Mach Learn
DOI 10.1007/s10994-008-5051-0

Aggregation by exponential weighting, sharp PAC-Bayesian bounds and sparsity

A. Dalalyan · A.B. Tsybakov

Received: 10 September 2007 / Revised: 17 March 2008 / Accepted: 21 March 2008

Abstract We study the problem of aggregation under the squared loss in the model of regression with deterministic design. We obtain sharp PAC-Bayesian risk bounds for aggregates defined via exponential weights, under general assumptions on the distribution of errors and on the functions to aggregate. We then apply these results to derive sparsity oracle inequalities.

Keywords Aggregation · Nonparametric regression · Oracle inequalities · Sparsity

1 Introduction

Aggregation with exponential weights is an important tool in machine learning. It is used for estimation, prediction with expert advice, in PAC-Bayesian settings and other problems. In this paper we establish a link between aggregation with exponential weights and sparsity. More specifically, we obtain a new type of oracle inequalities and apply them to show that the exponential weighted aggregate with a suitably chosen prior has a sparsity property.

We consider the regression model

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

where $x_1, \ldots, x_n$ are given non-random 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

Editors: Claudio Gentile, Nader H. Bshouty.

Research partially supported by the grant ANR-06-BLAN-0194 and by the PASCAL Network of Excellence.

A. Dalalyan (✉) · A.B. Tsybakov
LPMA, University of Paris 6, 4, Place Jussieu, 75252 Paris cedex 05, France
e-mail: dalalyan@ccr.jussieu.fr

A.B. Tsybakov
Laboratoire de Statistique, CREST, Timbre J340, 3, av. Pierre Larousse, 92240 Malakoff cedex, France

$(\Omega, \mathcal{F}, P)$ where $\Omega \subseteq \mathbb{R}$. The problem is to estimate the function f from the data $D_n = ((x_1, Y_1), \ldots, (x_n, Y_n))$.

Let $(\Lambda, \mathcal{A})$ be a measurable space and denote by $\mathcal{P}_\Lambda$ 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 $\lambda \mapsto f_\lambda(x)$ is measurable for all $x \in \mathcal{X}$, where $\mathbb{R}$ is 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 $\mathbf{Y} \triangleq (Y_1, \ldots, Y_n)$ and considered as frozen.

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

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

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 $\mathcal{P}_\Lambda$ and for $\beta > 0$ we set

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

with

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

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

Note that $\theta_\lambda(\mathbf{Y}) = \theta_\lambda(\beta, \pi, \mathbf{Y})$, so that $\hat{f}_n$ depends on two tuning parameters: the probability 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

$$Y_i = f_\lambda(x_i) + \xi_i' \tag{4}$$

where ξ_i' are i.i.d. normally distributed random variables 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 Yang 2001 for an overview of the subject). Most of the work has been focused on the important particular case where the set of estimators is finite, i.e., w.l.o.g. $\Lambda = \{1, \ldots, M\}$, 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 (Catoni 1999; 2004, Vovk 2001; Bunea 2005; Zhang 2006a, 2006b; Audibert 2004, 2006), partly in connection with 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 (Vovk 1990; Littlestone and Warmuth 1994; Cesa-Bianchi et al. 1997; Kivinen and Warmuth 1999; Cesa-Bianchi and Lugosi 2006); in this framework the results are proven 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.

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 (Catoni 1999, 2004; Yang 2000). In particular, (Catoni 1999) and (Catoni 2004) establish a sharp oracle inequality, i.e., an inequality with leading constant 1. This result is further refined in (Bunea 2005) and (Juditsky et al. 2008). In addition, (Juditsky et al. 2008) 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 is not meaningfully adapted to models with non-identically distributed observations.

For finite Λ, the aggregate $\hat{f}_n$ can be computed on-line. This, in particular, motivated its use for on-line prediction. Papers (Juditsky et al. 2005, 2008) 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, (Cesa-Bianchi et al. 2004; Juditsky et al. 2008) establish some links 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 obtain 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. Such an inequality has been pioneered in (Leung and Barron 2006) in a somewhat different setting. Namely, it is assumed in (Leung and Barron 2006) that Λ is a finite set, the errors ξ_i are Gaussian and f_λ are estimators constructed from the same sample D_n and satisfying some strong restrictions (essentially, these should be the projection estimators). The result of (Leung and Barron 2006) makes use of Stein's unbiased risk formula, and gives a very precise constant in the remainder term of the inequality. Inspection of the argument in (Leung and Barron 2006) shows that it can also be applied in the following special case of our setting: f_λ are arbitrary fixed functions, Λ is a finite set and the errors ξ_i are Gaussian.

The general line of our argument is to establish some PAC-Bayesian risk bounds (cf. (8), (10)) and then to derive sharp oracle inequalities by making proper choices of the probability measure p involved in those bounds (cf. Sects. 5, 7).

The main technical effort is devoted to the proof of the PAC-Bayesian bounds (Sects. 3, 4, 6). The results are valid for general Λ and arbitrary functions f_λ satisfying some mild conditions. Furthermore, we treat non-Gaussian errors ξ_i. For this purpose, we suggest three different approaches to prove the PAC-Bayesian bounds. The first one is based on integration by parts techniques that generalizes Stein's unbiased risk formula (Sect. 3). It is close in the spirit to (Leung and Barron 2006). This approach leads to most accurate results but it covers only a narrow class of distributions of the errors ξ_i. In Sect. 4 we introduce another techniques based on dummy randomization which allows us to obtain sharp risk bounds when the distributions of errors ξ_i are n-divisible. Finally, the third approach (Sect. 6) invokes the Skorokhod embedding and covers the class of all symmetric error distributions with finite moments of order larger than or equal to 2. Here the price to pay for the generality of the distribution of errors is in the rate of convergence that becomes slower if only smaller moments are finite.

In Sect. 7 we analyze our risk bounds in the important special case where f_λ is a linear combination of M known functions $\phi_1,\dots,\phi_M$ with the vector of weights $\lambda=(\lambda_1,\dots,\lambda_M)$: $f_\lambda=\sum_{j=1}^{M}\lambda_j\phi_j$. This setting is connected with the following three problems.

1. *High-dimensional linear regression.* Assume that the regression function has the form $f=f_{\lambda^*}$ where $\lambda^*\in\mathbb{R}^M$ is an unknown vector, in other words we have a linear regression model. During the last years a great deal of attention has been focused on estimation in such a linear model where the number of variables M is much larger than the sample size n. The

idea is that the effective dimension of the model is defined not by the number of potential parameters M but by the unknown number of non-zero components $M(\lambda^*)$ of vector λ^* that can be much smaller than n. In this situation methods like Lasso, LARS or Dantzig selector are used (Efron et al. 2004; Candes and Tao 2007). It is proved that if $M(\lambda^*) \ll n$ and if the dictionary $\{\phi_1, \dots, \phi_M\}$ satisfies certain conditions, then the vector λ^* and the function f can be estimated with reasonable accuracy (Greenshtein and Ritov 2004; Bunea et al. 2007a, 2007b; Candes and Tao 2007; Zhang and Huang 2008; Bickel et al. 2007). However, the conditions on the dictionary $\{\phi_1, \dots, \phi_M\}$ required to get risk bounds for the Lasso and Dantzig selector are quite restrictive. One of the consequences of our results in Sect. 7 is that a suitably defined aggregate with exponential weights attains essentially the same and sometimes even better behavior than the Lasso or Dantzig selector with no assumption on the dictionary, except for the standard normalization.

2. *Adaptive nonparametric regression.* Assume that f is a smooth function, and $\{\phi_1, \dots, \phi_M\}$ are the first M functions from a basis in $L_2(\mathbb{R}^d)$. If the basis is orthonormal, it is well-known that adaptive estimators of f can be constructed in the form $\sum_{j=1}^{M} \hat{\lambda}_j \phi_j$ where $\hat{\lambda}_j$ are appropriately chosen data-driven coefficients and M is a suitably selected integer such that $M \le n$ (cf., e.g., Nemirovski 2000; Tsybakov 2004). Our aggregation procedure suggests a more general way to treat adaptation covering the problems where the system $\{\phi_j\}$ is not necessarily orthonormal, even not necessarily a basis, and M is not necessarily smaller than n. In particular, the situation where $M \gg n$ arises if we want to deal with sparse functions f that have very few non-zero scalar products with functions from the dictionary $\{\phi_j\}$, but these non-zero coefficients can correspond to very high "harmonics". The results of Sect. 7 cover this case.

