Bernoulli **4**(3), 1998, 329–375

Minimum contrast estimators on sieves: exponential bounds and rates of convergence

LUCIEN BIRGÉ[1]* and PASCAL MASSART[2]

[1]*Laboratoire de Probabilités and URA CNRS, Boîte 188, Université Paris VI, 4 Place Jussieu, F-75252 Paris Cedex 05, France. E-mail: lb@jussieu.fr*
[2]*URA CNRS 743, Bât. 425. Université Paris Sud. Campus d'Orsay, F-91405 Orsay Cedex, France. E-mail: Pascal.Massart@math.u-psud.fr*

This paper, which we dedicate to Lucien Le Cam for his seventieth birthday, has been written in the spirit of his pioneering works on the relationships between the metric structure of the parameter space and the rate of convergence of optimal estimators. It has been written in his honour as a contribution to his theory. It contains further developments of the theory of minimum contrast estimators elaborated in a previous paper. We focus on minimum contrast estimators on sieves. By a 'sieve' we mean some approximating space of the set of parameters. The sieves which are commonly used in practice are D-dimensional linear spaces generated by some basis: piecewise polynomials, wavelets, Fourier, etc. It was recently pointed out that nonlinear sieves should also be considered since they provide better spatial adaptation (think of histograms built from any partition of D subintervals of [0, 1] as a typical example). We introduce some metric assumptions which are closely related to the notion of finite-dimensional metric space in the sense of Le Cam. These assumptions are satisfied by the examples of practical interest and allow us to compute sharp rates of convergence for minimum contrast estimators.

Keywords: empirical processes; finite-dimensional metric space; maximum likelihood estimation; minimum contrast estimators; nonparametric estimation; rates of convergence; sieves

1. Introduction

This paper (which originated from a question posed by Peter Bickel in the autumn of 1991) is devoted to further developments of the theory of *minimum contrast estimators* elaborated in Birgé and Massart (1993). Let $Z_1, \ldots, Z_n$ be independent and identically distributed with density $s \in L_\infty([0, 1], \mathrm{d}x)$, S be the linear span of some orthonormal system $\{\varphi_j | j = 1, \ldots, D\}$ and $\hat{s}$ be the projection estimator of s on S, as proposed by Cencov (1962) and defined by

$$\hat{s} = \sum_{j=1}^{D} \hat{\beta}_j \varphi_j, \qquad \text{with } \hat{\beta}_j = \frac{1}{n} \sum_{i=1}^{n} \varphi_j(Z_i).$$

*To whom correspondence should be addressed.

1350–7265

From straightforward computation we obtain

$$\mathrm{E}[\|s-\hat{s}\|_2^2] \leqslant \|s-s^*\|_2^2 + \frac{D}{n}\|s\|_\infty, \tag{1.1}$$

where s^* denotes the orthogonal projection of s onto S and $\|\cdot\|_p$ the L_p-norm with respect to Lebesgue measure. Assuming that we have to hand a family of approximating spaces S of s, an upper bound like (1.1) gives an idea of what should be an optimal choice for S (by minimizing the right-hand side of (1.1)). We also notice that $\hat{s}$ can be defined as the minimizer of the empirical criterion $-2n^{-1}\sum_{i=1}^n t(Z_i) + \|t\|_2^2$ when t varies in S. In that sense the projection estimator is our first example of what we call a *minimum contrast estimator* on the sieve S.

One purpose of this paper is to generalize the preceding upper bound (1.1) to further minimum contrast estimators including maximum likelihood estimators for densities and least-squares estimators for regression. More precisely, given n independent observations $Z_1, \ldots, Z_n$ with a joint distribution depending on an unknown function $s \in L_2(\mu)$, an approximation space S (the sieve) described by D parameters, such as some subset of a D-dimensional linear space or a neural net, and a contrast function $\overline{\gamma}$ (as defined in Birgé and Massart 1993), we consider the minimum contrast estimator $\hat{s}$ which is a minimizer over S of the empirical criterion $\gamma_n(t) = n^{-1}\sum_{i=1}^n \overline{\gamma}(Z_i, t)$. We shall prove, under proper assumptions, that, generally speaking, the behaviour of such estimators is described by

$$\mathrm{E}[\|s-\hat{s}\|_2^2] \leqslant Cd^2(s, S) + \mathscr{L}D/n, \qquad \text{where } d(s, S) = \inf_{t\in S} d(s, t), \tag{1.2}$$

d denoting the distance associated with the L_2-norm and C being a constant depending on the assumptions. Ideally, $\mathscr{L}$ should be bounded independently of n and D leading to the traditional squared bias plus variance upper bound for the risk, but the situation is slightly more complicated and $\mathscr{L}$ is either bounded or of order $\log n$ depending on the structure of the sieve S. A large part of the hard technical work in the proofs will be devoted to a proper control of $\mathscr{L}$.

Let us now explain the need to consider minimum contrast estimators on finite-dimensional sieves, which are typically chosen in order to approximate spaces of smooth functions, rather than on compact subsets of those spaces. It has long been well known (see, for example, Bahadur 1958) that the maximum likelihood estimator can behave very poorly and a simple illustration of this fact in the case of a translation family on the line could be given by the family generated by the density $f(x) = (1/6)[\mathbb{1}_{(0,1)}(|x|)|x|^{-1/2} + \mathbb{1}_{[1,\infty)}(|x|)x^{-2}]$. Such counterexamples usually involve families with unbounded likelihood ratios. More surprising was the fact, described in Birgé and Massart (1993), that even with uniformly bounded likelihood ratios one could get suboptimal rates of convergence (as compared to the minimax risk) for the maximum likelihood estimator when the size of the set of parameters is too large, which essentially means that the set of realizations of the contrast function on the set of parameters is not a Donsker class of functions in the sense of Dudley.

Solutions to the problems connected to the classical maximum likelihood estimator go back to Le Cam (1973) (see also Le Cam 1975; 1986, Section 16.5; or Le Cam and Yang 1990, Section 6.5) and involve discretization of the parameter space or related techniques.

Considering the example of the translation family above, one sees that maximizing the likelihood over a fine discretization of the line instead of the whole line would dramatically improve the situation. Actually most solutions to the problems connected with the nonparametric maximum likelihood estimator are related to the so-called *method of sieves*. The name and formalization of the method are due to Grenander (1981), although the idea of replacing a complicated parameter space by a more tractable one is clearly much older: Cencov's (1962) method of orthogonal series for density estimation is already a sieve technique.

Another advantage of considering a finite-dimensional sieve rather than the whole space of parameters in order to carry out the minimization of the contrast is that it can be more realistic from a numerical (computational) point of view. Classical finite-dimensional sieves are generated by a finite number of parameters and may be D-dimensional linear subspaces of some Hilbert space of functions such as piecewise polynomials, wavelets or Fourier expansions. More recently it was pointed out that some nonlinear sieves should also be considered, such as finite linear combinations of D sigmoidal functions which are studied in the neural network literature (see Barron 1994) or more simply histograms generated by any partition of $[0, 1]$ into D subintervals. In fact these sieves may behave better than linear sieves for approximating some families of functions (Barron 1994). Other examples of nonlinear sieve estimation (which are more sophisticated since they also involve an adaptive choice of D within a wavelet basis) are to be found in Donoho and Johnstone (1994; 1995) for the white-noise model and in Donoho *et al*. (1996) for density estimation. We shall also deal with nonlinear sieves of the above type (including histograms). In this case the function $\mathscr{L}$ in (1.2) is typically of order $\log(n/D)$ and we shall provide a lower bound for histograms which proves that the extra logarithmic factor, as compared to the linear D-dimensional case, is, in some sense, necessary.

Let us briefly review of some important results connected to sieve estimation. Many authors have used sieves or related methods in the past years. Apart from Grenander (see also Chow and Grenander 1985), let us first mention the pioneering work of Geman (1981) and Geman and Hwang (1982). In a series of papers starting with Stone (1990), Stone has extensively studied log-spline density estimation and spline regression (see, in particular, Stone 1994). Related results on regression with fixed design are to be found in Cox (1988). Cox (1988) and Stone (1990; 1994), working with linear sieves, obtained bounds of the type (1.2) with a bounded $\mathscr{L}$. Minimum contrast estimation on general sieves with a special emphasis on maximum likelihood estimation on nets or infinite-dimensional sieves has been recently studied by Shen and Wong (1994), Wong and Shen (1995) and Van de Geer (1995).

Our approach, as in Birgé and Massart (1993), was inspired by Van de Geer (1990) and based on the control of the fluctuations of the centred empirical contrast considered as a process indexed by the set S on which the minimization of the empirical contrast is performed. If we deal with the special case of projection estimators on linear sieves, the process is linear and we can use a simple technique based on a very powerful inequality due to Talagrand (1996). In general, the process is not linear and we must introduce metric-type assumptions on S to handle the fluctuations of this process. The point here is that metric properties are transformed in a controlled way when one takes the image of S by the

contrast while the linearity of S is destroyed if the centred contrast is not linear. This means that the use of metric properties of S does make sense even if S is linear.

Entropy with bracketing is a covering property which is classically used in the context of empirical process theory to derive maximal inequalities (see Dudley 1978; Ossiander 1987). This property was introduced in Birgé and Massart (1993) to study global minimum contrast estimators and is especially well suited to the study of nonparametric maximum likelihood estimators over a set of monotone functions such as the Grenander estimator. It is also the central tool for Shen and Wong (1994), Wong and Shen (1995) and Van de Geer (1995).

We shall dispense with entropy with bracketing in order to take advantage of the finite-dimensional structure of the sieves. We introduce a new metric property for the structure of the sieves which involves covering numbers related to both L_2- and L_∞-norms. This new notion is close but not directly comparable to L_2 with bracketing. The effect of using one metric assumption or another is reflected in the value of $\mathscr{L}$ in (1.2), the main problem being to decide whether $\mathscr{L}$ is bounded or of order $\log n$. In this respect we shall see that there is no systematic superiority of one metric property over the other. We postpone this discussion to the conclusion of the paper.

We shall actually prove substantially more than (1.2) and also derive exponential bounds for the fluctuations of the empirical criterion $\gamma_n(t)$. A first consequence is the possibility of bounding higher moments of $\|s - \hat{s}\|_2$. A deeper consequence of these exponential bounds is developed in Birgé and Massart (1997) and Barron *et al*. (1997). Indeed, these exponential bounds actually allow the construction and study of data-driven procedures for choosing an optimal S among some family of possible approximation spaces leading to adaptive procedures for nonparametric estimation and automatic methods for model selection. These important applications do justify the emphasis we put here on deriving those exponential bounds.

In Section 2 we explain our motivations for developing exponential bounds rather than limiting ourselves to evaluating quadratic risks. We do this in the simplest situation of projection estimators and show how exponential bounds, which are derived by the application of Talagrand's inequality, allow us to construct adaptive estimators. Section 3 is devoted to the study of various types of sieve and of their metric properties. These properties are exploited in Section 4 to treat various applications of our general framework to particular examples of minimum contrast estimators, namely projection and maximum likelihood estimators for densities and least-squares and minimum-L_1 estimators for regression functions. In Section 5 we develop the more general and abstract framework from which all our examples are obtained. Our conclusions are detailed in Section 6, as a number of remarks and comparisons with related work. Finally Section 7 contains the most technical proofs.

2. Motivations

In order to motivate our further developments let us concentrate on one of the simplest sieve methods of estimation, namely projection estimation on a finite-dimensional Fourier

expansion. Our purpose here is to provide a first simple illustration of our approach based on minimum contrast estimation. In particular, we want to insist on the comparison between direct computations (which are possible here since our estimator is explicit) and indirect computations that will later turn out to be more general and widely applicable. Moreover, we want to motivate our emphasis on exponential bounds, by showing how these bounds allow us to construct adaptive estimators which automatically choose the length of the expansion from the data, and to study the performance of those estimators.

Since we shall, for the most part, not deal with precise constants (by 'constant' we always mean a quantity which does not depend on n), in order to make our inequalities more transparent we shall from now on stick to the following convention: κ, κ', κ_1, ... will denote purely numerical constants. On the other hand, C, C', C_1, ... will denote constants that might depend on various parameters introduced in our assumptions. To emphasize this dependence with respect to some parameters, a and b say, we shall write $C(a, b)$. The value of these constants will usually be fixed within a given computation but change each time we start a new evaluation.

Assume that we are given n independent and identically distributed real variables $Z_1, \ldots, Z_n$ with common density s with respect to Lebesgue measure on $[0, 1]$ and that s belongs to $L_2([0, 1])$. Let $\varphi_0 = \mathbb{1}_{[0,1]}$, $\varphi_{2j}(x) = \sqrt{2}\cos(2\pi jx)$ and $\varphi_{2j-1}(x) = \sqrt{2}\sin(2\pi jx)$ for $j \geqslant 1$. We consider the D-dimensional linear space S spanned by $\{\varphi_\lambda | \lambda \in \Lambda_D\}$ where Λ_D is the set of integers smaller than D. The projection estimator relative to S is defined as

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

Let s^* be the orthogonal projection of s onto S; then, denoting by $\|\cdot\|$ the L_2-norm we have $\|s - \hat{s}\|^2 = \|s - s^*\|^2 + \|s^* - \hat{s}\|^2$. Denoting by ν_n the centred empirical operator

$$\nu_n(t) = \frac{1}{n}\sum_{i=1}^{n} t(Z_i) - \int st, \qquad \text{for all } t \in L_2([0, 1]),$$

we can rewrite $\|s^* - \hat{s}\|^2$ as $V^2 = \sum_{\lambda \in \Lambda_D} [\nu_n(\varphi_\lambda)]^2$, which yields

$$\|s - \hat{s}\|^2 = \|s - s^*\|^2 + V^2. \tag{2.1}$$

Since $\mathrm{E}[V^2] = n^{-1}\sum_{\lambda \in \Lambda_D} \mathrm{Var}(\varphi_\lambda)$, the quadratic risk is bounded by

$$\mathrm{E}[\|s - \hat{s}\|^2] = \|s - s^*\|^2 + \mathrm{E}[V^2] \leqslant \|s - s^*\|^2 + \frac{1}{n}\mathrm{E}\left[\sum_{\lambda \in \Lambda_D} \varphi_\lambda^2\right].$$

Restricting ourselves, as is natural, to odd values of D, we notice that

$$\left\| \sum_{\lambda \in \Lambda_D} \varphi_\lambda^2 \right\|_\infty = D \tag{2.2}$$

and therefore $\mathrm{E}[\sum_{\lambda \in \Lambda_D} \varphi_\lambda^2] \leqslant D$. This implies that

$$\mathrm{E}[\|s - \hat{s}\|^2] \leqslant \|s - s^*\|^2 + D/n, \tag{2.3}$$

and the first term clearly depends on the L_2 approximation properties of the Fourier basis with respect to the unknown s. The two terms in the upper bound are respectively non-increasing and increasing with respect to D and therefore one should choose D so as approximately to equate those terms in order to minimize the risk. A classical way of solving this problem is to put an a priori assumption on the smoothness of s. For instance, if $s = \sum_{\lambda \geqslant 0} \beta_\lambda \varphi_\lambda$ belongs to the Sobolev space W_2^α of the torus $\mathbb{R}/\mathbb{Z}$ with $\alpha \in \mathbb{N} - \{0\}$, i.e. $\|s^{(\alpha)}\|^2 < +\infty$, then it follows from Parseval's formula that

$$\|s^{(\alpha)}\|^2 = \sum_{j \geqslant 1} (2\pi j)^{2\alpha} (\beta_{2j-1}^2 + \beta_{2j}^2).$$

Since

$$\|s - s^*\|^2 = \sum_{j \geqslant (D+1)/2} (2\pi j)^{2\alpha} (\beta_{2j-1}^2 + \beta_{2j}^2)(2\pi j)^{-2\alpha},$$

we obtain

$$\|s - s^*\|^2 \leqslant [\pi(D+1)]^{-2\alpha} \|s^{(\alpha)}\|^2. \tag{2.4}$$

Choosing D approximately equal to $[n\|s^{(\alpha)}\|^2]^{1/(1+2\alpha)}$ gives

$$\mathrm{E}[\|s - \hat{s}\|^2] \leqslant \kappa \|s^{(\alpha)}\|^{2/(1+2\alpha)} n^{-2\alpha/(1+2\alpha)} \tag{2.5}$$

for some numerical constant κ.

We can actually recover (2.3) (up to multiplicative constants) by an indirect method which benefits from the fact that the projection estimator is a minimum contrast estimator: more precisely, if

$$\gamma_n(t) = -\frac{2}{n} \sum_{i=1}^n t(Z_i) + \|t\|^2,$$

one can easily check that $\hat{s}$ is the minimizer of γ_n on S. Starting from the identity $\|t - s\|^2 = \gamma_n(t) + 2\nu_n(t) + \|s\|^2$, we derive

$$\|t - s\|^2 = \|s - s^*\|^2 + \gamma_n(t) - \gamma_n(s^*) + 2\nu_n(t - s^*)$$

and, since $\gamma_n(\hat{s}) \leqslant \gamma_n(s^*)$,

$$\|\hat{s} - s\|^2 \leqslant \|s - s^*\|^2 + 2\nu_n(\hat{s} - s^*). \tag{2.6}$$

Recalling that $V^2 = \sum_{\lambda \in \Lambda_D} [\nu_n(\varphi_\lambda)]^2$, it follows by a standard duality argument that

$$V^2 = \sup_{t \in S} \frac{[\nu_n(t - s^*)]^2}{\|t - s^*\|^2}, \tag{2.7}$$

and therefore, by (2.6),

$$\|\hat{s} - s\|^2 \leqslant \|s - s^*\|^2 + 2\|\hat{s} - s^*\| V \leqslant \|s - s^*\|^2 + \tfrac{1}{2}\|\hat{s} - s^*\|^2 + 2V^2$$

which finally yields by Pyrthagoras's identity the following analogue of (2.1):

$$\|\hat{s} - s\|^2 \leqslant \|s - s^*\|^2 + 4V^2.$$

A first benefit of this indirect approach is equation (2.7) which allows us to compute higher-order moments of V (and therefore higher-order moments of $\|\hat{s} - s\|$) from the second moment using a powerful isoperimetric-type inequality due to Talagrand (see Theorem 1 below). A second benefit of this approach is to suggest a minimizing procedure which allows an automatic choice of D from the data.

Let us now assume that instead of one single sieve S we have at our disposal the whole family of sieves S_D corresponding to the different odd values of D. We can try to choose D in an optimal way. Actually the main defect of the above computation of the optimal value of D is its dependence with respect to some given Sobolev norm on the unknown density s. One would prefer a procedure of estimation ignoring any special feature of s and leading approximately to the same risk. Let us denote by s_D the orthogonal projection of s onto S_D and by $\hat{s}_D$ the minimum contrast estimator corresponding to S_D. Then the following method will solve this problem: we consider $\hat{D}$ minimizing $\gamma_n(\hat{s}_D) + 2n^{-1}D$ with respect to $D \leqslant n$ and define $\tilde{s} = \hat{s}_{\hat{D}}$. Arguing as above, one derives that

$$\|\tilde{s} - s\|^2 \leqslant \|s - s_D\|^2 + 2\nu_n(\tilde{s} - s_D) + 2n^{-1}(D - \hat{D}) \tag{2.8}$$

for all D. If $\hat{D}$ were fixed, one could again conclude by an analogue of (2.7) and take expectations. Since $\hat{D}$ can take all possible odd values between 1 and n it becomes crucial to use exponential bounds that can be summed over all possible values of $\hat{D}$. Let us first recall an important result of Talagrand (1996) (see also Ledoux 1996).

