The Annals of Statistics
2002, Vol. 30, No. 2, 325–396

RANDOM RATES IN ANISOTROPIC REGRESSION

BY M. HOFFMANN AND O. LEPSKI

Université Paris VII and Université de Provence

In the context of minimax theory, we propose a new kind of risk, normalized by a random variable, measurable with respect to the data. We present a notion of optimality and a method to construct optimal procedures accordingly. We apply this general setup to the problem of selecting significant variables in Gaussian white noise. In particular, we show that our method essentially improves the *accuracy of estimation*, in the sense of giving explicit improved confidence sets in L_2-norm. Links to adaptive estimation are discussed.

1. Introduction. Searching for significant variables is certainly one of the oldest and most popular problems in statistics. One of the simplest models where the issue of selecting significant variables was first stated mathematically is *linear regression*. A vast literature has been devoted to this topic since and different approaches have been proposed over the last forty years, both for estimation and for hypothesis testing. Among many authors, we refer to Akaike [1], Breiman and Freedman [3], Chernoff [5], Csiszar and Korner [6], Dychakov [10], Patel [42], Renyi [46], Freidlina [13], Meshalkin [35], Malyutov and Tsitovich [34], Schwarz [47] and Stone [48].

In classical parametric regression, if we consider a linear model, we first have to measure the possible gain of "searching for a limited number of significant variables." If the model comes from a specific field of application, then only an adequate description together with its solution is relevant. However, from a mathematical point of view, a theory of selecting significant variables does not lead—at least asymptotically—to a substantial improvement of the accuracy of estimation: in a regular parametric model, the classical $\sqrt{n}$ rate of convergence is not affected by the number of significant variables. (However, even in this setup, let us emphasize that "asymptotically" has to be understood as "up to a constant" and that the correct choice of significant variables may possibly improve this constant.)

If instead of a linear model we consider a nonparametric regression model, the search for significant variables becomes crucial for estimating the regression function: the rate of convergence explicitly depends on the set of significant variables.

Let us develop this statement with the following example of multivariate regression: suppose we observe $Z^{(n)} = (X_i, Y_i,\ i = 1, \ldots, n)$ in the model

$$Y_i = f(X_i) + \varepsilon_i, \qquad i = 1, \ldots, n, \tag{1}$$

Received July 1998; revised September 2001.
AMS 2000 subject classifications. 62G07, 62G10, 62G15.
Key words and phrases. Nonparametric estimation, minimax theory, random normalizing factors, anisotropic regression.

where f is a real-valued function defined on the unit cube $[0,1]^d$, the X_i are independent design points uniformly distributed in $[0,1]^d$ and the ε_i are uncorrelated zero-mean noise variables. We want to recover the signal f which belongs to the anisotropic class $\Sigma = \Sigma(\boldsymbol{\beta}, L)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$, $L > 0$, defined by

$$\Sigma(\boldsymbol{\beta}, L) = \left\{ f\colon [0,1]^d \to \mathbb{R} : \|f\|^2 + \sum_{i=1}^{d} \left\| \frac{\partial^{\beta_i}}{\partial x_i^{\beta_i}} f \right\|^2 \le L^2 \right\}, \tag{2}$$

where $\|f\| = (\int_{[0,1]^d} f(\mathbf{x})^2 \, d\mathbf{x})^{1/2}$ denotes the L_2-norm on $[0,1]^d$. See for instance [38]; the $\beta_i > 0$ measure the smoothness of f in the ith direction.

Given an estimator $\hat{f}_n = \hat{f}_n(\mathbf{x}, Z^{(n)})$, $\mathbf{x} \in [0,1]^d$, of f, we say, as usual in minimax theory, that the procedure $\hat{f}_n$ is *asymptotically optimal* w.r.t. Σ if

$$\limsup_{n\to\infty} \sup_{f\in\Sigma} E_f^n \{ \varphi_n^{-2}(\Sigma) \|\hat{f}_n - f\|^2 \} < \infty, \tag{3}$$

where the deterministic normalizing factor $\varphi_n(\Sigma) \to 0$ cannot be improved in order. The notation E_f^n denotes expectation w.r.t. the law of the observation $Z^{(n)}$. Mathematically, the message of statement (3) is clear. How do we interpret it statistically? What is the issue of the minimax theory in this context? First, we construct $\hat{f}_n$. Next, we have a normalizing factor $\varphi_n(\Sigma)$, which can be understood as an *accuracy of estimation*: the procedure $\hat{f}_n$ provides us with a confidence set (in the $\|\cdot\|$-norm) of size $\varphi_n(\Sigma)$; for any level $0 < \gamma < 1$, we can guarantee from (3) the existence of $C = C(\gamma, \Sigma)$ s.t. for n large enough,

$$\sup_{f\in\Sigma} P_f^n \{ \|\hat{f}_n - f\| \ge C\varphi_n(\Sigma) \} \le \gamma. \tag{4}$$

It is *essential* that the quantity $C\varphi_n(\Sigma)$ is known to the statistician if one wants to refer to the minimax risk as a notion of *accuracy*.

