Adapting to Unknown Smoothness by Aggregation of Thresholded Wavelet Estimators

Christophe Chesneau and Guillaume Lecué

University Pierre et Marie Curie, Paris VI

Abstract: We consider 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 it is optimal in the minimax sense over Besov balls under the L^2 risk, without any extra logarithm term.

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

AMS 1991 Subject Classification Primary: 62G05; Secondary: 62G07, 62G08, 68T05, 68Q32

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, it 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 inequality satisfied by aggregation methods. Such methods use an exponential weighting aggregation scheme, which has been studied, among others, by Augustin et al. (1997), Yang (2000), Catoni (2001), Leung and Barron (2006), Bunea and Nobel (2005) and Lecué (2005a,b,2006).

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 for the L^2 risk. All the proofs are postponed in the last section.

2. Oracle inequalities

2.1. Framework. Let $(\mathcal{Z},\mathcal{T})$ 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 $Z_1,\dots,Z_n$ 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 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 will 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 of the cases f^* will be equal to our aim $F(\pi)$. 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. It is denoted by

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

In order to illustrate this general statistical framework with a concrete problem, let us focus our attention on the 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 $\mathcal{F}$ 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 $f \in \mathcal{F}$,

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

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 improve the quality of estimation in our framework. This assumption has been first introduced by Mammen and Tsybakov (1999), for the problem of discriminant analysis, and Tsybakov (2004), for the classification problem. With this assumption, 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 the 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. Let's 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), 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 procedure is defined by

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

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

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

2.3. Oracle Inequalities. In this Subsection we state an exact oracle inequality satisfied by the AEW procedure in the general framework of the beginning of Section 2. From this exact oracle inequality, we deduce an oracle inequality in the density estimation framework. Now, let us introduce a quantity which is going to be our residual term in the exact oracle inequality. We define the quantity $\gamma = \gamma(n, M, \kappa, \mathcal{F}_0, \pi, Q)$ by

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

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

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

and

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

where the constant $c > 0$ appears in the margin assumption MA$(\kappa, c, \mathcal{F}_0)$ and K is considered in the following theorem.

Theorem 2.1 *Let us consider the general framework introduced in the beginning of Section 2. Let $M \geq 2$ be an integer. Let $\mathcal{F}_0$ 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).

Now, we give a corollary of Theorem 2.1 in the density estimation framework.

Corollary 2.2 *Let us consider the density estimation framework. Assume that the 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*

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

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

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 the purposes of this paper, 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 $\phi^{per}_{j,k}(x) = \sum_{l\in\mathbb{Z}} \phi_{j,k}(x-l), \psi^{per}_{j,k}(x) = \sum_{l\in\mathbb{Z}} \psi_{j,k}(x-l)$, 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 $\zeta = \{\phi^{per}_{\tau,k}, k = 0,...,2^\tau - 1;\ \psi^{per}_{j,k},\quad j = \tau,...,\infty,\ k = 0,...,2^j-1\}$ constitutes an orthonormal basis of $L^2([0,1])$. In what follows, the superscript "*per*" will be suppressed from the notations for convenience. 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).

Now, let us define the main function spaces of the study. Let $L \in (0,\infty)$, $s \in (0,\infty)$, $p \in [1,\infty)$ 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 $B^s_{p,q}(L)$ 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

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

j_1 is an integer satisfying $(n/\log n) \leq 2^{j_1} < 2(n/\log n)$, $\lambda = (\lambda_\tau, ...\lambda_{j_1})$ 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

$$|\Upsilon_u(x) - y|^2 \leq C_1\left(|\min(y, C_2 u)|^2 + |x-y|^2 \mathbb{1}_{\{|x-y|\geq 2^{-1}u\}}\right). \tag{3.3}$$

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

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

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. 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

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

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.

