Adapting to Unknown Smoothness by Aggregation of Thresholded Wavelet Estimators.

Christophe Chesneau, Guillaume Lecué

Adapting to Unknown Smoothness by Aggregation of Thresholded Wavelet Estimators.

Christophe Chesneau and Guillaume Lecué
Université Paris VI

Abstract

We study the performances of an adaptive procedure based on a convex combination, with data-driven weights, of term-by-term thresholded wavelet estimators. For the bounded regression model, with random uniform design, and the nonparametric density model, we show that the resulting estimator is optimal in the minimax sense over all Besov balls under the L^2 risk, without any logarithm factor.

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. A well-known example is the hard thresholded estimator introduced by [21]. If we observe n statistical data and if the unknown function f has an expansion of the form $f = \sum_j \sum_k \beta_{j,k}\psi_{j,k}$ where $\{\psi_{j,k},\ j,k\}$ is a wavelet basis and $(\beta_{j,k})_{j,k}$ is the associated wavelet coefficients, then the term-by-term wavelet thresholded method consists in three steps:

1. a linear step corresponding to the estimation of the coefficients $\beta_{j,k}$ by some estimators $\hat{\beta}_{j,k}$ constructed from the data,

2. a non-linear step consisting in a thresholded procedure $T_\lambda(\hat{\beta}_{j,k})\mathbb{1}_{\left\{|\hat{\beta}_{j,k}|\geq\lambda_j\right\}}$ where $\lambda=(\lambda_j)_j$ is a positive sequence and $T_\lambda(\hat{\beta}_{j,k})$ denotes a certain transformation of the $\hat{\beta}_{j,k}$ which may depend on λ,

3. a reconstruction step of the form $\hat{f}_\lambda=\sum_{j\in\Omega_n}\sum_k T_\lambda(\hat{\beta}_{j,k})\mathbb{1}_{\left\{|\hat{\beta}_{j,k}|\geq\lambda_j\right\}}\psi_{j,k}$ where Ω_n is a finite set of integers depending on the number n of data.

Naturally, the performances of $\hat{f}_\lambda$ strongly depend on the choice of the threshold λ. For the standard statistical models (regression, density,...), the most common choice is the universal threshold introduced by [21]. It can be expressed in the form: $\lambda^*=(\lambda_j^*)_j$ where $\lambda_j^*=c\sqrt{(\log n)/n}$ where $c>0$ denotes a large enough constant. In the literature, several technics have been proposed to determine the 'best' adaptive threshold. There are, for instance, the RiskShrink and SureShrink methods (see [20, 21]), the cross-validation methods (see [45], [53] and [31]), the methods based on hypothesis tests (see [1] and [2]), the Lepski methods (see [33]) and the Bayesian methods (see [17] and [3]). Most of them are described in detailed in [45] and [4].

In the present paper, we propose to study the performances of an adaptive wavelet estimator based on a convex combination of $\hat{f}_\lambda$'s. In the framework of nonparametric density estimation and bounded regression estimation with random uniform design, we prove that, in some sense, it is at least as good as the term-by-term thresholded estimator $\hat{f}_\lambda$ defined with the 'best' threshold λ. In particular, we show that this estimator is optimal, in the minimax sense, over all Besov balls under the L^2 risk. The proof is based on a non-adaptive minimax result proved by [19] and some powerful oracle inequality satisfied by aggregation methods. There are two steps in our approach. A first step, called the training step, where non-adaptive thresholded wavelet estimators are constructed for different thresholds. A second step, called learning step, where an aggregation scheme is worked out to realize the adaptation to the smoothness.

The exact oracle inequality of Section 2 is given in a general framework. Two aggregation procedures satisfy this oracle inequality. The well known ERM (for Empirical Risk Minimization) procedure (cf. [51], [38] and references therein) and an exponential weighting aggregation scheme, which has been studied, among others, by [5], [8], [40], [41] and [39]. There is a recursive version of this scheme studied by [13], [54], [35] and [36]. In the sequential prediction problem, weighted average predictions

with exponential weights have been widely studied (cf. e.g. [52] and [15]). A recent result of [42] shows that the ERM procedure is suboptimal for strictly convex losses (which is the case for density and regression estimation when the integrated squared risk is used). Thus, in our case it is better to combine the $\hat{f}_\lambda$'s, for λ lying in a grid, using the aggregation procedure with exponential weights than using the ERM procedure. Moreover, from a computation point of view the aggregation scheme with exponential weights does not require any minimization step contrarily to the ERM procedure.

The paper is organized as follows. Section 2 presents general oracle inequalities satisfied by two aggregation methods. Section 3 describes the main procedure of the study and investigates its minimax performances over Besov balls for the L^2 risk. All the proofs are postponed in the last section.

2 Oracle Inequalities

2.1 Framework

Let $(\mathcal{Z}, \mathcal{T})$ 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". It means that there exists $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$ such that the risk of an estimate $f \in \mathcal{F}$ of $F(\pi)$ is of the form

$$A(f) = \mathbb{E}\left[Q(Z, f)\right].$$

In what follows, we present several statistical problems which can be written in this way. If the minimum over all f in $\mathcal{F}$

$$A^* \stackrel{\text{def}}{=} \min_{f \in \mathcal{F}} A(f)$$

is achieved by at least one function, we denote by f^* a minimizer in $\mathcal{F}$. In this paper

we will assume that $\min_{f\in\mathcal{F}} A(f)$ 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 of the cases f^* will be equal to our aim $F(\pi)$ up to some known additive terms. We don't know the risk A, since π is not available from the statistician, thus, instead of minimizing A over $\mathcal{F}$ we consider an empirical version of A constructed from the observations D_n. The main interest of such a framework is that we have access to an empirical version of $A(f)$ for any $f \in \mathcal{F}$. It is denoted by

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

We exhibit three statistical models having the previous form of estimation.

Bounded Regression: Take $\mathcal{Z} = \mathcal{X} \times [0,1]$, where $(\mathcal{X}, \mathcal{A})$ is a measurable space, $Z = (X, Y)$ a couple of random variables on $\mathcal{Z}$, with probability distribution π, such that X takes its values in $\mathcal{X}$ and Y takes its values in $[0,1]$. We assume that the conditional expectation $\mathbb{E}[Y|X]$ exists. In the regression framework, we want to estimate the regression function

$$f^*(x) = \mathbb{E}\left[Y|X = x\right], \ \forall x \in \mathcal{X}.$$

Usually, the variable Y is not an exact function of X. Given is an input $X \in \mathcal{X}$, we are not able to predict the exact value of the output $Y \in [0,1]$. This issue can be seen in the regression framework as a noised estimation. It means that in each spot X of the input set, the predicted label Y is concentrated around $\mathbb{E}\left[Y|X\right]$ up to an additional noise with null mean denoted by ζ. The regression model can then be written as

$$Y = \mathbb{E}\left[Y|X\right] + \zeta.$$

Take $\mathcal{F}$ the set of all measurable functions from $\mathcal{X}$ to $[0,1]$. Define $||f||^2_{L^2(P^X)} = \int_{\mathcal{X}} f^2(x) dP^X(x)$ for all functions f in $L^2(\mathcal{X}, \mathcal{A}, P^X)$ where P^X is the probability measure of X. Consider

$$Q((x,y), f) = (y - f(x))^2, \tag{2}$$

for any $(x,y) \in \mathcal{X} \times \mathbb{R}$ and $f \in \mathcal{F}$. Pythagore's Theorem yields

$$A(f) = \mathbb{E}\left[Q((X,Y), f)\right] = ||f^* - f||^2_{L^2(P^X)} + \mathbb{E}\left[\zeta^2\right].$$

Thus f^* is a minimizer of $A(f)$ and $A^* = \mathbb{E}[\zeta^2]$.

Density estimation: Let $(\mathcal{Z}, \mathcal{T}, \mu)$ be a measured space. Let Z be a random variable with values in $\mathcal{Z}$ and denote by π its probability distribution. We assume that π is absolutely continuous w.r.t. to μ and denote by f^* one version of the density. Consider $\mathcal{F}$ the set of all density functions on $(\mathcal{Z}, \mathcal{T}, \mu)$. We consider

$$Q(z, f) = -\log f(z),$$

for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. We have

$$A(f) = \mathbb{E}\left[Q(Z, f)\right] = K(f^*|f) - \int_{\mathcal{Z}} \log(f^*(z)) d\pi(z).$$

Thus, f^* is a minimizer of $A(f)$ and $A^* = -\int_{\mathcal{Z}} \log(f^*(z)) d\pi(z)$.

