Probab. Theory Relat. Fields **113**, 301–413 (1999)

**Probability Theory** and Related Fields


# Risk bounds for model selection via penalization

**Andrew Barron[1,⋆], Lucien Birgé[2,⋆⋆], Pascal Massart[3,⋆⋆⋆]**

[1] Department of Statistics, Yale University, P.O. Box 208290, New Haven, CT 06520-8290, USA. e-mail: barron@stat.yale.edu

[2] URA 1321 "Statistique et modèles aléatoires", Laboratoire de Probabilités, boîte 188, Université Paris VI, 4 Place Jussieu, F-75252 Paris Cedex 05, France. e-mail: lb@ccr.jussieu.fr

[3] URA 743 "Modélisation stochastique et Statistique", Bât. 425, Université Paris Sud, Campus d'Orsay, F-91405 Orsay Cedex, France. e-mail: massart@stats.matups.fr

Received: 7 July 1995 / Revised version: 1 November 1997

**Abstract.** Performance bounds for criteria for model selection are developed using recent theory for sieves. The model selection criteria are based on an empirical loss or contrast function with an added penalty term motivated by empirical process theory and roughly proportional to the number of parameters needed to describe the model divided by the number of observations. Most of our examples involve density or regression estimation settings and we focus on the problem of estimating the unknown density or regression function. We show that the quadratic risk of the *minimum penalized empirical contrast estimator* is bounded by an index of the accuracy of the sieve. This accuracy index quantifies the trade-off among the candidate models between the approximation error and parameter dimension relative to sample size.

If we choose a list of models which exhibit good approximation properties with respect to different classes of smoothness, the estimator can be simultaneously minimax rate optimal in each of those classes. This is what is usually called *adaptation*. The type of classes of smoothness in which one gets adaptation depends heavily on the list of models. If too many models are involved in order to get accurate approximation of many wide classes of functions simultaneously, it may happen that the estimator is only approx-

⋆ Work supported in part by the NSF grant ECS-9410760,
⋆⋆ and URA CNRS 1321 "Statistique et modèles aléatoires",
⋆⋆⋆ and URA CNRS 743 "Modélisation stochastique et Statistique".
*Key words and phrases*: Penalization – Model selection – Adaptive estimation – Empirical processes – Sieves – Minimum contrast estimators

imately adaptive (typically up to a slowly varying function of the sample size).

We shall provide various illustrations of our method such as penalized maximum likelihood, projection or least squares estimation. The models will involve commonly used finite dimensional expansions such as piecewise polynomials with fixed or variable knots, trigonometric polynomials, wavelets, neural nets and related nonlinear expansions defined by superposition of ridge functions.

*Mathematics subject classifications (1991):* Primary 62G05, 62G07; secondary 41A25

## Contents

## 1. Introduction

### 1.1. What is this paper about?

The purpose of this paper is to provide a general method for estimating an unknown function $s$ on the basis of $n$ observations and a finite or countable family of *models* $S_m$, $m \in \mathcal{M}_n$, using an *empirical model selection criterion*. Here, by "model" we have in mind any possible space of finite dimension $D_m$ (in a sense that will be made precise later on and includes the classical case where $S_m$ is linear). We do not mean that $s$ belongs to any of the models, although this might be the case. Therefore we shall always think of a model $S_m$ as an *approximate model* for the true $s$ with controlled complexity and this is the reason why we shall use alternatively the term *sieve* introduced by Grenander (1981) in connection with approximation theory.

For each model $S_m$ we build an estimator $\hat{s}_{m,n}$ which minimizes some *empirical contrast function* $\gamma_n$ over the set $S_m$. The precise nature of the sampling model will be discussed later. It suffices for now to think of regression and density estimation problems in which, for each candidate function $t$, the empirical contrast $\gamma_n(t)$ is, respectively, the empirical average squared error or $(1/n)$ times the minus logarithm of likelihood.

Denoting by $R_{m,n}(s) = \mathbb{E}_s[d^2(s, \hat{s}_{m,n})]$ the risk at $s$ of the estimator $\hat{s}_{m,n}$ (where $d$ denotes some convenient distance) an *ideal model* should minimize $R_{m,n}(s)$ when $m$ varies. Nevertheless, even if $s$ belongs to some $S_{m_0}$, this "true" model can be far from being "ideal" (in the preceding sense). Think of a polynomial fitting of a regression curve with 100 observations when the true $s$ is a polynomial of degree 50.

Since $s$ is unknown, one cannot determine such an ideal model exactly. Therefore one would like to find a *model selection procedure* $\hat{m}$, based on the data, such that the risk of the resulting estimator $\hat{s}_{\hat{m},n}$ is equal to the minimal risk $\inf_{m \in \mathcal{M}_n} R_{m,n}(s)$. This program is too ambitious and we shall content ourselves to consider, instead of the minimal risk, some accuracy index of the form

$$a_n(s) = \inf_{m \in \mathcal{M}_n} \left\{ d^2(s, S_m) + \mathrm{pen}_{m,n} \right\} = \inf_{m \in \mathcal{M}_n} \left\{ \inf_{t \in S_m} d^2(s, t) + \mathrm{pen}_{m,n} \right\}$$

which majorizes the minimal risk and to provide a model selection procedure $\hat{m}$ such that the risk of $\hat{s}_{\hat{m},n}$ achieves the accuracy index up to some constant independent of $n$ which means that

$$\mathbb{E}_s \left[ d^2(s, \hat{s}_{\hat{m},n}) \right] \leq C(s) a_n(s) \quad \text{for all } n \ . \tag{1.1}$$

The procedure $\hat{m}$ is defined by the minimization over $\mathcal{M}_n$ of the *penalized empirical contrast* $\{\gamma_n(\hat{s}_{m,n}) + \mathrm{pen}_{m,n}\}$. More precisely it follows from the analysis of Birgé and Massart (1998) that the risk $R_{m,n}(s)$ is typically of

order $d^2(s, S_m) + D_m/n$. The penalty term $\text{pen}_{m,n}$ then generally takes the form $\kappa L_m D_m/n$ where $\kappa$ is an absolute constant and $L_m \geq 1$ is a weight that satisfies a condition of the type

$$\sum_{m \in \mathcal{M}_n} \exp[-L_m D_m] \leq 1 \quad .$$

The penalty term takes into account both the difficulty to estimate within the model $S_m$ (role of $D_m$) and the additional noise due to the size of the list of models (role of $L_m$) and derives from exponential probability bounds for the empirical contrast. It follows from (1.1) and our choice of the penalty that, for any $s$,

$$\mathbb{E}_s \left[ d^2(s, \hat{s}_{\hat{m},n}) \right] \leq C(s) \inf_{m \in \mathcal{M}_n} \left\{ d^2(s, S_m) + \frac{\kappa L_m D_m}{n} \right\} \quad . \tag{1.2}$$

Although we emphasized the fact that $s$ need not belong to any $S_m$, the bound (1.2) also makes sense in the parametric case. More precisely, if one starts from a finite collection of models $\{S_m\}_{m \in \mathcal{M}}$ which does not depend on $n$ and fix $L_m = 1$ for all $m$, one finds, whenever $s$ belongs to some $S_{m_0}$, that the risk of $\hat{s}_{\hat{m},n}$ is of order $n^{-1}$ as expected for this parametric framework.

More generally, the bound (1.2) permits the reduction of the problem of investigation of the performance of the estimator (to within certain constant multipliers) to an investigation of the approximation capabilities of the sieves. Here we have in mind a variety of possible function classes and the accuracy index will be evaluated for each. Since it is not known to which subsets of functions the target $s$ belongs, it is a merit of the accuracy index and indeed a merit of the minimum penalized empirical contrast estimator $\hat{s}_{\hat{m},n}$ in many cases that the maximum of the accuracy index $a_n(s)$ on certain subclasses of functions is within a constant factor of the minimax optimal value for the risk on these subclasses. For typical choices of models, the target function $s$ is a cluster point, that is $d(s, S_m)$ tends to zero for some subsequence of models, and the accuracy index quantifies the rate of convergence in a way that is naturally tied to the dimension of the models and the sample size through the penalty term. As a consequence of the accuracy index, there exists many situations where model selection provides estimators $\hat{s}_{\hat{m},n}$ which are (at least approximately) simultaneously minimax over a family of classes of functions, usually balls with respect to the seminorms of the classical spaces of smooth functions. Such estimators are then called (approximately) *adaptive*. We shall now go further into details to describe our work and relate our results to the existing literature on model selection and adaptive estimation.

### 1.2. Model selection

Historically, one can consider that model selection begins with the works of Mallows (1973) and Akaike (1973) although classical $t$ or $F$ tests and Bayes tests were long used for model selection. Actually, Daniel and Wood (1971, p. 86) already mention the $C_p$ criterion for variable selection in regression as described by Mallows in a conference dating back to (1964). Our model selection criteria can be viewed as extensions of Mallows' and Akaike's. In order to describe the heuristics underlying Mallows' approach, and more generally model selection based on penalization, let us consider here a typical and historically meaningful example, namely model selection for linear regression with fixed design.

Let us consider observations $Y_1, \ldots, Y_n$ such that $Y_i = s(x_i) + W_i$ where the $W_i$'s are centered independent identically distributed variables with variance one and the $x_i$'s are deterministic values in some space $\mathcal{X}$. We want to estimate the function $s$ defined on $\mathcal{X}$ from the $Y_i$'s and measure the error of estimation in terms of the distance derived from the Euclidean norm $\|t\| = [n^{-1} \sum_{i=1}^n t(x_i)^2]^{1/2}$. We consider a family of linear models $\{S_m\}_{m \in \mathcal{M}_n}$ (finite dimensional spaces of functions on $\mathcal{X}$), each model $S_m$ being of dimension $D_m$. Let $s_m$ be the orthogonal projection of $s$ onto $S_m$ and $\hat{s}_{m,n}$ be the least squares estimator of $s$ relatively to $S_m$. The risk of $\hat{s}_{m,n}$ is equal to

$$\mathbb{E}_s \left[ \|\hat{s}_{m,n} - s\|^2 \right] = \|s - s_m\|^2 + D_m/n \ .$$

Since $\|s - s_m\|^2 = \|s\|^2 - \|s_m\|^2$, the ideal model is given by the minimization of $-\|s_m\|^2 + D_m/n + n^{-1} \sum_{i=1}^n Y_i^2$. Let us consider the normalized residual sum of squares $n^{-1} \sum_{i=1}^n Y_i^2 - \|\hat{s}_{m,n}\|^2$. Since $\|\hat{s}_{m,n}\|^2 - D_m/n$ is an unbiased estimator of $\|s_m\|^2$, an unbiased estimator of the ideal criterion to minimize is $n^{-1} \sum_{i=1}^n Y_i^2 - \|\hat{s}_{m,n}\|^2 + 2D_m/n$ which is precisely Mallows' $C_p$. If we set

$$\gamma_n(t) = \frac{1}{n} \sum_{i=1}^n [Y_i - t(x_i)]^2$$

we notice that $\hat{s}_{m,n}$ is the minimizer of $\gamma_n$ over $S_m$ and that $\gamma_n(\hat{s}_{m,n}) = n^{-1} \sum_{i=1}^n Y_i^2 - \|\hat{s}_{m,n}\|^2$. Therefore Mallows' $C_p$ is a minimum penalized empirical contrast criterion in our sense with $\text{pen}_{m,n} = 2D_m/n$. This procedure is expected to work when the variables $\|\hat{s}_{m,n}\|^2$ concentrate around their expectations uniformly with respect to $m$. This is not clear at all when the cardinality of $\mathcal{M}_n$ is large as compared to $n$. Since the practical use of Mallow's $C_p$ criterion is for a fixed sample size it is a natural question to wonder whether the criterion will work for a given value of the cardinality of $\mathcal{M}_n$ as a function of $n$.

This particular problem has been studied by Shibata (1981) for Gaussian errors and Li (1987) under suitable moment assumptions on the errors (see also Polyak and Tsybakov 1990 for sharper moment conditions in the Fourier case). One can in particular deduce from these works that if the family of models $\{S_m\}_{m\in\mathcal{M}_n}$ is nested and each model has a dimension bounded by $n$, the heuristics of Mallows $C_p$ is validated in the sense that the selected index $\hat{m}$ provides an estimator $\hat{s}_{\hat{m},n}$ such that asymptotically the risk $\mathbb{E}_s[\|s-\hat{s}_{\hat{m},n}\|^2]$ is equivalent to $\inf_{m\in\mathcal{M}_n}\mathbb{E}_s[\|s-\hat{s}_{m,n}\|^2]$. It is worth noticing that this asymptotic equivalence holds provided that $s$ does not belong to any of the $S_m$'s.

Apart from Mallows' $C_p$ classical empirical penalized criteria for model selection include AIC, BIC, and MDL criteria proposed by Akaike (1973), Schwarz (1978), and Rissanen (1978 and 1983), respectively. They differ from the structure of the penalties involved, which are based on asymptotic, Bayesian or information-theoretic considerations and concern various empirical criteria such as maximum likelihood and least squares.

For our approach to model selection, the penalty term is motivated solely on the basis of what sorts of statistical risk bounds we can obtain. This conceptual point of view has been previously developed by Barron and Cover (1991) in their attempt to provide a global approach to model selection. Using a class of discretized models Barron and Cover (1991) or Barron (1991) prove risk bounds for complexity regularization criteria which in some cases include AIC, BIC, and MDL. The work by Barron and Cover is for criteria that possess a minimum description length interpretation and the discretization reduces the choice to a countable set of candidate functions $t$ with penalty $L(t)/n$ satisfying $\sum_t 2^{-L(t)} \leq 1$ as required for lengths of uniquely decodable codes. There these authors developed an approximation index called the index of resolvability that is a precursor to our accuracy index $a_n(s)$ and they establish comparable risk bounds for Hellinger distance in density estimation. The main innovation here, as compared to Barron and Cover (1991), is that we do not require that the models should be discrete. This supposes a lot of additional work.

The technical approach in this paper is in the spirit of Vapnik (1982). His method of "empirical minimization of the risk" also heavily relies on an analysis of the behavior of an empirical contrast based on empirical process theory and his method of "structural minimization of the risk" is related to a model selection criterion which parallels ours. We use here the tools developed in Birgé and Massart (1998). This makes a difference between Vapnik's approach and ours both in the formulation of the empirical process conditions and techniques. In particular, the introduction of recent isoperimetric inequalities by Talagrand (1994 and 1996) in the case of projection estimators on linear spaces, which has proved its efficiency in Birgé

and Massart (1997) and more recently in Baraud (1997), allows to obtain, in some cases, precise numerical evaluations of the penalty terms and to justify, even from a non-asymptotic point of view, Mallows' $C_p$, relaxing some restrictions imposed by Shibata (1981) and Li (1987). However, in general, penalty terms that satisfy our conditions may be different from those which are used in the familiar criteria. For instance we might have to consider heavier penalty terms if necessary in order to take into account the complexity of the family $\mathscr{M}_n$.

As to the implementation of minimum penalized contrast procedures, to be honest, we feel that this paper is merely a starting point which does not directly provide practical devices. However it is already possible to make a few remarks about implementation. The numerical value of the penalty function can be fixed in some cases as mentioned above. Also, as shown in Birgé and Massart (1997), the minimization procedure, even if the number of models is large, can be rather simple in some particular cases of interest since it is partly explicitly solvable, leading for instance to threshold or related estimators.

### *1.3. Sieve methods and approximation theory*

Let us recall that, for a given sieve $S$ of dimension $D$, $d^2(s, S) + D/n$ typically represents the order of magnitude of the risk $R_n(s)$ of a minimum contrast estimator $\hat{s}_n$ measured by the mean integrated squared error between $s$ and $\hat{s}_n$. The terms $d^2(s, S)$ and $D/n$ correspond to the bias squared and variance components, respectively. Given some prior information on $s$ (for instance an upper bound for some smoothness norm) one can, from approximation theory, choose a family $\{S_m\}_{m\in\mathscr{M}_n}$ of finite dimensional sieves such that $s$ is a cluster point of their union. If we select a sieve $S_{m_n}$ in the family according to the presumed property of the target function, rather than adaptively selected on the basis of data, what we study would fall under the general heading of analysis of sieves for function estimation. The choice of $S_{m_n}$ is determined by a particular trade-off between the variance and an upper bound for the bias squared. This method can lead to minimax risk computations. For instance, let us assume that $s$ belongs to some Sobolev ball $\mathbb{S}_\theta$ where $\theta$ is some known parameter which characterizes this ball. Approximation theory provides privileged families of sieves like spaces of piecewise polynomials with fixed or variable knots or trigonometric polynomials or wavelet expansions with optimal approximation properties with respect to those balls. Such a suitable choice of the list of sieves $S_m, m \in \mathscr{M}_n$ can typically guarantee that for given $n$ and $\theta$ the minimax risk $R_n(\theta)$ satisfies

$$R_n(\theta) = \inf_{\tilde{s}_n} \sup_{s\in\mathbb{S}_\theta} \mathbb{E}_s\left[d^2(s, \tilde{s}_n)\right] \geq C_1(\theta) \inf_{m\in\mathscr{M}_n} \left[\sup_{s\in\mathbb{S}_\theta} d^2(s, S_m) + \frac{D_m}{n}\right] \tag{1.3}$$

where $\tilde{s}_n$ is an arbitrary estimator. Such inequalities can in general be obtained by combining results in approximation theory with classical lower bounds on the minimax risk available in various contexts (density estimation, regression, white noise). Some references, among many others, are Bretagnolle and Huber (1979), Ibragimov and Khas'minskii (1980 and 1981), Nemirovskii (1985), Birgé (1983 and 1986), Donoho and Johnstone (1998). Therefore if $m(n, \theta)$ is a value of $m$ which minimizes $\sup_{s\in\mathbb{S}_\theta} d^2(s, S_m)+D_m/n$, the resulting minimum contrast estimator on the sieve $S_{m(n,\theta)}$ is typically minimax (up to some constant independent of $n$) on $\mathbb{S}_\theta$. The rates of convergence for sieves methods, as introduced by Grenander (1981), have been studied by several authors: Cencov (1982), Grenander and Chow (1985), Cox (1988), Stone (1990 and 1994), Barron and Sheu (1991), Haussler (1992), McGaffrey and Gallant (1994), Shen and Wong (1994), and Van de Geer (1995).

The main drawback of the preceding approach is connected with the prior assumption on the unknown $s$ which is not attractive for practical use although those estimators are relevant for minimax risk computations. As a matter of fact, Stone pointed out that his own works on sieves methods (mainly devoted to splines) were first steps towards data driven methods of nonparametric estimation. More precisely he had in view to provide some theoretical justifications for MARS (see Friedman 1991). The mathematical analysis of sequences of finite-dimensional models is at the heart of the techniques that we put to use in our study of adaptive methods of model selection. The point here is that a mere control of the quadratic risk on each sieve is far from being sufficient for achieving our program, as described in Section 1.1. Much more will be needed here and we shall have to make use of the exponential inequalities for the fluctuations of an empirical contrast on a sieve established in Birgé and Massart (1998).

We wish to allow a general framework of sieves characterized by their metric dimension and approximation properties. The examples we study typically involve linear combinations of a family of basis functions $\{\varphi_\lambda\}_{\lambda\in\Lambda}$, which are parameterized by an index $\lambda$ that is either discrete or continuous valued. In the discrete index case we have in mind examples of models based on Fourier series, wavelets, polynomials and piecewise polynomials with a discrete set of knot locations. Here the issue is the adaptive selection of the number of terms including all terms up to some total or the issue may be which subset of terms provides approximately the best estimate. In the first case there is only one sieve of each dimension and in the second there may be exponentially many candidate models as a function of dimension. The choice of whether subsets are taken has an impact on what types of tradeoffs are possible between bias and variance and on what types of penalty terms are permitted. In both cases the penalty term will be proportional

to the number of terms in the models, but in the latter case there is an additional logarithmic penalty factor that is typically necessary to realize approximately the best subset among exponentially many choices without substantial overfit. In contrast the use of fixed sets of terms typically allows for a penalty term with no logarithmic factors, but as we shall quantify (in the absence of subset selection) there can be less ability to realize a small statistical risk.

In the continuous index case we have in mind flexible nonlinear models including neural nets, trigonometric models with estimated frequencies, piecewise linear "hinged hyperplane" models and other piecewise polynomials with continuously parameterized knot locations. In these cases we write $\phi_w$ instead of $\varphi_\lambda$ for the terms that are linearly combined, where $w$ is a continuous vector-valued parameter. Not surprisingly, if the terms $\phi_w$ depend smoothly on $w$, the behavior of these nonlinear models is comparable to what is achieved in the discretized index set case with subset selection. We find that these nonlinear models have metric dimension properties that we can bound, but they lack the homogeneity of metric dimension satisfied by linear models with a fixed set of terms. The effect is that once again logarithmic factors arise in the penalty term and in the risk bounds. The advantage due to parsimony of the nonlinear models or the subset selection models is made especially apparent in the case of inference of functions with a high input dimension. In high dimensions, the exponential number of terms in linear models without subset selection precludes their practical use.

### *1.4. From model selection to adaptation*

Let us now consider the possible connections between our approach and adaptive estimation from the minimax point of view. As a matter of fact the adaptive properties of nonparametric estimators obtained from discrete model selection were already pointed out and studied by Barron and Cover (1991) for a number of classes of functions including Sobolev classes of log-densities without prior knowledge of which orders of smoothness and which norm bounds are satisfied by the target function. To recover the Barron and Cover result as a special case of our general density estimation results, set each model here to be a single function in their countable list. Barron (1991) extended the discretized model approach to deal also with complexity regularization for least squares regression and other bounded loss functions and applied it to artificial neural network models (see Barron 1994) . Let us also mention that the present paper is a companion to the paper by two of us (Birgé and Massart 1997) which explores the role of adaptive estimation for projection estimators of densities using linear models. Applications are given there for wavelet estimation and connections are established with

thresholding of wavelet coefficients and cross-validation criteria. More recently, Yang and Barron (1998) have got some results similar to ours for the particular case of log-density models.

Let us now provide a mathematical content to what we mean here by adaptation. Given a family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ of sets of functions we recall that the minimax risk over $\mathbb{S}_\theta$ is given by

$$R_n(\theta) = \inf_{\tilde{s}_n} \sup_{s\in\mathbb{S}_\theta} \mathbb{E}_s\left[d^2(s,\tilde{s}_n)\right]$$

where $\tilde{s}_n$ is an arbitrary estimator. We shall call a sequence of estimators $(\tilde{s}_n)_{n\geq 1}$ *adaptive in the minimax sense* if for every $\theta \in \Theta$ there exists a constant $C(\theta)$ such that

$$\sup_{s\in\mathbb{S}_\theta} \mathbb{E}_s\left[d^2(s,\tilde{s}_n)\right] \leq C(\theta) R_n(\theta) \quad .$$

If, for instance, one wants to give a precise meaning to the problem of estimating a function $s$ of unknown smoothness, one can assume that $s$ belongs to one of a large collection of balls such as Sobolev balls of variable index of smoothness and radius. Our purpose is to point out the connection between model selection via penalization as described previously and adaptation in the minimax sense. Starting from (1.2) and assuming that $L_m = L$ for all $m$ and $n$ and that $C(s)$ is bounded by $C_2(\theta)$ uniformly for $s \in \mathbb{S}_\theta$, one derives that

$$\sup_{s\in\mathbb{S}_\theta} \mathbb{E}_s\left[d^2(s,\hat{s}_{\hat{m},n})\right] \leq C_3(\theta) \inf_{m\in\mathcal{M}_n}\left[\sup_{s\in\mathbb{S}_\theta} d^2(s,S_m) + \frac{D_m}{n}\right] \quad .$$

If the family $\{S_m\}_{m\in\mathcal{M}_n}$ has convenient approximation properties with respect to the family $\{S_\theta\}_{\theta\in\Theta}$ such that (1.3) holds, it will follow that $\hat{s}_{\hat{m},n}$ is adaptive with respect to the family $\{S_\theta\}_{\theta\in\Theta}$ in the minimax sense.

We shall actually devote a large part of the paper to the illustration of this principle on various examples. For most of the illustrations that we shall consider one can take either $L_m$ as a constant $L$ or as $\log n$. In the latter case we shall get adaptation up to a slowly varying function of $n$. Moreover, in the first case, we shall also discuss the precise dependency of the ratio $C_3(\theta)/C_1(\theta)$ with respect to $\theta$ and sometimes show that it is bounded independently of $\theta$.

There is a huge amount of recent literature devoted to adaptive estimation and we postpone to Section 5 a discussion about the connections between model selection and adaptive estimation including a comparison between our approach to adaptation and the already existing methods and results.

The structure of the paper is described in the Table of Contents. Let us only mention that Sections 4, 7 and 8 are clearly more technical and

can be skipped at first reading. A first and particularly simple illustration of what we want to do and of the ideas underlying our approach is given in Section 2 which provides a self-contained introduction to our method while Section 3 provides an overview of its application to various situations. Section 5 does not contain any new result but is devoted to some detailed discussion, based on the examples of Sections 2 and 3, about the connections between adaptation and model selection.

## 2. A glimpse of the essentials

In order to give an idea of the way our approach to minimum penalized empirical contrast estimation works, let us describe it in the simplest framework we know, namely Gaussian regression on a fixed design. Its simplicity allows us to give a short and self-contained proof of an upper bound involving the accuracy index, for the risk of penalized least squares estimators. The main issue here is to enlighten the connection between the concentration of measure phenomenon and the choice of the penalty function for model selection.

### *2.1. Model selection in a toy framework*

In the Gaussian regression framework we observe $n$ random variables

$$Y_i = s(x_i) + W_i$$

where the $x_i$'s are known and the $W_i$'s are independent identically distributed standard normal. Identifying any function $t$ defined on the set $\mathcal{X} = \{x_1, \ldots, x_n\}$ to a vector $t = (t_1, \ldots, t_n)^T \in \mathbb{R}^n$ by setting $t_i = t(x_i)$, we define a scalar product and a norm on $\mathbb{R}^n$ by

$$\langle t, u \rangle = \frac{1}{n} \sum_{i=1}^{n} t(x_i) u(x_i) \quad \text{and} \quad \|t\|^2 = \frac{1}{n} \sum_{i=1}^{n} t(x_i)^2 \ . \tag{2.1}$$

We introduce a countable family $\{S_m\}_{m \in \mathcal{M}_n}$ of linear models, $S_m$ being of dimension $D_m$ and for each $m$ we consider the least squares estimator $\hat{s}_m$ on $S_m$ which is a minimizer with respect to $t \in S_m$ of

$$\gamma_n(t) = \|t\|^2 - 2\langle Y, t \rangle \quad \text{where} \quad Y = (Y_1, \ldots, Y_n)^T \ .$$

Then we choose a prior family of weights $\{L_m\}_{m \in \mathcal{M}_n}$ with $L_m \geq 1$ for each $m$, such that

$$\sum_{m\in\mathcal{M}_n} \exp[-L_m D_m] \leq \Sigma < +\infty \ . \tag{2.2}$$

Our aim is to prove the following

**Theorem 1** *Let* $\mathrm{pen}(m)$ *be defined on* $\mathcal{M}_n$ *by* $\mathrm{pen}(m) = \kappa L_m D_m/n$ *for a suitable constant* $\kappa$ *and the weights* $L_m$ *satisfy* (2.2). *Let* $\hat{s}_m$ *be the minimizer of* $\gamma_n(t)$ *for* $t \in S_m$ *and* $\hat{s}_{\hat{m}}$ *be the minimizer among the family* $\{\hat{s}_m\}_{m\in\mathcal{M}_n}$ *of the penalized criterion* $\gamma_n(\hat{s}_m) + \mathrm{pen}(m)$. *Then* $\hat{s}_{\hat{m}}$ *satisfies*

$$\mathbb{E}_s\left[\|s - \hat{s}_{\hat{m}}\|^2\right] \leq \kappa' \inf_{m\in\mathcal{M}_n} \left\{d^2(s, S_m) + \mathrm{pen}(m)\right\} + \kappa''\Sigma n^{-1} \ , \tag{2.3}$$

*where* $d^2(s, S_m) = \inf_{t\in S_m} \|s - t\|^2$ *and* $\kappa', \kappa''$ *are numerical constants.*

*Remark*: The following proof uses $\kappa = 24$ leading to $\kappa' = 3$ and $\kappa'' = 32$, which is obviously far from optimal as follows from Li (1987) or Baraud (1997). The result actually holds, for instance, with $\kappa = 2$ as in Mallow's $C_p$ but a proof leading to better values of the constants would be longer, involve additional technicalities and also use more specific properties of the framework. Since we want here to give a short and intuitive proof, in the spirit of the subsequent results given in the paper for different frameworks, we prefer to sacrifice optimality to simplicity and readability and put the emphasis on the main ideas to be used in the sequel without the specific tricks which are required for optimizing the constants.

*Proof*: We start with the identity

$$\|t - s\|^2 = \gamma_n(t) + 2\langle W, t\rangle + \|s\|^2 \quad \text{where } W = (W_1, \ldots, W_n)^T$$

and notice that, by definition, for any given $m \in \mathcal{M}_n$

$$\gamma_n(\hat{s}_{\hat{m}}) + \mathrm{pen}(\hat{m}) \leq \gamma_n(s_m) + \mathrm{pen}(m)$$

where $s_m$ denotes the orthogonal projection of $s$ onto $S_m$. Combining these two formulas we get

$$\|s - \hat{s}_{\hat{m}}\|^2 \leq \|s - s_m\|^2 + \mathrm{pen}(m) - \mathrm{pen}(\hat{m}) + 2\langle W, (\hat{s}_{\hat{m}} - s_m)\rangle \ . \tag{2.4}$$

Let $m$ be fixed. Given some $m' \in \mathcal{M}_n$, we introduce the Gaussian process $\{Z(t)\}_{t\in S_{m'}}$ defined by

$$Z(t) = \frac{\langle W, (t - s_m)\rangle}{w(m', t)} \quad \text{where } w(m', t) = \|t - s\|^2 + \|s - s_m\|^2 + \frac{x_{m'}}{2n} \ ,$$

$x_{m'}$ being some positive number to be chosen later. As a consequence of Cirel'son, Ibragimov and Sudakov's inequality (see Cirel'son, Ibragimov and Sudakov 1976 and, for more details about Gaussian concentration inequalities, Ledoux 1996).

$$\mathbb{P}_s\left[\sup_{t\in S_{m'}} Z(t) \geq E + \lambda\right] \leq \exp\left[-\frac{\lambda^2}{2\sigma^2}\right] \quad \text{for any } \lambda > 0 \tag{2.5}$$

provided that $E \geq \mathbb{E}[\sup_{t\in S_{m'}} Z(t)]$ and $\sup_{t\in S_{m'}} \mathrm{Var}(Z(t)) \leq \sigma^2$. Let us first notice that

$$w(m', t) \geq \frac{1}{2}\left[\|t - s_m\|^2 + \frac{x_{m'}}{n}\right] \geq \|t - s_m\|\left(\frac{x_{m'}}{n}\right)^{1/2} \tag{2.6}$$

and that for any function $u$, $\mathrm{Var}(\langle W, u\rangle) = n^{-1}\|u\|^2$. Then $\mathrm{Var}(Z(t)) = n^{-1}\|t - s_m\|^2 w^{-2}(m', t)$ which immediately yields that we can take $\sigma^2 = x_{m'}^{-1}$ in (2.5). On the other hand, expanding $t - s_m$ on an orthonormal basis $(\psi_1, \ldots, \psi_N)$ of $S_m + S_{m'}$ with $N \leq D_m + D_{m'}$, one gets by Cauchy-Schwarz inequality that

$$Z^2(t) \leq \|t - s_m\|^2 w^{-2}(m', t) \sum_{j=1}^{N} \langle W, \psi_j\rangle^2$$

and it follows from (2.6) and Jensen's inequality that we can take $E = [(D_m + D_{m'})/x_{m'}]^{1/2}$ in (2.5). If $\lambda$ is given by $\lambda^2 = 2(x + L_{m'}D_{m'})/x_{m'}$ where $x$ is any positive number we derive that

$$\lambda + E \leq \sqrt{2}\left(\frac{D_m + D_{m'} + 2x + 2L_{m'}D_{m'}}{x_{m'}}\right)^{1/2} \leq \frac{1}{4}$$

if $x_{m'} = 32(D_m + 2x + 3L_{m'}D_{m'})$. It then follows that

$$\mathbb{P}_s\left[Z(\hat{s}_{m'}) \geq 1/4\right] \leq \mathbb{P}_s\left[\sup_{t\in S_{m'}} Z(t) \geq 1/4\right] \leq \exp(-L_{m'}D_{m'})\exp(-x)$$

and therefore summing up those inequalities with respect to $m'$ that

$$\mathbb{P}_s\left[\sup_{m'\in\mathcal{M}_n} \frac{\langle W, (\hat{s}_{m'} - s_m)\rangle}{w(m', \hat{s}_{m'})} \geq \frac{1}{4}\right] \leq \Sigma\exp(-x) \ . \tag{2.7}$$

This implies from the definitions of $w$ and $x_{m'}$ that except on a set of probability bounded by $\Sigma e^{-x}$

$$\begin{aligned} 4\langle W, (\hat{s}_{\hat{m}} - s_m)\rangle &\leq w(\hat{m}, \hat{s}_{\hat{m}}) \\ &\leq \|s - \hat{s}_{\hat{m}}\|^2 + \|s - s_m\|^2 \\ &\quad + 16n^{-1}(D_m + 2x + 3L_{\hat{m}}D_{\hat{m}}) \ . \end{aligned}$$

Coming back to (2.4), this implies that

$$\|s-\hat{s}_{\hat{m}}\|^2 \leq 3\|s-s_m\|^2+2\mathrm{pen}(m)-2\mathrm{pen}(\hat{m})+16n^{-1}(D_m+2x+3L_{\hat{m}}D_{\hat{m}}) \ .$$

The choice $\kappa = 24$ entails the cancellation of $\mathrm{pen}(\hat{m})$, showing that, since $L_m \geq 1$

$$\|s-\hat{s}_{\hat{m}}\|^2 \leq 3\|s-s_m\|^2 + (8/3)\mathrm{pen}(m) + 32n^{-1}x$$

apart from a set of probability bounded by $\Sigma \mathrm{e}^{-x}$. Setting

$$V = \left(\|s-\hat{s}_{\hat{m}}\|^2 - 3\|s-s_m\|^2 - (8/3)\mathrm{pen}(m)\right) \vee 0$$

we get

$$\mathbb{E}_s\left[\|s-\hat{s}_{\hat{m}}\|^2\right] \leq 3\|s-s_m\|^2 + (8/3)\mathrm{pen}(m) + \mathbb{E}_s[V]$$

and $\mathbb{P}_s[V \geq 32x/n] \leq \Sigma \exp(-x)$. Integrating with respect to $x$ implies that $\mathbb{E}_s[V] \leq 32\Sigma/n$ which yields (2.3) since $m$ is arbitrary. □

### 2.2. *Variable selection*

We want to provide here a typical application of Theorem 1. Let us assume that we are given some (large) orthonormal system $\{\varphi_1, \ldots, \varphi_N\}$ in $\mathbb{R}^n$ with respect to the norm (2.1). We want to get an estimate of $s$ of the form $\tilde{s} = \sum_{\lambda \in m} \hat{\beta}_\lambda \varphi_\lambda$ where $m$ is some suitable subset of $\{1, 2, \ldots, N\}$. Let us first recall that if $m$ is given, the projection estimator $\hat{s}_m$ over $S_m = \mathrm{Span}\{\varphi_\lambda \mid \lambda \in m\}$, which is the minimizer with respect to $t \in S_m$ of the criterion $\gamma_n(t)$, is given by

$$\hat{s}_m = \sum_{\lambda \in m} \hat{\beta}_\lambda \varphi_\lambda \quad \text{with} \quad \hat{\beta}_\lambda = \langle Y, \varphi_\lambda \rangle$$

and that $\gamma_n(\hat{s}_m) = -\sum_{\lambda \in m} \hat{\beta}_\lambda^2$. Elementary computations show that

$$\mathbb{E}_s\left[\|s-\hat{s}_m\|^2\right] = d^2(s, S_m) + |m|/n \ .$$

Unfortunately, since $s$ is unknown we do not know how to choose $m$ in an optimal way in order to minimize $d^2(s, S_m) + |m|/n$. In order to select $m$ from the data, let us describe two simple strategies (among many others).

i) *Ordered variable selection.* In this case we select the "variables" $\varphi_\lambda$ in natural order which means that we restrict ourselves to $m_k = \{\varphi_\lambda \mid 1 \leq \lambda \leq k\}$, letting $k$ vary from 1 to $N$. In such a case one can take $L_m = 1$, $\Sigma = 0.6$, $\mathrm{pen}(m_k) = \kappa k/n$ and get a penalized least squares estimator $\hat{s}_{\hat{k}}$ where

$\hat{k}$ is the minimizer of $\kappa k/n - \sum_{\lambda \in m_k} \hat{\beta}_\lambda^2$. By Theorem 1, the risk of this estimator is bounded by

$$\mathbb{E}_s \left[ \|s - \hat{s}_{\hat{k}}\|^2 \right] \leq \bar{\kappa} \inf_{1 \leq k \leq N} \left\{ d^2(s, S_{m_k}) + k/n \right\}$$

for a suitable numerical constant $\bar{\kappa}$. One should notice here that $N$ does not enter the bound and can therefore be infinite and that we get the optimal risk among our family apart from the constant factor $\bar{\kappa}$. Note that this optimality is with respect to the best that can be achieved among the class of ordered variable selection models.

ii) *Complete variable selection.* Here we take $m$ to be any nonvoid subset of $\{1, 2, \ldots, N\}$. Since the number of such subsets with a given cardinality $D$ is $\binom{N}{D} < (eN/D)^D$ by Lemma 6 one can choose $L_m = 1 + \log N$ for all $m$ and $\Sigma = 1.3$. The resulting value $\hat{m}$ is then obtained by minimizing $\kappa(1 + \log N)|m|/n - \sum_{\lambda \in m} \hat{\beta}_\lambda^2$. It is easily seen that this amounts to select the values of $\lambda$ such that $\hat{\beta}_\lambda^2 > \kappa(1 + \log N)/n$ which means that

$$\hat{m} = \left\{ \lambda \;\middle|\; |\hat{\beta}_\lambda| > \left[ \frac{\kappa(1 + \log N)}{n} \right]^{1/2} \right\} \quad .$$

Therefore $\hat{s}_{\hat{m}}$ is a threshold estimator as studied by Donoho and Johnstone (1994a). Moreover by Theorem 1, there exists a constant $\bar{\kappa}$ such that

$$\mathbb{E}_s \left[ \|s - \hat{s}_{\hat{m}}\|^2 \right] \leq \bar{\kappa} \inf_m \left\{ d^2(s, S_m) + |m|(\log N)/n \right\} \quad .$$

If $N$ is independent of $n$, we only loose a constant as compared to the ideal estimator; if $N$ grows as a power of $n$, we only loose a $\log n$ factor as compared to the optimal risk for the class of all subset models, as in Donoho and Johnstone (1994a). This is the price to pay for complete variable selection among a large family but what is gained can be vastly superior in the approximation versus dimension tradeoff in the risk.

**Conclusion:** The simplicity of treatment of the preceding example is mainly due to the fact that the centered empirical contrast $2\langle W, t \rangle$ is a Gaussian linear process, acting on a finite dimensional linear space. The same treatment could be applied as well to penalized projection estimation for the white noise setting. Unfortunately the treatment of other empirical contrast functions or of nonlinear models requires that several technical difficulties be overcome.

- If we set here $\ell_n(s, t) = \mathbb{E}_s[\gamma_n(t) - \gamma_n(s)]$, then $\ell_n(s, t) = \|s - t\|^2$. In a non-Gaussian framework, one has to deal with a general empirical contrast function $\gamma_n$ and the analogue of (2.4) becomes

$$\ell_n(s, \hat{s}_{\hat{m}}) \leq \ell_n(s, s_m) + \left[\gamma_n^0(s_m) - \gamma_n^0(\hat{s}_{\hat{m}})\right] + \text{pen}(m) - \text{pen}(\hat{m})$$

where $\gamma_n^0(t) = \gamma_n(t) - \mathbb{E}_s[\gamma_n(t)]$. Pure $\mathbb{L}_2$-assumptions are not enough to control the fluctuations of the centered empirical contrast (the bracketed term) involved in this inequality. This motivates the introduction of $\mathbb{L}_\infty$-type assumptions on our models in the next section. Moreover, the structure of the exponential bounds that we use is connected to Bernstein's inequality rather than a subgaussian type inequality. We also would like to point out the status of the distance $d$ which has to be closely connected to the empirical contrast and chosen not too small in order to provide an appropriate control of the fluctuations of $\gamma_n^0$ and not too large in order that $d^2(s, t)$ be controlled by $\ell_n(s, t)$.

- In the most favorable case of the projection density estimator on linear models, one can mimic the preceding proof, replacing the concentration inequality (2.5) by Cirel'son, Ibragimov and Sudakov by an inequality of Talagrand (1996). The point here is that the linearity of the model and of $\gamma_n^0(t)$ as a function of $t$ allows to use Cauchy-Schwarz inequality as we did before to control the expectation of the supremum of the process involved. This point of view is developed in Birgé and Massart (1997) for projection density estimation and Baraud (1997) for non-Gaussian regression.
- More generally, in the nonlinear context, one has to deal with suitable modifications of the entropy methods introduced by Dudley (1978) to build the required exponential inequalities. Such results are collected in Proposition 7 below which is mainly based on Theorem 5 and Proposition 3 of Birgé and Massart (1998). Moreover, in the case of maximum likelihood estimation, we have to modify the initial empirical process in order to keep its fluctuations under control at the price of additional difficulties to get an analogue of inequality (2.4).

## 3. Main results with some illustrations

### *3.1. The minimum penalized empirical contrast estimation method*

We wish to analyze various functional estimation problems (density estimation, regression estimation, ...) that we describe precisely below. A common statistical framework covering all these examples is as follows. We observe $n$ random variables, $Z_1, \ldots, Z_n$ which, in the context of this paper, are assumed to be independent. These variables are defined on some measurable space $(\Omega, \mathcal{A})$ and take their values on some measurable space $(\mathcal{Z}, \mathcal{U})$. The space $(\Omega, \mathcal{A})$ is equipped with a family of probabilities $\{\mathbb{P}_s\}_{s\in\mathcal{S}}$ where $\mathcal{S}$ is a subset of some $\mathbb{L}_2$-space, $\mathbb{L}_2(\mu)$. Note that both $\mu$ and $\mathcal{S}$ can

depend on $n$, the same being true for each probability $\mathbb{P}_s$ but we do not make this dependence appear in the notation for the sake of simplicity since those quantities will be fixed (independent of $n$) in most applications. We denote by $\mathbb{E}_s$ the expectation with respect to probability $\mathbb{P}_s$, by $\mathbb{P}_n$ the empirical distribution of the $Z_i$'s and by $\nu_n = \mathbb{P}_n - \mathbb{E}_s \circ \mathbb{P}_n$ the centered empirical measure. The space $\mathbb{L}_2(\mu)$ is equipped with the distance $d$ induced by the norm $\|\cdot\| = \|\cdot\|_2$. More generally for $1 \leq p \leq \infty$, the norm in $\mathbb{L}_p(\mu)$ is denoted by $\|\cdot\|_p$.