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 following 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 satisfying (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 following 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 $\Lambda_n = \{0, ..., \lceil \log n \rceil\}$ and, for any $u \in \Lambda_n$,

$$w^{(l)}(\hat{f}^t_{m,u}) = \exp\Big(-lA^{(l)}(\hat{f}^t_{m,u})\Big) / \sum_{\gamma \in \Lambda_n} \exp\Big(-lA^{(l)}(\hat{f}^t_{m,\gamma})\Big),$$

where $A^{(l)}(f) = (1/l)\sum_{i=m+1}^{n} Q(Z_i, f)$ is the empirical risk constructed from the l last observations, for any function f and for the choice of a loss function Q defined in (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. Performances of the multi-thresholding estimator

4.1 Main result. Theorem 4.3 below investigates the minimax performances of the multi-thresholding estimator defined in (3.6) under the L^2 risk over Besov balls in the density estimation framework.

Theorem 4.3 *Let us consider the problem of estimating a density function f^* 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,*

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

Let us 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 over Besov balls under the 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 which 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 BlockShrink method and BlockJS method investigated by Cai and Chicken (2005).

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	$n^{-2s/(2s+1)}$	$n^{-2s/(2s+1)}$

As we notice 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. This kind of estimators are not easy to compute compare 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: preliminaries. First of all, let us recall the notations of the general framework introduced in the beginning of Section 2. 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.l.o.g., that it is achieved by an element f^* in $\mathcal{F}$ and, for any $f \in \mathcal{F}$, the empirical risk $A_n(f) = (1/n) \sum_{i=1}^n Q(Z_i, f)$. Now, let us consider the convex set $\mathcal{C}$ defined by

$$\mathcal{C} = \Big\{ (\theta_1, \ldots, \theta_M) : \theta_j \geq 0, \forall j = 1, \ldots, M, \text{ and } \sum_{j=1}^M \theta_j = 1 \Big\}. \tag{5.1}$$

For any $\theta \in \mathcal{C}$, we define the functions $\tilde{A}(\theta)$ and $\tilde{A}_n(\theta)$ by

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

The first function is the linear version of the risk A. The second is the empirical version of this risk.

We are now in position to explain the form of the exponential weights described by (2.4). By virtue of the Lagrange method of optimization, we find that the exponential weights $w = (w^{(n)}(f_j))_{1\leq j\leq M}$ are the unique solution of the minimization problem

$$\min_{(\theta_1,\dots,\theta_M)\in\mathcal{C}}\Big\{\tilde{A}_n(\theta) + (1/n)\sum_{j=1}^{M}\theta_j\log\theta_j\Big\}, \tag{5.2}$$

where we use the convention $0\log 0 = 0$. Take $\hat{\jmath}\in\{1,\dots,M\}$ such that $A_n(f_{\hat{\jmath}}) = \min_{j=1,\dots,M} A_n(f_j)$. If e_j denotes the vector in $\mathcal{C}$ with 1 for j-th coordinate and 0 elsewhere, then, by (5.2), the vector of exponential weights w satisfies

$$\tilde{A}_n(w) + (1/n)\sum_{j=1}^{M} w^{(n)}(f_j)\log w^{(n)}(f_j) \leq \tilde{A}_n(e_{\hat{\jmath}}).$$

Using the fact that $\sum_{j=1}^{M} w^{(n)}(f_j)\log(Mw^{(n)}(f_j)) \geq 0$ (because this is the Kullback-Leibler divergence between the weights w and the uniform weights), we obtain

$$\tilde{A}_n(w) \leq \tilde{A}_n(e_{\hat{\jmath}}) + \log M/n. \tag{5.3}$$

Now, observe that a linear function achieves its maximum over a convex polygon at one of the vertices of the polygon. Thus, for $j_0\in\{1,\dots,M\}$ such that $\tilde{A}(e_{j_0}) = \min_{j=1,\dots,M}\tilde{A}(e_j)$ $(= \min_{j=1,\dots,M} A(f_j))$, 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 $\mathcal{C}$. We define $\hat{w}$ by either:

$$\hat{w} = w \qquad \text{or} \qquad \hat{w} = e_{\hat{\jmath}}. \tag{5.4}$$

According to (5.3), we have

$$\tilde{A}_n(\hat{w}) \leq \min_{j=1,\dots,M}\tilde{A}_n(e_j) + \log M/n \leq \tilde{A}_n(e_{j_0}) + \log M/n. \tag{5.5}$$

This inequality, justified by the form of our weights, will be at the heart of the proof. Now, let us set two auxiliary lemmas.

