Statistica Sinica **19** (2009), 1407-1417

ADAPTING TO UNKNOWN SMOOTHNESS BY AGGREGATION OF THRESHOLDED WAVELET ESTIMATORS

Christophe Chesneau and Guillaume Lecué

LMNO, Caen, and University Pierre et Marie Curie, Paris VI

Abstract: We consider a multi-wavelet thresholding method for nonparametric estimation. An adaptive procedure based on a convex combination of weighted term-by-term thresholded wavelet estimators is proposed. By considering the density estimation framework, we prove that this procedure is optimal in the minimax sense over Besov balls under the L^2 risk, without an extra logarithm term.

Key words and phrases: Aggregation, density estimation, margin, oracle inequalities, threshold estimators, wavelets.

1. Introduction

Wavelet shrinkage methods have been very successful in nonparametric function estimation. They provide estimators that are spatially adaptive and (near) optimal over a wide range of function classes. Standard approaches are based on the term-by-term thresholds. The well-known examples are the hard and soft thresholded estimators introduced by Donoho and Johnstone (1995). The performances of such constructions are truly dependent of the choice of the threshold. In the literature, several techniques have been proposed to determine the 'best' adaptive threshold. There are, for instance, the RiskShrink and SureShrink methods (see Donoho and Johnstone (1995)), the cross-validation methods (see, for instance, Nason (1995) and Jansen (2001)), the methods based on hypothesis tests (see, for instance, Abramovich, Benjamini, Donoho and Johnstone (2006)), the Lepski methods (see Juditsky (1997)) and the Bayesian methods (see, for instance, Abramovich, Sapatinas and Silverman (1998)).

In the present paper, we propose to study the performances of a new adaptive wavelet estimator based on a convex combination of weighted local thresholding estimators (hard, soft, non negative garotte, ...). In the framework of nonparametric density estimation, we prove that, in some sense, this is at least as good as the term-by-term thresholded estimator defined with the 'best' threshold. In particular, we prove that the proposed estimator is optimal, in the minimax sense, over Besov balls under the L^2 risk. The proof is based on a non-adaptive

minimax result proved by Delyon and Juditsky (1996), and some powerful oracle inequalities satisfied by aggregation methods. Such methods use an exponential weighting aggregation scheme, that has been studied by, among others, Augustin, Buckland and Burnham (1997) Yang (2000), Catoni (2001), Leung and Barron (2006), Bunea and Nobel (2005), and Lecué (2005, 2006, 2007a,b).

The paper is organized as follows. Section 2 presents general oracle inequalities satisfied by the aggregation scheme using exponential weights. Section 3 describes the main procedure of the study and investigates its minimax performances over Besov balls under L^2 risk. Proofs are postponed to the last section.

2. Oracle Inequalities

2.1. Framework

Let $(\mathcal{Z}, \mathcal{T})$ be a measurable space. Denote by $\mathcal{P}$ the set of all probability measures on $(\mathcal{Z}, \mathcal{T})$. Let F be a function from $\mathcal{P}$ with values in an algebra $\mathcal{F}$. Let Z be a random variable with values in $\mathcal{Z}$ and denote by π its probability measure. Let D_n be a family of n i.i.d. observations $Z_1, \ldots, Z_n$ having the common probability measure π. The probability measure π is unknown. Our aim is to estimate $F(\pi)$ from the observations D_n.

In our estimation problem, we assume that we have access to an "empirical risk". This means that there exists $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$ such that the risk of an estimator $f \in \mathcal{F}$ of $F(\pi)$ is of the form $A(f) = \mathbb{E}\left[Q(Z, f)\right]$. If the infimum $A^* = \inf_{f \in \mathcal{F}} A(f)$ is achieved by at least one function, we denote by $f^* \in \mathcal{F}$ such a minimizer. In this paper we assume that $\inf_{f \in \mathcal{F}} A(f)$ is achievable, otherwise we replace f^* by f_n^*, an element in $\mathcal{F}$ satisfying $A(f_n^*) \leq \inf_{f \in \mathcal{F}} A(f) + n^{-1}$.

