AN ALTERNATIVE POINT OF VIEW ON LEPSKI'S METHOD

LUCIEN BIRGÉ

Université Paris VI

Lepski's method is a method for choosing a "best" estimator (in an appropriate sense) among a family of those, under suitable restrictions on this family. The subject of this paper is to give a nonasymptotic presentation of Lepski's method in the context of Gaussian regression models for a collection of projection estimators on some nested family of finite-dimensional linear subspaces. It is also shown that a suitable tuning of the method allows to asymptotically recover the best possible risk in the family.

AMS subject classifications: Primary: 62G07, Secondary 62G20.
Keywords and phrases: Adaptation, Lepski's method, Mallows' C_p, optimal selection of estimators.

1 Introduction

The aim of this paper is threefold. First we want to emphasize the importance of what is now called "Lepski's method", which appeared in a series of papers by Lepski (see Lepskii, 1990, 1991, 1992a and b). Then we shall present this method from an alternative point of view, different from the one initially developed by Lepski. Finally we shall introduce some generalization of the method and use it to prove some nice properties of it which, as far as we know, have not yet been considered, even by its initiator.

Let us first give a brief and simplified account of the classical method of Lepski. This method has been described in its general form and in great details in Lepskii (1991) and the interested reader should of course have a look at this milestone paper. Here we shall content ourselves to consider the problem within the very classical "Gaussian white noise model". According to Ibragimov and Has'minskii (1981, p.5), it has been initially introduced as a statistical model by Kotel'nikov (see Kotel'nikov, 1959). Since then, it has been extensively studied by many authors from the former Soviet Union (see for instance Ibragimov and Has'minskii, 1981, Pinsker, 1980, Efroimovich and Pinsker, 1984) and more recently by Donoho and Johnstone (1994a and b, 1995, 1996) and Birgé and Massart (1999), among many other references. Although not at all confined to this framework, the method has been often considered in the context of the Gaussian white noise model for the sake of simplicity. This model can be described by a stochastic differential equation of the form

$$(1.1) \qquad dY_\varepsilon(t) = s(t)\,dt + \varepsilon\,dW(t), \qquad \varepsilon > 0, \quad 0 \le t \le 1,$$

where $s \in \mathbb{L}_2([0,1])$ and W is a standard Brownian motion originating from 0. One wants to estimate the unknown function s using estimators $\hat{s}(\varepsilon)$, i.e. measurable functions of Y_ε and ε. By "estimator", Lepski actually means a family $\{\hat{s}(\varepsilon)\}$ of estimators depending on the parameter ε which is assumed to be small enough. In order to measure the performances of such estimators, a classical way is to fix some distance d on $\mathbb{L}_2([0,1])$ (or some pseudo-distance if $d(s,t) = 0$ does not necessarily imply that $s = t$ in $\mathbb{L}_2([0,1])$), some number $q \geq 1$ and define the risk of the estimator at s as $\mathbb{E}_s[d^q(s,\hat{s}(\varepsilon))]$. The point of view chosen by Lepski is then definitely minimax and asymptotic. He considers a family of parameter sets $\{\mathcal{S}_\theta\}_{\theta\in\Theta}$ and uniform rates of convergences of estimators over those parameter sets. For a given estimator $\hat{s}$, he defines its rate $r[\hat{s},\theta]$ on $\mathcal{S}_\theta$ and the minimax rate $r_M[\theta]$ on $\mathcal{S}_\theta$ given respectively by

$$r[\hat{s},\theta](\varepsilon) = \sup_{s\in\mathcal{S}_\theta} \mathbb{E}_s\left[d^q(s,\hat{s}(\varepsilon))\right] \qquad \text{and} \qquad r_M[\theta] = \inf_{\hat{s}} r[\hat{s},\theta],$$

where the infimum is taken over all possible estimators. Comparing estimators then amounts to comparing their rates, the rate r being better than the rate r' ($r \preceq r'$) if and only if $\limsup_{\varepsilon\to 0} r(\varepsilon)/r'(\varepsilon) < +\infty$ and two rates being equivalent ($r \asymp r'$) if $r \preceq r'$ and $r' \preceq r$. An estimator $\hat{s}$ is "rate asymptotically minimax" on $\mathcal{S}_\theta$, and therefore optimal from this point of view, if $r[\hat{s},\theta] \asymp r_M[\theta]$.

The problem that Lepski considers in his papers is the following: starting from a family of rate asymptotically minimax estimators $\{\hat{s}_\theta\}_{\theta\in\Theta}$, how can one get adaptation over the family $\{\mathcal{S}_\theta\}_{\theta\in\Theta}$, i.e. build a new estimator $\tilde{s}$ which is simultaneously rate asymptotically minimax over all the sets $\mathcal{S}_\theta$, i.e. satisfies $r[\tilde{s},\theta] \asymp r_M[\theta]$ for all $\theta \in \Theta$. Let us give a brief and rough account of his solution, rephrasing and simplifying his assumptions in the following way (see Lepskii, 1991 for the precise ones). Lepski's assumptions are essentially equivalent to

1. Θ is a bounded subset of $\mathbb{R}^+$;

2. the family $\{\mathcal{S}_\theta\}_{\theta\in\Theta}$ is nondecreasing with respect to θ;

3. the minimax rates $r_M[\theta]$ are, in a suitable sense, continuous with respect to θ;

4. for each $\theta \in \Theta$, one has available a rate asymptotically minimax estimator $\hat{s}_\theta$ on $\mathcal{S}_\theta$;

5. for ε small enough and each $\theta \in \Theta$, $d^q(s,\hat{s}_\theta(\varepsilon))$ is suitably concentrated around its expectation.

Lepski then chooses, for each ε, a suitable finite discretization $\theta_1 < \ldots < \theta_{n(\varepsilon)}$ of Θ and, given some large enough constant K, defines $\hat{\theta}(\varepsilon) = \theta_{\hat{\jmath}}(\varepsilon)$ where

$$\hat{\jmath} = \inf\{j \leq n(\varepsilon) \,|\, d^q(\hat{s}_{\theta_j}(\varepsilon), \hat{s}_{\theta_k}(\varepsilon)) \leq Kr[\hat{s}_{\theta_k}, \theta_k](\varepsilon) \quad \text{for all } k \in (j, n(\varepsilon)]\}.$$

He shows that $\tilde{s} = \hat{s}_{\hat{\theta}}$ is simultaneously rate asymptotically minimax over all the sets $\mathcal{S}_\theta$.

This problem of asymptotic adaptation can also be considered from a quite different point of view: if $s \in \mathcal{S} = \cup_{\theta\in\Theta}\mathcal{S}_\theta$, there exists a smallest value $\theta(s)$ of θ such that $s \in \mathcal{S}_\theta$ and, since we have therefore no idea of the behaviour of the risk $\mathbb{E}_s\left[d^q(s, \hat{s}_\theta(\varepsilon))\right]$ for $\theta < \theta(s)$, among the estimators at hand, $\hat{s}_{\theta(s)}$ can be considered as the best estimator for estimating s, among the family of estimators $\{\hat{s}_\theta\}_{\theta\in\Theta}$. From this point of view, the problem to be solved is to find a best estimator in a family of such estimators and it still makes sense without any reference to the minimax and even to the family $\{\mathcal{S}_\theta\}_{\theta\in\Theta}$. It can also be considered from a purely nonasymptotic point of view and set up as follows. Given Model (1.1) with a known value of ε and an unknown value of s, a family of estimators $\{\hat{s}_\theta(\varepsilon)\}_{\theta\in\Theta}$ and some loss function ℓ, is it possible to design a method for selecting an "almost best" estimator in the family? More precisely, assuming that $s \in \mathcal{S} \subset \mathbb{L}_2([0,1])$, does there exist a constant C, independent of ε and $s \in \mathcal{S}$ and a random selection procedure $\hat{\theta}$ based on Y_ε such that the estimator $\tilde{s} = \hat{s}_{\hat{\theta}}$ satisfies

$$(1.2) \quad \mathbb{E}_s[\ell(s, \tilde{s}(\varepsilon))] \leq C \inf_{\theta\in\Theta} \mathbb{E}_s[\ell(s, \hat{s}_\theta(\varepsilon))] \quad \text{for all } s \in \mathcal{S} \text{ and } \varepsilon > 0.$$

This is precisely the problem we shall deal with in this paper by a suitable modification of Lepski's initial recipe. In order to allow an easier understanding of our method and avoid technicalities, we shall stick to the Gaussian white noise model and restrict our study to the case of a family $\{\hat{s}_\theta(\varepsilon)\}_{\theta\in\Theta}$ of projection estimators over a nested family of finite-dimensional linear subspaces S_θ of $\mathbb{L}_2([0,1])$ indexed by some subset Θ of $\mathbb{N}$. We shall show that (1.2) actually holds with $\mathcal{S} = \mathbb{L}_2([0,1])$ and that one can even take C arbitrarily close to 1 when ε goes to zero under some suitable restrictions on s.

