Risk Bounds for Model Selection via Penalization

Andrew Barron* Yale University

Lucien Birgé† Université Paris VI

Pascal Massart‡ Université Paris Sud

June 1995

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 regression or density estimation settings and we focus on the problem of estimating the unknown density or regression function. We show that the quadratic risk of the *penalized minimum 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 approximately adaptive (typically up to a slowly varying function of the sample size).

We shall provide various illustrations of our methods 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.

*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"

AMS 1991 subject classifications. Primary 62G05, 62G07; secondary 41A25.

Key words and phrases. Penalization, Model selection, Adaptive estimation, Empirical processes, Sieves, Minimum contrast estimators.

Contents

1 Introduction

To estimate an unknown function we use a sequence of finite-dimensional models and a model selection criterion that takes the form of a penalized empirical loss or contrast function. The function estimate is obtained by minimizing the empirical loss for each model and then selecting the model that optimizes the criterion. The unknown function may or may not be in one of the finite-dimensional models. In the latter case we think of the finite-dimensional models as providing an approximation. The purpose of the model selection criterion is to make a data adaptive choice of model that achieves approximately the best trade-off between approximation error (bias) and additional statistical estimation error (variance) without advance knowledge of which models achieve the best trade-off. Thus we are led to develop upper and lower bounds on the total statistical risk of the estimated function in the model selection context. The aim is to determine suitable forms of criteria and to determine the extent to which it is possible to achieve the desired objectives of adaptive inference.

We allow a general framework of models characterized by their metric dimension properties. The examples we study typically involve linear combinations of a family of basis functions $\{\varphi_\lambda\}$, which are parameterized by an index λ 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 model 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 trade-offs 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 statististical 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 ϕ_w instead of ϕ_λ for the terms that are linearly combined, where w is a continuous vector-valued parameter. Not surprisingly, if the terms ϕ_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.

Let the models S_m indexed by $m \in \mathcal{M}$ be a countable collection of subsets of functions in some given $\mathbb{L}^2$ space. The estimator of the unknown function s takes the form $\hat{s}_n = \hat{s}_{\hat{m},n}$ where $\hat{m} = argmin_m\{\gamma_n(\hat{s}_{m,n}) + pen_{m,n}\}$. Here $\gamma_n(\hat{s}_{m,n})$ is the minimum of the empirical loss $\gamma_n(t)$ for functions t in S_m based on a sample of size n and $pen_{m,n}$ is a penalty function. The precise nature of the sampling model and the penalty function 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 loss $\gamma_n(t)$ is, respectively, the empirical average squared error or $(1/n)$ times the minus logarithm of likelihood. In our examples the models S_m are finitely parameterized and D_m is the parameter dimension. The penalty term is typically of the form $L_m D_m/n$. Here the multiplicative factor L_m takes on one of three typical forms: it is constant (as permitted for certain linear models with a fixed number of models per dimension), it is of the order of $\log n$ (as permitted for certain nonlinear models), or it is the logarithm of some characteristic of the model m (such as the number of terms in the complete model from which m is selected as a subset). As we shall discuss below, the criteria we analyze bear resemblance to and sometimes include as special cases other familiar model selection criteria. A difference here is that the penalty term is motivated solely on the basis of what sorts of statistical risk bounds we can obtain and not on the basis of other information-theoretic or Bayesian considerations.

An important role is played by an index of the accuracy of the models for approximation of the target function s, relative to sample size n. This accuracy index is defined by

$$a_n(s) = \inf_m \left\{d^2(s, S_m) + pen_{m,n}\right\}$$

where $d(s, S_m) = \inf_{t \in S_m} d(s,t)$ is an $\mathbb{L}^2$-distance from s to the subspace S_m [a related index of resolvability for estimators based on the minimum description length principle is in Barron and Cover (1991) and Barron (1991)]. In the density estimation case we will regard s and t as square roots of densities so that d becomes a Hellinger distance between the density s^2 and its approximation; in the regression case d^2 is an integrated squared error of approximation, integrating with respect to an average distribution of the regression inputs. For a given m, $d^2(s, S_m) + D_m/n$ represents the order of magnitude of the risk measured by the mean integrated squared error between s and $\hat{s}_{m,n}$ [at least for certain linear models, see Birgé and Massart (1994a)], the terms $d^2(s, S_m)$ and D_m/n corresponding to the bias squared and variance components, respectively. The penalty $pen_{m,n}$ overestimates D_m/n within a factor L_m which takes into account the additional noise due to simultaneous estimation in a possibly very large number of models and, roughly speaking, the larger this number, the larger the L_m's.

A special role in the analysis is played by the dimension D_{m_n} of models S_{m_n} that minimize the accuracy index. When the penalty term is of order D_m/n (i.e., L_m is a constant) we have that the accuracy index $a_n(s)$, the approximation error $d^2(s, S_{m_n})$, and the ratio D_{m_n}/n of dimension to sample size are asymptotically of the same order. The choice m_n (depending on s) corresponds to the size of model that minimizes this risk.

It is desired that the estimator $\hat{s}_n$ chosen on the basis of the model selection criterion without knowledge of the function s (and without knowledge of its regularity properties) should achieve a value for $\mathbb{E}[d^2(s, \hat{s}_n)]$ that is nearly as good as the the accuracy index $a_n(s)$. Among the results given in this paper are general conditions on the stochastic sampling setting, the candidate models, the contrast function, and the penalty function such that the statistical risk is bounded by the accuracy index

$$\mathbb{E}[d^2(s, \hat{s}_n)] \leq C a_n(s)$$

where $\hat{s}_n$ is the minimizer of the penalized empirical contrast criterion, and the constant C is given in terms of quantities arising in the conditions.

The relationship between the statistical risk and the accuracy index permits the statistician to reduce the problem of investigation of the performance of the estimator (to within certain constant multipliers) to an investigation of the approximation capabilities of the models. 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 $\hat{s}_n$ 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. That is, the penalized minimum contrast estimator is said to be simultaneously *mininax rate optimal* or said to be *adaptive* to such subclasses of functions. See Section 4 for further discussion of *Adaptation* via *Model Selection*.

The present paper is a companion to the paper by two of us (Birgé and Massart 1994b) which explores the role of adaptive estimation for projective estimators of densities using linear models. There a recent and very powerful empirical process inequality by Talagrand (1994) is put to use. Applications are given there for wavelet estimation and connections are established with thresholding of wavelet coefficients and cross-validation criteria.

In the case that a sequence of models m_n is preselected according to presumed properties 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. Sieve methods are forced to lock in at a particular (sometimes suboptimal) trade-off between bias and variance. Nevertheless, the mathematical analysis of sequences of finite-dimensional sieve models is at the heart of the techniques that we put to use in our study of adaptive methods of Model Selection. In particular we will make frequent use of the results of Birgé and Massart (1994a). There, rate of convergence results for sieve estimation are presented, building on empirical process bounds developed first for minimum contrast estimation in Birgé and Massart (1993). Other work on rates of convergence for sieves is in Cencov (1982), Vapnik (1982), Cox (1988), Stone (1990), Barron and Sheu (1991), McGaffrey and Gallant (1994), Haussler (1992), Shen and Wong (1994), and Van de Geer (1993). The results we give here for selections from a set of models can be specialized to give rate results for sieves by considering the degenerate case that there is only one model that the criterion is permitted to select for each sample size n; for the most part that simply recreates the results covered in Birgé and Massart (1994a).

Similarities are apparent in the form of the criteria we study here to the familiar AIC, C_p, BIC, and MDL criteria proposed by Akaike (1973), Mallows (1973), Schwarz (1978), and Rissanen (1978 and 1983), respectively, and the method of structural minimization of risk due to Vapnik (1982). In some cases our analysis provides consistency and rate of convergence results for these criteria. However, in general, penalty terms that satisfy our conditions may be different from those which are used in the familiar criteria.

Shibata (1981) and Li (1987) prove asymptotic optimality properties of the mean squared error for models selected by criteria related to the AIC in the context of linear least squares regression with a fixed design and subexponential growth restrictions on the number of candidate models as a function of dimension. In that important, but restrictive setting, they show that with AIC the choice of a constant value of $L_m = 2\sigma^2$ in the penalty, where σ^2 is an assumed known or consistently estimated homogeneous variance of independent errors, the risk of the estimated model is asymptotically optimal in the sense that it coincides asymptotically (to within a factor converging to one) with the mean squared error achievable if the optimal sequence of models m_n were known. Shibata assumed a Gaussian error model. Specialization of our general results to this setting does not determine the best constant in the penalty and identifies asymptotic optimality only to within a constant factor. However, our results permit relaxation of the fixed design setting and also allow other error distributions with a finite moment generating function. Li's result allows more general error distributions (with moments to a prescribed order) and he requires polynomial rather than merely subexponential bounds on the number of candidate models as a function of dimension. Both Shibata and Li assume the analog of $\inf_m d(s, S_m) = 0$, but they require that s not be contained in any of the S_m's. Our results also relax that requirement while showing asymptotic optimality to within a constant factor.

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 $\hat{s}_n$ which are (at least approximately) simultaneously minimax over a family of compact classes of functions, usually classical smoothness classes. Such estimators are then called (approximately) *adaptive*. There exists a huge amount of litterature devoted to adaptive estimation for non-compact classes of functions of known smoothness. Here adaptation takes place with respect to the radius of the ball in some Sobolev space via various penalization methods [see Wahba (1990) and the references therein and Van de Geer (1990) who introduces empirical process techniques]. The first example we know of adaptation to unknown smoothness via Model Selection is to be found in Efroimovich and Pinsker (1984). In this paper the authors obtain the exact asymptotic minimax risk over a class of Sobolev ellipsoids for the white noise model. Although these results are not presented as a Model Selection method, it turns out that it may be considered as such [see the discussion in Birgé and Massart (1994b), Section 5, Example 2]. Following this seminal work, there have been various extensions to different curve estimation problems in the papers by Efroimovich (1985), Efroimovich and Pinsker (1986) and Polyak and Tsybakov (1990) (with a different Model Selection criterion for this last paper).

Approximate adaptation over more complicated classes of functions with inhomogeneous smoothness (providing spatial adaptation) and based on wavelet thresholding, which may be viewed as Model Selection [see Birgé and Massart (1994b), Section 2.1.3] is to be found in a series of papers by Donoho and Johnstone and Kerkyacharian and Picard [see Donoho, Johnstone, Kerkyacharian and Picard (1995) and the references therein].

The first attempt providing a global approach to Model Selection in order to derive risk bounds using a class of abstract models is to be found in Barron and Cover (1991) or Barron (1991). They 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 with discretization of the parameters of the models reducing 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 they developed an approximation index called the index of resolvability that is a precursor to our accuracy index and they establish comparable risk bounds for Hellinger distance in density estimation and show that for a number of classes of functions (both parametric and nonparametric) the criterion provides convergence rates that are either simultaneous minimax rate optimal or optimal to within logarithmic factors. In particular, discretizations based on $\mathbb{L}^\infty$-metric entropy properties of Sobolev classes of log-densities, together with adaptive selections of the order of smoothness and of the value of the Sobolev norm are shown there to be simultaneously minimax rate optimal 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)], where the nonlinear models gave rise to logarithmic factors in the penalty not present in the idealized models for Sobolev classes. However, the description length method does necessitate a factor that is ultimately of logarithmic order when the target function is in one of the finite-dimensional models as demonstrated by the minimax bounds in Clarke and Barron (1994).

Our work has been inspired by the ideas of Barron and Cover (1991), the main methods and technical tools being developed in Birgé and Massart (1994a). As in Vapnik (1982), they heavily rely on empirical process theory. Vapnik's (1982) method of structural minimization of empirical risk bares many resemblances with ours. The major difference between Vapnik's approach and ours is in the formulation of the empirical process conditions and techniques. Finally we should mention that Yang (who is a student of one of us) has recently got some results similar to ours for the particular case of log-density models [Yang (1995)].

The structure of the paper is as follows: the next section is devoted to the description of the framework underlying the various problems and estimation strategies that we want to handle. We introduce several types of contrast functions and families of approximating models ("sieves"). For each situation we also give a formal result and a simple application of it. Section 3 develops the preceding examples when they require a more sophisticated treatment and also introduces further ones in order to illustrate the power and flexibility of our method. In Section 4, we shall discuss the connections between this technique of Model Selection and Adaptive Estimation. In Section 5 we state all the main theorems in an abstract framework and develop their proofs. Finally Section 6 collects a number of technical lemmas, mainly directed towards approximation theory.

2 Presentation of the framework and main results

The framework is the one used in Birgé and Massart (1994a). We are given $2n$ independent random variables $X_1, \ldots, X_n$ and $W_1, \ldots, W_n$ with values on measurable spaces $\mathcal{X}$ and $\mathcal{W}$ respectively and we observe the variables $Z_i = f(X_i, W_i, s)$ with values on the space $\mathcal{Z}$. The unknown function s which is the parameter to be estimated from the observations may also control the distribution of the X_i's and W_i's. We denote by $\mathbb{P}$ the joint distribution of all the variables (X_i, W_i, Z_i), by $\mathbb{E}$ the expectation with respect to probability $\mathbb{P}$, by $\mathbb{P}_n$ the empirical distribution of the Z_i's and by $\nu_n = \mathbb{P}_n - \mathbb{E} \circ \mathbb{P}_n$ the centered empirical measure. s is assumed to belong to some given space $\mathbb{L}^2(\mu_n)$ with the distance d_n induced by the norm $\|\cdot\| = \|\cdot\|_2$. More generally for $1 \leq p \leq \infty$, the $\mathbb{L}^p$-norm is denoted by $\|\cdot\|_p$.

We introduce the family of *sieves* S_m indexed by $m \in \mathcal{M}_n$ as a countable collection of subsets (the models) of some $\mathcal{S}_n \subset \mathbb{L}^2(\mu_n)$. These sieves play the role of approximating spaces for the true unknown value s of the parameter which might or might not be included in one of them. Typically, S_m will be 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 shall also omit it, from now on, in μ_n and d_n since those quantities will be fixed (independent of n) in most applications. In order to define our estimator we choose a function γ defined on $\mathcal{Z} \times \mathcal{S}_n$ [generally a *contrast function* as defined in Birgé and Massart (1993)] and a *penalty function* $pen(m)$, 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 $\varepsilon \geq 0$ be given, a *penalized minimum contrast estimator* will be defined as follows:

Definition 1 *A* penalized minimum contrast estimator *(relative to the collection of sieves $(S_m)_{m \in \mathcal{M}_n}$) is any parameter value $\hat{s}$ in $\cup_{m \in \mathcal{M}_n} S_m$ with $\hat{s} \in S_{\hat{m}}$ such that*

$$\gamma_n(\hat{s}) + pen(\hat{m}) \leq \inf_{m \in \mathcal{M}_n} \inf_{t \in S_m} \gamma_n(t) + pen(m) + \varepsilon,$$

where $\gamma_n(t) = n^{-1} \sum_{i=1}^n \gamma(Z_i, t)$.

In order to simplify the presentation we shall assume throughout the paper that $\varepsilon = 0$. The choice $\varepsilon = 1/n$ would lead to similar results.

2.1 Some classical contrast functions

Maximum likelihood density estimation: We observe i.i.d. variables $X_1, \ldots, X_n$ of density s^2 with respect to μ. The W_i's play no role here and $Z_i = X_i$. From now on $\mathcal{S}$ will denote the set of nonnegative elements of norm 1 in $\mathbb{L}^2(\mu)$ (which means that their squares are probability densities). We take $\mathcal{S}_n = \mathcal{S}$ and the choice of the contrast function $\gamma(z,t) = -\log t(z)$ leads to *penalized maximum likelihood estimators.*

Projection estimators for density estimation: We assume that μ is a probability measure and that the unknown density of the observations $X_1, \ldots, X_n$ $(Z_i = X_i)$ belongs to $\mathbb{L}^2(\mu)$. It can therefore be written $\mathbb{1}+s$ where s is orthogonal to the constant function $\mathbb{1}$. The contrast is given by $\gamma(z,t) = \|t\|^2 - 2t(z)$ and $\mathcal{S}_n$ is orthogonal to $\mathbb{1}$.

If S_m is a linear space 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

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

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 i.i.d. since we want to include the fixed design models in our framework. s is the unknown parameter which is supposed to belong to the Hilbert space $\mathbb{L}^2(\mu_n)$ where μ_n denotes the average distribution of the X_i's: $\mu_n = n^{-1}\sum_{i=1}^{n} R_i$. This distribution actually depends on n in the fixed design case 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.

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 models and contrast functions have been described in greater detail in Birgé and Massart (1993) and Birgé and Massart (1994a). We therefore refer the reader to these papers for more information.

2.2 Structure of the sieves

The value $pen(m)$ of the penalty function is essentially connected with the number D_m of parameters which are necessary to describe the elements of the sieve S_m. A general definition of D_m will appear in Section 5 and we shall here content ourselves with the presentation of two cases which are known to be of practical interest.

2.2.1 Linear models

For each $m \in \mathcal{M}_n$ we are given a finite-dimensional linear subspace $\bar{S}_m$ of $\mathbb{L}^2(\mu)$ with an orthonormal basis $\{\varphi_{\lambda,m}\}_{\lambda\in\Lambda_m}$ with $D_m = |\Lambda_m|$ and S_m is a subset of $\bar{S}_m$. In order to simplify the notations we shall write in the sequel $\{\varphi_\lambda\}_{\lambda\in\Lambda_m}$ instead of $\{\varphi_{\lambda,m}\}_{\lambda\in\Lambda_m}$, keeping in mind that φ_λ may depend on m. We assume that $\|\varphi_\lambda\|_\infty < +\infty$ for all $\lambda \in \Lambda_m$ and define for $\beta \in \mathbb{R}^{\Lambda_m}$, $|\beta|_\infty = \sup_{\lambda\in\Lambda_m} |\beta_\lambda|$ and $|\beta|_2^2 = \sum_{\lambda\in\Lambda_m} \beta_\lambda^2$. Next we define

$$\bar{r}_m = \frac{1}{\sqrt{D_m}} \sup_{\beta\neq 0} \frac{\|\sum_{\lambda\in\Lambda_m} \beta_\lambda\varphi_\lambda\|_\infty}{|\beta|_\infty} \quad \text{and} \quad \Phi_m = \frac{1}{\sqrt{D_m}} \sup_{\beta\neq 0} \frac{\|\sum_{\lambda\in\Lambda_m} \beta_\lambda\varphi_\lambda\|_\infty}{|\beta|_2}. \tag{2.1}$$

It follows from this definition that

$$\Phi_m \leq \bar{r}_m \leq \sqrt{D_m}\,\Phi_m \tag{2.2}$$

and from Lemma 6 of Birgé and Massart (1994a) that

$$\Phi_m = \frac{1}{\sqrt{D_m}} \|\sum_{\lambda\in\Lambda_m} \varphi_\lambda^2\|_\infty^{1/2}. \tag{2.3}$$

In opposition with what happens for the white noise model, the $\mathbb{L}^2$-structure of the sieves is not sufficient to guarantee a good behavior of the empirical contrast function γ_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 sieves. A control of the growth of $\bar{r}_m$ or Φ_m provides such connections.

2.2.2 Nonlinear Models

Here we have in mind a variety of models that include single hidden layer sigmoidal networks [see Barron (1993), (1994)], sparce trigonometric models, certain multivariate wavelet models as in Hornik et al. (1994), 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 on $\mathbb{R}^{q'}$, where the functions ϕ_w satisfy the Lipschitz property

$$|\phi_w(x) - \phi_{w'}(x)| \leq |w - w'|_1 \quad \text{for all } x \in [-1,1]^q, \tag{2.4}$$

$|\cdot|_1$ 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

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

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 ψ is a fixed univariate function with Lipschitz constant 1, and $w = (a, b)$ with $a \in \mathbb{R}^q$, $b \in \mathbb{R}$, and $q' = q + 1$. Then the Lipschitz property (2.4) for ϕ_w holds. The cases mentioned above are of this ridge expansion form. For the neural net case ψ 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 ψ 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 ψ 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 $\psi(z) = e^{-z}$, 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 ψ 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 $\phi_w(x) = \psi_{w_1}(x_1) \cdots \psi_{w_q}(x_q)$ 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 (1989).

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 ψ 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 the ϕ 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.

2.3 Examples

In order to keep the presentation of our results simple, we shall now concentrate on linear models and return to the nonlinear models in Section 3.2.

Uniformly bounded basis: When $\|\varphi_\lambda\|_\infty \leq \Phi$ for all $\lambda \in \Lambda_m$ and all m, then $\Phi_m \leq \Phi$ by (2.3). Subsets of the trigonometric basis in $\mathbb{L}^2([0,2\pi], dx)$ provide typical examples of uniformly bounded bases.

Wavelet expansions: Let us consider an orthonormal wavelet basis $\{\varphi_{j,k} \,|\, j \geq 0, k \in \mathbb{Z}^q\}$ 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:

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

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 $\{(j,k) \,|\, 1 \leq k \leq 2^{jq}M\}$. The relevant Λ_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, when $\Lambda_m = \cup_{j=0}^{J_m} \Lambda(j)$, $\bar{r}_m$ is uniformly bounded and so is Φ_m. The most relevant applications of such expansions have been studied extensively in Birgé and Massart (1994b).

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 $t \in \mathbb{L}^2([0,1],dx)$:

$$t = \beta_{-1,1}\varphi_{-1,1} + \sum_{j\geq 0}\sum_{k=1}^{2^j}\beta_{j,k}\varphi_{j,k}, \tag{2.5}$$

where $\varphi_{-1,1} = 1\!\!1_{[0,1]}$, $\psi = 1\!\!1_{[0,1/2]} - 1\!\!1_{]1/2,1]}$ and $\varphi_{j,k}(x) = 2^{j/2}\psi(2^j(x-(k-1)2^{-j}))$. We set $\Lambda(-1) = \{(-1,1)\}$ and for $j \geq 0$ $\Lambda(j) = \{(j,k) \,|\, 1 \leq k \leq 2^j\}$. If $\Lambda_m = \cup_{j=0}^m \Lambda(j)$ we see from (2.3) that $\Phi_m = 1$. To bound $\bar{r}_m$ we first notice that for $j \geq 0$

$$\|\sum_{k=1}^{2^j}\beta_{j,k}\varphi_{j,k}\|_\infty \leq 2^{j/2}\sup_k|\beta_{j,k}|. \tag{2.6}$$

Therefore

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

It will 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 $\mathcal{P}_i$ a partition of $[0,1]$ into $D(i)$ intervals. A linear space $\bar{S}_m$ of piecewise polynomials is characterized by $m = (r, \mathcal{P}_1, \ldots, \mathcal{P}_q)$ 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 $\mathcal{P} = \otimes_{i=1}^q \mathcal{P}_i$. This results in $D_m = (r+1)^q \prod_{i=1}^q D(i)$.

Starting with the orthogonal basis of the Legendre polynomials $Q_j, j \in \mathbb{N}$ in $\mathbb{L}^2([-1,1],dx)$, we notice that the following properties hold [see Whittaker and Watson (1927) pp. 302-305 for details]

$$|Q_j(x)| \leq 1 \quad \text{for all } x \in [-1,1], \quad Q_j(1) = 1, \quad \text{and} \quad \int_{-1}^1 Q_j^2(t)dt = \frac{2}{2j+1}.$$

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

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

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

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

Then taking Λ_m as the set of those (R,j)'s such that $R \in \mathcal{P}$ and $j \in \mathcal{J}$ we get from (2.1)

$$\bar{r}_m^2 \leq \frac{(r+1)^{2q}(2r+1)^q}{D_m \inf_{R\in\mathcal{P}} Vol(R)} = [(r+1)(2r+1)]^q\left[\inf_{R\in\mathcal{P}} Vol(R)\prod_{i=1}^q D(i)\right]^{-1}. \tag{2.7}$$

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

$$\bar{r}_m \leq [(r+1)(2r+1)]^{q/2} \tag{2.8}$$

Polynomials on a sphere and other eigenspaces of the Laplacian: Let $\mathbb{S}^q$ be the unit euclidean sphere of $\mathbb{R}^{q+1}$, μ 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 θ_j. Then $\{\mathbb{1}\} \cup \{\varphi_\lambda, \lambda \in \Lambda\}$ where $\Lambda = \cup_{j \geq 0}\Lambda(j)$ is an orthonormal basis of $\mathbb{L}^2(\mu)$. Defining $\Lambda_m = \cup_{j=0}^m \Lambda(j)$, for any integer $m \geq 0$ 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]

$$\sum_{\lambda \in \Lambda(j)} \varphi_\lambda^2(x) \equiv |\Lambda(j)|.$$

In such a case it follows from (2.3) that $\Phi_m = 1$ for any integer m.