Theorem 1 (Talagrand's inequality). *Let $Z_1, \ldots, Z_n$ be n independent and identically distributed variables and $\mathscr{F}$ a countable family of functions the absolute values of which are uniformly bounded by some constant b. Let*

$$\Sigma^2 = \frac{1}{n}\mathrm{E}\left[\sup_{f\in\mathscr{F}} \sum_{i=1}^{n} f^2(Z_i)\right].$$

There exists a universal constant κ_1 such that, for any positive ξ,

$$\mathrm{P}\left[\sup_{f\in\mathscr{F}} |\nu_n(f)| \geqslant E + \xi\right] \leqslant 3\exp\left[\frac{-n\kappa_1\xi^2}{\Sigma^2 + b\xi}\right] \qquad \text{if } E \geqslant \mathrm{E}\left[\sup_{f\in\mathscr{F}} |\nu_n(f)|\right]. \tag{2.9}$$

We shall actually prove in Section 7, as a consequence of Talagrand's inequality, that, for any D,

$$\sum_{D'=1}^{n} \mathrm{P}_s\left[\sup_{t\in S_{D'}} \frac{\nu_n(t - s_D)}{\|t - s_D\|} \geqslant (2\eta + 1)\left(\frac{D \vee D'}{n}\right)^{1/2} + \frac{x}{\sqrt{n}}\right] \leqslant C(\eta, \|s\|)\exp\left[-\kappa\frac{(\eta \wedge 1)x}{1 + \|s\|}\right] \tag{2.10}$$

from which we shall derive, using (2.8), that, apart from a set with a probability bounded by the right-hand side of (2.10),

$$\|\tilde{s} - s\|^2 \leqslant \kappa'[\|s - s_D\|^2 + n^{-1}(D + x^2)]. \tag{2.11}$$

Finally, introducing the power $q \geqslant 1$ and integrating with respect to x leads, since D is arbitrary, to

$$\mathrm{E}[\|\tilde{s} - s\|^q] \leqslant C'(\|s\|, q) \inf_{D \leqslant n} [\|s - s_D\|^q + (D/n)^{q/2}]. \tag{2.12}$$

If one wants to deal with other minimum contrast estimators like maximum likelihood estimators then two problems occur. First, the direct approach used at the beginning of this section is no longer possible since we do not have an explicit formula for the estimator. Second, one needs exponential inequalities which are analogues of (2.10), and unfortunately Talagrand's inequality is only really useful for linear contrast functions. The main purpose of this paper is to solve those two problems by systematically developing what we called the 'indirect approach' and the relevant exponential inequalities. These inequalities will follow from the use of two main techniques, namely entropy methods with chaining or isoperimetric methods (Talagrand's inequality). We will use them here to derive upper bounds for the risk of minimum contrast estimators on a given sieve. Penalized estimators (analogues of $\tilde{s}$) and their adaptive properties are studied in Birgé and Massart (1997) and Barron *et al*. (1997), using in an essential way the exponential bounds which are proved in the present paper.

3. The sieves

The sieves that we shall consider here are essentially the classical spaces one would use in approximation theory, such as trigonometric polynomials, wavelet expansions, piecewise polynomials, neural nets, etc., and these sieves are supposed to provide a good approximation to the unknown function s to be estimated. In practice, one does not work with one single sieve S but with a whole collection of them since one wants to adjust the choice of the sieve to the number n of observations to hand. In order to obtain uniform results with respect to n we shall consider collections of sieves with some uniformity properties. Therefore in the following when we speak of a constant, we shall mean a constant which is the same for all elements of a given collection of sieves and which is, in particular, independent of n. From now on, we denote by $|\Lambda|$ the cardinality of the set Λ.

3.1. Linear sieves

3.1.1. Two useful indices

In many situations, S is a subset of the linear D-dimensional subspace $\overline{S}$ of $L_2 \cap L_\infty(\mu)$ spanned by some orthonormal basis $\{\varphi_\lambda | \lambda \in \Lambda_D\}$ with $|\Lambda_D| = D$. For a given linear sieve, we introduce two indices, $\bar{r}$ and Φ, describing the relationships between its L_2 and L_∞ structures. They will be involved in our upper bounds for the risk of minimum contrast estimators on this sieve. For $1 \leqslant p \leqslant \infty$ we shall denote by $|\cdot|_p$ the usual l_p-norm in $\mathbb{R}^{\Lambda_D}$ and by $\|\cdot\|_p$ the L_p-norm on $\overline{S}$. Let

$$\Phi = \frac{1}{\sqrt{D}} \sup_{t \in \bar{S}, t \neq 0} \frac{\|t\|_\infty}{\|t\|_2}, \qquad \bar{r}_\varphi = \frac{1}{\sqrt{D}} \sup_{\beta \in \mathbb{R}^{\Lambda_D}, \beta \neq 0} \frac{\|\sum_{\lambda \in \Lambda_D} \beta_\lambda \varphi_\lambda\|_\infty}{|\beta|_\infty} \tag{3.1}$$

and $\bar{r}$ be the infimum of $\bar{r}_\varphi$ when the basis $\{\varphi_\lambda | \lambda \in \Lambda_D\}$ varies in the set of all possible orthonormal bases of $\bar{S}$. It follows from this definition that

$$\Phi \leqslant \bar{r} \leqslant \Phi\sqrt{D}. \tag{3.2}$$

Moreover, we have the following lemma.

Lemma 1. *Let $\bar{S}$ be a D-dimensional linear subspace of $L_2 \cap L_\infty(\mu)$ with basis $\{\varphi_\lambda | \lambda \in \Lambda_D\}$; then*

$$\Phi\sqrt{D} = \sup_{t \in \bar{S}, t \neq 0} \frac{\|t\|_\infty}{\|t\|_2} = \left\| \sum_{\lambda \in \Lambda_D} \varphi_\lambda^2 \right\|_\infty^{1/2}. \tag{3.3}$$

Proof. The proof follows from the fact that, for any $x \in \bar{S}$, one has

$$\left[\sum_{j=1}^{D} \varphi_j^2(x) \right]^{1/2} = \sup_{\beta_\lambda \in \mathbb{R}^{\Lambda_D}, \beta_\lambda \neq 0} \frac{|\sum_{\lambda \in \Lambda_D} \beta_\lambda \varphi_\lambda(x)|}{|\beta|_2}. \qquad \square$$

As we shall see later, the value of $\bar{r}$ will have some influence on the value of $\mathscr{L}$ in (1.2), large values of $\bar{r}$ together with large values of D (as compared to n) resulting in values of $\mathscr{L}$ of order $\log n$.

3.1.2. Examples

Trigonometric expansions. Using the trigonometric basis defined at the beginning of Section 2, we consider the sieve S generated by the D first elements of the basis. It then follows from (3.3) that Φ^2 is bounded by $(D+1)/D \leqslant 2$ and therefore that $\bar{r} \leqslant \sqrt{2D}$.

Polynomials. Let S be the linear space of polynomials on [0, 1] with degree bounded by $D-1$. It follows from Remark 1 of Barron and Sheu (1991, p. 1362) that $\Phi = \sqrt{D}$ and therefore from (3.2) that $\bar{r} \leqslant D$.

Localized bases. The orthonormal system $\{\varphi_j = \sqrt{D}\mathbb{1}_{[(j-1)/D, j/D)}\}_{1 \leqslant j \leqslant D}$ that generates regular histograms on [0, 1] is typical of this case. For the linear span of $\{\varphi_j\}_{1 \leqslant j \leqslant D}$ one can immediately check that $\bar{r} \leqslant 1$. This means, in particular, that $\bar{r}$ remains uniformly bounded over the class of all regular histograms. This property is shared by further related families of sieves. For piecewise polynomials on [0, 1] with regular partitions and degree bounded by m, using the Legendre basis on each piece of the partition, one can check that $\bar{r} \leqslant 2m+1$. We can alternatively consider an orthonormal wavelet system of $L_2(\mathbb{R}, \mathrm{d}x)$: $\{\phi_{j,\lambda} = 2^{j/2}\phi(2^j x - \lambda) | j \geqslant 0, \lambda \in \mathbb{Z}\}$, where ϕ is a compactly supported wavelet (for details, see Meyer 1990), and consider the linear sieve $\bar{S}$ generated by the family $\{\phi_{j,\lambda} | \lambda \in \Lambda_D\}$, where Λ_D is the set of indices λ such that the intersection between the support of $\phi(2^j x - \lambda)$

and (0, 1) is non-empty. In this case $M2^j \geqslant D = |\Lambda_D| \geqslant M'2^j$ for some constants M, M' depending on the size of the support of ϕ. It comes from Bernstein's inequality (see Meyer 1990, Chapter 2, Lemma 8) that if $t = \sum_{\lambda \in \Lambda_D} \beta_\lambda \phi_{j,\lambda}$ then $\|t\|_\infty \leqslant C(\phi)2^{j/2}|\beta|_\infty \leqslant C'(\phi)\sqrt{D}|\beta|_\infty$, where the constant C' depends only on the wavelet ϕ. This shows that in this case again $\bar{r} \leqslant C'(\phi)$ is bounded independently of j or D.

3.1.3. Metric interpretation

Apart from the special situation of projection estimators on linear sieves, we shall need for our proofs some metric properties of the sieve, instead of its linear features. In this case we therefore have to interpret the linear dimension and the index $\bar{r}$ in terms of metric characteristics of the sieve. We first recall the definition of a net:

Definition 1. *Given a set $\mathscr{B}$ and a distance d on some metric space containing $\mathscr{B}$, we say that a subset T of this metric space is an ε-net for $\mathscr{B}$ (with $\varepsilon > 0$) if for any point $u \in \mathscr{B}$ one can find a point $t \in T$ such that $d(u, t) \leqslant \varepsilon$. We say that the ε-entropy of $\mathscr{B}$ is bounded by $H(\varepsilon)$ if one can find an ε-net T for $\mathscr{B}$ with cardinality bounded by $\mathrm{e}^{H(\varepsilon)}$.*

Proposition 1. *Let $\bar{S}$ be a D-dimensional linear subspace of $L_2 \cap L_\infty$ with its index $\bar{r}$ defined above. Let $\mathscr{B}$ be any ball of radius σ in $\bar{S}$ and $0 < \delta < \sigma/5$. Then there exists a finite set $T \subset \mathscr{B}$ which is simultaneously a δ-net for $\mathscr{B}$ with respect to the L_2-norm and an $\bar{r}\delta$-net with respect to the L_∞-norm and such that $|T| \leqslant (6\sigma/\delta)^D$.*

The proof is based on the following lemma:

Lemma 2. *In $\mathbb{R}^D$, the maximal number of disjoint cubes of vertices $\delta'/\sqrt{D}$ that intersect a ball of radius σ is bounded by $(2\pi\mathrm{e})^{D/2}(1 + \sigma/\delta')^D$.*

Proof. An elementary computation using the exact formula for the volume of Euclidean balls and Stirling's formula with correction (see Feller 1968, p. 54) shows that the volume of a D-dimensional ball of radius σ is bounded by $(2\pi\mathrm{e}/D)^{D/2}(\pi D)^{-1/2}\sigma^D$. The result follows easily. □

Proof of Proposition 1. Let $\{\varphi_\lambda | \lambda \in \Lambda_D\}$ be an orthonormal basis for $\bar{S}$ such that $\bar{r}_\varphi \leqslant 1.1\bar{r}$. We shall repeatedly use the natural isometry between $\bar{S}$ with its basis $\{\varphi_\lambda | \lambda \in \Lambda_D\}$ and $\mathbb{R}^D$ with its canonical basis. If we consider in $\mathbb{R}^D$ the Euclidean ball $\mathscr{B}'$ of radius σ isometric to $\mathscr{B}$, we can cover it with cubes of vertices $\delta/(1.1\sqrt{D})$ and build a net $T' \subset \mathscr{B}'$ choosing one point in each cube. T' will be a δ-net for the Euclidean metric, a $[\delta/(1.1\sqrt{D})]$-net for the sup-norm metric and its cardinality can be bounded using Lemma 2. Then T is defined from T' through the isometry and it is both a δ-net for the L_2 metric and an $\bar{r}\delta$-net for the L_∞ metric in $\mathscr{B}$. The result follows since

$$(2\pi\mathrm{e})^{1/2}\left(1 + 1.1\frac{\sigma}{\delta}\right) = (2\pi\mathrm{e})^{1/2}\frac{\sigma}{\delta}\left(1.1 + \frac{\delta}{\sigma}\right) < 1.3(2\pi\mathrm{e})^{1/2}\frac{\sigma}{\delta} < 6\frac{\sigma}{\delta}.$$ □

3.2. Nonlinear sieves

3.2.1. Neural nets

We want to consider here the sieves connected with neural networks techniques as considered by Barron (1994). One starts with a sigmoidal function φ, i.e. a continuous bounded function on the real line such that $\varphi(x) \to -1$ when $x \to -\infty$ and $\varphi(x) \to 1$ when $x \to +\infty$. Moreover, one assumes that $\sup_x |\varphi(x)| \leqslant v$ and that for all real numbers x and y, $|\varphi(x) - \varphi(y)| \leqslant v|x - y|$. One can now, following the notation of Barron (1994), define a sieve of functions on $\mathbb{R}^k$ by

$$S(D', \tau, G) = \left\{ \sum_{j=1}^{D'} g_j \varphi(\langle a_j, x \rangle + b_j) + g_0 \right\}$$

with

$$a_j \in \mathbb{R}^k, |a_j|_1 \leqslant \tau; \qquad b_j \in \mathbb{R}, |b_j| \leqslant \tau;$$

$$g_j \in \mathbb{R}, \sum_{j=1}^{D'} |g_j| \leqslant G; \qquad g - G \leqslant g_0 \leqslant g + 5\tau + G \text{ for some given } g.$$

The approximation properties of this class are studied in Barron (1994) when the functions to be approximated are defined on the cube $[0, 1]^k$ with values between g and $g + 5\tau$. Barron proves that the number of functions needed to obtain a δ-net in $S(D', \tau, G)$ with respect to the uniform distance over the cube $[0, 1]^k$ is bounded by $[8vGe(1 + \tau)/\delta]^{D'(k+2)+1}$. Of course this net has similar approximation properties with respect to the L_2 distance on $[0, 1]^k$.

3.2.2. Histograms with bounded support

Another very interesting class of nonlinear sieves for approximating functions on a given compact hyperrectangle of $\mathbb{R}^k$ is the class of piecewise polynomials of bounded degree with a given number of pieces. To be more precise, let us assume that $k = 1$. The difference with piecewise polynomials on a fixed partition, which are a particular case of what we called 'linear sieves with a localized basis' is the fact that the number of bins in the partition is fixed but their lengths are arbitrary. These spaces are particularly interesting because of good approximation properties and also because one can easily restrict oneself to the positive part of the sieve when one wants to approximate non-negative functions such as densities, which is clearly more difficult in the case of wavelet approximation, for example. The metric structure of this inhomogeneous space is also very different from the structure of its linear analogue. In order to make our illustrations as simple as possible, we shall be content to deal with the case of histograms (piecewise polynomials of degree 0), the case of piecewise polynomials of higher degree being similar but more complicated.

Without loss of generality, we shall restrict ourselves to histograms on $[0, 1]$ with D pieces. This motivates the introduction of the following spaces:

$$\overline{\mathscr{H}}(D) = \left\{ \sum_{j=1}^{D} a_j \mathbb{1}_{[j-1/D.j/D)} \right\}$$

and, for $D \leqslant N$,

$$\mathscr{H}_N(D) = \left\{ \sum_{j=1}^{D} a_j \mathbb{1}_{[N_{j-1}/N, N_j/N)},\ 0 = N_0 \leqslant N_1 \leqslant \cdots \leqslant N_D = N,\ N_j \in \mathbb{N} \text{ for all } j \right\}.$$

To begin with it should be noticed that the metric structures of those spaces are substantially different, as are their approximation properties. $\overline{\mathscr{H}}(D)$ has a nicer metric structure since it is a D-dimensional linear sieve with a localized basis but its approximation properties are poor as compared to those of $\mathscr{H}_N(D)$. This is readily seen from the consideration of a spatially inhomogeneous density such as $(\mathbb{1}_{[0,1)} + N\mathbb{1}_{[0,1/N)})/2$. In order to obtain a good approximation of this density by a regular histogram we should take $D \asymp N$ instead of $D = 2$. On the other hand, the dimensional properties of $\mathscr{H}_N(D)$ with respect to the uniform metric deteriorate when N becomes large since, for instance, one can find $f \in \mathscr{H}_N(D)$ with $\|f\|_\infty = \sqrt{N}\|f\|_2$. As will become obvious later, one restricts the end-points of the underlying partition to the grid $\{j/N\}_{0 \leqslant j \leqslant N}$ in order to control the L_∞ properties of the sieve. Actually, $\mathscr{H}_N(D)$ is a union of some number of linear sieves and for each such linear sieve $\bar{r}$ can be computed and is seen to be bounded by $(N/D)^{1/2}$. Therefore the metric properties of $\mathscr{H}_N(D)$ can be analysed using the following lemma.

Lemma 3. *Let $\mathscr{H}$ be a finite union of at most K^D linear sieves, each of dimension bounded by D with an index $\bar{r}$ bounded by r'. Let $0 < \delta < \sigma/5$ and $\mathscr{B}$ be any ball of radius σ in $\mathscr{H}$. Then there exists a finite set $T \subset \mathscr{B}$ which is simultaneously a δ-net for $\mathscr{B}$ with respect to the L_2-norm and an $r'\delta$-net with respect to the L_∞-norm such that $|T| \leqslant (6K\sigma/\delta)^D$.*

Proof. According to Proposition 1, each linear component of $\mathscr{H}$ leads to a net of cardinality bounded by $(6\sigma/\delta)^D$. The union of those nets gives the required set T and its cardinality is therefore bounded by $K^D(6\sigma/\delta)^D$. □

It can easily be seen that the number of linear sieves (histograms built on a given partition) needed to build $\mathscr{H}_N(D)$ is bounded by

$$\sum_{j=0}^{D} \binom{N}{j} < \left(\frac{\mathrm{e}N}{D}\right)^D$$

(see Dudley 1984, Proposition 9.1.5, for an analogous result). Therefore one can take $K = \mathrm{e}N/D$ in the previous lemma and get a δ-net T for $\mathscr{H}_N(D)$ with cardinality bounded by $[(6\mathrm{e}N/D)(\sigma/\delta)]^D$ which is also an $(N/D)^{1/2}\delta$-net in L_∞ distance.

3.2.3. Metric characteristics of general sieves

We want to give a sense to the notion of dimension for a general sieve in such a way that it

coincides with the one we have for linear sieves. As already observed in Proposition 1, this notion is related to some entropy counts. The basic idea is to extend the notion of dimension related to entropy counts to nonlinear sieves. Difficulties arise then from the fact that in some cases (see the example of neural nets) one cannot guarantee that the entropy of a small ball is essentially smaller than the entropy of a big one. We shall therefore consider the following covering property which is satisfied by all the above examples.