Lemma 5.4 *Consider the framework introduced in the beginning of Section 2. Let $\mathcal{F}_0 = \{f_1,\dots,f_M\}$ 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, 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, for any positive numbers t, x and any integer n, we have:

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

The proof of Lemma 5.4 is postponed at the end of the proof of Theorem 2.1.

Lemma 5.5 *Let $\alpha \geq 1$ and $x, y > 0$. An integration by part yields*

$$\int_x^{+\infty} \exp\left(-yt^\alpha\right) dt \leq \exp(-yx^\alpha)/(\alpha y x^{\alpha-1}).$$

Proof of Theorem 2.1: technical details. Denote by $\tilde{A}_{\mathcal{C}}$ the minimum $\min_{\theta\in\mathcal{C}} \tilde{A}(\theta)$ where $\mathcal{C}$ is the set defined by (5.1). Using the following elementary inequality: for any $u \in \mathbb{R}$ and random variable $W \in]-\infty, K]$, we have $\mathbb{E}(W) = \mathbb{E}(W\mathbb{1}_{\{W<u\}} + W\mathbb{1}_{\{W\geq u\}}) \leq u + \int_0^K \mathbb{P}(W\mathbb{1}_{\{W\geq u\}} \geq \epsilon)d\epsilon = 2u + 2\int_{u/2}^{K/2} \mathbb{P}(W \geq 2\epsilon)d\epsilon$, we obtain:

$$\mathbb{E}[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}] \leq \mathbb{E}\left[\tilde{A}(\hat{w}) - \tilde{A}_{\mathcal{C}}\right] \leq 2u + 2\int_{u/2}^{K/2} \mathbb{P}\left[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right] d\epsilon, \quad (5.6)$$

where $\hat{w}$ is defined by (5.4).

Now, let us investigate the upper bound of the term $\mathbb{P}\left[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right]$.

Let us consider $\mathcal{D}$, the subset of $\mathcal{C}$ defined by

$$\mathcal{D} = \left\{\theta \in \mathcal{C} : \tilde{A}(\theta) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right\}.$$

If $\hat{w} \in \mathcal{D}$ then the inequality (5.5) implies the existence of $\theta \in \mathcal{D}$ 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, for any $\epsilon > 0$, we have

$$\begin{aligned}\mathbb{P}\left[\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right] &\leq \mathbb{P}\left[\inf_{\theta\in\mathcal{D}} \tilde{A}_n(\theta) - A_n(f^*) \leq \tilde{A}_n(e_{j_0}) - A_n(f^*) + \log M/n\right] \\ &\leq V_1 + V_2,\end{aligned}$$

where

$$V_1 = \mathbb{P}\left[\inf_{\theta\in\mathcal{D}} \tilde{A}_n(\theta) - A_n(f^*) < \tilde{A}_{\mathcal{C}} - A^* + \epsilon\right]$$

and

$$V_2 = \mathbb{P}\left[\tilde{A}_n(e_{j_0}) - A_n(f^*) \geq \tilde{A}_{\mathcal{C}} - A^* + \epsilon - \log M/n\right].$$

Let us investigate the upper bounds for V_1 and V_2, in turn.

The upper bound for V_1. We recall that $\tilde{A}_{\mathcal{C}}$ denotes the minimum $\min_{\theta\in\mathcal{C}} \tilde{A}(\theta)$. Assume that, for any $x > 0$, we have

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

Since, for any $\theta \in \mathcal{D}$, $\tilde{A}(\theta) - A^* \geq \tilde{A}_{\mathcal{C}} - A^* + 2\epsilon$, we obtain

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

Hence, for any $x > 0$, we can bound V_1 by

$$V_1 \leq \mathbb{P}\left[\sup_{\theta\in\mathcal{D}} \frac{\tilde{A}(\theta) - A^* - [\tilde{A}_n(\theta) - A_n(f^*)]}{\tilde{A}(\theta) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right]. \tag{5.7}$$