More generally, we can consider, instead of $\mathbb{S}^q$, a compact Riemannian manifold $\mathbb{M}$ of dimension q with its uniform distribution μ. 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 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

$$C_1(\mathbb{M}) D_m^{2/q} \leq \theta_m \leq C_2(\mathbb{M}) D_{m-1}^{2/q} < C_2(\mathbb{M}) D_m^{2/q}. \tag{2.9}$$

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

$$\sum_{j=0}^{\infty} \left[e^{-\theta_j t} \sum_{\lambda \in \Lambda(j)} \varphi_\lambda^2(x) \right] \leq C_3(\mathbb{M}) t^{-q/2} \tag{2.10}$$

for any positive t, any $x \in \mathbb{M}$ and some fixed positive constant $C_3(\mathbb{M})$. Applying (2.10) with $t = \theta_m^{-1}$ yields

$$\| \sum_{\lambda \in \Lambda_m} \varphi_\lambda^2 \|_\infty \leq e C_3(\mathbb{M}) \theta_m^{q/2}.$$

Combining this inequality with (2.9), we can derive from (2.3) that for any integer m, $\Phi_m^2 \leq \Phi^2(\mathbb{M}) = e C_3(\mathbb{M}) C_2(\mathbb{M})^{q/2}$ which implies that Φ_m is uniformly bounded as in the case of the sphere.

2.4 Some typical results

We now assume that the situation described at the beginning of Section 2.2 holds, i.e. S_m is a subset of a linear space $\bar{S}_m \in \mathbb{L}^2(\mu)$ of dimension D_m with Φ_m and $\bar{r}_m$

defined by (2.1). We shall also need, from now on, a number of different constants. In order to make our notations more transparent we shall hereafter systematically denote by the letter κ 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 the same notation will be used for different constants from one section to another, the form $C(\cdot, \cdots, \cdot)$ emphasizing the dependence of C on the others constants. Finally, the letter K will, most of the time, denote numerical constants, either to be chosen by the statistician or functions of such ones.

2.4.1 Maximum likelihood

We observe n i.i.d. variables $X_1, \ldots, X_n$ of density s^2 with respect to some *probability* measure μ. The model $S_m = \bar{S}_m \cap \mathcal{S}$ consist of those functions $t \in \bar{S}_m$ 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 μ 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.

$$K(u,v) = \int \log \frac{dP_u}{dP_v} dP_u \quad \text{if} \quad P_u \ll P_v \qquad \text{and} \qquad K(u,v) = +\infty \quad \text{otherwise.}$$

Theorem 1 *Let $\{L_m\}_{m \in \mathcal{M}_n}$ be a family of weights such that*

$$L_m \geq 1 \quad \textit{for all } m \in \mathcal{M}_n \qquad \textit{and} \qquad \sum_{m \in \mathcal{M}_n} \exp[-L_m D_m] = \Sigma < +\infty. \tag{2.11}$$

Let $pen(m)$ be such that

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

where κ_1 is a suitable positive numerical constant and let $\hat{s}$ be the penalized maximum likelihood estimator which is a minimizer with respect to $m \in \mathcal{M}_n$ and $t \in S_m$ of $-(1/n)\sum_{i=1}^n \log(t(X_i)) + pen(m)$ 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 $m \in \mathcal{M}_n$. Then

$$\mathbb{E}[d^2(s,\hat{s})] \leq \kappa_1' \left[\inf_{m \in \mathcal{M}_n} \{K(s, S_m) + pen(m)\} \wedge 1 \right]. \tag{2.12}$$

The upper bound in (2.12) 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 one will prove in Section 6 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 nonegative elements of norm 1 in $\bar{S}_m$. If there exists $s_m^+ \geq s$ in $\bar{S}_m$ then*

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

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

$$K(s, S_m) \wedge 1 \leq 12 \inf_{t \in \bar{S}_m} \|s - t\|_\infty^2. \tag{2.13}$$

Application to adaptive histograms We consider a family of sieves which are sets of piecewise polynomials on $[0,1]$ of degree 0, i.e., histograms, and take μ 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 $\mathcal{R}_n$ be the set of all regular partitions with at most n pieces and $\mathcal{G}_{n,N}$ be the set of all partitions with at most n pieces and endpoints belonging to the grid $\{j/N \mid 0 \leq j \leq N\}$. Then we define $\mathcal{M}_n = \mathcal{R}_n \cup (\cup_{N \geq 2} \mathcal{G}_{n,N})$ and choose $L_m = 1$ when $m \in \mathcal{R}_n$, $L_m = 2 + 3\log(N/D_m)$ when $m \in \mathcal{G}_{n,N}$. It follows from our study of piecewise polynomials that $\bar{r}_m \leq 1$ by (2.8) when $m \in \mathcal{R}_n$ and is bounded by $\sqrt{N/D_m}$ otherwise. Since the number of partitions of $\mathcal{G}_{n,N}$ with D pieces is bounded by $(eN/D)^D$ from Lemma 6 and clearly $D \leq N$, (2.11) is satisfied and we can apply Theorem 1 with

$$pen(m) = K_1(1 + \log 2) D_m/n \qquad \text{if } \ m \in \mathcal{R}_n, \tag{2.14}$$

$$pen(m) = K_1 \left[2 + 3\log\left(\frac{N}{D_m}\right) + \log\left(1 + \sqrt{\frac{N}{D_m}}\right)\right] \frac{D_m}{n} \qquad \text{if } \ m \notin \mathcal{R}_n \tag{2.15}$$

and $K_1 \geq \kappa_1$.

- If s is Hölderian of order α, i.e.

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

  $H > 0$ and $\alpha \in (0,1]$ being unknown, for each $m \in \mathcal{R}_n$, the $\mathbb{L}^\infty$-distance between s and S_m is bounded by $CHD_m^{-\alpha}$ and therefore by (2.13) $K(s, S_m) \wedge 1$ is bounded by $C'H^2 D_m^{-2\alpha}$. In that case (2.12) implies that the quadratic risk of our estimator is bounded by $C''H^{2/(2\alpha+1)} n^{-2\alpha/(2\alpha+1)}$. We shall see in Section 3.1.2 that even if H and α 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 $m \in \mathcal{G}_{n,N}$, (2.12) implies that the risk is bounded by $C\log(N/D_m)D_m/n$, which is of the usual parametric order $1/n$ 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 $\{m \in \cup_{N=2}^{n^l} \mathcal{G}_{n,N} \mid D_m = D\}$. On the other hand for a given value of D, $9 \leq D \leq n/5$, if follows from Corollary 1 of Birgé and Massart (1994a) 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 one estimator we achieve simultaneously the optimal $1/n$ rate for s in the parametric subfamilies, the optimal rate $n^{-2\alpha/(2\alpha+1)}$ for the α-Hölderian densities and within a logarithmic factor of this optimal rate for much less homogeneous functions with smoothness α. This will be illustrated in Section 3.2.1 below for densities with bounded α-variation.

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 ditribution function, automatically achieves a bound on the $\mathbb{L}^1$-risk which is analogous to (3.31) below with $\alpha = 1$ but without the $\log n$ factor [see Birgé (1989)].

2.4.2 Projection estimators

Although various adaptive properties of penalized projection estimators have been enlightened in Birgé and Massart (1994b), we shall develop here some new illustrations. The basic result is similar to Theorem 3 of Birgé and Massart (1994b). We recall that here μ is a probability measure. Since the true unknown density is $\mathbb{1} + s$, the space $\mathcal{S}_n$ is chosen to be a linear subspace of $\mathbb{L}^2(\mu)$ orthogonal to $\mathbb{1}$ with an orthonormal basis $\{\varphi_\lambda \mid \lambda \in \bar{\Lambda}_n\}$ and the family of finite dimensional linear spaces $\bar{S}_m$ is generated by a family $\{\Lambda_m\}_{m\in\mathcal{M}_n}$ of finite subsets of $\bar{\Lambda}_n$: for each $m \in \mathcal{M}_n$, $S_m = \bar{S}_m$ is the linear span of $\{\varphi_\lambda \mid \lambda \in \Lambda_m\}$ and $D_m = |\Lambda_m| \leq n$.

Theorem 2 *Assume that the family $\{\Lambda_m\}_{m\in\mathcal{M}_n}$ is totally ordered by inclusion and that the Φ_m's are uniformly bounded by Φ. Let κ_2 be a suitable numerical constant, $pen(m) \geq \kappa_2\Phi^2 D_m/n$ and $\hat{s}$ be the penalized projection estimator which is a minimizer with respect to m of $-\sum_{\lambda\in\Lambda_m}\hat{\beta}_\lambda^2 + pen(m)$, where $\hat{\beta}_\lambda = n^{-1}\sum_{i=1}^n \varphi_\lambda(X_i)$. Then*

$$\mathbb{E}[\|\hat{s} - s\|^2] \leq \kappa_2' \inf_{m\in\mathcal{M}_n} \{d^2(s, S_m) + pen(m)\} + \kappa_2'' \frac{[\Phi(1 + \|s\|_2)]^4}{n}. \tag{2.16}$$

Application to ellipsoids with unknown coefficients: We consider some orthonormal system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ of $\mathbb{L}^2(\mu)$ where $\Lambda = \cup_{j\in\mathbb{N}}\Lambda(j)$, each $\Lambda(j)$ being a finite set. Let us mention here two cases of particular interest to be studied in Section 3.

- μ is the uniform distribution on the torus $[0, 2\pi]$, $\Lambda(j) = \{2j; 2j+1\}$ for $j \geq 0$ and $\varphi_{2j}(x) = \sqrt{2}\cos((j+1)x)$, $\varphi_{2j+1}(x) = \sqrt{2}\sin((j+1)x)$.
- μ 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 2.3.

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

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

Let us define for each $m \in \mathbb{N}$, $\Lambda_m = \sum_{j=0}^m \Lambda(j)$ and $D_m = |\Lambda_m|$. Then $\mathcal{M}_n = \{m \geq 0 \mid D_m \leq n\}$. For the sake of simplicity, we shall limit ourselves in this section to the case $a_j(H,\alpha) = HD_j^{-\alpha}$ 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 $\cup_{H>0}\mathcal{E}(a(H,\alpha))$ is the set of periodic functions orthogonal to $\mathbb{1}$ belonging to the

Sobolev space W_2^α in the Fourier case and of functions orthogonal to $\mathbb{1}$ belonging to the Besov space $B_{2,2,\alpha}$ in the Haar case.

If s is an element of some ellipsoid $\mathcal{E}(a)$, it is immediate to see that $d^2(s, S_m) \leq a_m^2$. Therefore from Theorem 2 with $\Phi^2 = 2$, $K_2 \geq \kappa_2$ and $pen(m) = 2K_2D_m/n$ we get

$$\mathbb{E}\left[\|s - \hat{s}\|^2\right] \leq \kappa_2' \inf_{m \in \mathcal{M}_n} \left\{ H^2 D_m^{-2\alpha} + \frac{2K_2D_m}{n} \right\} + 4\kappa_2'' \frac{(1 + \|s\|_2)^4}{n}$$

and finally whatever the true unknown values of H and α,

$$\mathbb{E}\left[\|s - \hat{s}\|^2\right] \leq K_2' \left[\left(\frac{H}{n^\alpha} \right)^{2/(1+2\alpha)} + \frac{(1 + \|s\|_2)^4}{n} \right].$$

The discussion about the optimality properties of such a bound will be developed in Section 3. Related work using ellipsoids built on the Fourier basis are to be found in Efroimovich and Pinsker (1984) (for the white noise model) and (1986) (for the spectral density), Efroimovich (1985) (for density estimation) and Polyak and Tsybakov (1990) (for regression models), all the results, except for the first one, being restricted to Hilbert-Schmidt ellipsoids (i.e. $\alpha > 1/2$ in the above example).

2.4.3 Least squares estimators for smooth regression

We consider a regression model $Y_i = s(X_i) + W_i$ where the X_i's are i.i.d. with common distribution μ and the W_i's are i.i.d. and centered. We assume that $\|s\|_\infty$ is bounded by some known constant ξ and that all elements of $\mathcal{S}_n$ are bounded by ξ as well.

Theorem 3 *Assume that* $\mathbb{E}[e^{|W_1|/\xi'}] \leq 4$ *for some* $\xi' > 0$ *and let* $\{L_m\}_{m \in \mathcal{M}_n}$ *be a family of weights such that*

$$L_m \geq 1 \quad \textit{for all } m \in \mathcal{M}_n \qquad \textit{and} \qquad \sum_{m \in \mathcal{M}_n} \exp[-L_m D_m] = \Sigma < +\infty. \tag{2.17}$$

Let κ_3 *be a suitable numerical constant,*

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

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

$$\mathbb{E}[d^2(s, \hat{s})] \leq \kappa_3' \inf_{m \in \mathcal{M}_n} \left\{ d^2(s, S_m) + pen(m) \right\}. \tag{2.18}$$

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 $[0, 1]$, that the errors W_i satisfy the assumptions of Theorem 3 with a known constant ξ'. We consider simultaneously five different types of sieves indexed by the sets $\mathcal{M}^i$ with $1 \leq i \leq 5$ and take $\mathcal{M}_n = \cup_{1 \leq i \leq 5} \mathcal{M}^i$. Let us fix $r \in \mathbb{N}$. We define $\mathcal{M}^1$ to be the set of regular partitions of $[0, 1]$ and $\mathcal{M}_N^2$ to be the set of all partitions with endpoints belonging to the grid $\{j/N \mid 0 \leq j \leq N\}$. Then $\mathcal{M}^2 = \cup_{N\geq 2}\mathcal{M}^2_N$. 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 $\Lambda = \cup_{j\geq 0}\Lambda(j)$. We first consider the trigonometric basis with $\Lambda(0) = \{0\}$, $\Lambda(j) = \{2j-1; 2j\}$ for $j \geq 1$ and $\varphi_0 = 1\!\!1$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)$, $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)$. Then $\mathcal{M}^3 = \mathbb{N}$ and $\Lambda_m = \cup_{j\leq m}\Lambda(j)$. Finally we introduce a wavelet basis of regularity r as described in Section 2.3 with $q = A = 1$. An element m of $\mathcal{M}^4$ or $\mathcal{M}^5$ is a subset of the set of indices $\{(j,k) \mid 1 \leq k \leq 2^{jq}M,\ j \in \mathbb{N}\}$ 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 subsets which do not belong to $\mathcal{M}^4$ of the sets $\cup_{j=0}^J\Lambda(j)$ with $J \in \mathbb{N}$, i.e. all non-trivial subsets of the elements of $\mathcal{M}^4$.

If we assume that the true regression function s is bounded by $L\xi'$ where L is known, it is natural to restrict our sieves to sets of functions which are uniformly bounded by a constant $\xi = (L+1)\xi'$ for instance and therefore to choose $S_m = \bar{S}_m \cap \{t \mid \|t\|_\infty \leq \xi\}$ for each $m \in \mathcal{M}_n$. In order to describe the penalty function, it is enough to compute $\bar{r}_m$ and choose L_m for any m in order that condition (2.17) should be satisfied. It follows from Section 2.3 that $\bar{r}_m$ is uniformly bounded if m belongs to either $\mathcal{M}^1$ or $\mathcal{M}^4$. It follows from (2.2) that $\bar{r}_m \leq \sqrt{2D_m}$ if $m \in \mathcal{M}^3$. Finally for $m \in \mathcal{M}^2$, $\bar{r}_m \leq C\sqrt{N/D_m}$ and for $m \in \mathcal{M}^5$, $\bar{r}_m \leq C'\sqrt{2^J/D_m}$. As to L_m it can be choosen as 1 for $m \in \mathcal{M}^i$, $i = 1, 3$ or 4 and by Lemma 6 we can take $L_m = 2 + 3\log(N/D_m)$ for $m \in \mathcal{M}^2$ and $L_m = 2 + 3\log(M2^J/D_m)$ for $m \in \mathcal{M}^5$.

We have shown that Theorem 3 applies to this situation leading to the upper bound (2.18) for the risk. It is difficult to analyze this bound in general. Moreover the minimax point of view is especially inadequate 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 compact class of functions, which always reflect 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 numbers H and α with $\alpha = a + b$, $a \in \mathbb{N}$, $0 < b \leq 1$ the class $\mathcal{H}(H,\alpha)$ of functions f on $[0,1]$ with a derivatives and such that

$$|f^{(a)}(x) - f^{(a)}(y)| \leq H|x-y|^b \quad \text{for all } x, y \in [0,1]. \tag{2.19}$$

Recalling that $\mathcal{H}(H,\alpha)$ is included in the Besov space $B_{\alpha,\infty,\infty}$, it follows from Lemma 13 that for any positive ε one can find in each of the three collections $\mathcal{M}^i$, $i = 1, 3, 4$ an m such that $d(s, S_m) \leq \varepsilon$ and $D_m \leq C_i\varepsilon^{-1/\alpha}$ (with the additional assumption that s is periodic when $i = 3$ or that the support of s is included in $(0,1)$ when $i = 4$). Let us denote by $\bar{\bar{r}}_m$ the upper bound for $\bar{r}_m$ computed above and choose

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

with $K_3 \geq \kappa_3$. It then follows that for $i = 1, 3, 4$ the upper bound for the risk derived from Theorem3 takes the form

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

where it is required that $\alpha \geq 1/2$ if $i = 3$ because of the influence of $\bar{r}_m$. Since $C_i'(s)$ is uniformly bounded over the class $\mathcal{H}(H, \alpha)$ if L is given, our estimator clearly achieves the optimal rate of convergence in the minimax sense but does more than that since the numbers $C_i'(s)$ can vary substantially from one s to another and one i to another as well. 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 model) 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 (in a strong sense, see Section 4 below) estimation for Sobolev classes in a gaussian regression framework.

2.4.4 Least squares estimators for binary images

Let us now turn to a quite different situation essentially motivated by image analysis. The new framework is a model of the form $Y_i = s(X_i) + W_i$ where $s = \theta_f$ and $\|s\|_\infty \leq 1$. Typically $s = \theta_f$ will be the indicator function of a set the boundary of which is parametrized by the function f belonging to $\mathcal{G}_n$. For 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 consider the following framework: $\mathcal{G}_n$ is an interval of some space $\mathbb{L}^1(\mu')$ equipped with its natural distance d_1. More precisely, there exists two functions $F^+, F^- \in \mathbb{L}^1(\mu')$ such that $\mathcal{G}_n = \{g \in \mathbb{L}^1(\mu') \,|\, F^-(x) \leq g(x) \leq F^+(x)$ for all $x\}$. θ is a one-to-one mapping from $\mathcal{G}_n$ onto $\mathcal{S}_n \subset \mathbb{L}^2(\mu_n)$ where μ_n denotes the average distribution of the X_i's. S_m is the image of $G_m \subset \mathcal{G}_n$ by the mapping θ, where G_m is included in a linear subspace $\bar{G}_m$ with dimension D_m of $\mathbb{L}^1(\mu')$.

Theorem 4 *Assume 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*

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

- θ *is non-decreasing and maps* $\mathcal{G}_n$ *into* $\{t \,|\, \|t\|_\infty \leq 1\}$*;*

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

$$\Theta_1 \|h - g\|_1 \leq \|\theta_h - \theta_g\|^2 \qquad \textit{and} \qquad \|\theta_h - \theta_g\|_1 \leq \Theta_2 \|h - g\|_1; \tag{2.20}$$

- *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* $\bar{G}_m$ *with* $\|\varphi_\lambda\|_1 = 1$ *for all* λ *such that*

$$\sum_{\lambda \in \Lambda_m} |\beta_\lambda| \leq B_m'' \, \|\sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda\|_1 \quad \textit{for all } (\beta_\lambda) \in \mathbb{R}^{\Lambda_m}. \tag{2.21}$$

Let κ_4 be a suitable positive numerical constant,

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

and $\hat{f}$ be the penalized least squares estimator which is a minimizer with respect to $m \in \mathcal{M}_n$ and $g \in G_m$ of $n^{-1}\sum_{i=1}^n (Y_i - \theta_g(X_i))^2 + pen(m)$. Then

$$\Theta_1 \mathbb{E}[d_1(f, \hat{f})] \leq \kappa'_4 \inf_{m \in \mathcal{M}_n} \{\Theta_2 d_1(f, G_m) + pen(m)\}. \tag{2.22}$$

Application to binary images: Here $\mathcal{G}_n$ is the set of all measurable functions g from $[0,1]$ to $[0,1]$ and for each $g \in \mathcal{G}_n$ we define for $(x,y) \in [0,1]^2$, $\theta_g(x,y) = 1$ if $y \leq g(x)$ and 0 otherwise. Following Korostelev and Tsybakov (1993b) we consider the regression model $Y_i = \theta_f(X_i) + W_i$ and assume that μ_n is uniform on $[0,1]^2$ and μ' is uniform on $[0,1]$. f should be understood as the parametrization of a boundary fragment corresponding to some portion of a binary image in the plane. Then $\|\theta_h - \theta_g\|^2 = \|\theta_h - \theta_g\|_1 = \|h - g\|_1$ and we may take $\Theta_1 = \Theta_2 = 1$ in (2.20). Assuming that the errors W_i are either bounded by 1 or gaussian with variance smaller than 1 then $\xi' = 1$.

Let $\mathcal{R}(J)$ denote the regular partition of $[0,1]$ into J pieces and $\mathcal{M}_n$ be the set $\{(r, \mathcal{R}(J)) \,|\, J \geq 1,\, r \geq 0\}$. Following the definition of Section 2.3, we consider $\bar{G}_m$ to be the space of piecewise polynomials of degree not larger than r based on the regular partition $\mathcal{R}(J)$ if $m = (r, \mathcal{R}(J))$. Then $D_m = (r+1)J$ and we can take $L_m = 1$. Let us now turn to the verification of (2.21). We consider some orthonormal (with respect to $\mathbb{L}^2(\mu')$) basis $\psi_0, \ldots, \psi_r$ of the space of polynomials on $[0,1]$ with degree $\leq r$ and we define $\varphi_l = \alpha_l \psi_l$ with $\alpha_l \geq 1$ by $\|\varphi_l\|_1 = 1$. Then for any $(\beta_l) \in \mathbb{R}^{r+1}$

$$\begin{aligned}
\sum_{l=0}^r |\beta_l| &\leq \sum_{l=0}^r |\alpha_l \beta_l| \leq \left[(r+1)\sum_{l=0}^r |\alpha_l \beta_l|^2\right]^{1/2} = \sqrt{r+1}\,\|\sum_{l=0}^r \beta_l \varphi_l\|_2 \\
&\leq 2r\sqrt{r+1}\,\|\sum_{l=0}^r \beta_l \varphi_l\|_1
\end{aligned} \tag{2.23}$$

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

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

It then easily follows from (2.23) that (2.21) is satisfied with $B''_m = 2r\sqrt{r+1}$. We now define $G_m = \bar{G}_m \cap \mathcal{G}_n$. If $f \in \mathcal{G}_n$ belongs to some Besov space $B_{\alpha,\infty,\infty}$ where α is unknown, it follows from Lemma 13 in Section 6 that $\|f - g_m\|_\infty \leq \varepsilon = C(f,r)J^{-\alpha}$ for some $g_m \in \bar{G}_m$. Changing if necessary g_m into $(g_m + \varepsilon)/(1 + 2\varepsilon)$ and ε into 4ε we can assume that $g_m \in G_m$. Choosing $pen(m) = 4K_4[1 + \log(1 + 2r\sqrt{r+1})]$ with $K_4 \geq \kappa_4$, we then derive from Theorem 4 that $\mathbb{E}[d_1(f, \hat{f})]$ is bounded by $C'(f,\alpha)n^{-\alpha/(1+\alpha)}$.