The framework we use here is just the one we considered in Birgé and Massart (1999) for studying penalized least squares estimators. Since penalization can also be viewed as a method for selecting estimators, this allows us to make a parallel between these two methods. Indeed, under the assumptions we use here, they are essentially equivalent. A discussion of the relative merits of the two methods within a more general framework is beyond the scope of this paper. Let us merely mention that Lepski's method allows to handle more general loss functions, while penalization allows to deal with more general families of estimators.

Lepski's method has been put to use in various contexts and by several authors. Let us mention here the papers by Efroimovich and Low (1994), Lepski and Spokoiny (1995), Juditsky (1997), Lepski, Mammen and Spokoiny (1997), Lepski and Levit (1998), Tsybakov (1998) and Butucea (1999). Recently, Lepski has substantially improved his method by relaxing the monotonicity assumptions he previously imposed and which were in particular inadequate to deal with estimation of multidimensional functions with anisotropic smoothness. His new method, which he explained in a series of lectures (Lepski, 1998) could analogously be carried out in the context we use below. In order to keep our presentation simple and short, we shall dispense with this extension and content ourselves to present our point of view derived from the initial method from Lepski (1991).

The procedure for selecting an estimator among some family that we develop below is actually not exactly the original procedure proposed by Lepski, but rather some modification of it which is better suited to our non-asymptotic approach and avoids any reference to minimaxity. Nevertheless, the ideas underlying our construction definitely belong to Lepski.

2 Preliminary considerations

2.1 The problem at hand

The problem we want to deal with is the estimation of some unknown function $s \in \mathbb{L}_2([0,1])$ in the Gaussian white noise model (1.1). In order to accomplish this task, we have at our disposal a family of projection estimators $\{\hat{s}_m\}_{m\in\mathcal{M}}$ corresponding to some nested family $\{S_m\}_{m\in\mathcal{M}}$ of finite-dimensional linear subspaces of $\mathbb{L}_2([0,1])$ with respective positive dimensions D_m. Here $\mathcal{M} \subset \mathbb{N}$ is either $\mathbb{N}^\star = \mathbb{N}\setminus\{0\}$ or finite and equal to $[1;M]\cap\mathbb{N}$ and the sequence $(D_m)_{m\in\mathcal{M}}$ is strictly increasing. We recall that the projection estimator $\hat{s}_m$ onto S_m is derived from Y_ε by the formula

$$\hat{s}_m = \sum_{j=1}^{D_m}\left[\int_0^1 \psi_j(t)dY_\varepsilon(t)\right]\psi_j,$$

where $(\psi_1,\ldots,\psi_{D_m})$ is an arbitrary orthonormal basis of S_m.

Our purpose is then as follows: starting from the family of estimators $\{\hat{s}_m\}_{m\in\mathcal{M}}$, build a new one, denoted by $\tilde{s}$, function of those, of ε and of the sequence $(D_m)_{m\in\mathcal{M}}$ and such that

$$\mathbb{E}\left[\|\tilde{s}-s\|^2\right] \le C \inf_{m\in\mathcal{M}} \mathbb{E}\left[\|\hat{s}_m - s\|^2\right],$$

with a constant C independent of s and ε. Here and in the sequel, $\mathbb{E}$ denotes the expectation with respect to the distribution of the process Y_ε, as defined by (1.1).

Since the family $\{S_m\}_{m\in\mathcal{M}}$ is nested, one can always find an orthonormal basis $\varphi_1,\ldots,\varphi_j,\ldots$ of $\mathbb{L}_2([0,1])$ such that S_m is the linear span of $(\varphi_1,\ldots,\varphi_{D_m})$ for each $m\in\mathcal{M}$. Then, if $s=\sum_{j\geq 1}\beta_j\varphi_j$, it follows from (1.1) that, for all $m\in\mathcal{M}$,

$$\hat{s}_m=\sum_{j=1}^{D_m}\hat{\beta}_j\varphi_j \qquad \text{with } \hat{\beta}_j=\int_0^1\varphi_j(t)dY_\varepsilon(t)=\beta_j+\varepsilon Z_j, \tag{2.3}$$

where the random variables Z_j, $j\in\mathcal{M}$, are i.i.d. with distribution $\mathcal{N}(0,1)$.

2.2 Some properties of projection estimators

In order to describe some elementary properties of the projection estimators, it will be useful to introduce some notations. Setting $D_0=0$ and $D_\infty=+\infty$, we consider for all pairs (m,q) with $0\leq m\leq q\leq+\infty$ the quantities B_m^q, V_m^q and U_m^q given by

$$B_m^q=\varepsilon^{-2}\sum_{j=D_m+1}^{D_q}\beta_j^2,\quad V_m^q=\sum_{j=D_m+1}^{D_q}Z_j^2\quad\text{and}\quad U_m^q=\varepsilon^{-2}\sum_{j=D_m+1}^{D_q}\hat{\beta}_j^2,$$

with the convention that $\sum_{j=k}^{l}=0$ when $k>l$. Since the variables Z_j are i.i.d. $\mathcal{N}(0,1)$, it then follows that V_m^q has the distribution $\chi^2(D_q-D_m)$ of a chi square with D_q-D_m degrees of freedom and U_m^q the distribution $\chi'^2(D_q-D_m,\sqrt{B_m^q})$ of a non-central chi square with D_q-D_m degrees of freedom and noncentrality parameter $\sqrt{B_m^q}$. Therefore

$$\mathbb{E}[V_m^q]=D_q-D_m\qquad\text{and}\qquad\mathbb{E}[U_m^q]=D_q-D_m+B_m^q.$$

One then derives from (2.3) that

$$\|\hat{s}_m-s\|^2=\varepsilon^2\left[\sum_{j=1}^{D_m}Z_j^2\right]+\sum_{j>D_m}\beta_j^2=\varepsilon^2\left(V_0^m+B_m^\infty\right), \tag{2.4}$$

$$\mathbb{E}\left[\|\hat{s}_m-s\|^2\right]=\varepsilon^2\left(D_m+B_m^\infty\right),$$

and for any pair $(j,m)\in\mathcal{M}^2$ with $j<m$,

$$\|\hat{s}_m-\hat{s}_j\|^2=\varepsilon^2U_j^m\quad\text{and}\quad\mathbb{E}\left[\|\hat{s}_m-\hat{s}_j\|^2\right]=\varepsilon^2\left(B_j^m+D_m-D_j\right). \tag{2.5}$$

2.3 Optimal projection estimators

Given any sequence $(x_m)_{m\in\mathcal{M}}$ such that $\lim_{m\to+\infty}x_m=+\infty$ when $\mathcal{M}$ is infinite, one defines in a unique way

$$\begin{aligned}\operatorname{argmin}\{x_m,\, m\in\mathcal{M}\} &= \inf\left\{j\;\middle|\;x_j=\inf_{m\in\mathcal{M}}x_m\right\}\\ &= \inf\{j\,|\,x_j\leq x_m\ \text{for all } m>j\}.\end{aligned} \tag{2.6}$$

