.7}$$

Let us now check the lower bound. Let $T_n = n^\theta$, with $\theta > 0$ be given and let $\hat{f}_n^{L,\theta}$ be the corresponding estimator. In a first time, we have to show that for every $\theta > 0$ and $\varepsilon > 0$ fixed, the set

$$\tilde{M}(\theta,\varepsilon) := \left\{ f \in B_r^{s,\infty}([0,1]^d);\ \exists C > 0\ \forall n \in \mathbb{N},\ \mathbb{E}(\|\hat{f}_n^{L,\theta} - f\|_{L^p}^p) < Cn^{-(\alpha(s')+\varepsilon)p} \right\}$$

is a Borel Haar null set.

Taking into account that $\int |\psi_{j,k}|^p \sim 2^{jd(\frac{p}{2}-1)}$ we see that equation (4.1) is satisfied for $\theta \leq \frac{1}{2s'+d}$. We have thus two cases:

- if $\theta < \frac{1}{2s'+d}$ then from Proposition 4.2, $\tilde{M}(\theta,\varepsilon)$ is included in $B_p^{s'+\varepsilon,\infty}([0,1]^d)$. And from Proposition 4.5, we know that this set is a Haar null Borel set of $B_r^{s,\infty}([0,1]^d)$;
- if $\theta > \frac{1}{2s'+d}$ then

$$\begin{aligned}
\mathbb{E}(\|\hat{f}_n^L - f\|_p^p) &= \mathbb{E}\left(\int \left(\sum_{j\leq j_1(n)} \sum_k \frac{\varepsilon_{j,k}}{\sqrt{n}} \psi_{j,k}(x) + \sum_{j>j_1(n)} \sum_k c_{j,k}\psi_{j,k} \right)^p \right) \\
&\geq C\mathbb{E}\left(\sum_{j\leq j_1(n)} \sum_k (\frac{\varepsilon_{j,k}}{\sqrt{n}})^p \int \psi_{j,k}^p + \sum_{j>j_1(n)} \sum_k |c_{j,k}|^p \int \psi_{j,k}^p \right) \\
&\geq C \sum_{j\leq j_1(n)} \sum_k \mathbb{E}\left(\left(\frac{\varepsilon_{j,k}}{\sqrt{n}} \right)^p \right) 2^{dj(\frac{p}{2}-1)} \\
&\geq C \sum_{j\leq j_1(n)} \sum_k \left(\mathbb{E}\left(\left(\frac{\varepsilon_{j,k}}{\sqrt{n}} \right)^2 \right) \right)^{p/2} 2^{dj(\frac{p}{2}-1)} \\
&\geq Cn^{dp\theta/2-p/2} > n^{-\alpha(s')}.
\end{aligned}$$

Where the first inequality comes from the "superconcentration property" of wavelets, see Theorem 4.2 of [18], whereas the last one is the Hölder inequality [18]. Thus for $\theta > \frac{1}{2s'+d}$ and for any $\varepsilon > 0$, the set $\tilde{M}(\theta,\varepsilon)$ is an empty set.

We thus obtain that $\forall\theta > 0$ and $\forall\varepsilon > 0$, the set

$$\left\{ f \in B_r^{s,\infty}([0,1]^d);\ \exists C > 0\ \forall n \in \mathbb{N},\ \mathbb{E}(\|\hat{f}_n^{L,\theta} - f\|_{L^p}^p) < Cn^{-(\alpha(s')+\varepsilon)p} \right\}$$

is a Haar null set.

This set can also be written,

$$\left\{ f \in B_r^{s,\infty}([0,1]^d);\ \liminf_{n\to\infty} \frac{\log(\mathbb{E}(\|\hat{f}_n^{L,\theta} - f\|_{L^p}^p))}{-p\log n} > \alpha(s') + \varepsilon \right\}.$$

Taking the countable union of those sets over a dense sequence θ_n and a decreasing sequence $\varepsilon_n \to 0$, and the complementary we obtain that for almost every function in $B_r^{s,\infty}([0,1]^d)$,

$$\liminf_{n\to\infty} \frac{\log(\mathbb{E}(\|\hat{f}_n^L - f\|_{L^p}^p))}{-p\log n} \leq \alpha(s').$$

Which induces the expected result.