Covering Property M(η). *Given $\eta > 0$, we shall say that a subset S of $L_2 \cap L_\infty(\mu)$ satisfies Covering Property* M(η) *if there exist positive numbers D, B' and $r' \geqslant 1$, possibly depending on η, such that, for any $\delta \geqslant \eta$, $\sigma \geqslant 5\delta$ and any ball $\mathscr{B}$ of radius σ with respect to L_2, one can find a finite subset T of $\mathscr{B}$ which is simultaneously a δ-net of $\mathscr{B}$ for the L_2-norm and an $r'\delta$-net for the L_∞-norm such that $|T| \leqslant (B'\sigma/\delta)^D$.*

Covering Property M. *We shall say that a subset S of $L_2 \cap L_\infty(\mu)$ satisfies Covering Property* M *if it satisfies Covering Property* M(η) *for any positive η with values of D, B' and r' independent of η.*

Examples.

1. *Linear sieves*. If $\overline{S}$ is a D-dimensional linear subspace of $L_2 \cap L_\infty(\mu)$, it follows from Proposition 1 that $\overline{S}$ satisfies Covering Property M and B' can be taken as 6 while r' may be taken as $\bar{r}$.

2. *Neural nets*. Considering the sieve $S(D', \tau, G)$ defined in Section 3.2.2, we see from Barron's entropy computations that it satisfies Covering Property M(η) with $D = D'(k+2)+1$, $B' = 8vG\mathrm{e}(1+\tau)\eta^{-1/2}$ and $r' = 1$.

3. *Histograms*. Let us consider the class $\mathscr{H}_N(D)$. It follows from the computations of Section 3.2.2 that Covering Property M is satisfied with $B' = 6\mathrm{e}N/D$ and $r' = (N/D)^{1/2}$.

4. *ε-nets*. Assume that μ is a probability measure and consider some set of functions Θ which is totally bounded for the L_∞ metric. Let S_ε be an ε-net of Θ with minimal cardinality which means that $\log(|S_\varepsilon|) = H(\varepsilon)$, where H denotes the metric entropy of Θ. Then S_ε satisfies Covering Property M with $B' = r = 1$ and $D = H(\varepsilon)$.

Later we shall give several illustrations involving our first three examples and shall neglect the last one that we include here for the sake of completeness since it has been extensively studied in the literature on sieves (see, for instance, Shen and Wong 1994; Wong and Shen 1995; and Van de Geer 1995). We shall explain in Section 6 why we prefer not to dwell on this example.

We conclude this section by mentioning a related notion of finite-dimensional metric space. Following Le Cam (1975), let us say that a metric space (S, d) has an η-dimension D when any ball of radius 2σ can be covered by at most 2^D sets of diameter bounded by 2σ for any $\sigma \geqslant \eta$. The space is *finite-dimensional* (or *D-dimensional*) when one can choose D independently of η and *infinite-dimensional* if D tends to infinity when η goes to 0. In the first cast Le Cam has shown that, in some sense, D represents the number of parameters (in the sense of parametric estimation) when S is a space of square roots of densities and d is the L_2 distance and that the minimax quadratic risk is then of order D/n

for estimating the square root of a density belonging to S from n independent and identically distributed observations. Unfortunately, the estimators designed by Le Cam cannot be used for practical purposes. In that context, Covering Property M should be seen as an enforcement of Le Cam's notion of dimension that allows us to analyse the performance of a more practical estimator, namely the maximum likelihood estimator on S, as we shall see in Section 4.2 below.

4. Some illustrated results

Let $Z_1, \ldots, Z_n$ be independent random variables from some measurable space $(\Omega, \mathscr{A})$ to the measurable space $(\mathscr{Z}, \mathscr{C})$. The space $(\Omega, \mathscr{A})$ is equipped with a family of probabilities $\{P_s\}_{s\in\mathscr{S}}$ indexed by a set $\mathscr{S}$ of parameters which is included in $L_2(\mu)$, where μ is some positive measure (which may depend on n). We denote by E_s the expectation relative to P_s, by $\|\cdot\|$ the norm in $L_2(\mu)$ and by d the associated metric. Considering some function $\overline{\gamma}$ defined on $\mathscr{Z}\times\mathscr{T}$ where $\mathscr{S}\subset\mathscr{T}\subset L_2(\mu)$, we set $\gamma_n(t) = n^{-1}\sum_{i=1}^n \overline{\gamma}(Z_i, t)$ and we say that γ_n is an *empirical contrast function* if

$$E_s[\gamma_n(t)] \geqslant E_s[\gamma_n(s)], \qquad \text{for all } t\in\mathscr{T} \text{ and } s\in\mathscr{S}.$$

Let S be a subset of $\mathscr{T}$; a *minimum contrast estimator* relative to the sieve S is any measurable minimizer $\hat{s}(Z_1, \ldots, Z_n)$ on S of the function $\sum_{i=1}^n \overline{\gamma}(Z_i, t)$. More generally, we have the following definition.

Definition 2. *If ε is non-negative, an* ε-minimum contrast estimator *(relative to the sieve S) is any measurable $\hat{s}$ in S such that*

$$\gamma_n(\hat{s}) \leqslant \inf_{t\in S}\gamma_n(t) + \varepsilon, \qquad \textit{where } \gamma_n(t) = \frac{1}{n}\sum_{i=1}^n \overline{\gamma}(Z_i, t).$$

Since a large value of ε could result in bad behaviour by the estimator, we shall from now on assume that $\varepsilon \leqslant 1/n$. We emphasize the fact that s does not necessarily belong to S. We should think of S as a D-dimensional metric space (to be more precise, of S fulfilling Covering Property $M(n^{-1/2})$) which may depend on n (and therefore D as well).

4.1. Projection estimators for density estimation

In this case we observe n independent and identically distributed random variables $Z_1, \ldots, Z_n$ of density s with respect to μ, where s belongs to $L_2(\mu)$. For any $t\in L_2(\mu)$ we define

$$\gamma_n(t) = -\frac{2}{n}\sum_{i=1}^n t(Z_i) + \|t\|^2.$$

Since $E_s[\gamma_n(t) - \gamma_n(s)] = \|t-s\|^2$, γ_n is an empirical contrast function. When S is the D-

dimensional linear space spanned by some orthogonal system $\{\varphi_\lambda | \lambda \in \Lambda_D\}$ any t in S can be written as $\sum_{\lambda\in\Lambda_D}\alpha_\lambda\varphi_\lambda$ and

$$\gamma_n(t) = \sum_{\lambda\in\Lambda_D}\alpha_\lambda^2 - \frac{2}{n}\sum_{\lambda\in\Lambda_D}\alpha_\lambda\sum_{i=1}^{n}\varphi_\lambda(Z_i)$$

so that the minimum contrast estimator on S is simply the classical projection estimator of Cencov (1962):

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

The following theorem is based on the interpretation of $\hat{s}$ as a minimum contrast estimator and will be proved in Section 7 by an application of Talagrand's inequality.

Theorem 2. *Let $\hat{s}$ be the projection estimator on S based on n independent and identically distributed observations of density s. Assume that S is a D-dimensional linear space and define*

$$\Phi = \frac{1}{\sqrt{D}}\sup_{t\in S, t\neq 0}\frac{\|t\|_\infty}{\|t\|}, \qquad d(s, S) = \inf_{t\in S} d(s, t).$$

Then, for any $q \geqslant 1$, one has

$$\mathrm{E}_s[\|\hat{s}-s\|^q] \leqslant C(q)\left[d^q(s, S) + \left(\left(\Phi \wedge \|s\|_\infty^{1/2}\right)(D/n)^{1/2}\right)^q + \left(\frac{\|s\|}{\sqrt{n}}\left(1 \vee \Phi\|s\|_\infty^{-1/2}\right)\right)^q + \left(\frac{\Phi\sqrt{D}}{n}\right)^q\right].$$

Note that although this result might look obvious, as it is for $q = 2$, it is not that easy to derive it from (2.1) when $q > 2$. We recall that we have already bounded the index Φ for various sieves of interest in the preceding section. In particular, it is important to notice that when $\|s\|_\infty < +\infty$ and $D \leqslant n$ the influence of Φ in the upper bound will not affect the rates of convergence provided that $\Phi \leqslant \sqrt{D}$. In particular, if μ is Lebesgue measure on [0, 1] and S is the linear space of polynomials of degree bounded by $D-1$ on [0, 1], one gets an upper bound for the risk of the following form:

$$\mathrm{E}_s[\|\hat{s}-s\|^q] \leqslant C(q, \|s\|_\infty)[d^q(s, S) + (D/n)^{q/2}].$$

Applications

1. If s belongs to the Sobolev space W_2^α, $\alpha \in \mathbb{N} - \{0\}$, on the one-dimensional torus and S is spanned by the first D elements of the Fourier basis, we know from (2.4) that $d(s, S)$ is of order $D^{-\alpha}$ so that choosing D of the order of $n^{1/(1+2\alpha)}$ and noticing that $\Phi \leqslant \sqrt{2}$, we obtain that

$$\mathrm{E}_s[\|\hat{s}-s\|^q] \leqslant C(q, s)n^{-q\alpha/(2\alpha+1)},$$

which generalizes the bound (2.5).

2. Now let us assume that s belongs to the space $\mathrm{Lip}(\alpha, L_2)$, with $0 < \alpha \leqslant 1$, of those functions t defined on $[0, 1]$ which satisfy $\sup_{z>0}(z^{-\alpha}\omega(t, z)_2) = |t|^{(\alpha)} < +\infty$. Here the modulus of continuity ω_2 is given by

$$[\omega(t, z)_2]^2 = \int_0^{1-z} |t(x+z) - t(x)|^2 \, \mathrm{d}x.$$

Additional details are given in DeVore and Lorentz (1993, p. 51). Let S be the space of polynomials on $[0, 1]$ with degree bounded by $D - 1 \geqslant 1$. It then follows from Theorem 6.3 of DeVore and Lorentz (1993, p. 220) that

$$d(s, S) \leqslant \kappa\omega(s, (D-1)^{-1})_2 \leqslant \kappa|s|^{(\alpha)}(D-1)^{-\alpha}.$$

Since $\Phi \leqslant \sqrt{D}$, one obtains, for the projection estimator $\hat{s}$,

$$\mathrm{E}_s[\|\hat{s} - s\|^q] \leqslant C'(q, s)[(D-1)^{-q\alpha} + (D/n))^{q/2}].$$

A choice of D of the order $n^{1/(1+2\alpha)}$ leads to a risk bound of order $n^{-q\alpha/(2\alpha+1)}$.

4.2. Maximum likelihood estimation

4.2.1. An upper bound for the risk

In this section we assume that we observe n independent and identically distributed random variables $Z_1, \ldots, Z_n$ with common distribution P which is absolutely continuous with respect to the *probability* measure μ and we want to estimate $\mathrm{d}P/\mathrm{d}\mu$. It was pointed out by Le Cam a long time ago (see, for instance, Le Cam 1973; 1975; or 1986) that the natural distance to use as a risk function in density estimation is Hellinger distance defined by $h^2(u, v) = \frac{1}{2}\int(\sqrt{u} - \sqrt{v})^2$. Unfortunately, if s is a density the L_2 distance $d(s, S)$ cannot be easily transformed into the Hellinger distance $h(s, S)$ unless likelihood ratios are uniformly bounded as shown in Lemma 4.1 of Birgé (1983). In order to avoid too restrictive assumptions on the family of densities to hand, it is better to try to approximate in L_2-norm the root $\sqrt{\mathrm{d}P/\mathrm{d}\mu}$ of the true density by the set S rather than approximating $\mathrm{d}P/\mathrm{d}\mu$ itself in L_2-norm. We therefore define the parameter s as $\sqrt{\mathrm{d}P/\mathrm{d}\mu}$.

Let S be a subset of $L_2(\mu)$ such that any element t in S is a non-negative function with $\|t\| = 1$. This means that S is a set of square roots of density functions with respect to μ. Computing the maximum likelihood estimator over the set of densities corresponding to S amounts to minimizing the empirical contrast $\gamma_n(z, t) = -n^{-1}\sum_{i=1}^n \log t(Z_i)$. Since it would be confusing to use the same notation for the square root of a density and the corresponding distribution, we shall denote by P_t the distribution with density t^2 with respect to μ. We recall that $K(P, Q) = \int \log(\mathrm{d}P/\mathrm{d}Q)\,\mathrm{d}P$ is the Kullback–Leibler information number between P and Q (with $K(P, Q) = +\infty$ when P is not absolutely continuous with respect to Q). The following result will be proved in Section 7.

Theorem 3. *Assume that μ is a probability measure and that S satisfies Covering Property $\mathrm{M}(n^{-1/2})$, with $1 \leqslant D \leqslant n$. Let $\hat{s}$ be an n^{-1}-maximum likelihood estimator on S, which means that*

$$\frac{1}{n}\sum_{i=1}^{n} \log[\hat{s}(Z_i)] \leqslant \frac{1}{n}\sum_{i=1}^{n} \log[t(Z_i)] + \frac{1}{n},$$

for any $t \in S$. Then one can find a numerical constant κ such that if $K(s, S) = \inf_{t\in S} K(s, t)$, then

$$\mathrm{E}_s[\|s - \hat{s}\|^2] \leqslant \kappa[K(s, S) + \mathscr{L} D/n], \qquad \textit{where } \mathscr{L} = 1 + \log[B'(1 + r')].$$

Moreover,

$$K(s, S) \leqslant 2[1 + \log(\|s/s^*\|_\infty)]\|s - s^*\|^2, \qquad \textit{for any } s^* \in S. \tag{4.1}$$

Remarks. Wong and Shen (1995) and Van de Geer (1995) have obtained results which are similar but somehow different from various points of view. First, they both use bracketing covering assumptions instead of Covering Property M($n^{-1/2}$), which slightly influences the evaluations of the rate of convergence of the maximum likelihood estimator on a sieve. We shall comment on this in Section 6.2.

We would also like to point out that neither Wong and Shen nor Van de Geer provide integrated risk bounds but only bounds in probability. There are further differences. The bias term in Van de Geer's bound is of order $\inf_{s^*\in S}(\|s - s^*\|\,\|s/s^*\|_\infty)^2$ instead of $K(s, S)$. It is clear from (4.1) that our evaluation of the bias is sharper. On the other hand, the probability bound in Theorem 3 of Wong and Shen (1995) does not tend to zero as n goes to infinity if we use a fixed finite-dimensional sieve such as the space of histograms with D pieces, assuming that the true parameter belongs to the sieve. Moreover, their probability bound involves quantities of the form $\inf_{s^*\in S} \int[\log(s/s^*)]^2 s^2/K(s, S)$ which can be very large.

Let us now see how one can use (4.1) to bound $K(s, S)$ in some particular cases.

Bracketed approximation. When S is defined as the subset of elements of norm 1 of some cone S^+ of non-negative functions in $L_2(\mu)$ and there exists an element s^+ in S^+ such that $s^+ \geqslant s$ and $\|s - s^+\| \leqslant \delta$, we define $s^* \in S$ by $(s^*)^2 = (s^+)^2/[\int (s^+)^2\,\mathrm{d}\mu]$. Then (7.7) shows that $\|s - s^*\| \leqslant \delta$ and that the ratio $(s/s^*)^2$ is bounded by 3 provided that $\delta \leqslant 1/\sqrt{2}$. This means that in this case we can bound $K(s, S)$ by $3\inf_{s^+\in S^+, s^+\geqslant s}\|s - s^+\|^2$. This situation will be illustrated below by the case of approximation by histograms.

Modification of the sieve. If μ is a probability measure, one can always modify the sieve S in the following way: change each t in S into $\bar{t}$, with $\bar{t}^2 = [t^2 + 1/(2n)]/[1 + 1/(2n)]$. Then, by (7.1), $\|t - \bar{t}\| \leqslant n^{-1/2}$. If we denote by $\overline{S}$ the corresponding modification of S we can check that $\overline{S}$ still satisfies Covering Property M($n^{-1/2}$). Moreover, if $\|s\|_\infty$ is finite, $\|s/\bar{t}\|_\infty^2 \leqslant (1 + 2n)\|s\|_\infty^2$ and, by (4.1),

$$K(s, \overline{S}) \leqslant 4(1 + \log[(1 + 2n)\|s\|_\infty])(d^2(s, S) + n^{-1}).$$

This shows that, up to a small modification of the sieve, one can always control $K(s, S)$ by $(\log n)(d^2(s, S) + n^{-1})$. Such a result could not be derived from Van de Geer (1995) since the

ratio $\|s/\bar{t}\|_\infty$ would appear as a multiplicative factor in her computations, introducing an extra factor n instead of our $\log n$.

4.2.2. Application to histograms

Regular histograms. Let s be an element of the Hölder class $\mathcal{H}(\alpha)([0, 1])$ (where $0 < \alpha \leqslant 1$) with seminorm

$$|s|_\alpha = \sup_{0 \leqslant x < y \leqslant 1} \frac{|s(x) - s(y)|}{(y - x)^\alpha} < +\infty.$$

Consider the sieve $\overline{\mathcal{H}}(D)$ of regular histograms introduced in Section 3.2.2. One can define an upper approximation s^+ of s in $\overline{\mathcal{H}}(D)$ by

$$s^+ = \sum_{j=1}^{D} \left[\sup_{(j-1)/D \leqslant x < j/D} s(x) \right] \mathbb{1}_{[(j-1)/D, j/D)}.$$

It follows that $\|s^+ - s\| \leqslant \|s^+ - s\|_\infty \leqslant |s|_\alpha D^{-\alpha}$. Let S be the subset of positive elements of norm 1 of $\overline{\mathcal{H}}(D)$ then it follows from the above remark about bracketed approximation that $K(s, S) \leqslant 3|s|_\alpha^2 D^{-2\alpha}$. It then follows from Proposition 1 that $\overline{\mathcal{H}}(D)$ and therefore S satisfy Covering Property M with $B' = 6$ and $r' = \bar{r} = 1$ as mentioned in Section 3.1.2. We can therefore apply Theorem 3 to derive that if $\hat{s}$ is the maximum likelihood estimator on S, then

$$\mathrm{E}_s[\|s - \hat{s}\|^2] \leqslant \kappa[|s|_\alpha^2 D^{-2\alpha} + D/n],$$

where κ is an absolute constant. Choosing D of order $n^{1/(2\alpha+1)}$, we obtain a bound on the risk of order $n^{-2\alpha/(2\alpha+1)}$ which is known to be the right rate of convergence for Hölderian densities.

Remark. In this case it is easy to compute $\hat{s}^2$ explicitly (it is the empirical histogram) and to bound $\mathrm{E}_s[\|s^2 - \hat{s}^2\|^2]$ by a direct computation. It is not that clear that one could easily bound $\mathrm{E}_s[\|s - \hat{s}\|^2]$ in the same way without any restriction on s.