3. *Linear, convex or model selection type aggregation.* Assume now that $\phi_1, \dots, \phi_M$ are either some preliminary estimators of f constructed from a training sample independent of $(Y_1, \dots, Y_n)$ or some weak learners, and our aim is to construct an aggregate which is approximately as good as the best among $\phi_1, \dots, \phi_M$ or approximately as good as the best linear or convex combination of $\phi_1, \dots, \phi_M$. In other words, we deal with the problems of model selection (MS) type aggregation or linear/convex aggregation respectively (Nemirovski 2000; Tsybakov 2003). It is shown in (Bunea et al. 2007a) that a BIC type aggregate achieves optimal rates simultaneously for MS, linear and convex aggregation. This result is deduced in (Bunea et al. 2007a) from a sparsity oracle inequality (SOI), i.e., from an oracle inequality stated in terms of the number $M(\lambda)$ of non-zero components of λ. For a discussion of the concept of SOI we refer to (Tsybakov 2006). Examples of SOI are proved in (Koltchinskii 2006; Bunea et al. 2006, 2007a, 2007b; van de Geer 2006; Bickel et al. 2007) for the Lasso, BIC and Dantzig selector aggregates. Note that the SOI for the Lasso and Dantzig selector are not as strong as those for the BIC: they fail to guarantee optimal rates for MS, linear and convex aggregation unless $\phi_1, \dots, \phi_M$ satisfy some very restrictive conditions. On the other hand, the BIC aggregate is computationally feasible only for very small dimensions M. So, neither of these methods achieves both the computational efficiency and the optimal theoretical performance.

In Sect. 7 we propose a new approach to sparse recovery that realizes a compromise between the theoretical properties and the computational efficiency. We first suggest a general technique of deriving SOI from the PAC-Bayesian bounds, not necessarily for our particular aggregate $\hat{f}_n$. We then show that the exponentially weighted aggregate $\hat{f}_n$ with an appropriate prior measure π satisfies a sharp SOI, i.e., a SOI with leading constant 1. Its theoretical performance is comparable with that of the BIC in terms of sparsity oracle inequalities for the prediction risk. No assumption on the dictionary $\phi_1, \dots, \phi_M$ is required, except for the standard normalization. Even more, the result is sharper than the best available SOI for the

BIC-type aggregate (Bunea et al. 2007a), since the leading constant in the oracle inequality of (Bunea et al. 2007a) is strictly greater than 1. At the same time, similarly to the Lasso and Dantzig selector, our method is computationally feasible for moderately large dimensions M.

2 Some notation

In what follows we will often write for brevity θ_λ instead of $\theta_\lambda(\mathbf{Y})$. For any vector $\mathbf{z} = (z_1, \dots, z_n)^\top \in \mathbb{R}^n$ set

$$\|\mathbf{z}\|_n = \left(\frac{1}{n} \sum_{i=1}^{n} z_i^2 \right)^{1/2}.$$

Denote by $\mathcal{P}'_\Lambda$ the set of all measures $\mu \in \mathcal{P}_\Lambda$ such that $\lambda \mapsto f_\lambda(x)$ is integrable w.r.t. μ for $x \in \{x_1, \dots, x_n\}$. Clearly $\mathcal{P}'_\Lambda$ is a convex subset of $\mathcal{P}_\Lambda$. For any measure $\mu \in \mathcal{P}'_\Lambda$ 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 can write

$$\hat{f}_n = \bar{f}_{\theta \cdot \pi}.$$

3 A PAC-Bayesian bound based on unbiased risk estimation

In this section we prove our first PAC-Bayesian bound. An important element of the proof is an extension of Stein's identity which uses integration by parts. For this purpose we introduce the function

$$m_\xi(x) = -E[\xi_1 \mathbb{1}(\xi_1 \le x)] = -\int_{-\infty}^{x} z dF_\xi(z) = \int_x^{\infty} z dF_\xi(z),$$

where $F_\xi(z) = P(\xi_1 \le z)$ is the c.d.f. of ξ, $\mathbb{1}(\cdot)$ denotes the indicator function and the last equality follows from the assumption $E(\xi_1) = 0$. Since $E|\xi_1| < \infty$ the function m_ξ is well defined, non negative and satisfies $m_\xi(-\infty) = m_\xi(+\infty) = 0$. Moreover, m_ξ is increasing on $(-\infty, 0]$, decreasing on $[0, +\infty)$ and $\max_{x \in \mathbb{R}} m_\xi(x) = m_\xi(0) = \frac{1}{2}E|\xi_1|$. We will need the following assumption.

(A) *$E(\xi_1^2) = \sigma^2 < \infty$ and the measure $m_\xi(z)dz$ is absolutely continuous with respect to $dF_\xi(z)$ with a bounded Radon-Nikodym derivative, i.e., there exists a function $g_\xi : \mathbb{R} \to \mathbb{R}_+$ such that $\|g_\xi\|_\infty \triangleq \sup_{x \in \mathbb{R}} g_\xi(x) < \infty$ and*

$$\int_a^{a'} m_\xi(z)dz = \int_a^{a'} g_\xi(z) dF_\xi(z), \quad \forall a, a' \in \mathbb{R}.$$

Clearly, Assumption (A) is a restriction on the probability distribution of the errors ξ_i. Some examples where Assumption (A) is fulfilled are:

(i) If $\xi_1 \sim \mathcal{N}(0, \sigma^2)$, then $g_\xi(x) \equiv \sigma^2$.

(ii) If ξ_1 is uniformly distributed in the interval $[-b, b]$, then $m_\xi(x) = (b^2 - x^2)_+/(4b)$ and $g_\xi(x) = (b^2 - x^2)_+/2$.
(iii) If ξ_1 has a density function f_ξ with compact support $[-b, b]$ and such that $f_\xi(x) \geq f_{\min} > 0$ for every $x \in [-b, b]$, then assumption (A) is satisfied with $g_\xi(x) = m_\xi(x)/f_\xi(x) \leq E|\xi_1|/(2f_{\min})$.

We now give some examples where (A) is not fulfilled:

(iv) If ξ_1 has a double exponential distribution with zero mean and variance σ^2, then $g_\xi(x) = (\sigma^2 + \sqrt{2\sigma^2}|x|)/2$.
(v) If ξ_1 is a Rademacher random variable, then $m_\xi(x) = \mathbb{1}(|x| \leq 1)/2$, and the measure $m_\xi(x)dx$ is not absolutely continuous with respect to the distribution of ξ_1.

The following lemma can be viewed as an extension of Stein's identity (cf. Lehmann and Casella 1998).

Lemma 1 *Let $T_n(x, \mathbf{Y})$ be an estimator of $f(x)$ such that the mapping $\mathbf{Y} \mapsto T_n(\mathbf{Y}) \triangleq (T_n(x_1, \mathbf{Y}), \ldots, T_n(x_n, \mathbf{Y}))^\top$ is continuously differentiable and let us denote by $\partial_j T_n(x_i, \mathbf{Y})$ the partial derivative of the function $\mathbf{Y} \mapsto T_n(\mathbf{Y})$ with respect to the jth coordinate of $\mathbf{Y}$. If Assumption (A) and the following condition*

$$\begin{aligned} &\int_{\mathbb{R}} |y| \int_0^y |\partial_i T_n(x_i, f + \mathbf{z})| dz_i dF_\xi(y) < \infty, \quad i = 1, \ldots, n, \quad \text{or} \\ &\partial_i T_n(x_i, \mathbf{Y}) \geq 0, \quad \forall \mathbf{Y} \in \mathbb{R}^n,\ i = 1, \ldots, n, \end{aligned} \tag{5}$$

are satisfied where $\mathbf{z} = (z_1, \ldots, z_n)^\top$, $f = (f(x_1), \ldots, f(x_n))^\top$ then

$$E[\hat{r}_n(\mathbf{Y})] = E(\|T_n(\mathbf{Y}) - f\|_n^2),$$

where

$$\hat{r}_n(\mathbf{Y}) = \|T_n(\mathbf{Y}) - \mathbf{Y}\|_n^2 + \frac{2}{n} \sum_{i=1}^n \partial_i T_n(x_i, \mathbf{Y}) g_\xi(\xi_i) - \sigma^2.$$

Proof We have

$$\begin{aligned} E(\|T_n(\mathbf{Y}) - f\|_n^2) &= E\left[\|T_n(\mathbf{Y}) - \mathbf{Y}\|_n^2 + \frac{2}{n} \sum_{i=1}^n \xi_i (T_n(x_i, \mathbf{Y}) - f(x_i)) \right] - \sigma^2 \\ &= E\left[\|T_n(\mathbf{Y}) - \mathbf{Y}\|_n^2 + \frac{2}{n} \sum_{i=1}^n \xi_i T_n(x_i, \mathbf{Y}) \right] - \sigma^2. \end{aligned} \tag{6}$$