Then, given s, a best estimator in the family $\{\hat{s}_m,\ m\in\mathcal{M}\}$, i.e. one minimizing the quadratic risk $\mathbb{E}\left[\|\hat{s}_m-s\|^2\right]=\varepsilon^2\left(D_m+B_m^\infty\right)$ at s is $\hat{s}_{\overline{m}}$ with $\overline{m}=\operatorname{argmin}\{D_m+B_m^\infty,\ m\in\mathcal{M}\}$. More generally, given some number $\gamma>0$ one can define

$$\begin{aligned} J &= \operatorname{argmin}\{\gamma D_m+B_m^\infty,\ m\in\mathcal{M}\}\\ &= \inf\left\{j\,|\,B_j^\infty-B_m^\infty\le\gamma(D_m-D_j)\ \text{ for all } m>j\right\}. \end{aligned}\tag{2.7}$$

Since $B_j^\infty-B_m^\infty=B_j^m$, it follows from (2.5) that

$$J=\inf\left\{j\ \left|\,\mathbb{E}\left[\|\hat{s}_m-\hat{s}_j\|^2\right]\le(1+\gamma)\varepsilon^2(D_m-D_j)\ \text{ for all } m>j\right.\right\}.\tag{2.8}$$

On the other hand, by (2.7)

$$\gamma D_J+B_J^\infty\le\gamma D_{\overline{m}}+B_{\overline{m}}^\infty\le(1\vee\gamma)\left(D_{\overline{m}}+B_{\overline{m}}^\infty\right),\tag{2.9}$$

and therefore the risk of $\hat{s}_J$ satisfies

$$\begin{aligned} \varepsilon^{-2}\mathbb{E}\left[\|\hat{s}_J-s\|^2\right]=D_J+B_J^\infty &\le \left(1\vee\gamma^{-1}\right)\left(\gamma D_J+B_J^\infty\right)\\ &\le \left(\gamma\vee\gamma^{-1}\right)\left(D_{\overline{m}}+B_{\overline{m}}^\infty\right), \end{aligned}\tag{2.10}$$

which is equivalent to

$$\mathbb{E}\left[\|\hat{s}_J-s\|^2\right]\le\left(\gamma\vee\gamma^{-1}\right)\inf_{m\in\mathcal{M}}\mathbb{E}\left[\|\hat{s}_m-s\|^2\right].$$

One can therefore conclude that the risk of $\hat{s}_J$ remains within a factor $\gamma\vee\gamma^{-1}$ of the optimal risk. Of course, from this point of view, the best value of γ is clearly one. Nevertheless, in order to deal with more general versions of Lepski's estimators, it is interesting to consider general values of γ.

2.4 Some heuristics

Since the definition of J involves the sequence $(B_m^\infty)_{m\in\mathcal{M}}$ which depends on the true unknown parameter s, it cannot be computed but only estimated from the data. Our purpose here is to find an estimator $\hat{J}$ of J with the property that the quadratic risk of the estimator $\tilde{s}=\hat{s}_{\hat{J}}$ is close to the quadratic risk of $\hat{s}_J$. In order to define $\hat{J}$, one first observes that (2.8) gives a characterization of J in terms of the sequence of estimators $(\hat{s}_m)_{m\in\mathcal{M}}$ rather that in terms of the true unknown function s to be estimated. Lepski's method is actually based on this argument. Since J depends on the quantities $\|\hat{s}_j-\hat{s}_m\|^2$ through their expectations and we only have at hand a single realization of these variables, we have to replace these expectations by the random variables themselves and since these variables clearly fluctuate around their expectations and tend to be larger than them with a nonnegligible probability we have to suitably enlarge the bound $(1+\gamma)\varepsilon^2(D_m-D_j)$ in (2.8) in order to derive a sensible estimator $\hat{J}$ of J.

To see what should be added to $(1+\gamma)\varepsilon^2(D_m - D_j)$ in (2.8) we observe that J can equivalently be defined as the smallest index j such that

$$\begin{aligned}\|\hat{s}_j - \hat{s}_m\|^2 \ \leq\ & (1+\gamma)\varepsilon^2(D_m - D_j) \\ & + \left(\|\hat{s}_j - \hat{s}_m\|^2 - \mathbb{E}\left[\|\hat{s}_j - \hat{s}_m\|^2\right]\right), \ \text{ for all } m > j.\end{aligned}$$

Therefore a sensible definition of an estimator $\hat{J}$ of J is obtained by replacing in this formula $\|\hat{s}_j - \hat{s}_m\|^2 - \mathbb{E}\left[\|\hat{s}_j - \hat{s}_m\|^2\right]$ by some quantity which bounds it with a large enough probability. In order to derive such a quantity, we recall from (2.5) that $\varepsilon^{-2}\|\hat{s}_j - \hat{s}_m\|^2$ has a noncentral chi square distribution and appeal to Lemma 8.1 in the Appendix which controls its deviations from the mean. More precisely, it follows from this lemma with $D = D_m - D_J$ and $B = B_J^m \leq \gamma(D_m - D_J)$ that whatever $m > J$ and $x_{m,J} > 0$

$$\begin{aligned}\mathbb{P}\Big[\varepsilon^{-2}\|\hat{s}_J - \hat{s}_m\| \geq (1+\gamma)(D_m - D_J) \\ +2\sqrt{(1+2\gamma)(D_m - D_J)x_{m,J}} + 2x_{m,J}\Big] \leq \exp(-x_{m,J}).\end{aligned}$$

In order to control those deviations for all values of m simultaneously, one should require that the series $\sum_{m>J}\exp(-x_{m,J})$ be summable and suitably small. This suggests the following version of Lepski's estimator.

3 Construction and existence of our estimator

Let us now define our estimator recalling that the framework we use has been described in Section 2.1

Definition 3.1 *Given the increasing sequence $(D_m)_{m\in\mathcal{M}}$ of positive integers, the projection estimators $\{\hat{s}_m\}$ described by (2.3), a family of nonnegative numbers $(\lambda_m)_{m\in\mathcal{M}}$ such that*

$$\sum_{m\in\mathcal{M}} \exp(-\lambda_m) < +\infty \tag{3.11}$$

and a family of numbers $K_{m,j}$ defined for $m \geq 2$, $1 \leq j < m$ and satisfying

$$K_{m,j} \geq [(1+2\gamma)\lambda_m(D_m - D_j)]^{1/2} + \lambda_m, \quad \textit{for some } \gamma > 0, \tag{3.12}$$

we consider the random integer

$$\begin{aligned}\hat{J} = \inf\{j \in \mathcal{M} \mid & \|\hat{s}_j - \hat{s}_m\|^2 \\ & \leq \varepsilon^2[(1+\gamma)(D_m - D_j) + 2K_{m,j}] \textit{ for all } m > j\}.\end{aligned} \tag{3.13}$$

Our estimator is then given by $\tilde{s} = \hat{s}_{\hat{J}}$.

Remark: The convergence assumption (3.11) is quite analogous to the assumption $\sum_{m\in\mathcal{M}}\exp(-L_mD_m) < +\infty$ which appears in Assumption **B** p.70 of Birgé and Massart (1997) and in various places in Barron, Birgé and Massart (1999). Its aim is the same, as shown by the proof of the next proposition, namely to ensure that a large number of deviation inequalities be satisfied simultaneously.

One should observe that $\hat{J}$ is well-defined when $\mathcal{M} = [1; M]\cap\mathbb{N}$ is finite since then the set $\{m > M\}$ is empty and therefore $\hat{J} \le M$ (an empty restriction being always true). If $\mathcal{M}$ is infinite, one has to check that $\hat{J} < +\infty$ a.s. in order that $\tilde{s}$ be well-defined, which follows from the next proposition.