Remarks:

- The rates may seem unusual as compared to density estimation. It results from the fact that $\|\theta_h - \theta_g\|^2 = d_1(h,g)$ which leads to a risk expressed as the sum of a bias term and a variance term instead of the classical variance plus square bias. It comes from Korostelev and Tsybakov (1993b) that the rates correspond to the optimal rate (in the minimax sense) when α 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 ρ, ψ with $0 \leq \rho \leq 1$ and ψ belongs to the one-dimensional torus $\mathbb{T}$. We then define $\mathcal{G}_n$ as the set of functions g from $\mathbb{T}$ to $[0,1]$ and set $\theta_g(\rho \cos\psi, \rho \sin\psi) = 1\!\!1_{\{\rho^2 \leq g(\psi)\}}(\rho, \psi)$. Choosing μ as the uniform distribution on the disk and μ' as Lebesgue measure on $\mathbb{T}$ we can check that (2.20) is satisfied with $\Theta_1 = \Theta_2 = 1/2$.

2.4.5 A glimpse of the essentials

The proofs (see Section 5 below) are essentially based on two arguments:

- A control of the fluctuations of a weighted version of the process $\nu_n[\gamma(., s_m) - \gamma(., u)]$ when $u \in S_{m'}$ and s_m realizes (approximately) the distance between s and S_m. More precisely one looks for some exponential control of

$$\frac{\nu_n[\gamma(., s_m) - \gamma(., u)]}{d^2(s, u) \vee d^2(s_m, s) \vee pen(m) \vee pen(m')} \tag{2.24}$$

uniformly with respect to $u \in S_{m'}$ and $m' \in \mathcal{M}_n$. For a given value of m' such a control restricted to one sieve may be derived by using various empirical processes techniques: suitable modifications of the entropy methods introduced by Dudley (1978) or recent arguments by Talagrand (1994) based on the concentration of product measures. Such results are collected in Proposition 7 below which is mainly based on Theorems 1 and 2 of Birgé and Massart (1994a).

It should be noticed that an analogue of Proposition 7 could be easily obtained for least squares estimation in the white noise model using purely $\mathbb{L}^2$-metric conditions which are known to determine completely the behavior of a gaussian process. More than that, one could even for linear models calculate the constants involved in the exponential inequalities and derive from that explicit constraints on the penalty function. Most of the technicalities required in our proofs are due to the fact that we are not working with the white noise model.

In order to get a uniform result over the values of m' one needs a limitation on the number of sieves at hand which, in most cases, takes the following form:

S *There exists a family of weights* $\{L_m\}_{m \in \mathcal{M}_n}$ *such that*

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

If one wants to restrict to $L_m = 1$, it is necessary to limit the number of sieves at hand. In order to work with larger families of sieves, one has to choose bigger values of L_m. Since, roughly speaking, $pen(m)$ is of order $L_m D_m / n$, this leads to larger values of the penalty which will affect the risk.

- It follows from the definition of $\hat{s}$ that

$$\ell_n(s,\hat{s}) \leq \ell_n(s,s_m) + \nu_n[\gamma(.,s_m) - \gamma(.,u)] + pen(m) - pen(m')$$

  where $\ell_n(s,t) = (1/n)\sum_{i=1}^n \mathbb{E}[\gamma(Z_i,t) - \gamma(Z_i,s)]$. This inequality can be combined with the previous argument whenever ℓ_n is connected to d by $d^2(s,t) \asymp \ell_n(s,t)$ uniformly for $t \in \mathcal{S}_n$.

These arguments yield a final control of the risk of the form

$$\mathbb{E}[d^2(s,\hat{s})] \leq C \inf_{m\in\mathcal{M}_n} \left\{d^2(s,S_m) + pen(m)\right\}.$$

As a consequence, the price to pay for the use of a large family of sieves which implies large values of L_m is the effect of these values of L_m in the final risk since $pen(m)$ is roughly proportional to L_m.

As we have already seen, all our theorems are valid for values of $pen(m)$ larger than KL_mD_m/n in order to control (2.24). On the other hand, since $pen(m)$ appears as a summand of the upper bound on the risk it should be taken as small as possible. The ideal choice of $pen(m)$ should be the minimal value allowing a good control of (2.24) but this will not be provided by our theory. One can only notice that a moderate overestimation of the optimal penalty leads to a moderate overestimation of the risk but that underestimation of the penalty (think of the extremal case $pen(m) \equiv 0$) can lead to inconsistent estimation.

3 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 on a representative selection.

It should be noted that each particular family of sieves will be given for a particular type of contrast (maximum likelihood, projection or least squares regression) 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.

3.1 Nested families of sieves and analogues

By this we mean that the family of sieves 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 sieves with the same dimension D_m is rather small, at least small enough to ensure that the series $\sum_{m\in\mathcal{M}_n} \exp(-D_m) < +\infty$.

3.1.1 Ellipsoids with unknown coefficients

We shall here give 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 2.4.2. The ellipsoids are given by some orthonormal system $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ of $\mathbb{L}^2(\mu)$ where $\Lambda = \cup_{j\in\mathbb{N}}\Lambda(j)$, each $\Lambda(j)$ being a finite set. Furthermore $\int \varphi_\lambda d\mu = 0$ for all λ. We recall from Section 2.4.2 that for any non-increasing 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 $D_0 \leq n$) $\mathcal{M}_n = \{m \in \mathbb{N} \mid D_m \leq n\}$. We also assume that the Φ_m's are uniformly bounded by Φ and that $pen(m) = K_2\Phi^2 D_m/n$ with $K_2 \geq \kappa_2$. Then Theorem 2 holds with $d^2(s, S_m) \leq a_{m+1}^2$ leading to

$$\mathbb{E}\left[\|s-\hat{s}\|^2\right] \leq C(a_0)\inf_{m\in\mathcal{M}_n}\left\{a_{m+1}^2 + \frac{D_m}{n}\right\} \tag{3.1}$$

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

$$m(n) = min\left\{m \geq 0 \mid a_{m+1}^2 \leq \frac{D_m}{n}\right\} \tag{3.2}$$

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 $K \geq 1$:

$$D_{m(n)} \leq Kna_{m(n)}^2. \tag{3.3}$$

Therefore the convergence rate in (3.1) is of the order of $\sqrt{D_{m(n)}/n}$ by Lemma 5 of Section 6 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 $\bar{\mathcal{E}}(a) = \{u \in \mathcal{E}(a) \mid 1\!\!1 + u \geq 0\}$. The following proposition provides a lower bound for the minimax risk over $\bar{\mathcal{E}}(a)$. We shall then discuss on specific examples (Fourier, Haar and Sobolev ellipsoids) how far it is from the upper bound (3.1).

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

$$\sup_{\delta\in\mathcal{C}_m}\|\sum_{\lambda\in\Lambda_m}\delta_\lambda\varphi_\lambda\|_\infty \leq \Psi_m \tag{3.4}$$

and that $\Psi_{m(n)}^2 \leq n\Psi$. If $\Phi = \sup_{m\geq 0}\Phi_m$, any estimator $\tilde{s}$ satisfies

$$\sup_{s\in\bar{\mathcal{E}}(a)}\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \geq \kappa_{10}\frac{1\wedge na_0^2}{(\Phi^2/n)\vee\Psi}\sup_{m\in\mathbb{N}}\left\{\frac{D_m}{n}\wedge a_m^2\right\}. \tag{3.5}$$

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

$$\sup_{s\in\bar{\mathcal{E}}(a)}\mathbb{E}_s\left[\|s-\tilde{s}\|^2\right] \geq C(\Phi,\Psi,K_0)\left[\inf_{m\in\mathcal{M}_n}\left\{a_{m+1}^2 + \frac{D_m}{n}\right\}\wedge 1\right]. \tag{3.6}$$

If the set $\bar{\mathcal{E}}(a)$, because of the positivity requirement, is substantially smaller (from the point of view of the metric dimension) than $\mathcal{E}(a)$, there is no hope that our upper bound (3.1) be optimal since in designing it we essentially pretended that the whole of $\mathcal{E}(a)$ was the parameter space. The role of Ψ_m is to quantify this effect. If we take $\mathcal{C}_m = \{-1,+1\}^{\Lambda_m}$, Ψ_m is bounded by $r_m\sqrt{D_m}$.

Comments about the size of a_0: One should first observe that keeping a_0 bounded allows to keep $C(a_0)$ in (3.1) under control since it is a nondecreasing function of a_0 and therefore under the assumptions of Proposition 2 the bounds (3.1) and (3.6) 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 $1/\sqrt{n}$ 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 $1/n$. 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 $\Psi^2_{m(n)}/n$ is easily seen to be bounded and subsequently the rate provided by (3.1) is optimal for a fixed 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 2.3), $\Psi^2_{m(n)}/n \leq RD_{m(n)}/n \leq RKa_0^2$ by (3.3).
- Since Φ_m is bounded by Φ, by (2.2) we can always take $\Psi_m = \Phi D_m$. If moreover $\mathcal{E}(a)$ is Hilbert-Schmidt, i.e. a is such that $\sum_{j\geq 0}|\Lambda(j)|a_j^2 = \Sigma < +\infty$, then by monotonicity $D_m a_m^2 \leq \Sigma$ for any m. It follows from (3.3) that $D_{m(n)} \leq \sqrt{K\Sigma n}$ from which one derives that $\Psi^2_{m(n)} \leq \Phi^2 K\Sigma n$.

Fourier ellipsoids: When the ellipsoid is not Hilbert-Schmidt, the preceding argument breaks. We can still apply Proposition 2 with different sets $\mathcal{C}_m$. Recall that, in the case of the Fourier basis defined in Section 2.4.2, $D_m = 2(m+1)$. A classical result by Salem and Zygmund on random Fourier series [see Kahane (1985) Theorem 2 p.69] implies that there exists a subset $\mathcal{C}_m$ of $\{-1,+1\}^{\Lambda_m}$ of cardinality larger than 2^{D_m-1} such that (3.4) holds with $\Psi_m = \bar{\Psi}\sqrt{D_m\log(D_m)}$. If we assume that $a_j\sqrt{\log(j+2)}$ 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 (3.3) and so is $\Psi^2_{m(n)}/n$. Therefore (3.6) matches (3.1). 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. Similar lower bounds under the same restrictions $(\sup_j a_j^2\log(j+2) < +\infty)$ were found by Efroimovich and Pinsker (1981, 1982). They were actually able to compute not only a lower bound for the rate of convergence but even the exact asymptotic value of the minmax 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 Riemannian manifold $\mathbb{M}$ with dimension q and uniform distribution μ and recall from Section 2.3 that $\{\theta_j \mid j \geq 0\}$ is the set of eigenvalues of the Laplacian operator on $\mathbb{M}$ and $\{\varphi_\lambda \mid \lambda \in \Lambda(j)\}$ the set of eigenvectors corresponding to θ_j. We shall say that $s = \sum_{j\geq 0}\sum_{\lambda\in\Lambda(j)}\beta_\lambda\varphi_\lambda$ belongs to the Sobolev space $\mathcal{H}_\alpha(\mathbb{M})$ if and only if the coefficients $(\beta_\lambda)_{\lambda\in\Lambda}$ 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 $s \in \mathcal{E}(a)$, where $a_j = H\theta_j^{-\alpha/2}$ for all $j \geq 0$, with H and α unknown. It follows from our computations in Section 2.3 that the corresponding family $\{\Phi_m\}_{m\in\mathcal{M}_n}$ is uniformly bounded by some constant $\Phi(\mathbb{M})$ and therefore that Theorem 2 applies leading to the bound (2.16). In order to measure the effect of H on the risk we need to derive from (2.16) a sharper bound than (3.1). Choosing $pen(m) = K_2D_m/n$ with $K_2 \geq \kappa_2\Phi^2(\mathbb{M})$ we get from Theorem 2

$$\mathbb{E}[\|\hat{s}-s\|^2] \leq \kappa_2' \inf_{m\in\mathcal{M}_n}\left\{H^2\theta_{m+1}^{-\alpha} + \frac{K_2D_m}{n}\right\} + \kappa_2''\Phi^4(\mathbb{M})\frac{(1+\|s\|)^4}{n}. \tag{3.7}$$

We wish to know under which conditions this upper bound matches the lower bound (3.6) up to constants. In order to answer this question it is necessary to control $\|s\|$ when s belongs to the ellipsoid $\mathcal{E}(a)$. 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*

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

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

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

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 (2.3)

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

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

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

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 (2.9) that

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

Using (2.9) 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. $\square$

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 (3.10) 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*

$$\inf_{m\in\mathcal{M}_n}\left\{\frac{H^2}{\theta_{m+1}^\alpha}+\frac{D_m}{n}\right\} \leq 2\left[\frac{D_0}{n}+\left(\frac{C_2(\mathbb{M})\vee 1}{C_1(\mathbb{M})}\right)^{q/2}\left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)}\right] \tag{3.8}$$

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

$$\inf_{m\in\mathcal{M}_n}\left\{\frac{H^2}{\theta_{m+1}^\alpha}+\frac{D_m}{n}\right\} \geq \left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)}\left(\frac{C_1(\mathbb{M})}{C_2^{3/2}(\mathbb{M})\vee 1}\right)^q. \tag{3.9}$$

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

$$\frac{D_0\theta_1^\alpha\vee 1}{n} \leq H^2 \leq \left[\left(\frac{C_1(\mathbb{M})}{C_2(\mathbb{M})}\right)^{(2\alpha+q)/2} n^{2\alpha/q}\right]\wedge n^{(2\alpha-q)/(2q)}\wedge n, \tag{3.10}$$

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

$$\inf_{\tilde{s}}\sup_{s\in\bar{\mathcal{E}}(a)}\mathbb{E}_s[\|s-\tilde{s}\|^2] \geq C(\mathbb{M})\left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)} \tag{3.11}$$

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

$$\sup_{s\in\bar{\mathcal{E}}(a)}\mathbb{E}(\|\hat{s}-s\|^2) \leq C'(\mathbb{M})\left(\frac{H^q}{n^\alpha}\right)^{2/(2\alpha+q)}. \tag{3.12}$$

Proof: Let us first observe that it follows from Lemma 5 of Section 6 that

$$I = \inf_{m\in\mathcal{M}_n}\left\{\frac{H^2}{\theta_{m+1}^\alpha}+\frac{D_m}{n}\right\} \geq \sup_{m\geq 0}\left\{\frac{D_m}{n}\wedge\frac{H^2}{\theta_m^\alpha}\right\} = \frac{D_{m(n)}}{n}\wedge\frac{H^2}{\theta_{m(n)}^\alpha} \tag{3.13}$$

and

$$I \leq 2D_{m(n)}/n \qquad \text{provided that} \qquad D_{m(n)} \leq n \tag{3.14}$$

where $m(n)$ defined in (3.2) is given by

$$m(n) = \inf\{m\in\mathbb{N} : H^2\theta_{m+1}^{-\alpha} \leq D_m/n\}. \tag{3.15}$$

Assuming first that $m(n)\geq 1$ and noticing that (2.9) implies that

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

we derive from (3.15) that

$$\left(\frac{C_1(\mathbb{M})}{C_2(\mathbb{M})}\right)^{q/2}\frac{D_{m(n)}}{n} \leq H^2\theta_{m(n)}^{-\alpha} \leq \left(\frac{C_2(\mathbb{M})}{C_1(\mathbb{M})}\right)^\alpha\frac{D_{m(n)}}{n}. \tag{3.16}$$

Combining this with (2.9) we get

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

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

$$(nH^2)^{q/(2\alpha+q)}\left(\frac{C_1(\mathbb{M})}{C_2^2(\mathbb{M})}\right)^{q/2} \leq D_{m(n)} \leq (nH^2)^{q/(2\alpha+q)}\left(\frac{C_2(\mathbb{M})}{C_1(\mathbb{M})}\right)^{q/2}. \tag{3.17}$$

By (3.16) and (3.17) the lower bound in (3.13) 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 (3.9).

Turning our attention to (3.8), we see that it follows from (3.14) if $m(n) = 0$. When $m(n) \geq 1$ we derive from (3.17) 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 (3.8) then follows from (3.14) and (3.17).

We can now turn to a precise evaluation of the risk. Combining Lemma 1, (3.8) and (3.10) we see from (3.7) that the upper bound (3.12) holds for the risk of our estimator. On the other hand it follows from (2.3) that

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

and we can choose $\Psi_{m(n)} = \Phi^2(\mathbb{M})D_{m(n)}^2$. Consequently from (3.17) and (3.10) $\Psi_{m(n)}/n$ is bounded by a constant $C''(\mathbb{M})$. Then (3.6) combined with (3.9) imply the lower bound (3.11). □

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

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

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

$$\frac{1}{2} \leq 1\!\!1 + u \leq \frac{3}{2} \tag{3.18}$$

which a fortiori implies that $\mathcal{E}_m \subset \bar{\mathcal{E}}(a)$. It also follows from (3.18) that any pair (u, v) of elements of $\mathcal{E}_m$ satisfies

$$h^2(1\!\!1 + u, 1\!\!1 + v) \leq \frac{1}{4}\|u - v\|^2 \leq \frac{D_m}{N}$$

where h denotes the Hellinger distance and therefore the Kullback-Leibler information numbers between the probabilities corresponding to the elements of $\mathcal{E}_m$ are uniformly bounded [see inequality (6.4) of Birgé and Massart (1994a)] by $4.84D_m/N$. A classical combinatorial argument that we shall prove later for the sake of completeness [see

Lemma 8] ensures that there exists a subset $\mathcal{E}'_m$ of $\mathcal{E}_m$ of cardinality larger than $(1/2)\exp(D_m/3)$ such that

$$\|u-v\|^2 \geq 2\left[1-\sqrt{\frac{2}{3}}\right]\frac{D_m}{N} > .367\,\frac{D_m}{N} \quad \text{for all} \quad u,v\in\mathcal{E}'_m. \tag{3.19}$$

An application of Fano's Lemma [see Birgé (1986) page 279] shows that any estimator $\tilde{u}_m$ with values in $\mathcal{E}'_m$ satisfies $\sup_{u\in\mathcal{E}'_m}\mathbb{P}_u[\tilde{u}_m\neq u]\geq 1/4$ provided that

$$4.84\,\frac{nD_m}{N}+\log 2\leq\frac{3}{4}\log\left(\frac{1}{2}\exp\left(\frac{D_m}{3}\right)-1\right)$$

which is true since $D_m\geq 6$ and $N\geq 578n$. Since

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

one concludes with (3.19) that

$$\sup_{u\in\mathcal{E}'_m}\mathbb{E}_u\left[\|u-\tilde{u}_m\|^2\right]\geq\frac{.367}{4}\frac{D_m}{N}>\frac{D_m}{11N}.$$

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

$$1/(2N)\leq h^2(1\!\!1-\varphi_\lambda/\sqrt{N},1\!\!1+\varphi_\lambda/\sqrt{N})\leq 1/N$$

and Lemma 7 implies that

$$\sup_{u\in\mathcal{E}'_m}\mathbb{E}_u\left[\|u-\tilde{u}_m\|^2\right]\geq\frac{1}{4N}\left(1-\sqrt{\frac{2n}{N}}\right)>\frac{D_m}{23N}$$

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

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

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

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

Choosing $m=m(n)$, (3.5) follows from Lemma 5. (3.6) also follows from Lemma 5 provided that $D_{m(n)}\leq n$ which implies that $m(n)\in\mathcal{M}_n$. 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 (3.5) is a constant otherwise $m(n)-1\in\mathcal{M}_n$ and

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

Therefore (3.6) holds in both cases. $\square$

3.1.2 Densities with an unknown modulus of continuity

Let ω be a modulus of continuity i.e. is a subadditive continuous nondecreasing and nonnegative function defined on $[0,1]$ with $\omega(0)=0$ [see DeVore and Lorentz (1993) page 41 for details]. Let $\mathcal{F}_\omega$ denote the set of functions $s \in \mathcal{S}$ 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 $\mathcal{F}_\omega$ for some unknown ω. We want to show here that, using a penalized maximum likelihood procedure over the family of regular histograms, it is possible to estimate s without knowing ω as well (up to multiplicative constants) as if ω were known.

Let us choose $\mathcal{M}_n = \{2,\ldots,n\}$ 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

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

We want to apply Theorem 1. The family of sieves S_m, $m \geq 2$ satisfies (2.11) with $L_m = 1$ and $\bar{r}_m \leq 1$. It remains to control the bias term $K(s,S_m) \wedge 1$. Let s_m^+ be defined as follows:

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

Then $s_m^+ \geq s$ and using the fact that ω is nondecreasing one can check that $d(s,s_m^+) \leq \omega(1/m)$. It therefore comes from Proposition 1 and Theorem 1 that if $pen(m) = K_1 D_m/n$ with $K_1 \geq (1+\log 2)\kappa_1$,

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

where $\hat{s}$ denotes the penalized maximum likelihood estimator and one can conclude that

$$\mathbb{E}[d^2(s,\hat{s})] \leq 2 \wedge \left[K_1' \inf_{m \in \mathcal{M}_n} \left\{ \omega^2\left(\frac{1}{m}\right) + \frac{m}{n} \right\} \right] \tag{3.20}$$

since $d^2(s,\hat{s})$ is always bounded by 2.

Let us now find a lower bound for the minimax risk over $\mathcal{F}_\omega$. The proof will follow 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.

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

$$\sup_{s \in \mathcal{F}_\omega} \mathbb{E}_s\left[d^2(s,\tilde{s})\right] \geq \frac{\kappa_0' n\omega^2(1/2)}{1+n\omega^2(1/2)} \left[\inf_{m \in \mathcal{M}_n} \left\{ \omega^2\left(\frac{1}{m}\right) + \frac{m}{n} \right\} \wedge 1 \right]. \tag{3.21}$$

Proof: 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 shall define by v_0 and v_1 respectively the functions given by

$$\begin{aligned} v_0(x) &= \eta v(x)\mathbb{1}_{[0,2\delta)}(x) - v(x-2\delta)\mathbb{1}_{[2\delta,4\delta)}(x); \\ v_1(x) &= -v(x)\mathbb{1}_{[0,2\delta)}(x) + \eta v(x-2\delta)\mathbb{1}_{[2\delta,4\delta)}(x). \end{aligned}$$

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

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

Then $1\!\!1 + v_0$ and $1\!\!1 + v_1$ are nonnegative functions (since $v(\delta) \leq 1$) of norm 1. For any $\varepsilon \in \{0;1\}^m$ the function s_ε 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 $\mathcal{F}_\omega$ because of our choice of v. If all coordinates of ε and ε' 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 i.i.d. observations

$$\sup_{\varepsilon\in\{0;1\}^m} \mathbb{E}_{s_\varepsilon}\left[d^2(s_\varepsilon, \tilde{s})\right] \geq \frac{\beta}{16\delta}\left[1-\sqrt{2n\beta}\right] \qquad \text{with } \beta = \frac{\delta\omega^2(2\delta)(1+\eta)^2}{6}. \tag{3.22}$$

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

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

Then $\omega^2[1/(2m)] \leq 2$ 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 (3.22)

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

In order to derive (3.21) 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 $m_0 \in \mathcal{M}_n$. If $m(n) = 1$ taking $m_0 = 2$ gives (3.21). Otherwise if $\omega^2(1/(2n)) \leq 2$, $m(n) \leq n$ and $m(n) \in \mathcal{M}_n$. If $m(n) \geq 2$ we choose $m_0 = m(n)$. It then follows from the definition of $m(n) = m_0$ that

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

and from the subadditivity and monotonicity of ω [see (6.5) page 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 (3.21) again. Finally if $m(n) \geq n+1$ the same argument shows that $\omega^2(1/(2m(n)) > 1$ which concludes the proof. $\square$

Remarks:

- Comparing (3.21) with the upper bound in (3.20) one sees that the bounds match except when $n\omega^2(1/2)$ is too small which means that the whole of $\mathcal{F}_\omega$ is so close to the function 1 that a good procedure would be to ignore the observations and choose $\tilde{s} = 1$ as the estimator. This would result in a minimax risk of order $\omega^2(1/2)$ smaller than $1/n$. With the number of observations at hand, the parameter space $\mathcal{F}_\omega$ essentially behaves like a single point and the estimation problem is not really meaningful. Nevertheless it would be easy to check that a suitable modification of our estimator would solve the problem: just add the additional sieve $S_1 = \{1\!\!1\}$ with the penalty $pen(1) = 0$.