In most cases f^* will be $F(\pi)$. The risk A is unknown, instead of minimizing A over $\mathcal{F}$, we consider an empirical version of A constructed from the observations D_n. It is denoted by

$$A_n(f) = \frac{1}{n} \sum_{i=1}^{n} Q(Z_i, f). \tag{2.1}$$

In order to illustrate this general statistical framework with a concrete problem, let us focus our attention on nonparametric density estimation.

In the density estimation setup, $(\mathcal{Z}, \mathcal{T})$ is endowed with a finite measure μ and we assume that π is absolutely continuous w.r.t. to μ. One version of the density function of π w.r.t. μ is denoted by f^*. Consider $\mathcal{F}$ be the set of all density functions on $(\mathcal{Z}, \mathcal{T}, \mu)$. For any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$, the loss function considered is

$$Q(z, f) = \int_{\mathcal{Z}} |f(y)|^2 d\mu(y) - 2f(z). \tag{2.2}$$

We have, for any $f \in \mathcal{F}$,

$$A(f) = \mathbb{E}\left[Q(Z,f)\right] = \int_{\mathcal{Z}} |f(y)|^2 d\mu(y) - 2\int_{\mathcal{Z}} f(y)f^*(y)d\mu(y)$$
$$= \left\|f^* - f\right\|_2^2 - \int_{\mathcal{Z}} |f^*(y)|^2 d\mu(y).$$

Thus, the density function f^* is a minimizer of A over $\mathcal{F}$ and $A^* = -\int_{\mathcal{Z}} |f^*(y)|^2 d\mu(y)$.

Now, we introduce an assumption which improves the quality of estimation in our framework. This assumption was first introduced by Mammen and Tsybakov (1999) for the problem of discriminant analysis, and by Tsybakov (2004) for the classification problem. With it, parametric rates of convergence can be achieved, for instance, in the classification problem (cf. Tsybakov (2004) and Steinwart and Scovel (2007).

Margin Assumption (MA): *Let $\kappa \geq 1$, $c > 0$, and $\mathcal{F}_0$ be a subset of $\mathcal{F}$. We say that the probability measure π satisfies the margin assumption $MA(\kappa, c, \mathcal{F}_0)$ if, for any $f \in \mathcal{F}_0$, we have*

$$\mathbb{E}\left[|Q(Z,f) - Q(Z,f^*)|^2\right] \leq c(A(f) - A^*)^{1/\kappa}.$$

The margin assumption is linked to the convexity of the underlying loss. In density estimation with integrated squared risk, we can show that all probability measures π on $(\mathcal{Z}, \mathcal{T})$ absolutely continuous w.r.t. μ satisfy the margin assumption $\mathrm{MA}(1, 16B^2, \mathcal{F}_B)$ where $\mathcal{F}_B$ is the set of all non-negative functions $f \in L^2(\mathcal{Z}, \mathcal{T}, \mu)$ bounded by B. Other values for the margin parameter can be met in classification, for instance.

2.2. Aggregation procedures

We work with the notations introduced in the beginning of the previous subsection. The aggregation framework considered, among others, by Juditsky and Nemirovski (2000), Yang (2000), Nemirovski (2000), Tsybakov (2003), Leung and Barron (2006), and Birgé (2006), is the following. Take $\mathcal{F}_0$ a finite subset of $\mathcal{F}$, our aim is to mimic (up to an additive residual) the best function in $\mathcal{F}_0$ w.r.t. the risk A. For this, we consider the Aggregation with Exponential Weights aggregate (AEW) over $\mathcal{F}_0$. The resulting estimate is

$$\tilde{f}_n = \sum_{f \in \mathcal{F}_0} w^{(n)}(f) f, \tag{2.3}$$

where the exponential weights $w^{(n)}(f)$ are given by

$$w^{(n)}(f) = \frac{\exp\left(-nA_n(f)\right)}{\sum_{g \in \mathcal{F}_0} \exp\left(-nA_n(g)\right)}. \tag{2.4}$$