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

4.3. Thresholding algorithms

In this part, we take the estimation procedure given in Definition 2.2.

Let us turn out our attention to the minimax rate of convergence for this estimator. For this purpose, we write in the following

$$\tilde{\alpha}(s) = \begin{cases} \frac{2s}{2s+d} & \text{if } r > \frac{pd}{2s+d} \\ \frac{2(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d} & \text{else.} \end{cases} \tag{4.8}$$

The proof of the second point of Theorem 3.3 follows the same scheme as the previous one. In this case, the upper bound is given in [10]. Thus we know that for every function in $B_r^{s,\infty}([0,1]^d)$, and for all $1 < p < \infty$, there exists $C > 0$ such that

$$\mathbb{E}(\|\hat{f}_n^T - f\|_{L^p}^p) < C\left(\frac{n}{\log n}\right)^{-\tilde{\alpha}(s)p/2}.$$

In order to prove the lower bound, we use Proposition 4.4.

For every values of $\tilde{\alpha}$, let $0 < \varepsilon < 1 - \tilde{\alpha}$ be fixed, and $M(\varepsilon)$ be the set defined by

$$M(\varepsilon) = \left\{ f \in B_r^{s,\infty}([0,1]^d);\ \exists C > 0\ \forall n \in \mathbb{N},\ \mathbb{E}(\|\hat{f}_n^T - f\|_{L^p}^p) < C\left(\frac{n}{\log n}\right)^{-(\tilde{\alpha}(s)+\varepsilon)p/2} \right\}.$$

Thanks to Proposition 4.4, this set $M(\varepsilon)$ is embedded in $B_p^{\frac{\tilde{\alpha}+\varepsilon}{2},\infty} \cap W((1-\tilde{\alpha}-\varepsilon)p, p)$.

The end of the proof is based on the following proposition.

Proposition 4.6. *Let f be a given distribution. Let us define the weak scaling function of a distribution f by*

$$\forall p > 0 \quad \tilde{s}_f(p) = \sup\{\alpha :\ f \in W((1-\alpha)p, p)\}. \tag{4.9}$$

Let s and r be fixed such that $s - \frac{d}{r} > 0$. Outside a Haar-null set in $B_r^{s,\infty}([0,1]^d)$, we have:

$$\tilde{s}_f(p) = \begin{cases} \frac{2s}{2s+d} & \text{if } r > \frac{pd}{2s+d} \\ \frac{2(s-\frac{d}{r}+\frac{d}{p})}{2(s-\frac{d}{r})+d} & \text{else.} \end{cases} \tag{4.10}$$

The proof of this proposition is based on the "saturating" function defined in [15], which has the worst possible regularity in strong Besov spaces. More precisely we prove in the following that this function also "saturates" the corresponding weak Besov spaces.

Proof. In order to prove Proposition 4.6, let us prove that $W((1-\tilde{\alpha}-\varepsilon)p, p)$ is a Haar null Borel set in $B_r^{s,\infty}([0,1]^d)$. For this purpose, we define our transverse measure as the one-dimensional probe generated by the function g defined by its wavelet coefficients:

$$d_{j,k} = \frac{2^{-(s-\frac{d}{r}+\frac{d}{2})j} 2^{-\frac{d}{r}J}}{j^a}$$

where $a = 1 + \frac{3}{r}$ and $0 \le J \le j$ and $K \in \{0, \dots, 2^J - 1\}^d$ are such that

$$\frac{K}{2^J} = \frac{k}{2^j}$$

is an irreducible fraction. As it can be seen in Proposition 2 of [16], this function g belongs to $B_r^{s,\infty}([0,1]^d)$. Let $f \in B_r^{s,\infty}([0,1]^d)$ be an arbitrary function and consider the affine subset

$$M = \{\alpha \in \mathbb{R}\ \ f + \alpha g \in W((1-\tilde{\alpha}-\varepsilon)p, p)\}.$$