- One should keep in mind that although our computations were performed for $s \in \mathcal{F}_\omega$, the upper bounds results would make sense for any s and in particular if for some $m, s \in S_m$, the rate of convergence of our estimator will be the parametric one, i.e. $1/n$, since then $K(s, S_m) = 0$.

- In the Hölderian case considered in Section 2.4.1 the modulus of continuity is given by $\omega(x) = Hx^\alpha$ 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

$$|f(x) - f(y)| \leq H|x - y|^\alpha, \qquad \text{for all } \ x, y \in [0, 1], \tag{3.23}$$

  the resulting optimal rate of convergence when the loss is the square of the Hellinger distance will be $n^{-2\alpha/(2\alpha+1)}$ 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 (3.23) 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 (3.23) with α 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 absolutely 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.

3.1.3 Hölderian densities with unknown anisotropic smoothness

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

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

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

We again consider the problem of estimating an unknown element $s \in \mathcal{S}$ by a penalized maximum likelihood estimator method and we want to address the multidimensional case and 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 (1993 and 1995) or Goldenshluger and Nemirovski (1994)]. For the sake of simplicity we shall only consider densities with respect to Lebesgue measure μ 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 $\mathcal{F}(\underline{\alpha}, \underline{H})$ to be the subset of those $s \in \mathcal{S}$ such that the univariate functions $y \mapsto s(x_1, \ldots, x_{i-1}, y, x_{i+1}, \ldots, x_q)$ belong to $\mathcal{H}(H_i, \alpha_i)$ for all x and i where $\mathcal{H}(H, \alpha)$ has been defined by (2.19).

Following the notations of Section 2.3 a space of piecewise polynomials is characterized by its maximal degree r and a partition of $[0,1]^q$. Let $\mathcal{R}(N)$ denote the regular partition of $[0,1]$ with N pieces. We define $\mathcal{M}_n$ as the set of all $m = (r, \mathcal{R}(N_1), \ldots, \mathcal{R}(N_q))$ 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 $S_m = \bar{S}_m \cap \mathcal{S}$.

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

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

and assume that for any i

$$n^\alpha H_i^{2\alpha+q} \geq H^q. \tag{3.24}$$

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

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

Proof: We want to apply Theorem 1 to our model. Clearly (2.11) is satisfied with $L_m = 1$ and

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

since $\prod_{i=1}^q [(r+1)N_i] \geq r + (\sum_{i=1}^q N_i)/q$. It also follows from (2.8) that $\bar{r}_m \leq (2r+1)^q$ which justifies our choice for $pen(m)$.

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 (2.13). 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 $f \in \mathcal{F}(\underline{\alpha}, \underline{H})$ by a polynomial of degree $\leq r = \sup_{1 \leq i \leq q}(a_i)$ where a_i is the largest integer smaller than α_i. It follows from Dahmen, DeVore and Scherer (1980) Corollary 3.1 and Schumaker (1981) [see inequality (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

$$\omega_i(\delta_i) = \sup_x \sup_{|h_i| \le \delta_i} \left| \frac{\partial^{a_i}}{\partial x_i^{a_i}} f(x_1, \ldots, x_i + h_i, \ldots, x_q) - \frac{\partial^{a_i}}{\partial x_i^{a_i}} f(x_1, \ldots, x_i, \ldots, x_q) \right|.$$

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

$$\|f - P\|_\infty \le C'(q,r) \sum_{i=1}^q H_i \delta_i^{\alpha_i}. \tag{3.25}$$

Let us define

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

and N_i to be the integer such that $\delta_i^{-1} \le N_i < \delta_i^{-1} + 1$. It follows from (3.24) that $N_i \le 2/\delta_i$. Let $m = (r, \mathcal{R}(N_1), \ldots, \mathcal{R}(N_q))$ and $\bar{S}_m$ be the corresponding space of piecewise polynomials. (3.25) implies that there exists an element $\bar{s}_m \in \bar{S}_m$ such that

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

Therefore Theorem 1 and (2.13) imply (since $N_i \le 2/\delta_i$) that

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

and the conclusion follows from our choice of the δ_i's and η. □

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 $\underline{\alpha}$. 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 $\alpha \le \alpha_0$. Apart from the special situation of Fourier expansions in the white noise model [Efroimovich and Pinsker (1984)], we do not know of any other result of this type valid for any value of α.

3.1.4 Estimation of the support of a distribution

We observe n i.i.d. random variables $Z_1, \ldots, Z_n$ of unknown distribution μ which is absolutely continuous with respect to the uniform distribution on the unit disk $\mathbb{D}$ of $\mathbb{R}^2$ with a density bounded by some known constant Ψ and we want to estimate the indicator function s of the support Ω_s of μ. Some results in this direction, but from a non-adaptive point of view, can be found in Korostelev and Tsybakov (1993a). We define $\mathcal{S}_n$ as the set of indicator functions of measurable sets in the unit disk $\mathbb{D}$ and the contrast function by $\gamma(z,t) = -t(z)$. It satisfies

$$\mathbb{E}[\gamma(.,t) - \gamma(.,s)] = \|t - s\|_1 = \|t - s\|^2. \tag{3.26}$$

In order to define our sieves we restrict ourselves 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 $\theta_g(\rho \sin \psi, \rho \cos \psi) = 1\!\!1_{\{\rho^2 \leq \tilde{g}(\psi)\}}(\rho, \psi)$. The reader should notice here that we introduce an unusual parametrization of the boundary of the form $\rho^2 = \tilde{g}(\psi)$. We define μ' to be the uniform distribution on $\mathbb{T}$. Since the density of μ is bounded by Ψ, one can easily check that

$$\|\theta_{g_1} - \theta_{g_2}\|_1 \leq \frac{\Psi}{2}\|\tilde{g}_1 - \tilde{g}_2\|_1 \leq \frac{\Psi}{2}\|g_1 - g_2\|_1. \tag{3.27}$$

Let $\mathcal{G}_n$ be the set of measurable functions g from $\mathbb{T}$ to $\mathbb{R}$. Then θ maps $\mathcal{G}_n$ into $\mathcal{S}_n$. Let $\bar{G}_m$ be a linear subspace with dimension D_m of $\mathbb{L}^1(\mu')$ and $G_m \subset \{g \in \bar{G}_m \,|\, \|g\|_1 \leq 2\}$. S_m is defined as the image of $G_m \subset \mathcal{G}_n$ by the mapping θ. Then the following theorem to be proved in Section 5 holds

Theorem 5 *Let $\{L_m\}_{m \in \mathcal{M}_n}$ be a family of weights satisfying Assumption **S**. 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 $\bar{G}_m$ with $\|\varphi_\lambda\|_1 = 1$ for all λ and*

$$\sum_{\lambda \in \Lambda_m} |\beta_\lambda| \leq B''_m \|\sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda\|_1 \quad \textit{for all } (\beta_\lambda) \in \mathbb{R}^{\Lambda_m}. \tag{3.28}$$

Let κ_5 be a suitable positive numerical constant,

$$pen(m) \geq \kappa_5[L_m + \log(1 + nB''_m\Psi/D_m)]D_m/n$$

and $\hat{s}$ be the penalized minimum contrast estimator which is a minimizer with respect to $m \in \mathcal{M}_n$ and $t \in S_m$ of $pen(m) - n^{-1}\sum_{i=1}^n t(X_i)$. If Ω_s is starshaped with $s = \theta_f$ and $0 \leq f \leq 1$ then

$$\mathbb{E}[d^2(s, \hat{s})] \leq \kappa'_5 \inf_{m \in \mathcal{M}_n} [\Psi d_1(f, G_m) + pen(m)]$$

where d_1 denotes the $\mathbb{L}^1(\mu')$ distance.

Remark: Assuming without loss of generality that $0 \in G_m$, we see that $d_1(f, G_m) \leq \|f\|_1 \leq 1$ which shows that $G_m \subset \{g \,|\, \|g\|_1 \leq 2\}$ is a natural restriction.

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 $\varphi_0 = 1\!\!1$, $\varphi_{2j-1}(x) = (\pi/2)\cos(jx)$, $\varphi_{2j}(x) = (\pi/2)\sin(jx)$ 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 $G_m \subset \{g \in \bar{G}_m \,|\, \|g\|_1 \leq 2\}$. Then $D_m = 2m+1$ and we take $L_m = 1$. We can check (3.28) exactly as we did for (2.23), now using inequality (2.15) p.102 of De Vore and Lorentz (1993). We get

$$\sum_{\lambda \in \Lambda_m} |\beta_\lambda| \leq D_m \|\sum_{\lambda \in \Lambda_m} \beta_\lambda \varphi_\lambda\|_1$$

and therefore (3.28) is satisfied with $B''_m = D_m$. This leads to the choice $pen(m) = K_5[1 + \log(1 + n\Psi)](2m+1)/n$ with $K_5 \geq \kappa_5$. If g belongs to some Besov space $B_{\alpha,1,\infty}$ (see the precise definition in Lemma 13 below) of functions on $\mathbb{T}$, it follows from Lemma 13 that $d_1(f, G_m) \leq C(f)m^{-\alpha}$ 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 α 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 $\{B''_m\}_{m\in\mathcal{M}_n}$.

3.2 "Rich" families of sieves

By this we mean families for which the number of models of a given dimension D is so large that the summability condition **S** 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 $\sum_{m\in\mathcal{M}_n}\exp[-L_mD_m]$ 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.

3.2.1 Histograms with variable binwidths and spatial adaptation

Let's go back to maximum likelihood estimation with n i.i.d. 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 very rich family described in Section 2.4.1 with $\mathcal{M}_n=\mathcal{R}_n\cup(\cup_{N\geq 2}\mathcal{G}_{n,N})$. It has been mentioned already that the corresponding penalized maximum likelihood estimator had the right rate of convergence if the true s was α-Hölderian with index $\alpha\in(0,1]$. Let us now assume that s has a bounded α-variation with $0<\alpha\leq 1$ which means that

$$\sup_{k\geq 2}\ \sup_{x_1\leq\cdots\leq x_k}\ \sum_{j=2}^{k}|s(x_{j-1})-s(x_j)|^{1/\alpha}=J_\alpha(s)<+\infty \tag{3.29}$$

where the supremum is taken over all increasing sequence $x_1\leq\cdots\leq x_k$ of points in $[0,1]$. It follows from Lemma 12 and Proposition 1 that if $N\geq 2$ and $1\leq L\leq N$ there exists some $m\in\mathcal{G}_{n,N}$ such that $D_m\leq 2(N/L)^{1/(1+2\alpha)}+1$ and

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

Since one can only assume that $L_m\leq 2+3\log(N/D_m)\leq 2+3\log N$ and $\bar{r}_m\leq\sqrt{N/D_m}$, Theorem 1 implies that if $pen(m)$ is chosen as in (2.15),

$$\begin{aligned}\mathbb{E}[d^2(s,\hat{s})] \ &\leq\\ & K'_1\left[\inf_{N\geq 2}\ \inf_{1\leq L\leq N}\left\{J_\alpha^{2\alpha}(s)\left(\frac{L}{N}\right)^{2\alpha/(2\alpha+1)}+\frac{\log N}{n}\left(\frac{N}{L}\right)^{1/(2\alpha+1)}\right\}\wedge 1\right].\end{aligned} \tag{3.30}$$

Assuming that $J_\alpha^{2\alpha}(s)\geq 4/n$ we evaluate the bound at $N=[nJ_\alpha^{2\alpha}(s)]$. Then $L=J_\alpha^{-2\alpha}(s)N\log N/n$ satisfies $1\leq L\leq N$ since $N\geq 4$ and we get from (3.30) a risk bound of order

$$\left(\frac{J_\alpha(s)\log n}{n}\right)^{2\alpha/(2\alpha+1)} \wedge 1. \tag{3.31}$$

Remarks

- One should always keep in mind that whatever the true function s the right-hand side of (2.12) provides the best compromise, among all the histograms at hand, between $K(s, S_m)$ and $D_m(\log n)/n$ even when s does not belong to the particular smoothness classes considered above. But unless one makes precise assumptions on s it is not possible to compute the resulting rate of convergence.

- 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 many technicalities and for the sake of simplicity we shall not insist on this here.

3.2.2 Neural nets and related nonlinear models

We assume now the situation described in Section 2.2.2. Risk bounds for minimum penalized 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.

In keeping with the general framework of Section 2.1, 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 μ 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 i.i.d. centered errors and with target function s bounded by a known constant ξ. 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 (2.4) 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 ϕ_w is bounded by one (which handles most of the cases of interest) or that in some cases ϕ_0 is identically 0 so that then $|\phi_w(x)| \leq |w|_1$ by the Lipschitz condition.

Theorem 6 *Let $\{\phi_w : w \in \mathbb{R}^{q'}\}$ be a parameterized family of functions that satisfies the Lipschitz condition $\|\phi_w - \phi_{w'}\|_\infty \leq |w - w'|_1$ and suppose that $\|\phi_w\|_\infty \leq 1 \vee |w|_1$.*

- *For maximum likelihood density estimation we define*

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

and take

$$pen(m) \geq \kappa_6 \frac{D'q'}{n}\left[1 + \log\left(RH\left(1 + \frac{n}{D'q'}\right)\right)\right] \tag{3.32}$$

where κ_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 $D'(q'+1) \leq n$ and $t \in S_m$ of $-(1/n)\sum_{i=1}^n \log(t(X_i)) + pen(m)$, then

$$\begin{aligned} \mathbb{E}[d^2(s,\hat{s})] &\leq \kappa_6' \left[\inf_{D',H,R}\{K(s,S_m) + pen(m)\} \wedge 1\right] && (3.33) \\ &\leq \kappa_6'' \inf_{D',H,R}\{[d^2(s,\bar{S}_m) \vee n^{-2}][1+\log(n\|s\|_\infty)] + pen(m)\}. && (3.34) \end{aligned}$$

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

$$pen(m) \geq \kappa_7(\xi+\xi')^2 \frac{D'q'}{n}\left[1 + \log\left(RH\left(1+\frac{n}{D'q'}\right)\right)\right] \quad (3.35)$$

where κ_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 $(1/n)\sum_{i=1}^n (Y_i - t(X_i))^2 + pen(m)$, then

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

Statistical rate bounds using multivariate nonlinear additive ridge models: We now restrict to the case where the function ϕ_w is a ridge function on $\mathbb{R}^q$: $\phi_w(x) = \psi(a^Tx+b)$ 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 6.

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

$$s(x) = \int \exp\{ia^Tx\}\tilde{F}(da)$$

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 γ-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 α and the ridge function ψ satisfy the following constraints:

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

Then in each case, provided that m is such that $R \geq R(s)$ and $H \geq H_0$

$$d(s,\bar{S}_m) \leq c_{s,\alpha}\delta_H + R(s)/\sqrt{D'} \quad (3.36)$$

where δ_H does not depend on s and decreases at least polynomially fast with respect to H as H goes to infinity.

We now assume that ψ 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 6 with a value of $pen(m)$ of the order of the lower bound given by (3.32) or (3.35) and Proposition 6. Let $\hat{s}$ be the minimum penalized contrast estimator, taking the minimum over D', H, and R as in Theorem 6. 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 $1/\sqrt{D'}$. Then optimizing over D', we conclude that the risk $\mathbb{E}[d^2(s, \hat{s})]$ is of order $\log n/\sqrt{n}$ or $\sqrt{\log n/n}$ 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).

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 α 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. The principle dependence on the dimension is indirectly 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 $(1/D)^{1/q}$. 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 $\sqrt{1/D}$, comparable to the approximation rate of the much smaller subclass of functions that have bounds on all derivatives up to a certain high order.

3.2.3 Model selection with a bounded basis

We want to do density estimation using projection estimation as described in Section 2.4.2 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 μ. With such a basis (which is bonded by $\sqrt{2}$), assuming that we have a precise idea of an upper bound Φ' for $1\!\!1 + s$, we can apply Theorem 7 (with the Assumptions **LBB**) to be stated below. We shall look for a representation of s with a small number of parameters (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 smaller than 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 7 involves in this case a summability condition of the type $\sum_{m\in\mathcal{M}_n}\exp[-C(L_mD_m\wedge\sqrt{n})]$ which is more restrictive than **S**. We take $\Lambda_m = m$ and $L_m = l\log n\Phi'/\kappa$. Since $\left(\begin{array}{c} n^l \\ i \end{array}\right) \leq n^{li}/i!$ we can bound the first term of (5.18) below by

$$\sum_{i=1}^{K_n}\left(\begin{array}{c} n^l \\ i \end{array}\right)\exp(-li\log n) \leq \sum_{i=1}^{\infty}\frac{1}{i!}.$$

and the second term by

$$\sum_{i=1}^{K_n}\left(\begin{array}{c} n^l \\ i \end{array}\right)\exp\left(-\frac{\kappa\sqrt{n}}{2}\right) \leq en^{lK_n}\exp\left(-\frac{\kappa\sqrt{n}}{2}\right)$$

which is bounded from our choice of K_n. With the choice $pen(m) = K_6L_mD_m/n$ for a large enough constant K_6, our penalized projection estimator provides (up to a $\log n$ factor due to our choice of L_m) a risk which realizes the best trade-off between bias and variance among our family of sieves. 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 $\log n$ 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 (1994b).

4 Adaptation versus Model Selection

Although this terminology is widely used, we do not know of any general definition of *adaptation*. Usually one has to deal with a problem involving an unknown parameter θ (an index of regularity for instance) and a family of estimators $\hat{s}_m$ depending on a tuning parameter m (kernel density estimators with bandwidth m for instance) which should ideally be chosen as $m(n,\theta)$ when θ obtains. Since the true value of θ is unknown, one tries to find an estimator $\hat{m}$ of $m(n,\theta)$ based on the observations such that estimation based on $\hat{m}$ will be as good or almost as good as estimation based on the knowledge of the true $m(n,\theta)$. This is the general situation for cross-validation in kernel estimation when one wants to adjust the bandwidth.

To be more formal, let us say that one approach to adaptation would be as follows. 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 the density s (or s^2) or $X^{(n)} = \{X_{t,n}\}_{0\leq t\leq 1}$ is given by the white noise model

$$dX_{t,n} = s(t)\,dt + \frac{1}{n}\,dW_t$$

among other models (regression function, spectral density estimation,...). Given a loss function ℓ and an estimator $\hat{s}_n(X^{(n)})$ depending on the observation, the risk

$R_n(\hat{s}_n, \mathbb{S})$ of this estimator relative to some parameter space $\mathbb{S}$ and the minimax risk $R_n(\mathbb{S})$ over $\mathbb{S}$ are respectively defined by

$$R_n(\hat{s}_n, \mathbb{S}) = \sup_{u \in \mathbb{S}} \mathbb{E}_u[\ell(u, \hat{s}_n)] \qquad \text{and} \qquad R_n(\mathbb{S}) = \inf_{\hat{s}_n} R_n(\hat{s}_n, \mathbb{S}).$$

Let us now assume that we are given a family $\{\mathbb{S}_\theta\}_{\theta \in \Theta}$ of parameter spaces. One shall speak of adaptation if there exists some sequence $\tilde{s}_n$ of estimators independent of θ such that the ratios

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

are bounded independently of n by $C(\theta)$. One could as well require stronger properties of the sequence $\tilde{s}_n$ such as $C(\theta)$ being independent of θ or $\lim_{n \to +\infty} C_n(\theta) = 1$ [see for instance among others Efroimovitch (1985), Efroimovich and Pinsker (1984) and (1986), Golubev (1992), Polyak and Tsybakov (1990) or Lepskii (1991)]. One could also weaken this condition to approximate adaptation when $C_n(\theta)$ is a slowly varying function. See for instance, Lepskii (1992), Donoho, Johnstone, Kerkyacharian and Picard (1995), Goldenshluger and Nemirovski (1994) and the references therein. This means that, in a more or less strong sense, one can do as well not knowing to which $\mathbb{S}_\theta$ s belongs that knowing it. A very classical example could be the following : $\mathbb{S}_\theta$ is the set of all densities f on $[0,1]$ (with respect to Lebesgue measure) which satisfy a Lipschitz condition :

$$|f(x) - f(y)| \leq H|x - y|^\alpha, \qquad \text{forall } \; x, y \in [0,1]$$

and $\theta = (H, \alpha), H > 0, 0 < \alpha \leq 1$. One wants to estimate f using a kernel estimator of a given form but with bandwidth m to be chosen from the data.

One can also describe the preceding situation in a slightly different but equivalent way. One assumes that the true parameter s (density for instance) belongs to some non compact parameter space $\mathbb{S}$ on which the minimax risk is irrelevant (because it is too large or even infinite) but $\mathbb{S}$ can be viewed as a union of smaller compact sets $\mathbb{S}_\theta$ for which the minimax risk can be controlled. One will look for estimators $\tilde{s}_n$ such that $\mathbb{E}_s[\ell(\tilde{s}_n, s)] \leq C(\theta) R_n(\mathbb{S}_\theta)$ whenever s belongs to $\mathbb{S}_\theta$. 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 $\{\mathbb{S}_\theta\}$ if only $\mathbb{S}$ is given [think of $\mathbb{S} = \mathcal{C}([0,1])$]? 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? Not all families will do as shown by the following example: the observations $X_1, \ldots, X_n$ are i.i.d. with unknown density s belonging to $\mathbb{S} = \mathcal{C}([0,1])$ 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 clearly no hope to get an adaptive estimator in such a situation since the family $\{\mathbb{S}_\theta\}_{\theta \in \theta}$ is obviously too large.

A natural approach to solve the problem of adaptation is to consider a family of estimators $\{\hat{s}_{n,m}\}_{m \in \mathcal{M}_n}$, $\hat{s}_{n,m(n,\theta)}$ being tuned for estimation in $\mathbb{S}_\theta$. One then introduces a data-driven choice $\hat{m}$ of m in order to get the final estimator $\hat{s}_{n,\hat{m}}$. For instance one can build the required estimator $\hat{s}_{n,m}$ as a *minimum contrast estimator on a sieve* S_m using the method described in Birgé and Massart (1994a). This means that one chooses a convenient *contrast function* and for each θ an approximating space $S_{n,\theta}$ (the *sieve*). Setting $\{S_m\}_{m\in\mathcal{M}_n} = \{S_{n,\theta}\}_{\theta\in\Theta}$ we assume that the tuning $m = m(n,\theta)$ implies that the minimum contrast estimator on S_m which is defined by performing the minimization, over the space $S_{n,\theta}$, of the empirical expectation of the contrast function achieves approximately (up to constants...) the minimax risk on $\mathbb{S}_\theta$. The main problem is then again to define $\hat{m}$.

One should also keep in mind that, since the "true" parameter value s is definitely unknown, there is no reason that any of the "potentially true" models $\mathbb{S}_\theta$ should be more accurate that the best approximating one in the family $\{S_{n,\theta}\}_{\theta\in\Theta}$. Moreover, due to the classical trade-off between the bias and the random component of the error, it is often true that better performance is obtained using a smaller set $\mathbb{S}_\theta$ as the parameter set for the estimation procedure rather than the true one, $\mathbb{S}_{\theta'}$, which contains s but is much too big. Therefore, our approach is simply to forget about the true parameter space $\mathbb{S}$ and the "potentially true" models $\mathbb{S}_\theta$, but rather start with a nice family $\{S_m\}_{m\in\mathcal{M}_n}$ of approximate models. We do not postulate that the true s belongs to any of them, not even that they are good approximations to the truth. We can only hope that this will be the case, just as we would hope in a Bayesian setting that our chosen prior will be a good prior. Assuming that our approximating spaces S_m are finite-dimensional with respective dimensions D_m and that we use a quadratic loss based on some convenient distance d, the minimax risk $R_n(S_m)$ is usually of order D_m/n [this can be checked in many specific situations going from the classical regular parametric model to the finite-dimensional models, in the metric sense, introduced by Le Cam (1973), see also Le Cam (1986) and Le Cam and Yang (1990)]. In such a case, the best we can hope when estimating s with the approximate model S_m is to get a risk of order $d^2(s,S_m) + D_m/n$. This is actually what is achieved by the minimum contrast estimators on S_m (with possibly some extra $\log n$ factor in some cases) as proved in Birgé and Massart (1994a). In such a framework, an *Ideal Model Selection Procedure* would choose $m = m_n(s)$ as a minimizer of $d^2(s,S_m) + D_m/n$. This is, roughly speaking, what our penalized estimators tend to do (up to some extra $\log n$ factor in some situations), provided that the family $\{S_m\}_{m\in\mathcal{M}_n}$ has been well-chosen. Note that even if s belongs to some S_{m_0}, the index $m_n(s)$ of the *Ideal Model* might very well be different from m_0. This merely means that the additional bias $d^2(s,S_{m_n(s)})$ is smaller than the variance reduction $(D_{m_0} - D_{m_n(s)})/n$. This is, from a practical point of view, much more realistic than trying to guess to which S_m the true s belongs. This is even more obvious if s does not belong to any of the models in the list.