In the d-dimensional anisotropic regression context, we know (see, e.g., [24, 39, 40]) that, under some restrictions imposed on the noise ε_i, the minimax rate of convergence over $\Sigma = \Sigma(\boldsymbol{\beta}, L)$, that is, the smallest factor (in order) such that (3) is satisfied, is

$$\varphi_n(\Sigma) = n^{-\beta/(2\beta+1)}, \tag{5}$$

where the *effective smoothness* β is defined by the formula

$$\frac{1}{\beta} = \sum_{i=1}^{d} \frac{1}{\beta_i}. \tag{6}$$

This rate is attained by some kernel estimator which is the best possible one in view of (3). In particular, in the isotropic case $b = \beta_1 = \cdots = \beta_d$, the optimal rate of convergence is $n^{-b/(2b+d)}$. The factor d is a dimensional effect which severely limits the minimax approach. This pessimistic result for large d is the unavoidable payment to obtain an accuracy (4) uniformly over Σ.

Given a single experiment, one may legitimately suspect the "true" function to possess better approximation properties, that is, to lie in a smaller subset of Σ. For instance, f could be smoother, which means $f \in \Sigma(\boldsymbol{\lambda}, M)$, $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_d)$, $M > 0$, with

$$\frac{1}{\boldsymbol{\lambda}} := \sum_{i=1}^{d} \frac{1}{\lambda_i} < \frac{1}{\beta}.$$

If so, f can be estimated with the rate $n^{-\boldsymbol{\lambda}/(2\lambda+1)}$.

Another possibility is that f depends on a smaller number of variables. In this case the dimension of the problem should be reduced in some sense. More precisely, f may depend only on $s < d$ significant variables in a given direction

$$\mathbf{i}_s = (i_1, \ldots, i_s), \qquad 1 \le i_1 < i_2 < \cdots < i_s \le d.$$

This corresponds to the assumption that $f \in \Sigma(\mathbf{i}_s) \subset \Sigma$, where

$$\Sigma(\mathbf{i}_s) = \{ f \in \Sigma : f(x_1, \ldots, x_d) = F(x_{i_1}, \ldots, x_{i_s}) \},$$

and F is a function of s variables. In words, $\Sigma(\mathbf{i}_s)$ consists of the elements of Σ that only depend on the direction $\mathbf{i}_s$. From (5) and (6),

$$\varphi_n\big(\Sigma(\mathbf{i}_s)\big) = n^{-\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)},$$

where $1/\beta(\mathbf{i}_s) = \sum_{j \in \mathbf{i}_s} 1/\beta_j$. Note that $\varphi_n(\Sigma(\mathbf{i}_s))$ coincides with $n^{-\beta/(2\beta+1)}$ if we formally set $\beta_j = \infty$ for $j \notin \mathbf{i}_s$. As we see

$$\varphi_n\big(\Sigma(\mathbf{i}_s)\big)/\varphi_n(\Sigma) \to 0.$$

Note also that, in this anisotropic setting, the rate of convergence on $\Sigma(\mathbf{i}_s)$ depends not only on the number s of significant variables but also on the particular direction $(i_1, \ldots, i_s)$ on which F depends. For example, $d = 3$, $s = 2$ and $\Sigma = \Sigma(\boldsymbol{\beta}, L)$ with $\boldsymbol{\beta} = (1, 2, 3)$. Then $\beta(1, 3) = 3/4 > \beta(1, 2) = 2/3$.

To realize these ideas, the nonparametric community has put intense effort into *adaptive estimation* over the last decade. Among others, we mention the papers of Barron, Birgé and Massart [2], Delyon and Juditski [7], Donoho and Johnstone [8], Donoho, Johnstone, Kerkyacharian and Picard [9], Efromovich and Pinsker [12], Efromovich [11], Goldenshluger and Nemirovski [14], Golubev [15], Härdle and Marron [17], Hall, Kerkyacharian and Picard [20], [22], Lepski [25–28], Lepski and Spokoiny [30], [31], Lepski, Mammen and Spokoiny [32], Neumann and von Sachs [37] and Polyak and Tsybakov [45].

Consider a family $(\Sigma_j,\ j = 1, \ldots, N)$ (possibly N could depend on n; we discard this possibility for the moment) of subsets of Σ, where the optimal rates $\varphi_n(\Sigma_j)$ are "better" than $\varphi_n(\Sigma)$: $\varphi_n(\Sigma_j)/\varphi_n(\Sigma) \to 0$. In the anisotropic regression context, Σ_j could be a set of the type $\Sigma(\mathbf{i}_s)$ or $\Sigma(\boldsymbol{\lambda}, M)$, with $\boldsymbol{\lambda} > \beta$. To take this

refinement into account, the adaptive estimation paradigm proposes the following answer: define the *adaptive rate*

$$\psi_n(f) = \begin{cases} \varphi_n(\Sigma_j) & \text{if } f \in \Sigma_j, \\ \varphi_n(\Sigma) & \text{if } f \in \Sigma \setminus \Big(\bigcup_j \Sigma_j \Big). \end{cases}$$

If the Σ_j are not disjoint and $f \in \bigcup_j \Sigma_j$, we of course take $\psi_n(f) = \inf\{\varphi_n(\Sigma_j), j \in \mathcal{J}(f)\}$, where $\mathcal{J}(f) = \{j : f \in \Sigma_j\}$. We then look for an *adaptive estimator*, that is, a procedure $f_n^{\mathtt{adapt}}$ that satisfies

$$\limsup_{n\to\infty} \sup_{f\in\Sigma} E_f^n\{\psi_n^{-2}(f)\|f_n^{\mathtt{adapt}} - f\|^2\} < \infty. \tag{7}$$

The gain of the adaptive procedure is clear from a mathematical point of view, but what statistical interpretation can we give in terms of *accuracy of estimation* [particularly in the sense of (4)]? The improvement of the adaptive estimator is not observable, since we never know to which Σ_i the function f belongs. This becomes clear by looking at the risk defined by (7): the normalizing factor $\psi_n = \psi_n(f)$ depends on the unknown f. To paraphrase a common saying in the nonparametric community "you know adaptive estimators converge very fast if the function is very smooth (or has a prescribed complexity) but you can tell nothing about the estimated function itself." In other words, the issue of the adaptive approach is only the estimator $f_n^{\mathtt{adapt}}$. The impossibility of computing its accuracy is the unavoidable payment for the adaptive property.