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

Suppose that there exist two points α_1 and α_2 in M. Thus $f+\alpha_1 g-(f+\alpha_2 g)$ belongs to $W((1-\tilde{\alpha}-\varepsilon)p,p)$, and there exists $c>0$ such that

$$\|f+\alpha_1 g-(f+\alpha_2 g)\|_{W((1-\tilde{\alpha}-\varepsilon)p,p)}=\|(\alpha_1-\alpha_2)g\|_{W((1-\tilde{\alpha}-\varepsilon)p,p)}\leq c. \tag{4.11}$$

A fast calculation shows that

$$\forall \alpha>0,\quad \|\alpha g\|_{W(r,p)}=\alpha^r\|g\|_{W(r,p)}. \tag{4.12}$$

We thus just have to determine $\|g\|_{W(r,p)}$. Thanks to equation (4.2), this is equivalent to determine, for every $t>0$, the value of

$$2^{-(1-\tilde{\alpha}-\varepsilon)pt}\sum_{j\geq 0}2^{j(\frac{dp}{2}-d)}\sum_k \mathbb{1}_{\{d_{j,k}>2^{-t}\}}.$$

But by definition of g, we have,

$$\frac{2^{-(s-\frac{d}{r}+\frac{d}{2})j}2^{-\frac{d}{r}J}}{j^a}>2^{-t}\Rightarrow\left(s-\frac{d}{r}+\frac{d}{2}\right)j+\frac{d}{r}J\leq t,$$

which implies that

$$J\leq\frac{r}{d}t-\left(s-\frac{d}{r}+\frac{d}{2}\right)\frac{r}{d}j.$$

Note that the condition $J\geq 0$ implies also that j is limited by

$$j\left(s-\frac{d}{r}+\frac{d}{2}\right)\leq t.$$

We denote by $\tilde{t}=\frac{t}{s-\frac{d}{r}+\frac{d}{2}}$ and by $\tilde{\tilde{t}}=\frac{t}{s+\frac{d}{2}}$. Thus we have, for every $t>0$,

$$\begin{aligned}
\|g\|_{W((1-\tilde{\alpha}-\varepsilon)p,p)} &\geq 2^{-(1-\tilde{\alpha}-\varepsilon)pt}\sup_{0\leq j\leq\tilde{t}}2^{j(\frac{dp}{2}-d)}\sum_{J=0}^{j\wedge[\frac{r}{d}t-(s-\frac{d}{r}+\frac{d}{2})\frac{r}{d}j]}2^{dJ}\\
&\geq 2^{-(1-\tilde{\alpha}-\varepsilon)pt}\sup\left(\sup_{0\leq j\leq\frac{t}{s+\frac{d}{2}}}2^{j(\frac{dp}{2}-d)}\sum_{J=0}^{j}2^{dJ},\sup_{\frac{t}{s+\frac{d}{2}}+1\leq j\leq\tilde{t}}2^{j(\frac{dp}{2}-d)}\sum_{J=0}^{[\frac{r}{d}t-(s-\frac{d}{r}+\frac{d}{2})\frac{r}{d}j]}2^{dJ}\right)\\
&\geq\frac{2^{-(1-\tilde{\alpha}-\varepsilon)pt}}{2^d-1}\sup\left(\sup_{0\leq j\leq\tilde{\tilde{t}}}2^{\frac{dpj}{2}}(1-2^{-jd}),\sup_{\tilde{\tilde{t}}<j\leq\tilde{t}}2^{j(\frac{dp}{2}-d)}(2^{rt}2^{-jr(s+\frac{d}{2}-\frac{d}{r})}-1)\right).
\end{aligned}$$

