Aggregation by Exponential Weighting and Sharp Oracle Inequalities

Arnak S. Dalalyan and Alexandre B. Tsybakov

University of Paris 6,
4, Place Jussieu, 75252 Paris cedex 05, France

Abstract. In the present paper, we study the problem of aggregation under the squared loss in the model of regression with deterministic design. We obtain sharp oracle inequalities for convex aggregates defined via exponential weights, under general assumptions on the distribution of errors and on the functions to aggregate. We show how these results can be applied to derive a sparsity oracle inequality.

1 Introduction

Consider the regression model

$$Y_i = f(x_i) + \xi_i, \quad i = 1, \dots, n, \tag{1}$$

where $x_1, \dots, x_n$ are given elements of a set $\mathcal{X}$, $f : \mathcal{X} \to \mathbb{R}$ is an unknown function, and ξ_i are i.i.d. zero-mean random variables on a probability space $(\Omega, \mathcal{F}, P)$ where $\Omega \subseteq \mathbb{R}$. The problem is to estimate the function f from the data $D_n = ((x_1, Y_1), \dots, (x_n, Y_n))$.

Let $(\Lambda, \mathcal{A})$ be a probability space and denote by $\mathscr{P}_\Lambda$ the set of all probability measures defined on $(\Lambda, \mathcal{A})$. Assume that we are given a family $\{f_\lambda, \lambda \in \Lambda\}$ of functions $f_\lambda : \mathcal{X} \to \mathbb{R}$ such that the mapping $\lambda \mapsto f_\lambda$ is measurable, $\mathbb{R}$ being equipped with the Borel σ-field. Functions f_λ can be viewed either as weak learners or as some preliminary estimators of f based on a training sample independent of $\mathbf{Y} \triangleq (Y_1, \dots, Y_n)$ and considered as frozen.

We study the problem of aggregation of functions in $\{f_\lambda, \lambda \in \Lambda\}$ under the squared loss. Specifically, we construct an estimator $\hat{f}_n$ based on the data D_n and called *aggregate* such that the expected value of its squared error

$$\|\hat{f}_n - f\|_n^2 \triangleq \frac{1}{n} \sum_{i=1}^{n} \big(\hat{f}_n(x_i) - f(x_i)\big)^2$$

is approximately as small as the oracle value $\inf_{\lambda \in \Lambda} \|f - f_\lambda\|_n^2$.

In this paper we consider aggregates that are mixtures of functions f_λ with exponential weights. For a measure π from $\mathscr{P}_\Lambda$ and for $\beta > 0$ we set

$$\hat{f}_n(x) \triangleq \int_\Lambda \theta_\lambda(\beta, \pi, \mathbf{Y}) f_\lambda(x)\, \pi(d\lambda), \quad x \in \mathcal{X}, \tag{2}$$

N. Bshouty and C. Gentile (Eds.): COLT 2007, LNAI 4539, pp. 97–111, 2007.

with

$$\theta_\lambda(\beta,\pi,\mathbf{Y}) = \frac{\exp\big\{-n\|\mathbf{Y}-f_\lambda\|_n^2/\beta\big\}}{\int_\Lambda \exp\big\{-n\|\mathbf{Y}-f_w\|_n^2/\beta\big\}\pi(dw)} \tag{3}$$

where $\|\mathbf{Y}-f_\lambda\|_n^2 \triangleq \frac{1}{n}\sum_{i=1}^n \big(Y_i - f_\lambda(x_i)\big)^2$ and we assume that π is such that the integral in (2) is finite.

Note that $\hat{f}_n$ depends on two tuning parameters: the prior measure π and the "temperature" parameter β. They have to be selected in a suitable way.

Using the Bayesian terminology, $\pi(\cdot)$ is a prior distribution and $\hat{f}_n$ is the posterior mean of f_λ in a "phantom" model $Y_i = f_\lambda(x_i) + \xi_i'$, where ξ_i' are iid normally distributed with mean 0 and variance $\beta/2$.

The idea of mixing with exponential weights has been discussed by many authors apparently since 1970-ies (see [27] for a nice overview of the subject). Most of the work focused on the important particular case where the set of estimators is finite, i.e., w.l.o.g. $\Lambda = \{1,\dots,M\}$, and the distribution π is uniform on Λ. Procedures of the type (2)–(3) with general sets Λ and priors π came into consideration quite recently [9,8,3,29,30,1,2,25], partly in connection to the PAC-Bayesian approach. For finite Λ, procedures (2)–(3) were independently introduced for prediction of deterministic individual sequences with expert advice. Representative work and references can be found in [24,17,11]; in this framework the results are proved for cumulative loss and no assumption is made on the statistical nature of the data, whereas the observations Y_i are supposed to be uniformly bounded by a known constant. This is not the case for the regression model that we consider here.

We mention also related work on cumulative exponential weighting methods: there the aggregate is defined as the average $n^{-1}\sum_{k=1}^n \hat{f}_k$. For regression models with random design, such procedures are introduced and analyzed in [8], [9] and [26]. In particular, [8] and [9] establish a sharp oracle inequality, i.e., an inequality with leading constant 1. This result is further refined in [3] and [13]. In addition, [13] derives sharp oracle inequalities not only for the squared loss but also for general loss functions. However, these techniques are not helpful in the framework that we consider here, because the averaging device cannot be meaningfully adapted to models with non-identically distributed observations.

Aggregate $\hat{f}_n$ can be computed on-line. This, in particular, motivated its use for on-line prediction with finite Λ. Papers [13], [14] point out that $\hat{f}_n$ and its averaged version can be obtained as a special case of mirror descent algorithms that were considered earlier in deterministic minimization. Finally, [10] establishes an interesting link between the results for cumulative risks proved in the theory of prediction of deterministic sequences and generalization error bounds for the aggregates in the stochastic i.i.d. case.

In this paper we establish sharp oracle inequalities for the aggregate $\hat{f}_n$ under the squared loss, i.e., oracle inequalities with leading constant 1 and optimal rate of the remainder term. For a particular case, such an inequality has been pioneered in [16]. The result of [16] is proved for a finite set Λ and Gaussian errors. It makes use of Stein's unbiased risk formula, and gives a very precise constant in the remainder term of the inequality. The inequalities that we prove below are valid for general Λ and arbitrary functions f_λ satisfying some mild conditions. Furthermore, we treat non-Gaussian errors. We introduce new techniques of the proof based on dummy randomization which allows us to obtain the result for "n-divisible" distributions of errors ξ_i. We then apply the Skorokhod embedding to cover the class of all symmetric error distributions with finite exponential moments. Finally, we consider the case where f_λ is a linear combination of M known functions with the vector of weights $\lambda \in \mathbb{R}^M$. For this case, as a consequence of our main result we obtain a sparsity oracle inequality (SOI). We refer to [22] where the notion of SOI is introduced in a general context. Examples of SOI are proved in [15,5,4,6,23]. In particular, [5] deals with the regression model with fixed design that we consider here and proves approximate SOI for BIC type and Lasso type aggregates. We show that the aggregate with exponential weights satisfies a sharp SOI, i.e., a SOI with leading constant 1.

2 Risk Bounds for n-Divisible Distributions of Errors

The assumptions that we need to derive our main result concern essentially the probability distribution of the i.i.d. errors ξ_i.