The goal of this paper is to replace in (7) the unknown—and ideal—sequence $\psi_n(f)$ by a quantity $\hat{\rho}_n$ measurable w.r.t. the observation (data driven), and thus improve the accuracy of estimation in the sense of (4). To do so, we introduce the concept of *random normalizing factor* (RNF) and define its optimality. This notion entails the possibility to choose an optimal procedure $f_n^{\mathtt{new}}$. A pair $(\hat{\rho}_n, f_n^{\mathtt{new}})$ being optimal in this sense possesses the following properties:

1. $\hat{\rho}_n \le \varphi_n(\Sigma)$;
2. $\limsup_{n\to\infty} \sup_{f\in\Sigma} E_f^n\{\hat{\rho}_n^{-2}\|f_n^{\mathtt{new}} - f\|^2\} < \infty$;
3. for all $0 < \alpha < 1$, there exist deterministic $z_n(j)$, $j = 1, \dots, N$, that satisfy at least $z_n(j)/\varphi_n(\Sigma) \to 0$ and such that

$$\inf_{f\in\Sigma_j} P_f^n\{\hat{\rho}_n \le z_n(j)\} \ge 1 - \alpha.$$

Let us briefly discuss these points:

(i) We do not lose the minimax properties, since the first two points above guarantee *at least* an accuracy of the type (4).

(ii) If our guess of a simpler structure turns out to be true (i.e., $f \in \Sigma_j$ for some j), the third point shows that the RNF will be essentially smaller than $\varphi_n(\Sigma)$, with probability controlled by a free parameter α. This parameter represents a risk

level chosen by the statistician. It is possible to let $\alpha = \alpha_n \to 0$ depend on n. However, such a choice is a delicate issue which we will discuss below. This is so mainly because the sequence $z_n(j)$ actually depends on α.

(iii) The optimality mentioned above is linked to the best choice of the sequence $z_n(j)$, $j = 1, \dots, N$. For instance, applying the results of Section 3 to anisotropic regression with $\Sigma_j = \Sigma(\mathbf{i}_s)$ we arrive at

$$z_n(\mathbf{i}_s) \asymp \max\left\{\left(\frac{\sqrt{\ln(1/\alpha)}}{n}\right)^{2\beta/(4\beta+1)}, \left(\frac{1}{n}\right)^{\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)}\right\}.$$

This implies that the following hold: (1) The value of α affects the accuracy of estimation: the smaller α, the worse the accuracy. We will show that the choice $\alpha = \alpha_n = n^{-a}$, $a > 0$, will only reduce the accuracy by a factor of $\ln n$. (2) For any direction $\mathbf{i}_s$ such that $\beta(\mathbf{i}_s) < 2\beta$ and for any $\alpha = \alpha_n = n^{-a}$, $a > 0$, we have $z_n(\mathbf{i}_s) = n^{-\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)}$. This means that $z_n(\mathbf{i}_s)$ coincides with the rate of convergence on the set $\Sigma(\mathbf{i}_s)$. As shown in [29], this is the best possible improvement.

(iv) We prove in Section 2.3 that, for α_n converging to 0 fast enough, usually $\alpha_n = n^{-a}$ for an appropriate $a > 0$, the estimator $f_n^{\mathtt{new}}$ is adaptive in the sense of (7). To that extent, we have provided the minimax theory with a procedure possessing a new virtue without losing any previous step of the theory.

(v) It may well happen that $\hat{\rho}_n \ll \varphi_n(\Sigma)$, but $f \notin \bigcup_j \Sigma_j$. In this case, we still improve the accuracy of estimation. However, this suggests that f is somehow "close" to $\bigcup_j \Sigma_j$.

The paper is organized as follows. We present in Section 2 a general mathematical framework for improving the accuracy of estimation, based on the notion of RNF. In particular, we discuss in detail the concepts outlined in the Introduction. The proofs of the results stated in Section 2 are delayed until the Appendix. In Section 3, we apply our results to multivariate regression in an anisotropic Sobolev setup, formulating an answer to our original problem. For transparency, we state our results in the white noise model. The proofs are given in Section 4.

2. Random normalizing factors. In this section, we propose a new approach for improving the accuracy of an estimating procedure (in the sense given by (11) below). In principle, we could apply this approach to an arbitrary statistical model. It is therefore convenient to present the concept itself in terms of an abstract sequence of statistical experiments (see, e.g., [23]).

2.1. *Formal definitions.* We consider an experiment $\mathcal{E} = (\mathcal{X}^\varepsilon, \mathcal{B}^\varepsilon, P_f^\varepsilon, f \in \Sigma)_{\varepsilon>0}$ generated by an observation X_ε. The pair $(\mathcal{X}^\varepsilon, \mathcal{B}^\varepsilon)$ is a measurable space endowed with a family of probability measures $(P_f^\varepsilon,\ f \in \Sigma)$. The parameter space Σ is a bounded subset of a normed space $(V, \|\cdot\|)$ over the real field.