Instead of using the Kullback-Leiber loss, one can use the quadratic loss. For this setup, consider $\mathcal{F}$ the set $L^2(\mathcal{Z}, \mathcal{T}, \mu)$ of all measurable functions with an integrated square. Define

$$Q(z, f) = \int_{\mathcal{Z}} f^2 d\mu - 2f(z), \tag{3}$$

for any $z \in \mathcal{Z}$ and $f \in \mathcal{F}$. We have, for any $f \in \mathcal{F}$,

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

Thus, f^* is a minimizer of $A(f)$ and $A^* = -\int_{\mathcal{Z}} (f^*(z))^2 d\mu(z)$.

Classification framework: Let $(\mathcal{X}, \mathcal{A})$ be a measurable space. We assume that the space $\mathcal{Z} = \mathcal{X} \times \{-1, 1\}$ is endowed with an unknown probability measure π. We consider a random variable $Z = (X, Y)$ with values in $\mathcal{Z}$ with probability distribution π. We denote by P^X the marginal of π on $\mathcal{X}$ and $\eta(x) = \mathbb{P}(Y = 1|X = x)$ the conditional probability function of $Y = 1$ knowing that $X = x$. Denote by $\mathcal{F}$ the set of all measurable functions from $\mathcal{X}$ to $\mathbb{R}$. Let ϕ be a function from $\mathbb{R}$ to $\mathbb{R}$. For any $f \in \mathcal{F}$ consider the $\phi-$risk

$$A(f) = \mathbb{E}[Q((X, Y), f)],$$

where the loss is given by $Q((x, y), f) = \phi(yf(x))$for any $(x, y) \in \mathcal{X} \times \{-1, 1\}$.

Most of the time a minimizer f^* of the $\phi-$risk A over $\mathcal{F}$ or its sign is equal to the Bayes rule $f^*(x) = \text{Sign}(2\eta(x) - 1), \forall x \in \mathcal{X}$ (cf. [56]).

In this paper we obtain an oracle inequality in the general framework described at the beginning of this Subsection. Then, we use it in the density estimation and the bounded regression frameworks. For applications of this oracle inequality in the classification setup, we refer to [41] and [40].

Now, we introduce an assumption which improve the quality of estimation in our framework. This assumption has been first introduced by [43], for the problem of discriminant analysis, and [50], for the classification problem. With this assumption, parametric rates of convergence can be achieved, for instance, in the classification problem (cf. [50], [48]).

Margin Assumption(MA): *The probability measure π satisfies the margin assumption MA($\kappa, c, \mathcal{F}_0$), where $\kappa \geq 1, c > 0$ and $\mathcal{F}_0$ is a subset of $\mathcal{F}$ if*

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

for any function $f \in \mathcal{F}_0$.

In the bounded regression setup, it is easy to see that any probability distribution π on $\mathcal{X} \times [0,1]$ naturally satisfies the margin assumption $\mathrm{MA}(1, 16, \mathcal{F}_1)$, where $\mathcal{F}_1$ is the set of all measurable functions from $\mathcal{X}$ to $[0,1]$. In density estimation with the integrated squared risk, all probability measures π on $(\mathcal{Z}, \mathcal{T})$ absolutely continuous w.r.t. the measure μ with one version of its density a.s. bounded by a constant $B \geq 1$, 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.

Actually, the margin assumption is linked to the convexity of the underlying loss. In density and regression estimation it is naturally satisfied with the better margin parameter $\kappa = 1$, but, for non-convex loss (for instance in classification) this assumption does not hold naturally (cf. [42] for a discussion on the margin assumption and for examples of losses which does not satisfied naturally the margin assumption with parameter $\kappa = 1$).

2.2 Aggregation Procedures

Let's work with the notations introduced in the beginning of the previous Subsection. The aggregation framework considered, among others, by [34], [54], [13],[46], [49], [5], [6] 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 two aggregation procedures.

The Aggregation with Exponential Weights aggregate (**AEW**) over $\mathcal{F}_0$ is defined by

$$\tilde{f}_n^{(AEW)} \stackrel{\text{def}}{=} \sum_{f \in \mathcal{F}_0} w^{(n)}(f) f, \tag{4}$$

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

$$w^{(n)}(f) = \frac{\exp\left(-nA_n(f)\right)}{\sum_{g \in \mathcal{F}_0} \exp\left(-nA_n(g)\right)}, \quad \forall f \in \mathcal{F}_0. \tag{5}$$

We consider the Empirical Risk Minimization procedure **(ERM)** over $\mathcal{F}_0$ defined by

$$\tilde{f}_n^{(ERM)} \in \operatorname{Arg}\min_{f \in \mathcal{F}_0} A_n(f). \tag{6}$$

2.3 Oracle Inequalities

In this Subsection we state an exact oracle inequality satisfied by the ERM procedure and the AEW procedure (in the convex case) in the general framework of the beginning of Subsection 2.1. From this exact oracle inequality we deduce two others oracle inequalities in the density estimation and the bounded regression framework. We introduce a quantity which is going to be our residual term in the exact oracle inequality. We consider

$$\gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q) = \begin{cases} \left(\frac{\mathcal{B}(\mathcal{F}_0, \pi, Q)^{\frac{1}{\kappa}} \log M}{\beta_1 n}\right)^{1/2} & \text{if } \mathcal{B}(\mathcal{F}_0, \pi, Q) \geq \left(\frac{\log M}{\beta_1 n}\right)^{\frac{\kappa}{2\kappa-1}} \\ \left(\frac{\log M}{\beta_2 n}\right)^{\frac{\kappa}{2\kappa-1}} & \text{otherwise,} \end{cases}$$

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

$$\beta_1 = \min\Big(\frac{\log 2}{96cK}, \frac{3\sqrt{\log 2}}{16K\sqrt{2}}, \frac{1}{8(4c + K/3)}, \frac{1}{576c}\Big). \tag{7}$$

and

$$\beta_2 = \min\Big(\frac{1}{8}, \frac{3\log 2}{32K}, \frac{1}{2(16c + K/3)}, \frac{\beta_1}{2}\Big), \tag{8}$$

where the constant $c > 0$ appears in $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$.

Theorem 1. *Consider the general framework introduced in the beginning of Subsection 2.1. Let $\mathcal{F}_0$ denote a finite subset of M elements $f_1, \dots, f_M$ in $\mathcal{F}$, where $M \geq 2$ is an integer. Assume that the underlying probability measure π satisfies the margin assumption $MA(\kappa, c, \mathcal{F}_0)$ for some $\kappa \geq 1, c > 0$ and $|Q(Z, f) - Q(Z, f^*)| \leq K$ a.s., for any $f \in \mathcal{F}_0$, where $K \geq 1$ is a constant. The Empirical Risk Minimization procedure (6) satisfies*

$$\mathbb{E}[A(\tilde{f}_n^{(ERM)}) - A^*] \leq \min_{j=1,\dots,M} (A(f_j) - A^*) + 4\gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q).$$

Moreover, if $f \longmapsto Q(z, f)$ is convex for π-almost $z \in \mathcal{Z}$, then the AEW procedure satisfies the same oracle inequality as the ERM procedure.

Now, we give two corollaries of Theorem 1 in the density estimation and bounded regression framework.