Let us now introduce the key elements and notions that we need in the sequel.

**Definition 1** *Given some subset $\mathcal{T}$ of $\mathbb{L}_2(\mu)$ containing $\mathcal{S}$, an empirical contrast function $\gamma_n$ on $\mathcal{T}$ is defined for all $t \in \mathcal{T}$ as the empirical mean $\gamma_n(t) = n^{-1} \sum_{i=1}^n \gamma(Z_i, t)$ where $\gamma$ is a function defined on $\mathcal{Z} \times \mathcal{T}$ which satisfies*

$$\mathbb{E}_s[\gamma_n(t)] \geq \mathbb{E}_s[\gamma_n(s)] \quad \textit{for all} \quad s \in \mathcal{S} \quad \textit{and} \quad t \in \mathcal{T} \ .$$

We then introduce a countable collection of subsets $S_m$ of $\mathcal{T}$ (*models*) indexed by $m \in \mathcal{M}_n$. These models play the role of approximating spaces (*sieves*) for the true unknown value $s$ of the parameter which might or might not be included in one of them. Typically, $S_m$ is a subset of a finite-dimensional linear space. In order to make the notations simple we shall assume that everything which depends on $m \in \mathcal{M}_n$ might depend on $n$ but we omit this second index. We then consider a *penalty function* $\mathrm{pen}(m)$ which is a positive function on $\mathcal{M}_n$. We shall see later how to define this penalty function in order to get a sensible estimator. Let $\varepsilon_n \geq 0$ be given, a *minimum penalized empirical contrast estimator* is defined as follows:

**Definition 2** *Given some nonnegative number $\varepsilon_n$, an empirical contrast function $\gamma_n$, a collection of models $\{S_m\}_{m \in \mathcal{M}_n}$ and a penalty function* $\mathrm{pen}(\cdot)$ *on $\mathcal{M}_n$, an* $\varepsilon_n$-minimum penalized contrast estimator *is any estimator $\hat{s}$ in $\cup_{m \in \mathcal{M}_n} S_m$ with $\hat{s} \in S_{\hat{m}}$ such that*

$$\gamma_n(\hat{s}) + \mathrm{pen}(\hat{m}) \leq \inf_{m \in \mathcal{M}_n} \left\{ \inf_{t \in S_m} \gamma_n(t) + \mathrm{pen}(m) \right\} + \varepsilon_n \ . \tag{3.1}$$

*If $\varepsilon_n = 0$ we speak of a* minimum penalized contrast estimator.

As usual, by estimator we mean a measurable mapping from $(\mathcal{Z}, \mathcal{U})^{\otimes n}$ to the metric space $(\mathcal{T}, d)$ endowed with its Borel $\sigma$-algebra. If we omit the measurability problems, such an estimator is always defined provided that $\varepsilon_n > 0$ but might not be unique. Nevertheless, the following results

do apply to any solution of (3.1). In order to simplify the presentation we shall assume throughout the paper that $\hat{s}$ is well-defined for $\varepsilon_n = 0$. It turns out from our proofs that the choice $\varepsilon_n = n^{-1}$ would lead to the same risk bounds as those provided in the theorems below for the case $\varepsilon_n = 0$.

Some classical examples of minimum contrast estimation methods follow.

### 3.1.1. Maximum likelihood density estimation

We observe $n$ independent identically distributed variables $Z_1, \ldots, Z_n$ of density $s^2$ with respect to $\mu$. We define $\mathscr{T}$ to be the set of nonnegative elements of norm 1 in $\mathbb{L}_2(\mu)$ (which means that their squares are probability densities) and take $\mathscr{S} \subset \mathscr{T}$. The choice of the function $\gamma(z, t) = -\log t(z)$ leads to *maximum penalized likelihood estimators*.

### 3.1.2. Projection estimators for density estimation

We assume that $\mu$ is a probability measure and that the unknown density of the i.i.d. observations $Z_1, \ldots, Z_n$ belongs to $\mathbb{L}_2(\mu)$. It can therefore be written $\mathbb{1} + s$ where $s$ is orthogonal to the constant function $\mathbb{1}$. We take for $\mathscr{T}$ the subspace of $\mathbb{L}_2(\mu)$ which is orthogonal to $\mathbb{1}$ and derive the empirical contrast from $\gamma(z, t) = \|t\|^2 - 2t(z)$, $\mathscr{S}$ being chosen as any subset of those $t \in \mathscr{T}$ such that $\mathbb{1} + t \geq 0$. If $S_m$ is a linear subspace of $\mathscr{T}$ with an orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$, minimizing $\gamma_n(t)$ over $S_m$ leads to the classical projection estimator $\hat{s}_m$ on $S_m$ given by

$$\hat{s}_m = \sum_{\lambda \in \Lambda_m} \hat{\beta}_\lambda \varphi_\lambda \quad \text{with} \quad \hat{\beta}_\lambda = \frac{1}{n} \sum_{i=1}^{n} \varphi_\lambda(Z_i) \ .$$

### 3.1.3. Classical least squares regression

Observations are pairs $(X_i, Y_i) = Z_i$ with $Y_i = s(X_i) + W_i$ and the variables $X_i$ and $W_i$ are all independent with respective distributions $R_i$ and $Q_i$ (independent of $s$) but not necessarily independent identically distributed since we want to include the fixed design regression in our framework. In this case $\mathscr{S} \subset \mathscr{T} = \mathbb{L}_2(\mu)$ where $\mu$ denotes the average distribution of the $X_i$'s: $\mu = n^{-1} \sum_{i=1}^{n} R_i$. This distribution actually depends on $n$ in the case of a fixed design but not in the case of a random design. We assume that the errors $W_i$ are centered and choose $\gamma(z, t) = [y - t(x)]^2$. The resulting estimator is a penalized least squares estimator.

### 3.1.4. Minimum-$\mathbb{L}_1$ regression

We use the same regression framework as before, now assuming that the $W_i$'s are centered at their median and define $\gamma(z, t) = |y - t(x)|$.

These frameworks and related empirical contrast functions have been described in greater detail in Birgé and Massart (1993) and Birgé and Massart (1998). We therefore refer the reader to these papers for more information.

## *3.2. Examples of models*

In all our results, the value pen$(m)$ of the penalty function is, in particular, connected with the number $D_m$ of parameters which are necessary to describe the elements of the model $S_m$. A general definition of $D_m$ will appear in Section 6 and we shall here content ourselves with the presentation of two cases which are known to be of practical interest.

### 3.2.1. Linear models

By a "linear model" we mean a subset $S_m$ of some finite-dimensional linear subspace $\bar{S}_m$ of $\mathbb{L}_2 \cap \mathbb{L}_\infty(\mu)$ with dimension $D_m$. In opposition with what happens for Gaussian situations like the Gaussian regression on fixed design and the white noise setting, the $\mathbb{L}_2$-structure of the models is not sufficient to guarantee a good behavior of the empirical contrast function $\gamma_n$, which is essential for our purpose as we shall see later. More is needed, specifically some connections between the $\mathbb{L}_2$- and $\mathbb{L}_\infty$-structures of the models. It is the aim of the two following indices (indeed relative to $\bar{S}_m$) to quantify such connections. Firstly we set

$$\Phi_m = \frac{1}{\sqrt{D_m}} \sup_{t \in \bar{S}_m \setminus \{0\}} \frac{\|t\|_\infty}{\|t\|} \tag{3.2}$$

and denote by $\mathscr{F}_m$ the set of all orthonormal bases of $\bar{S}_m$. For any finite set $\Lambda$ and any $\beta \in \mathbb{R}^\Lambda$, we define $|\beta|_\infty = \sup_{\lambda \in \Lambda} |\beta_\lambda|$ and $|\beta|_2^2 = \sum_{\lambda \in \Lambda} \beta_\lambda^2$. We then notice that for any orthonormal basis $\varphi = \{\varphi_\lambda\}_{\lambda \in \Lambda_m} \in \mathscr{F}_m$

$$\Phi_m = \frac{1}{\sqrt{D_m}} \sup_{\beta \neq 0} \frac{\| \sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda \|_\infty}{|\beta|_2} = \frac{1}{\sqrt{D_m}} \left\| \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2 \right\|_\infty^{1/2}. \tag{3.3}$$

The second equality in (3.3) comes from Lemma 1 of Birgé and Massart (1998). Secondly we define

$$\bar{r}_m = \frac{1}{\sqrt{D_m}} \inf_{\varphi \in \mathscr{T}_m} \left\{ \sup_{\beta \neq 0} \frac{\| \sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda \|_\infty}{|\beta|_\infty} \right\} \quad . \tag{3.4}$$

It follows from (3.3) and this definition that

$$\Phi_m \leq \bar{r}_m \leq \sqrt{D_m} \Phi_m \quad . \tag{3.5}$$

Let us now detail a few examples of linear models and bound their indices.

**Uniformly bounded basis:** If one can find an orthonormal system $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ such that $\|\varphi_\lambda\|_\infty \leq \Phi$ for all $\lambda \in \Lambda$, if the elements of $\mathscr{M}_n$ are subsets of $\Lambda$ and $\bar{S}_m$ is the linear span of $\{\varphi_\lambda\}_{\lambda \in m}$, then $\Phi_m \leq \Phi$ by (3.3). Choosing $\mathscr{M}_n$ as a countable family of subsets of the trigonometric basis in $\mathbb{L}_2([0, 2\pi], dx)$ provides a typical example of this type.

**Wavelet expansions:** Let us consider an orthonormal wavelet basis $\{\varphi_{j,k} \mid j \geq 0, k \in \mathbb{Z}^q\}$ of $\mathbb{L}_2(\mathbb{R}^q, dx)$ (see Meyer 1990 for details) with the following conventions: $\varphi_{0,k}$ are translates of the father wavelet and for $j \geq 1$, the $\varphi_{j,k}$'s are affine transforms of the mother wavelet. One will also assume that these wavelets are compactly supported and have continuous derivatives up to some order $r$. Let $t \in \mathbb{L}_2(\mathbb{R}^q, dx)$ be some function with compact support in $(0, A)^q$. Changing the indexation of the basis if necessary we can write the expansion of $t$ on the wavelet basis as:

$$t = \sum_{j \geq 0} \sum_{k=1}^{2^{jq} M} \beta_{j,k} \varphi_{j,k} \quad ,$$

where $M \geq 1$ is a finite integer depending on $A$ and the size of the wavelet's supports. For any $j \in \mathbb{N}$, we denote by $\Lambda(j)$ the set of indices $\{(j, k) \mid 1 \leq k \leq 2^{jq} M\}$. The relevant $\Lambda_m$'s will be subsets of the larger sets $\cup_{j=0}^{J} \Lambda(j)$ for finite values of $J$ and we shall denote by $J_m$ the smallest $J$ such that this inclusion is valid. It comes from Bernstein's inequality (see Meyer 1990, Chapter 2, Lemma 8) that $\bar{r}_m \leq C(2^{qJ_m}/D_m)^{1/2}$ for some constant $C$. In particular, for all $\Lambda_m$'s of the form $\cup_{j=0}^{J_m} \Lambda(j)$, $\bar{r}_m$ is uniformly bounded and so is $\Phi_m$. The most relevant applications of such expansions have been studied extensively in Birgé and Massart (1997).

We also want to deal with wavelet expansions on the interval $[0, 1]$. Since the general case involves technicalities which are quite irrelevant to the subject of this paper, we shall content ourselves to deal with the simplest case of the Haar basis. Then the following expansion holds for any $t \in \mathbb{L}_2([0, 1], dx)$:

$$t = \beta_{-1,1}\varphi_{-1,1} + \sum_{j\geq 0}\sum_{k=1}^{2^j} \beta_{j,k}\varphi_{j,k} \quad , \tag{3.6}$$

where $\varphi_{-1,1} = \mathbb{1}_{[0,1]}$, $\psi = \mathbb{1}_{[0,1/2]} - \mathbb{1}_{]1/2,1]}$ and $\varphi_{j,k}(x) = 2^{j/2}\psi[2^j x - k + 1]$. We set $\Lambda(-1) = \{(-1, 1)\}$ and for $j \geq 0$ $\Lambda(j) = \{(j, k) \mid 1 \leq k \leq 2^j\}$. If $\Lambda_m = \cup_{j=0}^m \Lambda(j)$ we see from (3.3) that $\Phi_m = 1$. To bound $\bar{r}_m$ we first notice that for $j \geq 0$

$$\left\| \sum_{k=1}^{2^j} \beta_{j,k}\varphi_{j,k} \right\|_\infty \leq 2^{j/2} \sup_k |\beta_{j,k}| \quad . \tag{3.7}$$

Therefore

$$\bar{r}_m \leq \left[\sum_{j=0}^m 2^{j/2}\right]\left[\sum_{j=0}^m 2^j\right]^{-1/2} < 1 + \sqrt{2} \quad .$$

It may also be useful to choose $\Lambda_m = \cup_{j=-1}^m \Lambda(j)$ and then $\bar{r}_m < 2 + \sqrt{2}$.

**Piecewise polynomials:** We restrict our attention to piecewise polynomial spaces on a bounded rectangle in $\mathbb{R}^q$, which, without loss of generality, we take to be $[0, 1]^q$. Hereafter we denote by $\mathscr{P}_i$ a partition of $[0, 1]$ into $D(i)$ intervals. A linear space $\bar{S}_m$ of piecewise polynomials is characterized by $m = (r, \mathscr{P}_1, \ldots, \mathscr{P}_q)$ where $r$ is the maximal degree with respect to each variable of the polynomials involved. The elements $t$ of $\bar{S}_m$ are the functions on $[0, 1]^q$ which coincide with a polynomial of degree not greater than $r$ on each element of the product partition $\mathscr{P} = \otimes_{i=1}^q \mathscr{P}_i$. This results in $D_m = (r + 1)^q \prod_{i=1}^q D(i)$.

Let $\{Q_j\}_{j\in\mathbb{N}}$ be the orthogonal basis of the Legendre polynomials in $\mathbb{L}_2([-1, 1], dx)$, then the following properties hold for all $j \in \mathbb{N}$ (see Whittaker and Watson 1927, pp. 302–305 for details):

$$|Q_j(x)| \leq 1 \quad \text{for all} \quad x \in [-1, 1], \quad Q_j(1) = 1,$$

and

$$\int_{-1}^1 Q_j^2(t)\, dt = \frac{2}{2j+1} \quad .$$

Let us consider the hyperrectangle $R = \prod_{i=1}^q [a_i, b_i]$. For $j \in \mathscr{J} = \{0, \ldots, r\}^q$ we define

$$\varphi_{R,j}(x_1, \ldots, x_q) = \prod_{i=1}^q \left(\frac{2j_i + 1}{b_i - a_i}\right)^{1/2} Q_{j_i}\left(\frac{2x_i - a_i - b_i}{b_i - a_i}\right)\mathbb{1}_R(x_1, \ldots, x_q) \quad .$$

The family $\{\varphi_{R,j}\}_{j\in\mathcal{J}}$ provides an orthonormal basis for the space of polynomials on $R$ with degree bounded by $r$. If $H$ is a polynomial such that $H = \sum_{j\in\mathcal{J}} \beta_j \varphi_{R,j}$,

$$\|H\|_\infty \leq \left[(r+1)(2r+1)^{1/2}\right]^q [\mathrm{Vol}(R)]^{-1/2} |\beta|_\infty \ .$$

Then taking $\Lambda_m$ as the set of those $(R, j)$'s such that $R \in \mathcal{P}$ and $j \in \mathcal{J}$ we get from (3.4)

$$\bar{r}_m^2 \leq \frac{(r+1)^{2q}(2r+1)^q}{D_m \inf_{R\in\mathcal{P}} \mathrm{Vol}(R)} = [(r+1)(2r+1)]^q \left[\inf_{R\in\mathcal{P}} \mathrm{Vol}(R) \prod_{i=1}^{q} D(i)\right]^{-1} . \tag{3.8}$$

In particular, if $\mathcal{P}$ is a regular partition (all elements $R$ of $\mathcal{P}$ have the same volume),

$$\bar{r}_m \leq [(r+1)(2r+1)]^{q/2} \ . \tag{3.9}$$

**Polynomials on a sphere and other eigenspaces of the Laplacian:** Let $\mathbb{S}^q$ be the unit Euclidean sphere of $\mathbb{R}^{q+1}$, $\mu$ be the uniform distribution on the sphere and $0 < \theta_0 < \cdots < \theta_j < \cdots$ be the eigenvalues of the Laplace-Beltrami operator on $\mathbb{S}^q$. Let, for each $j \geq 0$, $\{\varphi_\lambda,\ \lambda \in \Lambda(j)\}$ be an orthonormal system of eigenfunctions associated with the eigenvalue $\theta_j$. Then $\{\mathbb{1}\} \cup \cup_{j\geq 0}\{\varphi_\lambda, \lambda \in \Lambda(j)\}$ is an orthonormal basis of $\mathbb{L}_2(\mu)$. Defining, for any integer $m \geq 0$, $\Lambda_m = \cup_{j=0}^m \Lambda(j)$ and $\bar{S}_m$ as the linear span of $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$, we get $D_m = |\Lambda_m|$, for $m \geq 0$. Actually these eigenvalues are given by explicit formulas (see for instance Berger, Gauduchon and Mazet 1971), the corresponding eigenfunctions are known to be harmonic zonal polynomials and one has (see Stein and Weiss 1971, p. 144)

$$\sum_{\lambda\in\Lambda(j)} \varphi_\lambda^2(x) \equiv |\Lambda(j)| \quad \text{for all } \ x \in \mathbb{S}^q \quad \text{and all } \ j \geq 0 \ .$$

In such a case it follows from (3.3) that $\Phi_m = 1$ for any integer $m$.

More generally, we can consider, instead of $\mathbb{S}^q$, a compact connected Riemannian manifold $\mathbb{M}$ of dimension $q$ with its uniform distribution $\mu$. The eigenfunctions of the Laplace-Beltrami operator provide an orthonormal basis of $\mathbb{L}_2(\mu)$ which is a multidimensional generalization of the Fourier basis. Of course no exact formula is available in this full generality but some asymptotic evaluation holds which is known as Weyl's formula (see Chavel 1984, p. 9). Keeping the same notations for the eigenvalues and eigenfunctions as above, defining $\Lambda(j)$, $\Lambda_m$ and $\bar{S}_m$ as in the case of the sphere and setting $D_{-1} = 1$, Weyl's formula ensures that there exists two positive constants $C_1(\mathbb{M})$ and $C_2(\mathbb{M})$ such that for any integer $m$

$$C_1(\mathbb{M})D_m^{2/q} \le \theta_m \le C_2(\mathbb{M})D_{m-1}^{2/q} < C_2(\mathbb{M})D_m^{2/q} \quad . \tag{3.10}$$

Moreover one can get the following control of the heat kernel (see Chavel 1984, inequality 55 p. 331):

$$\sum_{j=0}^{\infty}\Big[\mathrm{e}^{-\theta_j t}\sum_{\lambda\in\Lambda(j)}\varphi_\lambda^2(x)\Big] \le C_3(\mathbb{M})t^{-q/2} \tag{3.11}$$

for any positive $t$, any $x \in \mathbb{M}$ and some fixed positive constant $C_3(\mathbb{M})$. Applying (3.11) with $t = \theta_m^{-1}$ yields

$$\left\|\sum_{\lambda\in\Lambda_m}\varphi_\lambda^2\right\|_\infty \le eC_3(\mathbb{M})\theta_m^{q/2} \quad .$$

Combining this inequality with (3.10), we can derive from (3.3) that for any integer $m$, $\Phi_m^2 \le \Phi^2(\mathbb{M}) = eC_3(\mathbb{M})C_2(\mathbb{M})^{q/2}$ which implies that $\Phi_m$ is uniformly bounded as in the case of the sphere.

### 3.2.2. Nonlinear models

Here we have in mind a variety of models that include single hidden layer sigmoidal networks (see Barron 1993 and 1994), sparse trigonometric models, certain multivariate wavelet models as in Hornik et al. (1994) or Yukich et al. (1995) and piecewise linear "hinged hyperplane" models of Breiman (1993), for flexibly fitting a function of several variables. We take, for simplicity, the domain of the functions to be $[-1, 1]^q$. The models involve linear combinations of functions $\phi_w(x)$, continuously parameterized by a vector $w \in \mathbb{R}^{q'}$, where the functions $\phi_w$ satisfy the Lipschitz property

$$|\phi_w(x) - \phi_{w'}(x)| \le |w - w'|_1 \quad \text{for all } x \in [-1, 1]^q \quad , \tag{3.12}$$

$|\cdot|_1$ denoting the $l^1$-norm on $\mathbb{R}^{q'}$. The models $S_m$ are indexed by a triplet of positive integers $m = (D', H, R)$ and will be suitable modifications (via some clipping and renormalization) of the basic models

$$\bar{S}_m = \left\{ \sum_{j=1}^{D'} \beta_j \phi_{w_j}(x) \;\middle|\; \sum_{j=1}^{D'} |\beta_j| \le R \quad \text{and} \quad |w_j|_1 \le H \right.$$
$$\left. \text{for } 1 \le j \le D' \right\} \quad .$$

In such a case we can take $D_m = D'(q'+1)$, which is the parametric dimension of $\bar{S}_m$. Here the constraints $R$ and $H$ as well as $D'$ are included in the model index rather than fixed in advance, so that the metric entropy of each model can be controlled without advance knowledge of how large a value of $R$, $H$ or $D'$ is needed for the best model.

Of particular interest are the cases in which the terms in the model are $q$-dimensional ridge functions $\phi_w(x) = \psi(a^T x - b)$ where $\psi$ is a fixed univariate function with Lipschitz constant 1 and $w = (a, b)$ with $a \in \mathbb{R}^q$ and $b \in \mathbb{R}$ (then $q' = q + 1$). Then the Lipschitz property (3.12) holds for $\phi_w$. The cases mentioned above are of this ridge expansion form. For the neural net case $\psi$ is a sigmoidal function as in Barron (1993) (popular choices are the logistic, the hyperbolic tangent, and the linear ramp clipped at magnitude 1); for trigonometric sums $\psi$ is the cosine function and for the hinged hyperplane model $\psi(z) = z \vee 0$ to yield piecewise linear functions (see Breiman 1993). Hornik et al. (1994) and Yukich et al. (1995) take the activation function $\psi$ to be an arbitrary non-zero bounded function that is zero outside a bounded interval, which includes wavelet functions of ridge type. The Lipschitz condition used here holds for many (though not all) of these wavelets. A multivariate version of Proney's classic model can be developed with $\psi(z) = \mathrm{e}^{-z}$, where $z = a^T x + b$ is complex-valued with $a \in \mathbb{C}^q$, $b \in \mathbb{C}$, $x \in [0, 1]^q$ and all real parts of the coordinates of $a$ and $b$ taken to be nonnegative.

We are not restricted to ridge expansions here. For instance, radial basis function models with $\phi_w(x) = \psi(b|x - a|_1)$ are also of the required form when $\psi$ is a Lipschitz function such as $\psi(z) = \exp(-|z|)$ or $\exp(-z^2)$ and $b$ is bounded. This latter case leads to what Donoho calls the bump algebra. Tensor product expansions of the form $\phi_w(x) = \psi_{w_1}(x_1) \ldots \psi_{w_q}(x_q)$ for $x \in [-1, 1]^q$ satisfy the Lipschitz condition if the factors are built from a univariate Lipschitz function that is bounded by one (that is, $|\psi_{w_i}(x_i)| \leq 1$ and $|\psi_{w_i}(x_i) - \psi_{w_i'}(x_i)| \leq |w_i - w_i'|_1$ for $x_i \in [-1, 1]$). For instance, piecewise multilinear models correspond to $2\psi_{w_i}(x_i) = (x_i - w_i) \vee 0$ (with $w_i$ taken to be bounded by 1) as in the multivariate adaptive regression spline model of Friedman (1991).

Higher order piecewise polynomial ridge expansions and piecewise polynomial tensor products may also be handled with a slight modification of the framework, in which the linear combinations in $\bar{S}_m$ are built not just from one univariate $\psi$ function, but from several, such as $1$, $z$, $z^2$ and $(z \vee 0)^3$ in the cubic spline case. To simplify the discussion of the nonlinear models we have focussed attention on the case that $\phi$ is indexed by a continuous parameter rather than both discrete and continuous parameters. Multivariate piecewise polynomials will be explored as a subset selection problem using a grid rather than a continuum of possible knot locations in the next section.

### *3.3. The theorems and their applications*

In order to keep the presentation of our results simple, we now concentrate on linear models and return to the nonlinear models in Section 4.2. We assume that the situation described at the beginning of Section 3.2 holds, i.e. $S_m$ is a subset of a linear space $\bar{S}_m \subset \mathbb{L}_2(\mu)$ of dimension $D_m$ with $\Phi_m$ and $\bar{r}_m$ defined by (3.2) and (3.4) respectively. We shall also need, from now on, a number of different constants. Let us recall here that by "constant" we mean quantities that do not depend on $n$. In order to make our notations more transparent we shall hereafter systematically denote by the letter $\kappa$ as in $\kappa_1, \kappa', \ldots$ numerical constants, which do not depend on the various other constants involved in the assumptions. On the other hand, $C$ or $c$ denotes a constant depending on the former ones and possibly of $s$, the notation $C(\cdot, \cdots, \cdot)$ emphasizing the dependence of $C$ on the other constants. The same letter may be used for different constants from one section to another.

#### 3.3.1. Maximum likelihood estimators

We observe $n$ independent identically distributed variables $Z_1, \ldots, Z_n$ of density $s^2$ with respect to some *probability* measure $\mu$. The set of possible parameters $\mathcal{S}$ consists of those nonnegative functions $t$ for which $t^2$ is a probability density. To each $t \in \mathcal{S}$ corresponds a probability $P_t$ with density $t^2$ with respect to $\mu$ and $d(u, v)/\sqrt{2}$ is the Hellinger distance between the corresponding probabilities, i.e.

$$d^2(u, v) = \int \left( \sqrt{\frac{dP_u}{d\mu}} - \sqrt{\frac{dP_v}{d\mu}} \right)^2 d\mu \leq 2 \ .$$

We define analogously $K(u, v)$ to be the Kullback-Leibler information divergence between $P_u$ and $P_v$, i.e.

$$K(u, v) = \begin{cases} \int \log \left( \frac{dP_u}{dP_v} \right) dP_u & \text{if } P_u \ll P_v \ ; \\ +\infty \text{ otherwise} \ . \end{cases}$$

**Theorem 2** *Let $\{\bar{S}_m\}_{m \in \mathcal{M}_n}$ be a countable family of finite dimensional linear subspaces of $\mathbb{L}_2(\mu)$. For any $m \in \mathcal{M}_n$ we denote by $D_m$ the dimension of $\bar{S}_m$, by $\bar{r}_m$ the index defined by* (3.4) *and we set $S_m = \bar{S}_m \cap \mathcal{S}$. Let $\{L_m\}_{m \in \mathcal{M}_n}$ be a family of weights such that*

$$L_m \geq 1 \quad \textit{for all} \ \ m \in \mathcal{M}_n \quad \textit{and} \quad \sum_{m \in \mathcal{M}_n} \exp[-L_m D_m] \leq \Sigma < +\infty \ . \tag{3.13}$$

*Let* $\mathrm{pen}(m)$ *be such that*

$$\mathrm{pen}(m) \geq \kappa_1[L_m + \log(1 + \bar{r}_m)](D_m/n)$$

*where* $\kappa_1$ *is a suitable positive numerical constant and let* $\hat{s}$ *be the maximum penalized likelihood estimator which is a minimizer with respect to* $m \in \mathscr{M}_n$ *and* $t \in S_m$ *of* $-n^{-1}\sum_{i=1}^n \log[t(Z_i)] + \mathrm{pen}(m)$ *if* $t \in S_m$. *Define* $K(s, S_m) = \inf_{u \in S_m} K(s, u)$ *and assume that* $1 \leq D_m \leq n$ *for all* $m \in \mathscr{M}_n$. *Then whatever* $s \in \mathscr{S}$

$$\mathbb{E}_s\left[d^2(s, \hat{s})\right] \leq \kappa_1' \left[\inf_{m \in \mathscr{M}_n} \{(K(s, S_m) \wedge 1) + \mathrm{pen}(m)\} + \Sigma n^{-1}\right] . \quad (3.14)$$

The upper bound in (3.14) involves a bias term $K(s, S_m)$ where one would prefer $d^2(s, S_m)$. In many examples a natural way of deriving an approximation of $s$ by an element $s_m$ of $S_m$ is to normalize an upper approximation $s_m^+ \geq s$ in $\bar{S}_m$. More precisely we shall prove in Section 8 the following result:

**Proposition 1** *Assume that* $\bar{S}_m$ *is a linear space of functions in* $\mathbb{L}_2(\mu)$ *and* $S_m$ *is the set of nonnegative elements of norm* 1 *in* $\bar{S}_m$. *If there exists* $s_m^+ \geq s$ *in* $\bar{S}_m$ *then*

$$K(s, S_m) \wedge 1 \leq 3d^2(s, s_m^+)$$

*and if* $\mu$ *is a probability measure and* $\mathbb{1} \in \bar{S}_m$

$$K(s, S_m) \wedge 1 \leq 12 \inf_{t \in \bar{S}_m} \|s - t\|_\infty^2. \quad (3.15)$$

**Application to adaptive histograms:** We consider a family of sieves which are sets of piecewise polynomials of degree 0, i.e., histograms, on [0, 1] and take $\mu$ to be the Lebesgue measure. Here $m$ is a partition of [0, 1] which is a union of $D_m$ intervals, $\bar{S}_m$ is the space of piecewise constant functions on $m$ and $S_m$ is the set of nonnegative elements $t$ of $\bar{S}_m$ such that $\|t\| = 1$. Let $\mathscr{R}_n$ be the set of all regular partitions with at most $n$ pieces (this restriction being necessary since Theorem 2 requires that $D_m \leq n$) and $\mathscr{G}_{n,N}$ be the set of all irregular partitions with at most $n$ pieces and endpoints belonging to the grid $\{j/N \mid 0 \leq j \leq N\}$. Noticing that $\mathscr{G}_{n,N}$ is empty for $N = 1$ or 2, we define $\mathscr{M}_n = \mathscr{R}_n \cup (\cup_{N \geq 3} \mathscr{G}_{n,N})$ and choose $L_m = 1$ when $m \in \mathscr{R}_n$, $L_m = 2[1 + \log(N/D_m)]$ when $m \in \mathscr{G}_{n,N}$. It follows from our study of piecewise polynomials that $\bar{r}_m \leq 1$ by (3.9) when $m \in \mathscr{R}_n$ and is bounded by $(N/D_m)^{1/2}$ otherwise. One observes that the number of

partitions of $\mathcal{G}_{n,N}$ with $D$ pieces is bounded by $(eN/D)^D$ from Lemma 6 and that necessarily $1 < D < N$ (to avoid regular partitions). Therefore the following computations show that (3.13) is satisfied with $\Sigma = 1$:

$$\begin{aligned}
\sum_{m\in\mathcal{M}_n} \mathrm{e}^{-L_m D_m} &\le \sum_{j\ge 1} \mathrm{e}^{-j} + \sum_{N\ge 3}\sum_{j=2}^{N-1} \left(\frac{eN}{j}\right)^j \mathrm{e}^{-2j[1+\log(N/j)]} \\
&\le \frac{1}{e-1} + \sum_{N\ge 3}\sum_{j=2}^{N-1} \left(\frac{eN}{j}\right)^{-j} \\
&\le \frac{1}{e-1} + \sum_{j\ge 2} \left(\frac{e}{j}\right)^{-j} \sum_{N>j} N^{-j} \\
&\le \frac{1}{e-1} + \sum_{j\ge 2} \left(\frac{e}{j}\right)^{-j} \int_j^\infty x^{-j}\, dx \\
&= \frac{1}{e-1} + \sum_{j\ge 2} \frac{j}{j-1} \mathrm{e}^{-j} \quad .
\end{aligned}$$

Choosing $K \ge \kappa_1$, we can apply Theorem 2 with

$$\mathrm{pen}(m) = K(1+\log 2)(D_m/n) \quad \text{if } m \in \mathcal{R}_n \ , \tag{3.16}$$

$$\begin{aligned}
\mathrm{pen}(m) = K\left[2 + 2\log(N/D_m) + \log\left(1 + (N/D_m)^{1/2}\right)\right](D_m/n) \\
\text{if } m \notin \mathcal{R}_n \ .
\end{aligned} \tag{3.17}$$

- If $s$ is Hölderian of order $\alpha$, i.e.

$$|s(x) - s(y)| \le H|x-y|^\alpha \quad \text{for all } x, y \in [0,1] \ ,$$

  $H > 0$ and $\alpha \in (0,1]$ being unknown, for each $m \in \mathcal{R}_n$, the $\mathbb{L}_\infty$-distance between $s$ and $\bar{S}_m$ is bounded by $H(2D_m)^{-\alpha}$ and therefore by (3.15) $K(s, S_m) \wedge 1$ is bounded by $12H^2(2D_m)^{-2\alpha}$. In that case (3.14) implies that the quadratic risk of our estimator is bounded by $C(K)H^{2/(2\alpha+1)}n^{-2\alpha/(2\alpha+1)}$. We shall see in Section 4.1.2 that even if $H$ and $\alpha$ were known, one couldn't do better, from the minimax point of view, apart from the constant $C$.
- If $s$ belongs to some $S_m$ with $m \in \mathcal{G}_{n,N}$, (3.14) implies that the risk is bounded by $C'(K)\log(N/D_m)D_m/n$, which is of the usual parametric order $n^{-1}$ as $n \to \infty$ for each such $s$ and, for each given positive integer $l$, of order $(D/n)\log(n/D)$ uniformly in models with index in the set $\{m \in \cup_{N=3}^{2n^l} \mathcal{G}_{n,N} \mid D_m = D\}$. On the other hand for a given value of $D$, $9 \le$

$D \leq n/5$, it follows from Proposition 2 of Birgé and Massart (1998) that the minimax risk on this set is of the same order $(D/n)\log(n/D)$. This gives a sense in which the $\log n$ factor is a necessary price to pay when one compares the purely parametric problem of estimating a piecewise constant density on a known partition with $D$ pieces to the same problem with a completely unknown partition.

- The main advantage of including the families of irregular partitions in our construction is to allow *spatial adaptation*. With a single estimator we achieve simultaneously the optimal $n^{-1}$ rate for $s$ in the parametric subfamilies, the optimal rate $n^{-2\alpha/(2\alpha+1)}$ for the $\alpha$-Hölderian densities and within a logarithmic factor of this optimal rate for much less homogeneous functions with smoothness $\alpha$. This will be illustrated in Section 4.2.1 below for densities with bounded $\alpha$-variation.

3.3.2. Projection estimators

The basic result is similar to Theorem 3 of Birgé and Massart (1997) where a detailed study of some more specific examples involving Besov spaces is to be found. We recall that here $\mu$ is a probability measure and that the observations $Z_1, \ldots, Z_n$ have the same unknown density $\mathbb{1} + s$ with respect to $\mu$. Therefore the space $\mathscr{T}$ is chosen to be the linear subspace of $\mathbb{L}_2(\mu)$ orthogonal to $\mathbb{1}$.

**Theorem 3** *Assume that the family $\{S_m\}_{m\in\mathscr{M}_n}$ is a family of finite dimensional linear subspaces of $\mathscr{T} \cap \mathbb{L}_\infty(\mu)$ which is totally ordered by inclusion, that the dimension $D_m$ of $S_m$ is bounded by $n$ and that the index $\Phi_m$ defined by* (3.2) *is bounded by some constant $\Phi \geq 1$ for all $m \in \mathscr{M}_n$. Let $\hat{s}_m$ be the projection estimator on $S_m$ as defined in Section* 3.1.2, *$\kappa_2$ be a suitable numerical constant,* $\mathrm{pen}(m) \geq \kappa_2 \Phi^2 D_m/n$ *and $\hat{s}$ be the penalized projection estimator which is a minimizer with respect to $m \in \mathscr{M}_n$ of $-\|\hat{s}_m\|^2 + \mathrm{pen}(m)$. Then whatever $s \in \mathscr{T}$ such that $\mathbb{1} + s$ is a density*

$$\mathbb{E}_s\left[\|\hat{s} - s\|^2\right] \leq \kappa_2' \inf_{m\in\mathscr{M}_n}\left\{d^2(s, S_m) + \mathrm{pen}(m)\right\} + \kappa_2'' \frac{[\Phi(1 + \|s\|)]^4}{n} \quad . \tag{3.18}$$

**Application to ellipsoids with unknown coefficients:** We consider some orthonormal system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ in $\mathscr{T}$ where $\Lambda = \cup_{j\in\mathbb{N}}\Lambda(j)$, each $\Lambda(j)$ being a finite set. We limit ourselves here to the study of two cases of particular interest leaving the general case to Section 4.

- $\mu$ is the uniform distribution on the torus $[0, 2\pi]$, $\Lambda(j) = \{2j; 2j+1\}$ for $j \geq 0$ and $\varphi_{2j}(x) = \sqrt{2}\cos[(j+1)x]$, $\varphi_{2j+1}(x) = \sqrt{2}\sin[(j+1)x]$.

- $\mu$ is the Lebesgue measure on $[0, 1]$, $\Lambda(j) = \{(j, k) \mid 1 \leq k \leq 2^j\}$ for $j \geq 0$ and the $\varphi_{j,k}$ are the elements of the Haar basis described in Section 3.2.1.

For any non-increasing positive sequence $a = \{a_j\}_{j\geq 0}$ converging to zero we define the ellipsoid $\mathscr{E}(a)$ by

$$\mathscr{E}(a) = \left\{ \sum_{j\geq 0} \sum_{\lambda\in\Lambda(j)} \beta_\lambda \varphi_\lambda \,\middle|\, \sum_{j\geq 0} \left( a_j^{-2} \sum_{\lambda\in\Lambda(j)} \beta_\lambda^2 \right) \leq 1 \right\} .$$

Let us define for each $m \in \mathbb{N}$, $\Lambda_m = \sum_{j=0}^m \Lambda(j)$ and $D_m = |\Lambda_m|$. Then $\mathscr{M}_n = \{m \geq 0 \mid D_m \leq n\}$. For the sake of simplicity, we limit ourselves in this section to the case $a_j(H, \alpha) = H D_j^{-\alpha}$ with $H > 0$ and $\alpha > 0$ for the Fourier basis, $\alpha \in (0, 1]$ for the Haar basis. This case is of particular interest since it is well-known that $\cup_{H>0}\mathscr{E}(a(H, \alpha))$ is the set of periodic functions orthogonal to $\mathbb{1}$ belonging to the Sobolev space $W_2^\alpha$ in the Fourier case and of functions orthogonal to $\mathbb{1}$ belonging to the Besov space $B_{\alpha 2 2}$ in the Haar case. For the definition of those spaces, we refer to DeVore and Lorentz (1993, Chapter 2) and to the proof of Lemma 12 below.

If $s$ is an element of some ellipsoid $\mathscr{E}(a)$, it is immediate to see that $d^2(s, S_m) \leq a_{m+1}^2$. We also recall that in our examples, $\Phi_m$ is bounded by $\Phi$ with $\Phi^2 = 2$ for the trigonometric basis and $\Phi^2 = 1$ for the Haar basis. This allows to apply Theorem 3 with $\mathrm{pen}(m) = K_2 D_m/n$ and $K_2 \geq \Phi^2\kappa_2$ which implies that

$$\mathbb{E}_s\left[\|s - \hat{s}\|^2\right] \leq \kappa_2' \inf_{m\in\mathscr{M}_n} \left\{ H^2 D_{m+1}^{-2\alpha} + \frac{K_2 D_m}{n} \right\} + \kappa_2'' \Phi^4 \frac{(1 + \|s\|)^4}{n}$$

and finally whatever the true unknown values of $H$ and $\alpha$,

$$\mathbb{E}_s\left[\|s - \hat{s}\|^2\right] \leq C(K_2) \left(\frac{H}{n^\alpha}\right)^{2/(1+2\alpha)} + \kappa_2'' \Phi^4 \frac{(1 + \|s\|)^4}{n} .$$

The discussion about the optimality properties of such a bound will be developed in Section 4. Sharp asymptotic results using ellipsoids built on the Fourier basis are to be found in Efroimovich and Pinsker (1984) (for the white noise setting) and (1986) (for the spectral density), Efroimovich (1985) (for density estimation), all the results, except for the first one, being restricted to Hilbert-Schmidt ellipsoids (i.e. $\alpha > 1/2$ in the above examples).

### 3.3.3. Least squares estimators for smooth regression

We consider a regression framework $Y_i = s(X_i) + W_i$ where the $X_i$'s are independent identically distributed with common distribution $\mu$ and the

$W_i$'s are independent identically distributed and centered (with a distribution independent of $s$). The application of the following theorem requires the prior knowledge of some upper bound $\xi$ on $\|s\|_\infty$ since $\xi$ is involved in the construction of our estimator. This motivates the introduction of the set $\mathscr{T}_\xi = \{t \in \mathscr{T} = \mathbb{L}_2(\mu) \mid \|t\|_\infty \leq \xi\}$.

**Theorem 4** *Let $\xi$ and $\xi'$ be two positive numbers, assume that $\mathbb{E}[e^{|W_1|/\xi'}] \leq 4$ and let $\{L_m\}_{m\in\mathscr{M}_n}$ be a family of weights such that*

$$L_m \geq 1 \quad \textit{for all} \quad m \in \mathscr{M}_n \quad \textit{and} \quad \sum_{m\in\mathscr{M}_n} \exp[-L_m D_m] \leq \Sigma < +\infty \ . \tag{3.19}$$

*Assuming that $S_m \subset \mathscr{T}_\xi$ and recalling that $\bar{r}_m$ is defined by* (3.4) *for all $m \in \mathscr{M}_n$, there exists a suitable numerical constant $\kappa_3$ such that whenever*

$$\mathrm{pen}(m) \geq \kappa_3(\xi' + \xi)^2 \left[L_m + \log\left(1 + \bar{r}_m (D_m/n)^{1/2}\right)\right](D_m/n)$$

*the penalized least squares estimator $\hat{s}$ which is a minimizer with respect to $m \in \mathscr{M}_n$ and $t \in S_m$ of $n^{-1}\sum_{i=1}^n [Y_i - t(X_i)]^2 + \mathrm{pen}(m)$ satisfies*

$$\mathbb{E}_s\left[d^2(s,\hat{s})\right] \leq \kappa_3' \left[\inf_{m\in\mathscr{M}_n}\left\{d^2(s,S_m) + \mathrm{pen}(m)\right\} + \Sigma(\xi' + \xi)^2 n^{-1}\right] \tag{3.20}$$

*for all $s \in \mathscr{T}_\xi$.*

**Handling several bases simultaneously:** One of the advantages of model selection is to allow competition between various kinds of approximating spaces. In particular it is possible to use several bases at the same time to construct the penalized estimator. We now provide an illustration of this idea in the context of bounded regression. We assume that the regressors $X_i$ are uniformly distributed on $[0, 1]$ and that the errors $W_i$ satisfy the assumptions of Theorem 4 with a known constant $\xi'$. We consider simultaneously five different types of sieves indexed by the sets $\mathscr{M}^i$ with $1 \leq i \leq 5$ and take $\mathscr{M}_n = \cup_{1\leq i\leq 5}\mathscr{M}^i$. Let us fix $r \in \mathbb{N}$. We define $\mathscr{M}^1$ to be the set of regular partitions of $[0, 1]$ and $\mathscr{M}^2_N$ to be the set of all partitions with endpoints belonging to the grid $\{j/N \mid 0 \leq j \leq N\}$. Then $\mathscr{M}^2 = \cup_{N\geq 3}\mathscr{M}^2_N$. In both cases, $\bar{S}_m$ is the linear space of piecewise polynomials based on the partition $m$ with degree not larger than $r$. The other sieves in our collection are built from a basis $(\varphi_\lambda)_{\lambda\in\Lambda}$ of $\mathbb{L}_2([0, 1])$ with $\Lambda = \cup_{j\geq 0}\Lambda(j)$ and $\bar{S}_m$ is the linear span of $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$. We first consider the trigonometric basis with $\Lambda(0) = \{0\}$, $\Lambda(j) = \{2j - 1; 2j\}$ for $j \geq 1$ and $\varphi_0 = \mathbb{1}$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)$, $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)$. Then $\mathscr{M}^3 = \mathbb{N}$ and $\Lambda_m =$

$\cup_{j\le m}\Lambda(j)$. Finally we introduce a wavelet basis of regularity $r$ as described in Section 3.2.1 with $q = A = 1$. An element $m$ of $\mathcal{M}^4$ or $\mathcal{M}^5$ is a subset of the set of indices $\{(j,k)|1 \le k \le 2^j M,\ j \in \mathbb{N}\}$ and $\Lambda_m = m$. $\mathcal{M}^4$ is the set of all subsets of the form $m = \cup_{j=0}^{J}\Lambda(j)$ for $J \in \mathbb{N}$ and $\mathcal{M}^5$ is the collection of all finite subsets of $\Lambda$ which do not belong to $\mathcal{M}^4$. For each $m \in \mathcal{M}^5$ we define $J_m$ to be the smallest integer $J$ such that $m \subset \cup_{j=0}^{J}\Lambda(j)$.