Irregular histograms. Let us consider a function s with bounded variation $V(s)$ on $[0, 1]$ and the sieve S which is the subset of positive elements of norm 1 of $\mathcal{H}_n(D)$. As noticed in Section 3.2.3, S satisfies Covering Property M with $B' = 6\mathrm{e}n/D$ and $r' = (n/D)^{1/2}$. Moreover it follows from Corollary 1 of Barron *et al.* (1997) that one can find $s^+ \geqslant s$ in $\mathcal{H}_n(D)$ such that $\|s^+ - s\| \leqslant 5V(s)/D$ provided that $D \geqslant 3$. The above arguments therefore imply that, for a suitable absolute constant κ',

$$\mathrm{E}_s[\|s - \hat{s}\|^2] \leqslant \kappa'[V(s)^2 D^{-2} + (D \log n)/n].$$

Choosing D of order $(n/\log n)^{1/3}$, we obtain an upper bound for the risk of order $(n/\log n)^{-2/3}$.

Approximation properties of irregular histograms based on a grid. In order to apply Theorem 3 or Theorem 2 or any similar result to the sieve $\mathcal{H}_N(D)$ one is led to evaluate

quantities of the form $d(s, \mathscr{H}_N(D))$ but the usual results of approximation theory deal with histograms or, more generally, piecewise polynomials based on partitions with free end-points rather than end-points restricted to a grid of the type we use here. In order to compare the approximation properties of $\mathscr{H}_N(D)$ with those of histograms based on partitions with free end-points, the following elementary result might prove useful.

Lemma 4. *For any element f in the set*

$$\mathscr{H}(D, L) = \left\{ \sum_{j=1}^{D} a_j \mathbb{1}_{[b_{j-1}, b_j)},\ 0 = b_0 \leqslant b_1 \leqslant \cdots \leqslant b_D = 1,\ |a_j| < L,\ \textit{for } 1 \leqslant j \leqslant D \right\}$$

one can find an element g in $\mathscr{H}_N(D)$ with $\|f - g\|_2^2 \leqslant 2DL^2/N$.

Proof. Assume that $f = \sum_{j=1}^{D} a_j \mathbb{1}_{[b_{j-1}, b_j)}$ and define $g = \sum_{j=1}^{D} a_j \mathbb{1}_{[N_{j-1}/N, N_j/N)}$, where N_j is the integer closest to Nb_j. Since $|b_j - N_j/N| \leqslant (2N)^{-1}$, we obtain

$$\|f - g\|_2^2 \leqslant \sum_{j=1}^{D-1} (a_j - a_{j+1})^2/(2N) \leqslant 2DL^2/N. \qquad \square$$

Remark. It is clearly not necessary to assume that $|a_j| \leqslant L$, the condition $D^{-1}\sum_{j=1}^{D-1}(a_j - a_{j+1})^2 \leqslant 2L^2$ would lead to the same conclusion.

The meaning of this result is that when one replaces $d(s, \mathscr{H}(D, L))$ by $d(s, \mathscr{H}_n(D))$ one only loses a term of order $(D/n)^{1/2}$. Therefore since the bounds for the risk which appear in our theorems are of the form $C[d(s, \mathscr{H}_n(D)) + (D/n)^{1/2}]^q$, one could change $d(s, \mathscr{H}_n(D))$ into $d(s, \mathscr{H}(D, L))$ (which can be derived from classical results in approximation theory) without changing the order of magnitude of the bound but only the constant C.

4.2.3. Some lower bounds

Regular histograms. Let S be the subset of non-negative elements of norm 1 in $\overline{\mathscr{H}}(D)$ and assume that $s \in S$. It follows from Theorem 3 and our above evaluations of B' and r' ($B' = 6$ and $r' = 1$) that the maximal risk for $s \in S$ of the maximum likelihood estimator is bounded by $\kappa' D/n$. It follows from classical lower bounds methods (see, for instance, Assouad 1983; or Birgé 1986) that the minimax risk over S is bounded from below by $\kappa'' D/n$, which means that the maximum likelihood estimator is minimax on S, up to constants.

Irregular histograms. Let S be the subset of non-negative elements of norm 1 in $\mathscr{H}_{2n}(D)$ and $s \in S$. It follows from Theorem 3 and our above evaluations of B' and r' ($B' = 12\mathrm{e}n/D$ and $r' = (2n/D)^{1/2}$) that the maximal risk for $s \in S$ of the maximum likelihood estimator is bounded by $\kappa'(D/n)\log(1 + n/D)$. One might wonder whether the $\log(n/D)$ factor in the preceding bound is actually necessary or not. However, the presence of an extra $\log(n/D)$ factor is in some sense necessary when the sieve $\mathscr{H}_{2n}(D)$ is used, as shown by the following proposition to be proved in Section 7.

Proposition 2. *Assume that $D \geqslant 9$, $n \geqslant 5D$, $N \geqslant 1.4n$ and let $\mathscr{H}$ be the set of square roots of those densities in $\mathscr{H}_N(D)$ which are bounded by 2. For any estimator $\hat{s}_n$ based on n independent and identically distributed observations from some density s^2 with $s \in \mathscr{H}$, one has*

$$\sup_{s \in \mathscr{H}} \mathrm{P}_s \left[\|s - \hat{s}_n\|^2 > \kappa_1 \frac{D}{n} \log\left(\frac{n}{D}\right) \right] > \kappa_2, \tag{4.2}$$

where κ_1 and κ_2 are absolute constants.

Remark. The same result also holds if in (4.2) one replaces $\|s - \hat{s}_n\|$ by $\|s^2 - \hat{s}_n^2\|$, which is the L_2 distance between the densities themselves.

4.3. Regression framework

In this case we observe pairs $(X_i, Y_i) = Z_i$ with $Y_i = s(X_i) + W_i$ and the underlying variables (X_i, W_i) are independent with respective distributions $R_i \otimes Q_i$. The X_is are supposed to take values on a compact subset $\mathscr{X}$ of some Euclidean space and we denote by λ the Lebesgue probability measure on $\mathscr{X}$. We assume here that all the functions involved, i.e. s and the elements of S, belong to the Hilbert space $L_2(\mu)$, where μ is the average distribution of the X_is ($\mu = n^{-1}\sum_{i=1}^n R_i$) and that the norm $\|\cdot\|$ is the norm in $L_2(\mu)$. We also need the additional (and rather unpleasant) restriction that the sieve S is included in some L_∞ ball around s: there exists a constant H such that, for all $u \in S$, $\|s - u\|_\infty \leqslant H$. We introduce here two empirical contrast functions: $n^{-1}\sum_{i=1}^n [Y_i - t(X_i)]^2$, which corresponds to least-squares regression; and $n^{-1}\sum_{i=1}^n |Y_i - t(X_i)|$, which corresponds to minimum-L_1 regression. We denote by $\hat{s}_1$ and $\hat{s}_2$ the corresponding minimum contrast estimators on S, i.e. $\hat{s}_1$ is the least-squares estimator on S and $\hat{s}_2$ is the minimum-L_1 estimator on S. We can then prove the following theorem.

Theorem 4. *Assume that the sieve S satisfies Covering Property* $\mathrm{M}(n^{-1/2})$ *and that there exists a constant H such that, for all $u \in S$, $\|s - u\|_\infty \leqslant H$. Let $\mathscr{L} = 1 + \log B' + \log(1 + r'(D/n)^{1/2})$.*

(i) Assume that the errors W_i are centred random variables such that $\mathrm{E}_s[\exp \alpha |W_i|] \leqslant \Gamma$ for all i and suitable positive constants α and Γ. Then

$$\mathrm{E}_s[\|\hat{s}_1 - s\|^q] \leqslant C(H, q, \alpha, \Gamma)[d^q(s, S) + (\mathscr{L}D/n)^{q/2}]. \tag{4.3}$$

(ii) Assume that the errors W_i are independent and identically distributed with common distribution V with derivative v, that 0 is the medium of the distribution V and v is continuous and positive at 0. Then

$$\mathrm{E}_s[\|\hat{s}_2 - s\|^q] \leqslant C(H, q, v)[d^q(s, S) + (\mathscr{L}D/n)^{q/2}]. \tag{4.4}$$

Remarks.

1. The moment condition that we have used in case (i) (exponential moments for the W_is) is clearly too strong, at least in the case of a linear sieve, and could be

weakened, but at the price of a lot of additional technicalities that we do not wish to include here.

2. As in Nemirovskii *et al*. (1984) or Birgé and Massart (1993), one could introduce more general contrast functions and define

$$\gamma_n(t) = \frac{1}{n}\sum_{i=1}^{n} F[Y_i - t(X_i)],$$

where F is a function satisfying conditions (Ca), (Cc), (Cd), (Ce) of Birgé and Massart (1993, p. 125). Under such assumptions, one could derive from the general results of the next section an analogue of Theorem 4 for the corresponding minimum contrast estimator.

3. Checking Covering Property M($n^{-1/2}$) amounts to checking some entropy properties of the sieve S with respect to the metric induced by the norm in $L_2(\mu)$. The problem will then be quite different if this norm is equivalent to the usual $L_2(\lambda)$-norm or not. In the case of a random design, if μ and λ are mutually absolutely continuous with bounded densities, the two norms will be equivalent and classical approximation theory will generally do the job. For fixed design, μ is a discrete measure and it might be much more complicated to study the entropy properties of the classical spaces of approximation theory with respect to the norm in $L_2(\mu)$ (for related problems, see Van de Geer 1990).

We now consider some applications.

Neural nets. Assume that S is the neural net $S(D', \tau, G)$ described in Section 3.2.1. It follows that this sieve is a bounded subset of $L_\infty(\mu)$. If s itself belongs to $L_\infty(\mu)$ the boundedness property required in the above theorem is satisfied with a suitable constant H. It follows from Section 3.2.3 that Covering Property M($n^{-1/2}$) is also satisfied with $r' = 1$ and $B'(n^{-1/2}) = C_1 n^{1/2}$. Therefore, if the errors W_i do satisfy the properties required in part (i) of the theorem, one obtains

$$\mathrm{E}_s[\|\hat{s}_1 - s\|^q] \leqslant C[d^q(s, S) + (D' \log n/n)^{q/2}].$$

A similar bound holds for $\hat{s}_2$ if the errors satisfy the properties required in part (ii) of the theorem.

Trigonometric polynomials. Let us define $\overline{S}$ to be the linear span of the first D elements of the Fourier basis and assume that $\|s\|_\infty$ is bounded by a known constant H. Let S be the intersection of $\overline{S}$ and the L_∞ ball of radius $2H$ centred at zero and $\pi(s)$ be the orthogonal projection of s onto $\overline{S}$. If we assume that s belongs to the Sobolev space W_2^α of the one-dimensional torus for $\alpha \in \mathbb{N} - \{0\}$, which means that $\int[s^{(\alpha)}(x)]^2\,\mathrm{d}x < \infty$, then it follows from (7.12) and Theorem 2.3 of DeVore and Lorentz (1993, pp. 46, 205) that $\|s - \pi(s)\| \leqslant C_1\|s^{(\alpha)}\|D^{-\alpha}$. Using Theorem 3.4 of DeVore and Lorentz (1993, p. 181) we derive that

$$\inf_{t\in S}\|s-t\|_\infty \leqslant C_2\|s^{(\alpha)}\|D^{-\alpha+1/2}.$$

Using Lebesgue's lemma (DeVore and Lorentz 1993, p. 30), together with the fact that the norm of the operator π is bounded by the L_1-norm of the Dirichlet kernel which is itself not larger than $C_3 \log D$, we finally obtain

$$\|s-\pi(s)\|_\infty \leqslant C_4\|s^{(\alpha)}\|D^{-\alpha+1/2}\log D.$$

Therefore for large D, $\|s-\pi(s)\|_\infty \leqslant H$, $\pi(s)\in S$ and $d(s,S)\leqslant C_1\|s^{(\alpha)}\|D^{-\alpha}$. It follows from Sections 3.2.3 and 3.1.2 that in such a linear sieve case one can take B' as a numerical constant and r' as $(2D)^{1/2}$. Let us then choose $D=n^{1/(2\alpha+1)}$. Since $\alpha>1/2$, $\mathscr{L}$ is bounded and the estimator converges at a rate which is at least of order $n^{-2\alpha/(2\alpha+1)}$. This improves on Example 3 of Shen and Wong (1994, p. 593). They actually use weaker moment conditions on the W_is but do not obtain the right rates of convergence. The improvement we obtain is due to the introduction of the factor r' which relates L_2 and L_∞ approximations for nets.

5. A general approach leading to exponential inequalities

Let us recall that we observe n independent random variables Z_i, $1\leqslant i\leqslant n$, within the statistical framework described at the beginning of Section 4. The purpose here is to establish exponential bounds for the fluctuations of the centred empirical measure $\nu_n = \mathrm{P}_n - \mathrm{E}_s\circ \mathrm{P}_n$ acting on some class of functions $\gamma(\cdot,t)$, $t\in\mathscr{T}$, in order to analyse the behaviour of a minimum contrast estimator on a sieve $S\subset\mathscr{T}$, relative to some empirical contrast γ_n. A simple connection between γ and γ_n turns out to be

$$\nu_n[\gamma(\cdot,t)] = \gamma_n(t) - \mathrm{E}_s[\gamma_n(t)], \qquad \text{for any } t\in\mathscr{T} \text{ and } s\in\mathscr{S}. \tag{5.1}$$

Recalling that $\gamma_n(t) = n^{-1}\sum_{i=1}^n \overline{\gamma}(Z_i,t)$ this clearly occurs whenever $\gamma(\cdot,t)=\overline{\gamma}(\cdot,t)$ or $\gamma(\cdot,t)=\overline{\gamma}(\cdot,t)-\mathrm{E}_s[\gamma_n(t)]$. But, in order to deal with maximum likelihood estimation, it is useful to introduce more flexibility in the choice of γ.

We consider two sets of assumptions on the family $\{\gamma(\cdot,t)\}_{t\in S}$. The first says roughly that $\gamma(\cdot,t)$, as a function of t, behaves as a bounded Lipschitz function. The second tends to express that the metric structure of the sieve S is similar to the structure of the Euclidean space $\mathbb{R}^D$.

Assumption M1. *The observed random variables $Z_1,\ldots,Z_n$ can be written as $Z_i = f(s,X_i,W_i)$ for some function f and independent random variables $X_1,\ldots,X_n\in\mathscr{X}$ and $W_1,\ldots,W_n\in\mathscr{W}$. Moreover one can find positive numbers A, B, a_m, b_m and non-negative functions $M(w)$ and $\Delta(x,u,v)$ defined on $\mathscr{W}$ and $\mathscr{X}$ respectively such that, for each pair $(u,v)\in S^2$,*

$$|\gamma(z,u)-\gamma(z,v)|\leqslant M(w)\Delta(x,u,v)$$

and, for all $m\geqslant 2$,

$$\mathrm{E}_s[M^m(W_i)]\leqslant a_m A^m, \qquad \text{for all } i=1,\ldots,n, \tag{5.2}$$

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_s[\Delta^m(X_i, u, v)] \leqslant b_m \|u - v\|^2 B^{m-2}, \tag{5.3}$$

with either $a_m = 1$, $b_m = m!/2$, *for all* $m \geqslant 2$, *or* $b_m = 1$, $a_m = m!/2$, *for all* $m \geqslant 2$.

Remarks. It is useful to notice that:

(i) if, for any i, $M(W_i) = M$ is not random, then $A = M$ and $a_m = 1$;
(ii) we obtain (5.2) with $a_m = m!/2$ if we assume that, for all $i = 1, \ldots, n$,

$$\mathrm{E}_s[\exp(A^{-1}M(W_i))] \leqslant 3/2 + E_s[A^{-1}M(W_i)];$$

(iii) we obtain (5.3) with either $b_m = 1$ and $B = B_1$ or $b_m = m!/2$ and $B = B_1/3$ if

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_s[\Delta^2(X_i, u, v)] \leqslant \|u - v\|^2 \text{ and } \|\Delta(\cdot, u, v)\|_\infty \leqslant B_1, \qquad \text{for all } u, v \in S. \tag{5.4}$$

Assumption M2. *There exist two constants* B' *and* r *such that, for any* $\sigma \geqslant (D/n)^{1/2}$ *and* $0 < \delta < \sigma/5$, *one can find for any ball* $\mathscr{B} \subset S$ *of radius* σ *a finite* δ*-net* $T \subset \mathscr{B}$ *(which means that there exists a mapping* $\pi = \pi(\delta, \sigma)$ *from* $\mathscr{B}$ *to* T *such that* $d(u, \pi u) \leqslant \delta$, *for all* u *in* $\mathscr{B}$*) with*

$$|T| \leqslant (B'\sigma/\delta)^D, \qquad \text{for some } D \geqslant 1,\ B' \geqslant 1, \tag{5.5}$$

and

$$\sup_{u \in \pi^{-1}(t)} \|\Delta(\cdot, u, t)\|_\infty \leqslant r\delta, \qquad \text{for all } t \in T. \tag{5.6}$$

One can use Assumptions M1 and M2 to control the fluctuations of the centred empirical process $\nu_n[\gamma(\cdot, t)]$ in the following way.

Theorem 5. *Assume that* M1 *and* M2 *are satisfied and that we have fixed some positive number* τ. *Define* $\overline{\mathscr{S}}$, $\rho(\tau)$ *and* σ_D *by*

$$\overline{\mathscr{S}} = \frac{10\mathrm{e}}{2\mathrm{e} - 1}[\log B' + \log(4\sqrt{5} + 5) + \log(1 + r(D/n)^{1/2})], \tag{5.7}$$

$$\rho(\tau) = \frac{A}{\tau}\left[1 + \left(1 + \frac{B\tau}{2A}\right)^{1/2}\right], \qquad \sigma_D^2 = \frac{D}{n}\left[\left(\rho^2\left(\frac{3\tau}{4}\right)\overline{\mathscr{S}}\right) \vee 1\right].$$

Then one obtains, for any s^* *in* S,

$$\mathrm{P}_s\left[\sup_{u \in S} \frac{\nu_n[\gamma(\cdot, s^*) - \gamma(\cdot, u)]}{d^2(s^*, u) \vee \sigma^2} > \tau\right] < 3.03 \exp\left[-\frac{2n\sigma^2}{5\rho^2(3\tau/4)}\right], \qquad \text{for any } \sigma \geqslant \sigma_D. \tag{5.8}$$

The proof being rather technical, it will be deferred to Section 7.

We will now explain how Theorem 5 can be used to build risk bounds for minimum

contrast estimators. Let us recall that the minimum contrast estimator $\hat{s}$ which minimizes (up to n^{-1}) $\gamma_n(t) = n^{-1}\sum_{i=1}^n \overline{\gamma}(Z_i, t)$ over S satisfies

$$n^{-1}\sum_{i=1}^n \overline{\gamma}(Z_i, \hat{s}) \leqslant \sum_{i=1}^n n^{-1}\overline{\gamma}(Z_i, t) + n^{-1}, \qquad \text{for all } t \in S.$$

Setting $\ell(s, t) = \mathrm{E}_s[\gamma_n(t) - \gamma_n(s)]$, we see that

$$\ell(s, \hat{s}) \leqslant \ell(s, t) + \nu_n[\overline{\gamma}(\cdot, t) - \overline{\gamma}(\cdot, \hat{s})] + n^{-1}. \tag{5.9}$$

If either $\overline{\gamma}$ or its centred version satisfies Assumptions M1 and M2, then (5.1) holds with $\gamma = \overline{\gamma}$ (or the centred version) and we can combine (5.9) with (5.8) provided that one can find a suitable relation between ℓ and d. This motivates the introduction of the following assumption.