(A) There exist i.i.d. random variables $\zeta_1, \dots, \zeta_n$ defined on an enlargement of the probability space $(\Omega, \mathcal{F}, P)$ such that:

(A1) the random variable $\xi_1 + \zeta_1$ has the same distribution as $(1 + 1/n)\xi_1$,

(A2) the vectors $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ are independent.

Note that (A) is an assumption on the distribution of ξ_1. If ξ_1 satisfies (A1), then we will say that its distribution is *n-divisible.* We defer to Section 4 the discussion about how rich is the class of n-divisible distributions.

Hereafter, we will write for brevity θ_λ instead of $\theta_\lambda(\beta, \pi, \mathbf{Y})$. Denote by $\mathscr{P}'_\Lambda$ the set of all the measures $\mu \in \mathscr{P}_\Lambda$ such that $\lambda \mapsto f_\lambda(x)$ is integrable w.r.t. μ for $x \in \{x_1, \dots, x_n\}$. Clearly $\mathscr{P}'_\Lambda$ is a convex subset of $\mathscr{P}_\Lambda$. For any measure $\mu \in \mathscr{P}'_\Lambda$ we define

$$\bar{f}_\mu(x_i) = \int_\Lambda f_\lambda(x_i)\, \mu(d\lambda), \quad i = 1, \dots, n.$$

We denote by $\theta \cdot \pi$ the probability measure $A \mapsto \int_A \theta_\lambda\, \pi(d\lambda)$ defined on $\mathcal{A}$. With the above notation, we have $\hat{f}_n = \bar{f}_{\theta \cdot \pi}$.

We will need one more assumption. Let $L_\zeta : \mathbb{R} \to \mathbb{R} \cup \{\infty\}$ be the moment generating function of the random variable ζ_1, i.e., $L_\zeta(t) = E(e^{t\zeta_1})$, $t \in \mathbb{R}$.

(B) There exist a functional $\Psi_\beta : \mathscr{P}'_\Lambda \times \mathscr{P}'_\Lambda \to \mathbb{R}$ and a real number $\beta_0 > 0$ such that

$$\begin{cases} e^{(\|f - \bar{f}_{\mu'}\|_n^2 - \|f - \bar{f}_\mu\|_n^2)/\beta} \prod_{i=1}^n L_\zeta\left(\frac{2(\bar{f}_\mu(x_i) - \bar{f}_{\mu'}(x_i))}{\beta}\right) \le \Psi_\beta(\mu, \mu'), \\ \mu \mapsto \Psi_\beta(\mu, \mu') \text{ is concave and continuous in the total} \\ \text{variation norm for any } \mu' \in \mathscr{P}'_\Lambda, \\ \Psi_\beta(\mu, \mu) = 1, \end{cases} \tag{4}$$

for any $\beta \ge \beta_0$.

Simple sufficient conditions for this assumption to hold in particular cases are given in Section 4.

The next theorem presents a "PAC-Bayesian" type bound.

Theorem 1. *Let π be an element of $\mathscr{P}_\Lambda$ such that $\theta \cdot \pi \in \mathscr{P}'_\Lambda$ for all $\mathbf{Y} \in \mathbb{R}^n$ and $\beta > 0$. If assumptions (A) and (B) are fulfilled, then the aggregate $\hat{f}_n$ defined by (2) with $\beta \geq \beta_0$ satisfies the oracle inequality*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \leq \int \|f_\lambda - f\|_n^2 \, p(d\lambda) + \frac{\beta\,\mathcal{K}(p,\pi)}{n+1}, \quad \forall\, p \in \mathscr{P}_\Lambda, \tag{5}$$

where $\mathcal{K}(p,\pi)$ stands for the Kullback-Leibler divergence between p and π.

Proof. Define the mapping $\boldsymbol{H} : \mathscr{P}'_\Lambda \to \mathbb{R}^n$ by

$$\boldsymbol{H}_\mu = (\bar{f}_\mu(x_1) - f(x_1), \dots, \bar{f}_\mu(x_n) - f(x_n))^\top, \quad \mu \in \mathscr{P}'_\Lambda.$$

For brevity, we will write

$$\boldsymbol{h}_\lambda = \boldsymbol{H}_{\delta_\lambda} = (f_\lambda(x_1) - f(x_1), \dots, f_\lambda(x_n) - f(x_n))^\top, \quad \lambda \in \Lambda,$$

where δ_λ is the Dirac measure at λ (that is $\delta_\lambda(A) = \mathbb{1}(\lambda \in A)$ for any $A \in \mathcal{A}$ where $\mathbb{1}(\cdot)$ denotes the indicator function).

Since $E(\xi_i) = 0$, assumption (A1) implies that $E(\zeta_i) = 0$ for $i = 1, \dots, n$. On the other hand, (A2) implies that $\boldsymbol{\zeta}$ is independent of θ_λ. Therefore, we have

$$E\Big(\|\bar{f}_{\theta\cdot\pi} - f\|_n^2\Big) = \beta E \log \exp\Big\{\frac{\|\bar{f}_{\theta\cdot\pi} - f\|_n^2 - 2\boldsymbol{\zeta}^\top \boldsymbol{H}_{\theta\cdot\pi}}{\beta}\Big\} = S + S_1$$

where

$$S = -\beta E \log \int_\Lambda \theta_\lambda \exp\Big\{-\frac{\|f_\lambda - f\|_n^2 - 2\boldsymbol{\zeta}^\top \boldsymbol{h}_\lambda}{\beta}\Big\}\pi(d\lambda),$$

$$S_1 = \beta E \log \int_\Lambda \theta_\lambda \exp\Big\{\frac{\|\bar{f}_{\theta\cdot\pi} - f\|_n^2 - \|f_\lambda - f\|_n^2 + 2\boldsymbol{\zeta}^\top(\boldsymbol{h}_\lambda - \boldsymbol{H}_{\theta\cdot\pi})}{\beta}\Big\}\pi(d\lambda).$$

The definition of θ_λ yields

$$\begin{aligned} S = &-\beta E \log \int_\Lambda \exp\Big\{-\frac{n\|\mathbf{Y} - f_\lambda\|_n^2 + \|f_\lambda - f\|_n^2 - 2\boldsymbol{\zeta}^\top \boldsymbol{h}_\lambda}{\beta}\Big\}\pi(d\lambda) \\ &+ \beta E \log \int_\Lambda \exp\Big\{-\frac{n\|\mathbf{Y} - f_\lambda\|_n^2}{\beta}\Big\}\pi(d\lambda). \end{aligned} \tag{6}$$

Since $\|\mathbf{Y} - f_\lambda\|_n^2 = \|\boldsymbol{\xi}\|_n^2 - 2n^{-1}\boldsymbol{\xi}^\top \boldsymbol{h}_\lambda + \|f_\lambda - f\|_n^2$, we get

$$\begin{aligned} S = &-\beta E \log \int_\Lambda \exp\Big\{-\frac{(n+1)\|f_\lambda - f\|_n^2 - 2(\boldsymbol{\xi} + \boldsymbol{\zeta})^\top \boldsymbol{h}_\lambda}{\beta}\Big\}\pi(d\lambda) \\ &+ \beta E \log \int_\Lambda \exp\Big\{-\frac{n\|f - f_\lambda\|_n^2 - 2\boldsymbol{\xi}^\top \boldsymbol{h}_\lambda}{\beta}\Big\}\pi(d\lambda) \\ = &\ \beta E \log \int_\Lambda e^{-n\rho(\lambda)}\pi(d\lambda) - \beta E \log \int_\Lambda e^{-(n+1)\rho(\lambda)}\pi(d\lambda), \end{aligned} \tag{7}$$

where we used the notation $\rho(\lambda) = (\|f - f_\lambda\|_n^2 - 2n^{-1}\boldsymbol{\xi}^\top \boldsymbol{h}_\lambda)/\beta$ and the fact that $\boldsymbol{\xi}+\boldsymbol{\zeta}$ can be replaced by $(1+1/n)\boldsymbol{\xi}$ inside the expectation. The Hölder inequality implies that $\int_\Lambda e^{-n\rho(\lambda)}\pi(d\lambda) \le (\int_\Lambda e^{-(n+1)\rho(\lambda)}\pi(d\lambda))^{\frac{n}{n+1}}$. Therefore,

$$S \le -\frac{\beta}{n+1} E \log \int_\Lambda e^{-(n+1)\rho(\lambda)}\,\pi(d\lambda). \tag{8}$$