For $\mathbf{z} = (z_1, \ldots, z_n)^\top \in \mathbb{R}^n$ write $F_{\xi,i}(\mathbf{z}) = \prod_{j \neq i} F_\xi(z_j)$. Since $E(\xi_i) = 0$ we have

$$\begin{aligned} E[\xi_i T_n(x_i, \mathbf{Y})] &= E\left[\xi_i \int_0^{\xi_i} \partial_i T_n(x_i, Y_1, \ldots, Y_{i-1}, f(x_i) + z, Y_{i+1}, \ldots, Y_n) dz \right] \\ &= \int_{\mathbb{R}^{n-1}} \left[\int_{\mathbb{R}} y \int_0^y \partial_i T_n(x_i, f + \mathbf{z}) dz_i dF_\xi(y) \right] dF_{\xi,i}(\mathbf{z}). \end{aligned} \tag{7}$$

Condition (5) allows us to apply the Fubini theorem to the expression in squared brackets on the right hand side of the last display. Thus, using the definition of m_ξ and Assumption (A) we find

$$\begin{aligned}\int_{\mathbb{R}_+} y \int_0^y \partial_i T_n(x_i, f+\mathbf{z}) dz_i dF_\xi(y) &= \int_{\mathbb{R}_+} \int_{z_i}^\infty y dF_\xi(y) \partial_i T_n(x_i, f+\mathbf{z}) dz_i \\ &= \int_{\mathbb{R}_+} m_\xi(z_i) \partial_i T_n(x_i, f+\mathbf{z}) dz_i \\ &= \int_{\mathbb{R}_+} g_\xi(z_i) \partial_i T_n(x_i, f+\mathbf{z}) dF_\xi(z_i).\end{aligned}$$

A similar equality holds for the integral over $\mathbb{R}_-$. Thus, in view of (7), we obtain

$$E[\xi_i T_n(x_i, \mathbf{Y})] = E[\partial_i T_n(x_i, \mathbf{Y}) g_\xi(\xi_i)].$$

Combining the last display with (6) we get the lemma. □

Based on Lemma 1 we obtain the following bound on the risk of the exponentially weighted aggregate $\hat{f}_n$.

Theorem 1 *Let π be an element of $\mathcal{P}_\Lambda$ such that, for all $\mathbf{Y}' \in \mathbb{R}^n$ and $\beta > 0$, the mappings $\lambda \mapsto \theta_\lambda(\mathbf{Y}') f_\lambda^2(x_i)$, $i = 1, \dots, n$, are π-integrable. If Assumption (A) is fulfilled then the aggregate $\hat{f}_n$ defined by (2) with $\beta \geq 4\|g_\xi\|_\infty$ satisfies the inequality*

$$E(\|\hat{f}_n - f\|_n^2) \leq \int \|f_\lambda - f\|_n^2 p(d\lambda) + \frac{\beta \mathcal{K}(p, \pi)}{n}, \quad \forall p \in \mathcal{P}_\Lambda, \tag{8}$$

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

Proof We will now use Lemma 1 with $T_n = \hat{f}_n$. Accordingly, we write here $\hat{f}_n(x_i, \mathbf{Y})$ instead of $\hat{f}_n(x_i)$. Applying the dominated convergence theorem and taking into account the definition of $\theta_\lambda(\mathbf{Y})$ we easily find that the π-integrability of $\lambda \mapsto \theta_\lambda(\mathbf{Y}') f_\lambda^2(x_i)$ for all i, $\mathbf{Y}'$ implies that the mapping $\mathbf{Y} \mapsto \hat{f}_n(\mathbf{Y}) \triangleq (\hat{f}_n(x_1, \mathbf{Y}), \dots, \hat{f}_n(x_n, \mathbf{Y}))^\top$ is continuously differentiable. Simple algebra yields

$$\begin{aligned}\partial_i \hat{f}_n(x_i, \mathbf{Y}) &= \frac{2}{\beta} \left\{ \int_\Lambda f_\lambda^2(x_i) \theta_\lambda(\mathbf{Y}) \pi(d\lambda) - \hat{f}_n^2(x_i, \mathbf{Y}) \right\} \\ &= \frac{2}{\beta} \int_\Lambda (f_\lambda(x_i) - \hat{f}_n(x_i, \mathbf{Y}))^2 \theta_\lambda(\mathbf{Y}) \pi(d\lambda) \geq 0.\end{aligned}$$

Therefore, (5) is fulfilled for $T_n = \hat{f}_n$ and we can apply Lemma 1 which yields

$$E[\hat{r}_n(\mathbf{Y})] = E(\|\hat{f}_n(\mathbf{Y}) - f\|_n^2)$$

with

$$\hat{r}_n(\mathbf{Y}) = \|\hat{f}_n(\mathbf{Y}) - \mathbf{Y}\|_n^2 + \frac{2}{n} \sum_{i=1}^n \partial_i \hat{f}_n(x_i, \mathbf{Y}) g_\xi(\xi_i) - \sigma^2.$$

Since $\hat{f}_n(\mathbf{Y})$ is the expectation of f_λ w.r.t. the probability measure $\theta \cdot \pi$,

$$\|\hat{f}_n(\mathbf{Y}) - \mathbf{Y}\|_n^2 = \int_\Lambda \{\|f_\lambda - \mathbf{Y}\|_n^2 - \|f_\lambda - \hat{f}_n(\mathbf{Y})\|_n^2\}\theta_\lambda(\mathbf{Y})\pi(d\lambda).$$

Combining these results we get

$$\begin{aligned}\hat{r}_n(\mathbf{Y}) &= \int_\Lambda \left\{\|f_\lambda - \mathbf{Y}\|_n^2 - \sum_{i=1}^n \frac{(\beta - 4g_\xi(\xi_i))(f_\lambda(x_i) - \hat{f}_n(x_i, \mathbf{Y}))^2}{n\beta}\right\}\theta_\lambda(\mathbf{Y})\pi(d\lambda) - \sigma^2 \\ &\le \int_\Lambda \|f_\lambda - \mathbf{Y}\|_n^2 \theta_\lambda(\mathbf{Y})\pi(d\lambda) - \sigma^2,\end{aligned}$$

where we used that $\beta \ge 4\|g_\xi\|_\infty$. By definition of θ_λ,

$$-n\|f_\lambda - \mathbf{Y}\|_n^2 = \beta \log\theta_\lambda(\mathbf{Y}) + \beta \log\left[\int_\Lambda e^{-n\|\mathbf{Y} - f_w\|_n^2/\beta}\pi(dw)\right].$$

Integrating this equation over $\theta \cdot \pi$, using the fact that $\int_\Lambda \theta_\lambda(\mathbf{Y}) \log\theta_\lambda(\mathbf{Y})\pi(d\lambda) = \mathcal{K}(\theta \cdot \pi, \pi) \ge 0$ and convex duality argument (cf., e.g., Dembo and Zeitouni 1998, p. 264, or Catoni 2004, p. 160) we get

$$\begin{aligned}\hat{r}_n(\mathbf{Y}) &\le -\frac{\beta}{n}\log\left[\int_\Lambda e^{-n\|\mathbf{Y} - f_w\|_n^2/\beta}\pi(dw)\right] - \sigma^2 \\ &\le \int_\Lambda \|\mathbf{Y} - f_w\|_n^2 p(dw) + \frac{\beta\mathcal{K}(p, \pi)}{n} - \sigma^2\end{aligned}$$

for all $p \in \mathcal{P}_\Lambda$. Taking expectations in the last inequality we obtain (8). □

4 Risk bounds for n-divisible distributions of errors

In this section we present a second approach to prove sharp risk bounds of the form (8). The main idea of the proof consists in an artificial introduction of a "dummy" random vector $\boldsymbol{\zeta} \in \mathbb{R}^n$ independent of $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ and having the same type of distribution as $\boldsymbol{\xi}$. This approach will allow us to cover the class of distributions of ξ_i satisfying the following assumption.

(B) *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:*
(B1) *the random variable $\xi_1 + \zeta_1$ has the same distribution as $(1 + 1/n)\xi_1$,*
(B2) *the vectors $\boldsymbol{\zeta} = (\zeta_1, \dots, \zeta_n)$ and $\boldsymbol{\xi} = (\xi_1, \dots, \xi_n)$ are independent.*

If ξ_1 satisfies (B1), then we will say that its distribution is *n-divisible*.

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., $L_\zeta(t) = E(e^{t\zeta_1})$, $t \in \mathbb{R}$.