Assumption C′. *There exist two positive constants, k and k', such that*

$$kd^2(s, t) \leqslant \mathrm{E}_s[\gamma_n(t) - \gamma_n(s)] \leqslant k'd^2(s, t), \qquad \text{for any } t \in S \text{ and } s \in \mathscr{S}.$$

Unfortunately it can happen that neither the function $\overline{\gamma}$ nor its centred version satisfies Assumptions M1 and M2, which means that we have to consider more general functions γ. But still some analogue of (5.9) is needed with γ replacing $\overline{\gamma}$. Such a connection is forced by our next assumption.

Assumption C. *For any $s \in \mathscr{S}$ one can find some point $s^* \in S$ (depending on s), a non-negative random variable $U(s, s^*, Z_1, \ldots, Z_n)$ with finite second moment such that if t in S satisfies $\gamma_n(t) \leqslant \gamma_n(s^*) + n^{-1}$ then*

$$\nu_n[\gamma(\cdot, s^*) - \gamma(\cdot, t)] \geqslant kd^2(s, t) - U^2, \tag{5.10}$$

where k is some positive constant.

Remarks. One should think of s^* as minimizing the distance from s to S, i.e. $\|s - s^*\| \approx d(s, S)$ although formally s^* could be any point in S.

We can be precise as to the sense in which Assumption C extends C′. If γ is of the form $\gamma(z, t) = \overline{\gamma}(z, t) + \psi_1(t) + \psi_2(z)$, then $\nu_n[\gamma(\cdot, t) - \gamma(\cdot, u)] = \nu_n[\overline{\gamma}(\cdot, t) - \overline{\gamma}(\cdot, u)]$ for every t and u in S. This means that (5.9) holds with γ instead of $\overline{\gamma}$, and therefore if C′ holds then C is also satisfied for that choice of γ and any point $s^* \in S$ by setting $U^2 = k'd^2(s, s^*)$.

We now have all the necessary tools to compute the performance of our estimators. The results of Theorem 5, together with Assumption C, lead to various forms of control of the random term of the estimation error, while the bias term $d(s, s^*)$ is entirely taken care of by the approximation properties of the sieve S. In order to translate such results, which are given for fixed values of n and D, in terms of the usual 'rates of convergence', one must think of a sequence of such problems with choices of S and D depending on n in such a way that we achieve a balance between the bias and random terms. The resulting moment bounds can then be derived from the following corollary.

Corollary 1. *Assume that* M1, M2 *and* C *hold with* $\mathrm{E}[U^q] < \infty$, *for some integer* $q \geqslant 1$; *then*

$$\mathrm{E}_s[d^q(\hat{s}, s)] \leqslant C(q)[d^q(s, s^*) + (K(D\mathscr{L}/n)^{1/2})^q + k^{-q/2}\mathrm{E}_s[U^q]], \tag{5.11}$$

where

$$\mathscr{L} = \kappa[1 + \log[B'(1 + r(D/n)^{1/2})]], \qquad K = \left[\frac{16A}{3k}\left(1 + \left(1 + \frac{3Bk}{32A}\right)^{1/2}\right)\right] \vee 1.$$

When M1, M2 *and* C′ *hold, then*

$$\mathrm{E}_s[d^q(\hat{s}, s)] \leqslant C'(q, k, k', A, B)[d^q(s, s^*) + (D\mathscr{L}/n)^{q/2}].$$

Proof. Let us denote by $1 - p(\sigma)$ the probability that $\nu_n[\gamma(\cdot, s^*) - \gamma(\cdot, t)] \leqslant (k/4)(d^2(s^*, t) \vee \sigma^2)$ for all $t \in S$ and by $\hat{S}$ the subset of elements t in S such that $\gamma_n(t) \leqslant \gamma_n(s^*) + n^{-1}$. Then $\hat{s}$ belongs to $\hat{S}$ and it follows from Assumption C that with probability at least $1 - p(\sigma)$, for all t in $\hat{S}$,

$$(k/4)[2d^2(s, t) + 2d^2(s, s^*) + \sigma^2] \geqslant \nu_n[\gamma(\cdot, s^*) - \gamma(\cdot, t)] \geqslant kd^2(s, t) - U^2.$$

Then with probability at least $1 - p(\sigma)$,

$$d^2(s, t) \leqslant d^2(s, s^*) + (\sigma^2/2) + (2U^2/k)$$

and

$$d^q(s, t) \leqslant C_1(q)[d^q(s, s^*) + \sigma^q + (k^{-1}U^2)^{q/2}].$$

Let

$$V = \frac{d^q(s, t)}{C_1(q)} - d^q(s, s^*) - \left(\frac{U^2}{k}\right)^{q/2};$$

then $\mathrm{P}_s[V > \sigma^q] \leqslant p(\sigma)$ with $p(\sigma) \leqslant 1$, for $\sigma \leqslant \sigma_D$, and $p(\sigma)$ is bounded by the right-hand side of (5.8) with $\tau = k/4$ otherwise. It follows that

$$\mathrm{E}_s[V] = \int_0^\infty \mathrm{P}_s[V > t]\,\mathrm{d}t \leqslant \int_0^\infty p(t^{1/q})\,\mathrm{d}t = \int_0^\infty qx^{q-1}p(x)\,\mathrm{d}x.$$

Using the upper bounds for p we can conclude, since $\mathscr{L} \geqslant \overline{\mathscr{L}} \geqslant 1$ for a suitable constant κ, that

$$\mathrm{E}_s[V] \leqslant \sigma_D^q + C_2(q)\left[\frac{\rho(3k/16)}{\sqrt{n}}\right]^q \leqslant \left[\left(\rho^2\left(\frac{3k}{16}\right) \vee 1\right)\frac{\mathscr{L}D}{n}\right]^{q/2} + C_2(q)\left[\frac{\rho(3k/16)}{\sqrt{n}}\right]^q$$

and (5.11) follows since $K = \rho(3k/16) \vee 1$. When Assumption C′ holds, $U^2 = k'd^2(s, s^*) + 1/n$ and the bound follows. □

In the case of projection estimators on a linear sieve, the results of Theorem 5 can be improved thanks to the deep isoperimetric inequality due to Talagrand which is contained in

Theorem 1. The advantage of this approach is that it allows us to work with unbounded parameter sets when there exists a good connection between L_2- and L_∞-norms on S.

Proposition 3. *Assume that the observations $Z_1, \ldots, Z_n$ are independent and identically distributed with density s, that the sieve S is a subset of some D-dimensional linear subspace S_D of $L_2(\mu) \cap L_\infty(\mu)$ and that there exists a positive constant Φ such that $\|t\|_\infty \leqslant \Phi\sqrt{D}\|t\|$ for all $t \in S_D$. Let τ be some positive number and*

$$\sigma_D = (3/\tau)(\Phi \wedge \|s\|_\infty^{1/2})(D/n)^{1/2}.$$

There exists a universal constant κ such that, for any $\sigma \geqslant \sigma_D$, we have

$$\mathrm{P}_s\left[\sup_{u \in S} \frac{|\nu_n(u)|}{\|u\|^2 \vee \sigma^2} > \tau\right] \leqslant 3\exp\left[-\kappa n\left(\frac{\sigma^2\tau^2}{(\Phi\sqrt{D}\|s\|) \wedge \|s\|_\infty} \wedge \frac{\sigma\tau}{\Phi\sqrt{D}}\right)\right]. \tag{5.12}$$

The proof relies on Lemma 1 and on the following consequence of Theorem 1.

Corollary 2. *Let $Z_1, \ldots, Z_n$ be n independent and identically distributed variables and $\mathscr{F}$ a countable family of functions that are uniformly bounded by some constant b. Let $v = \sup_{f \in \mathscr{F}} \mathrm{E}[f^2(Z_1)]$. There exists a universal constant κ' such that, for any positive η and λ,*

$$\mathrm{P}\left[\sup_{f \in \mathscr{F}} |\nu_n(f)| \geqslant (1+\eta)E + \lambda\right] \leqslant 3\exp\left[-n\kappa'\left(\frac{\lambda^2}{v} \wedge \frac{(\eta \wedge 1)\lambda}{b}\right)\right] \tag{5.13}$$

if $E \geqslant \mathrm{E}[\sup_{f \in \mathscr{F}} |\nu_n(f)|]$.

Proof. Starting from (2.9) of Theorem 1 applied to $\xi = \lambda + \eta E$, we obtain from the inequality $\Sigma^2 \leqslant v + 8bE$ given in Ledoux (1996, p. 78) that

$$\frac{\xi^2}{\Sigma^2 + b\xi} \geqslant \frac{\lambda^2 + 2\eta\lambda E}{v + 8bE + b\lambda + b\eta E} \geqslant \frac{1}{3}\left[\frac{\lambda^2}{v} \wedge \frac{2\lambda\eta}{b(8+\eta)} \wedge \frac{\lambda^2}{b\lambda}\right] \geqslant \frac{1}{3}\left[\frac{\lambda^2}{v} \wedge \frac{2(\eta \wedge 1)\lambda}{9b}\right],$$

since $\eta/(8+\eta) \geqslant (\eta \wedge 1)/9$. Bound (5.13) follows immediately. □

Proof of Proposition 3. Let $\varphi_1, \ldots, \varphi_D$ be an orthonormal basis of S_D. We want to apply Corollary 2 to the family $\mathscr{F} = \{f_u | u \in S\}$, where $f_u = u/(\|u\| \vee \sigma)^2$. In principle, Corollary 2 only applies to countable families, but in our case u belongs to a finite-dimensional space which implies that the function $u \mapsto \nu_n(f_u)$ is continuous and therefore the value of $\sup_{f \in \mathscr{F}} |\nu_n(f)|$ will not change if we restrict ourselves to a countable and dense subset of S. Since $|f_u| \leqslant |u|/(\sigma\|u\|)$ the Cauchy–Schwarz inequality leads to

$$\left[\sum_{i=1}^n (f_u(Z_i) - \mathrm{E}_s[f_u(Z_i)])\right]^2 \leqslant \left(\frac{\|u\|}{\sigma\|u\|}\right)^2 \sum_{j=1}^D \left[\sum_{i=1}^n (\varphi_j(Z_i) - E_s[\varphi_j(Z_i)])\right]^2.$$

The variables $\varphi_j(Z_i) - \mathrm{E}_s[\varphi_j(Z_i)]$ being independent and centred, we obtain by Lemma 1

$$\mathrm{E}_s[\sup_{u\in S}|\nu_n(f_u)|^2] \leqslant \frac{1}{n\sigma^2}\sum_{j=1}^{D}\mathrm{E}_s[\varphi_j^2(Z_1)] \leqslant \frac{(\Phi^2 \wedge \|s\|_\infty)D}{n\sigma^2}$$

and by Jensen's inequality $\mathrm{E}_s[\sup_{f\in\mathscr{F}}|\nu_n(f)|] \leqslant \mathrm{E} = [(\Phi \wedge \|s\|_\infty^{1/2})/\sigma](D/n)^{1/2}$. Therefore $2E + \tau/3 \leqslant \tau$ provided that $\sigma \geqslant \sigma_D$ and we can bound the left-hand side of (5.12) by means of Corollary 2 with $\eta = 1$ and $\lambda = \tau/3$. The value of b derives from the fact that $\|f_u\|_\infty \leqslant \Phi\sqrt{D}/\sigma$. Finally,

$$v \leqslant \sup_{u\in S}\frac{\mathrm{E}_s[u^2]}{\sigma^2\|u\|^2} \leqslant \sup_{u\in S}\frac{\|u\|_\infty \mathrm{E}_s[|u|]}{\sigma^2\|u\|^2} \leqslant \frac{1}{\sigma^2}\|s\|\Phi\sqrt{D},$$

which can be improved to $v \leqslant \sigma^{-2}\|s\|_\infty$ when $\|s\|_\infty < +\infty$. The result follows. □

6. Conclusion

6.1. Why finite-dimensional sieves?

The basic idea of the method of sieves is to replace a complicated function space $\mathscr{S}$, to which the true parameter is supposed to belong, by a simpler space S_n (depending on the number n of observations) which is supposed to have good approximation properties with respect to $\mathscr{S}$. The choice of S_n is typically determined by two main constraints:

- There should be an optimal balance between the distance from S_n to $\mathscr{S}$ (bias) and the risk of the minimum contrast estimator on S_n. If this balance is suitably achieved one can hope that the estimator will achieve approximately the optimal rate of convergence on $\mathscr{S}$. More details about the connection between approximation theory and estimation rates are given in Birgé (1983).
- In order to obtain an effective estimation procedure the space S_n should be defined in such a way that the optimization procedure of the empirical contrast on S_n could be performed in practice.

In Birgé and Massart (1993) we studied the particular case $S_n = \mathscr{S}$. The rates of convergence of the resulting global minimum contrast estimators are computed there and expressed in terms of the metric entropy with bracketing of $\mathscr{S}$. When this entropy is too large these rates can be suboptimal. Moreover, apart from some particular cases (maximum likelihood on the space of monotone functions, for instance), the required global optimization procedures are in general unrealistic from the practical point of view. Replacing $\mathscr{S}$ by an approximating net or an infinite-dimensional sieve might look attractive from a theoretical point of view since it apparently provides more flexibility than considering only finite-dimensional sieves. Unfortunately, it leads to similar problems of optimization and therefore will not add much, from a practical point of view, to what is already known from Le Cam (1975) or Birgé (1983) who, by the way, also proposed impractical estimators but which typically achieve the optimal rates of convergence. On the other hand, a close look at classical references in approximation theory such as Birman and

Solomjak (1967), DeVore and Lorentz (1993) or DeVore *et al*. (1992) shows that many of the classical spaces used in approximation theory are finite-dimensional linear spaces or finite unions of such spaces. A typical example is provided by various collections of piecewise polynomials based on dyadic partitions. As a matter of fact it is often very difficult to exhibit explicit nets even if it is possible to bound their cardinality – see, for instance, the delicate constructions of Birman and Solomjak (1967). Moreover, when one is able to build such nets, these nets are, in all the cases we know, defined as unions of nets in finite-dimensional approximating spaces.

In fact classical linear approximation methods are based on a single linear approximating space of a given dimension (piecewise polynomials based on a regular partition, trigonometric polynomials, etc.) while more recent nonlinear approximation methods such as thresholding of wavelet-type expansions are based on approximation by unions of finite-dimensional linear spaces of bounded dimension.

6.2. L_2 brackets versus simultaneous L_2 and L_∞ coverings

The notion of L_2 entropy with bracketing that we borrowed from the theory of empirical processes to study global minimum contrast estimators in Birgé and Massart (1993) is especially well suited for infinite-dimensional spaces and more specifically for the space of monotone functions for which L_∞ entropy does not exist. Of course the same notion with the same method of proof can be used to extend these results to minimum contrast estimation on sieves, as is done in Shen and Wong (1994) and Van de Geer (1995).

The point here is that we essentially want to focus on sieves which are finite-dimensional linear spaces (or unions of such spaces). It turns out that since all sensible finite-dimensional approximation spaces are embedded in L_∞ one can always compute the L_∞ entropy of bounded sets of such spaces. The key point of the discussion is the comparison between L_2 entropy, L_2 entropy with bracketing and L_∞ entropy. When L_2 and L_∞ entropy are of the same order (with respect to the dimension D of the space), methods involving L_2 entropy with bracketing or L_∞ entropy are equivalent. Covering Property M, introduced in this paper, is a new covering notion. It is especially relevant when L_2 and L_∞ entropies are not of the same order. We now provide some simple examples showing that L_2 entropy with bracketing and Covering Property M are not directly comparable although one can derive a crude upper bound for L_2 entropy with bracketing from Covering Property M. To make the comparison easier we introduce the analogue of the Covering Property M for bracketing.

Covering Property B. *We shall say that a subset S of $L_2(\mu)$ satisfies Covering Property* B *if there exist numbers $D \geqslant 1$ and $B'' \geqslant 1$ such that, for any $\delta > 0$, $\sigma \geqslant 5\delta$ and any ball $\mathscr{B}$ of radius σ with respect to L_2, we can find a finite set I of indices and pairs of functions $t_i^- \leqslant t_i^+ \in L_2(\mu)$ with $\|t_i^+ - t_i^-\| \leqslant \delta$, for each $i \in I$ and such that for any element $t \in \mathscr{B}$ one can find $i \in I$ with $t_i^- \leqslant t \leqslant t_i^+$. Moreover, the cardinality of I is bounded by $[B''\sigma/\delta]^D$.*

6.2.1. Examples

***Histograms based on a partition of* [0, 1].** The following result is to be proved in Section 7.

Proposition 4. *Let $\mathscr{P}$ be a partition of* [0, 1] *by D intervals and $H(\mathscr{P})$ the D-dimensional linear space of piecewise constant functions on $\mathscr{P}$. Then*:

(i) $H(\mathscr{P})$ satisfies Covering Property M *with $r' = 1$ and*

$$B' = 5 \exp\left[-(2D)^{-1} \sum_{I \in \mathscr{P}} \log(D|I|)\right];$$

(ii) for any δ, σ, $0 < \delta < \sigma/5$, the number of brackets with L_2 diameter smaller than δ needed to cover an L_2 ball of radius σ is bounded by $(5\sigma/\delta)^D$.

When we consider the regular partition with D elements we get $B' = 5$. When we consider irregular partitions, B' may become arbitrarily large while the L_2 entropy with bracketing is not affected. This means that L_2 entropy with bracketing is preferable to Covering Property M for very irregular histograms, but we shall see later what are the consequences for rates of estimation.

Trigonometric polynomials. If S is the linear space generated by the first D elements of the Fourier basis, we have already seen that Covering Property M is satisfied with $B' = 6$ and $r' = \sqrt{2D}$. As far as we know, the only way of computing the L_2 entropy with bracketing in this case is to bound it crudely by the L_∞ entropy. This implies that the number of brackets with L_2 diameter smaller than δ needed to cover an L_2 ball of radius σ is bounded by $\kappa(\sqrt{2D}\sigma/\delta)^D$, which is much larger than the number given by Covering Property M.

6.2.2. Logarithmic factors in the risk

We will now discuss the presence of some logarithmic factors that appear in the upper bounds for the risk of minimum contrast estimators. Our concern is to discuss the effect of the assumptions (Covering Property B or Covering Property M) on the existence of these logarithmic factors and also to understand to what extent they are necessary or not.

Let us first observe that, whatever the method to be used, the quadratic risk (say, to be specific) can be split into two terms: a bias term which is proportional to the square of the distance between the sieve and the parameter; and a random term which is connected to the number of observations and the metric properties of the sieve. For a good choice of sieve the two terms are balanced, which means that the bias term should not be larger than the random term. On the other hand, the random term cannot be smaller than it is when the parameter belongs to the sieve (no bias term). Since the assumptions used only influence the random term, the main point is to discuss the effect of those assumptions on the risk when the parameter belongs to the sieve. We shall hereafter distinguish between two different situations: maximum likelihood estimation and regression.