On the other hand, if $\mathcal{M}_n$ (and therefore the family of S_m's) is independent of n, if Assumption **S** holds with $L_m \equiv 1$ and s belongs to some S_m, one gets the usual $1/n$ (parametric) rate of convergence of the quadratic risk. If s is very close to a model S_m of low dimension D_m, the same $1/n$-rate remains true for moderate values of n. Furthermore, as seen from the examples, a convenient choice of the collection of models also provides adaptation over large nonparametric sets of smooth functions. This means that the same method can perform simultaneously a selection among a family of parametric models and adaptation for nonparametric smoothness classes. It therefore provides a smooth and flexible link between parametric and nonparametric estimation.

As shown by our examples, a good choice of the family $\{S_m\}_{m\in\mathcal{M}_n}$ (with respect to the family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$) leads to *Adaptation* or *Approximate Adaptation* in the above sense, but *Model Selection* provides more flexibility since s is never assumed to belong to one of the S_m's and Assumption **S** provides a criterion for the existence of an optimal (if the L_m's are uniformly bounded) or nearly optimal (if they are bounded by some power of $\log n$) selection procedure. On the other hand, the choice of the family $\{S_m\}_{m\in\mathcal{M}_n}$ is not more arbitrary than the choice of a family $\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ of potentially true models. In both cases, the choice of the family ($\{\mathbb{S}_\theta\}_{\theta\in\Theta}$ or $\{S_m\}_{m\in\mathcal{M}_n}$), although it has a different meaning in each case, reflects our "a priori" information about s or our "belief" about the true state of nature, to put it in a Bayesian language.

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 exists many choices of weights L_m satisfying **S**. 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. It follows from our evaluations of the risk of the penalized minimum 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. Such a strategy has been illustrated by the example of adaptive histograms in Section 2.4.1.

This particular example gives us the opportunity to develop some connections between *Model Selection* and what is now called *Spatial Adaptation* [see for instance Donoho and Johnstone (1994a and 1995)]. Further illustrations are given in Birgé and Massart (1994b). A typical function with spatially homogeneous regularity is an element of some Hölder space $\mathcal{H}(H,\alpha)$ as defined by (2.19). In order to approximate such functions, piecewise polynomials based on a regular partition are particularly well suited. The simplest case occurs when $0 < \alpha \leq 1$ and one uses histograms. Let S_m be the space of regular histograms with m pieces, then for any p with $1 \leq p \leq \infty$ the following holds:

$$\sup_{t\in\mathcal{H}(H,\alpha)} \inf_{u\in S_m} \|t-u\|_p \leq C_p(H)m^{-\alpha}.$$

For our problem, the only relevant case is $p = 2$. Under such a restriction, the class of functions that can be approximated at the same rate by functions in S_m is actually larger than $\mathcal{H}(H,\alpha)$. It can actually be proved that this saturation class of the family of regular histograms is the Besov space $B_{\alpha,2,\infty}$ as proved in De Vore and Lorentz (1993) Theorem 2.4 page 358. It is important to notice that this saturation class depends in a crucial way on the norm ($\mathbb{L}^2$-norm in our case) used for the approximation. This means that $\mathcal{H}(H,\alpha)$ is well-approximated by the family $\{S_m\}_{m\geq 1}$ for any $\mathbb{L}^p$-norm, while the same $\mathbb{L}^2$-rate of approximation will still hold for functions of the larger space $B_{\alpha,2,\infty}$ which have a moderately inhomogeneous regularity. If the regularity of s is too spatially inhomogeneous, for instance if $s \in B_{\alpha,p,\infty} \setminus B_{\alpha,2,\infty}$ for $p < 2$, the family $\{S_m\}_{m\geq 1}$ will not have the desired $\mathbb{L}^2$-approximation properties but the larger class of histograms with D pieces based on irregular partition will do the job. Choosing a convenient partition (which should ideally depend on s) from the data is called *Spatial Adaptation.* If our list of sieves includes enough irregular partitions, one could recover the right bias (rate of approximation) $d^2(s, S_m) \le C(H)m^{-\alpha}$ for some well-chosen $m \in \mathcal{M}_n$. The price to pay is a much longer list $\mathcal{M}_n$ which leads to $L_m = \log n$, at least for most of the values of m instead of $L_m = 1$. This introduces an extra $\log n$ factor in the risk.

Finally, another advantage of the flexibility of the weights in *Model Selection* is the following: assume that we have at hand several lists of models $\mathcal{M}_{n,j}$ for $j \in J$, each one corresponding to a particular family $\{\mathbb{S}_\theta\}_{\theta\in\Theta_j}$ of potentially true models i.e. to some assumption on the "true state of nature". Rather than choosing one particular Θ_j or some family $\mathcal{M}_{n,j}$ one could just mix all the models in a larger list by a suitable modification of the weights. If the initial number of choices J is finite, one could even use the previous weights without change, or with minor changes that would not affect the rates of convergence.

5 General framework and main results

5.1 The assumptions

We shall now set up the general assumptions that are used in the proofs. They actually lead to more general versions of the theorems which were stated in Section 2. The presentation mainly follows Birgé and Massart (1994a) with a set of assumptions relative to a modified contrast function $\tilde{\gamma}_m$, possibly depending on the sieve and which is related to γ and d by mean of assumption **C** below. The family of sieves and the functions $\tilde{\gamma}_m$ might depend on n but, in order to get asymptotic results, we assume that all the constants involved in the following assumptions (unless otherwise stated) are independent of n. The dependence with respect to the sieves will be marked by subscripts m or m'.

The first assumption is essentially needed to get the final conclusion by an analogue of Lemma 1 of Birgé and Massart (1994b).

C (Closing argument) *There exists constants $k > 0, k_1 \ge 0$ and for each $m \in \mathcal{M}_n$ a distinguished point $s_m \in S_m$ and a random variable U_m with finite expectation (depending on s, s_m and D_m/n but not on t) such that for all $m, m' \in \mathcal{M}_n$ and all $t \in S_{m'}$ with $\gamma_n(t) + pen(m') \le \gamma_n(s_m) + pen(m)$ then*

$$\nu_n[\tilde{\gamma}_m(., s_m) - \tilde{\gamma}_{m'}(., t)] \ge 2k\left(d^2(s,t) - U_m^2 - k_1 D_{m'}/n\right) - pen(m) + pen(m'). \quad (5.1)$$

In most situations, if $s \in \mathcal{S}_n$, checking **C** amounts to check the following simpler assumption which implies **C** with $k = k'/2$, $k_1 = 0$ and $U_m^2 = (k''/k')d^2(s, s_m)$:

C' *The function $\tilde{\gamma}_m = \tilde{\gamma}$ is independent of m and there exists functions ψ_1 and ψ_2 such that $\tilde{\gamma}(z,t) = \gamma(z,t) + \psi_1(t) + \psi_2(z)$ and constants k', k'' such that*

$$k'd^2(s,t) \le \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[\gamma(Z_i,t) - \gamma(Z_i,s)] \le k''d^2(s,t) \quad \text{for all} \quad t \in \mathcal{S}_n.$$

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

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

The conclusion follows in this case since

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

The second set of assumptions will take different forms according to the structure of the contrast and sieves. In the standard situation the assumptions have a purely metric form which is the following

Lip (Lipschitz) *There exists measurable functions $\Delta_{m,m'}(.,u,v)$ defined on $\mathcal{X}$ for each pair $(u,v)\in S_m\times S_{m'}$ and $M(.)$ defined on $\mathcal{W}$ such that if $z=(w,x)$,*

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

and the moments of $\tilde{\gamma}$ can be bounded in one of the two following ways:

One can find positive constants A, B, E such that for all $j\geq 2, u$ in S_m, v in $S_{m'}$, either

i)

$$\|M(W_i)\|_\infty \leq A^j \qquad \textit{for all } \; i=1,\dots,n; \tag{5.2}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Delta^j_{m,m'}(X_i,u,v)] \leq \frac{j!}{2}B^{j-2}\left[d^2(u,v)+E\,\frac{D_m\vee D_{m'}}{n}1\!\!1_{\{m\neq m'\}}\right]. \tag{5.3}$$

or

ii)

$$\mathbb{E}[M^j(W_i)] \leq \frac{j!}{2}A^j \qquad \textit{for all } \; i=1,\dots,n; \tag{5.4}$$

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\Delta^2_{m,m'}(X_i,u,v)] \leq d^2(u,v); \quad \|\Delta_{m,m'}\|_\infty \leq B \quad \textit{and} \quad E=0. \tag{5.5}$$

Remarks :
i) We get (5.4) if we assume that $\mathbb{E}[\exp(M(W_i)/A)] \leq 3/2+\mathbb{E}[M(W_i)/A]$ for $i=1,\dots,n$.
ii) As we already noticed in Birgé and Massart (1994a) (5.3) can be deduced from $\mathbb{L}^2$ and $\mathbb{L}^\infty$ controls on Δ.
iii) If $m=m'$ (5.3) is merely (2.4) of Birgé and Massart (1994a).

The Assumption **M (Metric)** will take 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 $m \in \mathcal{M}_n$ one can find constants $B'_m \geq 1$ and $D_m \geq 1$ such that for each $\delta > 0$ and each ball $\mathcal{B}$ with radius $\sigma \geq 5\delta \vee (D_m/n)^{1/2}$ there exists a finite set $T = T(m, \delta, \mathcal{B}) \subset \mathcal{B}$ with*

$$|T| \leq (B'_m \sigma/\delta)^{D_m} \tag{5.6}$$

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

- ***$M_{2,\infty}$ ($\mathbb{L}^2/\mathbb{L}^\infty$ metric):*** *Assumption **Lip (i)** or **(ii)** holds, $d(u, \pi u) \leq \delta$ for all u in $\mathcal{B}$ and there exists r'_m independent of δ and $\mathcal{B}$ such that*

$$\sup_{u \in \pi^{-1}(t)} \|\Delta_{m,m}(., u, t)\|_\infty \leq r'_m \delta \qquad \text{for all } t \in T. \tag{5.7}$$

- ***$M_{1,[]}$ ($\mathbb{L}^1$ metric with bracketing):*** *Assumption **Lip (ii)** holds and for all $t \in T$ one can find a measurable function V_t such that for all $t \in T$ and all $x \in \mathcal{X}$*

$$\sup_{u \in \pi^{-1}(t)} \Delta_{m,m}(x, u, t) \leq V_t(x) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n \mathbb{E}[V_t(X_i)] \leq \delta^2. \tag{5.8}$$

In the linear situation, i.e. the case of a linear contrast acting on linear sieves, one can substantially simplify these assumptions. This situation essentially occurs when one deals with projection estimators and a detailed study of some more specific examples involving Besov spaces is to be found in Birgé and Massart (1994b). The first assumption describes the general linear set up.

L (Linear) *The space $\mathcal{S}_n$ is a linear subspace of $\mathbb{L}^2(\mu)$ with an algebraic basis $\{\varphi_\lambda \,|\, \lambda \in \bar{\Lambda}_n\}$ and the family of finite dimensional linear spaces $\bar{S}_m$ is generated by a family $\{\Lambda_m\}_{m \in \mathcal{M}_n}$ of subsets of $\bar{\Lambda}_n$: for each $m \in \mathcal{M}_n$, $\bar{S}_m$ is the linear span of $\{\varphi_\lambda \,|\, \lambda \in \Lambda_m\}$ and S_m some subset of $\bar{S}_m$. We denote by $D_m = |\Lambda_m|$ the cardinality of Λ_m which is the dimension of $\bar{S}_m$ and assume that there exists a positive constant Γ such that $\tilde{\gamma}_m(z, u) = -\Gamma u(z)$.*

Let us define the (possibly infinite) constants $\Phi_2(s)$ and $\Phi_\infty(s)$ by

$$\Phi_2(s) = \sup_{u,v \in \mathcal{S}_n} \frac{\mathbb{E}[|u(X_1) - v(X_1)|]}{\|u - v\|}; \qquad \Phi_\infty(s) = \sup_{u \in \mathcal{S}_n} \frac{\mathbb{E}[u^2(X_1)]}{\|u\|^2}.$$

Remark: If s is the common density of the X_i's, $\Phi_2(s)$ and $\Phi_\infty(s)$ are merely bounded by the $\mathbb{L}^2$- and $\mathbb{L}^\infty$-norms of s respectively.

Under this framework, we shall only consider two different situations: either all the elements φ_λ of the basis are uniformly bounded or the family of sieves S_m is nested. More involved applications of the linear framework are to be found in Birgé and Massart (1994b). This leads to the following sets of assumptions:

LBB (Linear with a bounded basis) ***L*** *holds,* $\Phi_\infty(s) < \Phi'$, $\sup_{\lambda\in\bar\Lambda_n} \|\varphi_\lambda\|_\infty \leq \Phi''$ *for some positive constants* Φ' *and* Φ'' *and*

$$\sum_{\lambda\in\bar\Lambda_n} \beta_\lambda^2 \leq A^2 \| \sum_{\lambda\in\bar\Lambda_n} \beta_\lambda\varphi_\lambda\|^2$$

for some positive constant A *and all* (β_λ).

An immediate consequence of **LBB** via Cauchy-Schwarz is that for all subsets Λ of $\bar\Lambda_n$

$$\|\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\|_\infty \leq \Phi|\Lambda|^{1/2}\|\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\| \qquad \text{with} \qquad \Phi = A\Phi''. \tag{5.9}$$

LN (Nested and linear) ***L*** *holds, the set* $\{\Lambda_m\}_{m\in\mathcal{M}_n}$ *is totally ordered by inclusion and when* $\{\Lambda_m\}\subset\{\Lambda_{m'}\}$ *then* $S_m \subset S_{m'}$*. Moreover* $\Phi_2(s) < +\infty$ *and there exists a positive constant* Φ *such that* $\Phi_m \leq \Phi$ *for all* m *where* Φ_m *is defined in (2.1).*

5.2 The Main Theorem

¿From Assumptions **M**, **LBB** and **LN**, one wants to derive various probability bounds for the fluctuations of the empirical process $\nu_n[\tilde\gamma(.,s_m)-\tilde\gamma(.,u)]$. They rely on similar results obtained in Birgé and Massart (1994a) for the case of a single sieve, i.e., when u is assumed to belong to S_m. For the sake of simplicity we shall not try, hereafter, to optimize the constants but rather present the results in a simplified form. The conclusions of Theorems 1 and 2 of Birgé and Massart (1994a) and similar arguments (for dealing with Assumption $\mathbf{M}_{1,[]}$) imply the following

Proposition 7 *Under each set of assumptions (**M**, **LBB** or **LN**) the following exponential inequality is satisfied for all* $t\in S_m$ *and any* $x\geq\sigma_m$ *and* $\tau>0$

$$\mathbb{P}\left[\sup_{u\in S_m}\frac{\nu_n[\tilde\gamma(.,t)-\tilde\gamma(.,u)]}{d^2(t,u)\vee x^2} > \tau\right] \leq 3.1\exp\left[-nh_m(x)\right] \tag{5.10}$$

where

- *under* ***M****:*

$$\sigma_m^2 = \left[\zeta^2\mathcal{L}'_m \vee 1 \vee E\right]\frac{D_m}{n} \quad \text{and} \quad h_m(x) = \left(\frac{x}{\zeta}\right)^2 \quad \text{with} \quad \zeta^2 = \frac{20}{9\tau^2}\left[32A^2+6AB\tau\right]$$

and

$$\mathcal{L}'_m = 2.5\log\left[12B'_m\left(1+r'_m\sqrt{D_m/n}\right)\right]$$

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

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

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

- *under* ***LBB*** *or* ***LN****:*

$$\sigma_m = \frac{\kappa'\Phi}{\tau'}\sqrt{\frac{D_m}{n}} \qquad \text{and} \qquad h_m(x) = \kappa\left(\frac{2\tau' x}{\Phi\sqrt{D_m}}\wedge\frac{\tau'^2x^2}{\Phi_\infty(s)\wedge\Phi\Phi_2(s)\sqrt{D_m}}\right)$$

where $\Phi = A\Phi''$ *under* ***LBB****,* $\kappa\leq 1/4$ *and* κ' *are positive constants and* $\tau'=\tau/\Gamma$.

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

$$\mathbb{P}\left[\sup_{u\in\mathcal{B}} \nu_n[\tilde{\gamma}(.,t) - \tilde{\gamma}(.,u)] > \tau\sigma^2\right] \leq 2\exp\left[-\frac{3n\sigma^2}{10\rho^2(\tau)}\right] \tag{5.11}$$

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

$$\rho^2(\tau) = \frac{16A^2}{\tau^2} + \frac{4AB}{\tau} \qquad \text{and} \qquad \mathcal{L}(\tau) = 5\log(B'_m\theta) \quad \text{with } \theta = \sqrt{B} \vee 2\sqrt{\frac{A}{\tau}} \vee 5.$$

Let us set $\rho = \rho(\tau)$, $\mathcal{L} = \mathcal{L}(\tau)$, $\delta = \sigma/\theta$ and $f_u = \tilde{\gamma}(.,t) - \tilde{\gamma}(.,u)$. Since $\sigma^2 \geq D_m/n$ by (5.6) 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_v with

$$\sup_{u\in\pi^{-1}(v)} \Delta_{m,m}(x,u,v) \leq V_v(x) \qquad \text{and} \qquad \frac{1}{n}\sum_{i=1}^n \mathbb{E}_s[V_v(X_i)] \leq \delta^2. \tag{5.12}$$

Since by (5.5) $\|\Delta_{m,m}\|_\infty \leq B$ we can assume without loss of generality that $\|V_v\|_\infty \leq B$. If $v = \pi(u)$, $|f_u - f_v| \leq MV_v$ and we get

$$\nu_n(f_u) \leq \frac{2}{n}\sum_{i=1}^n \mathbb{E}[M(W_i)V_v(X_i)] + \nu_n(f_v) + \nu_n(MV_v) \leq 2A\delta^2 + \nu_n(f_v) + \nu_n(MV_v) \tag{5.13}$$

by the independence of W_i and X_i, (5.12), (5.4) and Cauchy-Schwarz inequality.

Control of $\nu_n(f_v)$: From the independence of W_i and X_i, (5.4) and (5.5) with $d^2(t,v) \leq \sigma^2$ we get

$$\mathbb{E}[|\tilde{\gamma}(Z_i,t) - \tilde{\gamma}(Z_i,v)|^j] \leq \mathbb{E}[M^j(W_i)]\mathbb{E}[\Delta^j(X_i,t,v)] \leq \frac{j!}{2}A^jB^{j-2}\mathbb{E}[\Delta^2(X_i,t,v)]$$

and

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

Therefore Bernstein's Inequality [see Lemma 5 of Birgé and Massart (1994a)] implies that, if $\eta = \sigma\sqrt{2x} + Bx$

$$\mathbb{P}[\nu_n(f_v) > A\eta] \leq \exp(-nx). \tag{5.14}$$

Control of $\nu_n(MV_v)$: We use again the independence between W_i and X_i and (5.12) to get since $\|V_v\|_\infty \leq B$

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[M^j(W_i)V_v^j(X_i)] \leq \frac{j!}{2}A^2B\delta^2(AB)^{j-2}.$$

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

$$\mathbb{P}[\nu_n(MV_v) > A\eta] \leq \exp(-nx). \tag{5.15}$$

It follows from (5.13), (5.14) and (5.15) that

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

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

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

In order to get (5.11) it remains to check that $2A(\eta+\delta^2) \leq \tau\sigma^2$. This follows from our choices of θ and ρ which imply that $\delta^2 \leq \tau\sigma^2/(4A)$ and $\eta \leq \theta\sigma^2/(4A)$. Since $\mathcal{L} \geq 8$ which implies that $n\sigma^2/\rho^2 \geq 8$ we can derive from (5.11) that if $n\sigma^2 \geq [\mathcal{L}\rho^2(3\tau/4) \vee 1]D_m$,

$$\mathbb{P}\left[\sup_{u\in S_m} \frac{\nu_n[\tilde{\gamma}(.,t) - \tilde{\gamma}(.,u)]}{d^2(t,u) \vee \sigma^2} > \tau\right] \leq 3\exp\left[-\frac{2n\sigma^2}{5\rho^2(3\tau/4)}\right] \tag{5.16}$$

exactly as (2.9) is derived from (2.7) in Birgé and Massart (1994a), following the last lines of the proof of their Theorem 1.

We now want to derive (5.10) with the corresponding values of σ_m, $\mathcal{L}'_m$ and ζ. For Assumption **M** we use either (5.16) (under $\mathbf{M}_{1,[\,]}$) or (2.9) of Theorem 1 of Birgé and Massart (1994a), following their notations (under $\mathbf{M}_{2,\infty}$). In both cases one can choose

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

since one can always increase the value of σ_m. We therefore choose ζ^2 as an upper bound for $(5/2)\rho^2(3\tau/4)$ and $\mathcal{L}'_m$ as an upper bound for $2\mathcal{L}/5$. In the case of $\mathbf{M}_{2,\infty}$ we use the upper bound for $\mathcal{L}$ given in Theorem 1 of Birgé and Massart (1994a) with $\lambda = .757$ to get

$$\mathcal{L} \leq 6.13\left[\log B'_m\left(1 + r'_m\sqrt{D_m/n}\right)\right] + 15.14.$$

To deal with the linear case one should first notice that [since **LBB** implies that $\Phi_m \leq \Phi = A\Phi''$ by (5.9)] both assumptions **LBB** and **LN** imply **LU** of Birgé and Massart (1994a). Therefore their Theorem 2 (which involves two universal constants κ_1 and κ_2) applies. Taking $\kappa = \kappa_1/4$ and $\kappa' = 2\kappa_2$, we get the conclusions for **LBB** and **LN**. □

¿From those results and our various sets of assumptions one can derive the main theorem of this paper which is at the origin of all our developments and examples.