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

$$\begin{cases} e^{(\|f - \bar{f}_{\mu'}\|_n^2 - \|f - \bar{f}_\mu\|_n^2)/\beta} \prod_{i=1}^n L_\zeta\left(\frac{2(\bar{f}_\mu(x_i) - \bar{f}_{\mu'}(x_i))}{\beta}\right) \le \Psi_\beta(\mu, \mu'), \\ \mu \mapsto \Psi_\beta(\mu, \mu') \text{ is concave and continuous in the total} \\ \text{variation norm for any } \mu' \in \mathcal{P}'_\Lambda, \\ \Psi_\beta(\mu, \mu) = 1, \end{cases} \tag{9}$$

for any $\beta \ge \beta_0$.

We now discuss some sufficient conditions for assumptions (B) and (C). Denote by $\mathcal{D}_n$ the set of all probability distributions of ξ_1 satisfying assumption (B1). First, it is easy to see that all the zero-mean Gaussian or double exponential distributions belong to $\mathcal{D}_n$. Furthermore, $\mathcal{D}_n$ contains all the stable distributions. However, since the non-Gaussian stable distributions do not have second order moments, they do not satisfy (9). One can also check that the convolution of two distributions from $\mathcal{D}_n$ belongs to $\mathcal{D}_n$. Finally, note that the intersection $\mathcal{D} = \bigcap_{n\geq 1} \mathcal{D}_n$ is included in the set of all infinitely divisible distributions and is called the L-class (see Petrov 1995, 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 $\mathcal{D}_n$.

Assumption (C) can be readily checked when the moment generating function $L_\zeta(t)$ is locally sub-Gaussian, i.e., there exists a constant $c > 0$ such that the inequality $L_\zeta(t) \leq e^{ct^2}$ holds for sufficiently small values of t. Examples include all the zero-mean distributions with bounded support, the Gaussian and double-exponential distributions, etc. The validity of Assumption (C) for such distributions follows from Lemma 4 in the Appendix.

Theorem 2 *Let π be an element of $\mathcal{P}_\Lambda$ such that, for all $\mathbf{Y}' \in \mathbb{R}^n$ and $\beta > 0$, the mappings $\lambda \mapsto \theta_\lambda(\mathbf{Y}') f_\lambda(x_i)$, $i = 1, \ldots, n$, are π-integrable. If assumptions (B) and (C) are fulfilled, then the aggregate $\hat{f}_n$ defined by (2) with $\beta \geq \beta_0$ satisfies the inequality*

$$E(\|\hat{f}_n - f\|_n^2) \leq \int \|f_\lambda - f\|_n^2 p(d\lambda) + \frac{\beta \mathcal{K}(p, \pi)}{n+1}, \quad \forall p \in \mathcal{P}_\Lambda. \tag{10}$$

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

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

For brevity, we will write

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

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

Since $E(\xi_i) = 0$, assumption (B1) implies that $E(\zeta_i) = 0$ for $i = 1, \ldots, n$. On the other hand, (B2) implies that $\boldsymbol{\zeta}$ is independent of θ_λ. Therefore, we have

$$E(\|\bar{f}_{\theta\cdot\pi} - f\|_n^2) = \beta E \log \exp\left\{ \frac{\|\bar{f}_{\theta\cdot\pi} - f\|_n^2 - 2\boldsymbol{\zeta}^\top \boldsymbol{H}_{\theta\cdot\pi}}{\beta} \right\} = S + S_1 \tag{11}$$

where

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

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

The definition of θ_λ yields

$$S = -\beta E \log \int_\Lambda \exp\left\{-\frac{n\|\mathbf{Y} - f_\lambda\|_n^2 + \|f_\lambda - f\|_n^2 - 2\boldsymbol{\zeta}^\top \boldsymbol{h}_\lambda}{\beta}\right\}\pi(d\lambda) + \beta E \log \int_\Lambda \exp\left\{-\frac{n\|\mathbf{Y} - f_\lambda\|_n^2}{\beta}\right\}\pi(d\lambda). \tag{12}$$

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

$$\begin{aligned} S &= -\beta E \log \int_\Lambda \exp\left\{-\frac{(n+1)\|f_\lambda - f\|_n^2 - 2(\boldsymbol{\xi} + \boldsymbol{\zeta})^\top \boldsymbol{h}_\lambda}{\beta}\right\}\pi(d\lambda) \\ &\quad + \beta E \log \int_\Lambda \exp\left\{-\frac{n\|f - f_\lambda\|_n^2 - 2\boldsymbol{\xi}^\top \boldsymbol{h}_\lambda}{\beta}\right\}\pi(d\lambda) \\ &= \beta E \log \int_\Lambda e^{-n\rho(\lambda)}\pi(d\lambda) - \beta E \log \int_\Lambda e^{-(n+1)\rho(\lambda)}\pi(d\lambda), \end{aligned} \tag{13}$$

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,

$$S \le -\frac{\beta}{n+1} E \log \int_\Lambda e^{-(n+1)\rho(\lambda)}\pi(d\lambda). \tag{14}$$

Assume now that $p \in \mathcal{P}_\Lambda$ 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 (14) 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 $i = 1, \ldots, n$, we have $\beta E(\rho(\lambda)) = \|f_\lambda - f\|_n^2$, and using the Fubini theorem we find

$$S \le \int_\Lambda \|f_\lambda - f\|_n^2 p(d\lambda) + \frac{\beta\mathcal{K}(p, \pi)}{n+1}. \tag{15}$$

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,

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

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 $\Psi_\beta(\theta\cdot\pi, \theta\cdot\pi) = 1$. □

Another way to read the results of Theorems 1 and 2 is that, if the "phantom" Gaussian error model (4) with variance taken larger than a certain threshold value is used to construct the Bayesian posterior mean $\hat{f}_n$, then $\hat{f}_n$ is close on the average to the best prediction under the true model, even when the true data generating distribution is non-Gaussian.

We now illustrate application of Theorem 2 by an example. Assume that the errors ξ_i are double exponential, that is the distribution of ξ_i admits a density with respect to the Lebesgue measure given by

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

Aggregation under this assumption is discussed in (Yang 2003) where it is recommended to modify the weights (3) matching them to the shape of f_ξ. For such a procedure (Yang 2003) proves an oracle inequality with leading constant which is greater than 1. The next proposition shows that sharp risk bounds (i.e., with leading constant 1) can be obtained without modifying the weights (3).

Proposition 1 *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* (10).

Proof We apply Theorem 2. The characteristic function of the double exponential density is $\varphi(t) = 2/(2 + \sigma^2 t^2)$. 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 $L_\zeta(t) = \varphi_\zeta(-\mathrm{i}t)$, which yields

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

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 (C). Let β be larger than $4\sigma(1+1/n)\bar{L}$. Then for all $\mu, \mu' \in \mathcal{P}_\Lambda$ we have

$$\frac{2\big|\bar{f}_\mu(x_i) - \bar{f}_{\mu'}(x_i)\big|}{\beta} \le \frac{4\bar{L}}{\beta} \le \frac{n}{(n+1)\sigma}, \quad i = 1, \dots, n,$$

and consequently

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

This implies that

$$\exp\left(\frac{\|f - \bar{f}_{\mu'}\|_n^2 - \|f - \bar{f}_\mu\|_n^2}{\beta}\right)\prod_{i=1}^n L_\zeta\left(\frac{2(\bar{f}_\mu(x_i) - \bar{f}_{\mu'}(x_i))}{\beta}\right) \le \Psi_\beta(\mu, \mu'),$$

where

$$\Psi_\beta(\mu, \mu') = \exp\left(\frac{\|f - \bar{f}_{\mu'}\|_n^2 - \|f - \bar{f}_\mu\|_n^2}{\beta} + \frac{4\sigma^2(2n+1)\|\bar{f}_\mu - \bar{f}_{\mu'}\|_n^2}{n\beta^2}\right).$$

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 $\beta \ge (8+4/n)\sigma^2 + 2\sup_\lambda \|f - f_\lambda\|_n^2$ by virtue of Lemma 4 in the Appendix. Therefore, assumption (C) is fulfilled and the desired result follows from Theorem 2. □

An argument similar to that of Proposition 1 can be used to deduce from Theorem 2 that if the random variables ξ_i are i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$, then inequality (10) holds for every $\beta \ge (4+2/n)\sigma^2 + 2L^2$ (cf. Dalalyan and Tsybakov 2007). However, in this Gaussian framework we can also apply Theorem 1 that gives better result: essentially the same inequality (the only difference is that the Kullback divergence is divided by n and not by $n+1$) holds for $\beta \ge 4\sigma^2$, with no assumption on the function f.