In the former situation, following Theorem 3, we see (assuming that s belongs to the sieve) that the quadratic risk is bounded by $\kappa[1 + \log[B'(1 + r')]](D/n)$. On the other hand,

using Covering Property B would lead to a bound of order $(1+\log B'')D/n$. Let us concentrate here on the particular example of histograms. Assume first that S is the linear sieve $H(\mathscr{P})$ of histograms generated by a given partition $\mathscr{P}$. Proposition 4 implies that the risk is always under better control if we use Covering Property B rather than Covering Property M. Let us discuss how much better it can be. If the partition $\mathscr{P}$ is regular or close enough to regular, r' and B' are bounded (by Proposition 4) and therefore both assumptions lead to equivalent results. Otherwise if we set $\Delta = \inf_{I\in\mathscr{P}} |I|$, then $\log B' \leqslant -\frac{1}{2}\log[\Delta D] + \log 5$. Typically Δ will be of order $1/n$ in the worse case and we will not lose more than a $\log(n/D)$ factor.

If we use as S the space $\mathscr{H}_N(D)$ of all histograms with D pieces with end-points on a regular grid of mesh $1/N$ introduced in Section 3.2.2, we have seen that one can take $r' = \sqrt{N/D}$ and $B' = 6eN/D$. This leads to a bound on the risk of order $[\log(N/D)+1](D/n)$ which is optimal in the minimax sense if, for instance, $N = 2n$ (see Proposition 2). This means that in this case one does not improve the bound by introducing Covering Property B.

In the case of *least-squares regression*, Theorem 4 implies that the quadratic risk is bounded by

$$C[1+\log B' + \log(1+r'(D/n)^{1/2})](D/n)$$

while Covering Property B would lead to a bound of order $(1+\log B'')D/n$ as before. The two bounds are not directly comparable in general but one can always bound B'' by $B'r'$. It may happen that the best bound for B'' that one is able to derive is actually of order $B'r'$. In such a case Covering Property M leads to a better bound. A good example of such a situation is provided by the trigonometric polynomials for which $B' = 6$, $r' = \sqrt{2D}$ and $B'' = \kappa\sqrt{2D}$. Then Covering Property B leads to an extra $\log D$ factor in the risk which does not appear when one uses Covering Property M and $D \leqslant \sqrt{n}$. If D is much larger than $\sqrt{n}$ then both bounds become equivalent and we do not know in that case whether the extra $\log D$ factor is actually necessary or not.

From the study of a few typical examples we see that it is not at all clear which type of assumptions – L_2 entropy with bracketing as in Covering Property B or simultaneous L_2 and L_∞ coverings as in Covering Property M – is to be preferred even if the latter property leads to simpler proofs for deriving maximal inequalities. In particular, it avoids the use of adaptive truncation in the chaining arguments. There is no systematic superiority of one property over the other and the choice of appropriate assumptions clearly depends on the particular situation to hand. In any case, for all interesting examples we know about, the differences only involve $\log n$ factors. One could even imagine that other covering properties could lead to similar maximal inequalities. One such example (involving L_1 entropy with bracketing) is given in Barron *et al.* (1997).

6.3. The importance of Talagrand's inequality and exponential bounds

In Section 4.1 we have given a separate treatment of the case of projection estimators on linear sieves with a proof which is based on Talagrand's inequality rather than on covering

assumptions as our Theorem 5. The reason is that Theorem 5, for the contrast of projection, could be applied only for uniformly bounded sieves which definitely exclude linear sieves. Of course, as explained in Section 2, the proof of Theorem 2 is straightforward when $q = 2$ and the interest in using Talagrand's inequality is actually twofold: first, it allows us to deal with $q > 2$; second, it leads to exponential inequalities which are essential for penalization methods as shown in Section 2. More generally, the exponential bounds that we have derived in the present paper are crucial tools for both defining and studying penalized minimum contrast estimators which we use in Barron *et al*. (1997) for model selection. We emphasize the fact that the structure of those exponential bounds directly influences the size of the penalty terms involved in those methods. In particular, it is of a special interest to use Talagrand's inequality in the case of projection estimators since it leads to explicit values for the penalty terms. An illustration of this idea was given in Section 2.

7. Proofs

7.1. Proof of (2.10) and (2.11)

The proof of (2.10) is based on an application of Corollary 2 with $\mathscr{F} = \{f_t | t \in S_{D'}\}$, where $f_t = (t - s_D)/\|t - s_D\|$. As we have already mentioned in the proof of Proposition 3, the fact that $\mathscr{F}$ is not countable does not create any problem since, by continuity, one could restrict the definition to those t belonging to a dense countable subset of $S_{D'}$. We obtain by the Cauchy–Schwarz inequality

$$\mathrm{E}_s(f_t^2) = \frac{\int (t - s_D)^2 s}{\|t - s_D\|^2} \leqslant \frac{\|t - s_D\|_\infty \|t - s_D\| \, \|s\|}{\|t - s_D\|^2}.$$

Let $D'' = D \vee D'$. Then $t - s_D$ can be written as $\sum_{\lambda \in \Lambda_{D''}} a_\lambda \varphi_\lambda$ and it follows from the Cauchy–Schwarz inequality and (2.2) that

$$\|t - s_D\|_\infty = \left\| \sum_{\lambda \in \Lambda_{D''}} a_\lambda \varphi_\lambda \right\|_\infty \leqslant \|t - s_D\| \left\| \sum_{\lambda \in \Lambda_{D''}} \varphi_\lambda^2 \right\|_\infty^{1/2} \leqslant \|t - s_D\| \sqrt{D''}.$$

It follows that $\mathrm{E}_s(f_t^2) \leqslant \|s\| \sqrt{D''}$, which implies that one can choose $v = \|s\| \sqrt{D''}$. The same computation shows that one can take $b = \sqrt{D''}$ in Corollary 2. Let us now look for an upper bound on E.

$$\sup_{t \in S_{D'}} |\nu_n(f_t)| \leqslant \sup_{t \in S_{D''}} |\nu_n(f_t)| = \sup_{a \in \mathbb{R}^{D''}} \frac{|\sum_{\lambda \in \Lambda_{D''}} a_\lambda \nu_n(\varphi_\lambda)|}{|a|} = \left[\sum_{\lambda \in \Lambda_{D''}} \nu_n^2(\varphi_\lambda) \right]^{1/2}.$$

One then obtains by Jensen's inequality

$$\mathrm{E}_s\left[\sup_{t\in S_{D'}}|\nu_n(f_t)|\right] \leqslant \left[E_s\left[\sum_{\lambda\in\Lambda_{D''}}\nu_n^2(\varphi_\lambda)\right]\right]^{1/2} = \frac{1}{\sqrt{n}}\left[\sum_{\lambda\in\Lambda_{D''}}\mathrm{Var}_s(\varphi_j)\right]^{1/2}$$

$$\leqslant \frac{1}{\sqrt{n}}\left[\int\left(\sum_{\lambda\in\Lambda_{D''}}\varphi_j^2\right)s\right]^{1/2} \leqslant \frac{1}{\sqrt{n}}\left\|\sum_{\lambda\in\Lambda_{D''}}\varphi_j^2\right\|_\infty^{1/2} = \left(\frac{D''}{n}\right)^{1/2},$$

and we can choose $E=(D''/n)^{1/2}$. Let us now take $\lambda=(1/\sqrt{n})(\eta\sqrt{D''}+x)$ in (5.13). It follows that

$$\frac{\lambda^2}{v} \geqslant \frac{1}{nv}[\eta^2 D'' + 2x\eta\sqrt{D''}] = \frac{1}{n\|s\|}[\eta^2\sqrt{D''}+2x\eta] \geqslant \frac{1}{n\|s\|}[\eta^2\sqrt{D'}+2x\eta]$$

and

$$\frac{\lambda}{b} = \frac{1}{\sqrt{n}b}[\eta\sqrt{D''}+x] = \frac{1}{\sqrt{n}}\left[\eta+\frac{x}{\sqrt{D''}}\right] \geqslant \frac{1}{n}[\eta\sqrt{D'}+x],$$

since $D'\leqslant D''\leqslant n$. Then

$$n\kappa\left(\frac{\lambda^2}{v}\wedge\frac{(n\wedge 1)\lambda}{b}\right) \geqslant \kappa\frac{\eta\wedge 1}{1+\|s\|}(\eta\sqrt{D'}+x).$$

Finally,

$$\sum_{D'=1}^{n}\mathrm{P}_s\left[\sup_{t\in S_{D'}}|\nu_n(f_t)| \geqslant (1+2\eta)\left(\frac{D\vee D'}{n}\right)^{1/2}+\frac{x}{\sqrt{n}}\right]$$

$$\leqslant \sum_{D'=1}^{n}\exp\left[-\kappa\frac{\eta\wedge 1}{1+\|s\|}(\eta\sqrt{D'}+x)\right] \leqslant C(\eta,\|s\|)\exp\left[-\kappa\frac{\eta\wedge 1}{1+\|s\|}x\right].$$

Now, since the bound in (2.10) is valid for all D' simultaneously and all $t\in S_{D'}$ we can apply it with $D'=\hat{D}$ and $t=\tilde{s}$ which shows that, apart from a set of probability bounded by the right-hand side of (2.10), we have

$$2\nu_n(\tilde{s}-s_D) \leqslant 2[\|\tilde{s}-s\|+\|s-s_D\|]\left[(2\eta+1)\left(\frac{D\vee\hat{D}}{n}\right)^{1/2}+\frac{x}{\sqrt{n}}\right]$$

$$\leqslant \frac{2(2\eta+1)^2}{3}\left[(1+\eta)\|\tilde{s}-s\|^2+\left(1+\frac{1}{\eta}\right)\|s-s_D\|^2\right]$$

$$+\frac{3}{2(2\eta+1)^2}\left[\frac{4}{3}(2\eta+1)^2\frac{D\vee\hat{D}}{n}+4\frac{x^2}{n}\right].$$

Combining this with (2.8) for η small enough, we finally obtain (2.11). □

7.2. Proof of Theorem 2

Let s^* be the orthogonal projection of s onto S; then $d(s, S) = \|s - s^*\|$. As already shown in (2.6), $\|\hat{s} - s\|^2 \leqslant \|s - s^*\|^2 + 2\nu_n(\hat{s} - s^*)$. Using Proposition 3 we see that, up to a probability bounded by the right-hand side of (5.12), one obtains, for $\sigma \geqslant \sigma_D$, using Pythagoras's inequality

$$\|\hat{s} - s\|^2 \leqslant \|s - s^*\|^2 + 2\tau(\|\hat{s} - s^*\| \vee \sigma)^2 \leqslant \|s - s^*\|^2 + 2\tau\|\hat{s} - s\|^2 + 2\tau\sigma^2.$$

Choosing $\tau = 1/4$ we then get $\|\hat{s} - s\|^2 \leqslant 2\|s - s^*\|^2 + \sigma^2$. Denote by $p(\sigma)$ the right-hand side of (5.12) when $\sigma \geqslant \sigma_D$ and set $p(\sigma) = 1$ otherwise; then

$$\mathrm{P}_s[\|\hat{s} - s\|^q > C_1(q)(\|s - s^*\|^q + \sigma^q)] \leqslant p(\sigma),$$

from which one derives that

$$\mathrm{E}_s[\|\hat{s} - s\|^q] \leqslant C_1(q)\left(\|s - s^*\|^q + \int_0^\infty q x^{q-1} p(x)\,\mathrm{d}x\right).$$

It remains for us to bound the integral, which can be done in the following way since $\sigma_D = 12(\Phi \wedge \|s\|_\infty^{1/2})\sqrt{D/n}$:

$$\int_0^\infty x^{q-1} p(x)\,\mathrm{d}x \leqslant \sigma_D^q + 3\left[\int_{\sigma_D}^\infty x^{q-1} \exp\left(-\frac{\kappa_2 n x^2}{\Phi\sqrt{D}\|s\|}\right)\mathrm{d}x + \int_0^\infty x^{q-1}\exp\left(-\frac{\kappa_3 n x}{\Phi\sqrt{D}}\right)\mathrm{d}x\right]$$

$$\leqslant \left[12(\Phi \wedge \|s\|_\infty^{1/2})\left(\frac{D}{n}\right)^{1/2}\right]^q + 3n^{-q/2}\|s\|^q\left(1 \vee \frac{\Phi}{\|s\|_\infty^{1/2}}\right)^q \int_{12a}^\infty a^q y^{q-1}\,\mathrm{e}^{-\kappa_2 y^2}\,\mathrm{d}y$$

$$+ C_2(q)\left[\frac{\Phi\sqrt{D}}{n}\right]^q, \qquad \text{where } a^2 = \frac{(\Phi^2 \wedge \|s\|_\infty)\sqrt{D}}{\Phi\|s\|},$$

and the result follows since the last integral is bounded with respect to a. □

7.3. Some inequalities relating Hellinger distance and Kullback–Leibler information

Recalling that the Hellinger distance $h(P, Q)$ between two positive measures P and Q is defined by $h(P, Q) = \frac{1}{2}\int(\sqrt{dP} - \sqrt{dQ})^2$, we shall summarize a number of useful results and inequalities involving Hellinger distance and Kullback–Leibler information numbers in the following lemma.

Lemma 5. *Let P, Q, R be three probability measures and $\lambda \in [0, 1]$. Then the following inequalities are valid:*

$$h^2(P, \lambda Q + (1 - \lambda)R) \leqslant \lambda h^2(P, Q) + (1 - \lambda)h^2(P, R); \tag{7.1}$$

$$h\left(P, \frac{P+Q}{2}\right) > 0.29\, h(P, Q); \tag{7.2}$$

$$h^2(\lambda P + (1-\lambda)R, \lambda Q + (1-\lambda)R) \leqslant \lambda h^2(P, Q); \tag{7.3}$$

$$K(P, \lambda Q + (1-\lambda)R) \leqslant \lambda K(P, Q) + (1-\lambda)K(P, R); \tag{7.4}$$

$$2h^2(P, Q) \leqslant K(P, Q). \tag{7.5}$$

Finally, if $\|\mathrm{d}P/\mathrm{d}Q\|_\infty < +\infty$, *then*

$$\frac{K(P, Q)}{h^2(P, Q)} \leqslant \mathscr{K}\left(\left\|\frac{\mathrm{d}P}{\mathrm{d}Q}\right\|_\infty\right), \qquad \text{with } \mathscr{K}(x) = \frac{x(\log x - 1) + 1}{(x+1)/2 - \sqrt{x}} \leqslant 4 + 2\log x. \tag{7.6}$$

Moreover, if P^+ *is a finite positive measure which dominates* P *and such that* $\mathrm{d}P/\mathrm{d}P^+ \leqslant 1$, $\int \mathrm{d}P^+ = 1 + \alpha$ *and* $h(P, P^+) = h$, *defining the probability* Q *by* $\mathrm{d}Q = \mathrm{d}P^+/(1+\alpha)$ *we obtain*

$$h^2(P, Q) = \frac{h^2 - v(\alpha)}{(1+\alpha)^{1/2}} \quad \text{and } \alpha \leqslant v^{-1}(h^2), \qquad \text{where } v(x) = 1 + \frac{x}{2} - (1+x)^{1/2}. \tag{7.7}$$

Proof. Inequality (7.1) derives from convexity as well as (7.4), which is classical, while (7.2) is proved in Lemma 1 of Birgé and Massart (1993). Expressions (7.5) and (7.6) derive from the proof of Lemma 4.4 in Birgé (1983) and elementary calculus. As to (7.3), it is proved as follows using the fact that $\mathrm{d}P + \mathrm{d}Q \geqslant 2\sqrt{\mathrm{d}P\,\mathrm{d}Q}$:

$$\begin{aligned}\int [(\lambda\,\mathrm{d}P + (1-\lambda)\,\mathrm{d}R)(\lambda\,\mathrm{d}Q + (1-\lambda)\,\mathrm{d}R)]^{1/2} &\geqslant \int [\lambda^2\,\mathrm{d}P\,\mathrm{d}Q + 2\lambda(1-\lambda)\,\mathrm{d}R\sqrt{\mathrm{d}P\,\mathrm{d}Q} \\ &\quad + (1-\lambda)^2\,\mathrm{d}R^2]^{1/2} \\ &= \lambda \int \sqrt{\mathrm{d}P\,\mathrm{d}Q} + (1-\lambda).\end{aligned}$$

Finally, (7.7) follows from Lemma 10 of Wong and Shen (1992) and Jensen's inequality. □

7.4. Proof of Theorem 3

The proof derives from two auxiliary results, the first of which is elementary.

Lemma 6. *For all* $a \geqslant 0$ *and* $b \geqslant b_0 > 0$, *one has*

$$\log \frac{(a+\delta)^2 + b^2}{a^2 + b^2} \leqslant 2\log\left(1 + \frac{\delta}{b_0}\right) \leqslant \frac{2\delta}{b_0}. \tag{7.8}$$

Proof. The left-hand side of (7.8) is decreasing with respect to b and its maximum with

respect to a is obtained for $a = [(\delta^2 + 4b^2)^{1/2} - \delta]/2$ which leads to the maximum value

$$\log \frac{((\delta^2 + 4b_0^2)^{1/2} + \delta)^2 + 4b_0^2}{((\delta^2 + 4b_0^2)^{1/2} - \delta)^2 + 4b_0^2} = \log \frac{(\delta^2 + 4b_0^2)^{1/2} + \delta}{(\delta^2 + 4b_0^2)^{1/2} - \delta} = 2 \log \frac{(\delta^2 + 4b_0^2)^{1/2} + \delta}{2b_0}$$

$$\leqslant 2 \log \frac{2\delta + 2b_0}{2b_0}. \qquad \square$$

The next result is a two-sided version of the inequality given in Lemma 3.4 of Van de Geer (1995) which can be proved in the same way.

Lemma 7. *Let f, $\tilde{f}$, g_1, g_2 be densities with respect to some measure μ and $P_1 = g_1.\mu$, $P_2 = g_2.\mu$; then*

$$\mathrm{E}_f \left[\left| \frac{1}{2} \log \frac{\tilde{f} + g_1}{\tilde{f} + g_2} \right|^m \right] \leqslant \frac{m!}{2} h^2(P_1, P_2) \left\| \frac{f}{\tilde{f}} \right\|_\infty, \qquad \textit{for all } m \geqslant 2.$$

Proof of Theorem 3. Let s be the square root of the true density $\mathrm{d}P/\mathrm{d}\mu$, s^* a given point in S, and $\hat{s}$ a $(1/n)$-maximum likelihood estimator on S. Set $P^* = P_{s^*}$ and $\hat{P} = P_{\hat{s}}$. Since μ is a probability measure, one can define η by $\int (s^2 \vee \eta)\, \mathrm{d}\mu = 1 + D/n$ and $\mathrm{d}\tilde{P}/\mathrm{d}\mu = \tilde{s}^2 = (s^2 \vee \eta)/(1 + D/n)$. Using the fact that $D \leqslant n$, one can easily check that $\eta \geqslant D/n$, $\inf_x \tilde{s}^2(x) \geqslant \eta/2$,

$$h^2(P, \tilde{P}) = \tfrac{1}{2} \int (\sqrt{\mathrm{d}P} - \sqrt{\mathrm{d}\tilde{P}})^2 \leqslant \frac{D}{2n} \tag{7.9}$$

and

$$\left\| \frac{s}{\tilde{s}} \right\|_\infty^2 \leqslant 1 + \frac{D}{n} \leqslant 2. \tag{7.10}$$

We want to show that maximum likelihood estimation fits in with our general framework and apply Corollary 1. We therefore have to check Assumptions M1, M2 and C. Let us first recall that $\|u - v\|^2 = 2h^2(P_u, P_v)$. Since we cannot directly work with the function $\overline{\gamma}(z, t)$ we have to introduce here the function $\gamma(z, t) = -\log[(\tilde{s}^2 + t^2)/2](z)$ which is a slight modification of the one we used in our treatment of general maximum likelihood estimation (Birgé and Massart 1993). Taking $x = z$, we obtain $|\gamma(z, t) - \gamma(z, u)| = M\Delta(x, t, u)$ with

$$\Delta(x, t, u) = \frac{1}{2} \left| \log \frac{\tilde{s}^2(x) + t^2(x)}{\tilde{s}^2(x) + u^2(x)} \right|$$

and $M = A = 2$, from which we get (5.2). Bound (5.6) follows from Lemma 6 and the lower bound $\tilde{s} \geqslant \sqrt{\eta/2}$. Indeed, if $\|t - u\|_\infty \leqslant r'\delta$, $\|\Delta(\cdot, t, u)\|_\infty \leqslant \sqrt{2} r'\delta/\sqrt{\eta}$ and (5.6) holds with

$$r = \frac{\sqrt{2} r'}{\sqrt{\eta}} \leqslant r' \left(\frac{2n}{D} \right)^{1/2}. \tag{7.11}$$

Finally, Lemma 7 and (7.10) imply (5.3) with $B = 1$. Assumptions M1 and M2 are therefore satisfied.