Corollary 1. *Consider the bounded regression setup. Let $f_1, \dots, f_M$ be M functions on $\mathcal{X}$ with values in $[0, 1]$. Let $\tilde{f}_n$ denote either the ERM or the AEW procedure. For β_2 defined in (8) and for any $\epsilon > 0$, we have*

$$\mathbb{E}[||f^* - \tilde{f}_n||^2_{L^2(P^X)}] \leq (1 + \epsilon) \min_{j=1,\dots,M} (||f^* - f_j||^2_{L^2(P^X)}) + \frac{4 \log M}{\epsilon \beta_2 n}.$$

Corollary 2. *Consider the density estimation framework. Assume that the underlying density function f^* to estimate is bounded by $B \geq 1$. Let $f_1, \dots, f_M$ be M functions bounded from above and below by B. Let $\tilde{f}_n$ denote either the ERM or the AEW procedure. For β_2 defined in (8) and any $\epsilon > 0$, we have*

$$\mathbb{E}[||f^* - \tilde{f}_n||^2_{L^2(\mu)}] \leq (1 + \epsilon) \min_{j=1,\dots,M} (||f^* - f_j||^2_{L^2(\mu)}) + \frac{4 \log M}{\epsilon \beta_2 n}. \tag{9}$$

In both of the last Corollaries, the ERM and the AEW procedures can both be used to mimic the best f_j among the f_j's. Nevertheless, from a computational point of view the AEW procedure does not require any minimization step contrarily to the ERM procedure. Moreover, from a theoretical point of view the ERM procedure can not mimic the best f_j among the f_j's as fast as the cumulative aggregate with exponential weights (it is an average of AEW procedures). For a comparison between these procedures we refer to [42]. The constants of aggregation multiplying the residual term in Theorem 1 and in both of the following Corollaries come from the proof and are certainly not optimal. We did not make any serious attempt to optimize them.

3 Multi-thresholding wavelet estimator

In the present section, we propose an adaptive estimator constructed from aggregation technics and wavelet thresholding methods. For the density model and the regression model with uniform random design, 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 the purposes of this paper, we use the periodized wavelets bases on the unit interval. Let

$$\phi_{j,k} = 2^{j/2}\phi(2^j x - k), \qquad \psi_{j,k} = 2^{j/2}\psi(2^j x - k)$$

be the elements of the wavelet basis and

$$\phi_{j,k}^{per}(x) = \sum_{l\in\mathbb{Z}} \phi_{j,k}(x-l), \qquad \psi_{j,k}^{per}(x) = \sum_{l\in\mathbb{Z}} \psi_{j,k}(x-l),$$

there 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 $\zeta = \{\phi_{j,k}^{per}, k = 0,...,2^\tau - 1;\ \psi_{j,k}^{per},\quad j = \tau,...,\infty,\ k = 0,...,2^j-1\}$ constitutes an orthonormal basis of $L^2([0,1])$. In what follows, the superscript "*per*" will be suppressed from the notations for convenience. For any integer $l \geq \tau$, a square-integrable function f^* on $[0,1]$ can be expanded into a wavelet series

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

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 [44] and [18].

Now, let us define the main function spaces of the study. Let $M \in (0,\infty)$, $s \in (0,N)$, $p \in [1,\infty)$ and $q \in [1,\infty)$. Let us set $\beta_{\tau-1,k} = \alpha_{\tau,k}$. We say that a function f^* belongs to the Besov balls $B_{p,q}^s(M)$ if and only if the associated wavelet

coefficients satisfy

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

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 [44]).

3.2 Term-by-term thresholded estimator

In this Subsection, we consider the estimation of an unknown function f^* in $L^2([0,1])$ from a general situation. We only assume to have n observations gathered in the data set D_n from which we are able to estimate the wavelet coefficients $\alpha_{j,k}$ and $\beta_{j,k}$ of f^* in the basis ζ. We denote by $\hat{\alpha}_{j,k}$ and $\hat{\beta}_{j,k}$ such estimates. Finally, let us mention that all the constants of our study are independent of f^* and n.

Definition 1 (Term-by-term thresholded estimator). *Let j_1 be an integer satisfying $(n/\log n)\leq 2^{j_1}<2(n/\log n)$. For any integer $l\geq\tau$, let $\lambda=(\lambda_l,...\lambda_{j_1})$ be a vector of positive integers. Let us consider the estimator $\hat{f}_\lambda:D_n\times[0,1]\to\mathbb{R}$ defined by*

$$\hat{f}_\lambda(D_n,x)=\sum_{k=0}^{2^\tau-1}\hat{\alpha}_{\tau,k}\phi_{\tau,k}(x)+\sum_{j=\tau}^{j_1}\sum_{k=0}^{2^j-1}\Upsilon_{\lambda_j}(\hat{\beta}_{j,k})\psi_{j,k}(x),\tag{10}$$

where for all $u\in(0,\infty)$ the operator Υ_u is such that there exist two constants $C_1,C_2>0$ satisfying

$$|\Upsilon_u(x)-y|^2\leq C_1(\min(y,C_2u)^2+(|x-y|^2)\mathbb{1}_{\{|x-y|\geq 2^{-1}u\}}),\tag{11}$$

for any $x\in\mathbb{R}$ and $y\in\mathbb{R}$.

The inequality (11) 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 [21], [22] and [19]) and the non-negative garrote thresholding rule $\Upsilon_u^{NG}(x)=(x-u^2/x)\,\mathbb{1}_{\{|x|\geqslant u\}}$ (see [26]).

If we consider the minimax point of view over Besov balls under the integrated squared risk, then [19] makes the conditions on $\hat{\alpha}_{j,k}$, $\hat{\beta}_{j,k}$ and the threshold λ such that the estimator $\hat{f}_\lambda(D_n,.)$ defined by (10) is optimal for numerous statistical models. This result is recalled in Theorem 2 below.

Theorem 2 (Delyon and Juditsky (1996)). *Let us consider the general statistical framework described in the beginning of the present section. Suppose that the two following assumptions hold.*

- *Moments inequality: There exists a constant $C > 0$ such that, for any $j \in \{\tau - 1, ..., j_1\}$, $k \in \{0, ..., 2^j - 1\}$ and n large enough, we have*

$$\mathbb{E}(|\hat{\beta}_{j,k} - \beta_{j,k}|^4) \leq Cn^{-2}, \text{ where we take } \hat{\beta}_{\tau-1,k} = \hat{\alpha}_{\tau,k}. \tag{12}$$

- *Large deviation inequality: There exist two constants $C > 0$ and $\rho_* > 0$ such that, for any $a, j \in \{\tau, ..., j_1\}$, $k \in \{0, ..., 2^j - 1\}$ and n large enough, we have*

$$\mathbb{P}\left(2\sqrt{n}|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \rho_*\sqrt{a}\right) \leqslant C2^{-4a}. \tag{13}$$

Let us consider the term-by-term thresholded estimator $\hat{f}_{v_{j_s}}(D_n, .)$ defined by (10) with the threshold

$$v_{j_s} = (\rho_*(j - j_s)_+)_{j=\tau,...,j_1},$$

where j_s is an integer such that $n^{1/(1+2s)} \leq 2^{j_s} < 2n^{1/(1+2s)}$. Then, there exists a constant $C > 0$ such that, for any $p \in [1, \infty]$, $s \in (1/p, N]$, $q \in [1, \infty]$ and n large enough, we have:

$$\sup_{f \in B^s_{p,q}(L)} \mathbb{E}[\|\hat{f}_{v_{j_s}}(D_n, .) - f^*\|^2_{L^2([0,1])}] \leqslant Cn^{-2s/(2s+1)}.$$

The rate of convergence $V_n = n^{-2s/(1+2s)}$ is minimax for numerous statistical models, where s is a regularity parameter. For the density model and the regression model with uniform design, we refer the reader to [19] for further details about the choice of the estimator $\hat{\beta}_{j,k}$ and the value of the thresholding constant ρ_*. Starting from this non-adaptive result, we use aggregation methods to construct an adaptive estimator at least at good in the minimax sense as $\hat{f}_{v_{j_s}}(D_n, .)$.

3.3 Multi-thresholding estimator

Let us divide our 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 last remaining observations, where we take

$$l = \lceil n/\log n \rceil \text{ and } m = n - l.$$

The first subsample D_m, sometimes called "training sample", is used to construct a family of estimators (in our case this is thresholded estimators) and the second subsample $D^{(l)}$, called the "training sample", is used to construct the weights of the aggregation procedure.