Theorem 7 (Main Theorem) *Let m be a given element of $\mathcal{M}_n$, $\tau = k/8$ and $L_{m'} \geq 1$ be a weight defined for any $m' \in \mathcal{M}_n$. Assuming that one of the assumptions (**M**, **LBB** or **LN**) holds and that $\sigma_{m'}$ is given by Proposition 7, one defines $x_{m'}$ for any $m' \in \mathcal{M}_n$ by $nx^2_{m'} = \theta + n(\sigma_m^2 \vee \sigma_{m'}^2) \vee \lambda(L_{m'}D_{m'} \vee L_m D_m)$ where $\theta \geq 1$ and $\lambda > 0$ and the set $\Omega(\theta)$ by*

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

- *If **M** and **S** hold then*

$$\mathbb{P}[\Omega(\theta)] \leq 4.1\Sigma\exp(-\theta/\lambda) \qquad \text{with } \lambda = \frac{5(32A^2+6AB\tau)}{9\tau^2}. \tag{5.17}$$

- *If **LBB** holds, $\tau' = \tau/\Gamma$, $\lambda = \tau'^{-2}$, $n \geq \Gamma'^2 D_{m'}$ for all $m' \in \mathcal{M}_n$ and some $\Gamma' > 0$ and*

$$\sum_{m'\in\mathcal{M}_n} \exp\left[-\kappa\frac{L_{m'}D_{m'}}{\Phi'}\right] + |\mathcal{M}_n|\exp\left[-\kappa\frac{\sqrt{n}}{A\Phi''\sqrt{2}}\right] = \Sigma < +\infty \tag{5.18}$$

then

$$\mathbb{P}[\Omega(\theta)] \leq 4\Sigma\exp\left[-\kappa\sqrt{\theta}\left(\frac{\tau'\Gamma'}{A\Phi''\sqrt{2}} \wedge \frac{\tau'^2}{\Phi'}\right)\right]. \tag{5.19}$$

- *If **LN** holds, $\tau' = \tau/\Gamma$, $\lambda = 2/\tau'^2$ and $n \geq \Gamma'^2 D_{m'}$ for all $m' \in \mathcal{M}_n$, then*

$$\begin{aligned}\mathbb{P}[\Omega(\theta)] \quad &\leq \quad 4\exp\left[-\frac{\kappa\tau'\sqrt{\theta}}{\Phi\sqrt{2}}\left(\Gamma' \wedge \frac{\sqrt{2}}{\Phi_2(s)}\right)\right] \\ &\qquad \times \sum_{j=1}^{\infty}\exp\left[-\frac{\kappa\sqrt{j \vee D_m}}{\Phi}\left(\Gamma'^2 \wedge \frac{1}{\Phi_2(s)}\right)\right]. \end{aligned}\tag{5.20}$$

*Moreover, if Assumption **C** holds together with the following lower bound on $pen(m')$,*

$$pen(m') \geq k\left(\sigma^2_{m'} \vee \lambda\frac{L_{m'}D_{m'}}{n} + 2k_1\frac{D_{m'}}{n}\right) \quad \text{for all } m' \in \mathcal{M}_n, \tag{5.21}$$

for any integer $l \geq 1$ the risk of the penalized minimum contrast estimator $\hat{s}$ is bounded by

$$\begin{aligned}\mathbb{E}[d^l(s,\hat{s})] \quad &\leq \quad 2^{l/2+[(l-2)\vee 0]}\inf_{m\in\mathcal{M}_n}\left\{\mathbb{E}[U^l_m] + \left[\frac{pen(m)}{k}\right]^{l/2} + \left[\frac{d^2(s,s_m)}{2}\right]^{l/2}\right\} \\ &\qquad + Cn^{-l/2}. \end{aligned}\tag{5.22}$$

*Under **M** or **LBB**, the constant C only depends on l and the various constants involved in the upper bounds (5.17) or (5.19) but not on s. Under **LN**, C can be written as $C'[\Phi(\Phi_2(s) \vee 1)]^{l+2}$ and therefore depends on s through $\Phi_2(s)$ and on the other constants appearing in (5.20) through C'.*

*If **C'** holds together with (5.21) with $k = k'/2$ and $k_1 = 0$ then for all $\theta \geq 1$*

$$\mathbb{P}\left[d^2(s,\hat{s}) > \frac{1}{k'}\inf_{m\in\mathcal{M}_n}\{4pen(m) + (k'+2k'')d^2(s,s_m)\} + \frac{\theta}{n}\right] \leq \mathbb{P}[\Omega(\theta)] \tag{5.23}$$

where $\mathbb{P}[\Omega(\theta)]$ is bounded by (5.17), (5.19) or (5.20) according to the type of assumptions we use.

Proof: The first step in the proof is to show that whatever $m, m' \in \mathcal{M}_n$ and $x \geq \sigma_m \vee \sigma_{m'}$ the following exponential inequality is valid:

$$\mathbb{P}\left[\sup_{u\in S_{m'}}\frac{\nu_n[\tilde{\gamma}(.,s_m) - \tilde{\gamma}(.,u)]}{d^2(s,u) \vee d^2(s_m,s) \vee x^2} > k\right] \leq 4.1\exp\left[-nh_{m,m'}(x)\right] \tag{5.24}$$

where σ_m and $\sigma_{m'}$ are defined according to Proposition 7. As in Proposition 7, the function $h_{m,m'}$ will take different forms, to be specified below, according to our set of assumptions. Let us consider some point t in $S_{m'}$ and denote by d the quantity $d(s_m,t)$. One can apply Proposition 7 with m replaced by m' and S_m by $S_{m'}$. On the other hand if Assumption **Lip** holds, Bernstein's inequality [see Birgé and Massart (1994a) Lemma 5] leads to a bound of the form

$$\mathbb{P}\left[\frac{\nu_n[\tilde{\gamma}(.,s_m)-\tilde{\gamma}(.,t)]}{d^2\vee x^2}>\tau\right]\leq \exp\left(\frac{-\frac{1}{2}n\tau^2(d^2\vee x^2)^2}{v^2+c\tau(d^2\vee x^2)}\right)$$

provided that for all integers $j\geq 2$

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

We only have to identify v^2 and c. From (5.2) and (5.3) or (5.4) and (5.5) it can be seen that $v^2=A^2[d^2+E(D_m\vee D_{m'})/n]$ and $c=AB$. Since $x^2\geq E(D_m\vee D_{m'})/n$, $v^2\leq 2A^2(d^2\vee x^2)$ and finally

$$\mathbb{P}\left[\frac{\nu_n[\tilde{\gamma}(.,s_m)-\tilde{\gamma}(.,t)]}{d^2\vee x^2}>\tau\right]\leq \exp\left[\frac{-n\tau^2(d^2\vee x^2)}{4A^2+2\tau AB}\right].$$

In order to handle the linear cases we first notice from **L** that if $\tau'=\tau/\Gamma$

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

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

$$\mathbb{P}\left[\frac{\nu_n[\tilde{\gamma}(.,s_m)-\tilde{\gamma}(.,t)]}{d^2\vee x^2}>\tau\right]\leq \exp\left[\frac{-\frac{1}{2}n\tau'^2d^2x^2}{\mathbb{E}[(s_m-t)^2]+\frac{1}{3}\|s_m-t\|_\infty\tau' dx}\right]. \tag{5.25}$$

We first notice that in both cases **LBB** and **LN** $\|s_m-t\|_\infty\leq\Phi d\sqrt{D_m\vee D_{m'}}$. This follows from (5.9) with $\Phi=A\Phi''$ in the first case and the fact that $\Phi_m\vee\Phi_{m'}\leq\Phi$ in the nested case.

LBB: Now we use $\mathbb{E}[(s_m-t)^2]\leq d^2\Phi_\infty(s)\leq d^2\Phi'$ and bound (5.25) by

$$\exp\left[\frac{-\frac{1}{2}n\tau'^2d^2x^2}{\frac{1}{3}\Phi d\sqrt{D_m\vee D_{m'}}\tau' dx+d^2\Phi'}\right]\leq \exp\left[\frac{-n}{4}\left(\frac{\tau'^2x^2}{\Phi'}\wedge\frac{3\tau' x}{\Phi\sqrt{D_m\vee D_{m'}}}\right)\right].$$

LN: Since $\mathbb{E}[(s_m-t)^2]\leq d\Phi_2(s)\|s_m-t\|_\infty$ the right-hand side of (5.25) can be bounded by

$$\exp\left[\frac{-\frac{1}{2}n\tau'^2d^2x^2}{\Phi d\sqrt{D_m\vee D_{m'}}[d\Phi_2(s)+\frac{1}{3}\tau' dx]}\right]\leq \exp\left[\frac{-n}{4\Phi\sqrt{D_m\vee D_{m'}}}\left(\frac{\tau'^2x^2}{\Phi_2(s)}\wedge 3\tau' x\right)\right].$$

Putting these bounds together with inequality (5.10) and using the facts that $\kappa\leq 1/4$ and $\tau=k/8$ we get

$$\mathbb{P}\left[\sup_{u\in S_{m'}}\frac{\nu_n[\tilde{\gamma}(.,s_m)-\tilde{\gamma}(.,u)]}{d^2(t,u)\vee x^2\vee d^2(s_m,t)}>k/4\right]\leq 4.1\exp\left[-nh_{m,m'}(x)\right] \tag{5.26}$$

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

$$M: \quad h_{m,m'}(x) = x^2/\zeta^2 \tag{5.27}$$

$$LBB: \quad h_{m,m'}(x) = \kappa\left[\frac{\tau^2 x^2}{\Gamma^2\Phi'} \wedge \frac{2\tau x}{\Gamma A\Phi''\sqrt{D_m \vee D_{m'}}}\right]. \tag{5.28}$$

$$LN: \quad h_{m,m'}(x) = \frac{\kappa}{\Phi\sqrt{D_m \vee D_{m'}}}\left[\frac{\tau^2 x^2}{\Gamma^2\Phi_2(s)} \wedge \frac{2\tau x}{\Gamma}\right]; \tag{5.29}$$

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 in (5.26) leads to (5.24), since ε is arbitrary.

Let us now prove (5.17), (5.19) and (5.20). Since $x_{m'} > \sigma_m \vee \sigma_{m'}$ we can use (5.24) under each of our three sets of assumptions. In the first case, since $\lambda = \zeta^2$ we get from (5.27)

$$\begin{aligned} \mathbb{P}[\Omega(\theta)] &\leq 4.1\sum_{m'}\exp\left[-nx_{m'}^2/\lambda\right] \leq 4.1\sum_{m'}\exp\left[-(\lambda L_{m'}D_{m'}+\theta)/\lambda\right] \\ &\leq 4.1\exp[-\theta/\lambda]\sum_{m'}\exp\left[-L_{m'}D_{m'}\right] = 4.1\Sigma\exp[-\theta/\lambda] \end{aligned}$$

which is (5.17). In order to deal with the linear cases it will be useful to notice that

$$\sqrt{2n}x_{m'} \geq \sqrt{\lambda(L_m D_m \vee L_{m'}D_{m'})} + \sqrt{\theta} \tag{5.30}$$

and

$$2nx_{m'}^2 \geq \lambda(L_m D_m \vee L_{m'}D_{m'}) + 2\sqrt{2\theta\lambda(L_m D_m \vee L_{m'}D_{m'})}. \tag{5.31}$$

Under the second set of assumptions, setting $\tau' = \tau/\Gamma$ we get from (5.28)

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

and from (5.30) since $n \geq \Gamma'^2 D_{m'}$ and $L_{m'} \geq 1$ for all $m' \in \mathcal{M}_n$

$$\frac{nx_{m'}}{\sqrt{D_m \vee D_{m'}}} \geq \frac{\sqrt{n\lambda} + \Gamma'\sqrt{\theta}}{\sqrt{2}}. \tag{5.32}$$

Then since $\lambda = \tau'^{-2}$

$$\begin{aligned} \mathbb{P}[\Omega(\theta)] &\leq 4.1\sum_{m'}\exp\left[-\kappa\left(\frac{\tau'^2\theta + \tau'^2\lambda L_{m'}D_{m'}}{\Phi'} \wedge \frac{\tau'\Gamma'\sqrt{\theta} + \tau'\sqrt{n\lambda}}{A\Phi''\sqrt{2}}\right)\right] \\ &\leq 4.1\exp\left[-\kappa\left(\frac{\sqrt{\theta}\tau'\Gamma'}{A\Phi''\sqrt{2}} \wedge \frac{\theta\tau'^2}{\Phi'}\right)\right]\sum_{m'}\exp\left[-\kappa\left(\frac{L_{m'}D_{m'}}{\Phi'} \wedge \frac{\sqrt{n}}{A\Phi''\sqrt{2}}\right)\right] \end{aligned}$$

and (5.19) follows from (5.18). Finally when **LN** holds one gets from (5.29)

$$\mathbb{P}[\Omega(\theta)] \leq 4.1 \sum_{m'} \exp\left[-\frac{n\kappa}{\Phi\sqrt{D_m \vee D_{m'}}}\left(\frac{\tau'^2 x_{m'}^2}{\Phi_2(s)} \wedge \tau' x_{m'}\right)\right].$$

We modify the linear term as before with (5.32) and use the following relation

$$\frac{n x_{m'}^2}{\sqrt{D_m \vee D_{m'}}} \geq \frac{\lambda}{2}\sqrt{D_m \vee D_{m'}} + \sqrt{2\theta\lambda}$$

derived from (5.31) to deal with the quadratic term. Since $n \geq \Gamma'^2(D_m \vee D_{m'})$ we get

$$\begin{aligned}
&\mathbb{P}[\Omega(\theta)] \\
&\leq 4.1 \sum_{m'} \exp\left[-\frac{\kappa}{\Phi}\left(\frac{\tau'^2}{\Phi_2(s)}\left(\frac{\lambda}{2}\sqrt{D_m \vee D_{m'}} + \sqrt{2\theta\lambda}\right) \wedge \frac{\tau'\Gamma'\sqrt{\theta} + \tau'\sqrt{n\lambda}}{\sqrt{2}}\right)\right] \\
&\leq 4.1 \exp\left[-\frac{\kappa\sqrt{\theta}}{\Phi}\left(\frac{\tau'\Gamma'}{\sqrt{2}} \wedge \frac{\tau'^2\sqrt{2\lambda}}{\Phi_2(s)}\right)\right] \\
&\quad \times \sum_{m'} \exp\left[-\frac{\kappa\sqrt{D_m \vee D_{m'}}}{\Phi}\left(\frac{\tau'\Gamma'^2\sqrt{\lambda}}{\sqrt{2}} \wedge \frac{\tau'^2\lambda}{2\Phi_2(s)}\right)\right]. \qquad (5.33)
\end{aligned}$$

Since under **LN**, all $D_{m'}$ are different integers, the series converges and we get (5.20) from our choice of λ.

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

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

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

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

Since by (5.21) $x_{\hat{m}}^2 + 2k_1 D_{\hat{m}}/n \leq k^{-1}[pen(m) + pen(\hat{m})] + \theta/n$ the following inequality holds:

$$\mathbb{1}_{\Omega^c(\theta)} d^2(s,\hat{s}) \leq 2k^{-1}pen(m) + 2U_m^2 + d^2(s,s_m) + \theta/n. \qquad (5.34)$$

Let us now define

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

then for any $m \in \mathcal{M}_n$ and any positive integer l

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

It follows from (5.34) that if $\theta \geq 1$, $\mathbb{P}[V > \theta/n] \leq \mathbb{P}[\Omega(\theta)]$. Since we have proved that under each of our sets of assumptions $\mathbb{P}[\Omega(\theta)] \leq c_1\exp(-c_2\sqrt{\theta})$ for suitable constants c_1 and c_2, it follows that

$$\begin{aligned}
\mathbb{E}[V^{l/2}] &= n^{-l/2}\mathbb{E}[(nV)^{l/2}] = n^{-l/2}\int_0^\infty \mathbb{P}[nV > y^{2/l}]\,dy \\
&\leq n^{-l/2}\left[1 + \int_1^\infty c_1\exp\left(-c_2 y^{1/l}\right)dy\right] \leq n^{-l/2}\left[1 + c_1 c_2^{-l} l!\right].
\end{aligned}$$

which proves (5.22). It follows from (5.17), (5.19) and (5.20) that c_1 and c_2 depend on the constants introduced in the assumptions but not on s except under **LN**. In this latter case, $1/c_2$ can be written as $c_3\Phi(\Phi_2(s) \vee 1)$ and comparing the series in (5.20) with an integral one easily checks that $c_1 \leq c_4(\Phi(\Phi_2(s) \vee 1))^2$ which gives the structure of C in this case.

Finally (5.23) follows from (5.34) and the fact that **C'** implies **C** with $k = k'/2$, $k_1 = 0$ and $U_m^2 = (k''/k')d^2(s, s_m)$. $\square$

5.3 Application to maximum likelihood estimation

We now want to show how one can apply Theorem 7 to maximum likelihood estimation. In order to apply the general theory to specific situations, **M** will often follow from the simpler Assumption $\mathbf{M'}_{2,\infty}$:

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

$$|T| \leq (B'_m\sigma/\delta)^{D_m}. \tag{5.35}$$

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

$$\sup_{u \in \pi^{-1}(t)} \|u - t\|_\infty \leq r_m\delta \qquad \textit{for all } t \textit{ in } T. \tag{5.36}$$

The following is a generalized version of Theorem 1.

Theorem 8 *Assume that μ is a probability, that* ***M'***$_{2,\infty}$ *holds with $D_m \leq n$ for each m and that the family $\{S_m, m \in \mathcal{M}_n\}$ satisfies Assumption **S**. Define η_m by $\int(s^2 \vee \eta_m)\, d\mu = 1 + D_m/n$ and $pen(m) \geq \kappa_8(L_m + \mathcal{L}_m)D_m/n$ where*

$$\mathcal{L}_m = \log\left[B'_m\left(1 + r_m\sqrt{\frac{7D_m}{4n\eta_m}}\right)\right] + 1$$

and κ_8 is a suitable positive numerical constant. Let $\hat{s}$ be a minimizer with respect to $m' \in \mathcal{M}_n$ and $t \in S_{m'}$ of $-\sum_{i=1}^n \log[t(X_i)] + n\, pen(m')$. Then

$$\eta_m \geq D_m/n \qquad \textit{and} \qquad \mathbb{E}[d^2(s, \hat{s})] \leq \kappa'_8 \inf_{m \in \mathcal{M}_n} \{K(s, S_m) + pen(m)\}. \tag{5.37}$$

Remark: One could of course get similar results under the slightly more general assumption that $D_m \leq Kn$. We shall only deal with the case $K = 1$ for the sake of simplicity .

Proof: Since it follows rather closely the lines of the proof of Theorem 4 of Birgé and Massart (1994a) we shall omit some details. Let us first notice that, since μ is a probability measure, η_m is well-defined and larger than D_m/n and introduce the auxiliary density $\tilde{s}_m^2 = (s^2 \vee \eta_m)/(1 + D_m/n)$. Then

$$\left\|\frac{s}{\tilde{s}_m}\right\|_\infty^2 \leq 1 + \frac{D_m}{n} \leq 2 \qquad \text{and} \qquad \inf_x \tilde{s}_m^2(x) \geq \frac{\eta_m}{2} \geq \frac{D_m}{2n} \tag{5.38}$$

and it follows from (6.1) below that

$$d^2(s, \tilde{s}_m) \le \frac{D_m}{n}. \tag{5.39}$$

In order to apply Theorem 7 we define for any $m \in \mathcal{M}_n$ and $t \in S_m$ the function

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

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

Lemma 2 *Let f, g, g_1, g_2 be densities with respect to some measure μ then for all $j \ge 2$*

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

Proof : The bound involving $\|f/g\|_\infty$ has been proved in Birgé and Massart (1994a), Proposition 2. For the other part we use (6.15) of Birgé and Massart (1994a) to show that when $g_1 \ge g_2$,

$$\begin{aligned}
\frac{1}{j!}\left(\frac{1}{2}\log\frac{g+g_1}{g+g_2}\right)^j &\le \frac{1}{j!}\left(\frac{1}{2}\log\frac{g_1}{g_2}\right)^j \le \sqrt{\frac{g_1}{g_2}} - 1 - \log\left(\sqrt{\frac{g_1}{g_2}}\right) \\
&\le \frac{1}{7}\left(\sqrt{\frac{g_1}{g_2}} - \sqrt{\frac{g_2}{g_1}}\right)^2 \\
&= \frac{1}{7}(\sqrt{g_1} - \sqrt{g_2})^2\left(\frac{\sqrt{g_1} + \sqrt{g_2}}{\sqrt{g_1 g_2}}\right)^2 \le \frac{4}{7}\frac{(\sqrt{g_1} - \sqrt{g_2})^2}{g_1 \wedge g_2}.
\end{aligned}$$

A symetric result holds when $g_2 \ge g_1$ and integration with respect to $f\mu$ gives the result. $\square$

Lemma 3 *Assume that f, g_1, g_2, s_1^2 and s_2^2 are densities with respect to the probability measure μ and that $\|f/s_i^2\|_\infty \le 2$ for $i = 1, 2$. For any integer $j \ge 2$ one has*

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

Proof: Successive applications of Lemma 2 give

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

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

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

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

Then (5.2) is satisfied and

$$h^2(\tilde{s}_m^2,\tilde{s}_{m'}^2)\le 4h^2(s^2,\tilde{s}_m^2)\vee 4h^2(s^2,\tilde{s}_{m'}^2)=2d^2(s,\tilde{s}_m)\vee 2d^2(s,\tilde{s}_{m'});$$

hence by (5.39)

$$h^2(\tilde{s}_m^2,\tilde{s}_{m'}^2)\le 2\frac{D_m\vee D_{m'}}{n}1\!\!1_{\{m\neq m'\}}.$$

An application of Lemma 3, which is valid because of (5.38), leads to

$$\mathbb{E}\left[\Delta_{m,m'}^j(X_i,u,v)\right]=\left(\frac{4}{A}\right)^j\frac{4}{7}\frac{j!}{2}\frac{1}{2}\left[d^2(u,v)+16\frac{D_m\vee D_{m'}}{n}1\!\!1_{\{m\neq m'\}}\right].$$

The choice $A=4\sqrt{2/7}$ gives (5.3) with $B=\sqrt{7/2}$ and $E=16$. (5.6) holds by assumption and the lower bound on $\tilde{s}_m$ in (5.38) together with Lemma 3 of Birgé and Massart (1994a) imply that whenever $\|t-u\|_\infty\le r_m\delta$ for t and u in S_m, then

$$\|\Delta_{m,m}(x,t,u)\|_\infty\le 2A^{-1}r_m\delta\sqrt{2/\eta_m}.$$

Therefore (5.7) holds with $r'_m=r_m\sqrt{7/(4\eta_m)}\le r_m\sqrt{(7n)/(4D_m)}$ by (5.38). The value of $\mathcal{L}_m$ follows from the value of $\mathcal{L}'_m$ given in Proposition 7 after a suitable modification of the multiplicative constant which can be included in κ_8. It remains to check Assumption **C**. We proceed as in Birgé and Massart (1994a). If $t\in S_{m'}$ is such that $\gamma_n(t)+pen(m')\le\gamma_n(s_m)+pen(m)$ by convexity

$$\sum_{i=1}^n\log\frac{t^2+\tilde{s}_{m'}^2}{2}(X_i)\ge\sum_{i=1}^n\left[\log[\tilde{s}_{m'}(X_i)]+\log[s_m(X_i)]\right]+pen(m')-pen(m)$$

and since by (5.38) $\log\tilde{s}_{m'}\ge\log s-(1/2)\log(1+D_{m'}/n)\ge\log s-D_{m'}/(2n)$,

$$\begin{aligned}
&\nu_n[\tilde{\gamma}_m(.,s_m)-\tilde{\gamma}_{m'}(.,t)]\\
&\quad\ge\ \mathbb{P}_n\left[\log\frac{2ss_m}{\tilde{s}_m^2+s_m^2}\right]-\mathbb{E}\left[\log\frac{\tilde{s}_{m'}^2(X)+t^2(X)}{2s^2(X)}-\log\frac{\tilde{s}_m^2(X)+s_m^2(X)}{2s^2(X)}\right]\\
&\qquad-\frac{D_{m'}}{2n}+pen(m')-pen(m)\\
&\quad\ge\ \mathbb{P}_n\left[\log\frac{2s^2}{\tilde{s}_m^2+s_m^2}\right]-\frac{1}{2}\mathbb{P}_n\left[\log\frac{s^2}{s_m^2}\right]-\frac{D_{m'}}{2n}-K\left(s,\sqrt{\frac{\tilde{s}_m^2+s_m^2}{2}}\right)\\
&\qquad+K\left(s,\sqrt{\frac{\tilde{s}_{m'}^2+t^2}{2}}\right)+pen(m')-pen(m).
\end{aligned}$$

Since for any positive θ

$$d^2\left(s,\sqrt{\frac{s^2+t^2}{2}}\right)\le(1+\theta)\,d^2\left(s,\sqrt{\frac{\tilde{s}_{m'}^2+t^2}{2}}\right)+\left(1+\frac{1}{\theta}\right)d^2\left(\sqrt{\frac{s^2+t^2}{2}},\sqrt{\frac{\tilde{s}_{m'}^2+t^2}{2}}\right)$$

and recalling the correspondence $d^2(u,v)=2h^2(P_u,P_v)$, one derives from (6.1) and (6.3) of Lemma 4 of Birgé and Massart (1994a), that

$$\begin{aligned}
d^2\left(s,\sqrt{\frac{\tilde{s}_{m'}^2+t^2}{2}}\right)&\ge\frac{1}{1+\theta}\,d^2\left(s,\sqrt{\frac{s^2+t^2}{2}}\right)-\frac{1}{\theta}\,d^2\left(\sqrt{\frac{s^2+t^2}{2}},\sqrt{\frac{\tilde{s}_{m'}^2+t^2}{2}}\right)\\
&\ge\frac{0.29^2}{1+\theta}\,d^2(s,t)-\frac{1}{2\theta}\,d^2(s,\tilde{s}_{m'}).
\end{aligned}$$

Choosing $\theta = 1/2$ we deduce from (6.2) below that

$$K\left(s, \sqrt{\frac{\tilde{s}_{m'}^2 + t^2}{2}}\right) \geq \frac{2}{3}.29^2\, d^2(s,t) - d^2(s, \tilde{s}_{m'}).$$

It follows from the concavity of the logarithm that

$$2K\left(s, \sqrt{\frac{\tilde{s}_m^2 + s_m^2}{2}}\right) \leq K(s, \tilde{s}_m) + K(s, s_m) \leq (2 + \log 2)\, d^2(s, \tilde{s}_m) + K(s, s_m)$$

since (5.38) implies by (6.2) that $K(s, \tilde{s}_m) \leq (2 + \log 2) d^2(s, \tilde{s}_m)$. Putting all these bounds together with (5.39) we get

$$\begin{aligned}
&\nu_n[\tilde{\gamma}_m(., s_m) - \tilde{\gamma}_{m'}(., t)] \\
&\quad \geq \frac{2}{3}.29^2\, d^2(s,t) + pen(m') - pen(m) - \frac{3D_{m'}}{2n} - \left(1 + \frac{\log 2}{2}\right)\frac{D_m}{n} \\
&\qquad - \frac{1}{2}K(s, s_m) - \mathbb{P}_n\left[\frac{1}{2}\log\frac{s^2}{s_m^2} - \log\frac{2s^2}{\tilde{s}_m^2 + s_m^2}\right].
\end{aligned}$$

Since $.29^2/3 > .028$ and

$$\mathbb{E}\left[\frac{1}{2}\log\frac{s^2}{s_m^2}(X_i) - \log\frac{2s^2}{\tilde{s}_m^2 + s_m^2}(X_i)\right] \leq \frac{1}{2}K(s, s_m)$$

we finally see that **C** holds with $k = .028$, $k_1 = 3/(4k)$ and

$$2k\mathbb{E}[U_m^2] < \left(1 + \frac{\log 2}{2}\right) + K(s, s_m).$$

The application of Theorem 7 leads to inequality (5.37) since $d^2(s, s_m) \leq K(s, s_m)$ and $d^2(s, \hat{s}) \leq 2$. $\square$

Proof of Theorem 1 We can now derive Theorem 1 from Theorem 8. It is enough to check the properties (5.35) and (5.36) 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 $\mathcal{L}_m$ by $\bar{\kappa}[1 + \log(1 + \bar{r}_m)]$ and the result follows from a suitable modification of the constants since $L_m \geq 1$.