2.3. Oracle inequalities

In this subsection we state an exact oracle inequality satisfied by the AEW procedure in the general framework at the beginning of Section 2. From this exact oracle inequality, we deduce an oracle inequality in the density estimation framework. Define the quantity $\gamma = \gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q)$ by

$$\gamma = \begin{cases} \left(\frac{\mathcal{B}^{1/\kappa} \log M}{\beta_1 n}\right)^{1/2} & \text{if } \mathcal{B} \geq \left(\frac{\log M}{\beta_1 n}\right)^{\kappa/(2\kappa-1)}, \\ \left(\frac{\log M}{\beta_2 n}\right)^{\kappa/(2\kappa-1)} & \text{otherwise.} \end{cases} \tag{2.5}$$

Here $\mathcal{B} = \mathcal{B}(\mathcal{F}_0, \pi, Q) = \min_{f \in \mathcal{F}_0} (A(f) - A^*)$, $\kappa \geq 1$ is the margin parameter, π is the underlying probability measure, Q is the loss function,

$$\beta_1 = \min\left(\frac{\log 2}{96cK}, \frac{3\log 2}{16K\sqrt{2}}, \left(8(4c + \frac{K}{3})\right)^{-1}, (576c)^{-1}\right) \tag{2.6}$$

$$\beta_2 = \min\left(8^{-1}, \frac{3\log 2}{32K}, \left(2(16c + \frac{K}{3})\right)^{-1}, \frac{\beta_1}{2}\right), \tag{2.7}$$

where the constant $c > 0$ appears in the margin assumption $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$, and K surfaces below.

Theorem 2.1. *In the general framework introduced at the beginning of Section 2, $M \geq 2$ be an integer and $\mathcal{F}_0$ denote a finite subset of M elements $f_1, \ldots, f_M$ in $\mathcal{F}$. Assume that the underlying probability measure π satisfies the margin assumption $MA(\kappa, c, \mathcal{F}_0)$ for some $\kappa \geq 1, c > 0$. Assume that $f \longmapsto Q(z, f)$ is convex for π-almost $z \in \mathcal{Z}$ and, for any $f \in \mathcal{F}_0$, there exists a constant $K \geq 1$ such that $|Q(Z, f) - Q(Z, f^*)| \leq K$. Then, the AEW procedure $\tilde{f}_n$ defined by (2.3) satisfies*

$$\mathbb{E}\left[A(\tilde{f}_n) - A^*\right] \leq \min_{j=1,\ldots,M} \{A(f_j) - A^*\} + 4\gamma,$$

where $\gamma = \gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q)$ is defined by (2.5).

Corollary 2.2. *Assume an underlying density function f^* to estimate is bounded by $B > 0$. Let $M \geq 2$ be an integer. Let $f_1, \ldots, f_M$ be M functions such that $\|f_j\|_\infty \leq B, \forall j = 1, \ldots, M$. For β_2 defined in (2.7) and any $\epsilon > 0$, the AEW procedure $\tilde{f}_n$ defined by (2.3) satisfies*

$$\mathbb{E}\left[\left\|\tilde{f}_n - f^*\right\|_2^2\right] \leq (1 + \epsilon) \min_{j=1,\ldots,M} \left\{\left|f^* - f_j\right|_2^2\right\} + \frac{4\log M}{\epsilon \beta_2 n}. \tag{2.8}$$

Thus, the AEW procedure mimics the best f_j among the f_j's, up to a residual term which can be very small according to the value of M. A similar result can

be found in Yang (2000, 2001), where a randomized aggregate using exponential weights w.r.t. the Kullback-Leiber loss satisfies an oracle inequality like (2.8) with a 2 in front of the main term $\min_{j=1,\dots,M} \|f^* - f_j\|_2^2$.

3. Multi-thresholding Wavelet Estimator

In this section, we propose an adaptive estimator constructed from aggregation techniques and wavelet thresholding methods. For the density model, we show that it is optimal in the minimax sense over a wide range of function spaces.