It only remains to check Assumption C. Assuming that $\gamma_n(t) \leqslant \gamma_n(s^*) + 1/n$, we see that $\sum_i \log t(Z_i) \geqslant \sum_i \log s^*(Z_i) - 1$, which implies by convexity

$$\sum_{i=1}^{n} \log\left[\frac{t^2 + \tilde{s}^2}{2}\right](Z_i) \geqslant \sum_{i=1}^{n} \log[\tilde{s}(Z_i)s^*(Z_i)] - 1$$

and therefore, since by (7.10) $\log \tilde{s} \geqslant \log s - D/(2n)$,

$$\begin{aligned}
\nu_n[\gamma(\cdot, s^*) - \gamma(\cdot, t)] &\geqslant \mathrm{P}_n\left[\log \frac{2\tilde{s}s^*}{\tilde{s}^2 + (s^*)^2}\right] - \frac{1}{n} \\
&\quad - \mathrm{E}_s\left[\log \frac{\tilde{s}^2(Z_1) + t^2(Z_1)}{2s^2(Z_1)} - \log \frac{\tilde{s}^2(Z_1) + (s^*)^2(Z_1)}{2s^2(Z_1)}\right] \\
&\geqslant \mathrm{P}_n\left[\log \frac{2s^2}{\tilde{s}^2 + (s^*)^2}\right] - \frac{1}{2}\mathrm{P}_n\left[\log \frac{s^2}{(s^*)^2}\right] - \frac{D+2}{2n} \\
&\quad + K\left(P, \frac{\tilde{P} + P_t}{2}\right) - K\left(P, \frac{\tilde{P} + P^*}{2}\right).
\end{aligned}$$

In order to conclude we use inequalities (7.2) and (7.3) from Lemma 5 to obtain

$$\begin{aligned}
0.29h(P, P_t) < h\left(P, \frac{P + P_t}{2}\right) &\leqslant h\left(P, \frac{\tilde{P} + P_t}{2}\right) + h\left(\frac{P + P_t}{2}, \frac{\tilde{P} + P_t}{2}\right) \\
&\leqslant h\left(P, \frac{\tilde{P} + P_t}{2}\right) + \frac{1}{\sqrt{2}} h(P, \tilde{P})
\end{aligned}$$

and therefore

$$[0.29h(P, P_t)]^2 \leqslant 2h^2\left(P, \frac{\tilde{P} + P_t}{2}\right) + h^2(P, \tilde{P})$$

and by (7.5)

$$K\left(P, \frac{\tilde{P} + P_t}{2}\right) \geqslant 2h^2\left(P, \frac{\tilde{P} + P_t}{2}\right) \geqslant 0.29^2 h^2(P, P_t) - h^2(P, \tilde{P}).$$

Then (7.10) and (7.6) imply that $K(P, \tilde{P}) < 4.6h^2(P, \tilde{P})$ and by (7.4)

$$K\left(P, \frac{\tilde{P} + P^*}{2}\right) \leqslant \frac{1}{2}K(P, \tilde{P}) + \frac{1}{2}K(P, P^*) < 2.3h^2(P, \tilde{P}) + \frac{1}{2}K(P, P^*).$$

Finally, since $K(P, Q) \geqslant 0$ for any Q,

$$\mathrm{E}_s\left[\frac{1}{2}\log\frac{s^2}{(s^*)^2}(Z_1) - \log\frac{2s^2}{\tilde{s}^2+(s^*)^2}(Z_1)\right] \leqslant \frac{1}{2}K(P,\, P^*). \tag{7.12}$$

Putting all inequalities together, we finally see that Assumption C will be satisfied with $k = 0.29^2/2$ and

$$U^2 = \frac{1}{2}\mathrm{P}_n\left[\log\frac{s^2}{(s^*)^2}\right] - \mathrm{P}_n\left[\log\frac{2s^2}{\tilde{s}^2+(s^*)^2}\right] + \frac{D+2}{2n} + 3.3h^2(P,\, \tilde{P}) + \frac{1}{2}K(P,\, P^*).$$

It then follows from (7.9) and (7.12) that

$$\mathrm{E}_s[U^2] \leqslant \frac{4.3D+2}{2n} + K(P,\, P^*)$$

which, together with Corollary 1, gives the bound on the quadratic risk of $\hat{s}$ since s^* is arbitrary. Finally (4.1) follows from (7.6). □

Remark. The value of $\mathscr{L}$ in Theorem 3 is derived from the upper bound on r given in (7.11) but it also follows from (7.11) that Theorem 3 actually holds with

$$\mathscr{L} = 1 + \log B' + \log\left[1 + r'\left(\frac{2D}{n\eta}\right)^{1/2}\right].$$

We know that $\eta \geqslant D/n$, but it can actually be much larger than D/n. For instance, when s is bounded away from zero η can be chosen independently of n and $\mathscr{L}$ will behave as $1 + \log B' + \log[1 + r'(D/n)^{1/2}]$. Then r' plays here the same role as r in Theorem 5. All intermediate situations are possible according to the form of s and even if r' is unbounded, $r'[D/(\eta n)]^{1/2}$ can be bounded if D/n is small enough.

7.5. Proof of Theorem 4

We again have to check Assumptions M1, M2 and C. In the case of least-squares regression, we choose

$$\gamma(z,\, t) = [y - t(x)]^2 = [s(x) + w - t(x)]^2.$$

Then

$$\begin{aligned}|\gamma(z,\, u) - \gamma(z,\, v)| &= |u(x) - v(x)|\,|2w + 2s(x) - [u(x) + v(x)]| \\ &\leqslant 2|u(x) - v(x)|[|w| + H].\end{aligned}$$

We can therefore take $\Delta(x,\, u,\, v) = |u(x) - v(x)|$ and $M(w) = 2(|w| + H)$. The moment condition on the W_is implies (5.2) with $A = A(\alpha,\, \Gamma,\, H)$ and (5.3) follows with $B = 2H$ from the boundedness of $\|u - v\|$. Assumption M2 is satisfied because of Covering Property $\mathrm{M}(n^{-1/2})$. Finally, Assumption C′ follows from

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_s[\gamma(Z_i, t) - \gamma(Z_i, s)] = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_s[(s-t)^2(X_i)] = \|t-s\|^2.$$

In the case of minimum-L_1 regression one sets

$$\gamma(z, t) = |s(x) + w - t(x)|, \qquad \Delta(x, u, v) = |u(x) - v(x)|, \qquad M = 1.$$

Assumption M1 clearly follows as before with $A = 1$ and $B = 2H$, and M2 from Covering Property M($n^{-1/2}$). It remains to check Assumption C′ and we will follow here the proofs and notation of Birgé and Massart (1993, pp. 125–127). With G defined in Assumption Ce of Birgé and Massart (1993), the decomposition

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

and the computations of Birgé and Massart (1993), together with our boundedness assumptions, show that for some constants C_1 and C_2

$$C_1\|s-t\|_2^2 \leqslant \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_s[G(W_i, s(X_i) - t(X_i))] \leqslant C_2\|s-t\|_2^2,$$

which gives Assumption C′. □

7.6. Bernstein's inequality

In order to obtain the relevant exponential inequalities needed for the proof of Theorem 5 we shall repeatedly need the following version of Bernstein's inequality. This version is not exactly standard because of its assumptions (moment controls on $|Z_i|$ instead of $|Z_i - \mathrm{E}[Z_i]|$) and its conclusion (unusual form of the bound $\exp(-nx)$). We give here a sketch of the proof since we were unable to find it in the literature. For proofs of the classical inequalities, see Uspensky (1937).

Lemma 8. *Let $Z_1, \ldots, Z_n$ be independent random variables satisfying the moments conditions*

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}[|Z_i^m|] \leqslant \frac{m!}{2} v^2 c^{m-2} \qquad \text{for all } m \geqslant 2, \tag{7.13}$$

for some positive constants v and c. Then, for any positive ε and $S_n = \sum_{i=1}^n Z_i$,

$$\mathrm{P}[S_n - \mathrm{E}(S_n) \geqslant n\varepsilon] \leqslant \exp(-nx) \leqslant \exp\left(\frac{-n\varepsilon^2/2}{v^2 + c\varepsilon}\right), \tag{7.14}$$

where x is defined by the equation $\varepsilon = v\sqrt{2x} + cx$.

Proof. Let us put $\mu_i = \mathrm{E}(Z_i)$ and $r_i = \mathrm{E}[\mathrm{e}^{\lambda Z_i}] - 1 - \lambda\mu_i$ for $\lambda > 0$; then

$$\log \mathrm{E}[\mathrm{e}^{\lambda(Z_i - \mu_i)}] = -\lambda\mu_i + \log(1 + \lambda\mu_i + r_i) \leqslant r_i$$

and the bounds (7.13) lead to

$$\sum_{i=1}^{n} \log \mathrm{E}[\mathrm{e}^{\lambda(Z_i-\mu_i)}] \leqslant \sum_{i=1}^{n} \mathrm{E}[\mathrm{e}^{\lambda Z_i}] - n - \lambda \mathrm{E}(S_n) \leqslant \frac{nv^2\lambda^2}{2(1-c\lambda)}.$$

An application of the classical exponential inequality

$$\mathrm{P}\left[\sum_{i=1}^{n} Y_i \geqslant n\varepsilon\right] \leqslant \exp\left[\inf_{y\geqslant 0}\left(-ny\varepsilon + \sum_{i=1}^{n} \log \mathrm{E}[\mathrm{e}^{yY_i}]\right)\right]$$

therefore implies that $\mathrm{P}[S_n - \mathrm{E}(S_n) \geqslant n\varepsilon] \leqslant \exp[-nh(\varepsilon)]$, where

$$h(\varepsilon) = \sup_{\lambda>0}\left\{\lambda\varepsilon - \frac{v^2\lambda^2}{2(1-c\lambda)}\right\}.$$

The supremum is achieved for $\lambda = c^{-1}[1 - v(2\varepsilon c + v^2)^{-1/2}]$, and since $(1+t)^{-1/2} \leqslant 1 + t/2$ we find that

$$h(\varepsilon) = \frac{\varepsilon^2}{\varepsilon c + v^2 + v^2(1 + 2\varepsilon c/v^2)^{1/2}} \geqslant \frac{\varepsilon^2}{2c\varepsilon + 2v^2}.$$

Moreover, it is easy to check that $h(\varepsilon) = x$, which gives (7.14). □

7.7. Proof of Theorem 5

Let us define $\mathscr{L}'$ by the implicit equation

$$\mathscr{L}' = 5\log[B'(5 + 4r(D\mathscr{L}'/n)^{1/2})], \tag{7.15}$$

which clearly defines $\mathscr{L}'$ in a unique way. We first want to prove that if $\mathscr{B}(s^*, \sigma)$ denotes the L_2 ball of centre s^* and radius σ, then

$$\mathrm{P}_s[\sup_{u\in\mathscr{B}(s^*,\sigma)} \nu_n[\gamma(\cdot, s^*) - \gamma(\cdot, u)] > \tau\sigma^2] \leqslant 2.1\exp\left[-\frac{3n\sigma^2}{10\rho^2(\tau)}\right] < 2.1\exp\left[\frac{-12D}{5}\right] \tag{7.16}$$

provided that $n\sigma^2 \geqslant D(\mathscr{L}'\rho^2(\tau) \vee 1)$. Let us set $\rho = \rho(\tau)$, $\sigma = \rho(DL/n)^{1/2}$ with $L \geqslant \mathscr{L}' \vee \rho^{-2}$, $\theta = 5 + 4r(DL/n)^{1/2}$, $\delta_k = 2^{-k}\sigma/\theta$ for $k \in \mathbb{N}$, $\mathscr{B} = \mathscr{B}(s^*, \sigma)$ and $f_u = \gamma(\cdot, s^*) - \gamma(\cdot, u)$. Since $\sigma^2 \geqslant D/n$, by Assumption M2 we can assume the existence of δ_k-nets $T_k = T(\delta_k, \sigma)$ with respective cardinalities e^{H_k},

$$H_k = D\log(B'\sigma/\delta_k) = D\log(B'2^k\theta), \tag{7.17}$$

and given some point u in $\mathscr{B}$ we can find a sequence $\{t_k\}_{k\geqslant 0}$ with $t_k \in T_k$ such that, according to Assumptions M1 and M2,

$$\frac{1}{n}\sum_{i=1}^{n} \mathrm{E}_s[\Delta^2(X_i, u, t_k)] \leqslant \delta_k^2 \qquad \text{and} \qquad \|\Delta(\cdot, u, t_k)\|_\infty \leqslant r\delta_k. \tag{7.18}$$

Setting $f_k = f_{t_k}$, we can use the decomposition $f_u = f_0 + \sum_{k=0}^{+\infty}(f_{k+1} - f_k)$, since f_k converges to f_u when k goes to infinity because of (7.18), and obtain the bound

$$P_s[\sup_{u\in\mathscr{B}} \nu_n(f_u) > \tau\sigma^2] \leqslant \sum_{t_0\in T_0} P_s[\nu_n(f_0) > A\eta] + \sum_{k=0}^{+\infty} \sum_{t_k, t_{k+1}} P_s[\nu_n(f_{k+1} - f_k) > A\eta_k] = P_1 + P_2$$

provided that the (as yet not chosen) parameters η and η_k, $k \geqslant 0$, satisfy the relation

$$\eta + \sum_{k=0}^{+\infty} \eta_k \leqslant \tau\sigma^2/A. \tag{7.19}$$

Control of $\nu_n(f_0)$. From the independence of W_i and X_i, (5.2) and (5.3) with $m = 2$, we obtain

$$E_s[|\gamma(Z_i, s^*) - \gamma(Z_i, t_0)|^m] \leqslant E_s[M^m(W_i)]E_s[\Delta^m(X_i, s^*, t_0)] \leqslant a_m A^m E_s[\Delta^m(X_i, s^*, t_0)]$$

and

$$\frac{1}{n}\sum_{i=1}^{n} E_s[|\gamma(Z_i, s^*) - \gamma(Z_i, t_0)|^m] \leqslant a_m A^m b_m \sigma^2 B^{m-2} = \frac{m!}{2}(A\sigma)^2(AB)^{m-2}.$$

Therefore Bernstein's inequality (7.14) implies that, if $\eta = \sigma\sqrt{2x} + Bx$, then

$$P_s[\nu_n(f_0) > A\eta] \leqslant \exp(-nx). \tag{7.20}$$

Control of $\nu_n(f_{k+1} - f_k)$. Since

$$|\gamma(Z_i, t_{k+1}) - \gamma(Z_i, t_k)| \leqslant M(W_i)[\Delta(X_i, u, t_k) + \Delta(X_i, u, t_{k+1})],$$

we use (5.2) and (7.18) to obtain

$$\begin{aligned} E_s[|\gamma(Z_i, t_{k+1}) - \gamma(Z_i, t_k)|^m] &\leqslant E_s[M^m(W_i)]E_s[(\Delta(X_i, u, t_k) + \Delta(X_i, u, t_{k+1}))^m] \\ &\leqslant a_m A^m E_s[(\Delta(X_i, u, t_k) + \Delta(X_i, u, t_{k+1}))^2](\|\Delta(\cdot, u, t_k)\|_\infty \\ &\quad + \|\Delta(\cdot, u, t_{k+1})\|_\infty)^{m-2}, \end{aligned}$$

and

$$\begin{aligned} \frac{1}{n}\sum_{i=1}^{n} E_s[|\gamma(Z_i, t_{k+1}) - \gamma(Z_i, t_k)|^m] &\leqslant [a_m A^m][2(\delta_k^2 + \delta_{k+1}^2)(r\delta_k + r\delta_{k+1})^{m-2}] \\ &\leqslant \frac{m!}{2}\left[\frac{5}{2}\delta_k^2 A^2\right]\left[A\frac{3r\delta_k}{2}\right]^{m-2}, \end{aligned}$$

from which we deduce that if $\eta_k = \delta_k(\sqrt{5x_k} + 1.5rx_k)$, then

$$P_s[\nu_n(f_{k+1} - f_k) > A\eta_k] \leqslant \exp(-nx_k). \tag{7.21}$$

Control of $\sup_{u\in\mathscr{B}} \nu_n(f_u)$. Since $\theta > 5$ and $L \geqslant 5\log(B'\theta) > 8$, as can easily be seen from the definition of $\mathscr{L}'$, (7.17) implies that $H_k \leqslant D(L/5 + k\log 2)$, for all $k \in \mathbb{N}$. Therefore (7.20), (7.21), $D \geqslant 1$ and $DL = n\sigma^2/\rho^2 > 8D \geqslant 8$ imply the probability bound in (7.16), since choosing

$$x = \frac{DL}{2n} \qquad \text{and} \qquad x_k = \frac{D}{n}\left[\frac{2L}{5} + (2k+1)\log 2 + \frac{3(k+1)L}{10}\right],$$

we obtain

$$\mathrm{P}_1 \leqslant \exp(H_0 - nx) \leqslant \exp(-3DL/10) < \exp(-12D/5)$$

and

$$\mathrm{P}_2 \leqslant \sum_{k=0}^{\infty} \exp(H_k + H_{k+1} - nx_k) \leqslant \sum_{k=0}^{\infty} \exp[-3(k+1)DL/10]$$

$$\leqslant \frac{\exp(-3DL/10)}{1 - \exp(-3DL/10)} < 1.1\exp(-3DL/10).$$

Since $\log 2 \leqslant (L\log 2)/8$, nx_k is bounded by $(0.474k + 0.787)DL$ and numerical computation leads to

$$\sum_{k=0}^{\infty} \eta_k < \frac{\sigma}{\theta}\left[5\left(\frac{DL}{n}\right)^{1/2} + 4r\frac{DL}{n}\right] = \frac{\sigma^2}{\rho}.$$

Since $\eta = (\rho + B/2)(\sigma^2/\rho^2)$ and $\tau\rho^2/A = 2\rho + B/2$, the constraint (7.19) is satisfied.