Proposition 3.1 *Under the conditions of Definition 3.1, $\hat{J} < +\infty$ a.s.*

Proof We only have to study the case $\mathcal{M} = \mathbb{N}^\star$. Let us consider the subset $\mathcal{J} = \{J_0; J_1; \ldots\}$ of $\mathcal{M}$ of those indices j which satisfy $B_j^\infty + \gamma D_j \le B_m^\infty + \gamma D_m$ for all $m > j$. By definition,

$$J_0 = \operatorname{argmin}\left\{B_j^m + \gamma(D_m - D_j),\ m \in \mathcal{M}\right\} = J,$$

as defined by (2.7), and for $k \ge 0$,

$$J_{k+1} = \operatorname{argmin}\left\{j \in \mathcal{M},\ j > J_k \;\middle|\; B_j^\infty + \gamma D_j \le B_m^\infty + \gamma D_m \quad \text{for all } m > j\right\}.$$

Moreover, since $D_m \to +\infty$ when $m \to +\infty$, $\mathcal{J}$ is infinite. Let now $j \in \mathcal{J}$ and $m > j$. Then $B_j^m \le \gamma(D_m - D_j)$ and it follows from (3.12) that

$$K_{m,j} \ge \sqrt{\left(D_m - D_j + 2B_j^m\right)\lambda_m} + \lambda_m.$$

Consequently

$$(1+\gamma)(D_m-D_j)+2K_{m,j} \ge D_m-D_j+B_j^m+2\sqrt{(D_m - D_j + 2B_j^m)\lambda_m}+2\lambda_m.$$

Let us now set, for $j \in \mathcal{J}$,

$$F_{j,m} = \left\{\|\hat{s}_j - \hat{s}_m\|^2 > \varepsilon^2[(1+\gamma)(D_m - D_j) + 2K_{m,j}]\right\},$$

and

$$A_j = \bigcap_{m>j} F_{j,m}^c. \tag{3.14}$$

Since U_j^m has the distribution $\chi'^2\left(D_m - D_j, \sqrt{B_j^m}\right)$, it follows from Lemma 8.1 that

$$\mathbb{P}[F_{j,m}] = \mathbb{P}\left[U_j^m > (1+\gamma)(D_m - D_j) + 2K_{m,j}\right] \le e^{-\lambda_m}. \tag{3.15}$$

Then, by (3.13), $\hat{J} \leq j$ on A_j and therefore $\{\hat{J} > j\} \subset A_j^c = \bigcup_{m>j} F_{j,m}$. We conclude from (3.15) that, for any $j \in \mathcal{J}$,

$$\text{(3.16)} \qquad \mathbb{P}\left[\hat{J} > j\right] \leq \mathbb{P}[A_j^c] \leq \sum_{m>j} \mathbb{P}[F_{j,m}] \leq \Sigma_j = \sum_{m>j} e^{-\lambda_m},$$

which implies by (3.11) that $\mathbb{P}\left[\hat{J} > j\right]$ converges to zero when j tends to infinity in $\mathcal{J}$. ∎

4 The performance of our estimator

Let us first set the assumptions we shall need to prove our results, recalling that the numbers λ_m and $K_{m,j}$ have been given in Definition 3.1.

Assumption 4.1

1. *$(D_m)_{m\in\mathcal{M}}$ is a strictly increasing sequence of positive integers such that, if $\mathcal{M}$ is infinite, $\sup_{m\geq 1} D_{m+1}/D_m < +\infty$.*
2. *There exists some integer $p \geq 2$ such that*

$$\text{(4.17)} \qquad \sum_{j\geq 1} D_j \left(\sum_{m>j} e^{-\lambda_m}\right)^{1-1/p} < +\infty.$$

3. *The numbers $K_{m,j}$ satisfy (3.12) and*

$$\text{(4.18)} \qquad \sup_{m\geq 2}\left[D_m^{-1}\left(\sup_{1\leq j<m} K_{m,j}\right)\right] < +\infty.$$

Let us first observe that, apart from the fact that it should not grow faster than exponentially, the sequence $(D_m)_{m\in\mathcal{M}}$ can be fairly arbitrary. In practice, one typically encounters two situations. Either $D_m = u + v(m-1)$ (trigonometric type expansions) or $D_m = u + v^{m-1}$ (wavelet type expansions) for some suitable nonnegative constants u and v.

The numbers $K_{m,j}$ have to satisfy simultaneously (3.12) and (4.18) and it is not at first sight obvious to choose the λ_m's in such a way that this is possible when $\mathcal{M}$ is infinite. The following proposition gives some hints for a proper choice of the parameters involved in our construction.

Proposition 4.1 *Assume that $\alpha > 3$, $m_0 \geq 1$ and*

$$\text{(4.19)} \qquad \lambda_m \geq \alpha \log D_m \quad \textit{for } m \geq m_0,$$

then (4.17) holds.

Proof Recalling from (3.16) that $\Sigma_j = \sum_{m>j}\exp(-\lambda_m)$, we consider some integer $p \geq 2$ such that $\alpha > 3 + 2/(p-1)$. We want to prove that $\sum_{j\geq 1} D_j \Sigma_j^{1-1/p} < +\infty$. By (4.19) and the convexity of the function $x \mapsto x^{-\alpha}$ one gets for $j \geq m_0$

$$\Sigma_j = \sum_{m>j} e^{-\lambda_j} \leq \sum_{m>j} D_m^{-\alpha} \leq \int_{D_j+1/2}^{+\infty} x^{-\alpha}\,dx = \frac{1}{\alpha-1}\left(D_j + \frac{1}{2}\right)^{-\alpha+1}$$

and it follows that $D_j\Sigma_j^{1-\frac{1}{p}} < D_j^{-\frac{(p-1)(\alpha-1)-p}{p}}$ for $j \geq m_0$. Since $\alpha > 3 + 2/(p-1)$, the series $\sum_{j\geq 1} D_j\Sigma_j^{1-1/p}$ converges and (4.17) is satisfied. ■

Let us observe that (4.19) is in particular compatible with a choice of numbers $K_{m,j}$ satisfying for some positive constants $A \geq a > 0$,

$$aD_m \leq K_{m,j} \leq AD_m \quad \text{for all } m \geq 2,\ 0 < j < m, \tag{4.20}$$

which ensures that (4.18) holds. In particular, the original method of Lepski is based on the choice

$$\hat{J} = \inf\left\{j \in \mathcal{M} \mid \|\hat{s}_j - \hat{s}_m\|^2 \leq K\varepsilon^2 D_m \text{ for all } m > j\right\}$$

with a suitably large constant K. Choosing $K > 1$ and $0 < \gamma < K - 1$ leads then to

$$(K-1-\gamma)D_m < 2K_{m,j} = (K-1-\gamma)D_m + (1+\gamma)D_j < KD_m,$$

which is (4.20). Such a choice is therefore compatible with (3.12) and (4.19) for suitable values of the parameters λ_m. In particular, the classical Lepski's method with a choice of $K > 1$ satisfies our assumptions. This is not true anymore if $K \leq 1$ and one could prove, in the same way that we proved lower bounds for the penalty term in Birgé and Massart (1999), that $K < 1$ could lead to inconsistent estimators when ε converges to zero. One shall not insist on this here. On the other hand, if $K > 1$, the following theorem applies.

Theorem 4.1 *Under the above assumptions, there exists some constant C depending only on the various parameters involved in the construction of the estimator, but neither on ε, nor on s and such that*

$$\mathbb{E}\left[\|\tilde{s} - s\|^2\right] \leq C \inf_{m\in\mathcal{M}} \mathbb{E}\left[\|\hat{s}_m - s\|^2\right]. \tag{4.21}$$

If we fix the values of the various parameters involved in the construction of our estimator, C can then be taken as a universal constant. For instance, the particular choice of $\lambda_m = 4\log(D_m+1)$, $p = 4$, $\gamma = 1$ and

$K_{m,j} = [3\lambda_m D_m]^{1/2} + \lambda_m$ together with the assumption that $D_{m+1} \leq 2D_m$ for all m satisfies our requirements and, although this particular choice of the parameters has nothing special, it can cope with almost all practical situations.

5 Proof of Theorem 4.1

For the sake of simplicity we shall prove it below only under the assumption that $M = \mathbb{N}^\star$. Only minor modifications in Section 5.5 below are needed to handle the finite case.