5 Model selection with finite or countable Λ

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

Theorem 3 *Let either assumptions of Theorem 1 or those of Theorem 2 be satisfied and let Λ be countable. Then for any $\beta \ge \beta_0$ the aggregate $\hat{f}_n$ satisfies the inequality*

$$E(\|\hat{f}_n - f\|_n^2) \le \inf_{j \ge 1}\left(\|f_j - f\|_n^2 + \frac{\beta \log \pi_j^{-1}}{n}\right)$$

where $\beta_0 = 4\|g_\xi\|_\infty$ when Theorem 1 is applied. In particular, if $\pi_j = 1/M$, $j = 1, \dots, M$, we have, for any $\beta \ge \beta_0$,

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

Proof For a fixed integer $j_0 \geq 1$ we apply Theorems 1 or 2 with p being the Dirac measure: $p(\lambda = j) = \mathbb{1}(j = j_0)$, $j \geq 1$. This gives

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

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. □

Theorem 3 generalizes the result of (Leung and Barron 2006) where the case of finite Λ and Gaussian errors ξ_i is treated. For this case it is known that the rate of convergence $(\log M)/n$ in (16) cannot be improved (Tsybakov 2003; Bunea et al. 2007a). Furthermore, for the examples (i)–(iii) of Sect. 3 (Gaussian or bounded errors) and finite Λ, inequality (16) is valid with no assumption on f and f_λ. Indeed, when Λ is finite the integrability conditions are automatically satisfied. Note that, for bounded errors ξ_i, oracle inequalities of the form (16) are also established in the theory of prediction of deterministic sequences (Vovk 1990; Littlestone and Warmuth 1994; Cesa-Bianchi et al. 1997; Kivinen and Warmuth 1999; Cesa-Bianchi and Lugosi 2006). However, those results require uniform boundedness not only of the errors ξ_i but also of the functions f and f_λ. What is more, the minimal allowed values of β in those works depend on an upper bound on f and f_λ which is not always available. The version of (16) based on Theorem 1 is free of such a dependence.

6 Risk bounds for general distributions of errors

As discussed above, assumption (B) restricts the application of Theorem 2 to models with n-divisible errors. We now show that this limitation can be dropped. The main idea of the proof of Theorem 2 was to introduce a dummy random vector $\boldsymbol{\zeta}$ independent of $\boldsymbol{\xi}$. However, the independence property is stronger than what we really need in the proof of Theorem 2. Below we come to a weaker condition invoking a version of the Skorokhod embedding (a detailed survey on this subject can be found in (Obloj 2004)).

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.

First, we describe a version of Skorokhod's construction that will be used below, cf. (Revuz and Yor 1999, Proposition II.3.8).

Lemma 2 *Let $\xi_1, \ldots, \xi_n$ be i.i.d. symmetric random variables on $(\Omega, \mathcal{F}, P)$. Then there exist i.i.d. random variables $\zeta_1, \ldots, \zeta_n$ 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) *$E(\zeta_i|\boldsymbol{\xi}) = 0$, $i = 1, \ldots, n$,*
(c) *for any $\lambda > 0$ and for any $i = 1, \ldots, n$, we have*

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

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

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

and Lemma 3 in the Appendix with $x = \lambda\xi_i/n$ and $\alpha_0 = 2n+2$. □

We now state the main result of this section.

Theorem 4 *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*

$$E(\|\hat{f}_n - f\|_n^2) \le \int_\Lambda \|f_\lambda - f\|_n^2 p(d\lambda) + \frac{\beta\mathcal{K}(p,\pi)}{n+1} + R_n, \quad \forall p \in \mathcal{P}_\Lambda, \tag{17}$$

where the residual term R_n is given by

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

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

Proof We slightly modify the proof of Theorem 2. We now consider a dummy random vector $\boldsymbol{\zeta} = (\zeta_1, \ldots, \zeta_n)$ as in Lemma 2. Note that for this $\boldsymbol{\zeta}$ relation (11) remains valid: in fact, it suffices to condition on $\boldsymbol{\xi}$, to use Lemma 2(b) and the fact that θ_λ is measurable with respect to $\boldsymbol{\xi}$. Therefore, with the notation of the proof of Theorem 2, we have $E(\|\hat{f}_n - f\|_n^2) = S + S_1$. Using Lemma 2(a) and acting exactly as in the proof of Theorem 2 we get that S is bounded as in (15). Finally, as shown in the proof of Theorem 2 the term S_1 satisfies

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

According to Lemma 2(c),

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

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

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

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

In view of Theorem 4, to get the bound (10) it suffices to show that the remainder term R_n is non-positive under some assumptions on the errors ξ_i. More generally, we may derive somewhat less accurate inequalities than (10) by proving that R_n is small enough. This is illustrated by the following corollaries.

Corollary 1 *Let the assumptions of Theorem 4 be satisfied and let $|\xi_i| \le B$ almost surely where B is a finite constant. Then the aggregate $\hat{f}_n$ satisfies inequality (10) 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 4 be satisfied and suppose that $E(e^{t|\xi_i|^\kappa}) \le B$ for some constants $t > 0$, $\kappa > 0$, $B > 0$. Then for any $n \ge e^{1/\kappa}$ and any $\beta \ge 4(1+1/n)(2(\log n)/t)^{2/\kappa} + 2L^2$ we have*

$$E(\|\hat{f}_n - f\|_n^2) \le \int_\Lambda \|f_\lambda - f\|_n^2 p(d\lambda) + \frac{\beta \mathcal{K}(p,\pi)}{n+1} + \frac{16BL^2(n+1)(2\log n)^{2/\kappa}}{n^2 \beta t^{2/\kappa}}, \quad \forall p \in \mathcal{P}_\Lambda. \tag{18}$$

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

$$E(\|\hat{f}_n - f\|_n^2) \le \min_{j=1,\dots,M} \|f_j - f\|_n^2 + \frac{\beta \log M}{n+1} + \frac{16BL^2(n+1)(2\log n)^{2/\kappa}}{n^2 \beta t^{2/\kappa}}. \tag{19}$$

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

$$\begin{aligned} R_n &\le \frac{4(n+1)}{n\beta} \sup_{\lambda\in\Lambda, \mu\in\mathcal{P}'_\Lambda} \|f_\lambda - \bar{f}_\mu\|_n^2 \sum_{i=1}^n E(\xi_i^2 - \alpha)_+ \\ &\le \frac{16L^2(n+1)}{\beta} E(\xi_1^2 - \alpha)_+ \end{aligned} \tag{20}$$

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

Corollary 3 *Assume that $\sup_{\lambda\in\Lambda} \|f - f_\lambda\|_\infty \le L$ and the errors ξ_i are symmetric with $E(|\xi_i|^s) \le B$ for some constants $s \ge 2$, $B > 0$. Then for any $\alpha_0 > 0$ and any $\beta \ge 4(1+1/n)\alpha_0 n^{2/(s+2)} + 2L^2$ we have*

$$E(\|\hat{f}_n - f\|_n^2) \le \int_\Lambda \|f_\lambda - f\|_n^2 p(d\lambda) + \frac{\beta \mathcal{K}(p,\pi)}{n+1} + \bar{C} n^{-s/(s+2)}, \quad \forall p \in \mathcal{P}_\Lambda,$$

where $\bar{C} > 0$ is a constant that depends only on s, L, B and α_0.

Proof Set $\alpha = \alpha_0 n^{2/(s+2)}$. In view of the inequality $(f_\lambda(x_i) - \bar{f}_{\theta\cdot\pi}(x_i))^2 \le 4\sup_{\lambda\in\Lambda} \|f - f_\lambda\|_\infty^2$, the remainder term of Theorem 4 can be bounded as follows:

$$R_n \le \frac{16L^2(n+1)}{n^2\beta} \sum_{i=1}^n E(\xi_i^2 - \alpha)_+ \le \frac{4L^2}{\alpha} E(\xi_1^2 - \alpha)_+.$$

To complete the proof, it suffices to notice that $E(\xi_1^2 - \alpha)_+ = E(\xi_1^2 \mathbb{1}(\xi_1^2 > \alpha)) \le E(|\xi_1|^s)/\alpha^{s/2-1}$ by the Markov inequality. □

Corollary 2 shows that if the tails of the distribution of errors have exponential decay and if β is of the order $(\log n)^{2/\kappa}$, then the rate of convergence in the bound (19) is of the order $(\log n)^{\frac{2}{\kappa}}(\log M)/n$. The residual R_n in Corollary 2 is of a smaller order than this rate and can be made even further smaller by taking $\alpha = (u(\log n)/t)^{2/\kappa}$ with $u > 2$. For $\kappa = 1$, comparing Corollary 2 with the risk bounds obtained in (Catoni 1999; Juditsky et al. 2008) for an averaged algorithm in i.i.d. random design regression, we see that an extra $\log n$ multiplier appears. It is noteworthy that this deterioration of the convergence rate does not occur if only the existence of finite (power) moments is assumed. In this case, the result of Corollary 3 provides the same rates of convergence as those obtained under the analogous moment conditions for model selection type aggregation in the i.i.d. case (cf. Juditsky et al. 2008; Audibert 2006).