Remark 1. *From a theoretical point of view we can take $m = l$ which means that we use as many observations for the estimation step as for the learning step. But, in practice it is better to use a greater part of the observations for the construction of the estimators and the last observations for the aggregation procedure, because if the basis estimators that we aggregate, are not good, then the obtained aggregate is likely to be as bad as the prior estimators. Another interesting thing is that we can split the whole sample D_n in many different ways. For instance we can take m observations randomly in D_n to form the training subsample and the last remaining observations for the learning subsample. We can also take an average of different aggregates constructed from different splits of the initial sample D_n and by a simple argument of convexity it is easy to prove that the averaged aggregate has a better risk than the others aggregates constructed only from one split.*

Definition 2. *Let us consider the term-by-term thresholded estimator described in (10). Assume that we want to estimate a function f^* from $[0,1]$ with values in $[a,b]$. Consider the projection function*

$$h_{a,b}(y) = \max(a, \min(y, b)), \forall y \in \mathbb{R}. \tag{14}$$

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

$$\tilde{f}_n(x) = \sum_{u \in \Lambda_n} w^{(l)}(h_{a,b}(\hat{f}_{v_u}(D_m, .)))h_{a,b}(\hat{f}_{v_u}(D_m, x)), \tag{15}$$

where $\Lambda_n = \{0, ..., \log n\}$, $v_u = (\rho(j-u)_+)_{j=\tau,...,j_1}, \forall u \in \Lambda_n$ and ρ is a positive constant depending on the model worked out and

$$w^{(l)}(h_{a,b}(\hat{f}_{v_u}(D_m, .))) = \frac{\exp\left(-lA^{(l)}(h_{a,b}(\hat{f}_{v_u}(D_m, .)))\right)}{\sum_{\gamma \in \Lambda_n} \exp\left(-lA^{(l)}(h_{a,b}(\hat{f}_{v_\gamma}(D_m, .)))\right)}, \quad \forall u \in \Lambda_n,$$

where $A^{(l)}(f) = \frac{1}{l}\sum_{i=m+1}^{n} Q(Z_i, f)$ *is the empirical risk constructed from the* l *last observations, for any function* f *and for the choice of a loss function* Q *depending on the model considered (cf. (2) and (3) for examples).*

The principle of the construction of the multi-thresholding estimator $\tilde{f}_n$ is to use aggregation technics to easily construct an adaptive optimal estimator of f^*. It 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 element in Λ_n depending on the regularity of f^* such that the non-adaptive estimator $\hat{f}_{v_u}(D_m, .)$ is optimal in the minimax sense (see Theorem 2), the multi-thresholding estimator is optimal independently of the regularity of f^*.

4 Performances of the multi-thresholding estimator

This section is devoted to the minimax performances of the multi-thresholding estimator defined in (15) under the $L^2([0,1])$ risk over Besov balls. Firstly, we consider the framework of the density model. Secondly, we focus our attention on the bounded regression with uniform random design. Finally, we compare these results with some well-known wavelet thresholded procedures.

4.1 Density model

In the density estimation model, Theorem 3 below investigates rates of convergence achieved by the multi-thresholding estimator (defined by (15)) under the $L^2([0,1])$ risk over Besov balls.

Theorem 3. *Let us consider the problem of estimating* f^* *from the density model. Assume that there exists* $B \geq 1$ *such that the underlying density function* f^* *to estimate is bounded by* B*. Let us consider the multi-thresholding estimator defined in (15) where we take* $a = 0, b = B$*,* ρ *such that*

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

and

$$\hat{\alpha}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \phi_{j,k}(X_i), \qquad \hat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} \psi_{j,k}(X_i). \tag{16}$$

Then, there exists a constant $C > 0$ such that

$$\sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[\|\tilde{f}_n - f^*\|^2_{L^2([0,1])}] \leqslant C n^{-2s/(2s+1)},$$

for any $p \in [1, \infty]$, $s \in (p^{-1}, N]$, $r \in [1, \infty]$ and integer n.

The rate of convergence $V_n = n^{-2s/(1+2s)}$ is minimax over $B^s_{p,q}(L)$. Further details about the minimax rate of convergence over Besov balls under the $L^2([0,1])$ risk for the density model can be found in [19] and [29]. For further details about the density estimation via adaptive wavelet thresholded estimators, see [23], [19] and [47]. See also [30] for a practical study.

4.2 Bounded regression

In the framework of the bounded regression model with uniform random design, Theorem 4 below investigates the rate of convergence achieved by the multi-thresholding estimator defined by (15) under the $L^2([0,1])$ risk over Besov balls.

Theorem 4. *Let us consider the problem of estimating the regression function f^* in the bounded regression model with random uniform design. Let us consider the multi-thresholding estimator (15) with ρ such that*

$$\frac{\rho^2}{8 + (8\rho/(3\sqrt{2}))(\|\psi\|_\infty + 1)} \geq 4(\log 2)$$

and

$$\hat{\alpha}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i \phi_{j,k}(X_i), \qquad \hat{\beta}_{j,k} = \frac{1}{n}\sum_{i=1}^{n} Y_i \psi_{j,k}(X_i). \tag{17}$$

Then, there exists a constant $C > 0$ such that, for any $p \in [1, \infty]$, $s \in (p^{-1}, N]$, $q \in [1, \infty]$ and integer n, we have

$$\sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[\|\tilde{f}_n - f^*\|^2_{L^2([0,1])}] \leqslant C n^{-2s/(2s+1)}.$$

The rate of convergence $V_n = n^{-2s/(1+2s)}$ is minimax over $B^s_{p,q}(L)$. The multi-thresholding estimator has better minimax properties than several other wavelet estimators developed in the literature. To the authors's knowledge, the result obtained, for instance, by the hard thresholded estimator (see [21]), by the global wavelet block thresholded estimator (see [37]), by the localized wavelet block thresholded estimator (see [9, 12, 10], [28, 27], [24, 25], [16] and [11]) and, in particular, the penalized Blockwise Stein method (see [14]) are worse than the one obtained by the multi-thresholding estimator and stated in Theorems 3 and 4. This is because, on the difference of those works, we obtain the optimal rate of convergence without any extra logarithm factor.

In fact, the multi-thresholding estimator has similar minimax performances than the empirical Bayes wavelet methods (see [55] and [32]) and several term-by-term wavelet thresholded estimators defined with a random threshold (see [33] and [7]).

Finally, it is important to mention that the multi-thresholding estimator does not need any minimization step and is relatively easy to implement.

5 Proofs

Proof of Theorem 1. We recall the notations of the general framework introduced in the beginning of Subsection 2.1. Consider a loss function $Q : \mathcal{Z} \times \mathcal{F} \longmapsto \mathbb{R}$, the risk $A(f) = \mathbb{E}[Q(Z, f)]$, the minimum risk $A^* = \min_{f \in \mathcal{F}} A(f)$, where we assume, w.o.l.g, that it is achieved by an element f^* in $\mathcal{F}$ and the empirical risk $A_n(f) = (1/n) \sum_{i=1}^n Q(Z_i, f)$, for any $f \in \mathcal{F}$. The following proof is a generalization of the proof of Theorem 1 in [39].

We first start by a 'linearization' of the risk. Consider the convex set

$$\mathcal{C} = \left\{ (\theta_1, \ldots, \theta_M) : \theta_j \geq 0 \text{ and } \sum_{j=1}^M \theta_j = 1 \right\}$$

and define the following functions on $\mathcal{C}$

$$\tilde{A}(\theta) \stackrel{\text{def}}{=} \sum_{j=1}^M \theta_j A(f_j) \text{ and } \tilde{A}_n(\theta) \stackrel{\text{def}}{=} \sum_{j=1}^M \theta_j A_n(f_j)$$

which are linear versions of the risk A and its empirical version A_n.

Using the Lagrange method of optimization we find that the exponential weights $w \stackrel{\text{def}}{=} (w^{(n)}(f_j))_{1\le j\le M}$ are the unique solution of the minimization problem

$$\min\Big(\tilde{A}_n(\theta) + \frac{1}{n}\sum_{j=1}^{M}\theta_j \log\theta_j : (\theta_1,\ldots,\theta_M)\in\mathcal{C}\Big),$$

where we use the convention $0\log 0 = 0$. Take $\hat{\jmath}\in\{1,\ldots,M\}$ such that $A_n(f_{\hat{\jmath}}) = \min_{j=1,\ldots,M} A_n(f_j)$. The vector of exponential weights w satisfies

$$\tilde{A}_n(w) \le \tilde{A}_n(e_{\hat{\jmath}}) + \frac{\log M}{n},$$

where e_j denotes the vector in $\mathcal{C}$ with 1 for j-th coordinate (and 0 elsewhere).