In particular, $\sup_{f\in\Sigma}\|f\|\le Q$ for some constant $Q>0$. Asymptotics are taken as $\varepsilon\to 0$. For the regression model discussed in the Introduction, $1/\sqrt{n}$ plays the role of ε and Σ is the smoothness class (2) of real-valued functions defined on $[0,1]^d$; thus

$$V=\left\{f\colon [0,1]^d\to\mathbb{R},\ \|f\|=\left(\int_{[0,1]^d} f(\mathbf{x})^2\,d\mathbf{x}\right)^{1/2}<\infty\right\}.$$

For any subset $\tilde{\Sigma}\subseteq\Sigma$, we define the risk of an arbitrary estimator $\tilde{f}_\varepsilon=\tilde{f}_\varepsilon(X^\varepsilon)$ by

$$R_\varepsilon\bigl(\tilde{f}_\varepsilon,\tilde{\Sigma},\varphi_\varepsilon(\tilde{\Sigma})\bigr)=\sup_{f\in\tilde{\Sigma}} E_f^\varepsilon\bigl\{\varphi_\varepsilon^{-p}(\tilde{\Sigma})\|\tilde{f}_\varepsilon-f\|^p\bigr\},\qquad p\ge 2, \tag{8}$$

where $\varphi_\varepsilon(\tilde{\Sigma})>0$ is a deterministic normalizing factor. The factor $\varphi_\varepsilon(\tilde{\Sigma})$ is called the *minimax rate of convergence* (MRC) on $\tilde{\Sigma}$ if the following hold:

(i) $\liminf_{\varepsilon\to 0}\inf_{\tilde{f}_\varepsilon} R_\varepsilon(\tilde{f}_\varepsilon,\tilde{\Sigma},\varphi_\varepsilon(\tilde{\Sigma}))>0$, where the infimum is taken over all estimators;

(ii) $\limsup_{\varepsilon\to 0} R_\varepsilon(\hat{f}_\varepsilon,\tilde{\Sigma},\varphi_\varepsilon(\tilde{\Sigma}))<\infty$,

for some estimator $\hat{f}_\varepsilon$, called *asymptotically optimal* on $\tilde{\Sigma}$ *in the minimax sense* (cf. [23]). Later, we will assume that the minimax rate of convergence $\varphi_\varepsilon(\Sigma)$ exists and is known a priori.

Now, let us be given a family of subsets $\Sigma_1,\Sigma_2,\dots,\Sigma_N$ of Σ, that is, $\Sigma_i\subset\Sigma$, for all $i=1,\dots,N$. It is assumed to be known that for each $i=1,\dots,N$ there exists an estimator $\hat{f}_\varepsilon^{(i)}$ such that the following hold:

(iii) $\limsup_{\varepsilon\to 0} R_\varepsilon(\hat{f}_\varepsilon^{(i)},\Sigma_i,\varphi_\varepsilon(\Sigma_i))<\infty$;

(iv) $\varphi_\varepsilon(\Sigma_i)/\varphi_\varepsilon(\Sigma)\to 0$ as $\varepsilon\to 0$.

Suppose now that the statistician has a *guess* based on some a priori qualitative analysis of the model that the parameter f may actually belong to one of the Σ_i. Then, from (i) and (ii), there is some hope of improving the rate $\varphi_\varepsilon(\Sigma)$ in the following way:

Introduce the family Ω_ε of observable normalizing factors (so-called random normalizing factors) defined as the class

$$\Omega_\varepsilon=\bigl\{\rho_\varepsilon\in(0,\varphi_\varepsilon(\Sigma)]\colon \rho_\varepsilon \text{ is a random variable measurable w.r.t. } X^\varepsilon\bigr\}.$$

For an arbitrary $\rho_\varepsilon\in\Omega_\varepsilon$ and an estimator $\tilde{f}_\varepsilon$, introduce the risk

$$R_\varepsilon^{(r)}\bigl(\tilde{f}_\varepsilon,\Sigma,\rho_\varepsilon\bigr)=\sup_{f\in\Sigma} E_f^\varepsilon\bigl\{\rho_\varepsilon^{-p}\|\tilde{f}_\varepsilon-f\|^p\bigr\}. \tag{9}$$

The superscript (r) is put here to emphasize the random character of the normalizing factor. Suppose that there exist both a random normalizing factor $\hat{\rho}_\varepsilon\in\Omega_\varepsilon$ and an estimator f_ε^* such that, for some $M=M(\Sigma)$,

$$\limsup_{\varepsilon\to 0} R_\varepsilon^{(r)}(f_\varepsilon^*,\Sigma,\hat{\rho}_\varepsilon)\le M<\infty. \tag{10}$$

Clearly, for all $0 < \gamma < 1$, we have

$$P_f^{\varepsilon}\left\{\|f_{\varepsilon}^* - f\| \geq \left(\frac{M}{\gamma}\right)^{1/p} \hat{\rho}_{\varepsilon}\right\} \leq \gamma \tag{11}$$

as $\varepsilon \to 0$ and we can treat $\hat{\rho}_{\varepsilon}$ as an *accuracy of estimation* provided by the estimator f_{ε}^*. This yields a confidence set for f. Note that it is essential in our approach that the quantities $\hat{\rho}_{\varepsilon}$ and $M = M(\Sigma)$ are known. Let us mention two fundamental facts:

1. If such $\hat{\rho}_{\varepsilon}$ and f_{ε}^* exist, then f_{ε}^* is asymptotically optimal on Σ in the minimax sense as $\hat{\rho}_{\varepsilon} \in \Omega_{\varepsilon}$ by definition.
2. Since the random normalizing factor $\hat{\rho}_{\varepsilon}$ depends on the observation (i.e., on the function f itself and *not only* on the whole class Σ), we can hope to get some extra knowledge of f, rather than simply "$f \in \Sigma$." In particular, we may try to find a $\hat{\rho}_{\varepsilon}$ which, with some positive probability, is smaller than the rate of convergence $\varphi_{\varepsilon}(\Sigma)$. In this sense, some improvement of the accuracy of estimation is achievable.