3.1. Wavelets and Besov balls

We consider an orthonormal wavelet basis generated by dilation and translation of a compactly supported "father" wavelet ϕ and a compactly supported "mother" wavelet ψ. For our purposes, we use the periodized wavelets bases on the unit interval. Let $\phi_{j,k}(x) = 2^{j/2}\phi(2^j x - k), \psi_{j,k}(x) = 2^{j/2}\psi(2^j x - k)$ be the elements of the wavelet basis and $\phi_{j,k}^{per}(x) = \sum_{l\in\mathbb{Z}} \phi_{j,k}(x-l), \psi_{j,k}^{per}(x) = \sum_{l\in\mathbb{Z}} \psi_{j,k}(x-l)$, their periodized versions, defined for any $x \in [0,1]$, $j \in \mathbb{N}$ and $k \in \{0,\dots,2^j-1\}$. There exists an integer τ such that the collection ζ defined by $\zeta = \{\phi_{\tau,k}^{per}, k = 0,\dots,2^\tau - 1;\ \psi_{j,k}^{per},\quad j = \tau,\dots,\infty,\ k = 0,\dots,2^j-1\}$ constitutes an orthonormal basis of $L^2([0,1])$. In what follows, the superscript "*per*" will be suppressed from the notations for convenience. A square-integrable function f^* on $[0,1]$ can be expanded into a wavelet series

$$f^*(x) = \sum_{k=0}^{2^\tau-1} \alpha_{\tau,k}\phi_{\tau,k}(x) + \sum_{j=l}^{\infty}\sum_{k=0}^{2^j-1} \beta_{j,k}\psi_{j,k}(x),$$

where $\alpha_{j,k} = \int_0^1 f^*(x)\phi_{j,k}(x)dx$ and $\beta_{j,k} = \int_0^1 f^*(x)\psi_{j,k}(x)dx$. Further details on wavelet theory can be found in Meyer (1990) and Daubechies (1992).

Let $L \in (0,\infty)$, $s \in (0,\infty)$, $p \in [1,\infty)$ and $q \in [1,\infty)$, with $\beta_{\tau-1,k} = \alpha_{\tau,k}$. We say that a function f^* belongs to the Besov balls $B_{p,q}^s(L)$ if and only if there exists $L^* > 0$ such that the associated wavelet coefficients satisfy

$$\Big[\sum_{j=\tau-1}^{\infty}\Big[2^{j(s+1/2-1/p)}\Big(\sum_{k=0}^{2^j-1}|\beta_{j,k}|^p\Big)^{1/p}\Big]^q\Big]^{1/q} \le L^*, \qquad \text{if } q \in [1,\infty),$$

with the usual modification if $q = \infty$. We work with the Besov balls because of their exceptional expressive power; for a particular choice of parameters s, p and q, they contain the Hölder and Sobolev balls (see, for instance, Meyer (1990)).

3.2. Term-by-term thresholded estimator

In this subsection, we consider the estimation of an unknown density function f^* in $L^2([0,1])$.

A term-by-term thresholded wavelet estimator is given by

$$\hat{f}_\lambda(D_n,x)=\sum_{k=0}^{2^\tau-1}\hat{\alpha}_{\tau,k}\phi_{\tau,k}(x)+\sum_{j=\tau}^{j_1}\sum_{k=0}^{2^j-1}\Upsilon_{\lambda_j}(\hat{\beta}_{j,k})\psi_{j,k}(x), \tag{3.1}$$

where

$$\hat{\alpha}_{\tau,k}=\frac{1}{n}\sum_{i=1}^{n}\phi_{\tau,k}(X_i) \text{ and } \hat{\beta}_{j,k}=\frac{1}{n}\sum_{i=1}^{n}\psi_{j,k}(X_i), \tag{3.2}$$

j_1 is an integer satisfying $(n/\log n)\leq 2^{j_1}<2(n/\log n)$, $\lambda=(\lambda_\tau,\ldots,\lambda_{j_1})$ is a vector of positive integers and, for any $u>0$, the operator Υ_u is such that, for any $x,y\in\mathbb{R}$, there exist two constants $C_1,C_2>0$ satisfying

$$|\Upsilon_u(x)-y|^2\leq C_1\left(|\min(y,C_2u)|^2+|x-y|^2\mathbb{1}_{\{|x-y|\geq 2^{-1}u\}}\right). \tag{3.3}$$