If, for any $x > 0$, 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

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

Let us notice that, for any numbers $a_1, \ldots, a_M$ and positive numbers $b_1, \ldots, b_M$, we have $\sum_{j=1}^M a_j / \sum_{j=1}^M b_j \leq \max_{j=1,\ldots,M}(a_j/b_j)$. It follows that

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

where $\tilde{A}_{\mathcal{C}} = \min_{j=1,\dots,M} A(f_j)$ (which is equal to the $\tilde{A}_{\mathcal{C}}$ previously defined).

Now, we use the relative concentration inequality of Lemma 5.4 to obtain

$$\begin{aligned}
&\mathbb{P}\left[\max_{j=1,\dots,M} \frac{A(f_j) - A^* - (A_n(f_j) - A_n(f^*))}{A(f_j) - A^* + x} > \frac{\epsilon}{\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x}\right] \\
&\leq \; M\left(1 + \frac{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}{n(\epsilon x)^2}\right)\exp\left(-\frac{n(\epsilon x)^2}{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}\right) \\
&\quad + M\left(1 + \frac{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}{3n\epsilon x}\right)\exp\left(-\frac{3n\epsilon x}{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}\right). \qquad (5.8)
\end{aligned}$$

Putting (5.7) and (5.8) together, for any $x > 0$, we obtain:

$$\begin{aligned}
V_1 \;\leq\; & M\left(1 + \frac{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}{n(\epsilon x)^2}\right)\exp\left(-\frac{n(\epsilon x)^2}{4c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)^2 x^{1/\kappa}}\right) \\
& + M\left(1 + \frac{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}{3n\epsilon x}\right)\exp\left(-\frac{3n\epsilon x}{4K(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon + x)}\right). \qquad (5.9)
\end{aligned}$$

The upper bound for V_2. Using the margin assumption $\mathrm{MA}(\kappa, c, \mathcal{F}_0)$ to upper bound the variance term and applying Bernstein's inequality (cf. Massart (2006)), for any $\epsilon > \log M/n$, we get

$$V_2 \leq \exp\left(-\frac{n(\epsilon - (\log M)/n)^2}{2c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} + (2K/3)(\epsilon - \log M/n)}\right), \qquad (5.10)$$

Combining the obtained upper bounds of V_1 with $x = \tilde{A}_{\mathcal{C}} - A^* + 2\epsilon$ and V_2, then, for any $\log M/n < \epsilon < K/2$, we have

$$\begin{aligned}
&\mathbb{P}\left(\tilde{A}(\hat{w}) > \tilde{A}_{\mathcal{C}} + 2\epsilon\right) \leq V_1 + V_2 \\
&\leq \exp\left(-\frac{n(\epsilon - \log M/n)^2}{2c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} + (2K/3)(\epsilon - \log M/n)}\right) \\
&\quad + M\left(1 + \frac{16c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}{n\epsilon^2}\right)\exp\left(-\frac{n\epsilon^2}{16c(\tilde{A}_{\mathcal{C}} - A^* + 2\epsilon)^{1/\kappa}}\right) \\
&\quad + M\left(1 + \frac{8K}{3n\epsilon}\right)\exp\left(-\frac{3n\epsilon}{8K}\right).
\end{aligned}$$

It follows from (5.6) that, for any $2\log M/n < u < K/2$, we have

$$\mathbb{E}[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}] \leq 2u + 2\int_{u/2}^{K/2} [T_1(\epsilon) + M(T_2(\epsilon) + T_3(\epsilon))]\, d\epsilon, \qquad (5.11)$$

where the quantities $T_1(\epsilon)$, $T_2(\epsilon)$ and $T_3(\epsilon)$ are defined by

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

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

and

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

Now, let us investigate the upper bounds of $\int_{u/2}^1 T_1(\epsilon)d\epsilon$, $\int_{u/2}^1 T_2(\epsilon)d\epsilon$ and $\int_{u/2}^1 T_3(\epsilon)d\epsilon$, in turn. We distinguish two cases: the case where $\tilde{A}_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$ and the case where $\tilde{A}_{\mathcal{C}} - A^* < (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. Let us recall that β_1 is defined in (2.6).