Finding an estimator f_{ε}^* and a RNF $\hat{\rho}_{\varepsilon} \in \Omega_{\varepsilon}$ such that

$$\limsup_{\varepsilon \to 0} R_{\varepsilon}^{(r)}(f_{\varepsilon}^*, \Sigma, \hat{\rho}_{\varepsilon}) < \infty$$

will provide an improvement of the accuracy of estimation if

$$\forall i = 1, \ldots, N, \qquad \liminf_{\varepsilon \to 0} \inf_{f \in \Sigma_i} P_f^{\varepsilon}\{\hat{\rho}_{\varepsilon} < \varphi_{\varepsilon}(\Sigma)\} > 0; \tag{12}$$

otherwise, we would not gain anything new. The issue we now address is how to describe, in a favorable situation like (12), an optimal improvement taking into account the fact that we believe that $f \in \Sigma_i$ for some $i \in \{1, \ldots, N\}$ and how to construct an optimal procedure accordingly. Let us mention that in [29] the case $N = 1$ was considered. However, a nontrivial extension is needed for the case $N \geq 1$. Note also that we could develop a theory by letting $N = N_{\varepsilon}$ grow to infinity as $\varepsilon \to 0$, but a fixed N will be sufficient for the generalization level intended here (and the application to significant variables in particular).

Let $0 < \delta < 1$ be some given number and let us fix a function α_{ε} assumed to be small, such that $0 < \alpha_{\varepsilon} \leq 1 - \delta$ for all $\varepsilon \in (0, 1)$. The function α_{ε} is arbitrary and fixed by the statistician. We want to guarantee that if actually $f \in \Sigma_i$ for some i, then we can provide some improvement with confidence $1 - \alpha_{\varepsilon}$, uniformly over Σ_i. We thus introduce the following definition.

DEFINITION 1 (Characteristic of ρ_{ε}). The characteristic of $\rho_{\varepsilon} \in \Omega_{\varepsilon}$ is the deterministic sequence $(x_{\varepsilon}(\rho_{\varepsilon}, i), i = 1, \ldots, N)$ of functions

$$x_{\varepsilon}(\rho_{\varepsilon}, i) = \inf\Big\{x \in (0, \varphi_{\varepsilon}(\Sigma)] : \inf_{f \in \Sigma_i} P_f^{\varepsilon}\{\rho_{\varepsilon} \leq x\} \geq 1 - \alpha_{\varepsilon}\Big\}. \tag{13}$$

Note that the $x_\varepsilon(\rho_\varepsilon, i)$ depend on α_ε. The sequence of deterministic factors $(x_\varepsilon(\rho_\varepsilon, i),\ i = 1, \dots, N)$ measures the improvement rate that ρ_ε provides uniformly over each subset Σ_i, with prescribed probability (given by α_ε). We naturally derive the following criterion of comparison between RNFs.

DEFINITION 2 (Optimality of random normalizing factors). The RNF $\rho_\varepsilon^* \in \Omega_\varepsilon$ is α-optimal (asymptotically optimal) w.r.t. the family $(\Sigma_i,\ i = 1, \dots, N)$ if the following two conditions are fulfilled:

(i) There exists an estimator f_ε^* such that

$$\limsup_{\varepsilon \to 0} R_\varepsilon^{(r)}\bigl(f_\varepsilon^*, \Sigma, \rho_\varepsilon^*\bigr) < \infty. \tag{14}$$

(ii) If there exist $\rho_\varepsilon \in \Omega_\varepsilon$ and $j \in \{1, \dots, N\}$ such that

$$\frac{x_\varepsilon(\rho_\varepsilon, j)}{x_\varepsilon(\rho_\varepsilon^*, j)} \to 0 \quad \text{as } \varepsilon \to 0,$$

then

$$\liminf_{\varepsilon \to 0} \inf_{\hat f_\varepsilon} R_\varepsilon^{(r)}\bigl(\hat f_\varepsilon, \Sigma, \rho_\varepsilon\bigr) = +\infty, \tag{15}$$

where the infimum is taken over all estimators.

DEFINITION 3. Let ρ_ε^* be an α-optimal RNF w.r.t. $(\Sigma_i,\ i = 1, \dots, N)$. Then any estimator providing (14) in Definition 2 above is called α-adaptive.

REMARK 1. As we will see later, an optimal RNF ρ_ε^* can often be constructed as a random variable taking $N + 1$ values $x_\varepsilon(\rho_\varepsilon^*, i)$, $i = 0, \dots, N$, with the notational convention $x_\varepsilon(\rho_\varepsilon^*, 0) = \varphi_\varepsilon(\Sigma)$. The value $x_\varepsilon(\rho_\varepsilon^*, 0)$ cannot be improved in order because it is the minimax rate of convergence over Σ. The values $x_\varepsilon(\rho_\varepsilon^*, i)$ cannot be improved in order due to (15). Both these facts together with (14) explain why ρ_ε^* is called α-optimal.