Assume now that $p \in \mathscr{P}_\Lambda$ is absolutely continuous with respect to π. Denote by ϕ the corresponding Radon-Nikodym derivative and by Λ_+ the support of p. Using the concavity of the logarithm and Jensen's inequality we get

$$\begin{aligned} -E \log \int_\Lambda e^{-(n+1)\rho(\lambda)}\pi(d\lambda) &\le -E \log \int_{\Lambda_+} e^{-(n+1)\rho(\lambda)}\pi(d\lambda) \\ &= -E \log \int_{\Lambda_+} e^{-(n+1)\rho(\lambda)}\phi^{-1}(\lambda)\,p(d\lambda) \\ &\le (n+1)E \int_{\Lambda_+} \rho(\lambda)\,p(d\lambda) + \int_{\Lambda_+} \log\phi(\lambda)\,p(d\lambda). \end{aligned}$$

Noticing that the last integral here equals to $\mathcal{K}(p,\pi)$ and combining the resulting inequality with (8) we obtain

$$S \le \beta E \int_\Lambda \rho(\lambda)\,p(d\lambda) + \frac{\beta\,\mathcal{K}(p,\pi)}{n+1}.$$

Since $E(\xi_i) = 0$ for every $i = 1, \dots, n$, we have $\beta E(\rho(\lambda)) = \|f_\lambda - f\|_n^2$, and using the Fubini theorem we find

$$S \le \int_\Lambda \|f_\lambda - f\|_n^2\,p(d\lambda) + \frac{\beta\,\mathcal{K}(p,\pi)}{n+1}\,. \tag{9}$$

Note that this inequality also holds in the case where p is not absolutely continuous with respect to π, since in this case $\mathcal{K}(p,\pi) = \infty$.

To complete the proof, it remains to show that $S_1 \le 0$. Let $E_{\boldsymbol{\xi}}(\cdot)$ denote the conditional expectation $E(\cdot|\boldsymbol{\xi})$. By the concavity of the logarithm,

$$S_1 \le \beta E \log \int_\Lambda \theta_\lambda E_{\boldsymbol{\xi}} \exp\Big\{\frac{\|\bar{f}_{\theta\cdot\pi} - f\|_n^2 - \|f_\lambda - f\|_n^2 + 2\boldsymbol{\zeta}^\top(\boldsymbol{h}_\lambda - \boldsymbol{H}_{\theta\cdot\pi})}{\beta}\Big\}\pi(d\lambda).$$

Since $f_\lambda = \bar{f}_{\delta_\lambda}$ and $\boldsymbol{\zeta}$ is independent of θ_λ, the last expectation on the right hand side of this inequality is bounded from above by $\Psi_\beta(\delta_\lambda, \theta\cdot\pi)$. Now, the fact that $S_1 \le 0$ follows from the concavity and continuity of the functional $\Psi_\beta(\cdot, \theta\cdot\pi)$, Jensen's inequality and the equality $\Psi_\beta(\theta\cdot\pi, \theta\cdot\pi) = 1$.

Remark. Another way to read the result of Theorem 1 is that, if the probabilistic "phantom" Gaussian error model is used to construct $\hat{f}_n$, with variance taken larger than a certain threshold value, then the Bayesian posterior mean under the true model is close in expectation to the best prediction, even when the true data generating distribution does not have Gaussian errors, but errors of more general type.

3 Model Selection with Finite or Countable Λ

Consider now the particular case where Λ is countable. W.l.o.g. we suppose that $\Lambda = \{1, 2, \ldots\}$, $\{f_\lambda, \lambda \in \Lambda\} = \{f_j\}_{j=1}^\infty$ and we set $\pi_j \triangleq \pi(\lambda = j)$. As a corollary of Theorem 1 we get the following sharp oracle inequalities for model selection type aggregation.

Theorem 2. *Assume that π is an element of $\mathscr{P}_\Lambda$ such that $\theta \cdot \pi \in \mathscr{P}'_\Lambda$ for all $\mathbf{Y} \in \mathbb{R}^n$ and $\beta > 0$. Let assumptions (A) and (B) be fulfilled and let Λ be countable. Then for any $\beta \geq \beta_0$ the aggregate $\hat{f}_n$ satisfies the inequality*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \leq \inf_{j \geq 1} \bigg(\|f_j - f\|_n^2 + \frac{\beta \log \pi_j^{-1}}{n+1}\bigg).$$

In particular, if $\pi_j = 1/M$, $j = 1, \ldots, M$, we have

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \leq \min_{j=1,\ldots,M} \|f_j - f\|_n^2 + \frac{\beta \log M}{n+1}. \tag{10}$$

Proof. For a fixed integer $j_0 \geq 1$ we apply Theorem 1 with p being the Dirac measure: $p(\lambda = j) = \mathbb{1}(j = j_0)$, $j \geq 1$. This gives

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \leq \|f_{j_0} - f\|_n^2 + \frac{\beta \log \pi_{j_0}^{-1}}{n+1}.$$

Since this inequality holds for every j_0, we obtain the first inequality of the proposition. The second inequality is an obvious consequence of the first one.

Remark. The rate of convergence $(\log M)/n$ obtained in (10) is optimal rate of model selection type aggregation when the errors ξ_i are Gaussian [21,5].

4 Checking Assumptions (A) and (B)

In this section we give some sufficient conditions for assumptions (A) and (B). Denote by $\mathcal{D}_n$ the set of all probability distributions of ξ_1 satisfying assumption (A1). First, it is easy to see that all zero-mean Gaussian or double-exponential distributions belong to $\mathcal{D}_n$. Furthermore, $\mathcal{D}_n$ contains all stable distributions. However, since non-Gaussian stable distributions do not have second order moments, they do not satisfy (4). One can also check that the convolution of two distributions from $\mathcal{D}_n$ belongs to $\mathcal{D}_n$. Finally, note that the intersection $\mathcal{D} = \cap_{n \geq 1} \mathcal{D}_n$ is included in the set of all infinitely divisible distributions and is called the L-class (see [19], Theorem 3.6, p. 102).

However, some basic distributions such as the uniform or the Bernoulli distribution do not belong to $\mathcal{D}_n$. To show this, let us recall that the characteristic function of the uniform on $[-a, a]$ distribution is given by $\varphi(t) = \sin(at)/(\pi at)$. For this function, $\varphi((n+1)t)/\varphi(nt)$ is equal to infinity at the points where

$\sin(nat)$ vanishes (unless $n=1$). Therefore, it cannot be a characteristic function. Similar argument shows that the centered Bernoulli and centered binomial distributions do not belong to $\mathcal{D}_n$.