The inequality (3.3) holds for the hard thresholding rule $\Upsilon_u^{hard}(x)=x\mathbb{1}_{\{|x|\geqslant u\}}$, the soft thresholding rule $\Upsilon_u^{soft}(x)=sign(x)(|x|-u)\mathbb{1}_{\{|x|\geqslant u\}}$ (see Donoho and Johnstone (1995), Donoho, Johnstone, Kerkyacharian and Picard (1995) and Delyon and Juditsky (1996)), and the non-negative garrote thresholding rule $\Upsilon_u^{NG}(x)=\left(x-u^2/x\right)\mathbb{1}_{\{|x|\geqslant u\}}$ (see Gao (1998)).

In Delyon and Juditsky (1996), it is proved that, for the threshold $\lambda=(\rho\sqrt{(j-j_s)_+/n})_{j=\tau,\ldots,j_1}$ where j_s is an integer such that $n^{1/(1+2s)}<2^{j_s}\leq 2n^{1/(1+2s)}$ and ρ satisfying

$$\rho^2\geq 4(\log 2)(8B+(\frac{8\rho}{3\sqrt{2}})(\|\psi\|_\infty+B)), \tag{3.4}$$

the term-by-term thresholded wavelet estimator $\hat{f}_\lambda(D_n,.)$ achieves the minimax rate of convergence $n^{-2s/(1+2s)}$ over $B^s_{p,q}(L)$. In this study, we use aggregation methods to construct an adaptive estimator at least as good, in the minimax sense, as this non-adaptive estimator.

3.3. Multi-thresholding estimator

Divide the observations D_n into two disjoint subsamples D_m, of size m, made of the first m observations and $D^{(l)}$, of size l, made of the remaining observations, where we take $l=\lceil n/\log n\rceil$ and $m=n-l$. The first subsample D_m, sometimes called "training sample", is used to construct a family of estimators (in our case, the thresholded estimators) and the second subsample $D^{(l)}$, called the "training sample", is used to construct the weights of the aggregation procedure.

Assume that we want to estimate a density function f^* from $[0,1]$ bounded by B. For any $y\in\mathbb{R}$, we consider the projection function

$$h_B(y)=\max(0,\min(y,B)). \tag{3.5}$$

For any $u > 0$, we consider the truncated estimator

$$\hat{f}^t_{m,u}(x) = h_B(\hat{f}_{v_u}(D_m, x)),$$

where $v_u = (\rho\sqrt{(j-u)_+/n})_{j=\tau,\dots,j_1}$ and ρ satisfies (3.4).

We define the *multi-thresholding estimator* $\tilde{f}_n : [0,1] \to [0,B]$ at a point $x \in [0,1]$ by the aggregate

$$\tilde{f}_n(x) = \sum_{u\in\Lambda_n} w^{(l)}(\hat{f}^t_{m,u})\hat{f}^t_{m,u}(x), \tag{3.6}$$

where $\Lambda_n = \{0,\dots,\lceil \log n\rceil\}$ and, for any $u \in \Lambda_n$,

$$w^{(l)}(\hat{f}^t_{m,u}) = \frac{\exp\left(-lA^{(l)}(\hat{f}^t_{m,u})\right)}{\sum_{\gamma\in\Lambda_n}\exp\left(-lA^{(l)}(\hat{f}^t_{m,\gamma})\right)}.$$

Here $A^{(l)}(f) = (1/l)\sum_{i=m+1}^{n} Q(Z_i, f)$ is the empirical risk constructed from the l last observations, for any function f and for the choice of a loss function Q defined at (2.2).

The multi-thresholding estimator $\tilde{f}_n$ realizes a kind of "adaptation to the threshold" by selecting the best threshold v_u for u describing the set Λ_n. Since we know that there exists an integer j_* in Λ_n, depending on the regularity of f^*, such that the non-adaptive estimator $\hat{f}_{v_{j_*}}(D_m, .)$ is minimax (see Delyon and Juditsky (1996)), the multi-thresholding estimator is minimax independently of the regularity of f^*. Moreover, the cardinality of Λ_n is only $\lceil \log n\rceil$, thus $\tilde{f}_n$ does not require the construction of too many estimators.