REMARK 2. By definition, $\rho_\varepsilon^* \le \varphi_\varepsilon(\Sigma)$ for all $\varepsilon \in (0, 1)$ and any α_ε. Therefore from (14) any α-adaptive estimator is asymptotically optimal on the set Σ w.r.t. the risk (8). It means that, by considering risks of the type (9), we cover the framework of the standard minimax approach.

REMARK 3. In principle, we can use random normalizing factors for any norm $\|\cdot\|$ in the risk function (9). We will see in Section 3 a successful application in L_2-norm. Analogous improvements could be obtained in L_p-norm, $1 \le p < \infty$, but they lie beyond the scope of the paper. However, if we define the minimax risk in uniform norm L_∞, it is impossible to improve substantially the accuracy, in the sense that any optimal ρ_ε^* will be of order $\varphi_\varepsilon(\Sigma)$. This phenomenon is closely related to the work of Low [33] on nonparametric confidence intervals and follows from his result.

2.2. *Canonical construction of RNF.* Whenever a specific model is considered, it is often the case that one can construct an optimal RNF with respect to a single subset $\Sigma_1 \subset \Sigma$ (see [29] for specific examples), that is, in the case $N = 1$. Such optimal RNFs usually take only two values $\varphi_\varepsilon(\Sigma)$ and $\varphi_\varepsilon(\alpha_\varepsilon)$, the latter value $\varphi_\varepsilon(\alpha_\varepsilon)$ corresponding to the improvement obtained when one believes that $f \in \Sigma_1$. We now address the following task: construct an optimal RNF w.r.t. the family of subsets $(\Sigma_i,\ i = 1, \dots, N)$ when one can solve separately the problem of finding an optimal RNF w.r.t. a single Σ_i, for $i = 1, \dots, N$. This construction will be exploited later, in Section 3, in the multidimensional white noise model.

We consider the following assumption: for each $i \in \{1, \dots, N\}$ there exist $0 < \varphi_{\varepsilon,i}(\alpha_\varepsilon) \le \varphi_\varepsilon(\Sigma)$, random variables $\rho^*_{\varepsilon,i} \in \Omega_\varepsilon$ with values in $\{\varphi_\varepsilon(\Sigma), \varphi_{\varepsilon,i}(\alpha_\varepsilon)\}$ and estimators $f^*_{\varepsilon,i}$ possessing the following three properties:

$P_1(i)$ $\inf_{f\in\Sigma_i} P^\varepsilon_f\{\rho^*_{\varepsilon,i} = \varphi_{\varepsilon,i}(\alpha_\varepsilon)\} \ge 1 - \alpha_\varepsilon$.

$P_2(i)$ $\limsup_{\varepsilon\to 0} \sup_{f\in\Sigma} E^\varepsilon_f\{(\rho^*_{\varepsilon,i})^{-p} \| f^*_{\varepsilon,i} - f\|^p\} \le M^*_i < \infty$.

$P_3(i)$ For all $\rho_\varepsilon \in \Omega_\varepsilon$ with values in $\{\varphi_\varepsilon(\Sigma), a_\varepsilon\}$ satisfying

$$\frac{a_\varepsilon}{\varphi_{\varepsilon,i}(\alpha_\varepsilon)} \to 0 \quad \text{as } \varepsilon \to 0$$

and

$$\inf_{f\in\Sigma_i} P^\varepsilon_f\{\rho_\varepsilon = a_\varepsilon\} \ge 1 - \alpha_\varepsilon,$$

we have

$$\liminf_{\varepsilon\to 0} \inf_{\hat f_\varepsilon} \sup_{f\in\Sigma} E^\varepsilon_f\{\rho_\varepsilon^{-p} \| \hat f_\varepsilon - f\|^p\} = +\infty,$$

where the infimum is taken over all estimators.

We claim that if such an assumption is granted, we can find a recipe to construct an α-optimal RNF and an α-adaptive estimator w.r.t. the family $(\Sigma_i,\ i = 1, \dots, N)$ in a canonical way. Define i^* by the formula

$$\rho^*_{\varepsilon,i^*} = \inf_{i=1,\dots,N} \rho^*_{\varepsilon,i}$$

and put

$$\rho^*_\varepsilon = \rho^*_{\varepsilon,i^*}, \qquad f^*_\varepsilon = f^*_{\varepsilon,i^*}. \tag{16}$$

PROPOSITION 1. *Suppose that ρ^*_ε and f^*_ε are defined by* (16). *Then ρ^*_ε is α-optimal w.r.t. $(\Sigma_i,\ i = 1, \dots, N)$ and f^*_ε is α-adaptive.*

The following result gives a bound on the constant M^* of the risk of f^*_ε depending on N. It shows that, under mild conditions, M^* behaves no worse than the "worst" of the M^*_i. Define, for $i = 1, \dots, N$,

$$\xi_{\varepsilon,i}(f) = (\rho^*_{\varepsilon,i})^{-1} \| f^*_{\varepsilon,i} - f\|.$$

COROLLARY 1. *Assume that* ρ_ε^* *and* f_ε^* *are defined by* (16) *and that, for* $i = 1, \dots, N$, *there exist* $0 < M_i < \infty$ *such that*

$$\lim_{\varepsilon \to 0} \sup_{f \in \Sigma} E_f^\varepsilon \{ [\xi_{\varepsilon,i}(f)]^p 1_{\xi_{\varepsilon,i}(f) \geq M_i} \} = 0. \tag{17}$$

Then

$$\limsup_{\varepsilon \to 0} \sup_{f \in \Sigma} E_f^\varepsilon \{ (\rho_\varepsilon^*)^{-p} \| f_\varepsilon^* - f \|^p \} \leq \Big(\sup_{i=1,\dots,N} M_i \Big)^p.$$

Finally, we state the following obvious result, namely that a pair of an α-optimal RNF and an α-adaptive estimator w.r.t. $(\Sigma_i,\ i = 1, \dots, N)$ automatically provides N α-optimal RNFs and N α-adaptive estimators w.r.t. each set Σ_i, for $i = 1, \dots, N$.