If we assume that the true regression function $s$ is bounded by $L\xi'$ where $L$ is known, it is natural to restrict the sieves to sets of functions which are uniformly bounded by a constant $\xi = (L+1)\xi'$ for instance and therefore to choose $S_m = \bar{S}_m \cap \{t \mid \|t\|_\infty \le \xi\}$ for each $m \in \mathcal{M}_n$. In order to describe the penalty function, it is enough to bound $\bar{r}_m$ and choose $L_m$ for any $m$ in order that condition (3.19) should be satisfied. It follows from Section 3.2.1 that $\bar{r}_m$ is uniformly bounded if $m$ belongs to either $\mathcal{M}^1$ or $\mathcal{M}^4$. It follows from (3.5) that $\bar{r}_m \le \sqrt{2D_m}$ if $m \in \mathcal{M}^3$. Finally for $m \in \mathcal{M}^2$, $\bar{r}_m \le C(N/D_m)^{1/2}$ and for $m \in \mathcal{M}^5$, $\bar{r}_m \le C'(2^{J_m}/D_m)^{1/2}$. As to $L_m$ it can be chosen as 1 for $m \in \mathcal{M}^i$, $i = 1, 3$ or 4 and by Lemma 6 we can take $L_m = 2[1+\log(N/D_m)]$ for $m \in \mathcal{M}^2$ and $L_m = 2[1+\log(M2^{J_m}/D_m)]$ for $m \in \mathcal{M}^5$. Elementary computations similar to those we performed in Section 3.3.1 for histograms show that $\Sigma$ can then be taken as a numerical constant.

This is a situation where Theorem 4 applies leading to the upper bound (3.20) for the risk. It is difficult to analyze this bound in general. Moreover the minimax point of view is especially inadequate here since the interest of introducing such a rich family of sieves is to have more opportunity to approximate well a given $s$ by a sieve of low dimension rather than consider a uniform approximation over some large class of functions, which always reflects the worst case in the class. Nevertheless one can still evaluate the maximal risk over some suitable classes of smooth functions. Let us for instance consider for any positive number $\alpha$ with $\alpha = a + b,\ a \in \mathbb{N},\ 0 < b \le 1$ the class $\mathcal{H}_\alpha$ of functions $s$ on $[0,1]$ with $a$ derivatives and such that

$$\sup_{x,y\in[0,1]} \frac{|s^{(a)}(x) - s^{(a)}(y)|}{|x-y|^b} = |s|^{(\alpha)} < +\infty \ . \tag{3.21}$$

Recalling that $\mathcal{H}_\alpha$ is included in the Besov space $B_{\alpha\infty\infty}$, it follows from Lemma 12 below that for any positive $\varepsilon$ one can find in each of the three collections $\mathcal{M}^i$, $i = 1, 3, 4$ an $m$ such that when $s \in \mathcal{H}_\alpha$, there exists some point $\bar{s}_m \in \bar{S}_m$ such that $\|s - \bar{s}_m\|_\infty \le \varepsilon/2$ and $D_m \le C_i(|s|^{(\alpha)}/\varepsilon)^{1/\alpha}$ (with the additional assumption that $s$ is periodic when $i = 3$ or that the support of $s$ is included in $(0,1)$ when $i = 4$ and that $r \ge a$ when $i = 1$ or 4). Setting $s_m = \xi\bar{s}_m/(\xi + \varepsilon/2)$ we see that $s_m \in S_m$ and $\|s - s_m\| \le \varepsilon$ which implies that $d(s, S_m) \le \varepsilon$. Let us denote by $\tilde{r}_m$ the upper bound for $\bar{r}_m$ computed above and choose

$$\text{pen}(m) = K_3(\xi' + \xi)^2 \left[ L_m + \log\left(1 + \tilde{r}_m \left(\frac{D_m}{n}\right)^{1/2}\right)\right] \frac{D_m}{n}$$

with $K_3 \geq \kappa_3$. It then follows that for $i = 1, 3, 4$ the upper bound for the risk derived from Theorem 4 takes the form

$$C(K_3) \inf_{m \in \mathcal{M}^i} \left\{ d^2(s, S_m) + (\xi' + \xi)^2 \left[ 1 + \log\left(1 + \tilde{r}_m \left(\frac{D_m}{n}\right)^{1/2}\right)\right] \frac{D_m}{n} \right\}$$
$$\leq C'(K_3, i) \inf_{\alpha \leq r+1} \left\{ \left(\frac{n}{(\xi + \xi')^2}\right)^{-2\alpha/(2\alpha+1)} \left(|s|^{(\alpha)}\right)^{2/(1+2\alpha)} \right\}$$

where it is required that $\alpha \geq 1/2$ if $i = 3$ because of the influence of $\tilde{r}_m$ but then $r = +\infty$. This means that our estimator achieves the optimal rate of convergence $n^{-2\alpha/(2\alpha+1)}$ for functions of smoothness $\alpha$ but that it actually does more than this since it also optimizes the bound among the possible values of $\alpha$. Moreover the introduction of the larger classes $\mathcal{M}^2$ and $\mathcal{M}^5$ allows to get better approximation for functions $s$ of spatially inhomogeneous smoothness at the modest price of an additional $\log n$ factor. One could even go further in this direction by including in the model a fixed finite number of different wavelet bases. Related work (for the white noise setting) dealing with the selection of one among a library of orthonormal basis is to be found in Donoho and Johnstone (1994b). It is also worth mentioning here the work by Golubev and Nussbaum (1992) on spline adaptive estimation for Sobolev classes in a Gaussian regression framework.

## 4. Further examples

In order to keep the paper to a reasonable size, we shall only develop a few applications of our methods in various contexts. These particular examples were chosen because of their ability to illustrate different approaches to adaptation and model selection and the necessary compromise between the complexity of the family of sieves and the desire to get low and, in some sense, optimal rates of convergence if the true underlying density is not too complicated. Many other examples could be developed along the same lines but we shall concentrate here on a representative selection.

It should be noted that each particular family of sieves will be given for a particular type of minimum contrast estimation procedure (maximum likelihood and projection for density estimation or least squares for regression settings) for the sake of simplicity. For instance it is natural to use sieves with good uniform approximation properties in the case of maximum likelihood in order to warrant positivity. Pure $\mathbb{L}_2$-approximation is more suited for projection. For regression our choice of bounded sieves derives naturally from

the assumptions needed but it is clear that the examples that we introduce for density estimation could also be used in the regression framework with an additional restriction of uniform boundedness on the family of sieves.

### *4.1. Nested families of models and analogues*

By this we mean that the family of models is a totally ordered family of linear spaces which implies that all numbers $D_m$ are different or that a similar situation holds: $D_m$ is an integer and the number of models with the same dimension $D_m$ is rather small, at least small enough to ensure that the series $\sum_{m\in\mathcal{M}_n} \exp(-D_m) \leq \Sigma < +\infty$ independently of $n$.

#### 4.1.1. Ellipsoids with unknown coefficients

We give here a detailed account of the properties of projection estimators when $s$ belongs to some ellipsoid $\mathcal{E}(a)$ with unknown coefficients as described in Section 3.3.2. The ellipsoids are given by some orthonormal system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ of $\mathbb{L}_2(\mu)$ where $\mu$ is a probability measure and $\Lambda = \cup_{j\in\mathbb{N}}\Lambda(j)$, each $\Lambda(j)$ being a finite set. Furthermore $\int \varphi_\lambda \, d\mu = 0$ for all $\lambda \in \Lambda$. We recall from Section 3.3.2 that, for any non-increasing positive sequence $a = \{a_j\}_{j\geq 0}$ converging to zero, $\mathcal{E}(a)$ is the set of functions of the form $\sum_{\lambda\in\Lambda} \beta_\lambda\varphi_\lambda$ such that $\sum_{j\geq 0}\sum_{\lambda\in\Lambda(j)}(\beta_\lambda/a_j)^2 \leq 1$, that $\Lambda_m = \cup_{j=0}^m \Lambda(j)$, $D_m = |\Lambda_m|$ and (provided that $D_0 \leq n$) $\mathcal{M}_n = \{m \in \mathbb{N} | D_m \leq n\}$. We also assume that the $\Phi_m$'s are uniformly bounded by some constant $\Phi$ and that $\text{pen}(m) = K_2 D_m/n$ with $K_2 \geq \kappa_2\Phi^2$. Then Theorem 3 holds with $d^2(s, S_m) \leq a_{m+1}^2$ leading to

$$\mathbb{E}_s\left[\|s - \hat{s}\|^2\right] \leq C(K_2, a_0) \inf_{m\in\mathcal{M}_n}\left\{a_{m+1}^2 + \frac{D_m}{n}\right\} \tag{4.1}$$

since at least $\|s\| \leq a_0$. Defining

$$m(n) = \inf\left\{m \geq 0 | a_{m+1}^2 \leq D_m/n\right\} \tag{4.2}$$

we see that, if $na_0^2 \geq D_0$ (which always holds for $n$ large enough) and the ratios $D_{m+1}/D_m$ are uniformly bounded (which will be the case in all the applications below), the following inequality holds for some constant $K \geq 1$:

$$D_{m(n)} \leq Kna_{m(n)}^2 \ . \tag{4.3}$$

Therefore the convergence rate of the right-hand side of (4.1) when $n \to +\infty$ is of the order of $D_{m(n)}/n$ by Lemma 15 of Section 8 below.

Let us try to see what would happen if the sequence $(a_j)_{j\geq 1}$ were known. This would mean that our parameter space would be restricted to the set

$\bar{\mathscr{E}}(a) = \{u \in \mathscr{E}(a) | \mathbb{1} + u \geq 0\}$ since $\mathbb{1} + u$ is a density. The following proposition provides a lower bound for the minimax risk over $\bar{\mathscr{E}}(a)$. We shall then discuss on specific examples (Fourier, Haar and Sobolev ellipsoids) how far it is from the upper bound (4.1).

**Proposition 2** *Let $n$ be given and assume that for all $m \geq 1$ there exists a subset $\mathscr{C}_m$ of the cube $\{-1, +1\}^{\Lambda_m}$ with $|\mathscr{C}_m| \geq 2^{D_m - 1}$ and*

$$\sup_{\delta \in \mathscr{C}_m} \left\| \sum_{\lambda \in \Lambda_m} \delta_\lambda \varphi_\lambda \right\|_\infty \leq \Psi_m \tag{4.4}$$

*with $\Psi^2_{m(n)} \leq n\Psi$ for all $n > 0$. If $\Phi = \sup_{m \geq 0} \Phi_m$, any estimator $\tilde{s}$ satisfies*

$$\sup_{s \in \bar{\mathscr{E}}(a)} \mathbb{E}_s \left[ \|s - \tilde{s}\|^2 \right] \geq \kappa_{10} \frac{1 \wedge n a_0^2}{(\Phi^2/n) \vee \Psi} \sup_{m \in \mathbb{N}} \left\{ \frac{D_m}{n} \wedge a_m^2 \right\} . \tag{4.5}$$

*Moreover if one assumes that $a_0^2 \geq K_0/n$ we get*

$$\sup_{s \in \bar{\mathscr{E}}(a)} \mathbb{E}_s \left[ \|s - \tilde{s}\|^2 \right] \geq C(\Phi, \Psi, K_0) \left[ \inf_{m \in \mathscr{M}_n} \left\{ a_{m+1}^2 + \frac{D_m}{n} \right\} \wedge 1 \right] . \tag{4.6}$$

Before we come to the proof of this Proposition, let us make a few comments and develop some applications. If the set $\bar{\mathscr{E}}(a)$, because of the positivity requirement, is substantially smaller than $\mathscr{E}(a)$, there is no hope that our upper bound (4.1) be optimal since in designing it we essentially pretended that the whole of $\mathscr{E}(a)$ was the parameter space. The role of $\Psi_m$ is to quantify this effect. If we take $\mathscr{C}_m = \{-1, +1\}^{\Lambda_m}$, $\Psi_m$ is bounded by $\bar{r}_m \sqrt{D_m}$ by (3.4).

**Comments about the size of $a_0$:** One should first observe that keeping $a_0$ bounded allows to keep $C(K_2, a_0)$ in (4.1) under control since it is a nondecreasing function of $a_0$ and therefore under the assumptions of Proposition 2 the bounds (4.1) and (4.6) do match. Then one notices that if $na_0^2$ is too small one gets into trouble which is not surprising since this means that the diameter of the ellipsoid, which is measured by $a_0$ is essentially smaller than $n^{-1/2}$ and that the simple estimator $\tilde{s} = 0$ would perform very well in this situation. This is then a completely degenerate problem where the optimal rate of convergence for the quadratic risk is smaller than the parametric rate $n^{-1}$. In order to avoid unnecessary complications in the treatment of the applications below we shall assume from now on that $n$ is large enough to ensure that $na_0^2 \geq D_0$.

**Straightforward applications:** We first present two examples for which $\mathscr{C}_m = \{-1, +1\}^{\Lambda_m}$ and $\Psi^2_{m(n)}/n$ is easily seen to be bounded independently of $n$. Subsequently the rate provided by (4.1) is optimal for a given value of $a_0$. We also assume that the ratios $D_{m+1}/D_m$ are uniformly bounded.

- If $\bar{r}_m$ is bounded by $R$ which is the case for the Haar ellipsoid (see Section 3.3.2), $\Psi^2_{m(n)}/n \leq R^2 D_{m(n)}/n \leq R^2 K a_0^2$ by (4.3).
- Since $\Phi_m$ is bounded by $\Phi$, by (3.5) we can always take $\Psi_m = \Phi D_m$. If moreover $\mathscr{E}(a)$ is Hilbert-Schmidt, i.e. $a$ is such that $\sum_{j\geq 0} |\Lambda(j)| a_j^2 = \Xi < +\infty$, then by monotonicity $D_m a_m^2 \leq \Xi$ for any $m$. It follows from (4.3) that $D_{m(n)} \leq (K\Xi n)^{1/2}$ from which one derives that $\Psi^2_{m(n)} \leq \Phi^2 K \Xi n$.

**Fourier ellipsoids:** When the ellipsoid is not Hilbert-Schmidt, the preceding argument breaks down. We can still apply Proposition 2 with different sets $\mathscr{C}_m$. Recall that, in the case of the Fourier basis defined in Section 3.3.2, $D_m = 2(m+1)$. A classical result by Salem and Zygmund on random Fourier series (see Kahane 1985, Theorem 2 p. 69) implies that there exists a subset $\mathscr{C}_m$ of $\{-1, +1\}^{\Lambda_m}$ of cardinality larger than $2^{D_m - 1}$ such that (4.4) holds with $\Psi_m = \bar{\Psi}[D_m \log(D_m)]^{1/2}$. If we assume that $a_j[\log(j+2)]^{1/2}$ is bounded (which is clearly a much weaker condition than $\sum a_j^2 < +\infty$), then $D_{m(n)} \log[2(m(n)+1)]/n$ is bounded via (4.3) and so is $\Psi^2_{m(n)}/n$. Therefore (4.6) matches (4.1). Note that when $a_j$ converges to zero more slowly than $(\log j)^{-1/2}$ the minimax risk, by the preceding arguments, is anyway at least of order $1/\log n$ which is dramatically slow. The same kind of results hold for multidimensional Fourier expansions for the same reasons.

Similar lower bounds under the same restrictions ($\sup_j a_j^2 \log(j+2) < +\infty$) were found by Efroimovich and Pinsker (1981, 1982). These authors were actually able to compute not only a lower bound for the rate of convergence but even the exact asymptotic value of the minimax risk for a given ellipsoid built on the Fourier basis, for the problems of density estimation and spectral density estimation.

**Sobolev ellipsoids on compact Riemannian manifolds:** We consider some compact connected Riemannian manifold $\mathbb{M}$ with dimension $q$ and uniform distribution $\mu$ and recall from Section 3.2.1 that $\{\theta_j | j \geq 0\}$ is the set of eigenvalues of the Laplacian operator on $\mathbb{M}$ and $\{\varphi_\lambda | \lambda \in \Lambda(j)\}$ the set of eigenvectors corresponding to $\theta_j$. We shall say that $s = \sum_{j\geq 0} \sum_{\lambda\in\Lambda(j)} \beta_\lambda \varphi_\lambda$ belongs to the Sobolev space $\mathscr{H}_\alpha(\mathbb{M})$ for some $\alpha > 0$ if and only if the coefficients $\beta_\lambda$ satisfy

$$\sum_{j\geq 0} \sum_{\lambda\in\Lambda(j)} \theta_j^\alpha \beta_\lambda^2 = H^2 < +\infty \ .$$

Therefore estimating the function $s$ of unknown smoothness (in the Sobolev sense) amounts to estimating $s \in \mathscr{E}(a)$, where $a_j = H\theta_j^{-\alpha/2}$ for all $j \geq 0$, with $H$ and $\alpha$ unknown. It follows from our computations in Section 3.2.1 that the corresponding family $\{\Phi_m\}_{m\in\mathscr{M}_n}$ is uniformly bounded by some constant $\Phi(\mathbb{M})$ and therefore that Theorem 3 applies leading to the bound (3.18). In order to measure the effect of $H$ on the risk we need to derive from (3.18) a sharper bound than (4.1). Choosing $\mathrm{pen}(m) = K_2 D_m/n$ with $K_2 \geq \kappa_2 \Phi^2(\mathbb{M})$ we get from Theorem 3

$$\mathbb{E}_s\left[\|\hat{s} - s\|^2\right] \leq \kappa_2' \inf_{m\in\mathscr{M}_n}\left\{H^2\theta_{m+1}^{-\alpha} + \frac{K_2 D_m}{n}\right\} + \kappa_2'' \Phi^4(\mathbb{M})\frac{(1+\|s\|)^4}{n} \quad . \tag{4.7}$$

We wish to know under which conditions this upper bound matches the lower bound (4.6) up to constants. In order to answer this question it is necessary to control $\|s\|$ when $s$ belongs to the ellipsoid $\mathscr{E}(a)$. Such a control is given in the following

**Lemma 1** *Let $\mathbb{1} + s = \mathbb{1} + \sum_{j\geq 0}\sum_{\lambda\in\Lambda(j)} \beta_\lambda \varphi_\lambda$ be a probability density on $\mathbb{M}$ such that*

$$\sum_{j\geq 0}\sum_{\lambda\in\Lambda(j)} \theta_j^\alpha \beta_\lambda^2 \leq H^2 \quad .$$

*Then, there exists some constant $C(\mathbb{M})$ (independent of $\alpha$ and $H$) such that*

$$\|s\|^2 \leq C(\mathbb{M})\left(D_0 \vee H^{2q/(2\alpha+q)}\right) \quad .$$

*Proof*: Since $\beta_\lambda = \int(\mathbb{1}+s)\varphi_\lambda\, d\mu$ and $\mathbb{1}+s$ is a probability density, Jensen's inequality implies that $\beta_\lambda^2 \leq \int(\mathbb{1}+s)\varphi_\lambda^2\, d\mu$ and then by (3.3)

$$\sum_{j\leq m}\sum_{\lambda\in\Lambda(j)} \beta_\lambda^2 \leq \left\|\sum_{j\leq m}\sum_{\lambda\in\Lambda(j)} \varphi_\lambda^2\right\|_\infty \leq \Phi_m^2 D_m \leq \Phi^2(\mathbb{M}) D_m \quad .$$

Since we also know that $\sum_{j>m}\sum_{\lambda\in\Lambda(j)} \beta_\lambda^2 \leq H^2\theta_{m+1}^{-\alpha}$ it follows that

$$\|s\|^2 \leq \inf_{m\geq 0}\left\{\Phi^2(\mathbb{M}) D_m + H^2\theta_{m+1}^{-\alpha}\right\} \quad .$$

Defining $m' = \inf\{m \in \mathbb{N} \mid H^2\theta_{m+1}^{-\alpha} \leq \Phi^2(\mathbb{M})D_m\}$, we get $\|s\|^2 \leq 2\Phi^2(\mathbb{M})D_{m'}$. Then, either $m' = 0$ and $\|s\|^2 \leq 2\Phi^2(\mathbb{M})D_0$, or $m' > 0$ which implies by (3.10) that

$$H^2\theta_{m'}^{-\alpha} > \Phi^2(\mathbb{M})D_{m'-1} \geq \Phi^2(\mathbb{M})(C_1(\mathbb{M})/C_2(\mathbb{M}))^{q/2} D_{m'} \quad .$$

Using (3.10) again we get

$$D_{m'} \leq \left(\frac{H^2}{\Phi^2(\mathbb{M})}\right)^{q/(2\alpha+q)} C_1(\mathbb{M})^{-q/2} C_2(\mathbb{M})^{q^2/(4\alpha+2q)}$$

and the conclusion follows. □

The next proposition gives a precise evaluation of the quantity $\inf_{m\in\mathcal{M}_n}\{H^2\theta_{m+1}^{-\alpha} + D_m/n\}$ which appears in both the upper and lower bounds of the risk. It allows to conclude that if $\alpha > q/2$ and (4.10) below holds, these bounds coincide up to some multiplicative constant depending only on the structure of the manifold $\mathbb{M}$.

**Proposition 3** *If* $H \leq [C_1(\mathbb{M})/C_2(\mathbb{M})]^{(2\alpha+q)/4} n^{\alpha/q}$ *then*

$$\inf_{m\in\mathcal{M}_n}\left\{\frac{H^2}{\theta_{m+1}^{\alpha}} + \frac{D_m}{n}\right\} \leq 2\left[\frac{D_0}{n} + \left(\frac{C_2(\mathbb{M})\vee 1}{C_1(\mathbb{M})}\right)^{q/2}\left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)}\right] \quad (4.8)$$

*and if* $H^2 > D_0\theta_1^{\alpha}/n$ *then*

$$\inf_{m\in\mathcal{M}_n}\left\{\frac{H^2}{\theta_{m+1}^{\alpha}} + \frac{D_m}{n}\right\} \geq \left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)}\left(\frac{C_1(\mathbb{M})}{C_2^{3/2}(\mathbb{M})\vee 1}\right)^{q} . \quad (4.9)$$

*Moreover if we assume that* $\alpha > q/2$ *and that*

$$\frac{D_0\theta_1^{\alpha}\vee 1}{n} \leq H^2 \leq \left[\left(\frac{C_1(\mathbb{M})}{C_2(\mathbb{M})}\right)^{(2\alpha+q)/2} n^{2\alpha/q}\right] \wedge n^{(2\alpha-q)/(2q)} \wedge n \ , \quad (4.10)$$

*the following inequalities hold for suitable constants* $C(\mathbb{M})$ *and* $C'(\mathbb{M})$ *depending only on the structure of* $\mathbb{M}$*:*

$$\inf_{\tilde{s}} \sup_{s\in\bar{\mathcal{E}}(a)} \mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \geq C(\mathbb{M})\left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)} \quad (4.11)$$

*for the lower bound on the minimax risk and for our penalized projection estimator* $\hat{s}$

$$\sup_{s\in\bar{\mathcal{E}}(a)} \mathbb{E}_s\left[\|\hat{s}-s\|^2\right] \leq C'(\mathbb{M})\left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)} . \quad (4.12)$$

*Proof*: Let us first observe that it follows from Lemma 15 of Section 8 that

$$I = \inf_{m \in \mathcal{M}_n} \left\{ \frac{H^2}{\theta_{m+1}^{\alpha}} + \frac{D_m}{n} \right\} \geq \sup_{m \geq 0} \left\{ \frac{D_m}{n} \wedge \frac{H^2}{\theta_m^{\alpha}} \right\} = \frac{D_{m(n)}}{n} \wedge \frac{H^2}{\theta_{m(n)}^{\alpha}} \quad (4.13)$$

and

$$I \leq 2D_{m(n)}/n \quad \text{provided that } D_{m(n)} \leq n \quad (4.14)$$

where $m(n)$ defined in (4.2) is given by

$$m(n) = \inf\{m \in \mathbb{N} \mid H^2 \theta_{m+1}^{-\alpha} \leq D_m/n\} \ . \quad (4.15)$$

Assuming first that $m(n) \geq 1$ and noticing that (3.10) implies that

$$\theta_{m+1} \leq \theta_m \frac{C_2(\mathbb{M})}{C_1(\mathbb{M})} \quad \text{and} \quad D_m \geq D_{m+1} \left( \frac{C_1(\mathbb{M})}{C_2(\mathbb{M})} \right)^{q/2} \ ,$$

we derive from (4.15) that

$$\left( \frac{C_1(\mathbb{M})}{C_2(\mathbb{M})} \right)^{q/2} \frac{D_{m(n)}}{n} \leq H^2 \theta_{m(n)}^{-\alpha} \leq \left( \frac{C_2(\mathbb{M})}{C_1(\mathbb{M})} \right)^{\alpha} \frac{D_{m(n)}}{n} \ . \quad (4.16)$$

Combining this with (3.10) we get

$$nH^2 C_1^{\alpha}(\mathbb{M}) C_2^{-2\alpha}(\mathbb{M}) \leq D_{m(n)}^{(2\alpha+q)/q} \leq nH^2 C_1^{-(2\alpha+q)/2}(\mathbb{M}) C_2^{q/2}(\mathbb{M}) \ .$$

Assuming without loss of generality that $C_2(\mathbb{M}) \geq 1$ we note that

$$\left( \frac{C_1(\mathbb{M})}{C_2^2(\mathbb{M})} \right)^{\alpha q/(2\alpha+q)} \geq \left( \frac{C_1(\mathbb{M})}{C_2^2(\mathbb{M})} \right)^{q} \quad \text{and} \quad \left( C_2^{q/2}(\mathbb{M}) \right)^{q/(2\alpha+q)} \leq C_2^{q/2}(\mathbb{M})$$

which implies that

$$\left(nH^2\right)^{q/(2\alpha+q)} \left( \frac{C_1(\mathbb{M})}{C_2^2(\mathbb{M})} \right)^{q/2} \leq D_{m(n)} \leq \left(nH^2\right)^{q/(2\alpha+q)} \left( \frac{C_2(\mathbb{M})}{C_1(\mathbb{M})} \right)^{q/2} \ . \quad (4.17)$$

By (4.16) and (4.17) the lower bound in (4.13) becomes

$$I \geq \frac{D_{m(n)}}{n} \left( \frac{C_1(\mathbb{M})}{C_2(\mathbb{M})} \right)^{q/2} \geq \left( \frac{H^q}{n^{\alpha}} \right)^{2/(2\alpha+q)} \left( \frac{C_1(\mathbb{M})}{C_2^{3/2}(\mathbb{M})} \right)^{q} \ .$$

If $H^2 > D_0 \theta_1^{\alpha}/n$ which ensures that $m(n) \geq 1$ we get the lower bound (4.9).

Turning our attention to (4.8), we see that it follows from (4.14) if $m(n) = 0$. When $m(n) \geq 1$ we derive from (4.17) that $D_{m(n)} \leq n$ and therefore $m(n) \in \mathcal{M}_n$ as soon as $H \leq [C_1(\mathbb{M})/C_2(\mathbb{M})]^{(2\alpha+q)/4} n^{\alpha/q}$ and (4.8) then follows from (4.14) and (4.17).

We can now turn to a precise evaluation of the risk. Combining Lemma 1, (4.8) and (4.10) we see from (4.7) that the upper bound (4.12) holds for the risk of our estimator. On the other hand it follows from (3.3) that

$$\sup_{\delta \in \mathscr{C}_m} \left\| \sum_{\lambda \in \Lambda_m} \delta_\lambda \varphi_\lambda \right\|_\infty^2 \leq \Phi^2(\mathbb{M}) D_m^2$$

and we can choose $\Psi_{m(n)} = \Phi^2(\mathbb{M}) D_{m(n)}^2$. Consequently from (4.17) and (4.10) $\Psi_{m(n)}/n$ is bounded by a constant $C''(\mathbb{M})$. Then (4.6) combined with (4.9) imply the lower bound (4.11). □

*Proof of Proposition 2:* For each $m \in \mathbb{N}$ such that $D_m \geq 6$ let us define

$$\mathscr{E}_m = \left\{ \frac{1}{\sqrt{N_m}} \sum_{\lambda \in \Lambda_m} \delta_\lambda \varphi_\lambda \mid \delta \in \mathscr{C}_m \right\} \quad \text{with} \quad N_m = 578n \vee 4\Psi_m^2 \vee D_m a_m^{-2} \ .$$

Since $N_m^{-1} D_m \leq a_m^2$, $\mathscr{E}_m \subset \mathscr{E}(a)$. Moreover $\Psi_m \leq \sqrt{N_m}/2$ and all elements $u$ of $\mathscr{E}_m$ therefore satisfy

$$\frac{1}{2} \leq \mathbb{1} + u \leq \frac{3}{2} \tag{4.18}$$

which a fortiori implies that $\mathscr{E}_m \subset \bar{\mathscr{E}}(a)$. It also follows from (4.18) that any pair $(u, v)$ of elements of $\mathscr{E}_m$ satisfies

$$h^2(\mathbb{1} + u, \mathbb{1} + v) = \frac{1}{2} \int \left( \sqrt{\mathbb{1} + u} - \sqrt{\mathbb{1} + v} \right)^2 d\mu \leq \frac{1}{4} \|u - v\|^2 \leq \frac{D_m}{N_m}$$

where $h$ denotes the Hellinger distance and therefore the Kullback-Leibler information numbers between the probabilities corresponding to the elements of $\mathscr{E}_m$ are uniformly bounded (see Inequality 7.6 of Birgé and Massart 1998) by $4.84 D_m/N_m$. A classical combinatorial argument that we shall prove later for the sake of completeness (see Lemma 8) ensures that there exists a subset $\mathscr{E}'_m$ of $\mathscr{E}_m$ of cardinality larger than $(1/2)\exp(D_m/3)$ such that for all $u, v \in \mathscr{E}'_m$

$$\|u - v\|^2 \geq 2\left[1 - \sqrt{2/3}\right](D_m/N_m) > 0.367\,(D_m/N_m) \ . \tag{4.19}$$

An application of Fano's Lemma (see Birgé 1986, p. 279 for a suitable version of it) shows that any estimator $\tilde{u}_m$ with values in $\mathscr{E}'_m$ satisfies $\sup_{u \in \mathscr{E}'_m} \mathbb{P}_u[\tilde{u}_m \neq u] \geq 1/4$ provided that

$$4.84\,\frac{nD_m}{N_m} + \log 2 \le \frac{3}{4}\log\left(\frac{1}{2}e^{D_m/3} - 1\right)$$

which is true since $D_m \ge 6$ and $N_m \ge 578n$. Since

$$\sup_{u\in\mathscr{E}'_m} \mathbb{E}_u\left[\|u-\tilde{u}_m\|^2\right] \ge \sup_{u\in\mathscr{E}'_m} \mathbb{P}_u[\tilde{u}_m \ne u] \inf_{u,v\in\mathscr{E}'_m} \|u-v\|^2$$

one concludes with (4.19) that

$$\sup_{u\in\mathscr{E}'_m} \mathbb{E}_u\left[\|u-\tilde{u}_m\|^2\right] \ge \frac{0.367}{4}\frac{D_m}{N_m} > \frac{D_m}{11N_m} \quad .$$

If $D_m \le 5$ we simply choose $\mathscr{E}'_m = \{-\varphi_\lambda N_m^{-1/2}, \varphi_\lambda N_m^{-1/2}\}$ for some $\lambda \in \Lambda_0$ with $N_m = 162n \vee 20\Phi^2 \vee a_0^{-2}$ . Since $N_m^{-1/2} \le a_0,\ \Phi(D_0/N_m)^{1/2} \le 1/2$ (recalling that $D_0 \le 5$) and $\|\varphi_\lambda\|_\infty \le \Phi\sqrt{D_0}$, $\mathscr{E}'_m \subset \bar{\mathscr{E}}(a)$ with $1/2 \le \mathbb{1} \pm \varphi_\lambda N_m^{-1/2} \le 3/2$. It follows that

$$(2N_m)^{-1} \le h^2(\mathbb{1} - \varphi_\lambda N_m^{-1/2}, \mathbb{1} + \varphi_\lambda N_m^{-1/2}) \le N_m^{-1}$$

and Lemma 7 implies that

$$\sup_{u\in\mathscr{E}'_m} \mathbb{E}_u\left[\|u-\tilde{u}_m\|^2\right] \ge \frac{1}{4N_m}\left[1 - \left(\frac{2n}{N_m}\right)^{1/2}\right] > \frac{D_m}{23N_m}$$

since $D_m \le 5$. If $\tilde{u}$ is an arbitrary estimator and $\tilde{u}_m$ its projection on $\mathscr{E}'_m$ one gets

$$\sup_{u\in\mathscr{E}'_m} \mathbb{E}_u\left[\|u-\tilde{u}\|^2\right] \ge \frac{1}{4}\sup_{u\in\mathscr{E}'_m} \mathbb{E}_u\left[\|u-\tilde{u}_m\|^2\right]$$

from which one derives in both cases ($D_m < 6$ or $D_m \ge 6$) from the values of $N_m$ that

$$\sup_{u\in\bar{\mathscr{E}}(a)} \mathbb{E}_u\left[\|u-\tilde{u}\|^2\right] \ge \frac{D_m}{4(6358n \vee 44\Psi_m^2 \vee 460\Phi^2 \vee 23a_0^{-2} \vee 11D_m a_m^{-2})} \quad .$$

Choosing $m = m(n)$, (4.5) follows from Lemma 15. (4.6) also follows from Lemma 15 provided that $D_{m(n)} \le n$ which implies that $m(n) \in \mathscr{M}_n$. The only delicate situation occurs when $D_{m(n)} > n$. If $D_{m(n)-1} > n$ then $a_m^2 > 1$ and the lower bound given by (4.5) is a constant otherwise $m(n) - 1 \in \mathscr{M}_n$ and

$$\inf_{m\in\mathscr{M}_n}\left\{a_{m+1}^2 + \frac{D_m}{n}\right\} \le 2a_{m(n)}^2 \quad .$$

Therefore (4.6) holds in both cases. □

4.1.2. Densities with an unknown modulus of continuity

Let $\omega$ be a modulus of continuity which means a subadditive continuous nondecreasing and nonnegative function defined on $[0, 1]$ such that $\omega(0) = 0$ (see DeVore and Lorentz 1993, p. 41 for details). Let $\mathscr{S}_\omega$ denote the set of functions $s \in \mathscr{S}$ such that

$$|s(x) - s(y)| \leq \omega(|x - y|) \quad \text{for all } x, y \in [0, 1] \ .$$

We assume that the true density $s^2$ is such that $s$ belongs to $\mathscr{S}_\omega$ for some unknown $\omega$. We want to show here that, using a maximum penalized likelihood procedure over the family of regular histograms, it is possible to estimate $s$ without knowing $\omega$ as well (up to multiplicative constants) as if $\omega$ were known.

Let us choose $\mathscr{M}_n = \{2, \ldots, n\}$ and define $S_m$ to be the set of regular non-negative histograms with $m$ pieces and $\mathbb{L}_2(\mu)$-norm equal to one so that an element of $S_m$ may be written as

$$\sum_{j=1}^{m} b_j \mathbb{1}_{[(j-1)/m, j/m)} \quad \text{with } b_j \geq 0 \quad \text{for } 1 \leq j \leq m \quad \text{and } \sum_{j=1}^{m} b_j^2 = m \ .$$

We want to apply Theorem 2. The family of sieves $S_m$, $m \geq 2$ satisfies (3.13) with $L_m = 1$, $\Sigma = 1/4$ and we have seen in Section 3.2.1 that $\bar{r}_m \leq 1$. It remains to control the bias term $K(s, S_m) \wedge 1$. Let $s_m^+$ be defined as follows:

$$s_m^+ = \sum_{j=1}^{m} b_j \mathbb{1}_{[(j-1)/m, j/m)} \quad \text{with } b_j = \sup_{(j-1)/m \leq x < j/m} s(x) \ .$$

Then $s_m^+ \geq s$ and using the fact that $\omega$ is nondecreasing one can check that $d(s, s_m^+) \leq \omega(1/m)$. It therefore comes from Proposition 1 and Theorem 2 that if $\text{pen}(m) = K_1 m/n$ with $K_1 \geq (1 + \log 2)\kappa_1$ and $\hat{s}$ denotes the maximum penalized likelihood estimator,

$$\mathbb{E}_s\left[d^2(s, \hat{s})\right] \leq \kappa_1' \left[\inf_{m \in \mathscr{M}_n} \left\{3\omega^2\left(\frac{1}{m}\right) + K_1 \frac{m}{n}\right\} + \frac{1}{4n}\right] \ .$$

Since $d^2(s, \hat{s})$ is bounded by 2 one can conclude that

$$\mathbb{E}_s\left[d^2(s, \hat{s})\right] \leq 2 \wedge \left[C(K_1) \inf_{m \in \mathscr{M}_n} \left\{\omega^2\left(\frac{1}{m}\right) + \frac{m}{n}\right\}\right] \ . \tag{4.20}$$

Let us now find a lower bound for the minimax risk over $\mathscr{S}_\omega$.

**Proposition 4** *The maximal risk of any estimator $\tilde{s}$ is bounded from below by*

$$\sup_{s\in\mathscr{S}_\omega} \mathbb{E}_s\left[d^2(s,\tilde{s})\right] \geq \frac{\kappa_0' n\omega^2(1/2)}{1+n\omega^2(1/2)}\left[\inf_{m\in\mathscr{M}_n}\left\{\omega^2\left(\frac{1}{m}\right)+\frac{m}{n}\right\}\wedge 1\right] . \tag{4.21}$$

*Proof*: It follows the lines of Birgé (1983, pp. 211–212) with the necessary modifications due to the fact that we work with square-roots of densities rather than densities. Let $m$ be a positive integer such that $\omega[1/(2m)] \leq 2$, $\delta = 1/(4m)$ and $v$ be the triangular function on $[0, 2\delta]$ given by $v(0) = v(2\delta) = 0$ and $v(\delta) = \omega(2\delta)/2$. We define by $v_0$ and $v_1$ respectively the functions

$$v_0(x) = \eta v(x)\mathbb{1}_{[0,2\delta)}(x) - v(x-2\delta)\mathbb{1}_{[2\delta,4\delta)}(x) \ ;$$

$$v_1(x) = -v(x)\mathbb{1}_{[0,2\delta)}(x) + \eta v(x-2\delta)\mathbb{1}_{[2\delta,4\delta)}(x) \ .$$

where $\eta \in (1/2, 1)$ is given by

$$(1+\eta^2)\int_0^\delta v^2(x)\,dx = 2(1-\eta)\int_0^\delta v(x)\,dx \ .$$

Then $\mathbb{1}+v_0$ and $\mathbb{1}+v_1$ are nonnegative functions (since $v(\delta) \leq 1$) of norm 1. For any $\varepsilon \in \{0; 1\}^m$ the function $s_\varepsilon$ defined by

$$s_\varepsilon(x) = 1 + \sum_{j=0}^{m-1}[\varepsilon_{j+1}v_0(x-4j\delta) + (1-\varepsilon_{j+1})v_1(x-4j\delta)]$$

is an element of $\mathscr{S}_\omega$ because of our choice of $v$. If all coordinates of $\varepsilon$ and $\varepsilon'$ match except for one, a straightforward calculation yields

$$d^2(s_\varepsilon, s_{\varepsilon'}) = 2(1+\eta)^2\int_0^{2\delta} v^2(x)\,dx = \delta\omega^2(2\delta)(1+\eta)^2/3 \ .$$

It follows from Assouad's Lemma (see Birgé 1986, p. 280) that for any estimator $\tilde{s}$ based on $n$ independent identically distributed observations

$$\sup_{\varepsilon\in\{0;1\}^m} \mathbb{E}_{s_\varepsilon}\left[d^2(s_\varepsilon,\tilde{s})\right] \geq \frac{\beta}{16\delta}\left[1-\sqrt{2n\beta}\right] \quad \text{with} \ \ \beta = \frac{\delta\omega^2(2\delta)(1+\eta)^2}{6} \ . \tag{4.22}$$

Let us choose $m = m(n)$ with

$$m(n) = \min\left\{ m \geq 1 \,\middle|\, \omega^2\left(\frac{1}{2m}\right) \leq 2\left(\frac{m}{n} \wedge 2\right)\right\}.$$

Then $\omega[1/(2m)] \leq 2$ as required and since $\eta \in (1/2, 1)$ then $(3/8)\delta\omega^2(2\delta) \leq \beta \leq 1/(3n)$ and therefore one derives from (4.22) that

$$\sup_{s \in \mathscr{S}_\omega} \mathbb{E}_s\left[d^2(s, \tilde{s})\right] \geq \frac{3}{128}\omega^2\left(\frac{1}{2m(n)}\right)\left[1 - (2/3)^{1/2}\right] \quad .$$

In order to derive (4.21) it is enough to bound the ratio

$$\left[\left(\omega^2\left(\frac{1}{m_0}\right) + \frac{m_0}{n}\right) \wedge 1\right] \Big/ \omega^2\left(\frac{1}{2m(n)}\right)$$

for a suitable $m_0 \in \mathscr{M}_n$. If $m(n) = 1$ taking $m_0 = 2$ gives (4.21). Otherwise $m(n) \geq 2$. If $\omega^2(1/(2n)) \leq 2$, then $m(n) \leq n$, hence $m(n) \in \mathscr{M}_n$ and we choose $m_0 = m(n)$. It then follows from the definition of $m(n) = m_0 \geq 2$ that

$$\omega^2\left(\frac{1}{2(m_0 - 1)}\right) > \frac{2(m_0 - 1)}{n} \geq \frac{m_0}{n}$$

and from the subadditivity and monotonicity of $\omega$ (see 6.5 p. 41 of DeVore and Lorentz 1993) that

$$\omega\left(\frac{1}{2m(n)}\right) \geq \frac{1}{2}\omega\left(\frac{1}{m_0}\right) \geq \frac{1}{2}\omega\left(\frac{1}{2m_0 - 2}\right)$$

which together imply (4.21) again. Finally if $m(n) \geq n + 1$ the same argument shows that $\omega^2(1/(2m(n)) > 2$ which concludes the proof. □

*Remarks*:

- Comparing (4.21) with the upper bound in (4.20) one sees that both bounds match except when $n\omega^2(1/2)$ is very small which means that the whole of $\mathscr{S}_\omega$ is so close to the function $\mathbb{1}$ that a good procedure would be to ignore the observations and choose $\tilde{s} = \mathbb{1}$ as the estimator. This would result in a minimax risk of order $\omega^2(1/2)$ smaller than $n^{-1}$. With the number of observations at hand, the parameter space $\mathscr{S}_\omega$ essentially behaves like a single point and the estimation problem is not really meaningful.
- One should keep in mind that although our computations were performed for $s \in \mathscr{F}_\omega$, the upper bound (3.14) makes sense for any $s$. In particular, if one can find some fixed $m_0 \in \mathscr{M}_n$ (at least for large values of $n$) such that $s \in S_{m_0}$, the rate of convergence of our estimator is the parametric one, i.e. $n^{-1}$, since then $K(s, S_{m_0}) = 0$.
- In the Hölderian case considered in Section 3.3.1 the modulus of continuity is given by $\omega(x) = Hx^\alpha$ with $0 < \alpha \leq 1$ resulting in the optimal rate $(H/n^\alpha)^{2/(2\alpha+1)}$.