4. Performance of the Multi-thresholding Estimator

4.1. Main result

Theorem 4.1 investigates the minimax performance of the multi-thresholding estimator defined in (3.6) under the L^2 risk over Besov balls in the density estimation framework.

Theorem 4.1. *Suppose the density function f^* is bounded by $B > 0$. For any $p \in [1,\infty]$, $s \in (p^{-1},\infty)$, and $q \in [1,\infty]$, there exists a constant $C > 0$, depending only on s, p and q, such that the multithresholding estimator $\tilde{f}_n$ defined in* (3.6) *satisfies, for n large enough,*

$$\sup_{f^*\in B^s_{p,q}(L)} \mathbb{E}\left[\|\tilde{f}_n - f^*\|_2^2\right] \le Cn^{-2s/(2s+1)}.$$

Recall that, for the density model, the rate of convergence $n^{-2s/(1+2s)}$ is minimax over $B^s_{p,q}(L)$. Further details about the minimax rate of convergence

Table 4.1. Rates of convergence achieved by various wavelet thresholding estimators for the density model under the L^2 risk over Besov balls $B^s_{p,q}(L)$.

	Rates of convergence over $B^s_{p,q}(L)$	
	$2 > p \geq 1$	$p \geqslant 2$
Local thresh	$(\ln n/n)^{2s/(2s+1)}$	$(\ln n/n)^{2s/(2s+1)}$
Block thresh	$(\ln n/n)^{2s/(2s+1)}$	$n^{-2s/(2s+1)}$,
Multi thresh	$n^{-2s/(2s+1)}$	$n^{-2s/(2s+1)}$

over Besov balls under L^2 risk for the density model can be found in Delyon and Juditsky (1996), and Härdle, Kerkyacharian, Picard and Tsybakov (1998).

4.2. Minimax comparison with other estimators

If we focus our attention on the density model, there are several types of estimators that enjoy good minimax performances under the L^2 risk over Besov balls. We distinguish the local thresholding estimators and the block thresholding estimators. The local thresholding estimators include the soft thresholding and the hard thresholding proposed by Donoho, Johnstone, Kerkyacharian and Picard (1996); the block thresholding estimators include the BlockShrink method and the BlockJS method investigated by Cai and Chicken (2005).

As seen in Table 4.1, the rates of convergence achieved by the Multithresholding estimator is better than those achieved by the local and block thresholding estimators, we gain a logarithmic term.

Finally, Yang (2000) also took the approach of combining procedures to obtain adaptive density estimators over Besov classes. He used exponential weights with respect to the Kullback-Leiber loss (in this case, exponential weights are related to the likelihood of the model (cf., Lecué (2005))). The resulting aggregate achieves the minimax rate of convergence over all Besov Balls $B^s_{p,q}(L)$ for any $s \in (p^{-1}, \infty)$. Nevertheless, the estimators aggregated in Yang (2000) are constructed by using a metric entropy argument and are not easily compared to the wavelet estimators that we used here.

Remark 4.1. In the bounded regression framework with random uniform design, we can construct an aggregate with exponential weights of term-by-term thresholded wavelet estimator achieving the minimax rate of convergence $n^{-2s/(2s+1)}$ over all Besov balls $B^s_{p,q}(L)$ for any $p \in [1, \infty]$, $s \in (p^{-1}, \infty)$ and $q \in [1, \infty]$.

5. Proofs

Proof of Theorem 2.1. For a detailed proof of this theorem, we refer the reader to the supplement file available at the following `http://www.stat.sinica.edu.tw/statistica`.