Merging this result with (4.11) together with (4.12), we obtain that, if there exist α_1 and α_2 in M then they satisfy that for every $t\geq 0$ and $0\leq j\leq\tilde{t}$,

$$|\alpha_1-\alpha_2|^{(1-\tilde{\alpha}-\varepsilon)p}\leq\inf\left(\frac{c2^{(1-\tilde{\alpha}-\varepsilon)pt}}{\sup_{0\leq j\leq\tilde{\tilde{t}}}2^{\frac{dpj}{2}}|1-2^{-jd}|},\frac{c2^{(1-\tilde{\alpha}-\varepsilon)pt}}{\sup_{\tilde{\tilde{t}}<j\leq\tilde{t}}2^{j(\frac{dp}{2}-d)}|2^{rt}2^{-jr(s+\frac{d}{2}-\frac{d}{r})}-1|}\right). \tag{4.13}$$

We have thus two cases:

- If $r>\frac{dp}{2s+d}$

$$\tilde{\alpha}=\frac{2s}{2s+d}.$$

But, if we take the first term,

$$\sup_{0\leq j\leq\tilde{\tilde{t}}}2^{\frac{dpj}{2}}|1-2^{-jd}|\sim 2^{\frac{tdp}{2s+d}},$$

we have

$$|\alpha_1-\alpha_2|^{(1-\tilde{\alpha}-\varepsilon)p}\leq c2^{-\varepsilon pt}; \tag{4.14}$$

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

- when $r \leq \frac{dp}{2s+d}$, and as $s > \frac{d}{r}$ we have necessarily $p > 2$ and we obtain

$$\tilde{\alpha} = \frac{2(s - \frac{d}{r} + \frac{d}{p})}{2(s - \frac{d}{r}) + d}.$$

In this case,

$$\sup_{\tilde{t} < j \leq \tilde{t}} 2^{j(\frac{dp}{2} - d)} |2^{rt} 2^{-jr(s + \frac{d}{2} - \frac{d}{r})} - 1| \sim 2^{\frac{td(p-2)}{2(s - \frac{d}{r}) + d}}.$$

And once again,

$$\forall t > 0 \quad |\alpha_1 - \alpha_2|^{(1 - \tilde{\alpha} - \varepsilon)p} \leq c 2^{-\varepsilon p t}. \tag{4.15}$$

As $1 - \tilde{\alpha} - \varepsilon > 0$, it can be deduced from equations (4.14) and (4.15) that for t large enough, M is of vanishing Lebesgue measure and $W((1 - \tilde{\alpha} - \varepsilon)p, p)$ is an Haar null set in $B_r^{s,\infty}([0,1]^d)$. □

Thanks to invariance under inclusion, we have obtained that for every $\varepsilon > 0$, the set of functions f in $B_r^{s,\infty}([0,1]^d)$ such that

$$\exists C > 0 \; \forall n \in \mathbb{N}, \; \mathbb{E}(\|\hat{f}_n^T - f\|_{L^p}^p) < C \sqrt{\frac{n}{\log n}}^{-(\alpha(s) + \varepsilon)p}$$

is a Haar null set.

Taking a countable union over a decreasing sequence $\varepsilon_n \to 0$, we obtain that for almost every function in $B_r^{s,\infty}([0,1]^d)$, we have

$$\liminf_{n \to \infty} \frac{\log(\mathbb{E}(\|\hat{f}_n^T - f\|_{L^p}^p))}{-p \log n} \leq \tilde{\alpha}(s).$$

References

[1] F. Autin, *Point de vue maxiset en estimation non paramétrique*. Ph.D. thesis, Université Paris 7 (2004).