7 Sparsity oracle inequalities with no assumption on the dictionary

In this section we assume that f_λ is a linear combination of M known functions $\phi_1, \dots, \phi_M$, where $\phi_j : \mathcal{X} \to \mathbb{R}$, with the vector of weights $\lambda = (\lambda_1, \dots, \lambda_M)$ that belongs to a subset Λ of $\mathbb{R}^M$:

$$f_\lambda = \sum_{j=1}^M \lambda_j \phi_j.$$

The set of functions $\{\phi_1, \dots, \phi_M\}$ is called the dictionary.

Our aim is to obtain sparsity oracle inequalities (SOI) for the aggregate with exponential weights $\hat{f}_n$. The SOI are oracle inequalities bounding the risk in terms of the number $M(\lambda)$ of non-zero components (sparsity index) of λ or similar characteristics. As discussed in Introduction, the SOI is a powerful tool allowing one to solve simultaneously several problems: sparse recovery in high-dimensional regression models, adaptive nonparametric regression estimation, linear, convex and model selection type aggregation.

For $\lambda \in \mathbb{R}^M$ denote by $J(\lambda)$ the set of indices j such that $\lambda_j \neq 0$, and set $M(\lambda) \triangleq \mathrm{Card}(J(\lambda))$. For any $\tau > 0$, $0 < L_0 \le \infty$, define the probability densities

$$q_0(t) = \frac{3}{2(1+|t|)^4}, \quad \forall t \in \mathbb{R}, \tag{21}$$

$$q(\lambda) = \frac{1}{C_0} \prod_{j=1}^M \tau^{-1} q_0\big(\lambda_j/\tau\big) \mathbb{1}(\|\lambda\| \le L_0), \quad \forall \lambda \in \mathbb{R}^M, \tag{22}$$

where $C_0 = C_0(\tau, M, L_0)$ is a normalizing constant such that q integrates to 1, and $\|\lambda\|$ stands for the Euclidean norm of $\lambda \in \mathbb{R}^M$.

In this section we choose the prior π in the definition of f_λ as a distribution on $\mathbb{R}^M$ with the Lebesgue density q: $\pi(d\lambda) = q(\lambda)d\lambda$. We will call it the *sparsity prior*.

Let us now discuss this choice of the prior. Assume for simplicity that $L_0 = \infty$ which implies $C_0 = 1$. Then the aggregate $\hat{f}_n$ based on the sparsity prior can be written in the form $\hat{f}_n = f_{\hat{\lambda}}$, where $\hat{\lambda} = (\hat{\lambda}_1, \dots, \hat{\lambda}_M)$ is the posterior mean in the "phantom" parametric model (4):

$$\hat{\lambda}_j = \int_{\mathbb{R}^M} \lambda_j \theta_n(\lambda) d\lambda, \quad j = 1, \dots, M,$$

with the posterior density

$$\theta_n(\lambda) = C \exp\big\{ -n\|\mathbf{Y} - f_\lambda\|_n^2/\beta + \log q(\lambda)\big\}$$
$$= C' \exp\left\{ -n\|\mathbf{Y} - f_\lambda\|_n^2/\beta - 4\sum_{j=1}^{M} \log(1 + |\lambda_j|/\tau) \right\}. \tag{23}$$

Here $C > 0, C' > 0$ are normalizing constants, such that $\theta_n(\cdot)$ integrates to 1. To compare our estimator with those based on the penalized least squares approach (BIC, Lasso, bridge), we consider now the posterior mode $\tilde{\lambda}$ of $\theta_n(\cdot)$ (the MAP estimator) instead of the posterior mean $\hat{\lambda}$. It is easy to see that $\tilde{\lambda}$ is also a penalized least squares estimator. In fact, it follows from (23) that the MAP estimator is a solution of the minimization problem

$$\tilde{\lambda} = \arg\min_{\lambda \in \mathbb{R}^M} \left\{ \|\mathbf{Y} - f_\lambda\|_n^2 + \frac{4\beta}{n} \sum_{j=1}^{M} \log(1 + |\lambda_j|/\tau) \right\}. \tag{24}$$

Thus, the MAP "approximation" of our estimator suggests that it can be heuristically associated with the penalty which is logarithmic in λ_j. In the sequel, we will choose τ very small (cf. Theorems 5 and 6 below). For such values of τ the function $\lambda_j \mapsto \log(1+|\lambda_j|/\tau)$ is very steep near the origin and can be viewed as a reasonable approximation for the BIC penalty function $\lambda_j \mapsto \mathbb{1}(\lambda_j \neq 0)$. The penalty $\log(1 + |\lambda_j|/\tau)$ is not convex in λ_j, so that the computation of the MAP estimator (24) is problematic, similarly to that of the BIC estimator. On the other hand, our posterior mean $\hat{f}_n$ is efficiently computable. Thus, the aggregate $\hat{f}_n$ with the sparsity prior can be viewed as a computationally feasible approximation to the logarithmically penalized least squares estimator or to the closely related BIC estimator. Interestingly, the results that we obtain below for the estimator $\hat{f}_n$ are valid under weaker conditions than the analogous results for the Lasso and Dantzig selector proved in (Bickel et al. 2007; Bunea et al. 2007b) and are sharper than those for the BIC (Bunea et al. 2007a) since we get oracle inequalities with leading constant 1 that are not available for the BIC.

Note that if we redefine q_0 as the double exponential density, the corresponding MAP estimator is nothing but the penalized least squares estimator with the Lasso penalty $\sim \sum_{j=1}^{M} |\lambda_j|$. More generally, if $q_0(t) \sim \exp(-|t|^\gamma)$ for some $0 < \gamma < 2$, the corresponding MAP solution is a bridge regression estimator, i.e., the penalized least squares estimator with penalty $\sim \sum_{j=1}^{M} |\lambda_j|^\gamma$ (Frank and Friedman 1993). The argument that we develop below can be easily adapted for these priors, but the resulting SOI are not as accurate as those that we obtain in Theorems 5 and 6 for the sparsity prior (21), (22). The reason is that the remainder term of the SOI is logarithmic in λ_j when the sparsity prior is used, whereas it increases polynomially in λ_j for the above mentioned priors.

We first prove a theorem that provides a general tool to derive the SOI from the PAC-Bayesian bound (8). Then we will use it to get the SOI in more particular contexts. Note that in this general theorem $\hat{f}_n$ is not necessarily an exponentially weighted aggregate defined by (2). It can be any $\hat{f}_n$ satisfying (8). The result of the theorem obviously extends to the case where a remainder term as R_n (cf. (17)) is added to the basic PAC-Bayesian bound (8).

Theorem 5 *Let $\hat{f}_n$ satisfy (8) 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*

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

where the residual term is

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

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

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

$$\int_\Lambda\|f_\lambda-f_{\lambda^*}\|_n^2p(d\lambda)\le\frac{1}{C_0C_{\lambda^*}}\sum_{j=1}^{M}\|\phi_j\|_n^2\int_{\mathbb{R}}\frac{w_j^2}{\tau}\,q_0\Big(\frac{w_j}{\tau}\Big)dw_j=\frac{\tau^2\sum_{j=1}^{M}\|\phi_j\|_n^2}{C_0C_{\lambda^*}}.$$

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

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

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:

$$\mathcal{K}(p,\pi)=\int_{\mathbb{R}^M}\log\Big(\frac{C_{\lambda^*}^{-1}q(\lambda-\lambda^*)}{q(\lambda)}\Big)p(d\lambda)\le4\sum_{j=1}^{M}\log(1+|\tau^{-1}\lambda_j^*|)-\log C_{\lambda^*}.$$

Easy computation yields $C_0\le1$. Therefore $C_{\lambda^*}\ge C_0C_{\lambda^*}\ge\exp(-\frac{2\tau^3M^{5/2}}{(\delta L_0)^3})$ and the desired result follows. □

Inspection of the proof of Theorem 5 shows that our choice of prior density q_0 in (21) is not the only possible one. Similar result can be readily obtained when $q_0(t)\sim|t|^{-3-\delta}$, as $|t|\to\infty$, for any $\delta>0$. The important point is that $q_0(t)$ should be symmetric, with finite second moment, and should decrease not faster than a polynomial, as $|t|\to\infty$.