PROPOSITION 2. *Let* ρ_ε^* *and* f_ε^* *be* α*-optimal RNF and* α*-adaptive estimators w.r.t. the family* $(\Sigma_i,\ i = 1, \dots, N)$. *Put, for* $i = 1, \dots, N$,

$$\hat{\rho}_{\varepsilon,i} = \begin{cases} x_\varepsilon(\rho_\varepsilon^*, i), & \text{if } \rho_\varepsilon^* \leq x_\varepsilon(\rho_\varepsilon^*, i), \\ \varphi_\varepsilon(\Sigma), & \text{if } \rho_\varepsilon^* > x_\varepsilon(\rho_\varepsilon^*, i), \end{cases}$$

and

$$\hat{f}_{\varepsilon,i} = f_\varepsilon^*.$$

Then for $i = 1, \dots, N$, *the pair* $(\hat{\rho}_{\varepsilon,i}, \hat{f}_{\varepsilon,i})$ *possesses the properties* $P_1(i)$, $P_2(i)$ *and* $P_3(i)$.

In fact, Propositions 1 and 2 show that conditions $P_1(i)$, $P_2(i)$ and $P_3(i)$, $i = 1, \dots, N$, are necessary and sufficient for ρ_ε^* defined in (16) to be α-optimal.

2.3. *Links to adaptive estimation.* We keep the framework of Section 2.1 and consider a family of subsets $(\Sigma_i,\ i = 1, \dots, N)$ of Σ satisfying (iii) and (iv) of Section 2.1. An *adaptive* estimator $f_\varepsilon^{(a)}$ (if it exists) satisfies

$$\forall i = 1, \dots, N, \qquad \limsup_{\varepsilon \to 0} \sup_{f \in \Sigma_i} E_f^\varepsilon \{ \varphi_\varepsilon^{-p}(\Sigma_i) \| \hat{f}_\varepsilon^{(a)} - f \|^p \} < \infty, \tag{18}$$

that is, achieves the optimal rate simultaneously over all the Σ_i, without the knowledge of Σ_i. Putting $\mathcal{I}(f) = \{i\colon\ f \in \Sigma_i\}$ and

$$\psi_\varepsilon(f) = \begin{cases} \inf\{\varphi_\varepsilon(\Sigma_i),\ i \in \mathcal{I}(f)\}, & \text{if } \mathcal{I}(f) \neq \varnothing, \\ \varphi_\varepsilon(\Sigma), & \text{if } \mathcal{I}(f) = \varnothing, \end{cases}$$

we obtain an equivalent characterization of $\hat{f}_\varepsilon^{(a)}$:

$$\limsup_{\varepsilon \to 0} \sup_{f \in \Sigma} E_f^\varepsilon \{ \psi_\varepsilon(f)^{-p} \| \hat{f}_\varepsilon^{(a)} - f \|^p \} < \infty. \tag{19}$$

Thus, the normalizing factor $\psi_\varepsilon(f)$ is essentially better than the minimax rate of convergence $\varphi_\varepsilon(\Sigma)$ if f belongs to some Σ_i. In this sense, the procedure $\hat{f}_\varepsilon^{(a)}$ is better than an estimator asymptotically optimal on Σ. However, we cannot take $\psi_\varepsilon(\cdot)$ as an accuracy of estimation in the sense of (11), since $\psi_\varepsilon(\cdot) = \psi_\varepsilon(f)$ depends on the unknown f. In that specific sense, the adaptive estimator $\hat{f}_\varepsilon^{(a)}$ does not improve the minimal accuracy of estimation $\varphi_\varepsilon(\Sigma)$. However, under further restrictions, we have that an α-adaptive estimator can be *adaptive* in the usual sense given by (18) or (19):

PROPOSITION 3. *Let f_ε^* be an α-adaptive estimator and let ρ_ε^* be an α-optimal RNF w.r.t. the family $(\Sigma_i,\ i = 1, \ldots, N)$. Supppose that the estimator*

$$\hat{f}_{\varepsilon,i} := f_\varepsilon^* 1_{\{\rho_\varepsilon^* \le x_\varepsilon(\rho_\varepsilon^*, i)\}}$$

is asymptotically optimal w.r.t. Σ_i for all $i = 1, \ldots, N$. If, moreover, $\alpha_\varepsilon = \mathcal{O}(\inf_{i \le N}\ \varphi_\varepsilon(\Sigma_i)^p)$ as $\varepsilon \to 0$, then f_ε^* is adaptive w.r.t. $(\Sigma_i,\ i = 1, \ldots, N)$ in the usual sense and satisfies in particular (18).

The proof of Proposition 3 is delayed until the Appendix.

REMARK 4. We will see in the next section that we can construct an α-adaptive estimator of this kind in the problem of significant variables in Gaussian white noise.

REMARK 5. Thus, under some restrictions, α-adaptive estimators provide us with adaptive estimators. We thus retrieve the classical results of adaptive estimation. The converse, that is, whether adaptive estimators can be α-adaptive (in connection with an appropriate α-optimal RNF) in general, remains an open question.