Proof of Corollary 2.2. In density estimation with integrated squared risk, any absolutely continuous probability measure π on $(\mathcal{Z}, \mathcal{T})$ satisfies the margin assumption $\mathrm{MA}(1, 16B^2, \mathcal{F}_B)$, where $\mathcal{F}_B$ is the set of all non-negative function $f \in L^2(\mathcal{Z}, \mathcal{T}, \mu)$ bounded by B. To complete the proof we use, for any $\epsilon > 0$,

$$\left[\frac{\mathcal{B}(\mathcal{F}_0, \pi, Q) \log M}{\beta_1 n}\right]^{1/2} \leq \epsilon \mathcal{B}(\mathcal{F}_0, \pi, Q) + \frac{\log M}{\beta_2 n \epsilon}.$$

Proof of Theorem 4.1. We apply Theorem 2.2, with $\epsilon = 1$, to the multi-thresholding estimator $\hat{f}_n$ defined in (3.6). Since $Card(\Lambda_n) = \lceil \log n \rceil$, $m \geq n/2$ and the density function f^* to estimate takes its values in $[0, B]$, conditionally on the first subsample D_m, we have

$$\begin{aligned}
&\mathbb{E}\left[\|f^* - \hat{f}_n\|_2^2 \,|D_m\right] \\
&\leq 2 \min_{u \in \Lambda_n} \left(\left\| f^* - h_B(\hat{f}_{v_u}(D_m, .)) \right\|_2^2 \right) + \frac{4(\log n)\log(\log n)}{\beta_2 n} \\
&\leq 2 \min_{u \in \Lambda_n} \left(\left\| f^* - \hat{f}_{v_u}(D_m, .) \right\|_2^2 \right) + \frac{4(\log n)\log(\log n)}{\beta_2 n}, \qquad (5.1)
\end{aligned}$$

where h_B is the projection function introduced in (3.5), and β_2 is given in (2.7). Now, for any $s > 0$, consider j_s an integer in Λ_n such that $n^{1/(1+2s)} \leq 2^{j_s} < 2n^{1/(1+2s)}$. A result proved by Delyon and Juditsky (1996), says that the local thresholding estimator defined with threshold $v_{j_s} = \rho\sqrt{(j - j_s)_+/n}$ satisfies

$$\sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}\left[\left\| f^* - \hat{f}_{v_{j_s}}(D_m, .) \right\|_2^2 \right] \leq C n^{-2s/(1+2s)}.$$

Therefore, for any $p \in [1, \infty]$, $s \in (1/p, \infty)$, $q \in [1, \infty]$ and n large enough, the previous inequality and (5.1) yield

$$\begin{aligned}
&\sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}\Big[\|\tilde{f} - f^*\|_2^2\Big] = \sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}\Big[\mathbb{E}\Big[\|\tilde{f} - f^*\|_2^2 \,|D_m\Big]\Big] \\
&\leq 2 \sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}\Big[\min_{u \in \Lambda_n} \Big\| f^* - \hat{f}_{v_u}(D_m, .) \Big\|_2^2 \Big] + \frac{4(\log n)\log(\log n)}{\beta_2 n} \\
&\leq 2 \sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}\left[\left\| f^* - \hat{f}_{v_{j_s}}(D_m, .) \right\|_2^2 \right] + \frac{4(\log n)\log(\log n)}{\beta_2 n} \leq C n^{-2s/(1+2s)}.
\end{aligned}$$

This completes the proof of Theorem 4.1.

Acknowledgement

We would like to thank the referee and an associate editor for their useful comments which have helped to improve the presentation of the paper.

References

Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.* **34**, 584-653.

Abramovich, F., Sapatinas, T. and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *J. Roy. Statist. Soc. Ser. B* **60**, 725-749.

Augustin, N. H., Buckland, S. T., and Burnham, K. P. (1997). Model selection: An integral part of inference. *Boimetrics* **53**, 603-618.

Birgé, L. (2006). Model selection via testing: an alternative to (penalized) maximum likelihood estimators. *Ann. Inst. H. Poincaré Probab. Statist.* **42**, 273-325.

Bunea, F., and Nobel, A. (2005). Online prediction algorithms for aggregation of arbitrary estimators of a conditional mean. Submitted to *IEEE Trans. Inform. Theory*.

Cai, T., and Chicken, E. (2005). Block thresholding for density estimation: local and global adaptivity. *J. Multivariate Anal.* **95**, 76-106.

Catoni, O. (2001). *Statistical Learning Theory and Stochastic Optimization.* Ecole d'été de Probabilités de Saint-Flour 2001, Lectures Notes in Mathematics. Springer, New York.