We now discuss two important cases of Theorem 1 where the errors ξ_i are either Gaussian or double exponential.

Proposition 1. *Assume that* $\sup_{\lambda\in\Lambda}\|f-f_\lambda\|_n \le L < \infty$. *If the random variables* ξ_i *are i.i.d. Gaussian* $\mathcal{N}(0,\sigma^2)$, $\sigma^2>0$, *then for every* $\beta \ge (4+2/n)\sigma^2+2L^2$ *the aggregate* $\hat f_n$ *satisfies inequality (5).*

Proof. If $\xi_i \sim \mathcal{N}(0,\sigma^2)$, assumption (A) is fulfilled with random variables ζ_i having the Gaussian distribution $\mathcal{N}(0,(2n+1)\sigma^2/n^2)$. Using the Laplace transform of the Gaussian distribution we get $L_\zeta(u)=\exp(\sigma^2u^2(2n+1)/(2n^2))$. Therefore, take

$$\Psi_\beta(\mu,\mu')=\exp\left(\frac{\|f-\bar f_{\mu'}\|_n^2-\|f-\bar f_\mu\|_n^2}{\beta}+\frac{2\sigma^2(2n+1)\|\bar f_\mu-\bar f_{\mu'}\|_n^2}{n\beta^2}\right).$$

This functional satisfies $\Psi_\beta(\mu,\mu)=1$, and it is not hard to see that the mapping $\mu\mapsto\Psi_\beta(\mu,\mu')$ is continuous in the total variation norm. Finally, this mapping is concave for every $\beta\ge(4+2/n)\sigma^2+2\sup_\lambda\|f-f_\lambda\|_n^2$ by virtue of Lemma 3 in the Appendix. Therefore, assumption (B) is fulfilled and the desired result follows from Theorem 1.

Assume now that ξ_i are distributed with the double exponential density

$$f_\xi(x)=\frac{1}{\sqrt{2\sigma^2}}\,e^{-\sqrt{2}|x|/\sigma},\quad x\in\mathbb{R}.$$

Aggregation under this assumption is discussed in [28] where it is recommended to modify the shape of the aggregate in order to match the shape of the distribution of the errors. The next proposition shows that sharp risk bounds can be obtained without modifying the algorithm.

Proposition 2. *Assume that* $\sup_{\lambda\in\Lambda}\|f-f_\lambda\|_n\le L<\infty$ *and* $\sup_{i,\lambda}|f_\lambda(x_i)|\le \bar L<\infty$. *Let the random variables* ξ_i *be i.i.d. double exponential with variance* $\sigma^2>0$. *Then for any* β *larger than*

$$\max\left(\left(8+\frac{4}{n}\right)\sigma^2+2L^2,\ 4\sigma\left(1+\frac{1}{n}\right)\bar L\right)$$

the aggregate $\hat f_n$ *satisfies inequality (5).*

Proof. We apply Theorem 1. The characteristic function of the double exponential density is $\varphi(t)=2/(2+\sigma^2t^2)$. Solving $\varphi(t)\varphi_\zeta(t)=\varphi((n+1)t/n)$ we get the characteristic function φ_ζ of ζ_1. The corresponding Laplace transform L_ζ in this case is $L_\zeta(t)=\varphi_\zeta(-it)$, which yields

$$L_\zeta(t)=1+\frac{(2n+1)\sigma^2t^2}{2n^2-(n+1)^2\sigma^2t^2}.$$

Therefore

$$\log L_\zeta(t) \le (2n+1)(\sigma t/n)^2, \quad |t| \le \frac{n}{(n+1)\sigma}\,.$$

We now use this inequality to check assumption (B). For all $\mu, \mu' \in \mathscr{P}_\Lambda$ we have

$$2\big|\bar f_\mu(x_i) - \bar f_{\mu'}(x_i)\big|/\beta \le 4\bar L/\beta, \quad i = 1, \dots, n.$$

Therefore, for $\beta > 4\sigma\big(1 + 1/n\big)\bar L$ we get

$$\log L_\zeta\bigg(2\big|\bar f_\mu(x_i) - \bar f_{\mu'}(x_i)\big|/\beta\bigg) \le \frac{4\sigma^2(2n+1)(\bar f_\mu(x_i) - \bar f_{\mu'}(x_i))^2}{n\beta^2}\,.$$

Thus, we get the functional Ψ_β of the same form as in the proof of Proposition 1, with the only difference that σ^2 is now replaced by $2\sigma^2$. Therefore, it suffices to repeat the reasoning of the proof of Proposition 1 to complete the proof.

5 Risk Bounds for General Distributions of Errors

As discussed above, assumption (A) restricts the application of Theorem 1 to models with "n-divisible" errors. We now show that this limitation can be dropped. Recall that the main idea of the proof of Theorem 1 consists in an artificial introduction of the dummy random vector $\boldsymbol{\zeta}$ independent of $\boldsymbol{\xi}$. However, the independence property is too strong as compared to what we really need in the proof of Theorem 1. Below we come to a weaker condition invoking a version of Skorokhod embedding (a detailed survey on this subject can be found in [18]).

For simplicity we assume that the errors ξ_i are symmetric, i.e., $P(\xi_i > a) = P(\xi_i < -a)$ for all $a \in \mathbb{R}$. The argument can be adapted to the asymmetric case as well, but we do not discuss it here.

We now describe a version of Skorokhod's construction that will be used below, cf. [20, Proposition II.3.8].

Lemma 1. *Let $\xi_1, \dots, \xi_n$ be i.i.d. symmetric random variables on $(\Omega, \mathcal{F}, P)$. Then there exist i.i.d. random variables $\zeta_1, \dots, \zeta_n$ defined on an enlargement of the probability space $(\Omega, \mathcal{F}, P)$ such that*

(a) $\boldsymbol{\xi} + \boldsymbol{\zeta}$ has the same distribution as $(1 + 1/n)\boldsymbol{\xi}$.
(b) $E(\zeta_i|\xi_i) = 0$, $i = 1, \dots, n$,
(c) for any $\lambda > 0$ and for any $i = 1, \dots, n$, we have

$$E(e^{\lambda\zeta_i}|\xi_i) \le e^{(\lambda\xi_i)^2(n+1)/n^2}.$$

Proof. Define ζ_i as a random variable such that, given ξ_i, it takes values ξ_i/n or $-2\xi_i - \xi_i/n$ with conditional probabilities $P(\zeta_i = \xi_i/n|\xi_i) = (2n+1)/(2n+2)$ and $P(\zeta_i = -2\xi_i - \xi_i/n|\xi_i) = 1/(2n+2)$. Then properties (a) and (b) are straightforward. Property (c) follows from the relation

$$E(e^{\lambda\zeta_i}|\xi_i) = e^{\frac{\lambda\xi_i}{n}}\bigg(1 + \frac{1}{2n+2}\big(e^{-2\lambda\xi_i(1+1/n)} - 1\big)\bigg)$$

and Lemma 2 in the Appendix with $x = \lambda\xi_i/n$ and $\alpha = 2n + 2$.

We now state the main result of this section.