Let $\epsilon > 0$. Denote by $\tilde{A}_{\mathcal{C}}$ the minimum $\min_{\theta\in\mathcal{C}}\tilde{A}(\theta)$. We consider the subset of $\mathcal{C}$

$$\mathcal{D} \stackrel{\text{def}}{=} \Big\{\theta\in\mathcal{C} : \tilde{A}(\theta) > \tilde{A}_{\mathcal{C}} + 2\epsilon\Big\}.$$

Let $x > 0$. If

$$\sup_{\theta\in\mathcal{D}}\frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(\theta) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} \le \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x},$$

then for any $\theta\in\mathcal{D}$, we have

$$\tilde{A}_n(\theta) - A_n(f^*) \ge \tilde{A}(\theta) - A^* - \frac{\epsilon(\tilde{A}(\theta) - A^* + x)}{(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)} \ge \tilde{A}_{\mathcal{C}} - A^* + \epsilon,$$

because $\tilde{A}(\theta) - A^* \ge \tilde{A}_{\mathcal{C}} - A^* + 2\epsilon$. Hence,

$$\begin{aligned}
&\mathbb{P}\left[\inf_{\theta\in\mathcal{D}}\Big(\tilde{A}_n(\theta) - A_n(f^*)\Big) < \tilde{A}_{\mathcal{C}} - A^* + \epsilon\right] \\
&\quad\le\ \mathbb{P}\left[\sup_{\theta\in\mathcal{D}}\frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(\theta) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right]. \qquad (18)
\end{aligned}$$

Observe that a linear function achieves its maximum over a convex polygon at one of the vertices of the polygon. Thus, for $j_0\in\{1,\ldots,M\}$ such that $\tilde{A}(e_{j_0}) = \min_{j=1,\ldots,M}\tilde{A}(e_j)$ $(= \min_{j=1,\ldots,M} A(f_j))$, we have $\tilde{A}(e_{j_0}) = \min_{\theta\in\mathcal{C}}\tilde{A}(\theta)$. We obtain the last inequality by linearity of $\tilde{A}$ and the convexity of $\mathcal{C}$. Let $\hat{w}$ denotes either the exponential weights w or $e_{\hat{\jmath}}$. According to (18), We have

$$\tilde{A}(\hat{w}) \le \min_{j=1,\ldots,M}\tilde{A}_n(e_j) + \frac{\log M}{n} \le \tilde{A}_n(e_{j_0}) + \frac{\log M}{n}$$

So, if $\tilde{A}(\hat{w}) > A_{\mathcal{C}} + 2\epsilon$ then $\hat{w} \in \mathcal{D}$ and thus, there exists $\theta \in \mathcal{D}$ such that $\tilde{A}_n(\theta) - \tilde{A}_n(f^*) \leq \tilde{A}_n(e_{j_0}) - \tilde{A}_n(f^*) + (\log M)/n$. Hence, we have

$$\begin{aligned}
\mathbb{P}\left[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right] &\leq \mathbb{P}\left[\inf_{\theta\in\mathcal{D}} \tilde{A}_n(\theta) - A_n(f^*) \leq \tilde{A}_n(e_{j_0}) - A_n(f^*) + \frac{\log M}{n}\right] \\
&\leq \mathbb{P}\left[\inf_{\theta\in\mathcal{D}} \tilde{A}_n(\theta) - A_n(f^*) < \tilde{A}_{\mathcal{C}} - A^* + \epsilon\right] \\
&\quad + \mathbb{P}\left[\tilde{A}_n(e_{j_0}) - A_n(f^*) \geq \tilde{A}_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\right] \\
&\leq \mathbb{P}\left[\sup_{\theta\in\mathcal{C}} \frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(f) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right] \\
&\quad + \mathbb{P}\left[\tilde{A}_n(e_{j_0}) - A_n(f^*) \geq \tilde{A}_{\mathcal{C}} - A^* + \epsilon - \frac{\log M}{n}\right].
\end{aligned}$$

If we assume that

$$\sup_{\theta\in\mathcal{C}} \frac{\tilde{A}(\theta) - A^* - (\tilde{A}_n(\theta) - A_n(f^*))}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x},$$

then, there exists $\theta^{(0)} = (\theta_1^{(0)}, \ldots, \theta_M^{(0)}) \in \mathcal{C}$, such that

$$\frac{\tilde{A}(\theta^{(0)}) - A^* - (\tilde{A}_n(\theta^{(0)}) - A_n(f^*))}{\tilde{A}(\theta^{(0)}) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}.$$

The linearity of $\tilde{A}$ yields

$$\frac{\tilde{A}(\theta^{(0)}) - A^* - (\tilde{A}_n(\theta^{(0)}) - A_n(f^*))}{\tilde{A}(\theta^{(0)}) - A^* + x} = \frac{\sum_{j=1}^M \theta_j^{(0)}[A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{\sum_{j=1}^M \theta_j^{(0)}[A(f_j) - A^* + x]}$$

and since, for any numbers $a_1, \ldots, a_M$ and positive numbers $b_1, \ldots, b_M$, we have

$$\frac{\sum_{j=1}^M a_j}{\sum_{j=1}^M b_j} \leq \max_{j=1,\ldots,M} \left(\frac{a_j}{b_j}\right),$$

then, we obtain

$$\max_{j=1,\ldots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{A_{\mathcal{F}_0} - A^* + 2\epsilon + x},$$

where $A_{\mathcal{F}_0} \stackrel{\text{def}}{=} \min_{j=1,\ldots,M} A(f_j)$ $(= \tilde{A}_{\mathcal{C}})$.

Now, we use the relative concentration inequality of Lemma 1 to obtain

$$\begin{aligned}
&\mathbb{P}\left[\max_{j=1,\dots,M}\frac{A(f_j)-A^*-(A_n(f_j)-A_n(f^*))}{A(f_j)-A^*+x}>\frac{\epsilon}{A_{\mathcal{F}_0}-A^*+2\epsilon+x}\right]\\
&\leq\ M\left(1+\frac{4c(A_{\mathcal{F}_0}-A^*+2\epsilon+x)^2x^{1/\kappa}}{n(\epsilon x)^2}\right)\exp\left(-\frac{n(\epsilon x)^2}{4c(A_{\mathcal{F}_0}-A^*+2\epsilon+x)^2x^{1/\kappa}}\right)\\
&\quad+M\left(1+\frac{4K(A_{\mathcal{F}_0}-A^*+2\epsilon+x)}{3n\epsilon x}\right)\exp\left(-\frac{3n\epsilon x}{4K(A_{\mathcal{F}_0}-A^*+2\epsilon+x)}\right).
\end{aligned}$$

Using the margin assumption $\mathrm{MA}(\kappa,c,\mathcal{F}_0)$ to upper bound the variance term and applying Bernstein's inequality, we get

$$\begin{aligned}
\mathbb{P}\Big[A_n(f_{j_0})\ -\ A_n(f^*)\geq A_{\mathcal{F}_0}-A^*+\epsilon-\frac{\log M}{n}\Big]&\\
\leq\exp\left(-\frac{n(\epsilon-(\log M)/n)^2}{2c(A_{\mathcal{F}_0}-A^*)^{1/\kappa}+(2K/3)(\epsilon-(\log M)/n)}\right),&
\end{aligned}$$

for any $\epsilon>(\log M)/n$. From now, we take $x=A_{\mathcal{F}_0}-A^*+2\epsilon$, then, for any $(\log M)/n<\epsilon<1$, we have

$$\begin{aligned}
\mathbb{P}\left(\tilde{A}(\hat{w})>A_{\mathcal{F}_0}+2\epsilon\right)&\leq\exp\left(-\frac{n(\epsilon-\log M/n)^2}{2c(A_{\mathcal{F}_0}-A^*)^{1/\kappa}+(2K/3)(\epsilon-(\log M)/n)}\right)\\
&+\ M\left(1+\frac{32c(A_{\mathcal{F}_0}-A^*+2\epsilon)^{1/\kappa}}{n\epsilon^2}\right)\exp\left(-\frac{n\epsilon^2}{32c(A_{\mathcal{F}_0}-A^*+2\epsilon)^{1/\kappa}}\right)\\
&+\ M\left(1+\frac{32}{3n\epsilon}\right)\exp\left(-\frac{3n\epsilon}{32}\right).
\end{aligned}$$