5.1 Basic inequality

It follows from (2.4) and the monotonicity of the sequence B_m^∞ that

$$\begin{aligned}
\varepsilon^{-2}\|\tilde{s} - s\|^2 &= \left(V_0^{\hat{J}} + B_{\hat{J}}^\infty\right) \\
&= \left(V_0^{\hat{J}} + B_{\hat{J}}^\infty\right) \mathbb{1}_{\{\hat{J} \leq J\}} + \left(V_0^{\hat{J}} + B_{\hat{J}}^\infty\right) \mathbb{1}_{\{\hat{J} > J\}} \\
&\leq \left(D_{\hat{J}} + B_{\hat{J}}^J + B_J^\infty\right) \mathbb{1}_{\{\hat{J} \leq J\}} + \left(V_0^{\hat{J}} - D_{\hat{J}}\right) \mathbb{1}_{\{\hat{J} \leq J\}} \\
&\quad + \left(V_0^J + V_J^{\hat{J}} + B_J^\infty\right) \mathbb{1}_{\{\hat{J} > J\}} \\
&\leq B_J^\infty + (1 \vee \gamma^{-1}) \left(B_{\hat{J}}^J + \gamma D_{\hat{J}}\right) \mathbb{1}_{\{\hat{J} \leq J\}} \\
&\quad + \left(V_0^{\hat{J}} - D_{\hat{J}}\right) \mathbb{1}_{\{\hat{J} \leq J\}} + V_0^J \mathbb{1}_{\{\hat{J} > J\}} + V_J^{\hat{J}} \mathbb{1}_{\{\hat{J} > J\}},
\end{aligned}$$

and therefore after integration

$$\begin{aligned}
(5.22)\quad \varepsilon^{-2}\mathbb{E}\left[\|\tilde{s} - s\|^2\right] &\leq B_J^\infty + \mathbb{E}\left[V_0^J \mathbb{1}_{\{\hat{J} > J\}}\right] \\
&\quad + (1 \vee \gamma^{-1})\,\mathbb{E}\left[\left(B_{\hat{J}}^J + \gamma D_{\hat{J}}\right) \mathbb{1}_{\{\hat{J} \leq J\}}\right] \\
&\quad + \mathbb{E}\left[\left|V_0^{\hat{J}} - D_{\hat{J}}\right| \mathbb{1}_{\{\hat{J} \leq J\}}\right] + \mathbb{E}\left[V_J^{\hat{J}} \mathbb{1}_{\{\hat{J} > J\}}\right].
\end{aligned}$$

We shall now bound successively each of the four expectations in the right-hand side of (5.23).

5.2 Control of the first expectation

Recalling that the set A_j is defined by (3.14), we see that it only depends on the random variables U_j^m for $m > j$ and therefore on the variables $\hat{\beta}_m$ for $m > j$ which implies that A_j is independent of V_0^j and therefore by (3.16)

$$\mathbb{E}\left[V_0^J \mathbb{1}_{\{\hat{J} > J\}}\right] \leq \mathbb{E}\left[V_0^J \mathbb{1}_{A_J^c}\right] = \mathbb{E}\left[V_0^J\right] \mathbb{P}\left[A_J^c\right] \leq D_J \Sigma_J.$$

5.3 Control of the second expectation

We want to bound $\mathbb{E}\left[\left(B_{\hat{J}}^J + \gamma D_{\hat{J}}\right) 1\!\!1_{\{\hat{J} \le J\}}\right]$. We first notice that if $J = 1$ and $\hat{J} \le J$, then $\hat{J} = J$ and a trivial bound is γD_J. Assuming now that $J \ge 2$, we define for $t > 0$

$$E_t = \bigcap_{0<m<J} \left\{ U_m^J > D_J - D_m + B_m^J - 2\sqrt{t\,(D_J - D_m + 2B_m^J)} \right\}.$$

Recalling that U_m^J has a distribution $\chi'^2\left(D_J - D_m, \sqrt{B_m^J}\right)$, we derive from (8.35) of Lemma 8.1 that

$$\begin{aligned} \mathbb{P}[E_t^c] &\le \sum_{0<m<J} \mathbb{P}\left[U_m^J \le D_J - D_m + B_m^J - 2\sqrt{t\,[D_J - D_m + 2B_m^J]}\right] \\ &\le (J-1)e^{-t}. \qquad (5.23) \end{aligned}$$

Since $U_{\hat{J}}^J \le (1+\gamma)(D_J - D_{\hat{J}}) + 2K_{J,\hat{J}}$ for $\hat{J} < J$ by the definition of $\hat{J}$ we derive that, on the set $E_t \cap \{\hat{J} < J\}$,

$$\gamma(D_J - D_{\hat{J}}) + 2K_{J,\hat{J}} \ge B_{\hat{J}}^J - 2\sqrt{2t\left[(D_J - D_{\hat{J}})/2 + B_{\hat{J}}^J\right]}.$$

Using the fact that $-2\sqrt{xy} = \left(\sqrt{x} - \sqrt{y}\right)^2 - x - y$ we then derive that, on the set $E_t \cap \{\hat{J} < J\}$, one has

$$\left(\sqrt{\left[(D_J - D_{\hat{J}})/2 + B_{\hat{J}}^J\right]} - \sqrt{2t} \right)^2 \le (\gamma + 1/2)(D_J - D_{\hat{J}}) + 2K_{J,\hat{J}} + 2t,$$

and therefore

$$\sqrt{\left[(D_J - D_{\hat{J}})/2 + B_{\hat{J}}^J\right]} \le \sqrt{2t} + \sqrt{(\gamma + 1/2)(D_J - D_{\hat{J}}) + 2K_{J,\hat{J}} + 2t}.$$

Squaring everything and using $\sqrt{x+y} \le \sqrt{x} + \sqrt{y}$ finally gives

$$B_{\hat{J}}^J \le \gamma(D_J - D_{\hat{J}}) + 2K_{J,\hat{J}} + 8t + \left[t(2\gamma+1)(D_J - D_{\hat{J}}) + 4tK_{J,\hat{J}}\right]^{1/2}.$$

Let us now set

$$a_J = D_J^{-1} \sup_{m<J} K_{J,m}. \qquad (5.24)$$

It then follows that $K_{J,\hat{J}} \le a_J D_J$. Therefore, if $t \ge 1$,

$$B_{\hat{J}}^J + \gamma D_{\hat{J}} \le (\gamma + 2a_J) D_J + \Delta t \quad \text{with } \Delta = \left(8 + \sqrt{(2\gamma + 1 + 4a_J)D_J}\right) \qquad (5.25)$$

on the set $E_t \cap \{\hat{J} \leq J\}$ since this inequality clearly also holds if $\hat{J} = J$. One then derives from (5.23) that

$$\mathbb{P}\left[\left(B_{\hat{J}}^{J} + \gamma D_{\hat{J}}\right) \mathbb{1}_{\{\hat{J} \leq J\}} > (\gamma + 2a_J)D_J + \Delta t\right] \leq (J-1)e^{-t} \quad \text{for } t \geq 1.$$

Integration with respect to t finally leads, for $J \geq 2$, to

$$\mathbb{E}\left[\left(B_{\hat{J}}^{J} + \gamma D_{\hat{J}}\right) \mathbb{1}_{\{\hat{J} \leq J\}}\right] \leq (\gamma + 2a_J)D_J + [1 + (1 \vee \log(J-1))]\Delta,$$

which clearly remains true when $J = 1$. Therefore whatever $J \geq 1$,

$$\begin{aligned}\mathbb{E}\left[\left(B_{\hat{J}}^{J} + \gamma D_{\hat{J}}\right) \mathbb{1}_{\{\hat{J} \leq J\}}\right] &\leq (\gamma + 2a_J)D_J \\ &\quad + \left(8 + \sqrt{(2\gamma + 1 + 4a_J)D_J}\right) \log(3J+5).\end{aligned}$$