Theorem 3. *Fix some $\alpha > 0$ and assume that $\sup_{\lambda\in\Lambda} \|f - f_\lambda\|_n \le L$ for a finite constant L. If the errors ξ_i are symmetric and have a finite second moment $E(\xi_i^2)$, then for any $\beta \ge 4(1 + 1/n)\alpha + 2L^2$ we have*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \int_\Lambda \|f_\lambda - f\|_n^2\, p(d\lambda) + \frac{\beta\,\mathcal{K}(p,\pi)}{n+1} + R_n, \quad \forall\, p \in \mathscr{P}_\Lambda, \qquad (11)$$

where the residual term R_n is given by

$$R_n = E^*\bigg(\sup_{\lambda\in\Lambda} \sum_{i=1}^{n} \frac{4(n+1)(\xi_i^2 - \alpha)(f_\lambda(x_i) - \bar{f}_{\theta\cdot\pi}(x_i))^2}{n^2\beta}\bigg)$$

and E^* denotes expectation with respect to the outer probability P^*.

Proof. In view of Lemma 1(b) the conditional expectation of random variable ζ_i given θ_λ vanishes. Therefore, with the notation of the proof of Theorem 1, we get $E(\|\hat{f}_n - f\|_n^2) = S + S_1$. Using Lemma 1(a) and acting exactly as in the proof of Theorem 1 we get that S is bounded as in (9). Finally, as shown in the proof of Theorem 1 the term S_1 satisfies

$$S_1 \le \beta E \log \int_\Lambda \theta_\lambda E_{\boldsymbol{\xi}} \exp\Big\{\frac{\|\bar{f}_{\theta\cdot\pi} - f\|_n^2 - \|f_\lambda - f\|_n^2 + 2\boldsymbol{\zeta}^\top(\boldsymbol{h}_\lambda - \boldsymbol{H}_{\theta\cdot\pi})}{\beta}\Big\}\pi(d\lambda).$$

According to Lemma 1(c),

$$E_{\boldsymbol{\xi}}\Big(e^{2\boldsymbol{\zeta}^T(\boldsymbol{h}_\lambda - \boldsymbol{H}_{\theta\cdot\pi})/\beta}\Big) \le \exp\Big\{\sum_{i=1}^{n} \frac{4(n+1)(f_\lambda(x_i) - \bar{f}_{\theta\cdot\pi}(x_i))^2\xi_i^2}{n^2\beta^2}\Big\}.$$

Therefore, $S_1 \le S_2 + R_n$, where

$$S_2 = \beta E \log \int_\Lambda \theta_\lambda \exp\Big(\frac{4\alpha(n+1)\|f_\lambda - \bar{f}_{\theta\cdot\pi}\|_n^2}{n\beta^2} - \frac{\|f - f_\lambda\|_n^2 - \|f - \bar{f}_{\theta\cdot\pi}\|_n^2}{\beta}\Big)\pi(d\lambda).$$

Finally, we apply Lemma 3 with $s^2 = 4\alpha(n+1)$ and Jensen's inequality to get that $S_2 \le 0$.

Corollary 1. *Let the assumptions of Theorem 3 be satisfied and let $|\xi_i| \le B$ almost surely where B is a finite constant. Then the aggregate $\hat{f}_n$ satisfies inequality (5) for any $\beta \ge 4B^2(1 + 1/n) + 2L^2$.*

Proof. It suffices to note that for $\alpha = B^2$ we get $R_n \le 0$.

Corollary 2. *Let the assumptions of Theorem 3 be satisfied and suppose that $E(e^{t|\xi_i|^\kappa}) \le B$ for some finite constants $t > 0$, $\kappa > 0$, $B > 0$. Then for any $n \ge e^{2/\kappa}$ and any $\beta \ge 4(1 + 1/n)(2(\log n)/t)^{1/\kappa} + 2L^2$ we have*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \int_\Lambda \|f_\lambda - f\|_n^2\, p(d\lambda) + \frac{\beta\,\mathcal{K}(p,\pi)}{n+1} \qquad (12)$$

$$+ \frac{16BL^2(n+1)(2\log n)^{2/\kappa}}{n^2\beta\, t^{2/\kappa}}, \quad \forall\, p \in \mathscr{P}_\Lambda.$$

In particular, if $\Lambda = \{1, \dots, M\}$ and π is the uniform measure on Λ we get

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \min_{j=1,\dots,M} \|f_j - f\|_n^2 + \frac{\beta \log M}{n+1} + \frac{16BL^2(n+1)(2\log n)^{2/\kappa}}{n^2\beta\, t^{2/\kappa}}. \tag{13}$$

Proof. Set $\alpha = (2(\log n)/t)^{1/\kappa}$ and note that

$$R_n \le \frac{4(n+1)}{n^2\beta} \sup_{\lambda\in\Lambda, \mu\in\mathscr{P}'_\Lambda} \|f_\lambda - \bar{f}_\mu\|_n^2 \sum_{i=1}^n E(\xi_i^2 - \alpha)_+ \le \frac{16L^2(n+1)}{n\beta} E(\xi_1^2 - \alpha)_+$$

where $a_+ = \max(0, a)$. For any $x \ge (2/(t\kappa))^{1/\kappa}$ the function $x^2 e^{-tx^\kappa}$ is decreasing. Therefore, for any $n \ge e^{2/\kappa}$ we have $x^2 e^{-tx^\kappa} \le \alpha^2 e^{-t\alpha^\kappa} = \alpha^2/n^2$, as soon as $x \ge \alpha$. Hence, $E(\xi_1^2 - \alpha)_+ \le B\alpha^2/n^2$ and the desired inequality follows.

Remark. Corollary 2 shows that if the tails of the errors have exponential decay and β is of the order $(\log n)^{1/\kappa}$ which minimizes the remainder term, then the rate of convergence in the oracle inequality (13) is of the order $(\log n)^{\frac{1}{\kappa}}(\log M)/n$. In the case $\kappa = 1$, comparing our result with the risk bound obtained in [13] for averaged algorithm in random design regression, we see that an extra $\log n$ multiplier appears. We conjecture that this deterioration is due to the technique of the proof and probably can be removed.

6 Sparsity Oracle Inequality

Let $\phi_1, \dots, \phi_M$ be some functions from $\mathcal{X}$ to $\mathbb{R}$. Consider the case where $\Lambda \subseteq \mathbb{R}^M$ and $f_\lambda = \sum_j \lambda_j \phi_j$, $\lambda = (\lambda_1, \dots, \lambda_M)$. For $\lambda \in \mathbb{R}^M$ denote by $J(\lambda)$ the set of indices j such that $\lambda_j \ne 0$, and set $M(\lambda) \triangleq Card(J(\lambda))$. For any $\tau > 0$, $0 < L_0 \le \infty$, define the probability densities

$$q_0(t) = \frac{3}{2(1+|t|)^4}, \;\; \forall t \in \mathbb{R},$$
$$q(\lambda) = \frac{1}{C_0} \prod_{j=1}^M \tau^{-1}\, q_0\big(\lambda_j/\tau\big)\mathbb{1}(\|\lambda\| \le L_0), \;\; \forall \lambda \in \mathbb{R}^M,$$

where $C_0 = C_0(\tau, M, L_0)$ is the normalizing constant and $\|\lambda\|$ stands for the Euclidean norm of $\lambda \in \mathbb{R}^M$.

Sparsity oracle inequalities (SOI) are oracle inequalities bounding the risk in terms of the sparsity index $M(\lambda)$ or similar characteristics. The next theorem provides a general tool to derive SOI from the "PAC-Bayesian" bound (5). Note that in this theorem $\hat{f}_n$ is not necessarily defined by (2). It can be any procedure satisfying (5).