- *The case* $\tilde{A}_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. Denote by $\mu(M)$ the unique solution of the equation $\mu_0 - 3M\exp(-\mu_0) = 0$. Then, clearly $(\log M)/2 \leq \mu(M) \leq \log M$. Take u such that $(n\beta_1 u^2)/(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} = \mu(M)$. Using the fact that $\tilde{A}_{\mathcal{C}} - A^* \geq (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$ and the definition $\mu(M)$, we get $u \leq \tilde{A}_{\mathcal{C}} - A^*$. Moreover, since $u \geq 4\log M/n$, we have

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

Using Lemma 5.5 and the inequality $u \leq \tilde{A}_{\mathcal{C}} - A^*$, we obtain

$$\int_{u/2}^{K/2} T_1(\epsilon)d\epsilon \leq \frac{8(4c+K/3)(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{nu}\exp\left(-\frac{nu^2}{8(4c+K/3)(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \tag{5.12}$$

Since $16c(\tilde{A}_{\mathcal{C}} - A^* + 2u) \leq nu^2$, Lemma 5.5 yields

$$\begin{aligned}\int_{u/2}^{K/2} T_2(\epsilon)d\epsilon \;&\leq\; 2\int_{u/2}^{(\tilde{A}_{\mathcal{C}}-A^*)/2} \exp\left(-\frac{n\epsilon^2}{64c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right)d\epsilon \\ &\quad +2\int_{(\tilde{A}_{\mathcal{C}}-A^*)/2}^{K/2} \exp\left(-\frac{n\epsilon^{2-1/\kappa}}{128c}\right)d\epsilon \\ &\leq\; \frac{2148c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{nu}\exp\left(-\frac{nu^2}{2148c(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \end{aligned}\tag{5.13}$$

Since $16(3n)^{-1} \le u \le \tilde{A}_{\mathcal{C}} - A^*$, we have

$$\int_{u/2}^{K/2} T_3(\epsilon) d\epsilon \le \frac{16K(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{3nu} \exp\left(-\frac{3nu^2}{16K(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right). \qquad (5.14)$$

From (5.11), (5.12), (5.13), (5.14) and the definition of u (and, a fortiori, $\mu(M)$), we obtain

$$\begin{aligned} \mathbb{E}\left[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}\right] &\le 2u + 6M\frac{(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}{n\beta_1 u} \exp\left(-\frac{n\beta_1 u^2}{(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa}}\right) \\ &= 4u \le 4\sqrt{(\tilde{A}_{\mathcal{C}} - A^*)^{1/\kappa} \log M/(n\beta_1)}. \end{aligned}$$

- *The case* $\tilde{A}_{\mathcal{C}} - A^* < (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$. We now choose u such that $n\beta_2 u^{(2\kappa-1)/\kappa} = \mu(M)$, where $\mu(M)$ denotes the unique solution of the equation $\mu_0 - 3M\exp(-\mu_0) = 0$ and β_2 is defined in (2.7). Using the fact that $\tilde{A}_{\mathcal{C}} - A^* < (\log M/(\beta_1 n))^{\kappa/(2\kappa-1)}$ and the definition of $\mu(M)$, we get $u \ge \tilde{A}_{\mathcal{C}} - A^*$ (since $\beta_1 \ge 2\beta_2$). Using the fact that $u > 4\log M/n$ and Lemma 5.5, we find

$$\int_{u/2}^{K/2} T_1(\epsilon) d\epsilon \le \frac{2(16c + K/3)}{nu^{1-1/\kappa}} \exp\left(-\frac{3nu^{2-1/\kappa}}{2(16c + K/3)}\right). \qquad (5.15)$$

Since $u \ge (128c/n)^{\kappa/(2\kappa-1)}$, Lemma 5.5 yields

$$\int_{u/2}^{K/2} T_2(\epsilon) d\epsilon \le \frac{256c}{nu^{1-1/\kappa}} \exp\left(-\frac{nu^{2-1/\kappa}}{256c}\right). \qquad (5.16)$$