We now explain how the result of Theorem 5 can be applied to improve the SOI existing in the literature. In our setup the values $x_1,\dots,x_n$ are deterministic. For this case, SOI for the BIC, Lasso and Dantzig selector are obtained in (Bunea et al. 2007a; Candes and

Tao 2007; Zhang and Huang 2008; Bickel et al. 2007). In those papers the random errors ξ_i are Gaussian. So, we will also focus on the Gaussian case, though similar corollaries of Theorem 5 are straightforward to obtain for other distributions of errors satisfying the assumptions of Sects. 3, 4 or 6.

Denote by Φ the Gram matrix associated with the family $(\phi_j)_{j=1,\dots,M}$, i.e., the $M \times M$ matrix with entries $\Phi_{j,j'} = n^{-1} \sum_{i=1}^n \phi_j(x_i)\phi_{j'}(x_i)$, $j, j' \in \{1, \dots, M\}$, and denote by $\mathrm{Tr}(\Phi)$ the trace of Φ. Set $\log_+ x = \max(\log x, 0)$, $\forall\, x > 0$.

Theorem 6 *Let $\hat{f}_n$ be defined by (2) with $\pi(d\lambda) = q(\lambda)d\lambda$ and $L_0 = \infty$. Let ξ_i be i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$ and assume that $\beta \geq 4\sigma^2$, $\mathrm{Tr}(\Phi) > 0$. Set $\tau = \frac{\sigma}{\sqrt{n\,\mathrm{Tr}(\Phi)}}$. Then for all $\lambda^* \in \mathbb{R}^M$ we have*

$$E(\|\hat{f}_n - f\|_n^2) \leq \|f_{\lambda^*} - f\|_n^2 + \frac{4\beta M(\lambda^*)}{n}\left(1 + \log_+\left\{\frac{\sqrt{n\,\mathrm{Tr}(\Phi)}}{M(\lambda^*)\sigma}|\lambda^*|_1\right\}\right) + \frac{\sigma^2}{n}$$

where $|\lambda^*|_1 = \sum_{j=1}^M |\lambda_j^*|$.

Proof To apply Theorem 5 with $L_0 = \infty$, we need to verify that $\hat{f}_n$ satisfies (8). This is indeed the case in view of Theorem 1. Thus we have

$$E(\|\hat{f}_n - f\|_n^2) \leq \|f_{\lambda^*} - f\|_n^2 + \frac{4\beta}{n} \sum_{j \in J(\lambda^*)} \log(1 + \tau^{-1}|\lambda_j^*|) + \tau^2\,\mathrm{Tr}(\Phi). \tag{25}$$

By Jensen's inequality, $\sum_{j\in J(\lambda^*)} \log(1 + \tau^{-1}|\lambda_j^*|) \leq M(\lambda^*)\log(1 + |\lambda^*|_1/(\tau M(\lambda^*)))$. Since $\log(1 + |\lambda^*|_1/(\tau M(\lambda^*))) \leq 1 + \log_+(|\lambda^*|_1/(\tau M(\lambda^*)))$, the result of the theorem follows from the choice of τ. □

Theorem 6 establishes a SOI *with leading constant 1* and *with no assumption on the dictionary*. Of course, for the inequality to be meaningful, we need a mild condition on the dictionary: $\mathrm{Tr}(\Phi) < \infty$. But this is even weaker than the standard normalization assumption $\|\phi_j\|_n^2 = 1$, $j = 1, \dots, M$. Note that a BIC type aggregate also satisfies a SOI similar to that of Theorem 6 with no assumption on the dictionary (cf. Bunea et al. 2007a), but with leading constant greater than 1. However, it is well-known that the BIC is not computationally feasible, unless the dimension M is very small (say, $M = 20$ in the uppermost case), whereas our estimator can be efficiently computed for much larger M.

The oracle inequality of Theorem 6 can be compared with the analogous SOI obtained for the Lasso and Dantzig selector under deterministic design (Bunea et al. 2007a; Bickel et al. 2007). Similar oracle inequalities for the case of random design $x_1, \dots, x_n$ can be found in (Bunea et al. 2007b; van de Geer 2006; Koltchinskii 2006). All those results impose heavy restrictions on the dictionary in terms of the coherence introduced in (Donoho et al. 2006) or other analogous characteristics that limit the applicability of the corresponding SOI, see the discussion after Corollary 4 below.

We now turn to the problem of high-dimensional parametric linear regression, i.e., to the particular case of our setting when there exists $\lambda^* \in \mathbb{R}^M$ such that $f = f_{\lambda^*}$. This is the framework considered in (Candes and Tao 2007; Zhang and Huang 2008) and also covered as an example in (Bickel et al. 2007). In these papers it was assumed that the basis functions are normalized: $\|\phi_j\|_n^2 = 1$, $j = 1, \dots, M$, and that some restrictive assumptions on the eigenvalues of the matrix Φ hold. We only impose a very mild condition: $\|\phi_j\|_n^2 \leq \phi_0$, $j = 1, \dots, M$, for some constant $\phi_0 < \infty$.

Corollary 4 *Let $\hat{f}_n$ be defined by (2) with $\pi(d\lambda) = q(\lambda)d\lambda$ and $L_0 = \infty$. Let ξ_i be i.i.d. Gaussian $\mathcal{N}(0, \sigma^2)$ random variables with $\sigma^2 > 0$ and assume that $\beta \geq 4\sigma^2$. Set $\tau = \frac{\sigma}{\sqrt{\phi_0 n M}}$. If there exists $\lambda^* \in \mathbb{R}^M$ such that $f = f_{\lambda^*}$ and $\|\phi_j\|_n^2 \leq \phi_0$, $j = 1, \ldots, M$, for some $\phi_0 < \infty$, we have*

$$E(\|\hat{f}_n - f\|_n^2) \leq \frac{4\beta}{n} M(\lambda^*) \left(1 + \log_+ \left\{ \frac{\sqrt{\phi_0 n M}}{M(\lambda^*)\sigma} |\lambda^*|_1 \right\}\right) + \frac{\sigma^2}{n}. \tag{26}$$

Proof is based on the fact that $\mathrm{Tr}(\Phi) = \sum_{j=1}^M \|\phi_j\|_n^2 \leq M\phi_0$ in (25).

Under the assumptions of Corollary 4, the rate of convergence of $\hat{f}_n$ is of the order $O(M(\lambda^*)/n)$, up to a logarithmic factor. This illustrates the sparsity property of the exponentially weighted aggregate $\hat{f}_n$: if the (unknown) number of non-zero components $M(\lambda^*)$ of the true parameter vector λ^* is much smaller than the sample size n, the estimator $\hat{f}_n$ is close to the regression function f, even when the nominal dimension M of λ^* is much larger than n. In other words, $\hat{f}_n$ achieves approximately the same performance as the "oracle" ordinary least squares that knows the set $J(\lambda^*)$ of non-zero components of λ^*. Note that similar performance is proved for the Lasso and Dantzig selector (Bunea et al. 2007a; Candes and Tao 2007; Zhang and Huang 2008; Bickel et al. 2007), however the risk bounds analogous to (26) for these methods are of the form $O\big(M(\lambda^*)(\log M)/(\kappa_{n,M} n)\big)$, where $\kappa_{n,M}$ is a "restricted eigenvalue" of the matrix Φ which is assumed to be positive (see Bickel et al. 2007 for a detailed account). This kind of assumption is violated for many important dictionaries, such as the decision stumps, cf. Bickel et al. 2007, and when it is satisfied the eigenvalues $\kappa_{n,M}$ can be rather small. This indicates that the bounds for the Lasso and Dantzig selector can be quite inaccurate as compared to (26).

Appendix

Lemma 3 *For any $x \in \mathbb{R}$ and any $\alpha_0 > 0$, $x + \log(1 + \frac{1}{\alpha_0}(e^{-x\alpha_0} - 1)) \leq \frac{x^2\alpha_0}{2}$.*

Proof On the interval $(-\infty, 0]$, the function $x \mapsto x + \log(1 + \frac{1}{\alpha_0}(e^{-x\alpha_0} - 1))$ is increasing, therefore it is bounded by its value at 0, that is by 0. For positive values of x, we combine the inequalities $e^{-y} \leq 1 - y + y^2/2$ (with $y = x\alpha_0$) and $\log(1+y) \leq y$ (with $y = 1 + \frac{1}{\alpha_0}(e^{-x\alpha_0} - 1)$). □

Lemma 4 *For any $\beta \geq s^2/n + 2\sup_{\lambda\in\Lambda} \|f - f_\lambda\|_n^2$ and for every $\mu' \in \mathcal{P}'_\Lambda$, the function*

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

is concave.