5.4 Control of the third expectation

Since V_0^m has a χ^2 distribution with D_m degrees of freedom, one can use Lemmas 8.3 and 8.4 to bound $\mathbb{E}\left[\left|V_0^{\hat{J}} - D_{\hat{J}}\right| \mathbb{1}_{\{\hat{J} \leq J\}}\right]$ in the following way:

$$\begin{aligned}\left(\mathbb{E}\left[\left|V_0^{\hat{J}} - D_{\hat{J}}\right| \mathbb{1}_{\{\hat{J} \leq J\}}\right]\right)^4 &\leq \sum_{m=1}^{J} \mathbb{E}\left[(V_0^m - D_m)^4\right] \\ &\leq 12 \sum_{m=1}^{J} \left(2D_m^2 + 3D_m\right) \\ &\leq 12 \sum_{i=1}^{D_J} \left(2i^2 + 3i\right) \\ &= 12D_J(D_J+1)\left[\frac{2D_J+1}{3} + \frac{3}{2}\right] \\ &< 8D_J(D_J+1)(D_J+3).\end{aligned}$$

5.5 Control of the fourth expectation

In order to control $\mathbb{E}\left[V_J^{\hat{J}} \mathbb{1}_{\{\hat{J} > J\}}\right]$ we introduce an increasing sequence $(I_k)_{k \geq 0}$ of elements of $\mathcal{J}$, starting with $I_0 = J$, to be defined below. One can therefore write, using the monotonicity of the sequence $(V_0^m)_{m>0}$ that

$$\begin{aligned}V_J^{\hat{J}} \mathbb{1}_{\{\hat{J} > J\}} &\leq \sum_{k \geq 0} V_J^{I_{k+1}} \mathbb{1}_{\{I_k < \hat{J} \leq I_{k+1}\}} = \sum_{k \geq 0} V_J^{I_{k+1}} \left(\mathbb{1}_{\{\hat{J} > I_k\}} - \mathbb{1}_{\{\hat{J} > I_{k+1}\}}\right) \\ &= \sum_{k \geq 0} \mathbb{1}_{\{\hat{J} > I_k\}} \left(V_J^{I_{k+1}} - V_J^{I_k}\right) = \sum_{k \geq 0} V_{I_k}^{I_{k+1}} \mathbb{1}_{\{\hat{J} > I_k\}}.\end{aligned}$$

Since V_j^k has a $\chi^2(D_k-D_j)$ distribution and $p \geq 2$, it follows from Lemma 8.4 that $\mathbb{E}\left[\left(V_j^k\right)^p\right] < (D_k - D_j + p - 1)^p$ for $k > j$ and we then derive from the Hölder Inequality and (3.16) that

$$\begin{aligned}
\mathbb{E}\left[V_J^{\hat{J}}\mathbb{1}_{\{\hat{J}>J\}}\right] &\leq \sum_{k\geq 0}\left(\mathbb{E}\left[\left(V_{I_k}^{I_{k+1}}\right)^p\right]\right)^{1/p}\left(\mathbb{P}\left[\hat{J} > I_k\right]\right)^{1-1/p} \\
(5.26) \qquad &\leq \sum_{k\geq 0}(D_{I_{k+1}} - D_{I_k} + p - 1)\Sigma_{I_k}^{1-1/p}.
\end{aligned}$$

We now have to specify how we choose I_k for $k > 0$. Let us introduce $K = \inf\left\{j \,\middle|\, B_j^\infty < \gamma(D_j - p + 2)\right\}$ which exists since D_m tends to infinity with m and define

$$I_1 = \operatorname{argmin}\left\{B_j^\infty + \gamma D_j \,|\, j \geq (J+1) \vee K\right\};$$

$$I_{k+1} = \operatorname{argmin}\left\{B_j^\infty + \gamma D_j \,|\, j \geq I_k + 1\right\} \quad \text{for } k \geq 1.$$

Now, if $I_k \geq K - 1$, according to the definitions of I_{k+1} and K,

$$B_{I_{k+1}}^\infty + \gamma D_{I_{k+1}} \leq B_{I_k+1}^\infty + \gamma D_{I_k+1} < \gamma(2D_{I_k+1} - p + 2),$$

and therefore

$$(5.27) \quad D_{I_{k+1}} + p - 1 \leq 2D_{I_k+1} \leq 2\delta_J D_{I_k} \quad \text{with } \delta_J = \sup_{m\geq J} D_{m+1}/D_m.$$

This inequality holds for all $k \geq 0$ if $I_0 = J \geq K - 1$. Otherwise $K - 1 > J > 0$ and (5.27) only holds for $k \geq 1$. Nevertheless, the same arguments then show that $B_{I_1}^\infty + \gamma D_{I_1} < \gamma(2D_K - p + 2)$. Therefore, using the definition of K one gets

$$\begin{aligned}
D_{I_1} + p - 1 &\leq 2D_K \leq 2\delta_J D_{K-1} \leq 2\delta_J\left(\gamma^{-1}B_{K-1}^\infty + p - 2\right) \\
&\leq 2\delta_J\left(\gamma^{-1}B_J^\infty + p - 2\right).
\end{aligned}$$

Therefore in both cases $D_{I_1} + p - 1 \leq 2\delta_J\left[\left(\gamma^{-1}B_J^\infty + p - 2\right) \vee D_J\right]$. It then follows from (5.26) that

$$\begin{aligned}
\mathbb{E}\left[V_J^{\hat{J}}\mathbb{1}_{\{\hat{J}>J\}}\right] &\leq 2\delta_J\left[\left(\gamma^{-1}B_J^\infty + p - 2\right) \vee D_J - (D_J/2)\right]\Sigma_J^{1-1/p} \\
&\quad + (2\delta_J - 1)\sum_{k\geq 1} D_{I_k}\Sigma_{I_k}^{1-1/p} \\
&\leq 2\delta_J\left[\left(\gamma^{-1}B_J^\infty + p - 2\right) \vee D_J - (D_J/2)\right]\Sigma_J^{1-1/p} \\
&\quad + (2\delta_J - 1)\sum_{m>J} D_m\Sigma_m^{1-1/p}.
\end{aligned}$$

5.6 Completing the proof

Putting all four bounds together leads to

$$\begin{aligned}\varepsilon^{-2}\mathbb{E}\left[\|\tilde{s}-s\|^2\right] \;\le\;& B_J^\infty + D_J\Sigma_J + \left(1\vee\gamma^{-1}\right)(\gamma+2a_J)D_J\\ &+\left(1\vee\gamma^{-1}\right)\left(8+\sqrt{(2\gamma+1+4a_J)D_J}\right)\log(3J+5)\\ &+\left[8D_J(D_J+1)(D_J+3)\right]^{1/4}\\ &+2\delta_J\left[\left(\gamma^{-1}B_J^\infty+p-2\right)\vee D_J-\frac{D_J}{2}\right]\Sigma_J^{1-1/p}\\ &+(2\delta_J-1)\sum_{m>J}D_m\Sigma_m^{1-1/p},\end{aligned}$$

which can also be written, since $D_J \ge 1$ as

$$\mathbb{E}\left[\|\tilde{s}-s\|^2\right] \le C_J\varepsilon^2\left(B_J^\infty+\gamma D_J\right), \tag{5.28}$$

with

$$\begin{aligned}C_J \;=\;& \left(1\vee\gamma^{-1}\right)\left(1+2\gamma^{-1}a_J\right)+\gamma^{-1}\Sigma_J\\ &+\gamma^{-1}\left(1\vee\gamma^{-1}\right)\left(8+\sqrt{2\gamma+1+4a_J}\right)D_J^{-1/2}\log(3J+5)\\ &+\gamma^{-1}\left[3D_J^{-1/4}+(2\delta_J-1)D_J^{-1}\sum_{m>J}D_m\Sigma_m^{1-1/p}\right]\\ &+2\delta_J\gamma^{-1}(p-1)\,\Sigma_J^{1-1/p}.\end{aligned} \tag{5.29}$$

On the other hand, it follows from (2.9) that

$$\varepsilon^2\left(B_J^\infty+\gamma D_J\right) \le (1\vee\gamma)\,\varepsilon^2\left(D_{\overline{m}}+B_{\overline{m}}^\infty\right) = (1\vee\gamma)\mathbb{E}\left[\|\hat{s}_{\overline{m}}-s\|^2\right],$$

and finally from the definition of $\overline{m}$ that

$$\mathbb{E}\left[\|\tilde{s}-s\|^2\right] \le C_J(1\vee\gamma)\inf_{m\in\mathcal{M}}\mathbb{E}\left[\|\hat{s}_m-s\|^2\right]. \tag{5.30}$$

The constant C_J depends on the various parameters involved in the construction of the estimator and on ε and s only through the parameter J. Moreover, it is bounded independently of J since by assumption, the sequences $(a_j)_{j\ge1}$, $(\Sigma_j)_{j\ge1}$ and $(\delta_j)_{j\ge1}$ are bounded and $\sum_{m>0}D_m\Sigma_m^{1-1/p} < +\infty$. This completes the proof of Theorem 4.1.