5.4 Other contrast functions

5.4.1 Projection estimators

We shall content ourselves to prove Theorem 2. A complete treatment of penalized projection estimators is contained in Birgé and Massart (1994b) and we shall merely recall that one should choose $\tilde{\gamma}_m(z,t) = -2t(z)$ which implies **C'** with $k' = k'' = 1$. Moreover the assumptions of Theorem 2 imply **LN** since in this case $\Phi_2(s) = \|s\|^2$. Theorem 7 applies with $\Gamma = 2$, $\Gamma' = 1$ and the result follows.

5.4.2 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 $\mathcal{X}$. The unknown parameter s is supposed to belong to the Hilbert space $\mathbb{L}^2(\mu_n)$ where μ_n denotes the average distribution of the X_i's, $\mu_n = n^{-1}\sum_{i=1}^n R_i$ and we always assume that $s \in \mathcal{S}_n$. 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 i.i.d. with common distribution Q and that the X_i's are either i.i.d. with common distribution μ (which is the random design model) or that the X_i's are given numbers x_i (which is the fixed design model). In the latter case, μ_n is the empirical measure of the x_i's but we shall omit the subscript n for the sake of simplicity, keeping in mind that in this case both the underlying space $\mathbb{L}^2(\mu)$ and the distance d depend on n and that the results, as in Van de Geer (1990), 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 $pen(m)$ to be chosen later, we shall consider either the penalized least squares estimator which is a minimizer with respect to $m \in \mathcal{M}_n$ and $t \in S_m$ of $n^{-1}\sum_{i=1}^n [Y_i - t(X_i)]^2 + pen(m)$ or the penalized minimum $\mathbb{L}^1$ estimator which is a minimizer with respect to $m \in \mathcal{M}_n$ and $t \in S_m$ of $n^{-1}\sum_{i=1}^n |Y_i - t(X_i)| + pen(m)$. Then the following result holds.

Theorem 9 *Assume that* $\boldsymbol{M'}_{2,\infty}$ *holds, that the family* $\{S_m, m \in \mathcal{M}_n\}$ *satisfies Assumption* $\boldsymbol{S}$ *and that both* s *and all the elements of* $\mathcal{S}_n$ *are uniformly bounded by some known constant* ξ. *Assume moreover that the distribution* Q *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 positive and continuous density around the median in the case of minimum* $\mathbb{L}^1$ *estimation.*

Define the penalty function as $pen(m) \geq \kappa_9 C(\xi, Q)(L_m + \mathcal{L}_m)D_m/n$ *where*

$$\mathcal{L}_m = \log\left[B'_m\left(1 + r_m\sqrt{D_m/n}\right)\right] + 1,$$

κ_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 penalized minimum contrast estimator. Then in both cases*

$$\mathbb{E}[d^2(s, \hat{s})] \leq \kappa'_9 \inf_{m\in\mathcal{M}_n}\{d^2(s, S_m) + C'(\xi, Q)pen(m)\}. \tag{5.40}$$

In the case of least squares estimation one can choose $C'(\xi, Q) = 1$ *and* $C(\xi, Q) = (\xi + \xi')^2$.

Proof: We shall actually prove a more general result. Following the framework given in Birgé and Massart (1993) Section 3.C we shall assume that the contrast function is given by $\gamma(z, t) = F(y - t(x))$ where F is a convex function with suitable properties connected to the distributions Q of the $W'i$'s, provided that s and all the elements t of the sieves 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 contrast functions $(y-t(x))^2$ and $|y-t(x)|$ satisfy these assumptions under the conditions of Theorem 9. Then, Proposition 1 of Birgé and Massart (1993) implies that (5.2) and (5.3) are satisfied with $\Delta_{m,m'}(x,t,u) = |t(x)-u(x)|$, $E=0$ and suitable constants A, B depending on Q and ξ. In the particular case of $F(x)=x^2$, one has

$$|\gamma(z,t)-\gamma(z,u)| = |t(x)-u(x)|\,|2w+2s(x)-(t(x)+u(x))| \leq 2|t(x)-u(x)|\,[|w|+\xi]\,.$$

We can therefore take $B=2\xi$ and $M(w)=2(|w|+\xi)$. Our moment condition on the W_i's implies that

$$\mathbb{E}\left[2^j(|W|+\xi)^j\right] \leq \frac{4^j j!}{2}\left[2\xi'+\xi\right]^j$$

and one can choose $A=4(2\xi'+\xi)$. If the metric assumption (5.36) holds then (5.6) and (5.7) are fulfilled with $r'_m = r_m$ since here $\Delta_{m,m'}(x,t,u) = |t(x)-u(x)|$ and therefore $\mathbf{M}_{2,\infty}$ is satisfied. In order to apply Theorem 7 it only remains 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

$$\mathbb{E}[\gamma(Z_i,t)-\gamma(Z_i,s)] = \mathbb{E}[G(W_i, s(X_i)-t(X_i))]$$

and that for suitable positive constants C_1 and C_2

$$C_1 h^2 \leq \mathbb{E}[G(W_i,h)] \leq C_2 h^2 \quad \text{for} \quad |h|\leq 2\xi.$$

In view of the independence between X_i and W_i, these relations imply **C'**. In the quadratic case ($F(x)=x^2$), $G(x,h)=h^2$ and therefore $C_1=C_2=1$ and **C'** is satisfied with $k'=k''=1$. The choice of C and C' is justified by Theorem 7 and our computations of A, B, k' and k''. □

Proof of Theorem 3: By Lemma 9 assumptions (5.6) and (5.7) are satisfied with $B'_m=5$ and $r_m=\bar{r}_m$ and therefore Theorem 9 implies Theorem 3 via some elementary computations since $L_m\geq 1$. □

Proof of Theorem 4 We want to derive it from Theorem 7. As we already checked in the proof of Theorem 9 the Assumption **Lip i)** is satisfied for the contrast function $\gamma(z,t)=(y-t(x))^2$ with $\Delta_{m,m'}(x,t,u)=|t(x)-u(x)|$, $M(w)=2(|w|+\xi)$, $E=0$, $A=4(2\xi'+\xi)$ and $B=2\xi$ where now $\xi=1$. Moreover **C'** is also satisfied with with $k'=k''=1$. There only remains to check the Assumption $\mathbf{M}_{1,[\,]}$. Let us consider some ball $\mathcal{B}$ of radius σ in S_m and some $\delta\leq\sigma/5$. From inequality (2.20) $\mathcal{B}$ is included in the image via θ 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 $\mathcal{I}$ of intervals of $\mathbb{L}^1(\mu')$-diameter $\leq\varepsilon$ with cardinality bounded by

$$(3eB''_m\Theta_2/\Theta_1)^{D_m}(\sigma^2/\delta^2)^{D_m}. \tag{5.41}$$

Truncating the intervals if necessary, we can assume, without loss of generality that these intervals are included in $\mathcal{G}_n$. Therefore the images of the elements of $\mathcal{I}$ via θ

are covering $\mathcal{B}$. Since θ is non-decreasing, for each interval $[g^-, g^+] \in \mathcal{I}$, $\theta([g^-, g^+]) \subset [\theta(g^-), \theta(g^+)]$ and by (2.20) the $\mathbb{L}_1(\mu)$-diameter of $[\theta(g^-), \theta(g^+)]$ is bounded by δ^2. Choosing t as any point in $[\theta(g^-), \theta(g^+)] \cap \mathcal{B}$ and defining $V_t = \theta(g^+) - \theta(g^-)$ we take for T the set of all those t's when $[g^-, g^+]$ varies in $\mathcal{I}$. We then define π to be the projection mapping $[\theta(g^-), \theta(g^+)] \cap \mathcal{B}$ on the point $t \in T \cap [\theta(g^-), \theta(g^+)]$. Since $\Delta_{m,m'}(x,t,u) = |t(x) - u(x)|$ (5.8) is fulfilled and (5.41) implies (5.6) with D_m replaced by $2D_m$ and $B'_m = (3eB''_m \Theta_2/\Theta_1)^{1/2}$. We can therefore apply Theorem 7 and get via elementary transformations of the constants (since $\Theta_2 \geq \Theta_1$)

$$\mathbb{E}[d^2(s, \theta_{\tilde{f}})] \leq \kappa'_6 \inf_{m \in \mathcal{M}_n} \left[d^2(s, S_m) + (\xi' + 1)^2 \left[L_m + \log(1 + \Theta_2 B''_m/\Theta_1) \right] \frac{D_m}{n} \right].$$

Using (2.20) and the fact that $\theta_g \leq 1$ we derive (2.22). $\square$

5.4.3 Estimating the support of a density

Proof of Theorem 5: The proof is based on the version of Theorem 7 involving the Assumption $\mathbf{M}_{1,[]}$. Let us check the relevant assumptions: first **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 $k' = k'' = 1$ by (3.26). In order to check $\mathbf{M}_{1,[]}$ we consider some ball $\mathcal{B}$ of radius $\sigma \geq \sqrt{D_m/n}$ in S_m and some $\delta \leq \sigma/5$. Let $\varepsilon = \delta^2/\Psi$. Assume first that $\varepsilon \leq 1$ and apply Lemma 11 with $R = 2$. We get a covering of G_m by a family $\mathcal{I}$ of intervals of $\mathbb{L}^1(\mu')$-diameter $\leq \varepsilon$ with

$$|\mathcal{I}| \leq (3eB''_m/\varepsilon)^{D_m}. \tag{5.42}$$

The images of the elements of $\mathcal{I}$ via θ are therefore covering S_m and a fortiori $\mathcal{B}$. Since θ is non-decreasing, for each interval $[g^-, g^+] \in \mathcal{I}$, θ maps $[g^-, g^+]$ into $[\theta(g^-), \theta(g^+)]$ and by (3.27) the $\mathbb{L}_1(\mu)$-diameter of $[\theta(g^-), \theta(g^+)]$ is bounded by δ^2. Choosing t as any point in $[\theta(g^-), \theta(g^+)] \cap \mathcal{B}$ and defining $V_t = \theta(g^+) - \theta(g^-)$ we take for T the set of all those t's when $[g^-, g^+]$ varies in $\mathcal{I}$. We then define π to be the projection mapping $[\theta(g^-), \theta(g^+)] \cap \mathcal{B}$ on the point $t \in T \cap [\theta(g^-), \theta(g^+)]$. Since $\Delta_{m,m'}(x,t,u) = |t(x) - u(x)|$ (5.8) is fulfilled. From (5.42) we deduce, since $\sigma^2 \geq D_m/n \wedge 1$, that

$$|T| \leq \left[3eB''_m \Psi \left(\frac{n}{D_m} \vee 1 \right) \right]^{D_m} \left(\frac{\sigma^2}{\delta^2} \right)^{D_m}. \tag{5.43}$$

If $\varepsilon > 1$, noticing that $\Psi \geq 1$ since μ and μ' are two probability measures, we now take $T = \{\mathbb{1}_{\mathbb{D}}\}$ and $V_{\mathbb{1}_{\mathbb{D}}} = \mathbb{1}_{\mathbb{D}}$. Then (5.8) is satisfied since $\delta^2 \geq \Psi \geq 1$ and (5.43) still holds. (5.6) follows with D_m replaced by $2D_m$ and $B'_m = (3eB''_m \Psi n/D_m)^{1/2} \vee 1$. We can therefore apply Theorem 7 and get the result via elementary transformations of the constants. $\square$

5.5 Analysis of nonlinear sieves

Here we use Theorems 8 and 9 to prove the risk bounds for nonlinear models stated in Theorem 6. 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 properties using the following Assumption **M"**. The use of a global entropy condition 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 δ. The logarithm of the number of balls that are needed is, for small δ, of order $-q\log\delta + C$ instead of $q\log\lambda + C$ whichis needed for the covering of a ball of radius $\lambda\delta$. This makes a serious difference when λ is not large.

M" *For each $m \in \mathcal{M}_n$ one can find constants $c_m \geq 1$ and $D_m \geq 1$ and for each $\delta > 0$ there is a finite set $T = T(m,\delta) \subset S_m$ with*

$$|T| \leq \left[c_m(1 \vee \delta^{-1})\right]^{D_m}$$

such that for every u in S_m there is a t in T with $\|u-t\|_\infty \leq \delta$.

Remark: Recalling that μ is a probability (and therefore $\|u-t\| \leq \|u-t\|_\infty$), we see that **M"** implies **M'**$_{2,\infty}$ with $r_m = 1$ and $B'_m = c_m[(1/5) \vee (n/D_m)^{1/2}]$. With the constraint $D_m \leq n$ (which we always assume for maximum likelihood estimation) this simplifies to $B'_m = c_m(n/D_m)^{1/2}$.

Proof of Theorem 6: Let us first notice that (3.33) implies (3.34). 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 (6.1) below

$$\|s - \tilde{s}_m\| \leq 2\|s - (s_m \vee n^{-1})\| \leq 2(\|s - s_m\| + n^{-1}).$$

Since $\|s/\tilde{s}_m\|_\infty \leq n\|s\|_\infty$, (6.2) implies that $K(s,\tilde{s}_m) \leq 2[1 + \log(n\|s\|_\infty)]d^2(s,\tilde{s}_m)$, hence the result.

We can now apply either Theorem 8 or Theorem 9 to get our conclusion provided that we are able to check the assumptions **S** and **M'**$_{2,\infty}$. We first choose $L_m = 1 + 2\log(RH)$. Then $D_m L_m \geq D' + 2\log H + 2\log R$ which implies **S**. We shall now content ourselves to check **M"** in order to derive **M'**$_{2,\infty}$. Let us first notice that it is enough to check **M"** for $\bar{S}_m$ instead of S_m. This is clear in the regression case since the clipping operation is a contraction which can only decrease the global entropy. For the maximum likelihood case, it follows from (6.1) that $\|(u/\|u\|) - (v/\|v\|)\| \leq 2\|u-v\|$ when $\|u\|$ and $\|v\| \geq 1/2$; therefore the entropy of S_m follows from the entropy of $\bar{S}_m$ that we shall now evaluate.

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 ε_1/q' that cover this ball is bounded by $[2e(H/\varepsilon_1 + 1)]^{q'}$ (see Lemma 10 below). 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 ε_2/D'. Our aim is to obtain a δ-net in the $\mathbb{L}_\infty$-norm for the family $\bar{S}_m = \{\sum_{j=1}^{D'} \beta_j \phi_{w_j}\}$, with $\sum_{j=1}^{D'} |\beta_j| \leq R$ and $|w_j|_1 \leq H$. We 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, $j = 1,\dots,D'$ and β' 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

$$
\begin{aligned}
|u(x) - t(x)| &\leq \left| \sum_{j=1}^{D'} \beta_j [\phi_{w_j}(x) - \phi_{w'_j}(x)] \right| + \left| \sum_{j=1}^{D'} (\beta_j - \beta'_j) \phi_{w'_j}(x) \right| \\
&\leq \sum_{j=1}^{D'} |\beta_j| |\phi_{w_j}(x) - \phi_{w'_j}(x))| + \sum_{j=1}^{D'} |\beta_j - \beta'_j| H \\
&\leq R\varepsilon_1 + H\varepsilon_2 = \delta,
\end{aligned}
$$

uniformly for x in $[-1,1]^q$. Consequently, assumption **M"** holds with dimension $D_m = D'(q'+1)$ equal to the parameter dimension and $c_m = 8eRH$, and hence $\mathbf{M'}_{2,\infty}$ holds with $B'_m = 8eRH[(n/D_m)^{1/2} \vee 1/5]$. It follows that we can apply either Theorem 8 or Theorem 9 and that in both cases

$$
(L_m + \mathcal{L}'_m) D_m / n \leq \kappa_{11} \left[1 + \log(RH) + \log[1 + n/(D'q')]\right] D'q'/n
$$

which gives the results. □

Remarks: The metric entropy calculations in the Proof of Theorem 6 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.

6 Appendix

We shall first give a proof for Proposition 1. It derives easily from the following

Lemma 4 *Let s^2 be a probability density with respect to μ and t be a function in $\mathbb{L}^2(\mu)$. Then if $t' = t/\|t\|$*

$$
\|s - t'\| \leq \|s - t\| + (1 - \|t\|) \vee 0 \leq 2\|s - t\|; \tag{6.1}
$$

if t^2 is a density then

$$
d^2(s,t) \leq K(s,t) \leq 2[1 + \log(\|s/t\|_\infty)] d^2(s,t), \tag{6.2}
$$

consequently if s^+ is such that $s^+ \geq s$, $s' = s^+/\|s^+\|$ and $\varepsilon = \|s - s^+\|$ then

$$
K(s,s') \leq 2[1 + \log(1 + \varepsilon)]\varepsilon^2. \tag{6.3}
$$

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 (6.1). (6.2) follows from (6.5) of Lemma 4 of Birgé and Massart (1994a) 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 (6.1) and (6.2) imply (6.3). □

Proof of Proposition 1: The first inequality is an immediate consequence of (6.3), considering separately the cases $\varepsilon < .6$ and $\varepsilon \geq .6$. In order to derive (2.13) one notices that if $\tilde{s}$ is such that $\|\tilde{s} - s\|_\infty = \varepsilon$ and μ 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 5 *Let $(a_j)_{j\geq 0}$ and $(b_j)_{j\geq 0}$ two sequences of nonnegative numbers such that the first sequence is nonincreasing, the second is nondecreasing and $a_j < b_j$ for j large enough. Then defining $m = \inf\{j \geq 0 \,|\, a_{j+1} \leq b_j\} < +\infty$ one gets $\sup_j\{a_j \wedge b_j\} = a_m \wedge b_m$ and*

$$\sup_{j\geq 0}\{a_j \wedge b_j\} \leq \inf_{j\geq 0}\{a_{j+1} + b_j\} \leq \inf_{0\leq j\leq m}\{a_{j+1} + b_j\} \leq 2(1 \vee b_0/a_0)\sup_{j\geq 0}\{a_j \wedge b_j\}.$$

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 $a_1 + b_0 \leq 2b_0 \leq 2(b_0/a_0 \vee 1)(a_0 \wedge b_0)$ and the result follows in both cases. □

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 *The following inequality holds for all $n \geq 1$ and $1 \leq D \leq n$:*

$$\sum_{j=0}^{D}\binom{n}{j} < \left(\frac{en}{D}\right)^D.$$

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 Σ 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 $S_{\mathcal{C}}$ be a finite set of densities with respect to μ indexed by $\mathcal{C} = \{0, 1\}^D$ and such that*

$$h^2(s_x, s_y) = \theta\sum_{i=1}^{D}\mathbb{1}_{x_i \neq y_i}.$$

Let $\hat{s}$ be an estimator with values in $S_{\mathcal{C}}$ based on n i.i.d. observations with density s. Then

$$\sup_{s\in S_{\mathcal{C}}}\mathbb{E}_s\left[h^2(s, \hat{s})\right] \geq \frac{D\theta}{2}\left[1 - \sqrt{2n\theta}\right].$$

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 $\mathcal{C}$ be a subset of cardinality $\theta 2^D$ of the cube $\{0,1\}^D$. For any $\eta \in (0,1)$ one can find a subset $\mathcal{C}'$ of $\mathcal{C}$ with cardinality larger than $\theta \exp(D\eta^2/2)$ such that for any two distinct points $x,\ y \in \mathcal{C}'$*

$$\sum_{i=1}^{D} 1\!\!1_{x_i \neq y_i} > D\frac{1-\eta}{2}.$$

Proof: Let $D(1-\eta)/2 = d$ and $\mathcal{C}'$ a maximal subset of $\mathcal{C}$ such that $\sum_{i=1}^{D} 1\!\!1_{x_i \neq y_i} > d$ for any pair $x, y \in \mathcal{C}'$. The number of points z such that $\sum_{i=1}^{D} 1\!\!1_{x_i \neq z_i} \leq k$ is bounded by $\sum_{j=0}^{k} \binom{D}{j}$. It follows from a covering argument that $|\mathcal{C}'| \sum_{j=0}^{[d]} \binom{D}{j} \geq \theta 2^D$. Let B_D be a binomial random variable with parameters D and $1/2$, then

$$|\mathcal{C}'| \geq \theta \left(\mathbb{P}\left[B_D \leq D\frac{1-\eta}{2}\right]\right)^{-1}$$

and the result follows from Hoeffding's inequality. □

Lemma 9 *Let $\{\varphi_\lambda\}_{\lambda\in\Lambda}$ be an orthonormal system in the metric space $(\mathbb{L}^2(\mu), d)$ with $|\Lambda| = D$ and $\bar{S}$ be the linear span of $\{\varphi_\lambda\}$. Let*

$$\bar{r} = \frac{1}{\sqrt{D}} \sup_{\beta \neq 0} \frac{\|\sum_{\lambda\in\Lambda} \beta_\lambda \varphi_\lambda\|_\infty}{|\beta|_\infty}.$$

For any positive δ one can find a countable set $T \subset \bar{S}$ and a mapping π from $\bar{S}$ to T with the following properties:

- *for any ball $\mathcal{B}$ with radius $\sigma \geq 5\delta$*

$$|T \cap \mathcal{B}| \leq (B'\sigma/\delta)^D \qquad \textit{with} \qquad B' < 5. \tag{6.4}$$

- *$d(u, \pi u) \leq \delta$ for all u in $\bar{S}$ and*

$$\sup_{u\in\pi^{-1}(t)} \|u - t\|_\infty \leq \bar{r}\delta \qquad \textit{for all } t \textit{ in } T. \tag{6.5}$$

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 $\tilde{T} = [(\delta/\sqrt{D})\mathbb{Z}]^D$. Considering the partition of $\mathbb{R}^D$ into cubes of vertices with length $\delta/\sqrt{D}$ 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 π is the image of $\tilde{\pi}$ by the natural isometry and clearly $d(u, \pi u) \leq \delta$. The definition of $\bar{r}$ implies (6.5). It follows from Lemma 1 of Birgé and Massart (1994a) that (6.4) holds with $B' = 1.2\sqrt{2\pi e}$. □

Lemma 10 *In $\mathbb{R}^D$, the number of disjoint cubes of vertices ε/D that intersect an l^1-ball of radius R is bounded by*

$$[2e(R/\varepsilon+1)]^D.$$

Proof: An elementary computation shows that the volume of a D-dimensional l^1-ball of radius ρ is equal to $2^D\rho^D/(D!)$. Since all the cubes of vertices ε/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. $\square$

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*

$$\sum_{\lambda\in\Lambda}|\beta_\lambda|\le B''\|\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\|_1$$

for some constant $B''\ge 1$. Given ε, R with $\varepsilon\le R/2$, any ball $\mathcal{B}\in\bar{V}$ of radius R may be covered by N intervals $[f^-,f^+]\subset\mathbb{L}^1$ with diameter $\|f^+-f^-\|_1\le\varepsilon$ and $N\le(3eB'')^D(R/\varepsilon)^D$.