- One usually works with smoothness conditions on the densities themselves rather than the square roots of the densities. For instance, if the densities satisfy a Hölder condition of the type

$$|f(x) - f(y)| \leq H|x - y|^{\alpha}, \quad \text{for all } x, y \in [0, 1] \ , \tag{4.23}$$

the resulting optimal rate of convergence when the loss is the square of the Hellinger distance is $n^{-2\alpha/(2\alpha+1)}$ provided that the family of densities that we consider is uniformly bounded away from zero, as proved in Birgé (1986). But under such a restriction, the modulus of continuity of $\sqrt{f}$ has the same form (4.23) with a different value of $H$, and the rate $n^{-2\alpha/(2\alpha+1)}$ also derives from our results. On the other hand, let us assume that $H$ is large enough to allow $f$ to be zero on some interval. Then the modulus of continuity of $\sqrt{f}$ still takes the form (4.23) with $\alpha$ replaced by $\alpha/2$ and $H$ by $\sqrt{H}$. The resulting rate is therefore $n^{-\alpha/(\alpha+1)}$ which is the optimal one in this situation as shown in Birgé (1986). If one uses Hellinger distance (which is the $\mathbb{L}_2$-distance between the square roots of the densities) as the loss function, it is natural to put the smoothness restrictions on the set of square roots of densities since one knows that the optimal rate of convergence will be determined by the entropy properties of this set with respect to the $\mathbb{L}_2$-distance.

### 4.1.3. Hölderian densities with unknown anisotropic smoothness

For the sake of simplicity we only considered in the preceding section the classes $\mathcal{S}_\omega$ but one could show, with some additional efforts, that a similar result holds if one replaces them by the more general classes:

$$\mathcal{S}_{a,\omega} = \{s \in \mathcal{S} \,||\, s^{(a)}(x) - s^{(a)}(y)| \leq \omega(|x - y|)\}$$

with $a \in \mathbb{N}$, $a \leq a_0$ and $\omega$ as before. The maximum penalized likelihood estimator reaches again the optimal rate of convergence over the whole family if one replaces the histograms by piecewise polynomials of degree at most $a_0$ in the preceding arguments.

Rather than pursuing in this direction, let us address the multidimensional case. We take this occasion to show that a prior upper bound on the smoothness of $s$ is unnecessary although such a restriction is usually assumed in similar works (see Lepskii, 1991, Donoho, Johnstone, Kerkyacharian and Picard, 1995 and 1996 or Goldenshluger and Nemirovskii, 1997). For the sake of simplicity we shall only consider densities with respect to Lebesgue measure $\mu$ on $[0, 1]^q$ and Hölderian moduli of continuity. For any $\underline{\alpha} = (\alpha_1, \ldots, \alpha_q)$ and $\underline{H} = (H_1, \ldots, H_q)$ belonging to $\mathbb{R}^q$ with

positive coordinates we define $\mathscr{S}(\underline{\alpha}, \underline{H})$ to be the subset of those $s \in \mathscr{S}$ such that the univariate functions $y \mapsto s(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_q)$ belong to $\mathscr{H}(H_i, \alpha_i)$ for all $x \in [0,1]^q$ and $1 \leq i \leq q$ where $\mathscr{H}(H, \alpha)$ is the set of functions $s$ such that $|s|^{(\alpha)}$, as defined by (3.21), is bounded by $H$.

Following the notations of Section 3.2.1, we characterize a space of piecewise polynomials by its maximal degree $r$ with respect to each variable and by a partition of $[0,1]^q$. Let $\mathscr{R}(N)$ denote the regular partition of $[0,1]$ with $N$ pieces. We define $\mathscr{M}_n$ as the set of all $m = (r, \mathscr{R}(N_1), \ldots, \mathscr{R}(N_q))$ with $r \in \mathbb{N}$, $\underline{N} = (N_1, \ldots, N_q) \in [\mathbb{N} - \{0\}]^q$ such that the dimension $D_m = (r+1)^q \prod_{i=1}^q N_i$ of the corresponding space $\bar{S}_m$ of piecewise polynomials is bounded by $n$. Then $S_m = \bar{S}_m \cap \mathscr{S}$ which means that we restrict ourselves to polynomials which are square roots of densities.

**Proposition 5** *Let $\hat{s}$ be the maximum penalized likelihood estimator defined by a penalty function* $\mathrm{pen}(m) = K_1[1 + \log(1 + (2r+1)^q)]D_m/n$ *with $K_1 \geq \kappa_1$. Given $\underline{\alpha}$ and $\underline{H}$ we define $\alpha$ and $H$ by*

$$\frac{q}{\alpha} = \sum_{i=1}^{q} \frac{1}{\alpha_i} \quad \textit{and} \quad H = \left[ \prod_{i=1}^{q} H_i^{1/\alpha_i} \right]^{\alpha/q}$$

*and assume that for any $i$*

$$n^{\alpha} H_i^{2\alpha+q} \geq H^q \ . \tag{4.24}$$

*Then there exists a constant $C(q, \sup_i \alpha_i)$ such that for all $s \in \mathscr{S}(\underline{\alpha}, \underline{H})$*

$$\mathbb{E}_s \left[ d^2(s, \hat{s}) \right] \leq C\left(q, \sup_i \alpha_i\right) \left[ \left( \frac{H^q}{n^{\alpha}} \right)^{2/(2\alpha+q)} + \frac{1}{n} \right] \ .$$

*Proof:* We want to apply Theorem 2. In order to show that (3.13) is satisfied with $L_m = 1$ we notice that

$$(r+1)^q \prod_{i=1}^{q} N_i \geq (r+1)^q - 1 + \prod_{i=1}^{q} N_i \geq (r+1)^q - 1 + q^{-1} \sum_{i=1}^{q} N_i$$

which implies that

$$\sum_{r, \underline{N}} \exp\left[ -(r+1)^q \prod_{i=1}^{q} N_i \right] \leq \left[ \sum_{r \geq 0} \exp\left[ -(r+1)^q + 1 \right] \right]$$

$$\begin{aligned}
&\times \prod_{i=1}^{q} \left[ \sum_{N_i \geq 1} \exp(-N_i/q) \right] \\
&= \mathrm{e} \left[ \sum_{j \geq 1} \exp\left(-j^q\right) \right] \left[ \sum_{j \geq 1} \exp(-j/q) \right]^q \\
&= \Sigma_q \quad < \quad +\infty.
\end{aligned}$$

It also follows from (3.9) that $\bar{r}_m \leq (2r+1)^q$ which justifies our choice for $\mathrm{pen}(m)$.

In order to bound $K(s, S_m)$ we shall provide a control of the $\mathbb{L}_\infty$-distance between $s$ and $\bar{S}_m$ and apply (3.15). Let us begin with a bound on the uniform approximation on a fixed hyperrectangle $\prod_{i=1}^{q}[y_i, y_i + \delta_i]$ of a function $f \in \mathscr{S}(\underline{\alpha}, \underline{H})$ by a polynomial of degree $\leq r = \sup_{1 \leq i \leq q}(a_i)$ where $a_i$ is the largest integer smaller than $\alpha_i$. It follows from Dahmen, DeVore and Scherer (1980, Corollary 3.1) and Schumaker (1981, 13.62 p. 517) that there exists a polynomial $P$ with degree $\leq r$ such that

$$\|f - P\|_\infty \leq C'(q, r) \sum_{i=1}^{q} \delta_i^{a_i} \omega_i(\delta_i)$$

where $C'(q, r)$ is a constant independent of $f$ and the hyperrectangle and

$$\begin{aligned}
\omega_i(\delta_i) = \sup_x \sup_{|h_i| \leq \delta_i} \Bigg| &\frac{\partial^{a_i}}{\partial x_i^{a_i}} f(x_1, \ldots, x_i + h_i, \ldots, x_q) \\
&- \frac{\partial^{a_i}}{\partial x_i^{a_i}} f(x_1, \ldots, x_i, \ldots, x_q) \Bigg| \quad .
\end{aligned}$$

This implies from the definition of $\mathscr{S}(\underline{\alpha}, \underline{H})$ that

$$\|f - P\|_\infty \leq C'(q, r) \sum_{i=1}^{q} H_i \delta_i^{\alpha_i} \quad . \tag{4.25}$$

Let us set

$$\eta = \left( \frac{H^q}{n^\alpha} \right)^{1/(2\alpha+q)}, \quad \delta_i = \left( \frac{\eta}{H_i} \right)^{1/\alpha_i} \quad \text{for } 1 \leq i \leq q$$

and let $N_i$ be the integer such that $\delta_i^{-1} \leq N_i < \delta_i^{-1} + 1$. It follows from (4.24) that $\delta_i \leq 1$ and therefore $1 \leq N_i \leq 2/\delta_i$. Given $m = (r, \mathscr{R}(N_1), \ldots, \mathscr{R}(N_q)) \in \mathscr{M}_n$, (4.25) implies that there exists an element $\bar{s}_m \in \bar{S}_m$ such that

$$\|s - \bar{s}_m\|_\infty \leq C'(q,r) \sum_{i=1}^{q} H_i \delta_i^{\alpha_i} = qC'(q,r)\eta \ .$$

Therefore Theorem 2 and (3.15) imply (since $N_i \leq 2/\delta_i$) that

$$\frac{\mathbb{E}_s\left[d^2(s,\hat{s})\right]}{\kappa_1'} \leq K_1\left[1 + \log\left(1 + (2r+1)^q\right)\right] \frac{(r+1)^q}{n} \prod_{i=1}^{q} \frac{2}{\delta_i} + 12q^2 C'^2(q,r)\eta^2 + \frac{\Sigma_q}{n}$$

and the conclusion follows from our choice of the $\delta_i$'s and $\eta$. □

It follows from Ibragimov and Khas'minskii (1981) or Birgé (1986) that the rate $n^{-2\alpha/(2\alpha+q)}$ is the optimal rate of convergence for functions of anisotropic smoothness.

*Remark*: One should notice that our result holds without any restriction on $\underline{\alpha}$ and that, given $\underline{\alpha}$ and $\underline{H}$, (4.24) always holds for $n$ large enough. Even in the one-dimensional case with $\underline{\alpha} = \alpha$, the assumptions to be found in most papers dealing with adaptation are usually more restrictive, of the type $\alpha > 1/2$ or $\alpha \leq \alpha_0$. Apart from the special situation of Fourier expansions in the white noise setting (Efroimovich and Pinsker 1984), we do not know of any other result of this type valid for arbitrary values of $\alpha$.

### 4.1.4. Projection estimators on polynomials with variable degree

Let us assume now that the observations are drawn according to the unknown density $s$ on $[0, 1]$ belonging to the Besov space $B_{\alpha 2\infty}$ for some unknown $\alpha > 0$ and satisfying $\|s\|_\infty \leq \Phi^2$ where $\Phi^2 \geq 1$ is a known constant. We then define $\mathcal{M}_n$ to be the set of positive integers which are bounded by $n(\log n)^{-4}$ and $S_m$ as the linear space of polynomials of degree bounded by $m$ on $[0, 1]$. It then follows that $D_m = m + 1$ and also from Barron and Sheu (1991, Remark 1 p. 1362) that $\Phi_m \leq \sqrt{D_m}$. Let us set the penalty function to be $\text{pen}(m) = K\Phi^2 D_m/n$ where $K$ is a suitably large constant and let $\hat{s}$ be the penalized projection estimator. It then follows from Theorem 9 below, under the set of conditions **ii)**, that if $s_m$ is the orthogonal projection of $s$ onto $S_m$, the risk of the estimator is bounded by

$$\mathbb{E}_s\left[\|\hat{s} - s\|^2\right] \leq C(K,\Phi) \inf_{m \in \mathcal{M}_n} \left\{\|s - s_m\|^2 + D_m/n\right\} \ .$$

If we denote by $|s|_\alpha$ the Besov semi-norm of $s$ relative to $B_{\alpha 2\infty}$, it follows from DeVore and Lorentz (1993, Theorem 6.3 p. 220) that

$$\|s - s_m\| \leq C_\alpha |s|_\alpha m^{-\alpha}$$

which implies that

$$\mathbb{E}_s\left[\|\hat{s}-s\|^2\right] \leq C(\alpha,\Phi)|s|_\alpha^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)} \quad .$$

4.1.5. Least squares estimators for binary images

Let us now turn to a quite different situation essentially motivated by image analysis. The new framework is of the form $Y_i = s(X_i) + W_i$ where the $W_i$'s are independent identically distributed with a distribution independent of $s$ and $\mu$ denotes the average distribution of the $X_i$'s. We consider some measure $\mu'$ (typically Lebesgue measure on some interval) and some subset $\mathcal{G}$ of $\mathbb{L}_1(\mu')$. We also consider some one-to-one mapping $\chi : g \mapsto \chi_g$ from $\mathcal{G}$ into $\mathbb{L}_2(\mu)$. Since there is no ambiguity here, we denote by the same symbol $\|\cdot\|_p$ the norm in $\mathbb{L}_p(\mu)$ or in $\mathbb{L}_p(\mu')$ and by $d_1$ the distance in $\mathbb{L}_1(\mu')$. The problem is to estimate $s = \chi_f$ for some unknown $f \in \mathcal{G}$. We have in mind here the case where $\chi_f$ is the indicator function of a set the boundary of which is parametrized by the function $f$ belonging to $\mathcal{G}$. For such indicator functions, the square of the $\mathbb{L}_2$-distance is identical to the $\mathbb{L}_1$-distance which is actually the measure of the symmetric difference between the corresponding sets. In good cases (when those sets are epigraphs for instance) this symmetric difference corresponds to the $\mathbb{L}_1$-distance between the functions which parametrize the boundaries. It is therefore natural in such a situation to take the $\mathbb{L}_1(\mu')$-distance $d_1$ as loss function. We consider here a collection of models which are images via $\chi$ of a collection of linear models in $\mathcal{G}$.

**Theorem 5** *Let $\mathcal{G} = \{g \in \mathbb{L}_1(\mu') | F^-(x) \leq g(x) \leq F^+(x)$ for all $x\}$ where $F^+$, $F^- \in \mathbb{L}_1(\mu')$ and let $\chi$ be some non-decreasing mapping from $\mathcal{G}$ into $\{t \in \mathbb{L}_2(\mu) \mid \|t\|_\infty \leq 1\}$. We assume that, for each $m \in \mathcal{M}_n$, $S_m = \chi(G_m)$ where $G_m \subset \mathcal{G}$ is a subset of some linear subspace $\bar{G}_m$ with dimension $D_m$ of $\mathbb{L}_1(\mu')$ and that the following properties are satisfied*:

- $\mathbb{E}[e^{|W_1|/\xi'}] \leq 4$ *for some* $\xi' > 0$ *and* $\{L_m\}_{m\in\mathcal{M}_n}$ *is a family of weights such that*

$$L_m \geq 1 \quad \text{for all } m \in \mathcal{M}_n \quad \text{and} \quad \sum_{m\in\mathcal{M}_n} \exp[-L_m D_m] \leq \Sigma < +\infty \ ;$$

- *there exist two constants* $\Theta_1 \leq \Theta_2$ *independent of* $n$ *such that for all* $h, g \in \mathcal{G}$

$$\Theta_1\|h-g\|_1 \leq \|\chi_h - \chi_g\|^2 \quad \text{and} \quad \|\chi_h - \chi_g\|_1 \leq \Theta_2\|h-g\|_1 \ ; \quad (4.26)$$

- *for each* $m \in \mathcal{M}_n$ *one can find a linear basis* $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ *of* $\bar{G}_m$ *with* $\|\varphi_\lambda\|_1 = 1$ *for all* $\lambda \in \Lambda_m$ *and a constant* $B''_m \geq 1$ *such that*

$$\sum_{\lambda\in\Lambda_m}|\beta_\lambda| \le B''_m \left\|\sum_{\lambda\in\Lambda_m}\beta_\lambda\varphi_\lambda\right\|_1 \quad \textit{for all } (\beta_\lambda)\in\mathbb{R}^{\Lambda_m} \ . \tag{4.27}$$

*Let $\kappa_4$ be a suitable positive numerical constant,*

$$\mathrm{pen}(m) \ge \kappa_4(\xi'+1)^2[L_m + \log(1+\Theta_2 B''_m/\Theta_1) + \log(1+\xi')](D_m/n)$$

*and $\hat{f}$ be the penalized least squares estimator which is a minimizer with respect to $m\in\mathscr{M}_n$ and $g\in G_m$ of $n^{-1}\sum_{i=1}^n[Y_i-\chi_g(X_i)]^2+\mathrm{pen}(m)$. Then for all $f\in\mathscr{G}$ and $s=\chi_f$*

$$\Theta_1\mathbb{E}_s\big[d_1(f,\hat{f})\big] \le \kappa'_4\left[\inf_{m\in\mathscr{M}_n}\{\Theta_2 d_1(f,G_m)+\mathrm{pen}(m)\}+n^{-1}\Sigma(1+\xi')^2\right] \ . \tag{4.28}$$

**Application to binary images:** Here $\mathscr{G}$ is the set of all measurable functions $g$ from $[0,1]$ to $[0,1]$ and for each $g\in\mathscr{G}$ we define for $(x,y)\in[0,1]^2$, $\chi_g(x,y)=1$ if $y\le g(x)$ and 0 otherwise. Following Korostelev and Tsybakov (1993b) we consider the regression framework $Y_i=\chi_f(X_i)+W_i$ and assume that $\mu$ is uniform on $[0,1]^2$ and $\mu'$ is uniform on $[0,1]$. The function $f$ should be understood as the parametrization of a boundary fragment corresponding to some portion of a binary image in the plane. Then $\|\chi_h-\chi_g\|^2=\|\chi_h-\chi_g\|_1=\|h-g\|_1$ and we may take $\Theta_1=\Theta_2=1$ in (4.26). Assuming that the errors $W_i$ are either bounded by 1 or Gaussian with variance smaller than 1 we can take $\xi'=1$.

Let $\mathscr{R}(J)$ denote the regular partition of $[0,1]$ with $J$ pieces and $\mathscr{M}_n$ be the set $\{(r,J)|J\ge1,\ r\in\mathbb{N}\}$. Following the definition of Section 3.2.1, we consider $\bar{G}_m$ to be the space of piecewise polynomials of degree not larger than $r$ based on the regular partition $\mathscr{R}(J)$ if $m=(r,J)$. Then $D_m=(r+1)J$ and choosing $L_m=1$ we can take $\Sigma=1$. Let us now turn to the verification of (4.27). We consider some orthonormal (with respect to $\mathbb{L}_2(\mu')$) basis $\psi_0,\dots,\psi_r$ of the space of polynomials on $[0,1]$ with degree $\le r$ and we define $\varphi_l=\alpha_l\psi_l$ with $\alpha_l\ge1$ by $\|\varphi_l\|_1=1$. Then for any $(\beta_l)\in\mathbb{R}^{r+1}$

$$\begin{aligned}\sum_{l=0}^r|\beta_l| \le \sum_{l=0}^r|\alpha_l\beta_l| &\le \left[(r+1)\sum_{l=0}^r|\alpha_l\beta_l|^2\right]^{1/2}\\ &= (r+1)^{1/2}\left\|\sum_{l=0}^r\beta_l\varphi_l\right\|\\ &\le 2r(r+1)^{1/2}\left\|\sum_{l=0}^r\beta_l\varphi_l\right\|_1 \end{aligned}\tag{4.29}$$

where we successively used Cauchy-Schwarz inequality and DeVore and Lorentz (1993, Theorem 2.6 p. 102) about the relations between norms of polynomials. Starting from the basis $\{\varphi_l\}_{0\le l\le r}$ we build a linear basis $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ of $\bar{G}_m$ where $\Lambda_m = \{(j,l) \mid 1 \le j \le J,\ 0 \le l \le r\}$ such that $\|\varphi_\lambda\|_1 = 1$ for $\lambda \in \Lambda_m$ given by

$$\varphi_{j,l}(x) = J\varphi_l\left[J\left(x - \frac{j-1}{J}\right)\right] \quad .$$

It then easily follows from (4.29) that (4.27) is satisfied with $B_m'' = 2r(r+1)^{1/2}$. We finally define $G_m = \bar{G}_m \cap \mathscr{G}$.

Assume now that $f \in \mathscr{G}$ belongs to some Hölder space $\mathscr{H}_\alpha$ as defined by (3.21) where $\alpha > 0$ is unknown. Keeping in mind that $\mathscr{H}_\alpha$ is a subset of the Besov space $B_{\alpha\infty\infty}([0,1])$, it follows from Lemma 12 in Section 8 that $\|f - g_m\|_\infty \le \varepsilon = C(r)|s|^{(\alpha)}J^{-\alpha}$ for some $g_m \in \bar{G}_m$ with $m = (r, J)$ and $r > \alpha - 1$. Changing if necessary $g_m$ into $(g_m + \varepsilon)/(1+2\varepsilon)$ and $\varepsilon$ into $4\varepsilon$ we can assume that $g_m \in G_m$. Choosing

$$\mathrm{pen}(m) = K_4\left[1 + \log\left(1 + 2r(r+1)^{1/2}\right)\right](r+1)J/n \quad \text{with} \quad K_4 \ge 4\kappa_4$$

we derive from Theorem 5 that if $s = \chi_f$ then

$$\mathbb{E}_s\left[d_1(f,\hat{f})\right] \le C'(r)\left[|s|^{(\alpha)}\right]^{1/(1+\alpha)} n^{-\alpha/(1+\alpha)} \quad .$$

*Remarks*:

- The rates may seem unusual as compared to density estimation. It results from the fact that $\|\chi_h - \chi_g\|^2 = d_1(h,g)$ which leads to a risk expressed as the sum of a bias term and a variance term instead of the classical variance plus bias squared. It comes from Korostelev and Tsybakov (1993b, Theorem 3.3.2) that these rates are optimal (in the minimax sense) when $\alpha$ is known.
- One could consider analogously star-shaped images. In this case we describe a point of the Euclidean unit disk by its polar coordinates $\rho, \psi$ with $0 \le \rho \le 1$ and $\psi$ belonging to the one-dimensional torus $\mathbb{T}$. We then define $\mathscr{G}$ as the set of functions $g$ from $\mathbb{T}$ to $[0,1]$ and set $\chi_g(\rho\cos\psi, \rho\sin\psi) = \mathbb{1}_{\{\rho^2 \le g(\psi)\}}(\rho,\psi)$. Choosing $\mu$ as the uniform distribution on the disk and $\mu'$ as Lebesgue measure on $\mathbb{T}$ we can check that (4.26) is satisfied with $\Theta_1 = \Theta_2 = 1/2$.

### 4.1.6. Estimation of the support of a density

Let $\mu$ be the restriction to the unit disk $\mathbb{D} \subset \mathbb{R}^2$ of the Lebesgue measure on $\mathbb{R}^2$ and $s = \mathbb{1}_{\Omega_s}$ be the indicator function of some measurable subset $\Omega_s$

of $\mathbb{D}$. We observe $n$ independent identically distributed random variables $Z_1, \ldots, Z_n$ with density $fs$ with respect to $\mu$. Here $s$ and $f$ are unknown and we want to estimate $s$, assuming that $f$ satisfies $0 < a \leq f(x) \leq b$ for all $x \in \mathbb{D}$, $a \leq 1$ and $b$ being known constants. This estimation problem is considered in Korostelev and Tsybakov (1993a) where minimax rates of convergence on some smoothness classes are given. The novelty here is that, applying our model selection method, we construct adaptive estimators.

In order to estimate $s$ we define $\mathscr{T}$ as the set of indicator functions of measurable subsets of the unit disk $\mathbb{D}$ and consider the contrast function $\gamma(z, t) = -t(z) + (a/2)\int t\, d\mu$. Keeping in mind that $s$ and $t$ are indicator functions and setting $u = st = s \wedge t$ one gets

$$\begin{aligned}\mathbb{E}_s[\gamma(Z_1, t) - \gamma(Z_1, s)] &= \int (s-t) fs\, d\mu \\ &\quad + \frac{a}{2}\left[\int (t-u)\, d\mu - \int (s-u)\, d\mu\right] \\ &= \int (s-u)(f - a/2)\, d\mu + \frac{a}{2}\int (t-u)\, d\mu\end{aligned}$$

and

$$\|t - s\|^2 = \|t - s\|_1 = \int (s-u)\, d\mu + \int (t-u)\, d\mu \ .$$

We then derive from the bounds on $f$ that

$$(a/2)\|t - s\|^2 \leq \mathbb{E}_s[\gamma(Z_1, t) - \gamma(Z_1, s)] \leq (b - a/2)\|t - s\|^2 \ . \quad (4.30)$$

We also assume that the set $\Omega_s$ is starshaped and that its boundary is parametrized in polar coordinates $(\rho, \psi)$ by $\rho^2 = g_s(\psi)$ for $\psi$ belonging to the one-dimensional torus $\mathbb{T}$. The reader should notice here that we introduce an unusual parametrization of the boundary. It is therefore natural to restrict the models $S_m$ to starshaped subsets of the disk with a boundary parametrized in polar coordinates. More precisely, given a function $g$ from the torus $\mathbb{T}$ to $\mathbb{R}$ we set $\tilde{g}(x) = [0 \vee g(x)] \wedge 1$ and define the mapping $\chi$ from $\mathbb{R}^{\mathbb{T}}$ to $\mathscr{T}$ by $g \xrightarrow{\chi} \chi(g) = \chi_g$ given by $\chi_g(\rho \sin\psi, \rho\cos\psi) = \mathbb{1}_{\{\rho^2 \leq \tilde{g}(\psi)\}}(\rho, \psi)$. Denoting by $\mu'$ the Lebesgue measure on $\mathbb{T}$ (with $\mu'(\mathbb{T}) = 2\pi$) and by $d_1$ the distance in $\mathbb{L}_1(\mu')$, one gets

$$\|\chi_{g_1} - \chi_{g_2}\|_1 = \pi d_1(\tilde{g}_1, \tilde{g}_2) \leq \pi d_1(g_1, g_2) \quad \text{for all } \ g_1, g_2 \in \mathbb{L}_1(\mu') \ . \quad (4.31)$$

In order to define a family of models we start with a family $\{\bar{G}_m\}_{m \in \mathscr{M}_n}$ of finite dimensional linear subspaces of $\mathbb{L}_1(\mu')$ and denote by $D_m$ the dimension of $\bar{G}_m$. Given some positive constant $R$ we then set $G_m = \{g \in$

$\bar{G}_m \mid \int |g|\, d\mu' \leq R\}$ and define $S_m$ as the image of $G_m$ by the mapping $\chi$. Then the following theorem to be proved in Section 7 holds

**Theorem 6** *Assume that the family of models $\{S_m\}_{m\in\mathcal{M}_n}$ is defined as indicated before and that $\sup_{m\in\mathcal{M}_n} D_m \leq 25\pi bnR/2$. Let $\{L_m\}_{m\in\mathcal{M}_n}$ be a family of weights such that*

$$L_m \geq 1 \quad \textit{for all} \quad m \in \mathcal{M}_n \quad \textit{and} \quad \sum_{m\in\mathcal{M}_n} \exp[-L_m D_m] \leq \Sigma < +\infty \ .$$

*Assume that for each $m \in \mathcal{M}_n$ one can find a constant $B_m'' \geq 1$ and a linear basis $(\varphi_\lambda)_{\lambda\in\Lambda_m}$ of $\bar{G}_m$ such that $\|\varphi_\lambda\|_1 = 1$ for all $\lambda$ and*

$$\sum_{\lambda\in\Lambda_m} |\beta_\lambda| \leq B_m'' \Big\| \sum_{\lambda\in\Lambda_m} \beta_\lambda \varphi_\lambda \Big\|_1 \quad \textit{for all} \quad (\beta_\lambda) \in \mathbb{R}^{\Lambda_m} \ . \tag{4.32}$$

*Let $\kappa_5$ be a suitable positive numerical constant,*

$$\mathrm{pen}(m) \geq \frac{\kappa_5}{a}\left[L_m + \log\left(1 + \frac{nB_m'' Rb}{a^{1/2} D_m}\right)\right]\frac{D_m}{n}$$

*and $\hat{s}$ be the minimum penalized empirical contrast estimator which is a minimizer with respect to $m \in \mathcal{M}_n$ and $t \in S_m$ of $\mathrm{pen}(m) - n^{-1}\sum_{i=1}^n t(Z_i) + a\|t\|_1/2$. If $\Omega_s$ is starshaped with $s = \chi_{g_s}$ and $0 \leq g_s \leq 1$ then*

$$\mathbb{E}_s\left[d^2(s,\hat{s})\right] \leq \kappa_5'\left[\inf_{m\in\mathcal{M}_n}\left\{d_1(g_s, G_m) + a^{-1}\mathrm{pen}(m)\right\} + a^{-2}\Sigma n^{-1}\right]$$

*where $d_1$ denotes the distance in $\mathbb{L}_1(\mu')$.*

*Remark*: One can always take $R = 4\pi$. Indeed, since $0 \in G_m$, $d_1(g_s, G_m) \leq d_1(g_s, 0) \leq 2\pi$ which shows that taking $R > 4\pi$ cannot improve the distance $d_1(g_s, G_m)$.

A natural basis to be considered in this framework is the Fourier basis (correctly normalized in order to have $\|\varphi_\lambda\|_1 = 1$) defined by $\varphi_0 = \mathbb{1}/(2\pi)$, $\varphi_{2j-1}(x) = \cos(jx)/4$, $\varphi_{2j}(x) = \sin(jx)/4$ for $j \geq 1$. Defining $\Lambda(0) = \{0\}$, $\Lambda(j) = \{2j-1; 2j\}$ for $j \geq 1$ and $\Lambda_m = \sum_{j=0}^m \Lambda(j)$ we choose $\bar{G}_m$ to be the linear span of $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$ and $G_m = \{g \in \bar{G}_m \mid \|g\|_1 \leq 4\pi\}$. Then $D_m = 2m+1$ and we take $L_m = 1$. We can check (4.32) exactly as we did for (4.29). If $\alpha_0 = (2\pi)^{-1/2}$ and $\alpha_\lambda = \sqrt{\pi}/4$ for $\lambda \neq 0$ then $\alpha_\lambda^{-1} \leq \sqrt{2\pi}$ for all $\lambda$ and $\{\alpha_\lambda^{-1}\}_{\lambda\in\Lambda_m}$ is an orthonormal basis for $\bar{G}_m$. Then

$$\sum_{\lambda\in\Lambda_m}|\beta_\lambda| \le \sqrt{2\pi}\sum_{\lambda\in\Lambda_m}|\alpha_\lambda\beta_\lambda| \quad \le \quad \left[2\pi D_m \sum_{\lambda\in\Lambda_m}(\alpha_\lambda\beta_\lambda)^2\right]^{1/2}$$
$$= \sqrt{2\pi D_m}\left\|\sum_{\lambda\in\Lambda_m}\alpha_\lambda\beta_\lambda\left(\alpha_\lambda^{-1}\varphi_\lambda\right)\right\|$$
$$\le \sqrt{2\pi D_m}\left(\frac{D_m}{2\pi}\right)^{1/2}\left\|\sum_{\lambda\in\Lambda_m}\beta_\lambda\varphi_\lambda\right\|_1$$

where the last inequality comes from DeVore and Lorentz (1993, inequality 2.15 p. 102) and therefore (4.32) is satisfied with $B''_m = D_m$. This leads to the choice $\mathrm{pen}(m) = K_5 a^{-1}[1+\log(1+4\pi nba^{-1/2})](2m+1)/n$ with $K_5 \ge \kappa_5$. If $g_s$ belongs to some Besov space $B_{\alpha\,1\,\infty}$ (see the precise definition in Lemma 12 below) of functions on $\mathbb{T}$, it follows from Lemma 12 that $d_1(g_s, G_m) \le C(g_s)m^{-\alpha}$ and therefore $m = (n/\log n)^{1/(1+\alpha)}$ gives a rate of convergence of order $(\log n/n)^{\alpha/(1+\alpha)}$. By standard perturbation arguments of the type used in Proposition 4 one could show that this rate is optimal in the minimax sense, up to the $\log n$ factor, when $\alpha$ is known.

*Remark*: We considered here the Fourier basis for the sake of simplicity but one could use periodic wavelets as well (periodic wavelets are defined for instance in Daubechies 1992, Section 9.3). Such a localized basis would lead to a bounded family $\{B''_m\}_{m\in\mathcal{M}_n}$.

### 4.2. *"Rich" families of models*

By this we mean families for which the number of models of a given dimension $D$ is so large that the sumability condition

$$\sum_{m\in\mathcal{M}_n}\exp[-L_m D_m] \le \Sigma < +\infty \tag{4.33}$$

requires unbounded values of $L_m$. These families are much bigger than the preceding ones but we shall see from the examples that a modest increase of $L_m$ can provide much better approximation properties. A typical example is given by histograms with arbitrary binwidths compared to the histograms with equal binwidths considered above. The price to pay for this potentially better adequation of some of our models to the true value of $s$ is to be found in the requirement that the series $\sum_{m\in\mathcal{M}_n}\exp[-L_m D_m]$ should converge. This is not possible anymore if the $L_m$'s are bounded and we shall have to take some of the $L_m$'s of order $\log n$ which will result in the presence of an extra $\log n$ factor in the quadratic risk.

4.2.1. Histograms with variable binwidths and spatial adaptation

Let's go back to maximum likelihood estimation with $n$ independent identically distributed observations from an unknown density $s^2$ on $[0, 1]$. We choose for our family $\{S_m\}_{m\in\mathcal{M}_n}$ of approximating spaces the rich family described in Section 3.3.1 with $\mathcal{M}_n = \mathcal{R}_n \cup (\cup_{N\geq 3}\mathcal{G}_{n,N})$. It has been mentioned already that the corresponding maximum penalized likelihood estimator had the right rate of convergence if the true $s$ was $\alpha$-Hölderian with index $\alpha \in (0, 1]$. Let us now assume that $s$ has a bounded $\alpha$-variation with $0 < \alpha \leq 1$ which means that

$$\sup_{k\geq 2} \sup_{x_1\leq\cdots\leq x_k} \sum_{j=2}^{k} |s(x_j) - s(x_{j-1})|^{1/\alpha} = J_\alpha(s) < +\infty \tag{4.34}$$

where the supremum is taken over all increasing sequence $x_1 < \cdots < x_k$ of points in $[0, 1]$. It follows from Proposition 8 and Proposition 1 that if $1 \leq L \leq N$ there exists some $m \in \mathcal{G}_{n,N}\cup\mathcal{R}_n$ such that $D_m \leq 2(N/L)^{1/(1+2\alpha)}+1$ and

$$K(s, S_m) \wedge 1 \leq 9J_\alpha^{2\alpha}(s)\,(L/N)^{(2\alpha)/(2\alpha+1)} \quad .$$

Since one can only assume that $L_m \leq 2[1 + \log(N/D_m)] \leq 2[1 + \log N]$ and $\bar{r}_m \leq (N/D_m)^{1/2}$, Theorem 2 implies that if pen$(m)$ is chosen as in (3.17),

$$\mathbb{E}_s\left[d^2(s, \hat{s})\right] \leq C(K)\left[\inf_{N\geq 1}\inf_{1\leq L\leq N}\left\{J_\alpha^{2\alpha}(s)\left(\frac{L}{N}\right)^{2\alpha/(2\alpha+1)} + \frac{\log(N+1)}{n}\left(\frac{N}{L}\right)^{1/(2\alpha+1)}\right\}\wedge 1\right] .$$

Setting $J = J_\alpha^{2\alpha}(s) \vee n^{-1}$ we evaluate the bound at $N = [nJ]$ with $L = 1$ if $N = 1$ and $L = (3/2)(nJ)^{-1}N\log(N + 1)$ otherwise. Then $L$ satisfies $1 \leq L \leq N$ and we get a risk bound of the form

$$C'(K)\left[J^{1/(2\alpha+1)}\left(\frac{\log(nJ+1)}{n}\right)^{2\alpha/(2\alpha+1)}\right]\wedge 1 \quad \text{with } J = J_\alpha^{2\alpha}(s)\vee n^{-1} \quad . \tag{4.35}$$

*Remarks*:

- When $J_\alpha^{2\alpha}(s)$ is of the order of $n^{-1}$ or smaller, the risk bound is of order $n^{-1}$ as in the parametric case. Otherwise the risk is bounded by

$$C''\left[n^{-1}(\log n)J_\alpha(s)\right]^{2\alpha/(2\alpha+1)} \quad .$$

- One should always keep in mind that whatever the true function $s$ the right-hand side of (3.14) provides the best compromise, among all the histograms at hand, between $K(s, S_m)$ and $n^{-1}D_m \log(1 + N/D_m)$ even when $s$ does not belong to the particular smoothness classes considered above.
- It is worth mentioning here that when $s$ is decreasing on $[0, 1]$, the Grenander estimator, which is the derivative of the least concave majorant of the empirical distribution function, automatically achieves a bound on the $\mathbb{L}_1$-risk which is analogous to (4.35) with $\alpha = 1$ but without the $\log n$ factor (see Birgé 1989).
- In order to get better approximation properties for smoother densities, one could replace histograms by piecewise polynomials of degree $\leq r$. This is possible and would lead to various rates of convergence for various smoothness classes (not necessarily homogeneous) at the price of additional technicalities. For the sake of simplicity we shall not insist on this here.

### 4.2.2. Neural nets and related nonlinear models

We assume now the situation described in Section 3.2.2. Risk bounds for minimum penalized empirical contrast estimators are stated for the models derived from $\bar{S}_m = \{\sum_{j=1}^{D'} \beta_j \phi_{w_j}(x)\}$ where $\sum_{j=1}^{D'} |\beta_j| \leq R$, $|w_j|_1 \leq H$, and the index $m = (D', H, R)$ is taken as a triplet of positive integers. We recall here that the number of free parameters in $\bar{S}_m$ which will play the role of $D_m$ is $D'(q' + 1)$.

In keeping with the general framework of Section 3.1, we consider the case of penalized likelihood density estimation with densities of the form $t^2(x)$ for $t$ in $S_m$. Here the densities are taken with respect to a given *probability* measure $\mu$ on $[-1, 1]^q$ and $s^2$ is the true probability density. The set $S_m$ is taken to be those functions in $\bar{S}_m$, the positive part of which has a norm at least $1/2$, clipped from below to be not smaller than $1/n$, with each divided by its norm in $\mathbb{L}_2(\mu)$. We also consider the case of penalized least squares regression with data of the form $Y_i = s(X_i) + W_i$ where the $W_i$'s are independent identically distributed centered errors and with target function $s$ bounded by a known constant $\xi$. We take advantage of this knowledge by taking the least squares estimates in $S_m$, where $S_m$ consists of the functions in $\bar{S}_m$, clipped to the range $[-\xi, \xi]$. Such clipping is done to satisfy a boundedness condition without adversely affecting the approximation and metric entropy properties of the models.

In addition to the Lipschitz condition (3.12) we require that $|\phi_w(x)| \leq 1 \vee |w|_1$ for $x$ in $[-1, 1]^q$. This condition is verified in the examples by noting either that $\phi_w$ is bounded by one (which handles most of the cases of

interest) or that in some cases $\phi_0$ is identically 0 so that then $|\phi_w(x)| \le |w|_1$ by the Lipschitz condition.

**Theorem 7** *Let $\{\phi_w : w \in \mathbb{R}^{q'}\}$ be a parameterized family of functions that satisfies the Lipschitz condition $\|\phi_w - \phi_{w'}\|_\infty \le |w - w'|_1$ and suppose that $\|\phi_w\|_\infty \le 1 \vee |w|_1$.*

- *For maximum likelihood density estimation we define*

$$S_m = \left\{ \frac{t \vee n^{-1}}{\|t \vee n^{-1}\|} \;\middle|\; t \in \bar{S}_m \;\text{ and }\; \|t \vee 0\| \ge \frac{1}{2} \right\}$$

*and take*

$$\mathrm{pen}(m) \ge \kappa_6 \frac{D'q'}{n} \left[ 1 + \log(RH) + \log\left( 1 + \frac{n}{D'q'} \right) \right] \tag{4.36}$$

*where $\kappa_6$ is a suitable numerical constant. We define $\hat{s}$ to be a minimizer with respect to positive integers $D'$, $H$, and $R$ which satisfy $D'(q'+1) \le n$ and with respect to $t \in S_m$ of $-n^{-1}\sum_{i=1}^n \log[t(X_i)] + \mathrm{pen}(m)$, then*

$$\mathbb{E}_s\left[d^2(s,\hat{s})\right] \le \kappa_6' \left[ \inf_{D',H,R} \{K(s,S_m) + \mathrm{pen}(m)\} \wedge 1 \right] \tag{4.37}$$

$$\le \kappa_6'' \inf_{D',H,R} \left\{ \left[ d^2(s,\bar{S}_m) \vee n^{-2} \right] \times [1 + \log(n\|s\|_\infty)] + \mathrm{pen}(m) \right\} . \tag{4.38}$$

- *For the regression case $Y_i = s(X_i) + W_i$ we assume that $s$ is bounded by a known constant $\xi$ and that $\mathbb{E}_s[e^{|W_1|/\xi'}] \le 4$. We define $S_m = \{[t \vee (-\xi)] \wedge \xi \mid t \in \bar{S}_m\}$ and choose*

$$\mathrm{pen}(m) \ge \kappa_7(\xi + \xi')^2 \frac{D'q'}{n} \left[ 1 + \log(RH) + \log\left( 1 + \frac{n}{D'q'} \right) \right] \tag{4.39}$$

*where $\kappa_7$ is a suitable numerical constant. We take $\hat{s}$ to be a minimizer with respect to positive integers $D'$, $H$, and $R$ and $t \in S_m$ of $n^{-1}\sum_{i=1}^n [Y_i - t(X_i)]^2 + \mathrm{pen}(m)$, then*

$$\mathbb{E}_s\left[d^2(s,\hat{s})\right] \le \kappa_7' \inf_{D',H,R} \left\{ d^2(s,\bar{S}_m) + \mathrm{pen}(m) \right\} .$$

**Statistical rate bounds using multivariate nonlinear additive ridge models:** We now restrict to the case where the function $\phi_w$ is a ridge function

on $\mathbb{R}^q$ : $\phi_w(x) = \psi(a^T x + b)$ where $w = (a, b)$ with $a \in \mathbb{R}^q$ and $b \in \mathbb{R}$. We first state bounds on nonlinear approximation and estimation using linear combinations of functions of ridge type using Fourier conditions on the target function (building on the work of Jones 1992, Barron 1993, Breiman 1993, Hornik et al. 1994 and Yukich et al. 1995). The proof will be given in Section 8.