Theorem 4. *Let $\hat{f}_n$ satisfy (5) with $\pi(d\lambda) = q(\lambda)\,d\lambda$ and $\tau \le \delta L_0/\sqrt{M}$ where $0 < L_0 \le \infty$, $0 < \delta < 1$. Assume that Λ contains the ball $\{\lambda \in \mathbb{R}^M : \|\lambda\| \le L_0\}$. Then for all λ^* such that $\|\lambda^*\| \le (1-\delta)L_0$ we have*

$$E\Big(\|\hat{f}_n - f\|_n^2\Big) \le \|f_{\lambda^*} - f\|_n^2 + \frac{4\beta}{n+1}\sum_{j\in J(\lambda^*)} \log(1 + \tau^{-1}|\lambda_j^*|) + R(M,\tau,L_0,\delta),$$

where the residual term is

$$R(M,\tau,L_0,\delta) = \tau^2 e^{2\tau^3 M^{5/2}(\delta L_0)^{-3}} \sum_{j=1}^M \|\phi_j\|_n^2 + \frac{2\beta\tau^3 M^{5/2}}{(n+1)\delta^3 L_0^3}$$

for $L_0 < \infty$ and $R(M,\tau,\infty,\delta) = \tau^2\sum_{j=1}^M \|\phi_j\|_n^2$.

Proof. We apply Theorem 1 with $p(d\lambda) = C_{\lambda^*}^{-1} q(\lambda - \lambda^*)\mathbb{1}(\|\lambda - \lambda^*\| \le \delta L_0)\,d\lambda$, where C_{λ^*} is the normalizing constant. Using the symmetry of q and the fact that $f_\lambda - f_{\lambda^*} = f_{\lambda-\lambda^*} = -f_{\lambda^*-\lambda}$ we get

$$\int_\Lambda \langle f_{\lambda^*} - f, f_\lambda - f_{\lambda^*}\rangle_n\, p(d\lambda) = C_{\lambda^*}^{-1}\int_{\|w\|\le\delta L_0} \langle f_{\lambda^*} - f, f_w\rangle_n\, q(w)\,dw = 0.$$

Therefore $\int_\Lambda \|f_\lambda - f\|_n^2\, p(d\lambda) = \|f_{\lambda^*} - f\|_n^2 + \int_\Lambda \|f_\lambda - f_{\lambda^*}\|_n^2\, p(d\lambda)$. On the other hand, bounding the indicator $\mathbb{1}(\|\lambda - \lambda^*\| \le \delta L_0)$ by one and using the identities $\int_{\mathbb{R}} q_0(t)\,dt = \int_{\mathbb{R}} t^2 q_0(t)\,dt = 1$, we obtain

$$\int_\Lambda \|f_\lambda - f_{\lambda^*}\|_n^2\, p(d\lambda) \le \frac{1}{C_0 C_{\lambda^*}} \sum_{j=1}^M \|\phi_j\|_n^2 \int_{\mathbb{R}} \frac{w_j^2}{\tau}\, q_0\Big(\frac{w_j}{\tau}\Big)\,dw_j = \frac{\tau^2\sum_{j=1}^M \|\phi_j\|_n^2}{C_0 C_{\lambda^*}}.$$

Since $1 - x \ge e^{-2x}$ for all $x \in [0, 1/2]$, we get

$$\begin{aligned} C_{\lambda^*}C_0 &= \frac{1}{\tau^M}\int_{\|\lambda\|\le\delta L_0} \Big\{\prod_{j=1}^M q_0\Big(\frac{\lambda_j}{\tau}\Big)\Big\}\,d\lambda \ge \frac{1}{\tau^M}\prod_{j=1}^M \Big\{\int_{|\lambda_j|\le\frac{\delta L_0}{\sqrt{M}}} q_0\Big(\frac{\lambda_j}{\tau}\Big)\,d\lambda_j\Big\} \\ &= \Big(\int_0^{\delta L_0/\tau\sqrt{M}} \frac{3dt}{(1+t)^4}\Big)^M = \Big(1 - \frac{1}{(1+\delta L_0\tau^{-1}M^{-1/2})^3}\Big)^M \\ &\ge \exp\Big(-\frac{2M}{(1+\delta L_0\tau^{-1}M^{-1/2})^3}\Big) \ge \exp(-2\tau^3 M^{5/2}(\delta L_0)^{-3}). \end{aligned}$$

On the other hand, in view of the inequality $1 + |\lambda_j/\tau| \le (1 + |\lambda_j^*/\tau|)(1 + |\lambda_j - \lambda_j^*|/\tau)$ the Kullback-Leibler divergence between p and π is bounded as follows:

$$\mathcal{K}(p,\pi) = \int_{\mathbb{R}^M} \log\Big(\frac{C_{\lambda^*}^{-1}q(\lambda-\lambda^*)}{q(\lambda)}\Big)\,p(d\lambda) \le 4\sum_{j=1}^M \log(1 + |\tau^{-1}\lambda_j^*|) - \log C_{\lambda^*}\,.$$

Easy computation yields $C_0 \le 1$. Therefore $C_{\lambda^*} \ge C_0 C_{\lambda^*} \ge \exp(-\frac{2\tau^3 M^{5/2}}{(\delta L_0)^3})$ and the desired result follows.

We now discuss a consequence of the obtained inequality in the case where the errors are Gaussian. Let us denote by Φ the Gram matrix associated to the family

$(\phi_j)_{j=1,\dots,M}$, i.e., $M \times M$ matrix with entries $\Phi_{j,j'} = n^{-1}\sum_{i=1}^n \phi_j(x_i)\phi_{j'}(x_i)$ for every $j, j' \in \{1,\dots,M\}$. We denote by $\lambda_{\max}(\Phi)$ the maximal eigenvalue of Φ. In what follows, for every $x > 0$, we write $\log_+ x = (\log x)_+$.

Corollary 3. *Let $\hat f_n$ be defined by (2) with $\pi(d\lambda) = q(\lambda)\,d\lambda$ and let $\tau = \frac{\delta L_0}{M\sqrt{n}}$ with $0 < L_0 < \infty$, $0 < \delta < 1$. Let ξ_i be i.i.d. Gaussian $\mathcal{N}(0,\sigma^2)$ with $\sigma^2 > 0$, $\lambda_{max}(\Phi) \le K^2$, $\|f\|_n \le \bar L$ and let $\beta \ge (4+2n^{-1})\sigma^2 + 2L^2$ with $L = \bar L + L_0 K$. Then for all $\lambda^* \in \mathbb{R}^M$ such that $\|\lambda^*\| \le (1-\delta)L_0$ we have*

$$E\big[\|\hat f_n - f\|_n^2\big] \le \|f_{\lambda^*} - f\|_n^2 + \frac{4\beta}{n+1}\Big[M(\lambda^*)\Big(1 + \log_+\Big\{\frac{M\sqrt{n}}{\delta L_0}\Big\}\Big) + \sum_{J(\lambda^*)} \log_+ |\lambda_j^*|\Big]$$
$$+ \frac{C}{nM^{1/2}\min(M^{1/2}, n^{3/2})}\,,$$

where C is a positive constant independent of n, M and λ^*.