Proof: Without loss of generality we take $\Lambda=\{1,\dots,D\}$ and consider some ball $\mathcal{B}$ in $\bar{V}$ centered at the origin and defined by

$$\mathcal{B}=\left\{\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\ \middle|\ \|\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\|_1\le R\right\}.$$

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 β_λ of any point in this ball satisfy

$$\sum_{\lambda\in\Lambda}|\beta_\lambda|\le B''\|\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\|_1\le B''R$$

$\mathcal{B}$ can be identified to a subset $\mathcal{B}'$ 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 ε/D with $N\le(3eB''R\varepsilon^{-1})^D$. Let $\mathcal{C}$ be the set of the N centers of these cubes. For each $c=(\beta_\lambda)\in\mathcal{C}$ 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 $(\alpha_\lambda)\in\mathcal{B}'$ there exists some $c=(\beta_\lambda)\in\mathcal{C}$ such that $|\alpha_\lambda-\beta_\lambda|\le\varepsilon/(2D)$ for all λ. It follows that

$$\left|\sum_{\lambda\in\Lambda}\alpha_\lambda\varphi_\lambda-\sum_{\lambda\in\Lambda}\beta_\lambda\varphi_\lambda\right|\le\sum_{\lambda\in\Lambda}|\alpha_\lambda-\beta_\lambda|\,|\varphi_\lambda|\le\varepsilon/(2D)\sum_{\lambda\in\Lambda}|\varphi_\lambda|.$$

Therefore $\sum_{\lambda\in\Lambda}\alpha_\lambda\varphi_\lambda\in I_c$ and the intervals $(I_c)_{c\in\mathcal{C}}$ cover $\mathcal{B}$. Moreover the $\mathbb{L}^1$-diameter of each I_c is bounded by

$$\frac{\varepsilon}{D}\sum_{\lambda\in\Lambda}\|\varphi_\lambda\|_1=\varepsilon. \qquad\square$$

Some results in approximation theory:

Lemma 12 *Let s be a function of bounded α-variation on $[0,1]$ with $0<\alpha\le 1$ which means that*

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

where the supremum is taken over all increasing sequence $x_1\le\cdots\le x_k$ of points in $[0,1]$. Let L be any number between 1 and some positive integer N. There exists a partition $\mathcal{P}$ 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 $\mathcal{P}$ such that:

$$D\le 2\left(\frac{N}{L}\right)^{1/(1+2\alpha)}+1 \qquad \text{and} \qquad \|s^+-s\|^2\le J_\alpha^{2\alpha}(s)\left[2\left(\frac{L}{N}\right)^{(2\alpha)/(2\alpha+1)}+\frac{L}{N}\right].$$

Proof: Let $J=J_\alpha(s)$ and for any interval I define $J(I)$ by

$$[JJ(I)]^\alpha=\sup_{y\in I}s(y)-\inf_{y\in I}s(y).$$

Consider a partition $\mathcal{P}$ 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 can check that

$$\|s^+-s\|^2=\sum_{j=1}^{D}\int_{I_j}[s^+(x)-s(x)]^2dx\le\sum_{j=1}^{D}|I_j|[JJ(I_j)]^{2\alpha}. \tag{6.6}$$

Let us now build by induction a partition $\mathcal{P}$ from an increasing sequence $x_0=0<\cdots<x_D=1$ in the following way. Starting with $x_0=0$, define $Nx_{j+1}=\sup\{i\le N \mid L\ge(i-Nx_j)J^{2\alpha}([x_j,i/N))\}$ 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 $J(I)\le 1$ and by construction

$$|I_j|J^{2\alpha}(I_j)\le\frac{L}{N} \qquad \text{for} \qquad 1\le j\le D.$$

Let $I_j^+=[x_{j-1},x_j+1/N)$. Then for $j<D$, $|I_j^+|J^{2\alpha}(I_j^+)>L/N$. Moreover $\sum_{j=1}^{D-1}|I_j^+|\le 2$ and $\sum_{j=1}^{D-1}J(I_j^+)\le 2$, 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 (6.6) that

$$\|s^+-s\|^2\le DJ^{2\alpha}\frac{L}{N}\le J^{2\alpha}\frac{L}{N}\left[2\left(\frac{N}{L}\right)^{1/(2\alpha+1)}+1\right]. \qquad \square$$

Lemma 13 *Let $\mathcal{A}$ be either the interval $[0,1]$ or the one-dimensional torus $\mathbb{T}$. Let s belong to the Besov space $B_{\alpha\, p\, \infty}(\mathcal{A})$ with $\alpha>0$, $1\le p\le\infty$ and let D be a positive integer.*

- *If $\mathcal{A}=[0,1]$ 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 $\mathcal{A}=\mathbb{T}$, let S_2 be the space of trigonometric polynomials on $\mathbb{T}$ with degree $\le D$;*

- *if $\mathcal{A} = [0,1]$, 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 $\{\varphi_\lambda\}$ is a wavelet basis of regularity r.*

Then, there exists constants $C_1(s), C_2(s)$ and $C_3(s)$ such that

$$d_p(s, S_i) \leq C_i(s) D^{-\alpha} \qquad \textit{for } i = 1, 2, 3$$

where d_p denotes the $\mathbb{L}^p$-distance with respect to the uniform distribution on $\mathcal{A}$.

Remark: We recall that $\mathcal{H}(H, \alpha) \subset B_{\alpha\,\infty\,\infty}([0,1])$ with equality when α is not an integer.

Proof: We recall, following DeVore and Lorentz (1993), that a function s belongs to the Besov space $B_{\alpha\,p\,\infty}(\mathcal{A})$ if $r = [\alpha] + 1$ and its r-th order difference given by

$$\Delta_h^r(s, x) = \sum_{k=0}^{r} \begin{pmatrix} r \\ k \end{pmatrix} (-1)^{r-k} s(x + kh)$$

satisfy

$$\sup_{y>0} y^{-\alpha} \omega_r(s, y)_p < +\infty, \qquad \text{where } \omega_r(s, y)_p = \sup_{0<h\leq y} \|\Delta_h^r(s, .)\|_p.$$

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 belongs to the Besov space $B_{\alpha\,p\,\infty}(\mathcal{A})$ if and only if

$$\sup_{j\geq 0} 2^{j(\alpha + \frac{1}{2} - \frac{1}{p})} \left(\sum_{\lambda \in \Lambda(j)} |\beta_\lambda|^p \right)^{1/p} = \|s\|_p < +\infty.$$

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

$$\|s - s_J\|_p^p \leq C'_p \sum_{j>J} \sum_{\lambda \in \Lambda(j)} 2^{j(p/2-1)} |\beta_\lambda|^p \leq C'_p \|s\|_p^p \sum_{j>J} 2^{-jp\alpha}$$

hence the result. $\square$

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), Barron (1993)]. For convenience we restate it here with a slight modification obtained by application of the triangle inequality.

Lemma 14 *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*

$$\|s - s_D\| \leq R/\sqrt{D} + \|s - t\|.$$

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 $\bar{co}\{\pm\phi_w \mid |w|_1 \leq H\}$ where $\bar{co}(\ )$ denotes the closure of the convex hull. Then Lemma 14 with $t = s_H$ provides some $s_m = s_{(D',H,R(s))}$ in $\bar{S}_m$ with

$$\|s - s_m\| \leq \|s - s_H\| + R(s)/\sqrt{D'} \tag{6.7}$$

Let $\tilde{F}(da) = e^{ib_a}F(da)$ denote the phase and magnitude factorization of the complex-valued measure $\tilde{F}$ with phase $|b_a| \leq \pi$. We recall that $s(x) = \int \exp\{ia^T x\}\tilde{F}(da)$, hence since s is real valued

$$s(x) = \int \cos(a^T x + b_a)F(da). \tag{6.8}$$

For trigonometric approximation we assume that $H \geq 2\pi$ and consider $s_H(x) = \int \cos(a^T x + b_a)1_{\{|a|_1 \leq H/2\}}F(da)$ for which the error is bounded by $|s(x) - s_H(x)| \leq \int 1_{\{|a|_1 \geq H/2\}}F(da) \leq c_{s,\alpha}(2/H)^\alpha$ by Markov's inequality. We recognize s_H as an element of $\bar{co}\{\cos(a^T x + b) \mid |a|_1 \leq H/2, |b| \leq H/2\}$ multiplied by a constant not greater than $c_{s,0}$, so that we get from (6.7) $d(s, S_m) \leq c_{s,\alpha}(2/H)^\alpha + c_{s,0}/\sqrt{D'}$ with $m = (D', H, c_{s,0})$ and (3.36) 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 ψ 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

$$s(x) = \frac{1}{\tilde{\psi}(h)} \int_{a\in\mathbb{R}^q} \tilde{F}(hda) \int_{|b+a^T x|\leq 1} \psi(a^T x + b)e^{-ihb}db.$$

Here we assume that $H \geq 4$ and let $s_H(x)$ be the real part of the same quantity with integration with respect to the vector a restricted to $|a|_1 \leq H' = H/2 - 1$. Since $|a^T x| \leq |a|_1$ the value of $|b|$ in the integral is bounded by $\leq H/2$ and $|a|_1 + |b| \leq H$. By assumption $H' \geq 1$ and the error $|s(x) - s_H(x)|$ is bounded by

$$|s(x) - s_H(x)| \quad = \quad \left| \frac{1}{\tilde{\psi}(h)} \int_{|a|_1 > H'} \int_{|b+a^T x|\leq 1} \psi(a^T x + b)e^{-ihb}\tilde{F}(hda)db \right|$$

$$
\begin{aligned}
&\le \frac{2\Psi}{|\tilde{\psi}(h)|}\int_{|a|_1>H'}(1+|a|_1)F(hda) \;\le\; \frac{4\Psi}{|\tilde{\psi}(h)|}\int_{|a|_1>H'}|a|_1 F(hda)\\
&\le \frac{4\Psi}{|h\tilde{\psi}(h)|}\int_{|a'|_1>|h|H'}|a'|_1 F(da') \;\le\; \frac{4\Psi c_{s,\alpha}}{|h|^\alpha|\tilde{\psi}(h)|H'^\alpha}
\end{aligned}
$$

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 (6.7) that $d(s, S_m) \le C_\psi[c_{s,\alpha}/H^\alpha + (c_{s,0} + c_{s,1})/\sqrt{D'}]$ and then (3.36) 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

$$
\begin{aligned}
\cos(b + a^T x) &= \cos(b) - a^T x \sin(b) + \int_0^{a^T x} \cos(t+b)(a^T x - t)\, dt\\
&= \cos(b) - a^T x \sin(b) + \int_0^{|a|_1} \cos(t+b)[(a^T x - t) \vee 0]\, dt\\
&\quad + \int_{-|a|_1}^{0} \cos(t+b)[(t - a^T x) \vee 0]\, dt\\
&= \cos(b) - a^T x \sin(b) + |a|_1^2 \int_0^1 \cos(|a|_1 u + b)\left[\left(\frac{a^T x}{|a|_1} - u\right) \vee 0\right] du\\
&\quad + |a|_1^2 \int_{-1}^{0} \cos(|a|_1 u + b)\left[\left(u - \frac{a^T x}{|a|_1}\right) \vee 0\right] du
\end{aligned}
$$

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 (6.8) that $s(x)$ is equal to $s(0) + (\bigtriangledown s(0))^T x$ 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

$$
a^T x = 2\left[\frac{1}{2}a^T x \vee 0 + \frac{1}{2}\left[-(-a^T x) \vee 0\right]\right].
$$

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 $R(s) \le |s(0)| + 2\|\bigtriangledown s(0)\| + 2c_{s,2}$. The approximation bound (3.36) then follows from (6.7) for all $H \ge 2$ with $\delta_H = 0$. This completes the proof of Proposition 6. $\square$

References

AKAIKE, H. (1973). 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.

ASSOUAD, P. (1983). Deux remarques sur l'estimation. *C. R. Acad. Sc. Paris Sér. I Math.* **296**, 1021-1024.

BARRON, A.R. (1991). Complexity regularization with applications to artificial neural networks. In *Nonparametric functional estimation* (G. Roussas, ed.). Kluwer, Dordrecht, 561-576.

BARRON, A. R. (1993). Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory* **39**, 930-945.

BARRON, A.R. (1994). Approximation and estimation bounds for artificial neural networks. *Machine Learning.* **14**, 115-133.

BARRON, A.R. and COVER, T.M. (1991). Minimum complexity density estimation. *IEEE Transactions on Information Theory* **37**, 1034-1054.

BARRON, A.R. and SHEU C.-H. (1991). Approximation of density functions by sequences of exponential families. *Ann. Statist.* **19**, 1347-1369.

BERGER, M., GAUDUCHON, P. and MAZET, E. (1971). *Le Spectre d'une Variété Riemannienne.* Lecture Notes in Mathematics 194. Springer, Berlin.

BIRGÉ, L. (1983). Approximation dans les espaces métriques et théorie de l'estimation. *Z. Wahrscheinlichkeitstheorie Verw. Geb.* **65**, 181-237.

BIRGÉ, L. (1986). On estimating a density using Hellinger distance and some other strange facts. *Probab. Th. Rel. Fields* **71** , 271-291.

BIRGÉ, L. (1989). The Grenander estimator: a non asymptotic approach. *Ann. Statist.* **17**, 1532-1549.

BIRGÉ, L. and MASSART, P. (1993). Rates of convergence for minimum contrast estimators. *Probab. Theory Relat. Fields* **97**, 113-150.

BIRGÉ, L. and MASSART, P. (1994a). Minimum contrast estimators on sieves. Technical Report. Université Paris-Sud.

BIRGÉ, L. and MASSART, P. (1994b). From model selection to adaptive estimation. Technical Report. Université Paris-Sud.

BIRMAN, M.S. and SOLOMJAK, M.Z. (1967). Piecewise-polynomial approximation of functions of the classes W_p. *Mat. Sbornik* **73**, 295-317.

BREIMAN, L. (1993). Hinging hyperplanes for regression, classification, and function approximation. *IEEE Transactions on Information Theory* **39**, 999-1013.

CENCOV, N.N. (1982). *Statistical Decision Rules and Optimal Inference.* Translations of Mathematical Monographs **53**, American Math. Society, Providence.

CHAVEL, I. (1984). *Eigenvalues in Riemannian Geometry.* Academic Press, Orlando.

CLARKE, B. S. and BARRON, A. R. (1994). Jeffreys' prior is asymptotically least favorable under entropy risk. *Journal of Statistical Planning and Inference* **41**, 37-60.

COX, D.D. (1988). Approximation of least squares regression on nested subspaces. *Annals of Statistics* **16**, 713-732.

DAHMEN, W., DeVORE, R.A. and SCHERER, K. (1980). Multidimensional spline approximation. *SIAM J. Numer. Anal.* **17**, 380-402.

DAUBECHIES, I. (1992). *Ten Lectures on Wavelets.* S.I.A.M., Philadelphia.

DeVORE, R.A. and LORENTZ, G.G. (1993). *Constructive Approximation.* Springer-Verlag, Berlin.

DONOHO, D.L. and JOHNSTONE, I.M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81**, 425-455.

DONOHO, D.L. and JOHNSTONE, I.M. (1994b). Ideal denoising in an orthonormal basis chosen from a library of bases. *C. R. Acad. Sc. Paris Sér. I Math.* **319**, 1317-1322.

DONOHO, D.L. and JOHNSTONE, I.M. (1995). Minimax estimation via wavelet shrinkage. *Annals of Statistics* **23**, to appear.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by wavelet thresholding. Technical report 426. Department of Statistics, Stanford University.

DONOHO, D.L., JOHNSTONE, I.M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. R. Statist. Soc.* B **57**, 301-369.

DUDLEY, R.M. (1978). Central limit theorems for empirical measures. *Ann. Probab.* **6** 899-929.

DUDLEY, R.M. (1984). A course on empirical processes. In *Ecole d'Eté de Probabilités de Saint-Flour XII - 1982.* Lecture Notes in Mathematics 1097, Springer, Berlin.

EFROIMOVICH, S.Yu. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30**, 557-568.

EFROIMOVICH, S.Yu. and PINSKER, M.S. (1981). Estimation of square-integrable density on the basis of a sequence of observations. *Probl. Inf. Transm.* **17**, 182-196.

EFROIMOVICH, S.Yu. and PINSKER, M.S. (1982). Estimation of square-integrable probability density of a random variable. *Probl. Inf. Transm.* **18**, 175-189.

EFROIMOVICH, S.Yu. and PINSKER, M.S. (1984). Learning algorithm for nonparametric filtering. *Automat. Remote Control* **11**, 1434-1440, translated from *Avtomatika i Telemekhanika* **11**, 58-65.

EFROIMOVICH, S.Yu. and PINSKER, M.S. (1986). Self-tuning algorithm for minimax nonparametric estimation of spectral density. *Probl. Inf. Transm.* **22**, 209-221.

GALLAGER, R.G. (1968). *Information Theory and Reliable Communication.* Wiley, New York.

GIROSI, F. and ANZELLOTTI, G. (1992). Convergence rates of approximation by translates. Artificial Intelligence Lab Technical Report 1288, Massachusetts Institute of Technology.

GOLDENSHLUGER, A. and NEMIROVSKI, A. (1994). On spatial adaptive estimation of nonparametric regression. Technical report. Faculty of Industrial Engineering and Management, Technion.

GOLUBEV, G.K. (1990). Quasi-linear estimates of signals in $\mathbb{L}_2$. *Probl. Inf. Transm.* **26**, 15-20.

GOLUBEV, G.K. (1992). Nonparametric estimation of smooth probability densities in $\mathbb{L}_2$. *Probl. Inf. Transm.* **28**, 44-54.

GOLUBEV, G.K. and NUSSBAUM, M. (1992). Adaptive spline estimates for nonparametric regression models. *Theory Probab. Appl.* **37**, 521-529.

HAUSSLER, D. (1991). Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *J. Combinatorial Theory* A **69**, 217-232.

HAUSSLER, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation* **100**, 78-150.

HORNIK, K., STINCHCOMBE, M. B., WHITE, H. and AUER, P. (1994). Degree of approximation results for feedforward networks approximating unknown mappings and their derivatives. *Neural Computation* **6**, 1262-1275.

IBRAGIMOV, I.A. and HAS'MINSKII, R.Z. (1981). *Statistical Estimation - Asymptotic Theory.* Springer, New York.

JONES, L.K. (1992). 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.

KAHANE, J.-P. (1985). *Some random series of functions*, 2nd ed. Cambridge University Press, Cambridge.

KOROSTELEV, A.P. and TSYBAKOV, A.B. (1993a). Estimation of the density support and its functionals. *Probl. Inf. Transm.* **29**, 1-15.

KOROSTELEV, A.P. and TSYBAKOV, A.B. (1993b). *Minimax Theory of Image Reconstruction.* Lecture Notes in Statistics 82, Springer-Verlag, New York.

Le CAM, L.M. (1973). Convergence of estimates under dimensionality restrictions. *Ann. Statist.* **1** , 38-53.

Le CAM, L.M. (1986). *Asymptotic Methods in Statistical Decision Theory.* Springer, New York.

Le CAM, L.M. and YANG, G.L. (1990). *Asymptotics in Statistics: Some Basic Concepts.* Springer, New York.

LEPSKII, O.V. (1991). Asymptotically minimax adaptive estimation I: Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36**, 682-697.

LEPSKII, O.V. (1992). Asymptotically minimax adaptive estimation II: Statistical model without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37**, 433-468.

LI, K.C. (1987). Asymptotic optimality for C_p, C_L, cross-validation, and generalized cross-validation: Discrete index set. *Ann. Statist.* **15**, 958-975.

McGAFFREY, D. F. and GALLANT, A. R. (1994). Convergence rates for single hidden layer feedforward networks. *Neural Networks* **7**, 147-158.

MALLOWS, C.L. (1973). Some comments on C_p. *Technometrics* **15**, 661-675.

MEYER, Y. (1990). *Ondelettes et Opérateurs I.* Hermann, Paris.

POLYAK, B.T. and TSYBAKOV, A.B. (1990). Asymptotic optimality of the C_p-test for the orthogonal series estimation of regression. *Theory Probab. Appl.* **35**, 293-306.

RISSANEN, J. (1978). Modeling by shortest data description. *Automatica* **14**, 465-471.

RISSANEN, J. (1983). A universal prior for integers and estimation by minimum description length. *Annals of Statistics* **11**, 416-431.

SCHUMAKER, L.L. (1981). *Spline Functions: Basic Theory.* Wiley, New York.

SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6**, 461-464.

SHEN, X. and WONG, W.H. (1994). Convergence rates of sieve estimates. *Ann. Statist.* **22**, 580-615.

SHIBATA, R. (1981). An optimal selection of regression variables. *Biometrika* **68**, 45-54.

STEIN, E. M. and WEISS, G. (1971). *Introduction to Fourier Analysis on Euclidean Spaces*, Princeton University Press, Princeton.

STONE, C.J. (1990). Large-sample inference for log-spline models. *Ann. Statist.* **18**, 717-741.

TALAGRAND, M. (1994). Sharper bounds for empirical processes. *Ann. Probab.* **22**, 28-76.

Van de GEER, S. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907-924.

Van de GEER, S. (1993). The method of sieves and minimum contrast estimators. Technical report. University of Leiden.

VAPNIK, V. (1982). *Estimation of Dependences Based on Empirical Data.* Springer, New York.

WAHBA, G. (1990). *Spline Models for Observational Data.* S.I.A.M., Philadelphia.

WHITTAKER, E.T. and WATSON, G.N. (1927). *A Course of Modern Analysis.* Cambridge

University Press, London.

YANG, Y. (1995). A property of model selection criteria. Preprint.

YUKICH, J. E., STINCHCOMBE, M. B. and WHITE, H. (1995). Sup norm approximation bounds for networks through probabilistic methods. To appear in *IEEE Transactions on Information Theory.*

Andrew BARRON
Department of Statistics
Yale University
Box 2179 Yale Station
New Haven, CT 06520 USA
e-mail: BARRON@BRANDY.STAT.YALE.EDU

Lucien BIRGÉ
URA 1321 “Statistique et modèles aléatoires”
L.S.T.A., boîte 158
Université Paris VI, 4 Place Jussieu
F-75252 Paris Cedex 05
France
e-mail: LB@CCR.JUSSIEU.FR

Pascal MASSART
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