**Proposition 6** *Let $s(x)$ be a real-valued function on $[-1, 1]^q$ with a Fourier representation*

$$s(x) = \int \exp\left[\mathrm{i}a^T x\right] \tilde{F}(da)$$

*with respect to a complex-valued measure $\tilde{F}$ for frequency vectors $a$ in $\mathbb{R}^q$. For any $\gamma \geq 0$, we denote by $c_{s,\gamma} = \int |a|_1^\gamma F(da)$ the $\gamma$-absolute moment of the Fourier magnitude distribution $F = |\tilde{F}|$. We assume that for certain $\alpha > 0$, $c_{s,\alpha} + c_{s,0}$ is finite and that $\alpha$ and the ridge function $\psi$ satisfy the following constraints*:

1. *Trigonometric approximation: $\psi(x) = \cos x$ and $\alpha > 0$;*
2. *Sigmoidal approximation: $\psi(x) \to \pm 1$ and approaches its limits at least polynomialy fast as $x \to \pm\infty$ and $\alpha = 1$;*
3. *Wavelet ridge approximation: $\psi(x)$ is a bounded function with compact support and $\alpha > 1$;*
4. *Hinged hyperplanes: $\psi(x) = x \vee 0$ and $\alpha = 2$.*

*Then in each case, there exists some constants $H_0$ and $R(s)$ such that*

$$d(s, \bar{S}_m) \leq c_{s,\alpha}\delta_H + R(s)(D')^{-1/2} \tag{4.40}$$

*provided that $m$ is such that $R \geq R(s)$ and $H \geq H_0$, where $\delta_H$ does not depend on $s$ and decreases at least polynomialy fast with respect to $H$ as $H$ goes to infinity.*

We now assume that $\psi$ is a Lipschitz function of one of the forms mentioned in Proposition 6 and bounded by 1 so that we can combine the conclusions of Theorem 7 with a value of $\mathrm{pen}(m)$ of the order of the lower bound given by (4.36) or (4.39) and Proposition 6. Let $\hat{s}$ be the minimum penalized empirical contrast estimator, taking the minimum over $D'$, $H$, and $R$ as in Theorem 7. To bound the accuracy index, under the conditions of Proposition 6, note that when $H$ is a convenient power of $D'$, $d(s, \bar{S}_m)$ is of order $(D')^{-1/2}$. Then optimizing over $D'$, we conclude that the risk $\mathbb{E}_s[d^2(s, \hat{s})]$ is of order $n^{-1/2} \log n$ or $(\log n/n)^{1/2}$ in each of the two cases respectively.

Though we are building on previous approximation results, as far as we are aware these are the first statistical rate bounds of this sort stated for

the trigonometric, ridge wavelet, and hinged hyperplane cases. Comparable rate results for the neural network regression case are in Barron (1994) (under the more stringent assumption that the response $Y$ is bounded and that optimization is taken over a discretized grid of parameter values) and in Modha and Masry (1996).

The regularity conditions on $s$ needed for the approximation controls are given in terms of integrability conditions on its Fourier transform. Since larger values of $\alpha$ correspond to more stringent conditions, the assumptions above are more general for the trigonometric model than for the others, which is natural given that the conditions are imposed on the Fourier spectrum. The point in considering the other models is to give some risk bounds for these popular models under reasonably well understood conditions. We note that for the classes of functions considered here, the approximation and estimation rates as exponents of $1/D'$ and $1/n$ are independent of the dimension $q$. Actually, the dependence on the dimension is indirect through the spectral norms $c_{s,\alpha}$. Conditions under which these norm are not excessively large are discussed in Barron (1993).

The key to achieving these advantageous rates for these functions is the adaptation of the nonlinear parameters $w_j$ to fit the target. In contrast linear approximation would be forced to specify a fixed basis without adaptation to the target function. Indeed, it is also shown in Barron (1993) that for the class of functions with a bound on $c_{s,1} + |s(0)|$, the best $\mathbb{L}_2$-approximation by a fixed $D$ term basis is not uniformly smaller than order $D^{-1/q}$. Thus, without adaptation, we approximate functions in this class no better than for the much larger class with a bound on the gradient. Whereas, with adaptation, we approximate functions in this class at rate $D^{-1/2}$, comparable to the approximation rate of the much smaller subclass of functions that have bounds on all derivatives up to a certain high order.

### 4.2.3. Model selection with a bounded basis

We want to do density estimation using projection estimators as described in Section 3.3.2 and assuming that the basis $\{\varphi_\lambda \mid \lambda \in \bar{\Lambda}_n\}$ is a finite subset of cardinality $n^l$ ($l$ being some fixed positive integer) of the Fourier basis on the torus $\mathbb{T}$ with uniform distribution $\mu$. With such a basis (which is orthonormal and bounded by $\sqrt{2}$) we can apply Theorem 9 Case i) to be stated in Section 6.3 below. We look for a representation of $s$ with a small number of parameters (as compared to the number of observations). This looks rather attractive if one thinks of the model selection point of view. Let us therefore define our family $\{S_m\}_{m\in\mathcal{M}_n}$ as follows. Assuming that $n \geq 4$ we define $\mathcal{M}_n$ to be the (finite) collection of nonvoid subsets $m$ of $\bar{\Lambda}_n$ of cardinality bounded by $K_n$ where $K_n$ is the smallest integer $\geq \sqrt{n}(\log n)^{-2}$.

The reason for bounding the cardinality of $m$ in such a way is that Theorem 9 involves in this case a quantity of the type $\sum_{m\in\mathcal{M}_n} \exp[-x(L_m D_m \wedge \sqrt{n})]$ for some small positive number $x$. In order to bound this quantity, we choose $\Lambda_m = m$ and $L_m = (\log n)^2$. Since $\binom{n^l}{i} \le n^{li}/i!$ we can bound the second term of (6.20) by

$$\sum_{i=1}^{K_n} \binom{n^l}{i} \exp\left[-x\mathrm{i}(\log n)^2\right] \le \sum_{i=1}^{\sqrt{n}} \frac{1}{i!} \exp\left[-\mathrm{i}\log n\,(x\log n - l)\right]$$

which is bounded independently of $n$ and the first term of (6.20), using Lemma 6, by

$$\sum_{i=1}^{K_n} \binom{n^l}{i} \exp\left(-x\sqrt{n}\right) \le \left(\frac{en^l}{K_n}\right)^{K_n} \exp\left(-x\sqrt{n}\right)$$

which is also bounded independently of $n$ from our choice of $K_n$. With the choice $\mathrm{pen}(m) = K_6 L_m D_m/n$ for a large enough constant $K_6$, the penalized projection estimator provides (up to a $(\log n)^2$ factor due to our choice of $L_m$) a risk which realizes the best trade-off between bias and variance among our family of models. Moreover it has the simple expression $\sum_{\lambda\in\hat{\Lambda}} \hat{\beta}_\lambda \varphi_\lambda$ where $\hat{\Lambda}$ is the set of indices corresponding to the at most $K_n$ largest empirical coefficients $\hat{\beta}_\lambda$ which are also larger than some threshold $C(\log n/n)^{1/2}$. This type of procedure could be useful to estimate a density which is known to have a small number of non-zero Fourier coefficients. It leads (up to a $(\log n)^2$ factor) to the right rate of estimation although one ignores what are the coefficients to be estimated. A much more detailed treatment of selection of subsets of a basis and its relationship to threshold estimators is to be found in Birgé and Massart (1997).

## 5. Adaptation and model selection

Although this terminology is widely used, we do not know of any "universal" definition of *adaptation*. On the contrary, one can find in the literature different notions of adaptation. This is one purpose of our discussion to analyze and compare the various points of view. We assume that the unknown element $s$ to be estimated is a function belonging to some functional space $\mathcal{S}$ (typically $\mathcal{S} = \mathbb{L}_p(\mu)$ for some $p \ge 1$) and that a loss function $\ell$ is given on $\mathcal{S}\times\mathcal{S}$ (typically $\ell$ is some power of the $\mathbb{L}_p$-distance). To be more formal, let us say that we observe $X^{(n)}$, the distribution of which depends on an unknown function $s$. We have here in mind examples such as

$X^{(n)} = (X_1, \ldots, X_n)$ is a sample of density $s$ (or $s^2$) or $X^{(n)} = \{X_{t,n}\}_{0 \le t \le 1}$ is given by the white noise setting

$$dX_{t,n} = s(t)\,dt + n^{-1/2}\,dW_t$$

where $W_t$ denotes a standard Brownian motion originating from 0, among other settings (regression function, spectral density estimation, ...). Given an estimator $\hat{s}_n(X^{(n)})$ depending on the observation, the risk $R_n(\hat{s}_n, s)$ of this estimator at point $s$ is given by

$$R_n(\hat{s}_n, s) = \mathbb{E}_s[\ell(s, \hat{s}_n)] \ .$$

The maximal risk of $\hat{s}_n$ over some parameter space $\mathbb{S}$ and the minimax risk over $\mathbb{S}$ are respectively defined by

$$R_n(\hat{s}_n, \mathbb{S}) = \sup_{u \in \mathbb{S}} R_n(\hat{s}_n, u) \quad \text{and} \quad R_n(\mathbb{S}) = \inf_{\tilde{s}_n} R_n(\tilde{s}_n, \mathbb{S})$$

where the infimum is taken with respect to all possible estimators $\tilde{s}_n$. One can distinguish between two main approaches to adaptation.

- One considers some collection $\{\mathbb{S}_\theta\}_{\theta \in \Theta}$ of subsets of $\mathcal{S}$ (typically a collection of balls for some smoothness semi-norms) and we look for estimators which are approximately minimax simultaneously on all the $\mathbb{S}_\theta$'s. This is what we shall call *adaptation in the minimax sense*.
- One considers a family of estimators $\hat{s}_m$ depending on a tuning parameter $m \in \mathcal{M}$ (for instance $m$ can be the bandwidth of a kernel estimator or an arbitrary subset of some finite dimensional basis for a projection estimator) and we look for a data driven choice $\hat{m} \in \mathcal{M}$ such that, whatever the true $s \in \mathcal{S}$, the risk of $\hat{s}_{\hat{m}}$ reaches approximately the minimal risk among the family of estimators $\{\hat{s}_m\}_{m \in \mathcal{M}}$. This is what we shall call *adaptation to the target function*.

Actually, from a constructive point of view, all the solutions to the first problem that we know rely on a data driven choice of a tuning parameter $m \in \mathcal{M}$ for some given family $\{\hat{s}_m\}_{m \in \mathcal{M}}$ in the situation when for each $\theta \in \Theta$ one can find $m(\theta)$ such that the estimator $\hat{s}_{m(\theta)}$ is approximately minimax for $\mathbb{S}_\theta$ (we shall below make precise what we exactly mean by "approximately"). This points out the close connection between the first approach and the second.

### *5.1. Adaptation in the minimax sense*

A simple example for density estimation is the following: the unknown function $s$ belongs to some subset

$$\mathbb{S}_{\alpha,H} = \left\{ s \in \mathbb{L}_2([0;1]) \;\middle|\; \int \left|s^{(\alpha)}(x)\right|^2 dx \leq H^2 \right\},$$

with $H > 0$ and $\alpha \in \mathbb{N} - \{0\}$, of the Sobolev space $W_2^\alpha([0;1])$. One wants to estimate $s$ using (for instance) a kernel estimator of a given form with bandwidth $m$. If $\alpha$ and $H$ were known, one would now how to choose the bandwidth $m$ optimally as a function of $\alpha$, $H$, $n$ in order to get a quadratic risk uniformly bounded over $\mathbb{S}_{\alpha,H}$ by $\kappa(Hn^{-\alpha})^{2/(1+2\alpha)}$ where $\kappa$ is a numerical constant (see Bretagnolle and Huber 1979). Apart from the constant $\kappa$ this is the minimax risk over $\mathbb{S}_{\alpha,H}$. If $\alpha$ and $H$ are unknown $m$ has to be chosen from the data and the problem is to determine whether it is possible or not to achieve the same risk (up to some numerical constant) whatever $\alpha$ and $H$. Adaptation means that, in a more or less strong sense, one can do as well not knowing to which $\mathbb{S}_{\alpha,H}$ $s$ belongs that knowing it.

One can consider (at least) three different approaches to adaptation in the minimax sense. Let us assume that we are given a family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ of parameter spaces and a sequence $(\tilde{s}_n)_{n\geq 1}$ of estimators independent of $\theta$. One considers the ratios

$$\frac{R_n(\tilde{s}_n, \mathbb{S}_\theta)}{R_n(\mathbb{S}_\theta)} = C_n(\theta) \quad .$$

- Historically, the first approach to adaptation in the minimax sense was introduced by Efroimovich and Pinsker (1984). This is an asymptotic point of view which amounts to show that one can find a suitable sequence $((\tilde{s}_n)_{n\geq 1}$ such that $\limsup_{n\to+\infty} C_n(\theta) = 1$ for any $\theta \in \Theta$. They proved such a result for the white noise setting when the family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ is the class of ellipsoids. In Efroimovich (1985) and Efroimovich and Pinsker (1986) they extended their results to other settings. Further results in this direction are to be found in Golubev (1992). In this case we shall speak of *exact asymptotic adaptation*.
- Another point of view introduced by Lepskii (1991) and Donoho and Johnstone (1995) consists in showing that $\limsup_{n\to+\infty} C_n(\theta) = C(\theta)$ for any $\theta$. In this case we shall speak of *asymptotic adaptation*. One can weaken this definition to *asymptotic adaptation up to* $\mathscr{L}(n)$ when $\limsup_{n\to+\infty} C_n(\theta)/\mathscr{L}(n) = C(\theta)$ and $\mathscr{L}(n)$ is a slowly varying function. In those two directions one can cite the papers by Lepskii (1992), Golubev and Nussbaum (1992), Goldenshluger and Nemirovskii (1997), Lepskii, Mammen and Spokoiny (1997), Donoho, Johnstone, Kerkyacharian and Picard (1995) and the references therein, Lepskii and Spokoiny (1995), Juditsky (1997), among many recent results.
- A third approach is to show that $C_n(\theta) \leq \mathscr{L}(n)C(\theta)$ where $\mathscr{L}(n)$ is a slowly varying function. When $\mathscr{L}(n) = 1$ we shall speak of *nonasymptotic adaptation* and this point of view has been extensively developed in

the preceding sections with some examples where $C(\theta)$ does not depend on $\theta$ and in Birgé and Massart (1997) as well. When $\mathscr{L}(n)$ goes to infinity with $n$ we shall speak of *adaptation up to* $\mathscr{L}(n)$ and various results in this direction are to be found in Donoho, Johnstone, Kerkyacharian and Picard (1996).

One should first notice that nonasymptotic adaptation implies asymptotic adaptation while the converse (even in the case of exact asymptotic adaptation) does not hold since nothing warrants that the convergence is uniform with respect to $\theta$. The presence of the function $\mathscr{L}(n)$ (typically some power of $\log n$) is sometimes necessary (see for instance Lepskii 1992 for pointwise estimation) and is sometimes connected to the choice of the estimation procedure.

The difficulty of finding adaptive estimators (in any sense) is partly connected to the difficulty of finding minimax estimators (up to constants) on each parameter space $\mathbb{S}_\theta$. It is now well-known that if we choose, for instance, the loss function $\ell(s,t)$ as $\|s-t\|^2$ where $\|\cdot\|$ denotes the norm in $\mathbb{L}_2([0,1],dx)$, it is more difficult to estimate a function in a ball of the Sobolev space $W_p^\alpha$ with $p<2$ than in $W_2^\alpha$. In $W_2^\alpha$, linear estimators based on any optimal linear approximation procedure will do the job while no linear estimator can achieve the optimal rate of convergence in the spaces $W_p^\alpha$ for $p<2$, as shown in Donoho and Johnstone (1994c). Nemirovskii (1985) was the first to provide estimation procedures achieving the minimax risk (up to constants) over the balls of those spaces. It is a merit of wavelets to produce simple estimation methods based on thresholding or shrinkage of the coefficients that also achieve these optimal rates of convergence. The introduction of wavelets to construct optimal estimators in this context is due to Johnstone, Kerkyacharian and Picard (1992) and Donoho and Johnstone (1998). The functions in $W_p^\alpha$ for $p<2$ have a nonhomogeneous smoothness (relatively to the $\mathbb{L}_2$-norm) and this is the reason why Donoho and Johnstone introduced the term of *spatially adaptive* for the optimal estimators in those spaces. Another attractive feature of wavelets comes from the fact that mild modifications of the preceding estimators lead to adaptive procedures in various senses as mentioned above.

Adaptation in the minimax sense requires the introduction of some "a priori" class of compact sets $\mathbb{S}_\theta$ (in most cases balls with respect to a family of semi-norms defining some smoothness restriction for $s$). This presentation clearly leads to various questions:

- What happens if the true $s$ does not belong to $\mathbb{S}=\cup_{\theta\in\Theta}\mathbb{S}_\theta$?
- How should one choose the family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ if only $\mathbb{S}$ is given (think of $\mathbb{S}=\mathscr{C}([0,1])$)? Clearly there is not only one choice.
- What type of property is required on the family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ in order to get adaptation in any of the above senses? Not all families will do as

shown by the following example: the observations $X_1, \ldots, X_n$ are independent identically distributed with unknown density $s$ belonging to $\mathbb{S} = \mathscr{C}([0, 1])$ and $\mathbb{S}_\theta$ is any regular (in the usual sense) parametric submodel with parameter space $[0, 1]$ and Fisher information bounded away from zero. If the loss function is the square of the Hellinger distance between densities, the minimax risk over $\mathbb{S}_\theta$ will be of order $1/n$. The set of $\mathbb{S}_\theta$'s, which is the set of all such parametric submodels, will cover $\mathbb{S}$ and there is no hope to get an adaptive estimator in such a situation since the family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ is too large.

As we already mentioned at the beginning of Section 5, when one wants to construct an adaptive estimator in the minimax sense relative to some family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$, one is lead to introduce some collection of estimators depending on a tuning parameter $m \in \mathscr{M}$ and then perform a data-driven choice of $m$. On the other hand, the real object of interest is the true function $s$ itself and the introduction of the family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ can be viewed as rather artificial. This motivates the second approach to adaptation.

### *5.2. Adaptation with respect to the target function and model selection*

The idea now is to forget about the introduction of a reference family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ and rather introduce a collection of estimators $\{\hat{s}_m\}_{m\in\mathscr{M}}$. The choice of such a family is in some sense arbitrary and plays the same role as the choice of the prior in Bayesian theory. In any case it is not more artificial than the choice of the family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$. The gain is that one focuses only on the target function and tries to select the tuning parameter $m$ in order to minimize the risk at $s$. More precisely, if the data driven selection procedure is given by $\hat{m}$, one wants to optimize in some sense the ratio

$$\rho_n(s) = \frac{R_n(\hat{s}_{\hat{m}}, s)}{\inf_{m\in\mathscr{M}} R_n(\hat{s}_m, s)} \quad .$$

Once again, one can consider various types of controls for $\rho_n(s)$.

- One can try to get $\lim_{n\to+\infty} \rho_n(s) = 1$. Such a program can be carried out for the quadratic risk by using methods related to cross-validation or Mallows' $C_p$. Among many such results let us cite for kernel estimation Hall (1983) and Stone (1984), Hall (1987) for projection estimation and for regression on fixed design Li (1987) and Polyak and Tsybakov (1990).
- One can alternatively look for results of the form $\limsup_{n\to+\infty} \rho_n(s) < +\infty$. Such a point of view appears in the numerous works on maximum penalized likelihood or penalized least squares estimators as described in Silverman (1982) or Wahba (1990).
- Finally, one can try to define $\hat{m}$ in such a way that $\sup_n[\rho_n(s)/\mathscr{L}(n)] = \rho(s) < +\infty$ where $\mathscr{L}(n)$ is a slowly varying function. This is the type of

results that one can find in Donoho and Johnstone (1994a) for estimators based on thresholding of empirical wavelet coefficients in the Gaussian regression with fixed design. In this case the function $\mathscr{L}(n)$ is a power of $\log n$.

Apart from kernel cross-validation the methods of adaptation to the target function studied by these authors can be viewed as occurences of model selection by minimization of a penalized empirical criterion. One should nevertheless distinguish between infinite dimensional and finite dimensional model selection which involve penalty terms of very different natures.

The approach to penalization which is developed at length in the monograph by Wahba (1990) represents another way of penalizing which is rather connected with the calculus of variation on infinite dimensional spaces and can be seen as a penalized version of the infinite dimensional sieve method of Chow and Grenander (1985). The interested reader can consult the impressive list of references in Wahba (1990). Another illustration is to be found in Van de Geer (1990) who introduces empirical process techniques for studying those estimators. Typically one considers a function $s$ belonging to $\mathbb{S}_{\alpha,H}$ where $H$ is unknown (but $\alpha$ is given!) and derives the optimal rate of convergence $n^{-2\alpha/(2\alpha+1)}$ (with respect to the quadratic loss function). The estimator is obtained by minimizing some empirical contrast function (least squares for fixed design regression as in Wahba, 1990 or maximum likelihood as in Silverman, 1982) with respect to $t$ belonging to the whole Sobolev space $W_2^\alpha$ with a penalty term proportional to $\|t^{(\alpha)}\|^2$. The penalty is used there to avoid a compactness assumption but this method requires $\alpha$ to be known. From this point of view it cannot achieve one of the main issues of this paper which is to estimate a function of unknown smoothness.

The other methods are all based on model selection over a family of finite dimensional spaces via the minimization of an empirical criterion involving a penalty term which is roughly proportional to the dimension. This is clearly the case in Li (1987) and Polyak and Tsybakov (1990 and 1993) who study penalized least squares estimators closely related to Mallows' $C_p$ as described in the introduction. This is also true for projection cross-validation (see Hall, 1987) which can be viewed as a (randomly) penalized projection estimation method as shown in Birgé and Massart (1997).

Even if this is not apparent at first sight, the method of hard thresholding used in Donoho and Johnstone (1994a) can actually be viewed as a penalized least squares method (this has been shown at least in the context of density estimation by Birgé and Massart, 1997).

We would finally like to emphasize an important fact concerning the connection between model selection methods and adaptation in the minimax sense. While all the methods we just described can potentially lead to adaptive estimators in the minimax sense on some collections of smooth-

ness classes, such results cannot directly be derived from the asymptotic risk bounds given by the previous authors because of the lack of uniformity with respect to $s$. On the contrary, the nonasymptotic bounds such as those obtained by Polyak and Tsybakov (1993) or Donoho and Johnstone (1994a) naturally lead to adaptation in the minimax sense (up to a slowly varying factor in the last case) by using the device described in the introduction.

### *5.3. Comparison with other adaptive methods*

There is some difficulty to compare directly our results with the existing literature since many of the results which are connected to ours are developed in the context of the white noise setting, or regression with fixed design or more general regression settings. This is in particular the case of Li (1987), Lepskii (1991) and Donoho and Johnstone (1994a) that we are analyzing below. On the other hand we do not study here the white noise setting at all and we study only very partially the regression on fixed design. On the contrary we have developed many results for density estimation using penalized projection (see also Birgé and Massart, 1997) or maximum likelihood estimators. Therefore, in our comparisons with other works, we are putting the emphasis on the methods and comment on the types of results which are obtained, taking for granted that the reader is aware of the analogy between those different settings. Anyway, the reader can look at Section 2 in order to find an illustration of our way of thinking of this analogy.

The main issue of our approach is to define a proper penalty term for general collections of models and various empirical contrast functions and to derive an upper bound for the resulting minimum penalized empirical contrast estimator $\hat{s}$. As we have seen this penalty can typically be written as $\kappa L_m D_m/n$ with

$$\sum_{m\in\mathcal{M}_n} \exp(-L_m D_m) \leq \Sigma \tag{5.1}$$

for some $\Sigma < +\infty$ and independent of $n$ and the resulting risk is bounded by

$$\mathbb{E}_s\left[\|s-\hat{s}\|^2\right] \leq \kappa' \inf_{m\in\mathcal{M}_n}\left\{d^2(s,S_m) + L_m D_m/n\right\} + \Sigma C(s)/n \ . \tag{5.2}$$

#### 5.3.1. Adaptation to the target function

This type of penalty covers two particular cases that we just mentioned above and which are both related to least squares estimation. If the number of models having a given dimension $D$ is subexponential, which typically

occurs when the models are nested, one can choose $L_m = L$ and get a penalty which only differs from the penalty in Mallows' $C_p$ by a constant factor. From that point of view, our results are non-asymptotic analogues of those by Li (1987). One can even, in this case, using our approach, go further in the analogy by identifying exactly the constants involved in the penalty term and get nonasymptotic risk bounds for Mallows' $C_p$ as in Baraud (1997).

If one considers a "rich" family of models in the sense that there are many models of the same dimension, one is lead, in order to satisfy the constraint (5.1) which is essential to prove that our method works, to take larger values for the $L_m$'s (at least for some of them). A typical situation of that type is the "variable selection" example that we give in Section 2 where we take $L_m = L \log n$. We also show there that the resulting penalized estimator is a threshold estimator as introduced in Donoho and Johnstone (1994a). Our risk bound, which includes an extra $\log n$ factor (as compared to the minimal risk among the family of models), due to the choice of $L_m$, is consistent with that of Donoho and Johnstone. More generally, the interest of considering a "rich" family of models is to get better approximation properties.

### 5.3.2. Adaptation in the minimax sense

Indeed, if one wants to approximate an arbitrary function $s$ in $W_p^\alpha$ for $p < 2$ one cannot content oneself with one single linear model per dimension (see for instance Pinkus, 1985). One is therefore led to introduce many linear models of the same dimension. This strategy is especially relevant when the sieves are linear spans of finite subsets of a localized basis (piecewise polynomials or wavelets, for instance). Such a point of view has been developed by Donoho, Johnstone, Kerkyacharian and Picard and an interesting review is to be found in their 1995 paper.

In the recent literature devoted to adaptive estimation in the minimax sense one can distinguish between two main methods: thresholding from empirical coefficients and what is now known as "Lepskii's method". They both rely on the data-driven selection of some tuning parameter for a family of estimators and have been developed for various loss functions. The idea introduced by Efroimovich and Pinsker (1984) and developed by Donoho, Johnstone, Kerkyacharian and Picard in various works is to use various strategies of thresholding in order to select a subset of some given basis. As quoted in Birgé and Massart (1997), the adaptive procedures based on "hard thresholding" can be viewed as penalized projection estimators and another illustration of this idea is given in Section 2. This interpretation can be of some help, for instance it allowed Birgé and Massart (1997) to avoid the dependence of the estimator with respect to unpleasant quantities

like the radius of the Sobolev balls which appear in Donoho, Johnstone, Kerkyacharian and Picard (1996).

"Lepskii's method" was first introduced in Lepskii's seminal works of 1991 and 1992 and applied to adaptive estimation for Sobolev or Hölderian balls in the white noise setting and Gaussian regression. It is rather difficult to describe it in a few lines. Let us only recall here that it is based on two main assumptions. Given some (almost) arbitrary loss function, one starts with a family of parameter sets $\{\mathbb{S}_\theta\}_{\theta\in\mathbb{R}^+}$ for which one knows that the corresponding family of minimax rates of convergence is totally ordered. Moreover one postulates the a priori existence of a family of estimators with adequate convergence properties (involving not only the rates of convergence but also specific shapes for the probability tails of the deviations) which have to be constructed in each particular case. The method uses an ordering of the rates which, at first glance, implies adaptation with respect to a real parameter rather than a multidimensional one. It should be mentioned however that subsequent elaborations of this method are to be found in Lepskii and Spokoiny (1995) and Lepskii, Mammen and Spokoiny (1997) where they relax this monotonicity restriction by using a local version of Lepskii's method for kernel estimators.

### 5.3.3. What's new here?

Let us first mention that one serious drawback of our approach, as compared to some of the other adaptation results described above, is that it forces us to use a particular loss function which is naturally connected to the empirical contrast function we choose. More precisely we mean that we can derive the risk of the estimator for powers of some particular distance (which is typically some $\mathbb{L}_2$-distance) and we do not know how to bound the risk for other loss functions.

We now want to emphasize the novelties brought by our approach from two points of view: the risk bounds and the estimation procedures. First, let us recall that all our risk bounds are systematically "nonasymptotic". As we already mentioned, the typical risk bound takes the form (5.2) and expresses the performance of our estimator at the target function. From this point of view, our results are quite different from results on cross-validation like those of Hall (1987) and Li (1987). In particular we do not require that the true function does not belong to any of the models and therefore our result is also valid for model selection in a parametric setting. There is actually no difference, in our approach, between the parametric and the nonparametric points of view.

Another advantage of nonasymptotic bounds like (5.2) is that they naturally lead to adaptation in the minimax sense on various families of compact

sets and in particular classical smoothness classes, via an adequate choice of the family of sieves. Moreover, given such a family, the resulting upper bound for the maximal risk over any set in the family is in many cases comparable (up to a universal constant) with the minimax risk over this set. Several such examples are given in Section 4.1 (see in particular Proposition 3 and inequalities 4.20 and 4.21). This makes a substantial difference with most of the typical results in this direction. Many results similar to some of ours are well-known from an asymptotic point of view for instance those concerning adaptation with respect to Hölder classes (Lepskii, 1991) or with respect to ellipsoids (Efroimovitch and Pinsker, 1984) . But as far as we know, the results which are stated here are new, as they are stated, since we systematically provide inequalities for a given number $n$ of observations which not only describe the rate of convergence but also make explicit the dependence of the constants with respect to the smoothness parameters or some feature of the unknown function to be estimated, up to universal numerical constants.

Let us now turn to the advantages of our method of estimation using model selection. We see two advantages of this method as compared to Lepskii's: first, our approach does not impose any ordering on the rates of convergence and therefore can handle adaptation in multivariate estimation problems where the smoothness is not homogeneous with respect to directions. Secondly our method does not rely on the existence of preliminary estimators but automatically provides the estimators and the adaptation procedure simultaneously.

May be that the main quality of our method is its considerable flexibility since we have the choice of both the family of models and the weights (provided that they satisfy 5.1). In our examples, we mainly discussed the situation of constants weights, either equal to $L$ or to $L \log n$. This strategy of penalization proportional to the dimension includes the "hard thresholding methods" as illustrated in Section 2 but more sophisticated choices of the weights are interesting. To illustrate this point of view let us assume that we have at our disposal a "very large" family of models $\{S_m\}_{m \in \mathcal{M}_n}$ in the sense that $\sum_m \exp(-D_m) = +\infty$ but the number of $m$'s with $D_m \leq j$ is finite for any integer $j$. It is clear that there exist many choices of weights $L_m$ satisfying the condition (5.1). In particular, it is always possible, in a list of models with a given dimension $D_m = D$, to take $L_m = 1$ for a bounded number of them. If we denote by $m_n(s)$ the best model for estimating $s$ with $n$ observations, that is the model leading to the minimal risk at $s$, it follows from our evaluations of the risk of the minimum penalized empirical contrast estimators that the smaller $L_{m_n(s)}$, the better this risk which means that $L_{m_n(s)}$ should ideally be 1. Since $m_n(s)$ is unknown, for each given dimension $D$ we tend to put small values of $L_m$ on the models of dimension

$D_m = D$ which we believe to be more accurate and large values of $L_m$ for those that we consider as unlikely. This is very similar to the choice of a prior distribution on the family of models. Actually both the choice of the family $\{S_m\}_{m\in\mathcal{M}_n}$ and of the family of weights $\{L_m\}_{m\in\mathcal{M}_n}$, reflect our "a priori" information about $s$ or our "belief" about the true state of nature, to put it in a Bayesian language. This idea has been illustrated in Section 3.3.1 where we have introduced a mixture of histograms based on regular or irregular partitions, the first ones being suitable for estimation of Hölderian densities and the second ones for densities with bounded $\alpha$-variation. More generally, if we have at hand several lists of models $\mathcal{M}_{n,j}$ for $j \in J$, one could just mix all the models in a larger list by a suitable modification of the weights.

## 6. A general theorem in an abstract framework

The purpose of this section is to establish risk bounds for minimum penalized contrast estimators, that is an analogue of Theorem 1, in a general setting. We then show in the next section that this theorem implies all the results that we have stated in Section 3 for each particular empirical contrast function. This research of generality leads us to introduce some assumptions which will probably appear rather obscure and very technical at the first reading. As quoted in the conclusion of Section 2 the main task here is to control the fluctuations of some empirical process connected to $\gamma$. A natural candidate is the centered empirical process $\nu_n[\gamma(\cdot, s_m) - \gamma(\cdot, t)]$. Unfortunately the unboundedness of the function $\gamma$ defining the empirical contrast $\gamma_n$ leads to difficulties for the control of this process and to overcome these difficulties we introduce a suitable modification $\tilde{\gamma}_m$ of $\gamma$ (which might be equal to $\gamma$ itself) on each model $S_m$. In most situations, this is a minor modification which leaves the centered empirical process invariant, but it can be more complicated as required for the treatment of maximum likelihood estimation.

The main issue is then to control a weighted version of the empirical process $\nu_n[\tilde{\gamma}_m(\cdot, s_m) - \tilde{\gamma}_{m'}(\cdot, t)]$ for $t \in S_{m'}$ by exponential bounds similar to (2.7). Unfortunately, there is not a single and canonical way to do that and the multiplicity of our assumptions reflects the many cases we want to handle and the variety of techniques which have been developed in the recent years. These assumptions describe in different ways the "massiveness" of the models which directly influences the size of the fluctuations of the empirical process.

### *6.1. Exponential bounds for the fluctuations of empirical processes*

In what follows, as we already mentioned at the beginning of Section 3.1, the unknown parameter $s$ is supposed to belong to some subset $\mathcal{S}$ of the

set $\mathcal{T}$ on which the function $\gamma(z,\cdot)$ is defined. We recall that $\mathcal{S}$ and every quantity which is indexed by some element $m$ or $m'$ in $\mathcal{M}_n$ can depend on $n$ although this is not emphasized by our notations. On the contrary, all the constants involved in the following assumptions (unless otherwise stated) are independent of $n$ and of $s \in \mathcal{S}$ but may depend on our choice of $\mathcal{S}$. Those constants are supposed to be known to the statistician and can therefore be used in the construction of estimators of $s$. The statistical framework that we use is the one described in Section 3.1.

For each $m \in \mathcal{M}_n$ we associate some number $D_m \geq 1$ (referred to as the *dimension* of $S_m$) and we introduce a function $\tilde{\gamma}_m$, defined on $\mathcal{X} \times S_m$ and measurable with respect to the first variable. In some situations we shall take $\tilde{\gamma}_m = \gamma$, otherwise the reader should think of $\tilde{\gamma}_m$ as a suitable modification of $\gamma$ with improved boundedness properties. It will be the role of Assumption **C** below to specify what kinds of modifications of $\gamma$ are allowed. The following assumptions are related to such a family of functions.

**Lip (Lipschitz)** *For any $s \in \mathcal{S}$, the observed random variables, $Z_1, \ldots, Z_n$, under the distribution $\mathbb{P}_s$, can be written as $Z_i = f(s, X_i, W_i)$ for some known function $f$. The random variables $X_1, \ldots, X_n$ take their values in $\mathcal{X}$, $W_1, \ldots, W_n$ take their values in $\mathcal{W}$, they are all independent and the distributions of the $W_i$'s are free with respect to $s$. Moreover there exists a nonnegative measurable function $M(\cdot)$ defined on $\mathcal{W}$ and for each $(m, m') \in \mathcal{M}_n \times \mathcal{M}_n$ and each pair $(u, v) \in S_m \times S_{m'}$ a nonnegative measurable function $\Delta_{m,m'}(\cdot, u, v)$ defined on $\mathcal{X}$ such that*

$$|\tilde{\gamma}_m(z,u) - \tilde{\gamma}_{m'}(z,v)| \leq M(w)\Delta_{m,m'}(x,u,v) \quad \text{for } z = (x,w) \ .$$

*Furthermore one can find positive constants $A$, $B$, $E$ such that for all $j \geq 2$, any $(m, m')$ in $\mathcal{M}_n \times \mathcal{M}_n$, $u$ in $S_m$, $v$ in $S_{m'}$, either*

**i)**

$$\|M(W_i)\|_\infty \leq A \quad \text{for all } \ i = 1, \ldots, n \tag{6.1}$$

*and*

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_s[\Delta^j_{m,m'}(X_i,u,v)] \leq \frac{j!}{2} B^{j-2}\left[d^2(u,v) + E\,\frac{D_m \vee D_{m'}}{n}\mathbb{1}_{\{m\neq m'\}}\right] \tag{6.2}$$

*or*

**ii)**

$$\mathbb{E}[M^j(W_i)] \leq \frac{j!}{2}A^j \quad \text{for all } \ i = 1, \ldots, n \tag{6.3}$$

*and*

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_s[\Delta^2_{m,m'}(X_i,u,v)] \le d^2(u,v), \quad \|\Delta_{m,m'}\|_\infty \le B \quad \textit{and} \quad E=0 \tag{6.4}$$

*holds.*

*Remarks* :

1. As we already noticed in Birgé and Massart (1998), (6.2) can be deduced from $\mathbb{L}_2$- and $\mathbb{L}_\infty$-controls on $\Delta$.
2. If $m = m'$ (6.2) is merely (5.3) of Birgé and Massart (1998).

The Assumption **M (Metric)** takes one of the two following forms corresponding to controls of covering numbers either related to $\mathbb{L}_2$- and $\mathbb{L}_\infty$-norms or to $\mathbb{L}_1$ with bracketing.

**M (Metric)** *For each $m \in \mathscr{M}_n$ one can find constants $B'_m \ge 1$ such that for each $\delta > 0$ and each ball $\mathscr{B} \subset S_m$ with radius $\sigma \ge 5\delta \vee (D_m/n)^{1/2}$ (with respect to the $\mathbb{L}_2$-distance), there exists a finite set $T = T(m,\delta,\mathscr{B}) \subset \mathscr{B}$ with*

$$|T| \le (B'_m\sigma/\delta)^{D_m} \tag{6.5}$$

*and a mapping $\pi = \pi(m,\delta,\mathscr{B})$ from $\mathscr{B}$ to $T$ such that one of the two following sets of properties is satisfied:*

- $\mathbf{M}_{2,\infty}$ **($\mathbb{L}_2/\mathbb{L}_\infty$ metric):** *Assumption* **Lip (i)** *or* **(ii)** *holds, $d(u,\pi u) \le \delta$ for all $u$ in $\mathscr{B}$ and there exists some $r'_m > 0$ independent of $\delta$ and $\mathscr{B}$ such that*

$$\sup_{u\in\pi^{-1}(t)} \|\Delta_{m,m}(\cdot,u,t)\|_\infty \le r'_m\delta \quad \textit{for all } t \in T \ . \tag{6.6}$$

- $\mathbf{M}_{1,[\,]}$ **($\mathbb{L}_1$ metric with bracketing):** *Assumption* **Lip (ii)** *holds and for all $t \in T$ one can find a measurable function $V_{m,t}$ such that for all $t \in T$, all $x \in \mathscr{X}$ and all $s \in \mathscr{S}$*

$$\sup_{u\in\pi^{-1}(t)} \Delta_{m,m}(x,u,t) \le V_{m,t}(x) \quad \textit{and} \quad \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_s[V_{m,t}(X_i)] \le \delta^2 \ . \tag{6.7}$$

*Remark*: One should notice here that **M** combines an assumption concerning the metric structure of each sieve viewed separately and Assumption **Lip**

which is also supposed to handle the correspondances between different sieves.

In the case of projection estimation one can substantially simplify these assumptions provided that the family of models satisfy the following linearity property:

**L (Linear)** *We assume that each model $S_m$ for $m \in \mathcal{M}_n$ is a subset of some $D_m$-dimensional linear subspace $\bar{S}_m$ of $\mathbb{L}_2(\mu) \cap \mathbb{L}_\infty(\mu)$.*

We are now in a position to state an exponential inequality for a weighted empirical process related to $\tilde{\gamma}_m$ which, except under Assumption $\mathbf{M}_{1,[\,]}$, has been proved in Theorem 5 and Proposition 3 of Birgé and Massart (1998) .

**Proposition 7** *Let the family of subsets $\{S_m\}_{m\in\mathcal{M}_n}$ of $\mathbb{L}_2(\mu)$ be given and for each $m$ a function $\tilde{\gamma}_m$ be defined on $\mathcal{X} \times S_m$ which is measurable with respect to the first variable. We assume that either*

- **M** *holds*
  *or*
- *we observe $n$ independent identically distributed random variables $Z_1, \ldots, Z_n$ with density $s \in \mathbb{L}_2(\mu)$, $\tilde{\gamma}_m(z,t) = -2t(z)$ and* **L** *holds.*

  *The following exponential inequality is then satisfied for any $m \in \mathcal{M}_n$, any $t \in S_m$ and any $\tau > 0$*

$$\mathbb{P}_s\left[\sup_{u\in S_m} \frac{\nu_n[\tilde{\gamma}_m(\cdot,t) - \tilde{\gamma}_m(\cdot,u)]}{d^2(t,u) \vee x^2} > \tau\right] \leq 3.1 \exp\left[-nh_m(x)\right] \quad \text{for all } x \geq \sigma_m \tag{6.8}$$

  *where*

- *under* **M**:

$$\sigma_m^2 = \left[\zeta^2 \mathcal{L}'_m \vee 1\right] \frac{D_m}{n} \quad \text{and} \quad h_m(x) = \left(\frac{x}{\zeta}\right)^2$$

  *with*

$$\zeta^2 = \frac{20}{9\tau^2}\left[32A^2 + 6AB\tau\right]$$

  *and*

$$\mathcal{L}'_m = 2.5 \log\left[14B'_m\left(1 + r'_m(D_m/n)^{1/2}\right)\right]$$

*whenever Assumption* $\mathbf{M}_{2,\infty}$ *holds or*

$$\mathscr{L}'_m = 2\log\left[B'_m\left(\sqrt{B} \vee 4[A/(3\tau)]^{1/2} \vee 5\right)\right]$$

*whenever* $\mathbf{M}_{1,[\,]}$ *holds*;

- *under* **L**:

$$\sigma_m = 6\frac{\Phi_m \wedge \|s\|_\infty^{1/2}}{\tau}\left(\frac{D_m}{n}\right)^{1/2}$$

*and*

$$h_m(x) = \kappa'\left[\frac{\tau x}{\Phi_m\sqrt{D_m}} \wedge \frac{\tau^2 x^2}{(\Phi_m\sqrt{D_m}\|s\|) \wedge \|s\|_\infty}\right]$$

*where* $\Phi_m$ *is given by* (3.2) *and* $\kappa'$ *is a positive constant.*

*Remarks*:

- It should be noticed that, since the bound (6.8) involves a single function $\tilde{\gamma}_m$, we do not use in Proposition 7 the full power of Assumption **Lip** (which deals with all pairs $(m, m')$).
- In the statement of Proposition 7, the notation $\mathbb{P}_s$ in bound (6.8) can be abusive since our sets of assumptions do not always warrant (we especially think of $\mathbf{M}_{1,[\,]}$) that the supremum involved in (6.8) is measurable. If some measurability problems occur, $\mathbb{P}_s$ should be understood as an outer probability which does not destroy anything in the proof of Proposition 7 since it only uses the subadditivity properties of $\mathbb{P}_s$.
- It is noticeable that, under Assumption $\mathbf{M}_{1,[\,]}$, the proof of (6.8) does not involve any chaining argument (while such an argument is necessary for the proof under Assumption $\mathbf{M}_{2,\infty}$). Such a device has been used by Pollard (1985) for providing simple proofs of uniform central limit theorems following an original idea by Huber (1967).