Proof. We apply Theorem 4 with $\Lambda = \{\lambda \in \mathbb{R}^M : \|\lambda\| \le L_0\}$. We need to check that $\hat f_n$ satisfies (5). This is indeed the case in view of Proposition 1 and the inequalities $\|f_\lambda - f\|_n \le \|f\|_n + \sqrt{\lambda^\top \Phi \lambda} \le \bar L + K\|\lambda\| \le L$. Thus we have

$$E\Big(\|\hat f_n - f\|_n^2\Big) \le \|f_{\lambda^*} - f\|_n^2 + \frac{4\beta}{n+1}\sum_{j\in J(\lambda^*)} \log(1 + \tau^{-1}|\lambda_j^*|) + R(M,\tau,L_0,\delta),$$

with $R(M,\tau,L_0,\delta)$ as in Theorem 4. One easily checks that $\log(1+\tau^{-1}|\lambda_j^*|) \le 1 + \log_+(\tau^{-1}|\lambda_j^*|) \le 1 + \log_+(\tau^{-1}) + \log_+(|\lambda_j^*|)$. Hence, the desired inequality follows from

$$\begin{aligned} R(M,\tau,L_0,\delta) &= \tfrac{(\delta L_0)^2}{M^2 n} e^{2M^{-3}n^{-3/2}M^{5/2}} \textstyle\sum_{j=1}^M \|\phi_j\|_n^2 + \tfrac{2\beta M^{5/2}}{(n+1)M^3 n^{3/2}} \\ &\le \tfrac{(\delta L_0)^2 M K^2 e^2}{M^2 n} + \tfrac{2\beta}{(n+1)M^{1/2}n^{3/2}} \le \tfrac{C}{nM^{1/2}\min(M^{1/2},n^{3/2})}\,. \end{aligned}$$

Remark. The result of Corollary 3 can be compared with the SOI obtained for other procedures [5,6,7]. These papers impose heavy restrictions on the Gram matrix Φ either in terms of the coherence introduced in [12] or analogous local characteristics. Our result is not of that kind: we need only that the maximal eigenvalue of Φ were bounded. On the other hand, we assume that the oracle vector λ^* belongs to a ball of radius $< L_0$ in ℓ_2 with known L_0. This assumption is not very restrictive in the sense that the ℓ_2 constraint is weaker than the ℓ_1 constraint that is frequently imposed. Moreover, the structure of our oracle inequality is such that we can consider slowly growing L_0, without seriously damaging the result.

References

1. Audibert, J.-Y.: Une approche PAC-bayésienne de la théorie statistique de l'apprentissage. PhD Thesis. University of Paris 6 (2004)
2. Audibert, J.-Y.: A randomized online learning algorithm for better variance control. In: COLT 2006. Proceedings of the 19th Annual Conference on Learning Theory. LNCS (LNAI), vol. 4005, pp. 392–407. Springer, Heidelberg (2006)

3. Bunea, F., Nobel, A.B.: Sequential Procedures for Aggregating Arbitrary Estimators of a Conditional Mean. Preprint Florida State University (2005), `http://www.stat.fsu.edu/~flori`
4. Bunea, F., Tsybakov, A.B., Wegkamp, M.H.: Aggregation and sparsity via ℓ_1-penalized least squares. In: Lugosi, G., Simon, H.U. (eds.) COLT 2006. LNCS (LNAI), vol. 4005, pp. 379–391. Springer, Heidelberg (2006)
5. Bunea, F., Tsybakov, A.B., Wegkamp, M.H.: Aggregation for gaussian regression. Annals of Statistics, to appear (2007), `http://www.stat.fsu.edu/~wegkamp`
6. Bunea, F., Tsybakov, A.B., Wegkamp, M.H.: Sparsity oracle inequalities for the Lasso, Submitted (2006)
7. Candes, E., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Annals of Statistics, to appear (2007)
8. Catoni, O.: Universal. aggregation rules with exact bias bounds. Preprint n.510, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7 (1999), `http://www.proba.jussieu.fr/mathdoc/preprints/index.html#1999`
9. Catoni, O.: Statistical Learning Theory and Stochastic Optimization. In: Ecole d'été de Probabilités de Saint-Flour 2001. Lecture Notes in Mathematics, Springer, Heidelberg (2004)
10. Cesa-Bianchi, N., Conconi, A., Gentile, G.: On the generalization ability of on-line learning algorithms. IEEE Trans. on Information Theory 50, 2050–2057 (2004)
11. Cesa-Bianchi, N., Lugosi, G.: Prediction, Learning, and Games. Cambridge University Press, New York (2006)
12. Donoho, D.L., Elad, M., Temlyakov, V.: Stable Recovery of Sparse Overcomplete Representations in the Presence of Noise. IEEE Trans. on Information Theory 52, 6–18 (2006)
13. Juditsky, A., Rigollet, P., Tsybakov, A.: Learning by mirror averaging. Preprint n, Laboratoire de Probabilités et Modèle aléatoires, Universités Paris 6 and Paris 7, (2005). n. 1034, `https://hal.ccsd.cnrs.fr/ccsd-00014097`
14. Juditsky, A.B., Nazin, A.V., Tsybakov, A.B., Vayatis, N.: Recursive aggregation of estimators via the Mirror Descent Algorithm with averaging. Problems of Information Transmission 41, 368–384 (2005)
15. Koltchinskii, V.: Sparsity in penalized empirical risk minimization, Submitted (2006)
16. Leung, G., Barron, A.: Information theory and mixing least-square regressions. IEEE Transactions on Information Theory 52, 3396–3410 (2006)
17. Littlestone, N., Warmuth, M.K.: The weighted majority algorithm. Information and Computation 108, 212–261 (1994)
18. Obloj, J.: The Skorokhod embedding problem and its offspring. Probability Surveys 1, 321–392 (2004)
19. Petrov, V.V.: Limit Theorems of Probability Theory. Clarendon Press, Oxford (1995)
20. Revuz, D., Yor, M.: Continuous Martingales and Brownian Motion. Springer, Heidelberg (1999)
21. Tsybakov, A.B.: Optimal rates of aggregation. In: Schölkopf, B., Warmuth, M. (eds.) Computational Learning Theory and Kernel Machines. LNCS (LNAI), vol. 2777, pp. 303–313. Springer, Heidelberg (2003)
22. Tsybakov, A.B.: Regularization, boosting and mirror averaging. Comments on "Regularization in Statistics", by Bickel, P., Li, B., Test 15, 303–310 (2006)
23. van de Geer, S.A.: High dimensional generalized linear models and the Lasso. Research report No.133. Seminar für Statistik, ETH, Zürich (2006)

24. Vovk, V.: Aggregating Strategies. In: Proceedings of the 3rd Annual Workshop on Computational Learning Theory, COLT1990, pp. 371–386. Morgan Kaufmann, San Francisco, CA (1990)
25. Vovk, V.: Competitive on-line statistics. International Statistical Review 69, 213–248 (2001)
26. Yang, Y.: Combining different procedures for adaptive regression. Journal of Multivariate Analysis 74, 135–161 (2000)
27. Yang, Y.: Adaptive regression by mixing. Journal of the American Statistical Association 96, 574–588 (2001)
28. Yang, Y.: Regression with multiple candidate models: selecting or mixing? Statist. Sinica 13, 783–809 (2003)
29. Zhang, T.: From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. Annals of Statistics, to appear (2007)
30. Zhang, T.: Information theoretical upper and lower bounds for statistical estimation. IEEE Transactions on Information Theory, to appear (2007)

A Appendix

Lemma 2. *For any $x \in \mathbb{R}$ and any $\alpha > 0$, $x + \log\big(1 + \frac{1}{\alpha}\big(e^{-x\alpha} - 1\big)\big) \le \frac{x^2\alpha}{2}$.*

Proof. On the interval $(-\infty, 0]$, the function $x \mapsto x + \log\big(1 + \frac{1}{\alpha}(e^{-x\alpha} - 1)\big)$ is increasing, therefore it is bounded by its value at 0, that is by 0. For positive values of x, we combine the inequalities $e^{-y} \le 1 - y + y^2/2$ (with $y = x\alpha$) and $\log(1 + y) \le y$ (with $y = 1 + \frac{1}{\alpha}(e^{-x\alpha} - 1)$).