If $\hat{w}$ denotes $e_{\hat{\jmath}}$ then, $\tilde{A}(\hat{w})=\tilde{A}(e_{\hat{\jmath}})=A(\tilde{f}^{(ERM)})$. If $\hat{w}$ denotes the vector of exponential weights w and if $f\longmapsto Q(z,f)$ is convex for π-almost $z\in\mathcal{Z}$, then, $\tilde{A}(\hat{w})=\tilde{A}(w)\geq A(\tilde{f}_n^{(AEW)})$. If $f\longmapsto Q(z,f)$ is assumed to be convex for π-almost $z\in\mathcal{Z}$ then, let $\tilde{f}_n$ denote either the ERM procedure or the AEW procedure, otherwise, let $\tilde{f}_n$ denote the ERM procedure $\tilde{f}_n^{(ERM)}$. We have for any $2(\log M)/n<u<1$,

$$\mathbb{E}[A(\tilde{f}_n)-A_{\mathcal{F}_0}]\leq\mathbb{E}\left[\tilde{A}(\hat{w})-A_{\mathcal{F}_0}\right]\leq 2u+2\int_{u/2}^1\left[T_1(\epsilon)+M(T_2(\epsilon)+T_3(\epsilon))\right]d\epsilon,\quad(19)$$

where

$$T_1(\epsilon)=\exp\left(-\frac{n(\epsilon-(\log M)/n)^2}{2c(A_{\mathcal{F}_0}-A^*)^{1/\kappa}+(2K/3)(\epsilon-(\log M)/n)}\right),$$

$$T_2(\epsilon) = \left(1 + \frac{16c(A_{\mathcal{F}_0} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\right) \exp\left(-\frac{n\epsilon^2}{16c(A_{\mathcal{F}_0} - A^* + 2\epsilon)^{1/\kappa}}\right)$$

and

$$T_3(\epsilon) = \left(1 + \frac{8K}{3n\epsilon}\right) \exp\left(-\frac{3n\epsilon}{8K}\right).$$

We recall that β_1 is defined in (7). Consider separately the following cases $(C1)$ and $(C2)$.

$(C1)$ The case $A_{\mathcal{F}_0} - A^* \geq ((\log M)/(\beta_1 n))^{\kappa/(2\kappa-1)}$.

Denote by $\mu(M)$ the unique solution of $\mu_0 = 3M\exp(-\mu_0)$. Then, clearly $(\log M)/2 \leq \mu(M) \leq \log M$. Take u such that

$$(n\beta_1 u^2)/(A_{\mathcal{F}_0} - A^*)^{1/\kappa} = \mu(M).$$

Using the definition of case (1) and of $\mu(M)$ we get $u \leq A_{\mathcal{F}_0} - A^*$. Moreover, $u \geq 4\log M/n$, then

$$\begin{aligned}\int_{u/2}^1 T_1(\epsilon)d\epsilon &\leq \int_{u/2}^{(A_{\mathcal{F}_0}-A^*)/2} \exp\left(-\frac{n(\epsilon/2)^2}{(2c+K/6)(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right) d\epsilon \\ &\quad + \int_{(A_{\mathcal{F}_0}-A^*)/2}^1 \exp\left(-\frac{n(\epsilon/2)^2}{(4c+K/3)\epsilon^{1/\kappa}}\right) d\epsilon.\end{aligned}$$

Using Lemma 2 and the inequality $u \leq A_{\mathcal{F}_0} - A^*$, we obtain

$$\int_{u/2}^1 T_1(\epsilon)d\epsilon \leq \frac{8(4c+K/3)(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{nu} \exp\left(-\frac{nu^2}{8(4c+K/3)(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right). \tag{20}$$

We have $16c(A_{\mathcal{F}_0} - A^* + 2u) \leq nu^2$ thus, using Lemma 2, we get

$$\begin{aligned}\int_{u/2}^1 T_2(\epsilon)d\epsilon &\leq 2\int_{u/2}^{(A_{\mathcal{F}_0}-A^*)/2} \exp\left(-\frac{n\epsilon^2}{64c(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right) d\epsilon \\ &\quad + 2\int_{(A_{\mathcal{F}_0}-A^*)/2}^1 \exp\left(-\frac{n\epsilon^{2-1/\kappa}}{128c}\right) d\epsilon \\ &\leq \frac{2148c(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{nu} \exp\left(-\frac{nu^2}{2148c(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right). \end{aligned} \tag{21}$$

We have $16(3n)^{-1} \leq u \leq A_{\mathcal{F}_0} - A^*$, thus,

$$\int_{u/2}^1 T_3(\epsilon)d\epsilon \leq \frac{16K(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{3nu} \exp\left(-\frac{3nu^2}{16K(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right). \tag{22}$$

From (20), (21), (22) and (19) we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 2u + 6M\frac{(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}{n\beta_1 u}\exp\left(-\frac{n\beta_1 u^2}{(A_{\mathcal{F}_0} - A^*)^{1/\kappa}}\right).$$

The definition of u leads to $\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 4\sqrt{\frac{(A_{\mathcal{F}_0} - A^*)^{1/\kappa}\log M}{n\beta_1}}$.

$(C2)$The case $A_{\mathcal{F}_0} - A^* \leq ((\log M)/(\beta_1 n))^{\kappa/(2\kappa-1)}$.

We now choose u such that $n\beta_2 u^{(2\kappa-1)/\kappa} = \mu(M)$, where $\mu(M)$ denotes the unique solution of $\mu_0 = 3M\exp(-\mu_0)$ and β_2 is defined in (8). Using the definition of case (2) and of $\mu(M)$ we get $u \geq A_{\mathcal{F}_0} - A^*$ (since $\beta_1 \geq 2\beta_2$). Using the fact that $u > 4\log M/n$ and Lemma 2, we have

$$\int_{u/2}^{1} T_1(\epsilon)d\epsilon \leq \frac{2(16c + K/3)}{nu^{1-1/\kappa}}\exp\left(-\frac{3nu^{2-1/\kappa}}{2(16c + K/3)}\right). \tag{23}$$

We have $u \geq (128c/n)^{\kappa/(2\kappa-1)}$ and using Lemma 2, we obtain

$$\int_{u/2}^{1} T_2(\epsilon)d\epsilon \leq \frac{256c}{nu^{1-1/\kappa}}\exp\left(-\frac{nu^{2-1/\kappa}}{256c}\right). \tag{24}$$

Since $u > 16K/(3n)$ we have

$$\int_{u/2}^{1} T_3(\epsilon)d\epsilon \leq \frac{16K}{3nu^{1-1/\kappa}}\exp\left(-\frac{3nu^{2-1/\kappa}}{16K}\right). \tag{25}$$

From (23), (24), (25) and (19) we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 2u + 6M\frac{\exp\left(-n\beta_2 u^{(2\kappa-1)/\kappa}\right)}{n\beta_2 u^{1-1/\kappa}}.$$

The definition of u yields $\mathbb{E}\left[A(\tilde{f}_n) - A_{\mathcal{F}_0}\right] \leq 4\left(\frac{\log M}{n\beta_2}\right)^{\frac{\kappa}{2\kappa-1}}$. This completes the proof.

Lemma 1. *Consider the framework introduced in the beginning of Subsection 2.1. Let $\mathcal{F}_0 = \{f_1, \ldots, f_M\}$ be a finite subset of $\mathcal{F}$. We assume that π satisfies MA($\kappa, c, \mathcal{F}_0$), for some $\kappa \geq 1, c > 0$ and $|Q(Z, f) - Q(Z, f^*)| \leq K$ a.s., for any $f \in \mathcal{F}_0$, where $K \geq 1$ is a constant. We have for any positive numbers t, x and any integer n*

$$\mathbb{P}\left[\max_{f\in\mathcal{F}}\frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x} > t\right]$$
$$\leq M\left(\left(1 + \frac{4cx^{1/\kappa}}{n(tx)^2}\right)\exp\left(-\frac{n(tx)^2}{4cx^{1/\kappa}}\right) + \left(1 + \frac{4K}{3ntx}\right)\exp\left(-\frac{3ntx}{4K}\right)\right).$$