*Proof*: Let us begin with Assumption $\mathbf{M}_{1,[\,]}$. We first want to prove that if $\mathscr{B}$ denotes the ball of radius $\sigma$ and center $t$, whatever $t \in S_m$

$$\mathbb{P}_s\left[\sup_{u\in\mathscr{B}} \nu_n[\tilde{\gamma}_m(\cdot,t) - \tilde{\gamma}_m(\cdot,u)] > \tau\sigma^2\right] \le 2\exp\left[-\frac{3n\sigma^2}{10\rho^2(\tau)}\right] \qquad (6.9)$$

provided that $n\sigma^2 \ge D_m[\mathscr{L}(\tau)\rho^2(\tau) \vee 1]$ where $\rho(\tau)$ and $\mathscr{L}(\tau)$ are defined by

$$\rho^2(\tau) = \frac{16A^2}{\tau^2} + \frac{4AB}{\tau}$$

and

$$\mathscr{L}(\tau) = 5\log(B'_m\theta) \quad \text{with } \theta = \sqrt{B} \vee 2(A/\tau)^{1/2} \vee 5 \ .$$

Let us set $\rho = \rho(\tau)$, $\mathscr{L} = \mathscr{L}(\tau)$, $\delta = \sigma/\theta$ and $f_u = \tilde{\gamma}_m(\cdot, t) - \tilde{\gamma}_m(\cdot, u)$. Since $\sigma^2 \geq D_m/n$, by (6.5) and $\mathbf{M}_{1,[\,]}$ we can assume the existence of $T$ with cardinality $e^H$, $H \leq D_m \log(B'_m\theta)$ and for each $v \in T$ there exists a random variable $V_{m,v}$ with

$$\sup_{u\in\pi^{-1}(v)} \Delta_{m,m}(x, u, v) \leq V_{m,v}(x) \quad \text{and} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_s[V_{m,v}(X_i)] \leq \delta^2 \ . \tag{6.10}$$

Since by (6.4) $\|\Delta_{m,m}\|_\infty \leq B$ we can assume without loss of generality that $\|V_{m,v}\|_\infty \leq B$. If $v = \pi(u)$, $|f_u - f_v| \leq MV_{m,v}$ and we get

$$\begin{aligned} \nu_n(f_u) &\leq \frac{2}{n}\sum_{i=1}^{n} \mathbb{E}_s[M(W_i)V_{m,v}(X_i)] + \nu_n(f_v) + \nu_n(MV_{m,v}) \\ &\leq 2A\delta^2 + \nu_n(f_v) + \nu_n(MV_{m,v}) \end{aligned} \tag{6.11}$$

by the independence of $W_i$ and $X_i$, (6.10), (6.3) and Cauchy-Schwarz inequality.

**Control of $\nu_n(f_v)$:** From the independence of $W_i$ and $X_i$, (6.3) and (6.4) with $d^2(t, v) \leq \sigma^2$ we get

$$\begin{aligned} \mathbb{E}_s\big[|\tilde{\gamma}_m(Z_i, t) - \tilde{\gamma}_m(Z_i, v)|^j\big] &\leq \mathbb{E}_s\big[M^j(W_i)\big]\mathbb{E}_s\big[\Delta^j_{m,m}(X_i, t, v)\big] \\ &\leq \frac{j!}{2}A^j B^{j-2}\mathbb{E}_s\big[\Delta^2_{m,m}(X_i, t, v)\big] \end{aligned}$$

and

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_s\left[\left(A^{-1}|\tilde{\gamma}_m(Z_i, t) - \tilde{\gamma}_m(Z_i, v)|\right)^j\right] \leq \frac{j!}{2}\sigma^2 B^{j-2} \ .$$

Therefore Bernstein's inequality (see Birgé and Massart 1998, Lemma 8) implies that, if $\eta = \sigma\sqrt{2x} + Bx$

$$\mathbb{P}_s[\nu_n(f_v) > A\eta] \leq \exp(-nx) \ . \tag{6.12}$$

**Control of $\nu_n(MV_{m,v})$:** We use again the independence between $W_i$ and $X_i$ and (6.10) to get since $\|V_{m,v}\|_\infty \leq B$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_s\left[A^{-j}M^j(W_i)V^j_{m,v}(X_i)\right] \leq \frac{j!}{2}B\delta^2 B^{j-2} \ .$$

Therefore, since $B\delta^2 \leq \sigma^2$ Bernstein's inequality implies that

$$\mathbb{P}_s[\nu_n(MV_{m,v}) > A\eta] \leq \exp(-nx) \ . \tag{6.13}$$

It follows from (6.11), (6.12) and (6.13) that

$$\mathbb{P}_s\left[\sup_{u\in\mathscr{B}} \nu_n(f_u) > 2A(\eta+\delta^2)\right] \leq 2\exp(H-nx) \ .$$

Since $n\sigma^2 \geq \rho^2[5D_m \log(B'_m\theta)] \geq 5\rho^2 H$, choosing $x = \sigma^2/(2\rho^2)$ we get $H \leq 2nx/5$ and therefore

$$\mathbb{P}_s\left[\sup_{u\in\mathscr{B}} \nu_n(f_u) > 2A(\eta+\delta^2)\right] \leq 2\exp\left[-\frac{3n\sigma^2}{10\rho^2}\right] \ .$$

In order to get (6.9) it remains to check that $2A(\eta+\delta^2) \leq \tau\sigma^2$. This follows from our choices of $\theta$ and $\rho$ which imply that $\delta^2 \leq \tau\sigma^2/(4A)$ and $\eta \leq \tau\sigma^2/(4A)$.

Setting $\bar{\mathscr{L}} = \mathscr{L}(3\tau/4)$ and observing that $\bar{\mathscr{L}} \geq 8$ which implies that $n\sigma^2/\rho^2 \geq 8$ we can derive from (6.9) that if $n\sigma^2 \geq [\bar{\mathscr{L}}\rho^2(3\tau/4) \vee 1]D_m$,

$$\mathbb{P}_s\left[\sup_{u\in S_m} \frac{\nu_n[\tilde{\gamma}_m(.,t) - \tilde{\gamma}_m(.,u)]}{d^2(t,u)\vee\sigma^2} > \tau\right] \leq 3\exp\left[-\frac{2n\sigma^2}{5\rho^2(3\tau/4)}\right] \tag{6.14}$$

exactly as (5.8) is derived from (7.16) in Birgé and Massart (1998), following the last lines of the proof of their Theorem 5.

We now want to derive (6.8) with the corresponding values of $\sigma_m$, $\mathscr{L}'_m$ and $\zeta$. For Assumption **M** we use either (6.14) (under $\mathbf{M}_{1,[\,]}$) or (5.8) of Theorem 5 of Birgé and Massart (1998), following their notations (under $\mathbf{M}_{2,\infty}$). In both cases one can choose for $\sigma_m$ any number such that

$$\sigma_m^2 \geq \frac{D_m}{n}\left[\left(\frac{5}{2}\rho^2\left(\frac{3\tau}{4}\right)\right)\left(\frac{2}{5}\bar{\mathscr{L}}\right) \vee 1\right] \ .$$

We therefore choose $\zeta^2$ as an upper bound for $(5/2)\rho^2(3\tau/4)$ and $\mathscr{L}'_m$ as an upper bound for $2\bar{\mathscr{L}}/5$. In the case of $\mathbf{M}_{2,\infty}$ we use the value of $\bar{\mathscr{L}}$ given in Theorem 5 of Birgé and Massart (1998) to get

$$\bar{\mathscr{L}} < 6.13 \log\left[14B'_m\left(1 + r'_m(D_m/n)^{1/2}\right)\right] \ .$$

To derive the result in the linear case it is enough to apply Proposition 3 of Birgé and Massart (1998), noticing that $\tilde{\gamma}_m(z,t) - \tilde{\gamma}_m(z,u) = 2(u-t)(z)$. □

### 6.2. A general theorem

From the previous exponential bounds, one can now derive the main theorem of this paper which is at the origin of all our developments and examples apart from those concerning projection estimators on linear sieves to be treated in the next section. In order to connect the fluctuation of some empirical process to the distance between the estimator and the true function $s$, we need an inequality similar to (2.4). This is precisely the role of the assumption which we call *Closing argument* and which is relative to a family of functions $\{\tilde{\gamma}_m\}_{m\in\mathcal{M}_n}$ where $\tilde{\gamma}_m$ is defined on $\mathcal{Z}\times S_m$ and measurable with respect to the first variable.

**C (Closing argument)** *For each $s\in\mathcal{S}$ and $m\in\mathcal{M}_n$ there exists a point $s_m\in S_m$ and a nonnegative random variable $U_m$ (depending on $s$, $s_m$ and $D_m/n$ but not on $t$) with finite second moment such that for all $m, m'\in\mathcal{M}_n$ and all $t\in S_{m'}$ satisfying $\gamma_n(t)+\mathrm{pen}(m')\le\gamma_n(s_m)+\mathrm{pen}(m)$ the following holds with suitable constants $k>0$ and $k_1\ge 0$ independent of $m$ and $n$ :*

$$\nu_n[\tilde{\gamma}_m(\cdot,s_m)-\tilde{\gamma}_{m'}(\cdot,t)]\ge 2k\big(d^2(s,t)-U_m^2-k_1D_{m'}/n\big)$$
$$-\mathrm{pen}(m)+\mathrm{pen}(m')\ . \tag{6.15}$$

Since $\gamma_n(t)=n^{-1}\sum_{i=1}^n\gamma(Z_i,t)$, a natural candidate for being a proper $\tilde{\gamma}_m$ is the function $\gamma$ itself. Indeed $\gamma$ satisfies (6.15) as soon as it satisfies the next assumption

**C'** *There exists two positive constants $k', k''$ such that for all $s\in\mathcal{S}$ and $t\in\mathcal{T}$,*

$$k'd^2(s,t)\le\frac{1}{n}\sum_{i=1}^n\mathbb{E}_s[\gamma(Z_i,t)-\gamma(Z_i,s)]\le k''d^2(s,t)\ .$$

Actually, under Assumption **C'**, not only $\gamma$ but also any function $\tilde{\gamma}$ of the form $\tilde{\gamma}(z,t)=\gamma(z,t)+\psi_1(t)+\psi_2(z)$ satisfies **C**. More precisely

**Lemma 2** *Assume that* **C'** *holds and define $\tilde{\gamma}(z,t)=\gamma(z,t)+\psi_1(t)+\psi_2(z)$, then Assumption* **C** *holds with $\tilde{\gamma}_m=\tilde{\gamma}$, $s_m$ an arbitrary point in $S_m$, $k=k'/2$, $k_1=0$ and $U_m^2=(k''/k')d^2(s,s_m)$.*

*Proof*: From **C'** one derives that for all $s\in\mathcal{S}$, $t\in\mathcal{T}$ and $m\in\mathcal{M}_n$

$$\frac{1}{n}\sum_{i=1}^n\mathbb{E}_s[\gamma(Z_i,t)-\gamma(Z_i,s_m)]\ge k'd^2(s,t)-k''d^2(s,s_m)\ .$$

Then (6.15) follows since

$$\begin{aligned}
&\nu_n[\tilde{\gamma}_m(\cdot, s_m) - \tilde{\gamma}_{m'}(\cdot, t)] \\
&\quad = \gamma_n(s_m) + \psi_1(s_m) - \gamma_n(t) - \psi_1(t) \\
&\qquad -\mathbb{E}_s[\gamma_n(s_m) + \psi_1(s_m) - \gamma_n(t) - \psi_1(t)] \\
&\quad \geq \mathrm{pen}(m') - \mathrm{pen}(m) + \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_s[\gamma(Z_i, t) - \gamma(Z_i, s_m)] \ . \qquad \square
\end{aligned}$$

At this stage one should have in mind that in the sequel, the functions $\tilde{\gamma}_m$ will be required to satisfy Assumption **Lip**. This motivates the introduction of modifications $\tilde{\gamma}_m$ of $\gamma$ even when **C'** is satisfied.

We can now state our main result.

**Theorem 8 (Main Theorem)** *Let $\gamma_n$ be some empirical contrast function according to Definition 1 and assume that we are given a family of models $\{S_m\}_{m\in\mathcal{M}_n}$ and a family of functions $\{\tilde{\gamma}_m\}_{m\in\mathcal{M}_n}$ satisfying* **C** *and* **M** *simultaneously. Moreover, consider a family of weights $\{L_m\}_{m\in\mathcal{M}_n}$ and some constant $\Sigma$ such that*

$$L_m \geq 1 \quad \textit{for all} \ \ m \in \mathcal{M}_n \quad \textit{and} \quad \sum_{m\in\mathcal{M}_n} \exp[-L_m D_m] \leq \Sigma \ . \tag{6.16}$$

*Let $A$, $B$, $E$, $k$, $k_1$ be the constants coming from Assumptions* **C** *and* **M** *and $\kappa$ be a suitable numerical constant. Set $\lambda = (A^2 + ABk)/(\kappa k^2)$, $\tau = k/8$ and $\sigma_m^2 = [\zeta^2 \mathscr{L}'_m \vee 1 \vee E](D_m/n)$ where $\zeta$ and $\mathscr{L}'_m$ are defined in Proposition 7. Consider some penalty function* $\mathrm{pen}(\cdot)$ *defined on the set $\mathcal{M}_n$ satisfying*

$$\mathrm{pen}(m) \geq k\left(\sigma_m^2 \vee \lambda\frac{L_m D_m}{n} + 2k_1\frac{D_m}{n}\right) \ , \tag{6.17}$$

*for all $m \in \mathcal{M}_n$. Then for any $l > 0$ and any $s \in \mathcal{S}$, the risk of the minimum penalized contrast estimator $\hat{s}$ as defined by Definition 2 is bounded by*

$$\begin{aligned}
\mathbb{E}_s\left[d^{2l}(s, \hat{s})\right] \leq C_1(l) \inf_{m\in\mathcal{M}_n}\Bigg\{&\mathbb{E}_s\left[U_m^{2l}\right] + \left[\frac{\mathrm{pen}(m)}{k}\right]^l \\
&+ \left[\frac{d^2(s, s_m)}{2}\right]^l + \Sigma C_1'(l)\left[\frac{\lambda}{n}\right]^l\Bigg\} \ ,
\end{aligned} \tag{6.18}$$

*where $s_m$ comes from Assumption* **C**.

### *6.3. Penalized projection estimators on linear models*

The preceding theorem is an all purpose one, but it can be substantially improved in the particular situation of projection estimation. If we want to apply Theorem 8 to this situation, we have to assume, in order to check Assumption **Lip**, that all the models be included in some ball of known radius of $\mathbb{L}_\infty(\mu)$. This requires, in order that the models have good approximation properties with respect to $s \in \mathscr{S}$ that an upper bound on the $\mathbb{L}_\infty$-norm of the elements of $\mathscr{S}$ be known as in the regression setting. We already mentioned that it is an unpleasant restriction. Fortunately, if the models are linear, this restriction can be relaxed in the case of projection estimation methods. Then we get the following

**Theorem 9** *Let $Z_1, \ldots, Z_n$ be $n$ independent identically distributed random variables with density $s \in \mathbb{L}_2(\mu)$ and $\{S_m\}_{m \in \mathscr{M}_n}$ be a family of models with the linearity property* **L**. *Define the projection empirical contrast $\gamma_n$ as*

$$\gamma_n(t) = \|t\|^2 - \frac{2}{n} \sum_{i=1}^{n} t(Z_i) \quad \text{for all } \ t \in \mathbb{L}_2(\mu)$$

*and consider a family of weights $\{L_m\}_{m \in \mathscr{M}_n}$ and a penalty function* $\mathrm{pen}(\cdot)$ *satisfying*

$$\mathrm{pen}(m) \geq \kappa'(\Phi^2 \vee L_m) D_m / n \quad \text{for all } \ m \in \mathscr{M}_n \ , \tag{6.19}$$

*where $\kappa'$ is a suitable numerical constant and $\Phi$ is defined below. We also assume that one of the three following sets of conditions hold*:

**i)** *$\mathscr{S} \subset \mathbb{L}_\infty$, there exists a constant $\Phi$ and for each $\bar{S}_m$ an orthonormal basis $\{\varphi_\lambda\}_{\lambda \in \Lambda_m}$ of $\bar{S}_m$ such that $\sup_{\lambda \in \Lambda_m} \|\varphi_\lambda\|_\infty \leq \Phi$. Moreover $\sup_{m \in \mathscr{M}_n} D_m \leq n(\Gamma')^{-2}$ for some $\Gamma' > 0$, $L_m \geq 1$ for all $m \in \mathscr{M}_n$ and for any $x > 0$ one can find a constant $\Sigma(x)$ such that*

$$|\mathscr{M}_n| \exp\left[-x\sqrt{n}\right] + \sum_{m \in \mathscr{M}_n} \exp[-x L_m D_m] \leq \Sigma(x) \ ; \tag{6.20}$$

**ii)** *there exists a constant $\Phi$ such that $\|s\|_\infty \leq \Phi^2$ for all $s \in \mathscr{S}$, $L_m = 1$ for all $m \in \mathscr{M}_n$ and one can find positive constants $\Gamma$, $\Gamma'$, $\Gamma_1$ and $\Gamma_2$ such that $\Phi_m \leq \Gamma\sqrt{D_m}$ for all $m \in \mathscr{M}_n$, $\sup_{m \in \mathscr{M}_n} D_m \leq n(\Gamma')^{-2}(\log n)^{-4}$ and*

$$|\{m \in \mathscr{M}_n \mid D_m = j\}| \leq \Gamma_1 j^{\Gamma_2} \quad \text{for } \ j \in \mathbb{N} \setminus \{0\} \ ;$$

**iii)** *there exists a positive constant* $\Phi$ *such that* $\Phi_m \leq \Phi$ *for all* $m$, $L_m = 1$ *for all* $m \in \mathscr{M}_n$ *and one can find some positive constants* $\Gamma'$, $\Gamma_1$ *and* $\Gamma_2$ *such that* $\sup_{m \in \mathscr{M}_n} D_m \leq n(\Gamma')^{-2}$ *and*

$$|\{m \in \mathscr{M}_n \mid D_m = j\}| \leq \Gamma_1 j^{\Gamma_2} \quad \textit{for} \quad j \in \mathbb{N} \setminus \{0\} \ .$$

*Then for any* $l > 0$ *the risk of the minimum penalized projection estimator is bounded by*

$$\mathbb{E}_s \left[ d^{2l}(s, \hat{s}) \right] \leq C_2(l) \inf_{m \in \mathscr{M}_n} \left\{ [\mathrm{pen}(m)]^l + d^{2l}(s, S_m) + C_2' n^{-l} \right\} \ . \quad (6.21)$$

- *Under* **i)** $C_2'$ *depends only on* $l$, $\Gamma'$, $\Phi$, $\|s\|_\infty$ *and* $\Sigma[\kappa/(\Phi \vee \|s\|_\infty)]$ *where* $\kappa$ *denotes some fixed numerical constant.*
- *Under* **ii)** $C_2'$ *only depends on* $l$, $\Phi$, $\Gamma$, $\Gamma'$, $\Gamma_1$ *and* $\Gamma_2$.
- *Under* **iii)**, $C_2'$ *can be written as*

$$C_2' = 1 + \Gamma_1 C_2''(l, \Gamma_2) \left[ \left( \|s\| \vee \Gamma'^{-1} \right) \Phi \right]^{2(l+\Gamma_2+1)} \ . \quad (6.22)$$

### 6.4. *Proof of Theorems 8 and 9*

Since both theorems have a similar structure and both proofs follow essentially the same lines, it is more convenient to give them together. In order to distinguish the different sets of assumptions we shall speak of the metric situation for the assumptions of Theorem 8 and of the linear situation, or more precisely of case **i)**, **ii)** or **iii)** for the assumptions of Theorem 9.

In the linear situation we first make the following remarks:

- the function $\gamma$ which defines the projection empirical contrast is $\gamma(z, t) = \|t\|^2 - 2t(z)$ which immediately implies that it satisfies **C'** with $k' = k'' = 1$.
- In order to choose the value of $\sigma_m$ in Proposition 7 we note that

$$\Phi_m \wedge \|s\|_\infty^{1/2} \leq \Phi \quad (6.23)$$

  which is clear for cases **ii)** and **iii)** and follows from (3.3) in case **i)** since then $\Phi_m$ is bounded by $\Phi$.
- If $S' = S_m + S_{m'}$ has a dimension $D' \leq D_m + D_{m'}$ and an index $\Phi'$ defined by (3.2), it then follows from (3.3) that

$$(\Phi')^2 D' \leq \Phi_m^2 D_m + \Phi_{m'}^2 D_{m'} \ . \quad (6.24)$$

It follows from the first remark that if, for each $m \in \mathscr{M}_n$, we define $\tilde{\gamma}_m(z,t) = -2t(z)$ and choose $s_m$ such that $d(s,s_m) \le 2d(s,S_m)$, then **C** holds with $k = 1/2$, $k_1 = 0$ and $U_m = 2d(s,S_m)$ according to Lemma 2. This means that **C** holds and that one can define $\tau = k/8$ for both theorems. In order to apply Proposition 7 we observe that in the linear situation one can always choose $\sigma_m = 96\Phi(D_m/n)^{1/2}$ since then $\tau = 1/16$ and (6.23) holds.

We first want to show that whatever $m, m' \in \mathscr{M}_n$ and $x \ge \sigma_m \vee \sigma_{m'}$ the following exponential inequality is valid:

$$\mathbb{P}_s\left[\sup_{u\in S_{m'}} \frac{\nu_n[\tilde{\gamma}_m(\cdot,s_m) - \tilde{\gamma}_{m'}(\cdot,u)]}{d^2(s,u)\vee d^2(s_m,s)\vee x^2} > k\right] \le 4.1\exp\left[-nh_{m,m'}(x)\right] \tag{6.25}$$

with $h_{m,m'}$ to be specified below. Given some point $t$ in $S_{m'}$, we start by an application of Proposition 7 with $m$ replaced by $m'$ and $S_m$ by $S_{m'}$ and get

$$\mathbb{P}_s\left[\sup_{u\in S_{m'}} \frac{\nu_n[\tilde{\gamma}_{m'}(\cdot,t) - \tilde{\gamma}_{m'}(\cdot,u)]}{d^2(t,u)\vee x^2} > \tau\right] \le 3.1\exp\left[-nh_{m'}(x)\right] \tag{6.26}$$

with $h_{m'}$ given by Proposition 7. We now set $d = d(s_m,t)$.

In the metric case, Assumption **Lip** holds and a suitable version of Bernstein's inequality (see Birgé and Massart 1998, Lemma 8) leads to a bound of the form

$$\mathbb{P}_s\left[\frac{\nu_n[\tilde{\gamma}_m(\cdot,s_m) - \tilde{\gamma}_{m'}(\cdot,t)]}{d^2\vee x^2} > \tau\right] \le \exp\left(\frac{-(n\tau^2/2)(d^2\vee x^2)^2}{v^2 + c\tau(d^2\vee x^2)}\right)$$

provided that for all integers $j \ge 2$

$$\frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_s\left[|\tilde{\gamma}_m(Z_i,s_m) - \tilde{\gamma}_{m'}(Z_i,t)|^j\right] \le \frac{j!}{2}v^2c^{j-2} \quad .$$

We only have to identify $v^2$ and $c$. From (6.1) and (6.2) or (6.3) and (6.4) it can be seen that $v^2 = A^2[d^2 + n^{-1}E(D_m \vee D_{m'})]$ and $c = AB$. Since $x^2 \ge n^{-1}E(D_m \vee D_{m'})$, $v^2 \le 2A^2(d^2\vee x^2)$ and finally

$$\mathbb{P}_s\left[\frac{\nu_n[\tilde{\gamma}_m(\cdot,s_m) - \tilde{\gamma}_{m'}(\cdot,t)]}{d^2\vee x^2} > \tau\right] \le \exp\left[\frac{-n\tau^2(d^2\vee x^2)}{4A^2 + 2\tau AB}\right] \quad .$$

In order to handle the linear cases we first notice that

$$\mathbb{P}_s\left[\frac{\nu_n[\tilde{\gamma}_m(\cdot,s_m) - \tilde{\gamma}_{m'}(\cdot,t)]}{d^2\vee x^2} > \tau\right] = \mathbb{P}_s\left[\frac{\nu_n[t(\cdot) - s_m(\cdot)]}{d^2\vee x^2} > \frac{\tau}{2}\right]$$

and Bernstein's inequality implies, since $d^2 \vee x^2 \geq dx$, that

$$\mathbb{P}_s\left[\frac{\nu_n[\tilde{\gamma}_m(\cdot, s_m) - \tilde{\gamma}_{m'}(\cdot, t)]}{d^2 \vee x^2} > \tau\right] \leq \exp\left[\frac{-n\tau^2 d^2 x^2/8}{\int (s_m - t)^2 s + \|s_m - t\|_\infty \tau dx/6}\right] . \tag{6.27}$$

It then follows from (6.24) that

$$\|s_m - t\|_\infty \leq \left(\Phi_m^2 D_m + \Phi_{m'}^2 D_{m'}\right)^{1/2} d \ .$$

Therefore, setting $\delta = (D_m \vee D_{m'})^{1/2}$ one gets for cases **i)** and **iii)** $\|s_m - t\|_\infty \leq \sqrt{2}\Phi\delta d$ and for case **ii)** $\|s_m - t\|_\infty \leq \sqrt{2}\Gamma\delta^2 d$. As to $\int (s_m - t)^2 s$ it can be bounded by $d^2\|s\|_\infty$ for case **i)**, by $(d\Phi)^2$ for case **ii)** and by $d\|s\|\|s_m - t\|_\infty \leq \Phi d^2\delta\|s\|$ for case **iii)**. Together with the elementary inequality $(a+b)^{-1} \geq (1/2)(a^{-1} \wedge b^{-1})$, these bounds lead to the following upper bounds for (6.27):

$$\exp\left[\frac{-n}{16}\left(\frac{\tau^2 x^2}{\|s\|_\infty} \wedge \frac{6\tau x}{\sqrt{2}\Phi\delta}\right)\right] \quad \text{in case } \mathbf{i)} \ ;$$

$$\exp\left[\frac{-n}{16}\left(\frac{\tau^2 x^2}{\Phi^2} \wedge \frac{6\tau x}{\sqrt{2}\Gamma\delta^2}\right)\right] \quad \text{in case } \mathbf{ii)} \ ;$$

$$\exp\left[\frac{-n}{16}\left(\frac{\tau^2 x^2}{\Phi\delta\|s\|} \wedge \frac{6\tau x}{\sqrt{2}\Phi\delta}\right)\right] \quad \text{in case } \mathbf{iii)} \ .$$

Putting these bounds together with inequality (6.26) we get

$$\mathbb{P}_s\left[\sup_{u \in S_{m'}} \frac{\nu_n[\tilde{\gamma}_m(., s_m) - \tilde{\gamma}_{m'}(., u)]}{d^2(t,u) \vee x^2 \vee d^2(s_m, t)} > \frac{k}{4}\right] \leq 4.1 \exp\left[-nh_{m,m'}(x)\right] \tag{6.28}$$

where $h_{m,m'}$ takes the following values:

$$h_{m,m'}(x) = \kappa \frac{k^2 x^2}{A^2 + ABk} \qquad \text{in the metric situation} \ ; \tag{6.29}$$

$$h_{m,m'}(x) = \kappa\left[\frac{x^2}{\|s\|_\infty} \wedge \frac{x\sqrt{2}}{\Phi\delta}\right] \qquad \text{in case } \mathbf{i)} \ ; \tag{6.30}$$

$$h_{m,m'}(x) = \kappa\left[\frac{x^2}{\Phi^2} \wedge \frac{x}{\Gamma\delta^2}\right] \qquad \text{in case } \mathbf{ii)} \ ; \tag{6.31}$$

$$h_{m,m'}(x) = \frac{\kappa}{\Phi\delta}\left[\frac{x^2}{\|s\|} \wedge x\right] \qquad \text{in case } \mathbf{iii)} \ . \tag{6.32}$$

Here $\kappa$ denotes some numerical constant, not necessarily the same in each case. Now, for any $\varepsilon > 0$, since $x > 0$, one can always choose $t$ in such a way that

$$d(s,t) \leq \left[(1+\varepsilon) \inf_{u \in S_{m'}} d(s,u)\right] \vee x$$

and get for any $u \in S_{m'}$

$$\begin{aligned} d(u,t) \vee d(s_m,t) &\leq d(s,t) + [d(u,s) \vee d(s_m,s)] \\ &\leq [(1+\varepsilon)d(u,s)] \vee x + [d(u,s) \vee d(s_m,s)] \\ &\leq (2+\varepsilon)[d(u,s) \vee d(s,s_m) \vee x] \ . \end{aligned}$$

Substitution of this inequality into (6.28) leads to (6.25), since $\varepsilon$ is arbitrary.

Now, recalling that $\sigma_m = 96\Phi(D_m/n)^{1/2}$ and setting $\lambda = 1$ in the linear situation, we fix some element $m$ in $\mathcal{M}_n$, and define $x_{m'}$ for any $m' \in \mathcal{M}_n$ by

$$x_{m'}^2 = \sigma_m^2 \vee \sigma_{m'}^2 \vee \left[\frac{\lambda}{n}(L_{m'}D_{m'} \vee L_m D_m)\right] + \frac{\theta}{n} \quad \text{with } \theta \geq 1.$$

We denote by $\Omega(\theta)$ the set

$$\Omega(\theta) = \left\{ \sup_{m' \in \mathcal{M}_n} \sup_{u \in S_{m'}} \frac{\nu_n[\tilde{\gamma}_m(\cdot, s_m) - \tilde{\gamma}_{m'}(\cdot, u)]}{d^2(s,u) \vee d^2(s_m,s) \vee x_{m'}^2} > k \right\}$$

and we want to bound $\mathbb{P}_s[\Omega(\theta)]$. Since $x_{m'} > \sigma_m \vee \sigma_{m'}$ we can use (6.25) to get

$$\mathbb{P}_s[\Omega(\theta)] \leq 4.1 \sum_{m' \in \mathcal{M}_n} \exp[-nh_{m,m'}(x_{m'})] \ . \tag{6.33}$$

In the metric situation, we get from (6.29)

$$\begin{aligned} \mathbb{P}_s[\Omega(\theta)] &\leq 4.1 \sum_{m'} \exp\left[-nx_{m'}^2/\lambda\right] \\ &\leq 4.1 \sum_{m'} \exp\left[-(\lambda L_{m'}D_{m'} + \theta)/\lambda\right] \\ &\leq 4.1 \exp\left[-\frac{\theta}{\lambda}\right] \sum_{m'} \exp\left[-L_{m'}D_{m'}\right] \\ &\leq 4.1\,\Sigma \exp\left[-\frac{\kappa k^2 \theta}{A^2 + ABk}\right] \ . \end{aligned} \tag{6.34}$$

In order to deal with the linear situation we first notice that

$$(2n)^{1/2}x_{m'} \geq (L_m D_m \vee L_{m'} D_{m'})^{1/2} + \theta^{1/2} \quad . \tag{6.35}$$

- In case **i)** we get from (6.30)

$$\mathbb{P}_s[\Omega(\theta)] \leq 4.1 \sum_{m'} \exp\left[-\kappa n \left(\frac{x_{m'}^2}{\|s\|_\infty} \wedge \frac{x_{m'}\sqrt{2}}{\Phi(D_m \vee D_{m'})^{1/2}}\right)\right]$$

and from (6.35) since $n \geq \Gamma'^2(D_m \vee D_{m'})$ and $L_{m'} \geq 1$ for all $m' \in \mathscr{M}_n$

$$\sqrt{2}nx_{m'}(D_m \vee D_{m'})^{-1/2} \geq \sqrt{n} + \Gamma'\sqrt{\theta}. \tag{6.36}$$

Then, recalling that $\Sigma(x)$ is given by (6.20), we derive

$$\begin{aligned}
\mathbb{P}_s[\Omega(\theta)] &\leq 4.1 \sum_{m'} \exp\left[-\kappa \left(\frac{\theta + L_{m'} D_{m'}}{\|s\|_\infty} \wedge \frac{\Gamma'\sqrt{\theta} + \sqrt{n}}{\Phi}\right)\right] \\
&\leq 4.1 \exp\left[-\kappa \left(\frac{\sqrt{\theta}\Gamma'}{\Phi} \wedge \frac{\theta}{\|s\|_\infty}\right)\right] \\
&\quad \times \sum_{m'} \exp\left[-\kappa \left(\frac{L_{m'} D_{m'}}{\|s\|_\infty} \wedge \frac{\sqrt{n}}{\Phi}\right)\right] \\
&\leq 4.1 \Sigma\left(\frac{\kappa}{\|s\|_\infty \vee \Phi}\right) \exp\left[-\kappa \left(\frac{\sqrt{\theta}\Gamma'}{\Phi} \wedge \frac{\theta}{\|s\|_\infty}\right)\right]. \quad (6.37)
\end{aligned}$$

- Under **ii)** one gets from (6.33) and (6.31)

$$\mathbb{P}_s[\Omega(\theta)] \leq 4.1 \sum_{m'} \exp\left[-\kappa n \left(\frac{x_{m'}^2}{\Phi^2} \wedge \frac{x_{m'}}{\Gamma(D_m \vee D_{m'})}\right)\right].$$

Using the bound on $D_{m'}$ we get from (6.35)

$$\begin{aligned}
\frac{nx_{m'}}{D_m \vee D_{m'}} &\geq \frac{1}{\sqrt{2}}\left[\Gamma'(\log n)^2 + \Gamma'^2 \log n^4 (\theta/n)^{1/2}\right] \\
&\geq \frac{\Gamma'(\log n)^2}{2\sqrt{2}}\left[1 + 1 \vee \Gamma'(\log n)^2 (\theta/n)^{1/2}\right]
\end{aligned}$$

and therefore, setting

$$\Xi = \frac{\theta + D_{m'}}{\Phi^2} \wedge \frac{\Gamma'(\log n)^2}{2\Gamma\sqrt{2}}\big(1 + 1 \vee \Gamma'(\log n)^2 (\theta/n)^{1/2}\big),$$

$$
\begin{aligned}
&\mathbb{P}_s\left[\Omega(\theta)\right] \\
&\quad\leq \quad 4.1\sum_{m'}\exp[-\kappa\,\Xi] \\
&\quad\leq \quad 4.1\exp\left[-\kappa\left(\frac{\Gamma'(\log n)^2}{2\Gamma\sqrt{2}}\left(\Gamma'(\log n)^2(\theta/n)^{1/2}\vee 1\right)\wedge\frac{\theta}{\Phi^2}\right)\right] \\
&\qquad\times\sum_{m'}\exp\left[-\kappa\left(\frac{D_{m'}}{\Phi^2}\wedge\frac{\Gamma'(\log n)^2}{2\Gamma\sqrt{2}}\right)\right] \\
&\quad\leq \quad 4.1\,\Sigma_J\exp\left[-\kappa\left(\frac{\Gamma'(\log n)^2}{2\Gamma\sqrt{2}}\left(\Gamma'(\log n)^2(\theta/n)^{1/2}\vee 1\right)\wedge\frac{\theta}{\Phi^2}\right)\right]
\end{aligned}
\tag{6.38}
$$

where by assumption $J\leq n(\Gamma')^{-2}(\log n)^{-4}$ and

$$
\Sigma_J=\sum_{j=1}^{J}\Gamma_1 j^{\Gamma_2}\exp\left[-\kappa\left(\frac{j}{\Phi^2}\wedge\frac{\Gamma'(\log n)^2}{2\Gamma\sqrt{2}}\right)\right]. \tag{6.39}
$$

- Under **iii)** one gets from (6.32)

$$
\mathbb{P}_s[\Omega(\theta)]\leq 4.1\sum_{m'}\exp\left[-\frac{n\kappa}{\Phi(D_m\vee D_{m'})^{1/2}}\left(\frac{x_{m'}^2}{\|s\|}\wedge x_{m'}\right)\right]\ .
$$

We modify the linear term as before with (6.36) and use the following inequality

$$
2nx_{m'}^2\geq D_m\vee D_{m'}+2[2\theta(D_m\vee D_{m'})]^{1/2}
$$

to deal with the quadratic term. Since $n\geq\Gamma'^2(D_m\vee D_{m'})$ we get, setting

$$
\Xi=\frac{(D_m\vee D_{m'})^{1/2}+2\sqrt{2\theta}}{\|s\|_2}\wedge\left(\Gamma'\sqrt{2\theta}+\sqrt{2n}\right),
$$

$$
\begin{aligned}
\mathbb{P}_s[\Omega(\theta)]&\leq 4.1\sum_{m'}\exp\left[-\frac{\kappa\,\Xi}{2\Phi}\right] \\
&\leq 4.1\exp\left[-\frac{\kappa\sqrt{\theta}}{\Phi\sqrt{2}}\left(\Gamma'\wedge\|s\|^{-1}\right)\right] \\
&\qquad\times\sum_{j=1}^{\infty}\Gamma_1 j^{\Gamma_2}\exp\left[-\frac{\kappa\sqrt{j}}{2\Phi}\left(\Gamma'\wedge\|s\|^{-1}\right)\right]\ .
\end{aligned}
\tag{6.40}
$$

Assuming that $\Omega^c(\theta)$ is true and recalling that **C** holds we can deduce that for any $m'$ and $u \in S_{m'}$ such that $\gamma_n(u) + \text{pen}(m') \leq \gamma_n(s_m) + \text{pen}(m)$,

$$d^2(s,u) + d^2(s_m,s) + x_{m'}^2 \geq k^{-1}\nu_n[\tilde{\gamma}_m(\cdot,s_m) - \tilde{\gamma}_{m'}(\cdot,u)]$$
$$\geq 2\left(d^2(s,u) - U_m^2 - k_1\frac{D_{m'}}{n}\right) - \frac{\text{pen}(m) - \text{pen}(m')}{k} \quad .$$

Therefore any minimum penalized contrast estimator $\hat{s} \in S_{\hat{m}}$ satisfies

$$\mathbb{1}_{\Omega^c(\theta)} d^2(s,\hat{s}) \leq d^2(s,s_m) + x_{\hat{m}}^2 + 2U_m^2 + 2k_1\frac{D_{\hat{m}}}{n} + \frac{\text{pen}(m) - \text{pen}(\hat{m})}{k} \quad .$$

It follows from (6.19) with $\kappa' = 96^2$ and our choice of $\sigma_m$ that (6.17) also holds in the linear situation. Therefore one has $x_{\hat{m}}^2 + 2k_1 D_{\hat{m}}/n \leq k^{-1}[\text{pen}(m) + \text{pen}(\hat{m})] + \theta/n$ which implies that

$$\mathbb{1}_{\Omega^c(\theta)} d^2(s,\hat{s}) \leq 2k^{-1}\text{pen}(m) + 2U_m^2 + d^2(s,s_m) + \theta/n \quad . \tag{6.41}$$

Let us now define

$$V = \left[d^2(s,\hat{s}) - 2k^{-1}\text{pen}(m) - 2U_m^2 - d^2(s,s_m)\right] \vee 0 \quad .$$

Then for any $m \in \mathcal{M}_n$ and any positive number $l$

$$\mathbb{E}_s\left[d^{2l}(s,\hat{s})\right] \leq 4^{(l-1)\vee 0}\left\{2^l\mathbb{E}_s\left[U_m^{2l}\right] + \left[2k^{-1}\text{pen}(m)\right]^l + d^{2l}(s,s_m) + \mathbb{E}_s\left[V^l\right]\right\} \quad .$$

It follows from (6.41) that if $\theta \geq 1$, $\mathbb{P}_s[V > \theta/n] \leq \mathbb{P}_s[\Omega(\theta)]$ and therefore

$$\mathbb{E}_s\left[V^l\right] = n^{-l}\mathbb{E}_s\left[(nV)^l\right] = n^{-l}\int_0^\infty \mathbb{P}_s\left[nV > y^{1/l}\right] dy$$
$$\leq n^{-l}\left[1 + \int_1^\infty \mathbb{P}_s\left[\Omega\left(y^{1/l}\right)\right] dy\right]$$

which, together with (6.34) proves (6.18). It should be noticed here that in the above computations, according to our remark following Proposition 7, the quantity $\mathbb{P}_s[\Omega(\theta)]$ should be understood as an outer probability if necessary. This has no effect in the above proof since $V$ is measurable from our measurability assumption on $\hat{s}$.

In the linear situation, $k = 1/2$ and $U_m = d(s, s_m)$. One then derives analogously from (6.41) that

$$\mathbb{E}_s\left[d^{2l}(s,\hat{s})\right] \leq 3^{(l-1)\vee 0}\left[4^l \mathrm{pen}(m)^l + 3^l d^{2l}(s,s_m) + n^{-l}(1+I)\right] \quad (6.42)$$

where

$$I = \int_1^\infty \mathbb{P}_s\left[\Omega\left(y^{1/l}\right)\right] dy \ .$$

It therefore remains to bound $I$ in each of the three linear cases. Under **i)**, by (6.37) $I$ is bounded by some constant depending only on $\|s\|_\infty$, $\Gamma'$, $\Phi$, $l$ and $\Sigma[\kappa/(\Phi \vee \|s\|_\infty)]$. Under **iii)** an upper bound for $I$ depends on $\rho = (\Gamma' \wedge \|s\|^{-1})/\Phi$. More precisely by (6.40) one gets

$$\mathbb{P}_s\left[\Omega\left(y^{1/l}\right)\right] \leq 4.1 \exp\left[-\frac{\kappa\rho}{\sqrt{2}} y^{1/(2l)}\right] \sum_{j=1}^\infty \Gamma_1 j^{\Gamma_2} \exp\left[-\frac{\kappa\rho}{2}\sqrt{j}\right]$$

which implies that (6.22) holds. Finally under **ii)** one can see from (6.39) that, due to the bound on $J$,

$$\Gamma_1^{-1}\Sigma_J \leq \left[\Gamma'^{-2}(\log n)^{-4}\right]^{\Gamma_2+1} \exp\left[-\log n\left(\frac{\kappa\Gamma'\log n}{2\sqrt{2}\Gamma} - (\Gamma_2+1)\right)\right] + \sum_{j=1}^\infty \exp\left[-\Phi^{-2}\kappa j + \Gamma_2 \log j\right]$$

is bounded independently of $n$. Then from (6.38)

$$(4.1\Sigma_J)^{-1}\int_1^\infty \mathbb{P}_s\left[\Omega\left(y^{1/l}\right)\right] dy \leq \int_1^\infty \exp\left[-\Phi^{-2}\kappa y^{1/l}\right] dy + \exp\left[-\log n\left(\frac{\kappa\Gamma'\log n}{2\sqrt{2}\Gamma} - l\right)\right] + \int_{n^l}^{+\infty} \exp\left[-\kappa\frac{\Gamma'^2(\log n)^4 y^{1/2l}}{2\sqrt{2}\Gamma\sqrt{n}}\right] dy \ .$$

Setting $y = n^l x$ in the last integral shows that it is bounded independently of $n$ and the conclusion follows. □

## 7. Proofs of the main results

### *7.1. Maximum likelihood estimation*

We now want to show how one can apply Theorem 8 to maximum likelihood estimation. The framework has been given in Section 3.3.1: we observe $n$ independent identically distributed random variables $Z_1, \ldots, Z_n$ of density $s^2$ with respect to the probability $\mu$ and we have at hand a family of models $S_m \subset \mathscr{S}$ where $\mathscr{S}$ is the set of nonnegative elements of norm 1 in $\mathbb{L}_2(\mu)$. In order to apply the general theory it is convenient to introduce Assumption $\mathbf{M'}_{2,\infty}$:

$\mathbf{M'}_{2,\infty}$ *For each $m \in \mathscr{M}_n$ one can find constants $B'_m \geq 1$, $D_m \geq 1$ and $r_m$ such that for each $\delta > 0$ and each ball $\mathscr{B} \subset S_m$ with radius $\sigma \geq 5\delta \vee (D_m/n)^{1/2}$ there exists a finite set $T = T(m, \delta, \mathscr{B}) \subset \mathscr{B}$ with*

$$|T| \leq (B'_m \sigma/\delta)^{D_m} \tag{7.1}$$

*and a mapping $\pi = \pi(m, \delta, \mathscr{B})$ from $\mathscr{B}$ to $T$ such that $d(u, \pi u) \leq \delta$ for all $u$ in $\mathscr{B}$ and*

$$\sup_{u \in \pi^{-1}(t)} \|u - t\|_\infty \leq r_m \delta \quad \text{for all } t \text{ in } T \ . \tag{7.2}$$

The following is a generalized version of Theorem 2.