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

$$\int_{u/2}^{K/2} T_3(\epsilon) d\epsilon \le \frac{16K}{3nu^{1-1/\kappa}} \exp\left(-\frac{3nu^{2-1/\kappa}}{16K}\right). \qquad (5.17)$$

Putting (5.11), (5.15), (5.16) and (5.17) together and using the definition of u (and, a fortiori, $\mu(M)$), we obtain

$$\mathbb{E}\left[A(\tilde{f}_n) - \tilde{A}_{\mathcal{C}}\right] \le 2u + 6M\frac{\exp\left(-n\beta_2 u^{(2\kappa-1)/\kappa}\right)}{n\beta_2 u^{1-1/\kappa}} = 4u \le 4(\log M/(n\beta_2))^{\kappa/(2\kappa-1)}.$$

This completes the proof of Theorem 2.1.

Proof of Lemma 5.4. 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 $Z_x(f)$ by

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

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

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

Lemma 5.5 completes the proof.

Proof of Corollary 2.2. In density estimation with the integrated squared risk, any probability measure π on $(\mathcal{Z}, \mathcal{T})$, absolutely continuous 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 $\epsilon > 0$, we have

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

Proof of Theorem 4.3. 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 to the first subsample D_m, we have

$$\begin{aligned}
&\mathbb{E}\left[\|f^* - \hat{f}_n\|_2^2 \;|D_m\right] \\
&\leq \; 2 \min_{u \in \Lambda_n} \left(||f^* - h_B(\hat{f}_{v_u}(D_m, .))||_2^2\right) + 4(\log n)\log(\log n)/(\beta_2 n) \\
&\leq \; 2 \min_{u \in \Lambda_n} \left(||f^* - \hat{f}_{v_u}(D_m, .)||_2^2\right) + 4(\log n)\log(\log n)/(\beta_2 n), \qquad (5.18)
\end{aligned}$$

where h_B is the projection function introduced in (3.5) and β_2 is given in (2.7). Now, for any $s > 0$, let us 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:

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

Therefore, for any $p \in [1, \infty]$, $s \in (1/p, \infty)$, $q \in [1, \infty]$ and n large enough, the previous inequality and (5.18) yield

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

This completes the proof of Theorem 4.3.

Acknowledgment. We would like to thank the referee and the 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(**2**):584–653.

Abramovich, F., Sapatinas, T., and Silverman, B.W., (1998). *Wavelet thresholding via a Bayesian approach. J. R. Statist. Soc. 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, 1997.

Birgé, L., (2006). *Model selection via testing: an alternative to (penalized) maximum likelihood estimators. Ann. Inst. H. Poincar Probab. Statist.* 42(3): 273–325, 2006.

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

Cai, T., and Chicken, E., (2005). *Block thresholding for density estimation: local and global adaptivity. Journal of Multivariate Analysis*, **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, N.Y., 2001.

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. Applied Computational Harmonic Analysis*, **3**:215–228.

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

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., and Picard, D., (1995). *Wavelet shrinkage: Asymptotia ? J. Royal 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(**2**):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.

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

Jansen, M., (2001). Noise reduction by wavelet thresholding, volume 161. Springer–Verlag, New York, lecture notes in statistics edition.

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

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

Lecué, G., (2005). *Lower bounds and aggregation in density estimation. Journal of Machine Learning Research*, 7(Jun): 971–981, 2005.

Lecué, G.,(2005). *Optimal rates of aggregation in classification.* To appear in *Bernoulli.*

Lecué, G., (2005). *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/.

Lecué, G., (2006). *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.

Leung, G., and Barron, A., (2006). *Information theory and mixing least-square regressions. IEEE Transactions on Information Theory*, 52 (**8**):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, N.Y..

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

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

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

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

Yang, Y., (2001). *Minimax rate adaptive estimation over continuous hyperparameters. IEEE Transaction on Information Theory*, 47:2081-2085.