6 Asymptotic optimality of a modified Lepski's method

Our purpose is now to understand what is going on when J goes to infinity. Since $D_J \ge J$ and $\sum_{m>0}D_m\Sigma_m^{1-1/p} < +\infty$, all the terms in C_J then converge to zero, except for the first one which involves a_J, defined by (5.24). The assumption (4.18) only implies that a_J is bounded but one could enforce it to

$$\limsup_m D_m^{-1}\left(\sup_{1\leq j<m} K_{m,j}\right) = 0. \tag{6.31}$$

Let us observe that such a requirement is perfectly feasible. The choice $\lambda_m = \mathcal{O}(D_m/\log m)$ when $m \to +\infty$ is clearly compatible with (4.19) and one can choose the numbers $K_{m,j}$ in such a way that they satisfy (3.12) together with $\sup_{1\leq j<m} K_{m,j} = \mathcal{O}(D_m/\log m)$ which implies (6.31). It is now easy to prove the following corollary.

Corollary 6.1 *Let us assume that $\mathcal{M}$ is infinite as well as the set $\{m \in \mathcal{M} \mid \int s\varphi_m \neq 0\}$. Choose the parameters λ_m and $K_{m,j}$ in order to satisfy the assumptions of Section 4 with (4.18) replaced by (6.31) and define the estimator $\tilde{s}$ as before. Then*

$$\limsup_{\varepsilon\to 0} \frac{\mathbb{E}\left[\|\tilde{s}-s\|^2\right]}{\inf_{m\in\mathcal{M}}\mathbb{E}\left[\|\hat{s}_m-s\|^2\right]} \leq \gamma \vee \gamma^{-1}. \tag{6.32}$$

Proof We already noticed that all the terms in C_J, as given by (5.29) converge to zero when J goes to infinity, except for the first one, which tends to $1 \vee \gamma^{-1}$ since $a_J \to 0$ by (6.31). We then remark that our assumption on s implies that $\varepsilon^2 B_m^\infty$ is bounded away from 0 independently of ε whatever $m \in \mathcal{M}$. Then $\mathbb{E}\left[\|\hat{s}_m - s\|^2\right] = \varepsilon^2(B_m^\infty + D_m)$ remains bounded away from 0 for fixed m when $\varepsilon \to 0$ while it can be made arbitrarily small provided that both ε and m are suitably chosen. This implies that $J \to +\infty$ when $\varepsilon \to 0$ and therefore $C_J \to 1 \vee \gamma^{-1}$ when $\varepsilon \to 0$. The conclusion then follows from (5.30). ■

One should observe here that (6.31) rules out the initial choice of Lepski for the parameters $K_{m,j}$ which implies that (4.20) holds.

7 Conclusion

In the framework we have chosen here, an older and very popular method for choosing an optimal estimator in our family is Mallows' C_p which actually amounts to choose $\tilde{s} = \hat{s}_{\hat{J}}$ with

$$\begin{aligned} \hat{J} &= \operatorname{argmin}\left\{-\|\hat{s}_j\|^2 + 2\varepsilon^2 D_j,\, j \in \mathcal{M}\right\} \\ &= \inf\left\{j \in \mathcal{M} \;\middle|\; \|\hat{s}_m - \hat{s}_j\|^2 \leq 2\varepsilon^2(D_m - D_j) \text{ for all } m > j\right\}, \end{aligned} \tag{7.33}$$

since $\|\hat{s}_m - \hat{s}_j\|^2 = \|\hat{s}_m\|^2 - \|\hat{s}_j\|^2$ for $m > j$. One should observe that it is also the estimator derived from our extension of Lepski's method with

$0 < \gamma < 1$ and $2K_{m,j} = (1-\gamma)(D_m - D_j)$. Unfortunately, such a choice of $K_{m,j}$ does not always satisfy (3.12) when $j = m-1$ and m is large since λ_m goes to infinity with m while $D_m - D_{m-1}$ may remain bounded. Nevertheless (3.12) will be satisfied with $\lambda_m = \alpha \log D_m$, as in Proposition 4.1 provided that $D_m \geq D_{m-1} + c \log D_m$ for some large enough c. In any case, it has been proved in Shibata (1981), that the estimator $\tilde{s} = \hat{s}_{\hat{J}}$ with $\hat{J}$ given by (7.33) satisfies (6.32) with $\gamma = 1$ and by Birgé and Massart (1999) that it also satisfies (4.21).

As to the consequences of Theorem 4.1, they have been developed at length in Birgé and Massart (1999) where an analogue of this result has been proved for penalized estimators. We therefore refer the interested reader to this paper for applications of this result, just mentioning here the following one. Assume that $\mathcal{M} = \mathbb{N}^\star$ and that $D_m = m$, which implies that S_m is the linear span of $\{\varphi_1, \ldots, \varphi_m\}$. Given a nonincreasing sequence $a = (a_m)_{m \geq 1}$ of numbers in $[0, +\infty]$ such that $a_1 > 0$ and $a_m \to 0$ when $m \to +\infty$, we denote by $\mathcal{E}(a)$ the ellipsoid defined by

$$\mathcal{E}(a) = \left\{ s = \sum_{j=1}^{+\infty} \beta_j \varphi_j \,\middle|\, \sum_{j=1}^{+\infty} \frac{\beta_j^2}{a_j^2} \leq 1 \right\},$$

with the convention that $0/0 = 0$, $x/0 = +\infty$ and $x/(+\infty) = 0$ for $x > 0$. Let $\tilde{s}$ be any estimator satisfying (4.21), then it follows from Section 7.2 of Birgé and Massart (1999) that $\tilde{s}$ is minimax, up to constants, over all such ellipsoids. More precisely, there exists some constant κ such that, whatever the sequence a satisfying the above requirements,

$$\sup_{s \in \mathcal{E}(a)} \mathbb{E}\left[\|\tilde{s} - s\|^2\right] \leq \kappa \left[1 \vee (\varepsilon/a_1)^2\right] \inf_{\hat{s}} \sup_{s \in \mathcal{E}(a)} \mathbb{E}\left[\|\hat{s} - s\|^2\right],$$

where the infimum is taken over all possible estimators.

8 Appendix

The following lemma is a generalization of Lemma 1 of Laurent and Massart (1998).