**Theorem 10** *Assume that $\mu$ is a probability and that the family of models $\{S_m\}_{m \in \mathscr{M}_n}$ satisfy the assumptions $\mathbf{M'}_{2,\infty}$ with $\sup_{m \in \mathscr{M}_n} D_m \leq n$ and that the weights $L_m$ satisfy* (6.16). *Define $\eta_m$ by $\int (s^2 \vee \eta_m)\, d\mu = 1 + D_m/n$ and* $\mathrm{pen}(m) \geq \kappa_8(L_m + \mathscr{L}_m) D_m/n$ *where*

$$\mathscr{L}_m = \log\left[B'_m\left(1 + r_m\left(\frac{D_m}{n\eta_m}\right)^{1/2}\right)\right] + 1 \leq \log\left[B'_m(1 + r_m)\right] + 1$$

*and $\kappa_8$ is a suitable positive numerical constant. Let $\hat{s}$ be a minimizer with respect to $m \in \mathscr{M}_n$ and $t \in S_m$ of $-n^{-1}\sum_{i=1}^n \log[t(Z_i)] + \mathrm{pen}(m)$. Then*

$$\mathbb{E}_s\left[d^2(s, \hat{s})\right] \leq \kappa'_8\left[\inf_{m \in \mathscr{M}_n}\{K(s, S_m) + \mathrm{pen}(m)\} + \Sigma n^{-1}\right] \ . \tag{7.3}$$

*Remark*: One could of course get a similar result under the slightly more general assumption that $D_m \leq Kn$. We restrict ourselves to the case $K = 1$ for the sake of simplicity .

*Proof*: Let us first notice that, since $\mu$ is a probability measure, $\eta_m$ is well-defined and larger than $D_m/n$. We introduce the auxiliary density $\tilde{s}_m^2 = (s^2 \vee \eta_m)/(1 + D_m/n)$. Then

$$\left\| \frac{s}{\tilde{s}_m} \right\|_\infty^2 \leq 1 + \frac{D_m}{n} \leq 2 \quad \text{and} \quad \inf_x \tilde{s}_m^2(x) \geq \frac{\eta_m}{2} \geq \frac{D_m}{2n} \ . \tag{7.4}$$

Since $(1 + x)^{-1/2} \geq 1 - x/2$ for $x > -1$ one derives that $\int s\tilde{s}_m \, d\mu \geq 1 - D_m/(2n)$ which implies that

$$d^2(s, \tilde{s}_m) \leq \frac{D_m}{n} \ . \tag{7.5}$$

In order to apply Theorem 8 we define for any $m \in \mathcal{M}_n$ and $t \in S_m$ the function

$$\tilde{\gamma}_m(z, t) = -\log \left[ \frac{\tilde{s}_m^2(z) + t^2(z)}{2} \right] \ .$$

We have to show that these functions satisfy Assumptions **M** and **C**. In order to check **M** we use the following Lemmas, recalling that the Hellinger distance $h(g_1, g_2)$ between two densities is given by $2h^2(g_1, g_2) = \int (\sqrt{g_1} - \sqrt{g_2})^2 \, d\mu$.

**Lemma 3** *Let $f, g, g_1, g_2$ be densities with respect to some measure $\mu$, then*

$$\mathbb{E}_f \left[ \left| \frac{1}{2} \log \left( \frac{g + g_1}{g + g_2} \right) \right|^j \right] \leq \frac{4}{7} \frac{j!}{2} h^2(g_1, g_2) \left[ \left\| \frac{f}{g} \right\|_\infty \wedge 4 \left\| \frac{f}{g_1 \wedge g_2} \right\|_\infty \right] \quad \textit{for all } j \geq 2 \ .$$

*Proof*: The bound involving $\|f/g\|_\infty$ has been proved in Birgé and Massart (1994, Proposition 2) (see also Van de Geer 1995, Lemma 3.3 for an analogous result). For the other part we notice that $x^j/j! \leq e^x - x - 1$ and $x - 1 - \log x \leq (x - x^{-1})^2/7$ to derive that when $g_1 \geq g_2$,

$$\frac{1}{j!} \left( \frac{1}{2} \log \frac{g + g_1}{g + g_2} \right)^j \leq \frac{1}{j!} \left( \log \sqrt{\frac{g_1}{g_2}} \right)^j \leq \sqrt{\frac{g_1}{g_2}} - 1 - \log \left( \sqrt{\frac{g_1}{g_2}} \right) \leq \frac{1}{7} \left( \sqrt{\frac{g_1}{g_2}} - \sqrt{\frac{g_2}{g_1}} \right)^2$$

$$= \frac{1}{7}\left(\sqrt{g_1}-\sqrt{g_2}\right)^2\left(\frac{\sqrt{g_1}+\sqrt{g_2}}{\sqrt{g_1 g_2}}\right)^2$$
$$\leq \frac{4}{7}\frac{\left(\sqrt{g_1}-\sqrt{g_2}\right)^2}{g_1 \wedge g_2} \quad .$$

A symmetric result holds when $g_2 \geq g_1$ and integration with respect to $f\mu$ gives the result. □

**Lemma 4** *Assume that $f,\ g_1,\ g_2, s_1^2$ and $s_2^2$ are densities with respect to the probability measure $\mu$ and that $\|f/s_i^2\|_\infty \leq 2$ for $i = 1, 2$. For any integer $j \geq 2$ one has*

$$\mathbb{E}_f\left[\left|\frac{1}{4}\log\left(\frac{s_1^2+g_1}{s_2^2+g_2}\right)\right|^j\right] \leq \frac{4}{7}\frac{j!}{2}\left[h^2(g_1,g_2)+4h^2(s_1^2,s_2^2)\right] \quad .$$

*Proof*: Successive applications of Lemma 3 give

$$\mathbb{E}_f\left[\left|\frac{1}{4}\log\left(\frac{s_1^2+g_1}{s_2^2+g_2}\right)\right|^j\right]$$
$$\leq \frac{1}{2}\mathbb{E}_f\left[\left|\frac{1}{2}\log\left(\frac{s_1^2+g_1}{s_2^2+g_1}\right)\right|^j\right] + \frac{1}{2}\mathbb{E}_f\left[\left|\frac{1}{2}\log\left(\frac{s_2^2+g_1}{s_2^2+g_2}\right)\right|^j\right]$$
$$\leq \frac{1}{2}\frac{4}{7}\frac{j!}{2}\left[4h^2(s_1^2,s_2^2)\left\|\frac{f}{s_1^2\wedge s_2^2}\right\|_\infty + h^2(g_1,g_2)\left\|\frac{f}{s_2^2}\right\|_\infty\right]$$

and the result follows since $\|f/s_i^2\|_\infty \leq 2$. □

According to the definition of $\tilde{\gamma}_m$ we can choose

$$\Delta_{m,m'}(x,u,v) = A^{-1}\left|\log\left(\frac{\tilde{s}_m^2(x)+u^2(x)}{\tilde{s}_{m'}^2(x)+v^2(x)}\right)\right| \quad \text{and} \quad M = A \quad .$$

Then (6.1) is satisfied and

$$h^2(\tilde{s}_m^2,\tilde{s}_{m'}^2) \leq 4\left[h^2(s^2,\tilde{s}_m^2)\vee h^2(s^2,\tilde{s}_{m'}^2)\right] = 2\left[d^2(s,\tilde{s}_m)\vee d^2(s,\tilde{s}_{m'})\right] \quad ,$$

hence by (7.5) $h^2(\tilde{s}_m^2,\tilde{s}_{m'}^2) \leq 2n^{-1}(D_m \vee D_{m'})\mathbb{1}_{\{m\neq m'\}}$. An application of Lemma 4, which is valid because of (7.4), leads to

$$\mathbb{E}_s\left[\Delta_{m,m'}^j(X_i,u,v)\right] = \left(\frac{4}{A}\right)^j\frac{2}{7}\frac{j!}{2}\left[d^2(u,v)+16\frac{D_m\vee D_{m'}}{n}\mathbb{1}_{\{m\neq m'\}}\right] \quad .$$

The choice $A = 4\sqrt{2/7}$ gives (6.2) with $B = \sqrt{7/2}$ and $E = 16$. Then (7.1) implies (6.5) and the lower bound on $\tilde{s}_m$ in (7.4) together with Lemma 6 of Birgé and Massart (1998) imply that whenever $\|t - u\|_\infty \leq r_m \delta$ for $t$ and $u$ in $S_m$, then

$$\|\Delta_{m,m}(x, t, u)\|_\infty \leq 2A^{-1} r_m \delta \sqrt{2/\eta_m} \ .$$

Therefore (6.6) holds with $r'_m = r_m \sqrt{7/(4\eta_m)} \leq r_m \sqrt{(7n)/(4D_m)}$ by (7.4). The value of $\mathscr{L}_m$ follows from the value of $\mathscr{L}'_m$ given in Proposition 7 after a suitable modification of the multiplicative constant which can be included in $\kappa_8$. It remains to check Assumption **C**. We proceed as in Birgé and Massart (1998). If $t \in S_{m'}$ is such that $\gamma_n(t) + \text{pen}(m') \leq \gamma_n(s_m) + \text{pen}(m)$ by the concavity of the logarithm

$$\frac{1}{n}\sum_{i=1}^{n} \log\left[\left(\frac{t^2 + \tilde{s}_{m'}^2}{2}\right)(X_i)\right] \geq \frac{1}{n}\sum_{i=1}^{n} \{\log[\tilde{s}_{m'}(X_i)] + \log[s_m(X_i)]\} + \text{pen}(m') - \text{pen}(m)$$

and since by (7.4) $\log \tilde{s}_{m'} \geq \log s - (1/2)\log(1 + D_{m'}/n) \geq \log s - D_{m'}/(2n)$,

$$\begin{aligned}
&\nu_n[\tilde{\gamma}_m(., s_m) - \tilde{\gamma}_{m'}(., t)] \\
&\geq \mathbb{P}_n\left[\log \frac{2ss_m}{\tilde{s}_m^2 + s_m^2}\right] - \mathbb{E}_s\left[\log \frac{\tilde{s}_{m'}^2(X) + t^2(X)}{2s^2(X)} - \log \frac{\tilde{s}_m^2(X) + s_m^2(X)}{2s^2(X)}\right] \\
&\quad - \frac{D_{m'}}{2n} + \text{pen}(m') - \text{pen}(m) \\
&\geq \mathbb{P}_n\left[\log \frac{2s^2}{\tilde{s}_m^2 + s_m^2}\right] - \frac{1}{2}\mathbb{P}_n\left[\log \frac{s^2}{s_m^2}\right] - \frac{D_{m'}}{2n} - K\left(s, \left[\frac{\tilde{s}_m^2 + s_m^2}{2}\right]^{1/2}\right) \\
&\quad + K\left(s, \left[\frac{\tilde{s}_{m'}^2 + t^2}{2}\right]^{1/2}\right) + \text{pen}(m') - \text{pen}(m) \ .
\end{aligned}$$

Since

$$d^2\left(s, \left[\frac{s^2 + t^2}{2}\right]^{1/2}\right) \leq \frac{3}{2} d^2\left(s, \left[\frac{\tilde{s}_{m'}^2 + t^2}{2}\right]^{1/2}\right) + 3\, d^2\left(\left[\frac{s^2 + t^2}{2}\right]^{1/2}, \left[\frac{\tilde{s}_{m'}^2 + t^2}{2}\right]^{1/2}\right)$$

and $d^2(u, v) = 2h^2(P_u, P_v)$, one derives from (7.2) and (7.3) of Lemma 5 of Birgé and Massart (1998), that

$$d^2\left(s, \left[\frac{\tilde{s}_{m'}^2 + t^2}{2}\right]^{1/2}\right) \geq \frac{2}{3} d^2\left(s, \left[\frac{s^2 + t^2}{2}\right]^{1/2}\right) - 2d^2\left(\left[\frac{s^2 + t^2}{2}\right]^{1/2}, \left[\frac{\tilde{s}_{m'}^2 + t^2}{2}\right]^{1/2}\right)$$
$$\geq \frac{2}{3} 0.29^2\, d^2(s, t) - d^2(s, \tilde{s}_{m'}) \ .$$

Choosing $k = 0.028 < 0.29^2/3$, we deduce from (8.8) below that

$$K\left(s, \left[\frac{\tilde{s}_{m'}^2 + t^2}{2}\right]^{1/2}\right) \geq 2kd^2(s, t) - d^2(s, \tilde{s}_{m'}) \ .$$

It follows from the concavity of the logarithm that

$$2K\left(s, \left[\frac{\tilde{s}_m^2 + s_m^2}{2}\right]^{1/2}\right) \leq K(s, \tilde{s}_m) + K(s, s_m)$$
$$\leq (2 + \log 2)\, d^2(s, \tilde{s}_m) + K(s, s_m)$$

since (7.4) implies by (8.8) that $K(s, \tilde{s}_m) \leq (2 + \log 2) d^2(s, \tilde{s}_m)$. Putting all these bounds together with (7.5) we get

$$\nu_n[\tilde{\gamma}_m(., s_m) - \tilde{\gamma}_{m'}(., t)]$$
$$\geq 2kd^2(s, t) + \text{pen}(m') - \text{pen}(m) - \frac{3D_{m'}}{2n} - \left(1 + \frac{\log 2}{2}\right)\frac{D_m}{n}$$
$$- \frac{1}{2} K(s, s_m) - \mathbb{P}_n\left[\frac{1}{2}\log\frac{s^2}{s_m^2} - \log\frac{2s^2}{\tilde{s}_m^2 + s_m^2}\right] \ .$$

Since

$$\mathbb{E}_s\left[\frac{1}{2}\log\frac{s^2}{s_m^2}(X_i) - \log\frac{2s^2}{\tilde{s}_m^2 + s_m^2}(X_i)\right] \leq \frac{1}{2} K(s, s_m)$$

we finally see that **C** holds with $k_1 = 3/(4k)$ and

$$2k\mathbb{E}_s[U_m^2] < \left(1 + \frac{\log 2}{2}\right)\frac{D_m}{n} + K(s, s_m) \ .$$

The application of Theorem 8 leads to inequality (7.3) since $d^2(s, s_m) \leq K(s, s_m)$. □

*Proof of Theorem 2*: We can now derive Theorem 2 from Theorem 10. It is enough to check the properties (7.1) and (7.2) on $\bar{S}_m$ rather than $S_m$ since it is a larger set and they immediately follow from Lemma 9 with $B' = 5$ and $r_m = \bar{r}_m$. One can therefore bound $\mathscr{L}_m$ by $\bar{\kappa}[1 + \log(1 + \bar{r}_m)]$ and the result follows from a suitable modification of the constants since $L_m \geq 1$.

### *7.2. Other penalized minimum contrast estimation procedures*

#### 7.2.1. Penalized projection estimation

We have to prove Theorem 3. Its assumptions imply that we can apply the case iii) of Theorem 9 with $\Gamma' = \Gamma_1 = 1$ and $\Gamma_2 = 0$ and Theorem 3 follows. A complete treatment of penalized projection estimators is contained in Birgé and Massart (1997).

#### 7.2.2. Penalized least squares and minimum $\mathbb{L}_1$ regression

We recall that one observes pairs $(X_i, Y_i) = Z_i$ with $Y_i = s(X_i) + W_i$ where the variables $X_i$ and $W_i$ are all independent with respective distributions $R_i$ and $Q_i$ independent of $s$ and the $X_i$'s are defined on a compact set $\mathscr{X}$. Here $s \in \mathscr{S} \subset \mathscr{T} \subset \mathbb{L}_2(\mu)$ where $\mu$ denotes the average distribution of the $X_i$'s, $\mu = n^{-1} \sum_{i=1}^n R_i$. We shall assume hereafter [although these assumptions could be weakened as in Birgé and Massart (1993) Section 3.C] that the $W_i$'s are independent identically distributed with common distribution $Q$ and that the $X_i$'s are either independent identically distributed with common distribution $\mu$ (which is the random design setting) or that the $X_i$'s are given numbers $x_i$ (which is the fixed design setting). In the latter case, $\mu$ is the empirical measure of the $x_i$'s and the results, as in Van de Geer (1995), are given in the form of controls of $d^2(s, \hat{s}) = n^{-1} \sum_{i=1}^n [s(x_i) - \hat{s}(x_i)]^2$. Given a penalty function $\mathrm{pen}(m)$ to be chosen later, we consider either the penalized least squares estimator which is a minimizer with respect to $m \in \mathscr{M}_n$ and $t \in S_m$ of $n^{-1} \sum_{i=1}^n [Y_i - t(X_i)]^2 + \mathrm{pen}(m)$ or the minimum penalized $\mathbb{L}_1$ estimator which is a minimizer with respect to $m \in \mathscr{M}_n$ and $t \in S_m$ of $n^{-1} \sum_{i=1}^n |Y_i - t(X_i)| + \mathrm{pen}(m)$. Then the following result holds.

**Theorem 11** *Assume that the family* $\{S_m, m \in \mathscr{M}_n\}$ *satisfies Assumption* **M'**$_{2,\infty}$, *that the weights* $L_m$ *satisfy* (6.16) *and that* $\|t\|_\infty \leq \xi$ *for any* $t \in \mathscr{T}$ *and some known constant* $\xi$. *Assume moreover that the distribution* $Q$ *of the* $W_i$*'s has one of the following properties*:

- *the errors* $W_i$ *are centered at their expectation and* $\mathbb{E}[e^{|W_1|/\xi'}] \leq 4$ *for some* $\xi' > 0$ *in the case of least squares estimation*;

- *the errors $W_i$ are centered at their median and have a distribution $Q$ with a density which is positive and continuous around the median in the case of minimum $\mathbb{L}_1$ estimation.*
  *Define the penalty function as* $\text{pen}(m) \geq \kappa_9 C(\xi, Q)(L_m + \mathscr{L}_m)D_m/n$ *where*

$$\mathscr{L}_m = \log\left[B'_m\left(1 + r_m(D_m/n)^{1/2}\right)\right] + 1 \ ,$$

  $\kappa_9$ *is a numerical constant and* $C(\xi, Q)$ *is a suitable constant which takes two different forms in the two cases considered above. Let* $\hat{s}$ *be the minimum penalized empirical contrast estimator. Then in both cases*

$$\mathbb{E}_s[d^2(s, \hat{s})] \leq \kappa'_9 \inf_{m \in \mathscr{M}_n}\left\{d^2(s, S_m) + C'(\xi, Q)\text{pen}(m)\right\} + C''(\xi, Q)\Sigma n^{-1} \ . \tag{7.6}$$

  *In the case of least squares estimation one can choose* $C'(\xi, Q) = 1$ *and* $C(\xi, Q) = C''(\xi, Q) = (\xi + \xi')^2$.

*Proof*: We shall actually prove a more general result. Following the framework given in Birgé and Massart (1993) Section 3.C we assume that $\gamma$ is given by $\gamma(z, t) = F[y - t(x)]$ where $F$ is a convex function with suitable properties connected to the distribution $Q$ of the $W'i$'s, provided that $s$ and all the elements $t$ of the models are uniformly bounded, which is our assumption. The required conditions on $F$ are given by Assumptions **Ca**, **Cc**, **Cd** and **Ce** of Birgé and Massart (1993) and it is also proved there that the two functions $[y - t(x)]^2$ and $|y - t(x)|$ satisfy these assumptions under the conditions of Theorem 11. We set $\tilde{\gamma}_m \equiv \gamma$ and apply Proposition 1 of Birgé and Massart (1993) which implies that (6.3) and (6.4) are satisfied with $\Delta_{m,m'}(x, t, u) = |t(x) - u(x)|$ and suitable constants $A$, $B$ depending on $Q$ and $\xi$. In the particular case of $F(x) = x^2$, one has

$$|\gamma(z, t) - \gamma(z, u)| = |t(x) - u(x)|\,|2w + 2s(x) - [t(x) + u(x)]|$$
$$\leq 2|t(x) - u(x)|\,[|w| + 2\xi] \ .$$

We can therefore take $B = 2\xi$ and $M(w) = 2(|w| + 2\xi)$. Our moment condition on the $W_i$'s implies that for every $j \geq 2$,

$$\mathbb{E}_s\left[2^j(|W| + 2\xi)^j\right] \leq 2^{2j-1}\left(\mathbb{E}_s\left[|W|^j\right] + 2^j\xi^j\right)$$
$$\leq \frac{4^j}{2}\left[4j!(\xi')^j + 2^j\xi^j\right]^j$$

and one can choose $A = 8(\xi' + \xi)$. If the metric assumption (7.2) holds then (6.6) is fulfilled with $r'_m = r_m$ since here $\Delta_{m,m'}(x, t, u) = |t(x) - u(x)|$ and therefore $\mathbf{M'}_{2,\infty}$ implies $\mathbf{M}_{2,\infty}$. In order to apply Theorem 8 it comes from Lemma 2 that it is enough to check **C'**. But, according to the notations and arguments of Birgé and Massart (1993) Section 3.C, there exists a function $G$ such that

$$\mathbb{E}_s[\gamma(Z_i, t) - \gamma(Z_i, s)] = \mathbb{E}_s[G(W_i, s(X_i) - t(X_i))]$$

and that for suitable positive constants $C_1$, $C_2$ and $h \in \mathbb{R}$,

$$C_1 h^2 \leq \mathbb{E}_s[G(W_i, h)] \leq C_2 h^2 \quad \text{for} \quad |h| \leq 2\xi \ .$$

In view of the independence between $X_i$ and $W_i$, these relations imply **C'**. In the quadratic case ($F(u) = u^2$), $G(w, h) = h^2$ and therefore $C_1 = C_2 = 1$ and **C'** is satisfied with $k' = k'' = 1$. The choice of $C$, $C'$ and $C''$ is justified by Theorem 8 and our computations of $A$, $B$, $k'$ and $k''$. □

*Proof of Theorem* 4: By Lemma 9 assumptions (6.5) and (6.6) are satisfied with $B'_m = 5$ and $r_m = \bar{r}_m$. Therefore Theorem 11 implies Theorem 4 via some elementary computations since $L_m \geq 1$. □

*Proof of Theorem* 5: We want to derive it from Theorem 8. As we already checked in the proof of Theorem 11, the Assumption **Lip ii)** is satisfied for the function $\gamma(z, t) = [y - t(x)]^2$ by setting $\tilde{\gamma}_m \equiv \gamma$ and $\Delta_{m,m'}(x, t, u) = |t(x) - u(x)|$, $M(w) = 2(|w| + 2\xi)$, $E = 0$, $A = 8(\xi' + \xi)$ and $B = 2\xi$ where now $\xi = 1$. Moreover **C'** (and therefore **C**) is also satisfied with $k' = k'' = 1$. There only remains to check the Assumption $\mathbf{M}_{1,[\,]}$. Let us consider some ball $\mathscr{B}$ of radius $\sigma$ in $S_m$ and some $\delta \leq \sigma/5$. From inequality (4.26), $\mathscr{B}$ is included in the image via $\theta$ of some $\mathbb{L}_1(\mu')$-ball of radius $R = \sigma^2/\Theta_1$. Applying Lemma 11 with $\varepsilon = \delta^2/\Theta_2$ we can cover this ball by a family $\mathscr{I}$ of intervals of $\mathbb{L}_1(\mu')$-diameter $\leq \varepsilon$ with cardinality bounded by

$$(3eB''_m\Theta_2/\Theta_1)^{D_m}(\sigma^2/\delta^2)^{D_m} \ . \tag{7.7}$$

Truncating the intervals if necessary, we can assume, without loss of generality that these intervals are included in $\mathscr{G}$. Therefore the images of the elements of $\mathscr{I}$ via $\chi$ are covering $\mathscr{B}$. Since $\chi$ is non-decreasing, for each interval $[g^-, g^+] \in \mathscr{I}$, $\chi([g^-, g^+]) \subset [\chi(g^-), \chi(g^+)]$ and by (4.26) the $\mathbb{L}_1(\mu)$-diameter of $[\chi(g^-), \chi(g^+)]$ is bounded by $\delta^2$. Choosing $t$ as any point in $[\chi(g^-), \chi(g^+)] \cap \mathscr{B}$ and defining $V_{m,t} = \chi(g^+) - \chi(g^-)$ we take for $T$ the set of all those $t$'s when $[g^-, g^+]$ varies in $\mathscr{I}$. We then define $\pi$ to be the projection mapping $[\chi(g^-), \chi(g^+)] \cap \mathscr{B}$ on the point $t \in T \cap [\chi(g^-), \chi(g^+)]$. Since $\Delta_{m,m}(x, t, u) = |t(x) - u(x)|$, (6.7) is fulfilled and (7.7) implies

(6.5) with $D_m$ replaced by $2D_m$ and $B'_m = (3eB''_m\Theta_2/\Theta_1)^{1/2}$. We can therefore apply Theorem 8. Elementary transformations of the constants (since $\Theta_2 \geq \Theta_1$) justify the choice of pen$(m)$ and we get

$$\mathbb{E}_s[d^2(s, \chi_{\hat{f}})] \leq \kappa'_6 \inf_{m\in\mathscr{M}_n} \left\{d^2(s, S_m) + \text{pen}(m) + (\xi' + 1)^2\Sigma n^{-1}\right\} \quad . \qquad \square$$

Using (4.26) and the fact that $\chi_g \leq 1$ we derive (4.28).

7.2.3. Estimating the support of a density

*Proof of Theorem 6*: The proof is based on the version of Theorem 8 involving the Assumption $\mathbf{M}_{1,[\,]}$. Let us check the relevant assumptions: first setting $\tilde{\gamma}_m(t, z) = -t(z)$ we see that **Lip ii)** is satisfied with $\Delta_{m,m'}(x, t, u) = |t(x) - u(x)|$, $M = A = 1$ and $B = 1$. Moreover **C'** is also satisfied with $k' = a/2$ and $k'' = b - a/2$ by (4.30). We now want to check **M** with $\mathbf{M}_{1,[\,]}$. Given some $m \in \mathscr{M}_n$, some ball $\mathscr{B}$ of radius $\sigma \geq \sqrt{D_m/n}$ in $S_m$ and some positive $\delta \leq \sigma/5$ we set $\varepsilon = \delta^2/(\pi b)$ and apply Lemma 11 to the space $\bar{G}_m$ which implies, since $G_m$ is a ball of radius $R$ in $\bar{G}_m$, that we can cover $G_m$ by a family $\mathscr{I}_\varepsilon$ of intervals of $d_1$-diameter $\varepsilon$ with

$$|\mathscr{I}_\varepsilon| \leq \left[\frac{3eB''_mR}{\varepsilon \wedge (R/2)}\right]^{D_m} = \left[3eB''_m\left(\frac{R\pi b}{\delta^2} \vee 2\right)\right]^{D_m} \quad . \qquad (7.8)$$

The images of the elements of $\mathscr{I}_\varepsilon$ via $\chi$ are therefore covering $S_m$. Since $\chi$ is non-decreasing, for each interval $[g^-, g^+] \in \mathscr{I}_\varepsilon$, $\chi$ maps $[g^-, g^+]$ into $[\chi(g^-), \chi(g^+)]$ and by (4.31) the $\|\chi(g^+) - \chi(g^-)\|_1 \leq \pi\varepsilon = \delta^2/b$. One can then build from $\chi(\mathscr{I}_\varepsilon)$ a partition $\mathscr{J}$ of $\mathscr{B}$ into sets $J$ such that each $J \subset [\chi(g^-), \chi(g^+)]$ for some pair $[g^-, g^+] \in \mathscr{I}_\varepsilon$ and $|\mathscr{J}| \leq |\mathscr{I}_\varepsilon|$. We now have to define the set $T(m, \delta, \mathscr{B})$, the mapping $\pi(m, \delta, \mathscr{B})$ and the family of functions $\{V_{m,t}\}_{t\in T}$. Given some $J \in \mathscr{J}$ with $J \subset [\chi(g^-), \chi(g^+)]$ we define $\pi(J) = t$ as any point in $J$, $V_{m,t} = \chi(g^+) - \chi(g^-)$ and we take for $T$ the set of all those $t$'s when $J$ varies in $\mathscr{J}$. It then follows from (7.8), since $\sigma^2 \geq (D_m/n) \vee (25\delta^2)$ and $D_m \leq 25\pi bnR/2$ that

$$|T| \leq \left[3eB''_m\left(\frac{R\pi b}{\delta^2} \vee 2\right)\right]^{D_m} \leq \left[\frac{3eB''_mRn\pi b}{D_m}\right]^{D_m}\left(\frac{\sigma^2}{\delta^2}\right)^{D_m}$$

which implies (6.5) with $D_m$ replaced by $2D_m$ and $B'_m = [3eB''_mRn\pi b/D_m]^{1/2}$. Since $\Delta_{m,m'}(x, u, t) = |t(x) - u(x)|$ and $\|V_{m,t}\|_1 \leq \delta^2/b$, (6.7) is fulfilled. We can therefore apply Theorem 8 and get the result via elementary transformations of the constants since $a \leq 1$. $\square$

### *7.3. Analysis of nonlinear models*

Here we use Theorems 10 and 11 to prove the risk bounds for nonlinear models stated in Theorem 7. Unlike linear models, the models treated here do not have homogeneous control of their $\mathbb{L}_2$-local metric entropy properties of the sort that condition **M** or $\mathbf{M'}_{2,\infty}$ is designed to handle best. In these inhomogeneous cases we will be content to check global $\mathbb{L}_\infty$-entropy which implies the presence of a logarithmic factor in the penalty term and therefore in the risk bounds because of the resulting large value of $B'_m$. A similar phenomenon occurs when one covers the unit ball in $\mathbb{R}^q$ by balls of radius $\delta$. The logarithm of the number of balls that are needed is, for small $\delta$, of order $-q\log\delta + C$ instead of $q\log\lambda + C$ which is needed for the covering of a ball of radius $\lambda\delta$. This makes a serious difference when $\lambda$ is not large.

*Proof of Theorem 7*: We can apply either Theorem 10 or Theorem 11 to get our conclusion provided that we are able to check (6.16) and Assumption $\mathbf{M'}_{2,\infty}$. Recalling that we have set $D_m = D'(q'+1)$ we choose $L_m = 1+2\log(RH)$. Then, $D_m L_m \geq D' + 2\log H + 2\log R$ which implies (6.16). We now have to check $\mathbf{M'}_{2,\infty}$ in each case. In order to do this we first investigate the metric properties of $\bar{S}_m$ which are described by the following

**Lemma 5** *Given three positive integers $D'$, $H$, $R$, the family $\{\phi_w \mid w \in \mathbb{R}^{q'}\}$ satisfying the assumptions of Theorem 7 and the space $\bar{S}_m = \{\sum_{j=1}^{D'} \beta_j \phi_{w_j}\}$ with $\sum_{j=1}^{D'} |\beta_j| \leq R$ and $|w_j|_1 \leq H$, one can find for any $\delta > 0$ a subset $T(\delta)$ of $\bar{S}_m$ with cardinality bounded by $[2e(2RH/\delta + 1)]^{D'(q'+1)}$ and such that for each $u \in \bar{S}_m$ there exists $t \in T(\delta)$ with $\|u - t\|_\infty \leq \delta$.*

*Proof*: Because of the Lipschitz condition on $\{\phi_w\}$, an $\mathbb{L}_\infty$-covering of $\{\phi_w \mid |w|_1 \leq H\}$ follows from a covering of the $l_1$-ball $\{w \mid |w|_1 \leq H\}$. In $\mathbb{R}^{q'}$ the number of disjoint cubes spaced at width $\varepsilon_1/q'$ that cover this ball is bounded by $[2e(H/\varepsilon_1 + 1)]^{q'}$ by Lemma 10. Then for each $w$ with $|w|_1 \leq H$ there is a $w'$ in the grid with $\|\phi_w - \phi_{w'}\|_\infty \leq |w - w'|_1 \leq \varepsilon_1$. In the same way in $\mathbb{R}^{D'}$ we cover $\{\beta \mid \sum_{j=1}^{D'} |\beta_j| \leq R\}$ using not more than $[2e(R/\varepsilon_2 + 1)]^{D'}$ cubes spaced at width $\varepsilon_2/D'$. Set $\varepsilon_1 = \delta/2R$ and $\varepsilon_2 = \delta/2H$ and use the cubical grids intersecting the $l^1$-balls as indicated above. Restricting vectors $w'_j$, $j = 1, \ldots, D'$ and $\beta'$ to these grids provides a finite set $T(\delta)$ of functions $\sum_{j=1}^{D'} \beta'_j \phi_{w'_j}$ of cardinality not more than $[2e(2RH/\delta + 1)]^{D'(q'+1)}$. Then for each $u = \sum_{j=1}^{D'} \beta_j \phi_{w_j}$ in $\bar{S}_m$ there is a $t = \sum_{j=1}^{D'} \beta'_j \phi_{w'_j}$ in $T(\delta)$ with

$$|u(x) - t(x)| \leq \left| \sum_{j=1}^{D'} \beta_j \left[ \phi_{w_j}(x) - \phi_{w'_j}(x) \right] \right| + \left| \sum_{j=1}^{D'} \left( \beta_j - \beta'_j \right) \phi_{w'_j}(x) \right|$$

$$\leq \sum_{j=1}^{D'} |\beta_j| \left| \phi_{w_j}(x) - \phi_{w'_j}(x) \right| + \sum_{j=1}^{D'} |\beta_j - \beta'_j| H$$
$$\leq R\varepsilon_1 + H\varepsilon_2 = \delta \ ,$$

uniformly for $x$ in $[-1, 1]^q$ which proves the lemma. □

Returning to the proof of Theorem 7, let us first consider the regression setting. In this case, Lemma 5 applies to $S_m$ as well as $\bar{S}_m$ since the clipping operation which maps $\bar{S}_m$ onto $S_m$ is a contraction with respect to the $\mathbb{L}_\infty$-norm. Then $\mathbf{M'}_{2,\infty}$ holds with dimension $D_m = D'(q' + 1)$ equal to the parameter dimension, $r_m = 1$ and $B'_m = 8eRH(n/D_m + 1)^{1/2}$.

For maximum likelihood density estimation, the situation is slightly more subtle. Define the norming operator $g$ from $\bar{S}_m$ to $S_m$ by $g(u') = (u' \vee n^{-1})\|u' \vee n^{-1}\|^{-1}$, fix $\delta' = [\delta/(9RH)] \wedge 1/6$ and define $T \subset S_m$ to be the image by $g$ of $\{t \in T(\delta') \mid \|t \vee 0\| \geq 1/3\}$. Now, given $u \in S_m$ there exists $u' \in \bar{S}_m$ and $t' \in T(\delta')$ with $u = g(u')$, $\|u' \vee 0\| \geq 1/2$ and $\|u' - t'\|_\infty \leq \delta' \leq 1/6$. As a consequence $\|t' \vee 0\| \geq 1/3$ since $\mu$ is a probability, $g(t') \in T$ and $\|(u' \vee n^{-1}) - (t' \vee n^{-1})\|_\infty \leq \delta'$. Moreover

$$\|g(u') - g(t')\|_\infty \leq \frac{\|(u' \vee n^{-1})\| \, \|(u' \vee n^{-1}) - \left(t' \vee n^{-1}\right)\|_\infty}{\|u' \vee n^{-1}\| \, \|t' \vee n^{-1}\|}$$
$$+ \frac{\|u' \vee n^{-1}\|_\infty | \, \|u' \vee n^{-1}\| - \|t' \vee n^{-1}\| |}{\|u' \vee n^{-1}\| \, \|t' \vee n^{-1}\|}$$
$$\leq \frac{\|u' - t'\|_\infty}{\|t' \vee 0\|} \left(1 + \frac{\|u' \vee n^{-1}\|_\infty}{\|u' \vee 0\|}\right) \ .$$

Since $\|u' \vee 0\| \geq 1/2$, then $\|u' \vee n^{-1}\|_\infty \leq \|u'\|_\infty \leq RH$ and one concludes that $\|u - t\|_\infty \leq (6RH + 3)\delta' \leq \delta$. Since by Lemma 5

$$|T| \leq [2e(2RH/\delta' + 1)]^{D'(q'+1)} \leq [2e(18R^2H^2/\delta + 13RH)]^{D'(q'+1)} \ ,$$

we can again check that $\mathbf{M'}_{2,\infty}$ holds with $D_m = D'(q' + 1), r_m = 1$ and $B'_m = 62eR^2H^2(n/D_m + 1)^{1/2}$. It follows that we can apply either Theorem 10 or Theorem 11 and that in both cases

$$(L_m + \mathscr{L}_m)D_m/n \leq \kappa_{11} \left[1 + \log(RH) + \log[1 + n/(D'q')]\right] D'q'/n$$

which justifies our choices of the penalty terms. We finally notice that (4.37) implies (4.38). Indeed, we can restrict ourselves to the case $d(s, \bar{S}_m) < 1/2$. Choose $s_m \in \bar{S}_m$ with $d(s, s_m) \leq 1/2$ then $\tilde{s}_m = (s_m \vee n^{-1})/(\|s_m \vee n^{-1}\|)$

belongs to $S_m$ and by (8.7) below

$$\|s - \tilde{s}_m\| \leq 2 \left\| s - \left(s_m \vee n^{-1}\right) \right\| \leq 2 \left( \|s - s_m\| + n^{-1} \right) \quad .$$

Since $\|s/\tilde{s}_m\|_\infty \leq n\|s\|_\infty$, (8.8) implies that $K(s, \tilde{s}_m) \leq 2[1+\log(n\|s\|_\infty)]$ $d^2(s, \tilde{s}_m)$, and therefore (4.38). □

*Remarks*: The metric entropy calculations in the Proof of Theorem 7 are similar to those used in Barron (1993) in the special case of the sigmoids. But the risk bounds given there were for penalized least squares restricted to discretizations of the parameters and with less general error distributions than we permit here.

## 8. Appendix

### *8.1. Combinatorial and covering lemmas*

The following inequality appears without proof in Haussler (1991). It is very similar but not identical to Proposition 9.1.5 of Dudley (1984). Since we did not find a proof in the literature we include it for the sake of completeness.

**Lemma 6** *For all $n \geq 1$ and $1 \leq D \leq n$ one has*:

$$\sum_{j=0}^{D} \binom{n}{j} < \left(\frac{en}{D}\right)^D \quad .$$

*Proof*: Since the bound is larger than $2^n$ if $D \geq n/2$ we can assume that $x = D/n \in (0, 1/2)$. Let us denote by $\Sigma$ the sum to be bounded. Since $\Sigma = 2^n \mathbb{P}[N \leq D]$ where $N$ is a binomial random variable with parameter $1/2$, the Cramér-Chernoff inequality for the binomial implies that

$$\log \Sigma \leq n \log n - (n - D) \log(n - D) - D \log D$$

$$= D[\log(n/D) + (1 - x^{-1}) \log(1 - x)]$$

and it follows from elementary calculus that $(1 - x^{-1}) \log(1 - x) < 1$. □

**Lemma 7** *Let $S_{\mathscr{C}}$ be a finite set of densities with respect to $\mu$ indexed by $\mathscr{C} = \{0; 1\}^D$ and such that there exists a positive constant $\theta$ satisfying*

$$h^2(s_x, s_y) = \theta \sum_{i=1}^{D} \mathbb{1}_{x_i \neq y_i} \quad \textit{for all} \quad x, y \in \mathscr{C} \quad .$$

*Let $\hat{s}$ be any estimator with values in $S_{\mathscr{C}}$ based on $n$ independent identically distributed observations with density $s$. Then*

$$\sup_{s \in S_{\mathscr{C}}} \mathbb{E}_s \left[ h^2(s, \hat{s}) \right] \geq \frac{D\theta}{2} \left[ 1 - \sqrt{2n\theta} \right] \quad .$$

The proof is given in Birgé (1986) following the original treatment of Assouad (1983).

The following lemma is similar to what is usually called the Varshamov-Gilbert bound in information theory (see Gallager 1968).