Proof. We use a "peeling device". Let $x > 0$. For any integer j, we consider

$$\mathcal{F}_j = \{f \in \mathcal{F} : jx \leq A(f) - A^* < (j+1)x\}.$$

Define the empirical process

$$Z_x(f) = \frac{A(f) - A_n(f) - (A(f^*) - A_n(f^*))}{A(f) - A^* + x}.$$

Using Bernstein's inequality and margin assumption $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$ to upper bound the variance term, we have

$$\begin{aligned}
\mathbb{P}\left[\max_{f\in\mathcal{F}} Z_x(f) > t\right] &\leq \sum_{j=0}^{+\infty} \mathbb{P}\left[\max_{f\in\mathcal{F}_j} Z_x(f) > t\right] \\
&\leq \sum_{j=0}^{+\infty} \mathbb{P}\Big[\max_{f\in\mathcal{F}_j} A(f) - A_n(f) - (A(f^*) - A_n(f^*)) > t(j+1)x\Big] \\
&\leq M \sum_{j=0}^{+\infty} \exp\Big(-\frac{n[t(j+1)x]^2}{2c((j+1)x)^{1/\kappa} + (2K/3)t(j+1)x}\Big) \\
&\leq M\Big(\sum_{j=0}^{+\infty} \exp\Big(-\frac{n(tx)^2(j+1)^{2-1/\kappa}}{4cx^{1/\kappa}}\Big) + \exp\Big(-(j+1)\frac{3ntx}{4K}\Big)\Big) \\
&\leq M\Big(\exp\left(-\frac{nt^2x^{2-1/\kappa}}{4c}\right) + \exp\left(-\frac{3ntx}{4K}\right)\Big) \\
&\quad + M\int_1^{+\infty} \Big(\exp\left(-\frac{nt^2x^{2-1/\kappa}}{4c}u^{2-1/\kappa}\right) + \exp\left(-\frac{3ntx}{4K}u\right)\Big)du.
\end{aligned}$$

Lemma 2 completes the proof.

Lemma 2. *Let $\alpha \geq 1$ and $a, b > 0$. An integration by part yields*

$$\int_a^{+\infty} \exp\left(-bt^\alpha\right) dt \leq \frac{\exp(-ba^\alpha)}{\alpha b a^{\alpha-1}}$$

Proof of Corollaries 1 and 2. In the bounded regression setup, any probability distribution π on $\mathcal{X} \times [0,1]$ satisfies the margin assumption $\mathrm{MA}(1, 16, \mathcal{F}_1)$, where $\mathcal{F}_1$ is the set of all measurable functions from $\mathcal{X}$ to $[0,1]$. In density estimation with the integrated squared risk, any probability measure π on $(\mathcal{Z}, \mathcal{T})$, absolutely continuous w.r.t. the measure μ with one version of its density a.s. bounded by a constant

$B \geq 1$, 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 that for any $\epsilon > 0$,

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

and in both cases $f \longmapsto Q(z, f)$ is convex for any $z \in \mathcal{Z}$.

Proof of Theorem 3. We apply Theorem 2, with $\epsilon = 1$, to the multi-thresholding estimator $\hat{f}_n$ defined in (15). Since the density function f^* to estimate takes its values in $[0, B]$, $Card(\Lambda_n) = \log n$ and $m \geq n/2$, we have, conditionally to the first subsample D_m,

$$\begin{aligned}
&\mathbb{E}[\|f^* - \hat{f}_n\|^2_{L^2([0,1])} \ |D_m] \\
&\leq \ 2 \min_{u \in \Lambda_n} (||f^* - h_{0,B}(\hat{f}_{v_u}(D_m, .))||^2_{L^2([0,1])}) + \frac{4(\log n)\log(\log n)}{\beta_2 n} \\
&\leq \ 2 \min_{u \in \Lambda_n} (||f^* - \hat{f}_{v_u}(D_m, .)||^2_{L^2([0,1])}) + \frac{4(\log n)\log(\log n)}{\beta_2 n},
\end{aligned}$$

where $h_{0,B}$ is the projection function introduced in (14) and β_2 is given in (8). Now, for any $s > 0$, let us consider j_s an integer in Λ_n such that $n^{1/(1+2s)} \leq 2^{j_s} < 2n^{1/(1+2s)}$. Since the estimators $\hat{\alpha}_{j,k}$ and $\hat{\beta}_{j,k}$ defined by (16) satisfy the inequalities (12) and (13), Theorem 2 implies that, for any $p \in [1, \infty]$, $s \in (1/p, N]$, $q \in [1, \infty]$ and n large enough, we have

$$\begin{aligned}
&\sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[\|\tilde{f} - f^*\|^2_{L^2([0,1])}] = \sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[\mathbb{E}[\|\tilde{f} - f^*\|^2_{L^2([0,1])} \ |D_m]] \\
&\leq \ 2 \sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[\min_{u \in \Lambda_n} (||f^* - \hat{f}_{v_u}(D_m, .)||^2_{L^2([0,1])}] + \frac{4(\log n)\log(\log n)}{\beta_2 n} \\
&\leq \ 2 \sup_{f^* \in B^s_{p,q}(L)} \mathbb{E}[||f^* - \hat{f}_{v_{j_s}}(D_m, .)||^2_{L^2([0,1])}] + \frac{4(\log n)\log(\log n)}{\beta_2 n} \\
&\leq \ C n^{-2s/(1+2s)}.
\end{aligned}$$

This completes the proof of Theorem 3.

Proof of Theorem 4. The proof of Theorem 4 is similar to the proof of Theorem 3. We only need to prove that, for any $j \in \{\tau, ..., j_1\}$ and $k \in \{0, ..., 2^j - 1\}$, the estimators $\hat{\alpha}_{j,k}$ and $\hat{\beta}_{j,k}$ defined by (17) satisfy the inequalities (12) and (13). First

of all, let us notice that the random variables $Y_1\psi_{j,k}(X_1), ..., Y_n\psi_{j,k}(X_n)$ are i.i.d and that there $m-$th moment, for $m \geq 2$, satisfies

$$\mathbb{E}(|\psi_{j,k}(X_1)|^m) \leq \|\psi\|_\infty^{m-2} 2^{j(m/2-1)} \mathbb{E}(|\psi_{j,k}(X_1)|^2) = \|\psi\|_\infty^{m-2} 2^{j(m/2-1)}.$$

For the first inequality (cf. inequality (12)), Rosenthal's inequality (see [29, p.241]) yields, for any $j \in \{\tau, ..., j_1\}$,

$$\begin{aligned} \mathbb{E}(|\hat{\beta}_{j,k} - \beta_{j,k}|^4) &\leq C(n^{-3}\mathbb{E}(|Y_1\psi_{j,k}(X_1)|^4) + n^{-2}[\mathbb{E}(|Y_1\psi_{j,k}(X_1)|^2)]^2) \\ &\leq C\|Y\|_\infty^4 \|\psi\|_\infty^4 (n^{-3}2^{j_1} + n^{-2}) \leq Cn^{-2}. \end{aligned}$$

For second inequality (cf. inequality (13)), Bernstein's inequality yields

$$\mathbb{P}\Big(2\sqrt{n}|\hat{\beta}_{j,k} - \beta_{j,k}| \geq \rho\sqrt{a}\Big) \leq 2\exp\Big(-\frac{\rho^2 a}{8\sigma^2 + (8/3)M\rho\sqrt{a}/(2\sqrt{n})}\Big),$$

where $a \in \{\tau, ..., j_1\}$, $\rho \in (0, \infty)$,

$$\begin{aligned} M &= \|Y\psi_{j,k}(X) - \beta_{j,k}\|_\infty \leq 2^{j/2}\|Y\|_\infty\|\psi\|_\infty + \|f^*\|^2_{L^2([0,1])} \\ &\leq 2^{j_1/2}(\|\psi\|_\infty + 1) \leq 2^{1/2}(n/\log n)^{1/2}(\|\psi\|_\infty + 1), \end{aligned}$$

and

$$\sigma^2 = \mathbb{E}(|Y_1\psi_{j,k}(X_1) - \beta_{j,k}|^2) \leq \mathbb{E}(|Y_1\psi_{j,k}(X_1)|^2) \leq \|Y\|_\infty^2 \leq 1.$$

Since $a \leq \log n$, we complete the proof by seeing that for ρ large enough, we have

$$\exp\Big(-\frac{\rho^2 a}{8\sigma^2 + (8/3)M\rho\sqrt{a}/(2\sqrt{n})}\Big) \leq 2^{-4a}.$$

References

[1] F. Abramovich and Y. Benjamini. Adaptive thresholding of wavelet coefficients. *Computat. Stat. Data Anal.*, 22:351–361, 1996.

[2] F. Abramovich, Y. Benjamini, D.L. Donoho, and I.M. Johnstone. Adapting to unknown sparsity by controlling the false discovery rate. *Ann. Statist.*, 34(2):584–653, 2006.