Let us now show that $\mathscr{L}' \leqslant \overline{\mathscr{L}}$, where $\overline{\mathscr{L}}$ is given by (5.7). Starting from the inequality $a + b \leqslant (a/\lambda) \vee (b/(1-\lambda))$ for $a,\ b \geqslant 0$ and $0 < \lambda < 1$, one derives from (7.15) that

$$\frac{\mathscr{L}'}{5} \leqslant \log\left(\frac{5B'}{1-\lambda}\right) \vee \left[\log\left(\frac{4rB'}{\lambda}\left(\frac{5D}{n}\right)^{1/2}\right) + \frac{1}{2}\log\left(\frac{\mathscr{L}'}{5}\right)\right]. \tag{7.22}$$

Using the fact that $\log x \leqslant x/\mathrm{e}$ we get that $\log(\mathscr{L}'/5) \leqslant \mathscr{L}'/(5\mathrm{e})$ and from (7.22) we derive that

$$\mathscr{L}' \leqslant 5\log\left(\frac{5B'}{1-\lambda}\right) \vee \frac{10\mathrm{e}}{2\mathrm{e}-1}\log\left(\frac{4rB'}{\lambda}\left(\frac{5D}{n}\right)^{1/2}\right)$$

$$\leqslant \frac{10\mathrm{e}}{2\mathrm{e}-1}\left[\log B' + \left(\log\left(\frac{5}{1-\lambda}\right) \vee \log\left(\frac{4\sqrt{5}}{\lambda}\right)\right) + \log\left(1 + r\left(\frac{D}{n}\right)^{1/2}\right)\right].$$

Choosing $\lambda = 4/(4+\sqrt{5})$, we conclude that $\mathscr{L}' \leqslant \overline{\mathscr{L}}$ and therefore that (7.16) holds provided that $n\sigma^2 \geqslant D[\overline{\mathscr{L}}\rho^2(\tau) \vee 1]$.

In order to prove (5.8), let us now set $\lambda = 4/3$, $\rho = \rho(\tau/\lambda)$, $\sigma_0 = 0$, $\sigma_j^2 = \lambda^j\sigma^2$ for $j \geqslant 1$ and observe that the following is valid by (7.16) since $\sigma \geqslant \sigma_D \geqslant \sqrt{D/n}$ and $n\sigma^2/\rho^2 \geqslant 8$:

$$\mathrm{P}_s\left[\sup_{u \in S} \frac{\nu_n[\gamma(\cdot, s^*) - \gamma(\cdot, u)]}{d^2(s^*, u) \vee \sigma^2} > \tau\right]$$

$$\leqslant \sum_{j=0}^{+\infty} \mathrm{P}_s\left[\sup_{\sigma_j \leqslant d(s^*,u)<\sigma_{j+1}} \frac{\nu_n[\gamma(\cdot, s^*)-\gamma(\cdot, u)]}{\sigma_{j+1}^2/\lambda} > \tau\right]$$

$$\leqslant \sum_{j=0}^{+\infty} \mathrm{P}_s\left[\sup_{\mathscr{B}(s^*,\sigma_{j+1})} \nu_n[\gamma(\cdot, s^*)-\gamma(\cdot, u)] > \frac{\tau}{\lambda}\sigma_{j+1}^2\right] \leqslant 2.1\sum_{j=0}^{+\infty} \exp\left[-\frac{3n\lambda^{j+1}\sigma^2}{10\rho^2}\right]$$

$$\leqslant 2.1\exp\left[-\frac{3n\lambda\sigma^2}{10\rho^2}\right]\sum_{j=0}^{+\infty}\exp[-2.4\lambda(\lambda^j-1)] < 3.03\exp\left[-\frac{2n\sigma^2}{5\rho^2}\right],$$

which is (5.8). □

Remark. One should notice that the assumptions of Theorem 5 warrant that the suprema of empirical processes involved in the statement and proof of the theorem are measurable. Indeed our L_∞ assumptions in M2 ensure that these empirical processes are separable.

7.8. Proof of Proposition 2

The proof relies on the following result.

Lemma 9. *Let Ω be a finite set with M elements, λ be the counting measure on Ω, $C \leqslant M/2$ be a positive integer and $\mathscr{P}$ be the set of all subsets of cardinality C of Ω. We consider the distance δ on $\mathscr{P}$ given by*

$$\delta(A, B) = \tfrac{1}{2}\int |\mathbb{1}_A(x) - \mathbb{1}_B(x)|\, \mathrm{d}\lambda(x), \qquad \text{for all } A, B \in \mathscr{P}. \tag{7.23}$$

Given some integer q with $0 < q < C(M-C)/M$, let $\mathscr{M}$ be a maximal subset of $\mathscr{P}$ such that

$$\delta(A, B) > q, \qquad \text{for all } A, B \in \mathscr{M},\ A \neq B.$$

Then

$$|\mathscr{M}| \geqslant \left[1 - \frac{q^2}{(C-q)(M-C-q)}\right]\binom{M}{C}\left[\binom{C}{C-q}\binom{M-C}{q}\right]^{-1}.$$

Proof. Let us first observe that $\delta(A, B)$ can take any integer value between 0 and C. Clearly $\mathscr{M}$ exists since $q < C$. Since $\mathscr{M}$ is maximal, any element A in $\mathscr{P}\backslash\mathscr{M}$ satisfies $\delta(A, \mathscr{M}) \leqslant q$ and therefore $\mathscr{M}$ is a q-net for $\mathscr{P}$. This means that $\mathscr{P}$ can be covered by the balls of radius q centred on the elements of $\mathscr{M}$ and consequently that

$$|\mathscr{P}| = \binom{M}{C} \leqslant |\mathscr{M}|V, \tag{7.24}$$

where V denotes the cardinality of a ball of radius q. Now, given an element A in $\mathscr{P}$, the number of elements B such that $\delta(A, B) = j$ is given by

$$\binom{C}{C-j}\binom{M-C}{j}$$

and

$$\begin{aligned} V &= \sum_{j=0}^{q} \binom{C}{C-j}\binom{M-C}{j} \\ &\leqslant \binom{C}{C-q}\binom{M-C}{q} \sum_{j=0}^{q} \left[\frac{q^2}{(C-q+1)(M-C-q+1)}\right]^j \\ &\leqslant \left[1 - \frac{q^2}{(C-q)(M-C-q)}\right]^{-1} \binom{C}{C-q}\binom{M-C}{q} \end{aligned}$$

since the upper bound on q implies that the bracketed terms are smaller than 1. The conclusion then follows from (7.24). □

Proof of Proposition 2. Let us denote by $\mathrm{Int}[x]$ the integer part of the positive number x. Noticing that $4 \leqslant 4D/9 \leqslant D' < D/2$, we define $k = \mathrm{Int}[N \log(n/D')/(2.96n)]$ and $N' = \mathrm{Int}[N/k]$. Our assumptions imply that $1 \leqslant k < N/(6D)$ and therefore that $N' \geqslant 12D'$. We denote by $\mathscr{P}$ the set of all subsets of $\{1, \ldots, N'\}$ of cardinality D', and if $A \in \mathscr{P}$ we denote by f_A the element of $\mathscr{H}_N(D)$ defined by

$$f_A = (1-\eta)\mathbb{1}_{[0,1)} + \sum_{i \in A} \mathbb{1}_{[k(i-1)/N, ki/N)}, \qquad \text{where } 0 < \eta = \frac{kD'}{N} < \frac{1}{12}.$$

Clearly $\|f_A\|_\infty < 2$ for all $A \in \mathscr{P}$, and for any two sets A and B in $\mathscr{P}$ we have

$$\|f_A - f_B\|^2 = 2\frac{k}{N}\delta(A, B), \qquad h^2(f_A, f_B) = \frac{k}{N}\delta(A, B)(\sqrt{2-\eta} - \sqrt{1-\eta})^2$$

and

$$K(f_A, f_B) = \frac{k}{N}\delta(A, B)\left[(2-\eta)\log\frac{2-\eta}{1-\eta} + (1-\eta)\log\frac{1-\eta}{2-\eta}\right] = \frac{k}{N}\delta(A, B)\log\frac{2-\eta}{1-\eta}.$$

It finally follows from the bound on η that

$$h^2(f_A, f_B) = \lambda\frac{k}{N}\delta(A, B), \qquad \text{with } 0.171 < (\sqrt{2}-1)^2 \leqslant \lambda \leqslant 0.183, \tag{7.25}$$

and

$$K(f_A, f_B) = \left[\log\frac{2-\eta}{1-\eta}\right]\frac{k}{N}\delta(A, B) = \lambda'\frac{k}{N}\delta(A, B), \qquad \text{with } 0.69 < \log 2 \leqslant \lambda' \leqslant 0.74.$$

Let us now choose an integer q such that $2D'/5 \leqslant q \leqslant D'/2$, which always exists since $D' \geqslant 4$. It follows from Lemma 8 with $C = D'$ and $M = N'$ that one can find a subset $\mathscr{M}$ of $\mathscr{P}$ such that for any pair (A, B) in $\mathscr{M}$ with $A \neq B$, $\delta(A, B) > q \geqslant 2D'/5$ and

$$|\mathscr{M}| \geqslant \Sigma \binom{N'}{D'} \left[\binom{D'}{D'-q} \binom{N'-D'}{q} \right]^{-1},$$

with

$$\Sigma = 1 - \frac{q^2}{(D'-q)(N'-D'-q)} \geqslant 1 - \frac{D'^2/4}{(D'/2)(N'-3D'/2)} \geqslant \frac{20}{21},$$

since $q \leqslant D'/2$ and $N' \geqslant 12D'$. Some tedious computations involving Stirling's formula and the fact that $D' \geqslant 4$ and $N' \geqslant 12D'$ lead to

$$\log(|\mathscr{M}| - 1) > \frac{D'}{4}\left[1 + 2\log\left(\frac{N'}{D'}\right)\right].$$

Let us now consider an estimator $\hat{A}$ with values in $\mathscr{M}$ built on n independent and identically distributed observations from an unknown distribution f_A with A in $\mathscr{M}$. It follows from Fano's lemma – see, for instance, Lemma 2.7 of Birgé (1983) – that

$$\sup_{A \in \mathscr{M}} \mathrm{P}_{f_A}[\hat{A} \neq A)] \geqslant 1 - \frac{0.74nkD'/N + \log 2}{(D'/4)[1 + 2\log(N'/D')]} \geqslant 0.883 - \frac{2.96nk}{N[1 + 2\log(N'/D')]}.$$

Since $N/k > 6D \geqslant 54$, $N' > 54N/(55k) > 2.9n/\log(n/D')$ and therefore

$$\sup_{A \in \mathscr{M}} \mathrm{P}_{f_A}\left[\delta(\hat{A}, A) > \frac{2D'}{5}\right] \geqslant 0.883 - \frac{\log(n/D')}{1 + 2\log(2.9n/D') - 2\log[\log(n/D')]} > \tfrac{1}{3}$$

because the function $x/(0.5 + \log 2.9 + x - \log x)$ is bounded by 1.084 for $x \geqslant 1$. It then follows from (7.25) that

$$\sup_{A \in \mathscr{M}} \mathrm{P}_{f_A}\left[h^2(f_{\hat{A}}, f_A) > \kappa \frac{D}{n}\log\left(\frac{n}{D}\right)\right] \geqslant \sup_{A \in \mathscr{M}} \mathrm{P}_{f_A}\left[h^2(f_{\hat{A}}, f_A) > \frac{2D'}{5}\frac{\lambda k}{N}\right] > \tfrac{1}{3},$$

for a suitable constant κ since $k > N\log(n/D')/(5.92n)$ and $D' \geqslant 4D/9$. The conclusion then follows by standard arguments, and a similar result holds if we replace the distance h by the L_2 distance. □

7.9. Proof of Proposition 4

The proof of (i) is actually rather similar to the proof of Proposition 1. Let us consider the canonical isomorphism between $H(\mathscr{P})$ and $\mathbb{R}^D$ which is associated to the orthonormal basis $\{\varphi_I, I \in \mathscr{P}\}$ with $\varphi_I = |I|^{-1/2}\mathbb{1}_I$. The ball $\mathscr{B}$ of radius σ corresponds to a ball $\mathscr{B}'$ in $\mathbb{R}^D$. Introduce a covering of $\mathscr{B}'$ by hyperrectangles with side lengths $|I|^{1/2}\delta$ and pick a point in each hyperrectangle. Let us denote by T' the resulting set of points and by T its homologue in $H(\mathscr{P})$ (using the isomorphism). Then the following properties hold: first, the L_2 and L_∞ diameters of each hypercube are both equal to δ; second, the number of hypercubes needed to

cover $\mathscr{B}'$ is bounded by the number of such hypercubes that are contained in a ball of radius $\sigma + \delta$, which, using volume comparisons, is bounded by

$$\left(\frac{2\pi e}{D}\right)^{D/2}(\pi D)^{-1/2}\left(\frac{\sigma+\delta}{\delta}\right)^{D}\prod_{I\in\mathscr{P}}|I|^{-1/2} \leqslant \left[(2\pi e)^{1/2}\left(1+\frac{\delta}{\sigma}\right)\right]^{D}\left(\frac{\sigma}{\delta}\right)^{D}\left(\prod_{I\in\mathscr{P}}D|I|\right)^{-1/2}.$$

Then (i) follows since $(2\pi e)^{1/2}(1+\delta/\sigma) < 5$.

To prove (ii), we consider the set $T \in H(\mathscr{P})$ of functions of the form $(\delta/\sqrt{D})\sum_{I\in\mathscr{P}} a_I \varphi_I$, where $a_I \in \mathbb{Z}$ for all I. For any $t \in H(\mathscr{P})$ one can find a pair of functions $t^- \leqslant t \leqslant t^+$ with t^- and $t^+ \in T$, $t^+ = t^- + (\delta/\sqrt{D})\sum_{I\in\mathscr{P}}\varphi_I$ and $\|t^+ - t^-\| = \delta$. Using again the same isomorphism between $H(\mathscr{P})$ and $\mathbb{R}^D$, we see that a bracket $[t^-, t^+]$ corresponds to a hypercube of $\mathbb{R}^D$ of volume $(\delta/\sqrt{D})^D$ and the above argument shows that the number of such brackets needed to cover a ball of radius σ is bounded by $(5\sigma/\delta)^D$. □

Acknowledgements

We thank Andrew Barron and Emmanuel Rio for their helpful suggestions and corrections and, more especially, a referee who suffered greatly on account of an earlier version of this paper and provided us with many critical but very helpful comments.

References

Assouad, P. (1983) Deux remarques sur l'estimation. *C. R. Acad. Sci. Paris Sér. I Math.*, **296**, 1021–1024.

Bahadur, R.R. (1958) Examples of inconsistency of maximum likelihood estimates. *Sankhyá Ser. A*, **20**, 207–210.

Barron, A.R. (1994) Approximation and estimation bounds for artificial neural networks. *Mach. Learning*, **14**, 115–133.

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

Barron, A.R., Birgé, L. and Massart, P. (1997) Risk bounds for model selection via penalization. *Probab. Theory Related Fields.* To appear.

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. Theory Related Fields*, **71**, 271–291.

Birgé, L. and Massart, P. (1993) Rates of convergence for minimum contrast estimators. *Probab. Theory Related Fields*, **97**, 113–150.

Birgé, L. and Massart, P. (1997) From model selection to adaptive estimation. In D. Pollard, E. Torgersen and G. Yang (eds), *Festschrift for Lucien Le Cam: Research Papers in Probability and Statistics*, pp. 55–87. New York: Springer-Verlag.

Birman, M.S. and Solomjak, M.Z. (1967) Piecewise-polynomial approximation of functions of the classes W_p. *Mat. Sb.*, **73**, 295–317.

Cencov, N.N. (1962) Evaluation of an unknown distribution density from observations. *Soviet Math.*, **3**, 1559–1562.

Chow, Y.-S. and Grenander, U. (1985) A sieve method for the spectral density. *Ann. Statist.*, **13**, 998–1010.

Cox, D.D. (1988) Approximation of least squares regression on nested subspaces. *Ann. Statist.*, **16**, 713–732.

DeVore, R.A. and Lorentz, G.G. (1993) *Constructive Approximation*. Berlin: Springer-Verlag.

DeVore, R.A., Jawerth, B. and Popov, V. (1992) Compression of wavelets decompositions. *Amer. J. Math.*, **114**, 737–785.

Donoho, D.L. and Johnstone, I.M. (1994) Ideal spatial adaptation via wavelet shrinkage. *Biometrika*, **81**, 425–455.

Donoho, D.L. and Johnstone, I.M. (1995) Minimax estimation via wavelet shrinkage. *Ann. Statist.* To appear.

Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508–539.

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 *École d'Été de Probabilités de Saint-Flour XII – 1982*, Lecture Notes in Math. 1097. Berlin: Springer-Verlag.

Feller, W. (1968) *An Introduction to Probability Theory and its Applications, Vol. I*, 3rd edn. New York: Wiley.

Geman, S. (1981) Sieves for nonparametric estimation of densities and regression. Rep. Pattern Analysis, no. 99. DAM, Brown University.

Geman, S. and Hwang, C.-R. (1982) Nonparametric maximum likelihood estimation by the method of sieves. *Ann. Statist.*, **10**, 401–414.

Grenander, U. (1981) *Abstract Inference*. New York: Wiley.

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

Le Cam, L.M. (1975) On local and global properties in the theory or asymptotic normality of experiments. In M. Puri (ed.), *Stochastic Processes and Related Topics, Vol. 1*, pp. 13–54. New York: Academic Press.

Le Cam, L.M. (1986) *Asymptotic Methods in Statistical Decision Theory*. New York: Springer-Verlag.

Le Cam, L.M. and Yang, G.L. (1990) *Asymptotics in Statistics: Some Basic Concepts*. New York: Springer-Verlag.

Ledoux, M. (1996) On Talagrand's deviation inequalities for product measures. *ESAIM: Probab. Statist.*, **1**, 63–87. http//www.emath.fr/ps/.

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

Nemirovskii, A.S., Polyak, B.T. and Tsybakov, A.B. (1984) Signal processing by the nonparametric maximum-likelihood method. *Problems Inform. Transmission*, **20**, 177–192.

Ossiander, M. (1987) A central limit theorem under metric entropy with L_2 bracketing (1987). *Ann. Probab.*, **15**, 897–919.

Shen, X. and Wong, W.H. (1994) Convergence rates of sieve estimates, *Ann. Statist.*, **22**, 580–615.

Stone, C.J. (1982) Optimal rates of convergence for nonparametric regression. *Ann. Statist.*, **10**, 1040–1053.

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

Stone, C.J. (1994) The use of polynomial splines and their tensor products in multivariate function estimation. *Ann. Statist.*, **22**, 118–184.

Talagrand, M. (1996) New concentration inequalities in product spaces. *Invent. Math.*, **126**, 505–563.

Uspensky, J.V. (1937) *Introduction to Mathematical Probability.* New York: McGraw-Hill.

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

Van de Geer, S. (1995) The method of sieves and minimum contrast estimators. *Math. Methods Statist*., **4**, 20–38.

Wong, W.H. and Shen, X. (1992) Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. Technical report, University of Chicago.

Wong, W.H. and Shen, X. (1995) Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.*, **23**, 339–362.

Received December 1994 and revised May 1997.