Daubechies, I. (1992). *Ten Lectures on Wavelets*. CBMS-NSF Reg. Conf. Series in Applied Math. SIAM, Philadelphia.

Delyon, B., and Juditsky, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3**, 215–228.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **60**, 1200-1224.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1995). Wavelet shrinkage: Asymptotia ? *J. Roy. Statist. Soc. Ser. B* **57**, 301–369.

Donoho, D. L., Johnstone, I. M., Kerkyacharian, G., and Picard, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** : 508–539.

Gao, H. Y. (1998). Wavelet shrinkage denoising using the nonnegative garrote. *J. Comput. Graph. Statist.* **7**, 469-488.

Härdle, W., Kerkyacharian, G., Picard, D., and Tsybakov, A. (1998). Wavelet, Approximation and Statistical Applications. Volume **129** of *Lectures Notes in Statistics*. Springer Verlag, New York.

Jansen, M. (2001). Noise Reduction by Wavelet Thresholding, Volume **161** of *Lectures Notes in Statistics*. Springer Verlag, New York.

Juditsky, A. (1997). Wavelet estimators: adapting to unknown smoothness. *Math. Methods Statist.* **1**, 1-20.

Juditsky, A. and Nemirovski, A. (2000). Functional aggregation for nonparametric estimation. *Ann. Statist.* **28**, 681-712.

Lecué, G. (2005). Lower bounds and aggregation in density estimation. *J. Mach. Learn. Res.* **7**, 971-981.

Lecué, G. (2006). Optimal oracle inequality for aggregation of classifiers under low noise condition. *In Proceeding of the 19th Annual Conference on Learning Theory* **32**, 364–378.

Lecué, G. (2007a). Optimal rates of aggregation in classification. *Bernoulli* **13**, 1000-1022.

Lecué, G. (2007b). Simultaneous adaptation to the margin and to complexity in classification. *Ann. Statist.* **35**, 1698-1721.

Leung, G., and Barron, A. (2006). Information theory and mixing least-square regressions. *IEEE Trans. Inform. Theory* **52**, 3396-3410.

Mammen, E., and Tsybakov, A. (1999). Smooth discrimination analysis. *Ann. Statist.* **27**, 1808-1829.

Massart, P. (2006). Concentration Inequalities and Model Selection. Ecole d'été de Probabilités de Saint-Flour 2003. Lecture Notes in Mathematics, Springer.

Meyer, Y. (1990). *Ondelettes et Opérateurs.* Hermann, Paris.

Nason, G. P. (1995). *Choice of the Threshold Parameter in Wavelet Function Estimation*, volume **103**.

Nemirovski, A. (2000). *Topics in Non-parametric Statistics*, volume 1738 of *Ecole d'été de Probabilités de Saint-Flour* 1998*, Lecture Notes in Mathematics*. Springer, New York.

Steinwart, I., and Scovel, C. (2007). Fast rates for support vector machines using Gaussian kernels. *Ann. Statist.* **35**, 575-607.

Tsybakov, A. (2003). Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines*. (Edited by B. Schölkopf and M. Warmuth), 2777, 303-313. Springer, Heidelberg.

Tsybakov, A. (2004). Optimal aggregation of classifiers in statistical learning. *Ann. Statist.* **32**, 135-166.

Yang, Y. (2000). Mixing strategies for density estimation. *Ann. Statist.* **28**, 75-87.

Yang, Y. (2001). Minimax rate adaptive estimation over continuous hyper-parameters. *IEEE Trans. Inform. Theory* **47**, 2081-2085.

Université de Caen, LMNO, Campus II, Bâtiment Science 3, Bureau 332, 14032, CAEN, France.

E-mail: chesneau@math.unicaen.fr

Laboratoire d'Analyse et de Mathématiques Appliquées Université de Marne-la-Vallée 5, boulevard Descartes, Cité Descartes − Champs-sur-Marne, 77454 Marne-la-Vallée cedex 2, France.

E-mail: lecueguillaume@yahoo.fr

(Received April 2007; accepted September 2007)