Lemma 8.1 *Let X be a noncentral χ^2 variable with D degrees of freedom and noncentrality parameter $B^{1/2} \geq 0$, then for all $x > 0$,*

$$\mathbb{P}\left[X \geq (D+B) + 2\sqrt{(D+2B)x} + 2x\right] \leq \exp(-x), \tag{8.34}$$

and

$$\mathbb{P}\left[X \leq (D+B) - 2\sqrt{(D+2B)x}\right] \leq \exp(-x). \tag{8.35}$$

Proof Since we can write X as $\left(B^{1/2}+U\right)^2+V$ where U and V are independent with respective distributions $\mathcal{N}(0,1)$ and $\chi^2(D-1)$, the Laplace transform of X can be written as

$$\mathbb{E}\left[e^{tX}\right]=(1-2t)^{-D/2}\exp\left[\frac{Bt}{1-2t}\right]\quad \text{for } t<1/2,$$

which implies that

$$\begin{aligned}
&\log\left(\mathbb{E}\left[e^{t(X-D-B)}\right]\right)\\
&\quad = -D\left(\frac{1}{2}\log(1-2t)+t\right)+\frac{2t^2B}{1-2t}\\
&\quad \le (D+2B)\left[\left(-\frac{1}{2}\log(1-2t)-t\right)\vee\frac{t^2}{1-2t}\right]\\
(8.36)\qquad &\quad = (D+2B)\left[\frac{t^2\mathbb{1}_{\{1/2>t>0\}}}{1-2t}-\frac{\mathbb{1}_{\{t<0\}}[\log(1-2t)+2t]}{2}\right].
\end{aligned}$$

Then (8.34) follows from Lemma 8.2 below. On the other hand, for $z>0$,

$$\mathbb{P}[X\le D+B-z]\le\inf_{t<0}\mathbb{E}\left[e^{t(X-D-B+z)}\right],$$

and it follows from (8.36) that, for $t<0$

$$\log\left(\mathbb{E}\left[e^{t(X-D-B+z)}\right]\right)\le-\frac{[D+2B][\log(1-2t)+2t]}{2}+tz.$$

Setting $z=y(2B+D)$ with $0<y<1$, one observes that the minimum of the right-hand side is obtained for $2t=-y/(1-y)$ and therefore

$$\log(\mathbb{P}[X\le D+B-z])\le\frac{1}{2}[D+2B][\log(1-y)+y]\le-\frac{1}{4}[D+2B]y^2.$$

The result remains clearly true for $y\ge 1$ since $X\ge 0$ and (8.35) follows by setting $y=2x^{1/2}(2B+D)^{-1/2}$. ∎

Lemma 8.2 *Let X be a random variable such that*

$$\log\left(\mathbb{E}[\exp(tX)]\right)\le\frac{(at)^2}{1-bt}\quad \textit{for } 0<t<b^{-1},$$

where a and b are positive constants. Then

$$\mathbb{P}[X\ge 2a\sqrt{x}+bx]\le\exp(-x)\quad \textit{for all } x>0.$$

Proof For $z > 0$, $\mathbb{P}[X \geq z] \leq \exp[-h(z)]$ with

$$h(z) = \sup_{0<t<b^{-1}} \left\{ zt - \frac{(at)^2}{1-bt} \right\}$$

and the supremum is achieved for $t' = b^{-1}[1 - a(bz + a^2)^{-1/2}]$. Taking $z = bx + 2a\sqrt{x}$ gives $t' = \sqrt{x}/(a + b\sqrt{x})$ and

$$h\left(bx + 2a\sqrt{x}\right) = \frac{bx^{3/2} + 2ax}{a + b\sqrt{x}} - \frac{ax}{a + b\sqrt{x}} = x,$$

which allows to conclude. ■

Lemma 8.3 *Let $\mathcal{M}$ be a finite or countable set of indices, $\{X_m\}_{m\in\mathcal{M}}$ a set of nonnegative random variables indexed by $\mathcal{M}$, $\hat{m}$ a random variable with values in $\mathcal{M}$ and M some subset of $\mathcal{M}$, then, whatever $q > 1$,*

$$\mathbb{E}\left[X_{\hat{m}}\mathbb{1}_{\{\hat{m}\in M\}}\right] \leq \left(\sum_{m\in M} \mathbb{E}[X_m^q]\right)^{1/q} \mathbb{P}[\hat{m} \in M]^{1-1/q}.$$

Proof It follows from Hölder's Inequality that

$$\begin{aligned} \mathbb{E}\left[X_{\hat{m}}\mathbb{1}_{\{\hat{m}\in M\}}\right] &= \sum_{m\in M} \mathbb{E}[X_m\mathbb{1}_{\{\hat{m}=m\}}] \\ &\leq \left(\sum_{m\in M} \mathbb{E}[X_m^q]\right)^{1/q} \left(\sum_{m\in M} \mathbb{E}[\mathbb{1}_{\{\hat{m}=m\}}]\right)^{1-1/q}, \end{aligned}$$

which is the desired inequality. ■

Lemma 8.4 *If Y has a $\chi^2(n)$ distribution, then $\mathbb{E}\left[(Y-n)^4\right] = 12n(2n+3)$ and*

$$\mathbb{E}\left[Y^k\right] \leq (n+k-1)^k - 1 \quad \textit{for } k \in \mathbb{N},\ k \geq 2.$$

Proof It is well-known that $\mathbb{E}\left[Y^k\right] = \prod_{i=0}^{k-1}(n+2i)$. The first result then follows from elementary computations. As to the second, one derives from the strict concavity of the logarithm that

$$k^{-1}\log\left(\mathbb{E}\left[Y^k\right]\right) < \log\left(k^{-1}\sum_{i=0}^{k-1}(n+2i)\right) = \log(n+k-1)$$

and the conclusion follows since $\mathbb{E}\left[Y^k\right]$ is an integer. ■

REFERENCES

Barron, A.R., Birgé, L. and Massart, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Relat. Fields* **113**, 301-415.

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

Birgé, L. and Massart, P. (1999). Gaussian model selection. Technical Report.

Butucea, C. (1999). Doctoral Thesis. University Paris VI.

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

Donoho, D.L. and Johnstone, I.M. (1994b). Minimax risk over l_p-balls for l_q-error. *Probab. Theory Relat. Fields* **99**, 277-303.

Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *JASA* **90**, 1200-1224.

Donoho, D.L. and Johnstone, I.M. (1996). Neo-classical minimax problems, thresholding and adaptive function estimation. *Bernoulli* **2**, 39-62.

Efromovich, S.Yu. (1998). On global and pointwise adaptive estimation. *Bernoulli* **4**, 273-282.

Efroimovich, S.Yu. and Low, M.G. (1994). Adaptive estimates of linear functionals. *Probab. Theory Relat. Fields* **98**, 261-275.

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

Ibragimov, I.A. and Has'minskii, R.Z. (1981). *Statistical Estimation - Asymptotic Theory.* Springer-Verlag, New York.

Juditsky, A. (1997) Wavelet estimators: adapting to unknown smoothness. *Math. Methods of Statist.* **6**, 1-25.

Kotel'nikov, V. (1959). *The Theory of Optimum Noise Immunity.* McGraw Hill, New York.

Laurent, B. and Massart, P. (1998) Adaptive estimation of a quadratic functional by model selection. Technical report 98.81. Université Paris-Sud, Orsay.

Lepskii, O.V. (1990). On a problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35**, 454-466.

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

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

Lepskii, O.V. (1992b). On problems of adaptive estimation in white Gaussian noise. *Adv. in Soviet Math.* **12** , 87-106.

Lepski, O.V. (1998). Lectures given for the Paris-Berlin Seminar at Garchy.

Lepski, O.V. and Levit, B.Y. (1998). Adaptative minimax estimation of infinitely differentiable functions. *Math. Methods Statist.* **7**, 123-156.

Lepski, O.V., Mammen, E. and Spokoiny, V.G. (1997). Optimal spatial adaptation to inhomogeneous smoothness: An approach based on kernel estimates with variable bandwidth selectors. *Ann. Statist.* **25**, 929-947.

Lepski, O.V. and Spokoiny, V.G. (1995). Local adaptation to inhomogeneous smoothness: resolution level. *Math. Methods Statist.* **4**, 239-258.

Pinsker, M.S. (1980). Optimal filtration of square-integrable signals in Gaussian noise. *Problems of Information Transmission* **16**, 120-133.

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

Tsybakov, A.B. (1998). Pointwise and sup-norm sharp adaptive estimation of functions on Sobolev classes. *Ann. Statist.* **26**, 2420-2469.

Lucien Birgé
Laboratoire de Probabilités et Modèles Stochastiques
Boîte 188
Université Paris VI, 4 Place Jussieu
F-75252 Paris Cedex 05
FRANCE
lb@ccr.jussieu.fr