Proof Consider first the case where $\mathrm{Card}(\Lambda) = m < \infty$. Then every element of $\mathcal{P}_\Lambda$ 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 $\mu \in \mathcal{P}_\Lambda$, 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 $\tilde{\mu} \in \mathcal{P}_\Lambda$. Clearly, $\mathbf{M}$ is a symmetric and positive-semidefinite matrix. Moreover,

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

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

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

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

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

$$G\left(\frac{\mu+\tilde{\mu}}{2}\right) \ge \frac{G(\mu) + G(\tilde{\mu})}{2}, \quad \forall \mu, \tilde{\mu} \in \mathcal{P}'_\Lambda. \tag{27}$$

Fix now arbitrary $\mu, \tilde{\mu} \in \mathcal{P}'_\Lambda$. 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, $\mathcal{P}'_{\tilde{\Lambda}} = \mathcal{P}_{\tilde{\Lambda}}$. According to the first part of the proof, the functional

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

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

$$\tilde{G}\left(\frac{\nu+\tilde{\nu}}{2}\right) \ge \frac{\tilde{G}(\nu) + \tilde{G}(\tilde{\nu})}{2}, \quad \forall \nu, \tilde{\nu} \in \mathcal{P}_{\tilde{\Lambda}}.$$

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 (27). This completes the proof of the lemma. □

References

Audibert, J.-Y. (2004). *Une approche PAC-bayésienne de la théorie statistique de l'apprentissage*. PhD thesis. University of Paris 6.

Audibert, J.-Y. (2006). A randomized online learning algorithm for better variance control. In *Lecture notes in artificial intelligence: Vol. 4005*. *Proceedings of the 19th annual conference on learning theory, COLT 2006* (pp. 392–407). Heidelberg: Springer.

Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2007, submitted). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*. http://arxiv:0801.1158.

Bunea, F., Nobel, A.B. (2005). *Sequential procedures for aggregating arbitrary estimators of a conditional mean*, Preprint Florida State University. ttp://www.stat.fsu.edu/~flori.

Bunea, F., Tsybakov, A. B., & Wegkamp, M. H. (2006). Aggregation and sparsity via ℓ_1-penalized least squares. In *Lecture notes in artificial intelligence: Vol. 4005*. *Proceedings of 19th annual conference on learning theory, COLT 2006* (pp. 379–391). Springer: Heidelberg.

Bunea, F., Tsybakov, A. B., & Wegkamp, M. H. (2007a). Aggregation for Gaussian regression. *Annals of Statistics*, *35*, 1674–1697.

Bunea, F., Tsybakov, A. B., & Wegkamp, M. H. (2007b). Sparsity oracle inequalities for the Lasso. *Electronic Journal of Statistics*, *1*, 169–194.

Candes, E., & Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n (with discussion). *Annals of Statistics*, *35*, 2313–2404.

Catoni, O. (1999). *"Universal" aggregation rules with exact bias bounds*. Preprint n.510, Laboratoire de Probabilités et Modèles Aléatoires, Universités Paris 6 and Paris 7. http://www.proba.jussieu.fr/mathdoc/preprints/index.html$#$1999.

Catoni, O., (2004). Statistical learning theory and stochastic optimization. In *Lecture notes in mathematics*. *Ecole d'été de Probabilités de Saint-Flour 2001*, New York: Springer.

Cesa-Bianchi, N., & Lugosi, G. (2006). *Prediction, learning, and games*. New York: Cambridge University Press.

Cesa-Bianchi, N., Freund, Y., Haussler, D., Helmbold, D. P., Shapire, R., & Warmuth, M. (1997). How to use expert advice. *Journal of the ACM*, *44*, 427–485.

Cesa-Bianchi, N., Conconi, A., & Gentile, G. (2004). On the generalization ability of on-line learning algorithms. *IEEE Transactions on Information Theory*, *50*, 2050–2057.

Dalalyan, A., & Tsybakov, A. B. (2007). Aggregation by exponential weighting and sharp oracle inequalities. In N. H. Bshouty & C. Gentile (Eds.), *Lecture notes in artificial intelligence: Vol. 4539*. *Proceedings of the 20th annual conference on learning theory (COLT 2007)* (pp. 97–111). Berlin: Springer.

Dembo, A., & Zeitouni, O. (1998). *Large deviations techniques and applications*. New York: Springer.

Donoho, D. L., Elad, M., & Temlyakov, V. (2006). Stable recovery of sparse overcomplete representations in the presence of noise. *IEEE Transactions on Information Theory*, *52*, 6–18.

Efron, B., Hastie, T., Johnstone, I., & Tibshirani, R. (2004). Least angle regression. *Annals of Statistics*, *32*, 407–451.

Greenshtein, E., & Ritov, Y. (2004). Persistency in high dimensional linear predictor-selection and the virtue of over-parametrization. *Bernoulli*, *10*, 971–988.

Frank, I. E., & Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, *35*, 109–148.

Juditsky, A. B., Nazin, A. V., Tsybakov, A. B., & Vayatis, N. (2005). Recursive aggregation of estimators via the Mirror Descent Algorithm with averaging. *Problems of Information Transmission*, *41*, 368–384.

Juditsky, A., Rigollet, P., & Tsybakov, A. (2008, to appear). Learning by mirror averaging. *Annals of Statistics*. https://hal.ccsd.cnrs.fr/ccsd-00014097.

Kivinen, J., & Warmuth, M. K. (1999). Averaging expert predictions. In H. U. Simon & P. Fischer (Eds.), *Lecture notes in artificial intelligence: Vol. 1572*. *Proceedings of the fourth European conference on computational learning theory* (pp. 153–167). Berlin: Springer.

Koltchinskii, V. (2006, submitted). *Sparsity in penalized empirical risk minimization*.

Lehmann, E. L., & Casella, G. (1998). *Theory of point estimation*. New York: Springer.

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

Littlestone, N., & Warmuth, M. K. (1994). The weighted majority algorithm. *Information and Computation*, *108*, 212–261.

Nemirovski, A. (2000). Topics in non-parametric statistics. In *Lecture notes in mathematics: Vol. 1738*. *Ecole d'été de probabilités de saint-flour*, 1998. New York: Springer.

Obloj, J. (2004). The Skorokhod embedding problem and its offspring. *Probability Surveys*, *1*, 321–392.

Petrov, V. V. (1995). *Limit theorems of probability theory*. Oxford: Clarendon.

Revuz, D., & Yor, M. (1999). *Continuous martingales and Brownian motion*. Berlin: Springer.

Tsybakov, A. B. (2003). Optimal rates of aggregation. In B. Schölkopf & M. Warmuth (Eds.), *Lecture notes in artificial intelligence: Vol. 2777*. *Computational learning theory and kernel machines* (pp. 303–313). Heidelberg: Springer.

Tsybakov, A. B. (2004). *Mathématiques et applications: Vol. 41. Introduction à l'estimation non-paramétrique*. Berlin: Springer.

Tsybakov, A. B. (2006). Comments on "Regularization in statistics", by P. Bickel and B. Li. *Test*, *15*, 303–310.

van de Geer, S. A. (2006). *High dimensional generalized linear models and the Lasso* (Research report No. 133). Seminar für Statistik, ETH, Zürich.

Vovk, V. (1990). Aggregating strategies. In *Proceedings of the 3rd annual workshop on computational learning theory, COLT1990* (pp. 371–386). San Mateo: Morgan Kaufmann.

Vovk, V. (2001). Competitive on-line statistics. *International Statistical Review*, *69*, 213–248.

Yang, Y. (2000). Combining different procedures for adaptive regression. *Journal of Multivariate Analysis*, *74*, 135–161.

Yang, Y. (2001). Adaptive regression by mixing. *Journal of the American Statistical Association*, *96*, 574–588.

Yang, Y. (2003). Regression with multiple candidate models: selecting or mixing? *Statistica Sinica*, *13*, 783–809.

Zhang, T. (2006a). From epsilon-entropy to KL-complexity: analysis of minimum information complexity density estimation. *Annals of Statistics*, *34*, 2180–2210.

Zhang, T. (2006b). Information theoretical upper and lower bounds for statistical estimation. *IEEE Transactions on Information Theory*, *52*, 1307–1321.

Zhang, C.-H., & Huang, J. (2008, to appear). Model-selection consistency of the Lasso in high-dimensional regression. *Annals of Statistics*.