[2] Y. Benyamini and J. Lindenstrauss, *Geometric nonlinear functional analysis, Colloquium Publications,* vol. 1. American Mathematical Society (AMS) (2000).

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

[4] J.P.R. Christensen, On sets of Haar measure zero in Abelian Polish groups. *Isr. J. Math.* **13** (1972) 255–260.

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

[6] I. Daubechies, Orthonormal bases of compactly supported wavelets. *Comm. Pure Appl. Math.* **41** (1988) 909–996.

[7] P. Dodos, Dichotomies of the set of test measures of a Haar-null set. *Isr. J. Math.* **144** (2004) 15–28.

[8] D. Donoho and I. Johnstone, Minimax risk over l_p-balls for l_q-error. *Probab. Theory Relat. Fields* **99** (1994) 277–303.

[9] D. Donoho and I. Johnstone, Minimax estimation *via* wavelet shrinkage. *Ann. Stat.* **26** (1998) 879–921.

[10] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian and D. Picard, *Universal near minimaxity of wavelet shrinkage*. Festschrift for Lucien Le Cam, Springer, New York (1997) 183–218.

[11] A. Fraysse, Generic validity of the multifractal formalism. *SIAM J. Math. Anal.* **37** (2007) 593–607.

[12] B. Hunt, The prevalence of continuous nowhere differentiable function. *Proc. Am. Math. Soc.* **122** (1994) 711–717.

[13] B. Hunt, T. Sauer and J. Yorke, Prevalence: a translation invariant "almost every" on infinite dimensional spaces. *Bull. Am. Math. Soc.* **27** (1992) 217–238.

[14] I.A. Ibragimov and R.Z. Hasminski, *Statistical estimation, Applications of Mathematics,* vol. 16. Springer-Verlag (1981).

[15] S. Jaffard, Old friends revisited: The multifractal nature of some classical functions. *J. Fourier Anal. Appl.* **3** (1997) 1–22.

[16] S. Jaffard, On the Frisch-Parisi conjecture. *J. Math. Pures Appl.* **79** (2000) 525–552.

[17] G. Kerkyacharian and D. Picard, Density estimation by kernel and wavelets methods: optimality of Besov spaces. *Stat. Probab. Lett.* **18** (1993) 327–336.

[18] G. Kerkyacharian and D. Picard, Thresholding algorithms, maxisets and well-concentrated bases. *Test* **9** (2000) 283–344, With comments, and a rejoinder by the authors.

[19] G. Kerkyacharian and D. Picard, Minimax or maxisets? *Bernoulli* **8** (2002) 219–253.

[20] S. Mallat, *A wavelet tour of signal processing*. Academic Press, San Diego, CA (1998) xxiv.

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press

[21] Y. Meyer, *Ondelettes et opérateurs*. Hermann (1990).
[22] A.S. Nemirovskiĭ, B.T. Polyak and A.B. Tsybakov, The rate of convergence of nonparametric estimates of maximum likelihood type. *Problemy Peredachi Informatsii* **21** (1985) 17–33.
[23] M.S. Pinsker, Optimal filtration of square-integrable signals in Gaussian noise. *Probl. Infor. Transm.* **16** (1980) 52–68.
[24] V. Rivoirard, Maxisets for linear procedures, *Stat. Probab. Lett.* **67** (2004) 267–275.
[25] V. Rivoirard, Nonlinear estimation over weak Besov spaces and minimax Bayes method, *Bernoulli* **12** (2006) 609–632.
[26] E. Stein, *Singular integrals and differentiability properties of functions*. Princeton University Press (1970).
[27] A. Tsybakov, *Introduction to nonparametric estimation*. Springer Series in Statistics, Springer, New York (2009).
[28] A. Van der Vaart, *Asymptotic statistics*, *Cambridge Series in Statistical and Probabilistic Mathematics,* vol. 3. Cambridge University Press (1998).

https://doi.org/10.1051/ps/2012002 Published online by Cambridge University Press