Lemma 3. *For any $\beta \ge s^2/n + 2\sup_{\lambda \in \Lambda} \|f - f_\lambda\|_n^2$ and for every $\mu' \in \mathscr{P}'_\Lambda$, the function*

$$\mu \mapsto \exp\Big(\frac{s^2\|\bar f_{\mu'} - \bar f_\mu\|_n^2}{n\beta^2} - \frac{\|f - \bar f_\mu\|_n^2}{\beta}\Big)$$

is concave.

Proof. Consider first the case where $Card(\Lambda) = m < \infty$. Then every element of $\mathscr{P}_\Lambda$ can be viewed as a vector from $\mathbb{R}^m$. Set

$$\begin{aligned} Q(\mu) &= (1 - \gamma)\|f - f_\mu\|_n^2 + 2\gamma\langle f - f_\mu, f - f_{\mu'}\rangle_n \\ &= (1 - \gamma)\mu^T H_n^T H_n \mu + 2\gamma\mu^T H_n^T H_n \mu', \end{aligned}$$

where $\gamma = s^2/(n\beta)$ and H_n is the $n \times m$ matrix with entries $(f(x_i) - f_\lambda(x_i))/\sqrt{n}$. The statement of the lemma is equivalent to the concavity of $e^{-Q(\mu)/\beta}$ as a function of $\mu \in \mathscr{P}_\Lambda$, which holds if and only if the matrix $\beta\nabla^2 Q(\mu) - \nabla Q(\mu)\nabla Q(\mu)^T$ is positive-semidefinite. Simple algebra shows that $\nabla^2 Q(\mu) = 2(1-\gamma)H_n^T H_n$ and $\nabla Q(\mu) = 2H_n^T[(1 - \gamma)H_n\mu + \gamma H_n\mu']$. Therefore, $\nabla Q(\mu)\nabla Q(\mu)^T = H_n^T \mathbf{M} H_n$, where $\mathbf{M} = 4H_n\tilde\mu\tilde\mu^T H_n^T$ with $\tilde\mu = (1 - \gamma)\mu + \gamma\mu'$. Under our assumptions, β is larger than s^2/n, ensuring thus that $\tilde{\mu} \in \mathscr{P}_\Lambda$. Clearly, $\mathbf{M}$ is a symmetric and positive-semidefinite matrix. Moreover,

$$\begin{aligned}\lambda_{max}(\mathbf{M}) &\leq \mathrm{Tr}(\mathbf{M}) = 4\|H_n\tilde{\mu}\|^2 = \frac{4}{n}\sum_{i=1}^{n}\Big(\sum_{\lambda\in\Lambda}\tilde{\mu}_\lambda(f-f_\lambda)(x_i)\Big)^2 \\ &\leq \frac{4}{n}\sum_{i=1}^{n}\sum_{\lambda\in\Lambda}\tilde{\mu}_\lambda(f(x_i)-f_\lambda(x_i))^2 = 4\sum_{\lambda\in\Lambda}\tilde{\mu}_\lambda\|f-f_\lambda\|_n^2 \\ &\leq 4\max_{\lambda\in\Lambda}\|f-f_\lambda\|_n^2\end{aligned}$$

where $\lambda_{max}(\mathbf{M})$ is the largest eigenvalue of $\mathbf{M}$ and $\mathrm{Tr}(\mathbf{M})$ is its trace. This estimate yields the matrix inequality

$$\nabla Q(\mu)\nabla Q(\mu)^T \leq 4\max_{\lambda\in\Lambda}\|f-f_\lambda\|_n^2\, H_n^T H_n.$$

Hence, the function $e^{-Q(\mu)/\beta}$ is concave as soon as $4\max_{\lambda\in\Lambda}\|f-f_\lambda\|_n^2 \leq 2\beta(1-\gamma)$. The last inequality holds for every $\beta \geq n^{-1}s^2 + 2\max_{\lambda\in\Lambda}\|f-f_\lambda\|_n^2$.

The general case can be reduced to the case of finite Λ as follows. The concavity of the functional $G(\mu) = \exp\Big(\frac{s^2\|\bar{f}_{\mu'}-\bar{f}_\mu\|_n^2}{n\beta^2} - \frac{\|f-\bar{f}_\mu\|_n^2}{\beta}\Big)$ is equivalent to the validity of the inequality

$$G\Big(\frac{\mu+\tilde{\mu}}{2}\Big) \geq \frac{G(\mu)+G(\tilde{\mu})}{2}, \qquad \forall\ \mu,\tilde{\mu}\in\mathscr{P}'_\Lambda. \tag{14}$$

Fix now arbitrary $\mu,\tilde{\mu}\in\mathscr{P}'_\Lambda$. Take $\tilde{\Lambda} = \{1,2,3\}$ and consider the set of functions $\{\tilde{f}_\lambda, \lambda\in\tilde{\Lambda}\} = \{\bar{f}_\mu, \bar{f}_{\tilde{\mu}}, \bar{f}_{\mu'}\}$. Since $\tilde{\Lambda}$ is finite, $\mathscr{P}'_{\tilde{\Lambda}} = \mathscr{P}_{\tilde{\Lambda}}$. According to the first part of the proof, the functional

$$\tilde{G}(\nu) = \exp\left(\frac{s^2\|\bar{f}_{\mu'} - \bar{\tilde{f}}_\nu\|_n^2}{n\beta^2} - \frac{\|f-\bar{\tilde{f}}_\nu\|_n^2}{\beta}\right), \quad \nu\in\mathscr{P}_{\tilde{\Lambda}},$$

is concave on $\mathscr{P}_{\tilde{\Lambda}}$ as soon as $\beta \geq s^2/n + 2\max_{\lambda\in\tilde{\Lambda}}\|f-\tilde{f}_\lambda\|_n^2$, and therefore for every $\beta \geq s^2/n + 2\sup_{\lambda\in\Lambda}\|f-f_\lambda\|_n^2$ as well. (Indeed, by Jensen's inequality for any measure $\mu\in\mathscr{P}'_\Lambda$ we have $\|f-\bar{f}_\mu\|_n^2 \leq \int\|f-f_\lambda\|_n^2\mu(d\lambda) \leq \sup_{\lambda\in\Lambda}\|f-f_\lambda\|_n^2$.) This leads to

$$\tilde{G}\Big(\frac{\nu+\tilde{\nu}}{2}\Big) \geq \frac{\tilde{G}(\nu)+\tilde{G}(\tilde{\nu})}{2}, \qquad \forall\ \nu,\tilde{\nu}\in\mathscr{P}_{\tilde{\Lambda}}.$$

Taking here the Dirac measures ν and $\tilde{\nu}$ defined by $\nu(\lambda=j) = \mathbb{1}(j=1)$ and $\tilde{\nu}(\lambda=j) = \mathbb{1}(j=2)$, $j=1,2,3$, we arrive at (14). This completes the proof of the lemma.