**Lemma 8** *Let $\mathscr{C}$ be a subset of cardinality $\theta 2^D$ of the cube $\{0; 1\}^D$ with $0 < \theta \leq 1$. For any $\eta \in (0, 1)$ one can find a subset $\mathscr{C}'$ of $\mathscr{C}$ with cardinality larger than $\theta \exp(D\eta^2/2)$ such that for any two distinct points $x, \ y \in \mathscr{C}'$*

$$\sum_{i=1}^{D} \mathbb{1}_{x_i \neq y_i} > D \frac{1 - \eta}{2} \quad .$$

*Proof*: Let $D(1 - \eta)/2 = d$ and $\mathscr{C}'$ be a maximal subset of $\mathscr{C}$ such that $\sum_{i=1}^{D} \mathbb{1}_{x_i \neq y_i} > d$ for any pair $x, y \in \mathscr{C}'$. For each $x \in \mathscr{C}'$, the number of points $z \in \mathscr{C}$ such that $\sum_{i=1}^{D} \mathbb{1}_{x_i \neq z_i} \leq k$ is bounded by $\sum_{j=0}^{k} \binom{D}{j}$. It then follows from a covering argument that $|\mathscr{C}'| \sum_{j=0}^{[d]} \binom{D}{j} \geq \theta 2^D$. Let $B_D$ be a binomial random variable with parameters $D$ and $1/2$, then

$$|\mathscr{C}'| \geq \theta \left( \mathbb{P} \left[ B_D \leq D \frac{1 - \eta}{2} \right] \right)^{-1}$$

and the result follows from Hoeffding's inequality. □

**Lemma 9** *Let $\{\varphi_\lambda\}_{\lambda \in \Lambda}$ be a finite orthonormal system in $\mathbb{L}_2 \cap \mathbb{L}_\infty(\mu)$ with $|\Lambda| = D$ and $\bar{S}$ be the linear span of $\{\varphi_\lambda\}$. Let*

$$\bar{r} = \frac{1}{\sqrt{D}} \sup_{\beta \neq 0} \frac{\| \sum_{\lambda \in \Lambda} \beta_\lambda \varphi_\lambda \|_\infty}{|\beta|_\infty} \quad .$$

*For any positive $\delta$ one can find a countable set $T \subset \bar{S}$ and a mapping $\pi$ from $\bar{S}$ to $T$ with the following properties*:

- *for any ball $\mathscr{B}$ with radius $\sigma \geq 5\delta$*

$$|T \cap \mathscr{B}| \leq (B'\sigma/\delta)^D \quad \text{with} \quad B' < 5 \ . \tag{8.1}$$

- $\|u - \pi u\| \leq \delta$ *for all* $u$ *in* $\bar{S}$ *and*

$$\sup_{u \in \pi^{-1}(t)} \|u - t\|_\infty \leq \bar{r}\delta \quad \textit{for all } t \textit{ in } T \ . \tag{8.2}$$

*Proof*: Using the natural isometry between $\bar{S}$ and the Euclidean space $\mathbb{R}^D$ corresponding to the basis $\{\varphi_\lambda\}$ one defines $T$ as the image of $\tilde{T} = [(2\delta/\sqrt{D})\mathbb{Z}]^D$. Considering the partition of $\mathbb{R}^D$ into cubes of vertices with length $2\delta/\sqrt{D}$ centered on the points of $\tilde{T}$ we define the mapping $\tilde{\pi}$ from $\mathbb{R}^D$ onto $\tilde{T}$ such that $\tilde{\pi}(u)$ and $u$ belong to the same cube. Then $\pi$ is the image of $\tilde{\pi}$ by the natural isometry and clearly $\|u - \pi u\| \leq \delta$. The definition of $\bar{r}$ implies (8.2). It follows from Lemma 2 of Birgé and Massart (1998) that (8.1) holds with $B' = 1.1\sqrt{2\pi e}$. □

**Lemma 10** *In* $\mathbb{R}^D$, *the number of disjoint cubes of vertices* $\varepsilon/D$ *that intersect an* $l^1$*-ball of radius* $R$ *is bounded by* $[2e(R/\varepsilon + 1)]^D$.

*Proof*: An elementary computation (see Lemma 4.16 of Pisier 1989) shows that the volume of a $D$-dimensional $l^1$-ball of radius $\rho$ is equal to $2^D\rho^D/(D!)$. Since all the cubes of vertices $\varepsilon/D$ that intersect an $l^1$-ball of radius $R$ are included in an $l^1$-ball of radius $R + \varepsilon$, the required number is bounded by $(2D)^D(R/\varepsilon + 1)^D/D!$ and the result follows easily. □

**Lemma 11** *Let us consider a* $D$*-dimensional linear subspace* $\bar{V}$ *of* $\mathbb{L}_1(\mu)$. *We assume that there exists some basis* $(\varphi_\lambda)_{\lambda\in\Lambda}$ *of* $\bar{V}$ *with* $\|\varphi_\lambda\|_1 = 1$ *for all* $\lambda \in \Lambda$ *and some constant* $B'' \geq 1$ *such that*

$$\sum_{\lambda\in\Lambda} |\beta_\lambda| \leq B'' \left\| \sum_{\lambda\in\Lambda} \beta_\lambda\varphi_\lambda \right\|_1 \quad \textit{for all } (\beta_\lambda) \in \mathbb{R}^\Lambda \ .$$

*Given* $\varepsilon$ *and* $R$ *with* $0 < \varepsilon \leq R/2$, *any ball* $\mathscr{B} \in \bar{V}$ *of radius* $R$ *may be covered by* $N$ *intervals* $[f^-, f^+] \subset \mathbb{L}_1(\mu)$ *with diameter* $\|f^+ - f^-\|_1 \leq \varepsilon$ *and* $N \leq (3eB'')^D(R/\varepsilon)^D$.

*Proof*: Without loss of generality we take $\Lambda = \{1; \ldots; D\}$ and consider some ball $\mathscr{B}$ in $\bar{V}$ centered at the origin and defined by

$$\mathscr{B} = \left\{ \sum_{\lambda\in\Lambda} \beta_\lambda\varphi_\lambda \ \middle| \ \left\| \sum_{\lambda\in\Lambda} \beta_\lambda\varphi_\lambda \right\|_1 \leq R \right\} \ .$$

Using the standard linear isomorphism between $\bar{V}$ and $\mathbb{R}^D$ we may identify $(\beta_\lambda)_{\lambda\in\Lambda}$ with $\sum_{\lambda\in\Lambda} \beta_\lambda\varphi_\lambda$. Since the coefficients $\beta_\lambda$ of any point in $\mathscr{B}$ satisfy

$$\sum_{\lambda\in\Lambda}|\beta_\lambda| \le B'' \left\| \sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda \right\|_1 \le B''R$$

$\mathscr{B}$ can be identified to a subset $\mathscr{B}'$ of the $l_1$-ball with radius $RB''$ centered at the origin of $\mathbb{R}^D$. By Lemma 10 we can cover this ball by $N$ cubes with vertices of length $\varepsilon/D$ with $N \le (3eB''R\varepsilon^{-1})^D$. Let $\mathscr{C}$ be the set of the $N$ centers of these cubes. For each $c = (\beta_\lambda)_{\lambda\in\Lambda} \in \mathscr{C}$ we consider the interval

$$I_c = \left[ \sum_{\lambda\in\Lambda} \left( \beta_\lambda\varphi_\lambda - \frac{\varepsilon}{2D}|\varphi_\lambda| \right), \sum_{\lambda\in\Lambda} \left( \beta_\lambda\varphi_\lambda + \frac{\varepsilon}{2D}|\varphi_\lambda| \right) \right] .$$

For any $(\alpha_\lambda)_{\lambda\in\Lambda} \in \mathscr{B}'$ there exists some $c = (\beta_\lambda)_{\lambda\in\Lambda} \in \mathscr{C}$ such that $|\alpha_\lambda - \beta_\lambda| \le \varepsilon/(2D)$ for all $\lambda \in \Lambda$. It follows that

$$\left| \sum_{\lambda\in\Lambda}\alpha_\lambda\varphi_\lambda(x) - \sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda(x) \right| \le \sum_{\lambda\in\Lambda}|\alpha_\lambda - \beta_\lambda|\,|\varphi_\lambda(x)| \le \frac{\varepsilon}{2D}\sum_{\lambda\in\Lambda}|\varphi_\lambda(x)| .$$

Therefore $\sum_{\lambda\in\Lambda}\alpha_\lambda\varphi_\lambda \in I_c$ and the intervals $(I_c)_{c\in\mathscr{C}}$ cover $\mathscr{B}$. Moreover the $\mathbb{L}_1$-diameter of each $I_c$ is bounded by $\varepsilon D^{-1}\sum_{\lambda\in\Lambda}\|\varphi_\lambda\|_1 = \varepsilon$. □

### *8.2. Some results in approximation theory*

**Proposition 8** *Let $s$ be a function of bounded $\alpha$-variation on $[0, 1]$ with $0 < \alpha \le 1$ which means that*

$$\sup_{k\ge 2}\ \sup_{x_1<\cdots<x_k}\ \sum_{j=2}^{k}|s(x_j) - s(x_{j-1})|^{1/\alpha} = J_\alpha(s) < +\infty$$

*where the supremum is taken over all increasing sequences $x_1 < \cdots < x_k$ of points in $[0, 1]$. Let $L$ be any number between 1 and some positive integer $N$. There exists a partition $\mathscr{P}$ of $[0, 1]$ into $D$ intervals with endpoints belonging to the grid $\{i/N \mid 0 \le i \le N\}$ and a function $s^+ \ge s$ which is constant on the elements of $\mathscr{P}$ such that*

$$D \le 2\,(N/L)^{1/(1+2\alpha)} + 1$$

*and*

$$\|s^+ - s\|^2 \le J_\alpha^{2\alpha}(s)\left[2\,(L/N)^{(2\alpha)/(2\alpha+1)} + L/N\right] . \qquad (8.3)$$

*Proof*: Let $J = J_\alpha(s)$ and for any interval $I$ define $J(I) = J^{-1}[\sup_{y\in I} s(y) - \inf_{y\in I} s(y)]^{1/\alpha}$. Consider a partition $\mathscr{P}$ of $[0, 1]$ into $D$ intervals $I_1, \ldots, I_D$. If $s^+(x) = \sup_{y\in I_j} s(y)$ for $x \in I_j$, one gets

$$\|s^+ - s\|^2 = \sum_{j=1}^{D} \int_{I_j} [s^+(x) - s(x)]^2\, dx \le \sum_{j=1}^{D} |I_j| [J J(I_j)]^{2\alpha} \quad . \tag{8.4}$$

Let us now build by induction a partition $\mathscr{P}$ from an increasing sequence $x_0 = 0 < \cdots < x_D = 1$ in the following way. Starting with $x_0 = 0$, define

$$x_{j+1} = N^{-1} \sup \left\{ i \le N \;\middle|\; L \ge (i - N x_j) J^{2\alpha}([x_j, i/N)) \right\}$$

and stop the process when $x_{j+1} = 1$. Then $I_j = [x_{j-1}, x_j)$ is always nonvoid since $L \ge 1$ and $J(I) \le 1$ for all $I$. By construction $|I_j| J^{2\alpha}(I_j) \le L/N$ for $1 \le j \le D$ and, if we set $I_j^+ = [x_{j-1}, x_j + 1/N)$, then for $j < D$, $|I_j^+| J^{2\alpha}(I_j^+) > L/N$. Moreover $\sum_{j=1}^{D-1} |I_j^+| \le 2$ and by the definition of $J$, $\sum_{j=1}^{D-1} J(I_j^+) \le 2$. It then follows from Lemma 2.2 of Birman and Solomjak (1967) that $2^{-(2\alpha+1)} L/N \le (D-1)^{-(2\alpha+1)}$. Therefore it follows from (8.4) that

$$\|s^+ - s\|^2 \le D J^{2\alpha} \frac{L}{N} \le J^{2\alpha} \frac{L}{N} \left[ 2 \left( \frac{N}{L} \right)^{1/(2\alpha+1)} + 1 \right] \quad . \qquad \square$$

**Corollary 1** *Let $s$ be a function of bounded $\alpha$-variation on $[0, 1]$ with $0 < \alpha \le 1$, $N$ a positive integer and $D$ an integer such that $3 \le D \le 2N^{1/(1+2\alpha)} + 1$. Then one can find a partition $\mathscr{P}$ of $[0, 1]$ into $D$ intervals with endpoints belonging to the grid $\{i/N \mid 0 \le i \le N\}$ and a function $s^+ \ge s$ which is constant on the elements of $\mathscr{P}$ such that*:

$$\|s^+ - s\|^2 \le 3 \left( \frac{2D}{D-1} \right)^{2\alpha} J_\alpha^{2\alpha}(s) D^{-2\alpha} \le 27 J_\alpha^{2\alpha}(s) D^{-2\alpha} \quad .$$

*Proof*: Let $L$ be defined by $D = 2(N/L)^{1/(2\alpha+1)} + 1$. Then $1 \le L \le N$ and the preceding proposition applies showing that one can find a piecewise constant function $s^+$ based on a partition with $D' \le D$ intervals and such that (8.3) holds. Clearly $s^+$ can also be viewed as a piecewise constant function based on a partition with $D$ intervals. Moreover

$$\|s^+ - s\|^2 \le 3 J_\alpha^{2\alpha}(s) \left( \frac{L}{N} \right)^{(2\alpha)/(2\alpha+1)} = 3 J_\alpha^{2\alpha}(s) \left( \frac{2}{D-1} \right)^{2\alpha}$$

and the result follows since $2D/(2D-1) \le 3$. $\square$

**Lemma 12** *Let $\mathscr{A}$ be either the interval $[0, 1]$ or the one-dimensional torus $\mathbb{T}$. Let $s$ belong to the Besov space $B_{\alpha\, p\, \infty}(\mathscr{A})$ with $\alpha > 0,\ 1 \le p \le \infty$ and let $D$ be a positive integer.*

- *If $\mathscr{A} = [0, 1]$ and $r \in \mathbb{N} > \alpha - 1$, let $S_1$ be the space of piecewise polynomials of degree bounded by $r$ based on the regular partition with $D$ pieces;*
- *if $\mathscr{A} = \mathbb{T}$, let $S_2$ be the space of trigonometric polynomials on $\mathbb{T}$ with degree $\le D$;*
- *if $\mathscr{A} = [0, 1]$, $s$ has a compact support in $(0, 1)$ and $r \in \mathbb{N} > \alpha - 1$, let $S_3$ be the linear span of the set $\{\varphi_\lambda \mid \lambda \in \cup_0^J \Lambda(j)\}$ and $D = 2^J$ where $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ is a wavelet basis of regularity $r$.*

  *Then, there exists positive constants $C_i(\alpha)$ such that*

$$d_p(s, S_i) \le C_i(\alpha, p)|s|_{\alpha\, p} D^{-\alpha} \quad \text{for } i = 1, 2, 3$$

  *where $d_p$ denotes the $\mathbb{L}_p$-distance with respect to the uniform distribution on $\mathscr{A}$ and $|s|_{\alpha\, p}$ the semi-norm of $s$ in $B_{\alpha\, p\, \infty}(\mathscr{A})$.*

*Remark*: We recall that the Hölder space $\mathscr{H}_\alpha$ defined in Section 3.3.3 satisfies $\mathscr{H}_\alpha \subset B_{\alpha\, \infty\, \infty}([0, 1])$ with equality when $\alpha$ is not an integer.

*Proof*: We recall, following DeVore and Lorentz (1993), that a function $s$ belongs to the Besov space $B_{\alpha\, p\, \infty}(\mathscr{A})$ if its $r$-modulus of smoothness defined by $\omega_r(s, y)_p = \sup_{0<h\le y} \|\Delta_h^r(s, \cdot)\|_p$ where $\Delta_h^r(s, \cdot)$ denotes the $r$-th order differences given by

$$\Delta_h^r(s, x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} s(x + kh)$$

satisfies

$$\sup_{y>0} y^{-\alpha} \omega_r(s, y)_p = |s|_{\alpha\, p} < +\infty \quad \text{with } r = [\alpha] + 1 \ .$$

The required approximation properties are proved in DeVore and Lorentz (1993) page 359 for piecewise polynomials and page 205 for trigonometric polynomials; this gives the result for $i = 1$ or 2. If $i = 3$, it follows from Meyer (1990) Chapter 6, Section 10 that $s = \sum_{j\ge 0} \sum_{\lambda\in\Lambda(j)} \beta_\lambda \varphi_\lambda$ belongs to the Besov space $B_{\alpha\, p\, \infty}(\mathscr{A})$ if and only if

$$\sup_{j\ge 0} 2^{j(\alpha + \frac{1}{2} - \frac{1}{p})} \left( \sum_{\lambda\in\Lambda(j)} |\beta_\lambda|^p \right)^{1/p} = |\!|\!| s |\!|\!|_{\alpha\, p} < +\infty$$

with a semi-norm $\|\!|s|\!\|_{\alpha\, p}$ equivalent to $|s|_{\alpha\, p}$. Let $s_J$ be the orthogonal projection of $s$ onto $S_3$. It follows from Bernstein's inequality [see Meyer (1990) Chapter 2, Lemma 8] that

$$\|s - s_J\|_p^p \le C'_p \sum_{j>J} \sum_{\lambda\in\Lambda(j)} 2^{j(p/2-1)} |\beta_\lambda|^p \le C'_p \|\!|s|\!\|_{\alpha\, p}^p \sum_{j>J} 2^{-jp\alpha}$$

hence the result. □

Next we recall a simple approximation property of convex combinations of functions in $\mathbb{L}_2(\mu)$ which may be proved either by a random sampling or a greedy selection method (see Jones 1992 and Barron 1993). For convenience we restate it here with a slight modification obtained by application of the triangle inequality. Improvements in the approximation bound of the lemma are possible (see Makovoz 1996). These improvements become negligible in high dimensional settings so we shall stick with the simpler order $1/\sqrt{D}$ bounds here.

**Lemma 13** *Suppose $s$ and $t$ are given functions in $\mathbb{L}_2(\mu)$ with $t/R$ in the closure of the convex hull of a class of functions $\{\pm\phi_w\}$ in $\mathbb{L}_2(\mu)$ bounded by one, for some constant $R$ depending on $t$. Then there exists an approximation $s_D$ equal to $R$ times the convex combination of $D$ functions in the class such that*

$$\|s - s_D\| \le R/\sqrt{D} + \|s - t\| \ .$$

Several authors have recently put this lemma to use to prove approximation properties in some interesting contexts, especially in multivariate settings where it gives conditions for approximation at a dimension independent rate (see Jones 1992, Barron 1993, Breiman 1993, Girosi and Anzellotti 1992, Hornik et al. 1994, Yukich et al. 1995 for approximation based on Fourier analysis in the ridge function case and Girosi and Anzellotti 1992 for similar conclusions for approximation using radial basis functions). However, more is needed to ensure that accurate approximations can be achieved using a control $H$ on the parameters $w_j$ that is bounded by a polynomial in $D$ or $n$. Such a control is needed to prove Proposition 6.

*Proof of Proposition 6*: Our strategy is to give conditions for the existence of a function $s_H$ that is close to $s$ and such that for some $R(s)$ the function $s_H/R(s)$ is in $\bar{co}\{\pm\phi_w \mid |w|_1 \le H\}$ where $\bar{co}\{A\}$ denotes the closure of the convex hull of the set $A$. Then Lemma 13 with $t = s_H$ provides some $s_m = s_{(D',H,R(s))}$ in $\bar{S}_m$ with

$$\|s - s_m\| \le \|s - s_H\| + R(s)/\sqrt{D'} \ . \tag{8.5}$$

Let $\tilde{F}(da) = e^{ib_a} F(da)$ denote the phase and magnitude factorization of the complex-valued measure $\tilde{F}$ with phase $|b_a| \le \pi$. We recall that $s(x) = \int \exp\{ia^T x\}\tilde{F}(da)$, hence since $s$ is real valued

$$s(x) = \int \cos(a^T x + b_a) F(da) \quad . \tag{8.6}$$

For trigonometric approximation we assume that $H \ge 2\pi$ and consider $s_H(x) = \int \cos(a^T x + b_a) 1_{\{|a|_1 \le H/2\}} F(da)$ for which the error is bounded by $|s(x) - s_H(x)| \le \int 1_{\{|a|_1 \ge H/2\}} F(da) \le c_{s,\alpha}(2/H)^\alpha$ by Markov's inequality. We recognize $s_H$ as an element of $c\bar{o}\{\cos(a^T x + b) \mid |a|_1 \le H/2, |b| \le H/2\}$ multiplied by a constant not greater than $c_{s,0}$, so that we get from (8.5) $d(s, S_m) \le c_{s,\alpha}(2/H)^\alpha + c_{s,0}/\sqrt{D'}$ with $m = (D', H, c_{s,0})$ and (4.40) holds.

In the neural net case the Fourier components are related to convex combinations of sigmoids as shown in Barron (1993). The approximation bounds stated in the proposition are given there.

In the case of ridge wavelets the assumption is that $\|\psi\|_\infty = \Psi < +\infty$ and $\psi$ is zero outside a finite interval. For simplicity we take here the interval to be $[-1, 1]$. We use an integral representation in Hornik et al. (1994) and Yukich et al. (1995) to show that we may control $H$. First we pick any scalar value $h$ for which $\tilde{\psi}(h) = \int_{-1}^{1} e^{-ihz}\psi(z)\, dz$ is nonzero. Multiplying and dividing by $\tilde{\psi}(h)$ in the definition of $s$ and making a change in variables we get the integral representation

$$s(x) = \frac{1}{\tilde{\psi}(h)} \int_{a \in \mathbb{R}^q} \tilde{F}(h\, da) \int_{|b + a^T x| \le 1} \psi(a^T x + b) e^{-ihb}\, db \quad .$$

Here we assume that $H \ge 4$ and let $s_H(x)$ be the real part of the same quantity with integration with respect to the vector $a$ restricted to $|a|_1 \le H' = H/2 - 1$. Since $|a^T x| \le |a|_1$ the value of $|b|$ in the integral is bounded by $\le H/2$ and $|a|_1 + |b| \le H$. By assumption $H' \ge 1$ and the error $|s(x) - s_H(x)|$ is bounded by

$$\begin{aligned}
|s(x) - s_H(x)| &= \left| \frac{1}{\tilde{\psi}(h)} \int_{|a|_1 > H'} \int_{|b + a^T x| \le 1} \psi(a^T x + b) e^{-ihb} \tilde{F}(h\, da)\, db \right| \\
&\le \frac{2\Psi}{|\tilde{\psi}(h)|} \int_{|a|_1 > H'} (1 + |a|_1) F(h\, da) \\
&\le \frac{4\Psi}{|\tilde{\psi}(h)|} \int_{|a|_1 > H'} |a|_1 F(h\, da) \\
&\le \frac{4\Psi}{|h\tilde{\psi}(h)|} \int_{|a'|_1 > |h| H'} |a'|_1 F(da') \le \frac{4\Psi c_{s,\alpha}}{|h|^\alpha |\tilde{\psi}(h)| H'^{\alpha - 1}}
\end{aligned}$$

by the change of variable $a' = ha$ and Markov's inequality since $\alpha > 1$. In a similar manner we see that $s_H(x)$ is in $\bar{co}\{\psi(a^T x + b) \mid |a|_1 + |b| \le H\}$ multiplied by a constant not greater than $[2/\tilde{\psi}(h)][c_{s,0} + c_{s,1}/|h|]$. It then follows from (8.5) that $d(s, S_m) \le C_\psi[c_{s,\alpha}/H^{\alpha-1} + (c_{s,0} + c_{s,1})/\sqrt{D'}]$ and then (4.40) holds.

For the hinged hyperplanes of Breiman (1993) approximation bounds similar to what we need are in his paper. Here we give an integral representation that makes explicit that the approximation bound holds with $H$ as small as 2. We use Taylor's theorem with remainder to characterize each Fourier component $\cos(a^T x + b_a)$. Recalling that $|a^T x| \le |a|_1$ we have

$$\begin{aligned}
\cos(b + a^T x) &= \cos(b) - a^T x \sin(b) - \int_0^{a^T x} \cos(t + b)(a^T x - t)\, dt \\
&= \cos(b) - a^T x \sin(b) - \int_0^{|a|_1} \cos(t + b)[(a^T x - t) \vee 0]\, dt \\
&\quad - \int_{-|a|_1}^{0} \cos(t + b)[(t - a^T x) \vee 0]\, dt \\
&= \cos(b) - a^T x \sin(b) \\
&\quad - |a|_1^2 \int_0^1 \cos(|a|_1 u + b) \left[ \left( \frac{a^T x}{|a|_1} - u \right) \vee 0 \right] du \\
&\quad - |a|_1^2 \int_{-1}^{0} \cos(|a|_1 u + b) \left[ \left( u - \frac{a^T x}{|a|_1} \right) \vee 0 \right] du
\end{aligned}$$

where we have written separately the contributions from $t$ positive and $t$ negative and then changed variables from $t$ to $u = t/|a|_1$. Note that the functions in the last two integrals are in the closure of the convex hull of the functions of plus or minus bounded multiples of hinge functions. Integrating over the frequency vector $a$ according to $F(da)$ with $b = b_a$ equal to the phase, we get from (8.6) that $s(x)$ is equal to $s(0) + (\nabla s(0))^T x$ plus a function in $\bar{co}\{\pm[(\bar{a}^T x + \bar{b}) \vee 0] \mid |\bar{a}|_1 \le 1, |\bar{b}| \le 1\}$ times a constant which is not greater than $2c_{s,2}$. We note that trivially the constant 1 is a particular hinged hyperplane and that

$$a^T x = 2\left[ \frac{1}{2}\left(a^T x \vee 0\right) - \frac{1}{2}\left[(-a^T x) \vee 0\right] \right] .$$

It follows that $s/R(s)$ is in $\bar{co}\{\pm[(\bar{a}^T x + \bar{b}) \vee 0] \mid |\bar{a}|_1 \le 1, |\bar{b}| \le 1\}$ for $R(s) \le |s(0)| + 2\|\nabla s(0)\|_1 + 2c_{s,2}$. The approximation bound (4.40) then follows from (8.5) for all $H \ge 2$ with $\delta_H = 0$. This completes the proof of Proposition 6. □

### 8.3. Further technical results

We first give a proof for Proposition 1. It derives easily from the following

**Lemma 14** *Let $s^2$ be a probability density with respect to $\mu$ and $t$ be a function in $\mathbb{L}_2(\mu)$. Then if $t' = t/\|t\|$*

$$\|s - t'\| \leq \|s - t\| + (1 - \|t\|) \vee 0 \leq 2\|s - t\| \ ; \tag{8.7}$$

*if $t^2$ is a density then*

$$d^2(s,t) \leq K(s,t) \leq 2[1 + \log(\|s/t\|_\infty)]d^2(s,t) \ , \tag{8.8}$$

*consequently if $s^+$ is such that $s^+ \geq s$, $s' = s^+/\|s^+\|$ and $\varepsilon = \|s - s^+\|$ then*

$$K(s,s') \leq 2[1 + \log(1+\varepsilon)]\varepsilon^2 \ . \tag{8.9}$$

*Proof*: If $\|t\| \geq 1$, then

$$\|s - t\|^2 - \|s - t'\|^2 = (\|t\| - 1)\left(\|t\| + 1 - 2\int st/\|t\|\right)$$

and Cauchy-Schwarz inequality yields $\|s - t'\| \leq \|s - t\|$. If $\|t\| < 1$, then

$$\|s - t'\| \leq \|s - t\| + \|t' - t\| = \|s - t\| + \|t\|(1/\|t\| - 1)$$

which gives (8.7). Inequalities (8.8) follow from (7.6) of Lemma 5 of Birgé and Massart (1998) since $d/\sqrt{2}$ is the Hellinger distance. Noticing that $\|s^+\| \geq 1$ and $\|s/s'\|_\infty \leq \|s^+\| \leq 1 + \varepsilon$, one concludes that (8.7) and (8.8) imply (8.9). □

*Proof of Proposition* 1: The first inequality is an immediate consequence of (8.9), considering separately the cases $\varepsilon < 0.6$ and $\varepsilon \geq 0.6$. In order to derive (3.15) one notices that if $\tilde{s}$ is such that $\|\tilde{s} - s\|_\infty = \varepsilon$ and $\mu$ is a probability one can define $s^+ = (\tilde{s} + \varepsilon) \geq s$ and apply the preceding recipe since $\|s^+ - s\| \leq 2\varepsilon$. □

The next lemma is elementary but very useful to deal with ellipsoids:

**Lemma 15** *Let $(a_j)_{j\geq 0}$ and $(b_j)_{j\geq 0}$ two sequences of numbers in $[0, +\infty]$ which are respectively nonincreasing and nondecreasing and satisfy $a_j < b_j$ for $j$ large enough. Then, defining $m = \inf\{j \geq 0 \,|\, a_{j+1} \leq b_j\} < +\infty$ and $\theta = b_0/a_0$ (with the convention that $\theta = 1$ if $a_0 = b_0 = 0$ or $+\infty$), one gets $\sup_j\{a_j \wedge b_j\} = a_m \wedge b_m$ and*

$$\sup_{j\geq 0}\{a_j \wedge b_j\} \leq \inf_{j\geq 0}\{a_{j+1} + b_j\} \leq \inf_{0\leq j\leq m}\{a_{j+1} + b_j\} \leq 2(1\vee\theta)\sup_{j\geq 0}\{a_j \wedge b_j\} \ .$$

*Proof*: Notice first that when $0 \leq j < m$ one has $a_j \geq a_m > b_{m-1} \geq b_j$ which implies that $a_j \wedge b_j \leq a_m \wedge b_m$ and that a similar result holds for $j > m$. Considering separately the cases $j \leq k$ and $j \geq k+1$ one checks that $a_j \wedge b_j \leq a_{k+1} + b_k$ and the left-hand side inequality follows. If $m \geq 1$, $a_{m+1} + b_m \leq 2b_m$ and $a_m + b_{m-1} < 2a_m$, therefore $\inf_{0 \leq j \leq m}\{a_{j+1} + b_j\} \leq 2(a_m \wedge b_m)$. If $m = 0$ one gets $a_1 + b_0 \leq 2b_0 = 2(\theta \vee 1)(a_0 \wedge b_0)$ and the result follows in both cases. □

*Acknowledgements.* We would like to thank the Editors of P.T.R.F. for inviting us to present this paper in their journal and three diligent referees for their very useful suggestions and comments.

## References

Akaike, H.: Information theory and an extension of the maximum likelihood principle. In Proceedings 2nd International Symposium on Information Theory, P.N. Petrov and F. Csaki (Eds.). Akademia Kiado, Budapest, 267–281 (1973)

Assouad, P.: Deux remarques sur l'estimation. C. R. Acad. Sc. Paris Sér. I Math. **296**, 1021–1024 (1983)

Baraud, Y.: Model selection for regression on a fixed design. Technical Report #97.49, Université Paris-Sud (1997)

Barron, A.R.: Complexity regularization with applications to artificial neural networks. In Nonparametric Functional Estimation (G. Roussas, ed.). Kluwer, Dordrecht, 561–576 (1991)

Barron, A.R.: Universal approximation bounds for superpositions of a sigmoidal function. IEEE Transactions on Information Theory **39**, 930–945 (1993)

Barron, A.R.: Approximation and estimation bounds for artificial neural networks. Machine Learning. **14**, 115–133 (1994)

Barron, A.R., Cover, T.M.: Minimum complexity density estimation. IEEE Transactions on Information Theory **37**, 1034–1054 (1991)

Barron, A.R., Sheu C.-H.: Approximation of density functions by sequences of exponential families. Ann. Statist. **19**, 1347–1369 (1991)

Berger, M., Gauduchon, P. Mazet, E.: Le Spectre d'une Variété Riemannienne. Lecture Notes in Mathematics 194. Springer, Berlin (1971)

Birgé, L.: Approximation dans les espaces métriques et théorie de l'estimation. Z. Wahrscheinlichkeitstheorie Verw. Geb. **65**, 181–237 (1983)

Birgé, L.: On estimating a density using Hellinger distance and some other strange facts. Probab. Th. Rel. Fields **71** , 271–291 (1986)

Birgé, L.: The Grenander estimator: a non asymptotic approach. Ann. Statist. **17**, 1532–1549 (1989)

Birgé, L., Massart, P.: Rates of convergence for minimum contrast estimators. Probab. Theory Relat. Fields **97**, 113–150 (1993)

Birgé, L., Massart, P.: Minimum contrast estimators on sieves. Technical Report #94.34, Université Paris-Sud (1994)

Birgé, L., Massart, P.: From model selection to adaptive estimation. In Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics (D. Pollard, E. Torgersen and G. Yang, eds.), 55–87. Springer-Verlag, New York (1997)

Birgé, L., Massart, P.: Minimum contrast estimators on sieves: exponential bounds and rates of convergence. Bernoulli **4**, 329–375 (1998)
Birman, M.S., Solomjak, M.Z.: Piecewise-polynomial approximation of functions of the classes $W_p$. Mat. Sbornik **73**, 295–317 (1967)
Breiman, L.: Hinging hyperplanes for regression, classification, and function approximation. IEEE Transactions on Information Theory **39**, 999–1013 (1993)
Bretagnolle, J., Huber, C.: Estimation des densités: risque minimax. Z. Wahrscheinlichkeitstheorie verw. Gebiete **47**, 119–137 (1979)
Cencov, N.N.: Statistical Decision Rules and Optimal Inference. Translations of Mathematical Monographs **53**, American Math. Society, Providence (1982)
Chavel, I.: Eigenvalues in Riemannian Geometry. Academic Press, Orlando (1984)
Chow, Y.-S., Grenander, U.: A sieve method for the spectral density Ann. Statist. **13**, 998–1010 (1985)
Cirel'son, B.S., Ibragimov, I.A., Sudakov, V.N.: Norm of gaussian sample function. In Proceedings of the 3rd Japan-U.S.S.R. Symposium on Probability Theory, Springer Lecture Notes in Mathematics 550 pp. 20–41. Springer-Verlag, Berlin (1976)
Cox, D.D.: Approximation of least squares regression on nested subspaces. Ann. Statist. **16**, 713–732 (1988)
Dahmen, W., DeVore, R.A., Scherer, K.: Multidimensional spline approximation. SIAM J. Numer. Anal. **17**, 380–402 (1980)
Daniel, C., Wood, F.S.: Fitting Equations to Data. Wiley, New York (1971)
Daubechies, I.: Ten Lectures on Wavelets. S.I.A.M., Philadelphia (1992)
DeVore, R.A., Lorentz, G.G.: Constructive Approximation. Springer-Verlag, Berlin (1993)
Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**, 425–455 (1994a)
Donoho, D.L., Johnstone, I.M.: Ideal denoising in an orthonormal basis chosen from a library of bases. C. R. Acad. Sc. Paris Sér. I Math. **319**, 1317–1322 (1994b)
Donoho, D.L., Johnstone, I.M.: Minimax risk over $l_p$-balls for $l_q$-error. Probab. Theory Relat. Fields **99**, 277–303 (1994c)
Donoho, D.L., Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. JASA. **90**, 1200–1224 (1995)
Donoho, D.L., Johnstone, I.M.: Minimax estimation via wavelet shrinkage. Ann. Statist. **26**, 879–921 (1998)
Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Wavelet shrinkage: Asymptopia? J. R. Statist. Soc. B **57**, 301–369 (1995)
Donoho, D.L., Johnstone, I.M., Kerkyacharian, G., Picard, D.: Density estimation by wavelet thresholding. Ann. Statist. **24**, 508–539 (1996)
Dudley, R.M.: Central limit theorems for empirical measures. Ann. Probab. **6** 899–929 (1978)
Dudley, R.M.: A course on empirical processes. In Ecole d'Eté de Probabilités de Saint-Flour XII - 1982. Lecture Notes in Mathematics 1097, Springer, Berlin (1984)
Efroimovich, S.Yu.: Nonparametric estimation of a density of unknown smoothness. Theory Probab. Appl. **30**, 557–568 (1985)
Efroimovich, S.Yu., Pinsker, M.S.: Estimation of square-integrable density on the basis of a sequence of observations. Probl. Inf. Transm. **17**, 182–196 (1981)
Efroimovich, S.Yu., Pinsker, M.S.: Estimation of square-integrable probability density of a random variable. Probl. Inf. Transm. **18**, 175–189 (1982)
Efroimovich, S.Yu., Pinsker, M.S.: Learning algorithm for nonparametric filtering. Automat. Remote Control **11**, 1434–1440, translated from Avtomatika i Telemekhanika **11**, 58–65 (1984)

Efroimovich, S.Yu., Pinsker, M.S.: Self-tuning algorithm for minimax nonparametric estimation of spectral density. Probl. Inf. Transm. **22**, 209–221 (1986)
Friedman, J.: Multivariate adaptive regression splines (with discussion). Ann. Statist. **19**, 1–141 (1991)
Gallager, R.G.: Information Theory and Reliable Communication. Wiley, New York (1968)
Girosi, F., Anzellotti, G.: Convergence rates of approximation by translates. Artificial Intelligence Lab Technical Report 1288, Massachusetts Institute of Technology (1992)
Goldenshluger, A. and Nemirovski, A.: On spatially adaptive estimation of nonparametric regression. Math. Meth. Stat. **6**, 135–170 (1997)
Golubev, G.K.: Quasi-linear estimates of signals in $\mathbb{L}_2$. Probl. Inf. Transm. **26**, 15–20 (1990)
Golubev, G.K.: Nonparametric estimation of smooth probability densities in $\mathbb{L}_2$. Probl. Inf. Transm. **28**, 44–54 (1992)
Golubev, G.K., Nussbaum, M.: Adaptive spline estimates for nonparametric regression models. Theory Probab. Appl. **37**, 521–529 (1992)
Grenander, U.: Abstract inference. Wiley, New York (1981)
Hall, P.: Large-sample optimality of least-squares cross-validation in density estimation. Ann. Statist. **11**, 1156–1174 (1983)
Hall, P.: Cross-validation and the smoothing of orthogonal series density estimators. J. Mult. Analysis **21**, 207–237 (1987)
Haussler, D.: Sphere packing numbers for subsets of the Boolean $n$-cube with bounded Vapnik-Chervonenkis dimension. J. Combinatorial Theory A **69**, 217–232 (1991)
Haussler, D.: Decision theoretic generalizations of the PAC model for neural net and other learning applications. Information and Computation **100**, 78–150 (1992)
Hornik, K., Stinchcombe, M. B., White, H., Auer, P.: Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. Neural Computation **6**, 1262–1275 (1994)
Huber, P.J.: The behavior of maximum likelihood estimates under non-standard conditions. Proc. 5th Berkeley Symp. Math. Statist. Prob. Vol. 1, 221–233 (1967)
Ibragimov, I.A., Has'minskii, R.Z.: On estimate of the density function. Zap. Nauchn. Semin. LOMI **98**, 61–85 (1980)
Ibragimov, I.A., Has'minskii, R.Z.: On the non-parametric density estimates. Zap. Nauchn. Semin. LOMI **108**, 73–89 (1981)
Johnstone, I., Kerkyacharian, G., Picard, D.: Estimation d'une densité de probabilité par méthode d'ondelettes. C. R. Acad. Sc. Paris Sér. I Math. **315**, 211–216 (1992)
Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. Ann. Statist. **20**, 608–613 (1992)
Juditsky, A.: Wavelet estimators: adapting to unknown smoothness. Math. Methods of Statist **6**, 1–25 (1997)
Kahane, J.-P.: Some random series of functions, 2nd ed. Cambridge University Press, Cambridge (1985)
Korostelev, A.P., Tsybakov, A.B.: Estimation of the density support and its functionals. Probl. Inf. Transm. **29**, 1–15 (1993a)
Korostelev, A.P., Tsybakov, A.B.: Minimax Theory of Image Reconstruction. Lecture Notes in Statistics 82, Springer-Verlag, New York (1993b)
Le Cam, L.M.: Convergence of estimates under dimensionality restrictions. Ann. Statist. **1**, 38–53 (1973)
Le Cam, L.M.: Asymptotic Methods in Statistical Decision Theory. Springer, New York (1986)

Le Cam, L.M., Yang, G.L.: Asymptotics in Statistics: Some Basic Concepts. Springer, New York (1990)

Ledoux, M.: Isoperimetry and Gaussian analysis. In Lectures on Probability Theory and Statistics, Ecole d'Eté de Probabilités de Saint-Flour XXIV-1994 (P. Bernard, ed.), 165–294. Springer, Berlin (1996)

Lepskii, O.V.: Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. Theory Probab. Appl. **36**, 682–697 (1991)

Lepskii, O.V.: Asymptotically minimax adaptive estimation II: Statistical model without optimal adaptation. Adaptive estimators. Theory Probab. Appl. **37**, 433–468 (1992)

Lepskii, O.V., Mammen, E., Spokoiny, V.G.: Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. Ann. Statist. **25** , 929–947 (1997)

Lepskii, O.V., Spokoiny, V.G.: Local adaptation to inhomogeneous smoothness: resolution level. Math. Methods Statist. **4**, 239–258 (1995)

Li, K.C.: Asymptotic optimality for $C_p$, $C_L$, cross-validation, and generalized cross-validation: Discrete index set. Ann. Statist. **15**, 958–975 (1987)

McGaffrey, D.F., Gallant, A.R.: (1994). Convergence rates for single hidden layer feedforward networks. Neural Networks **7**, 147–158 (1987)

Makovoz, Y.: Random approximants and neural networks. J. Approx. Th. **85**, 98–109 (1996)

Mallows, C.L.: Some comments on $C_p$. Technometrics **15**, 661–675 (1973)

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

Modha, D.S., Masry, E.: Minimum complexity regression estimation with weakly dependent observations. IEEE Transactions on Information Theory **42**, 2133–2145 (1996)

Nemirovskii, A.S.: Nonparametric estimation of smooth regression function. Izv. Akad. Nauk. SSSR Tekhn. Kibernet. **3**. 50–60 (1985) (in Russian); Soviet. J. Comput. Systems Sci. **23**, 1–11 (1986) (in English)

Pisier, G.: The Volume of Convex Bodies and Banach Space Geometry. Cambridge University Press, Cambridge (1989)

Pollard, D.: New ways to prove central limit theorems. Econometric Theory **1**, 295–314 (1985)

Polyak, B.T., Tsybakov, A.B.: Asymptotic optimality of the $C_p$-test for the orthogonal series estimation of regression. Theory Probab. Appl. **35**, 293–306 (1990)

Polyak, B.T., Tsybakov, A.B.: A family of asymptotically optimal methods for choosing the estimate order in orthogonal series regression. Theory Probab. Appl. **37**, 471–481 (1993)

Rissanen, J.: Modeling by shortest data description. Automatica **14**, 465–471 (1978)

Rissanen, J.: A universal prior for integers and estimation by minimum description length. Annals of Statistics **11**, 416–431 (1983)

Schumaker, L.L.: Spline Functions: Basic Theory. Wiley, New York (1981)

Schwarz, G.: Estimating the dimension of a model. Ann. Statist. **6**, 461–464 (1978)

Shen, X., Wong, W.H.: Convergence rates of sieve estimates. Ann. Statist. **22**, 580–615 (1994)

Shibata, R.: An optimal selection of regression variables. Biometrika **68**, 45–54 (1981)

Silverman, B.W.: On the estimation of a probability density function by the maximum penalized likelihood method. Ann. Statist. **10**, 795–810 (1982)

Stein, E.M., Weiss, G.: Introduction to Fourier Analysis on Euclidean Spaces, Princeton University Press, Princeton (1971)

Stone, C.J.: An asymptotically optimal window selection rule for kernel density estimates. Ann. Statist. **12**, 1285–1297 (1984)

Stone, C.J.: Large-sample inference for log-spline models. Ann. Statist. **18**, 717–741 (1990)
Stone, C.J.: The use of polynomial splines and their tensor products in multivariate function estimation. Ann. Statist. **22**, 118–184 (1994)
Talagrand, M.: Sharper bounds for empirical processes. Ann. Probab. **22**, 28–76 (1994)
Talagrand, M.: New concentration inequalities in product spaces. Invent. Math. **126**, 505–563 (1996)
Van de Geer, S.: Estimating a regression function. Ann. Statist. **18**, 907–924 (1990)
Van de Geer, S.: The method of sieves and minimum contrast estimators. Math. Methods Statist. **4**, 20–38 (1995)
Vapnik, V.: Estimation of Dependences Based on Empirical Data. Springer, New York (1982)
Wahba, G.: Spline Models for Observational Data. S.I.A.M., Philadelphia (1990)
Whittaker, E.T., Watson, G.N.: A Course of Modern Analysis. Cambridge University Press, London (1927)
Yang, Y., Barron, A.R.: An asymptotic property of model selection criteria. IEEE Transactions on Information Theory **44**, 95–116 (1998)
Yukich, J.E., Stinchcombe, M.B., White, H.: Sup norm approximation bounds for networks through probabilistic methods. IEEE Transactions on Information Theory **41**, 1021–1027 (1995)