[3] F. Abramovich, T. Sapatinas, and B.W. Silverman. Wavelet thresholding via a Bayesian approach. *J. R. Statist. Soc. B*, 60:725–749, 1998.

[4] A. Antoniadis and J. Bigot. Wavelet Estimators in Nonparametric Regression: A Comparative Simulation Study. *J. Statist. Software*, 6(3):1–83, 2001.

[5] A. Barron and G. Leung. Information theory and mixing least-square regressions. *IEEE Transactions on Information Theory*, 52 (8):3396–3410, 2006.

[6] L. Birgé. Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Available at http://www.proba.jussieu.fr/mathdoc/textes/PMA-862.pdf, 2005.

[7] L Birgé and P. Massart. Gaussian model selection. *J. Eur. Math. Soc.*, 3:203–268, 2001.

[8] F. Bunea and A. Nobel. Online prediction algorithms for aggregation of arbitrary estimators of a conditional mean. Submitted to IEEE Transactions in Information Theory, 2005.

[9] T. Cai. On adaptivity of Blockshrink wavelet estimator over Besov spaces. *Technical Report, 97-05, Department of Statistics, Purdue University*, 1997.

[10] T. Cai. Adaptive wavelet estimation: a block thresholding and oracle inequality approach. *Ann. Stat.*, 27:898–924, 1999.

[11] T. Cai and E. Chicken. Block thresholding for density estimation: local and global adaptivity. *Journal of Multivariate Analysis*, 95:76–106, 2005.

[12] T. Cai and B.W. Silverman. Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya*, (63):127–148, 2001.

[13] O. Catoni. *Statistical Learning Theory and Stochastic Optimization.* Ecole d'été de Probabilités de Saint-Flour 2001, Lecture Notes in Mathematics. Springer, N.Y., 2001.

[14] L. Cavalier and A. Tsybakov. Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Meth. Statist.*, 10(3):247–282, 2001.

[15] N. Cesa-Bianchi and G. Lugosi. *Prediction, Learning, and Games*. Cambridge University Press, New York, 2006.

[16] E. Chicken. Nonparametric regression on random processes and design. *Florida State University Department of Statistics, Technical Report*, 2003.

[17] H. A. Chipman, E. Kolaczyk, and R. McCulloch. Adaptive Bayesian wavelet shrinkage. *J. Am. Statist. Ass.*, 92:1413–1421, 1997.

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

[19] B. Delyon and A. Juditsky. On minimax wavelet estimators. *Applied Computational Harmonic Analysis*, 3:215–228, 1996.

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

[21] D.L. Donoho and I.M. Johnstone. Adaptating to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995.

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

[23] D.L. Donoho, I.M. Johnstone, G. Kerkyacharian, and D. Picard. Density estimation by wavelet thresholding. *Ann. Statist.*, 24(2):508–539, 1996.

[24] S. Efromovich. Quasi-linear wavelet estimation. *J. Amer. Statist. Assoc.*, 94(445):189–204, 1999.

[25] S. Efromovich. Sharp linear and block shrinkage wavelet estimation. *Statist. Probab. Lett.*, 4(445):323–329, 2000.

[26] H. Gao. Wavelet shrinkage denoising using the nonnegative garrote. *J. Comput. Graph. Statist.*, 7:469–488, 1998.

[27] P. Hall, G. Kerkyacharian, and D. Picard. Block thresholding rules for curve estimation using kernel and wavelet methods. *Ann. Statist.*, 26:942–962, 1998.

[28] P. Hall, G. Kerkyacharian, and D. Picard. On the minimax optimality of block thresholded wavelet estimators. *Statist. Sinica*, 9(1):33–49, 1999.

[29] W. Härdle, G. Kerkyacharian, D. Picard, and A. Tsybakov. *Wavelet, Approximation and Statistical Applications*, volume 129 of *Lectures Notes in Statistics*. Springer Verlag, New York, 1998.

[30] D.R.M. Herrick, G.P. Nason, and B.W. Silverman. Some new methods for wavelet density estimation. *Sankhya Series A*, 63:394–411, 2001.

[31] M. Jansen. *Noise reduction by wavelet thresholding*, volume 161. Springer–Verlag, New York, lecture notes in statistics edition, 2001.

[32] I.M Johnstone and B.W. Silverman. Empirical bayes selection of wavelet thresholds. *Ann. Statist.*, 33(4):1700–1752, 1998.

[33] A. Juditsky. Wavelet estimators: adapting to unknown smoothness. *Math. Methods of Statistics*, (1):1–20, 1997.

[34] A. Juditsky and A. Nemirovski. Functional aggregation for nonparametric estimation. *Ann. Statist.*, 28(3):681–712, 2000.

[35] A.B. Juditsky, A.V. Nazin, A.B. Tsybakov, and N. Vayatis. Recursive aggregation of estimators via the mirror descent algorithm with averaging. *Problems of Information Transmission*, 41:368 – 384, 2005.

[36] A.B. Juditsky, Ph. Rigollet, and A.B. Tsybakov. Learning by mirror averaging. Preprint n.1034, Laboratoire de Probabilités et Modèle aléatoires, Universités Paris 6 and Paris 7 (available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2005), 2006.

[37] G. Kerkyacharian, D. Picard, and K. Tribouley. Lp adaptive density estimation. *Bernoulli*, 2:229–247, 1996.

[38] V. Koltchinskii. Local Rademacher Complexities and Oracle Inequalities in Risk Minimization. *Ann. Statist.*, 34(6):1–50, December 2006. 2004 IMS Medallion Lecture.

[39] G. Lecué. Optimal rates of aggregation in classification. Submitted to Bernoulli, 2005.

[40] G. Lecué. Simultaneous adaptation to the margin and to complexity in classification. To appear in Ann. Statist. Available at http://hal.ccsd.cnrs.fr/ccsd-00009241/en/, 2005.

[41] G. Lecué. Optimal oracle inequality for aggregation of classifiers under low noise condition. *In Proceeding of the 19th Annual Conference on Learning Theory, COLT 2006*, 32(4):364–378, 2006.

[42] G. Lecué. Suboptimality of Penalized Empirical Risk Minimization in Classification. Submitted, 2006.

[43] E. Mammen and A.B. Tsybakov. Smooth discrimination analysis. *Ann. Statist.*, 27:1808–1829, 1999.

[44] Y. Meyer. *Ondelettes et Opérateurs*. Hermann, Paris, 1990.

[45] G.P. Nason. *Choice of the Threshold Parameter in Wavelet Function Estimation*, volume 103. 1995.

[46] A. Nemirovski. *Topics in Non-parametric Statistics*, volume 1738 of *Ecole d'été de Probabilités de Saint-Flour 1998, Lecture Notes in Mathematics*. Springer, N.Y., 2000.

[47] D. Picard and K. Tribouley. Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.*, 28(1):298–335, 2000.

[48] I. Steinwart and C. Scovel. Fast Rates for Support Vector Machines using Gaussian Kernels. *Ann. Statist.*, 35(2), April 2007.

[49] A. B. Tsybakov. Optimal rates of aggregation. *Computational Learning Theory and Kernel Machines. B.Schölkopf and M.Warmuth, eds. Lecture Notes in Artificial Intelligence*, 2777:303–313, 2003. Springer, Heidelberg.

[50] A.B. Tsybakov. Optimal aggregation of classifiers in statistical learning. *Ann. Statist.*, 32(1):135–166, 2004.

[51] V. Vapnik. *Statistical Learning Theory.* John Wiley and Sons, New York, 1998.

[52] V.G. Vovk. Aggregating Strategies. *In: Proceedings of the 3rd Annual Workshop on Computational Learning Theory, COLT1990, CA: Morgan Kaufmann*, pages 371–386, 1990.

[53] N. Weyrich and G.T. Warhola. Wavelet shrinkage and generalized cross-validation for image denoising. *IEEE Trans. Im. Proc.*, 7:82–90, 1998.

[54] Y. Yang. Mixing strategies for density estimation. *Ann. Statist.*, 28(1):75–87, 2000.

[55] C.H. Zhang. General empirical Bayes wavelet methods and exactly adaptive minimax estimation. *Ann. Statist.*, 33:54–100, 2005.

[56] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Ann. Statist.*, 32(1):56–85, 2004.