2.4. *Links to nonparametric confidence sets.* Clearly, the concept of RNF is related to confidence sets: by (10) and (11), any suitable RNF $\hat{\rho}_\varepsilon$ yields a confidence set in the $\|\cdot\|$-norm as the ball with center f_ε^* and radius $(M/\gamma)^{1/p}\hat{\rho}_\varepsilon$. For all $\gamma \in]0, 1[$, this confidence set has coverage probability over the class Σ of at least $1 - \gamma$ as $\varepsilon \to 0$. The random radius of the confidence set is at least of order $\varphi_\varepsilon(\Sigma)$, but, with prescribed probability $1 - \alpha_\varepsilon$, this radius can be essentially better than $\varphi_\varepsilon(\Sigma)$ if we take for $\|\cdot\|$ the L_2-norm. This is developed in detail in the example of multivariate regression in Section 3 below, in a setting where one suspects the unknown signal to depend on fewer significant variables than those prescribed in the original model.

The severe limitation of adaptation for nonparametric confidence intervals (i.e., if one tries to construct a confidence interval for f at a particular point x_0) has been studied by Low [33]. Our approach is not in conflict with the lower bound

of Low if we consider, instead of pointwise, global error measurement such as the L_2-norm. See also Remark 3 in Section 2.1 above.

There exists another quite different approach for constructing nonparametric confidence intervals and bands: one starts by proving the asymptotic normality of a pivotal quantity, usually relying on a preliminary nonparametric estimator of the unknown function, and then uses bias correction. Several methods have been investigated in the literature: we refer to [18, 19, 36], based on kernel estimation, and more recently [44], based on nonlinear methods connected to wavelet thresholding. The accuracy of this latter approach can be measured in terms of the length of the interval and also the coverage error: for constructing a confidence interval I_n for a point x_0 based on n observations, one can seek an expansion of the quantity $P_f^n(f(x_0) \in I_n)$ as $n \to \infty$, a result we do not have here.

It is noteworthy that this latter approach significantly differs from the minimax setting, since a uniform result of the type (10) is lost. Moreover, it is developed for confidence intervals at a fixed point x_0 whereas we consider global loss measurement in $\|\cdot\|$-norm instead.

3. Anisotropic multidimensional white noise model. In this section, we apply the general results developed in the previous sections to the problem of estimating a nonparametric real-valued function on $[0,1]^d$, when we believe that the function may only depend on $s < d$ variables. The statistical model we consider is multidimensional white Gaussian noise (WGN). The motivation for choosing WGN is at least twofold:

1. It clarifies the mathematical transparency of the problem at hand and allows us to avoid technicalities and routine computations.
2. It is known that univariate regression (see [4]) and density estimation (see [16]) are asymptotically equivalent to white noise (for smooth enough parameter classes). We believe that the same type of equivalence is true for multivariate regression with random normalizing factors.

Consider the statistical experiment generated by the observation

$$X_\varepsilon = \big(X_\varepsilon(\mathbf{x}),\ \mathbf{x} = (x_1, \dots, x_d) \in [0,1]^d\big),$$

where

$$X_\varepsilon(d\mathbf{x}) = f(\mathbf{x})\, d\mathbf{x} + \varepsilon W(d\mathbf{x}). \tag{20}$$

The random process W is a standard d-dimensional Brownian sheet, ε is a noise level, and $f \in L_2([0,1]^d)$ is the unknown parameter of interest. In other words, for any $g \in L_2([0,1]^d)$, we are given

$$\int_{[0,1]^d} g(\mathbf{x}) X_\varepsilon(d\mathbf{x}) = \int_{[0,1]^d} f(\mathbf{x}) g(\mathbf{x})\, d\mathbf{x} + \varepsilon \xi(g),$$

where $\xi(g) \sim \mathcal{N}(0, \int_{[0,1]^d} g(\mathbf{x})^2\,d\mathbf{x})$. In the framework of Section 2, we consider $(V, \|\cdot\|) = L_2([0,1]^d)$ and the experiment $\mathcal{E} = (\mathcal{X}^\varepsilon, \mathcal{B}^\varepsilon, P_f^\varepsilon, f \in \Sigma)_{\varepsilon>0}$ generated by X_ε, where $\Sigma = \Sigma(\boldsymbol{\beta}, L)$, $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_d)$, $\beta_i > 0$, $i = 1, \ldots, d$, is the anisotropic Sobolev body defined as follows.

Let $(\phi_k,\ k \in \mathbf{N})$ be an orthonormal basis of $L_2([0,1])$. We require that

$$\int_{[0,1]} \phi_k(x)\,dx = \delta_{0k}, \tag{21}$$

where δ is the Kronecker symbol. For instance, we may consider the Fourier basis $\phi_0 = 1$, and for $k \neq 0$,

$$\phi_{2k}(x) = \sqrt{2}\cos(2\pi kx), \qquad \phi_{2k-1}(x) = \sqrt{2}\sin(2k\pi x)$$

but other choices are obviously possible. For a multiindex $\mathbf{k} = (k_1, \ldots, k_d) \in \mathbf{N}^d$, define

$$\phi_{\mathbf{k}}(\mathbf{x}) = \phi_{\mathbf{k}}(x_1, \ldots, x_d) = \phi_{k_1}(x_1) \cdots \phi_{k_d}(x_d).$$

The sequence $(\phi_{\mathbf{k}},\ \mathbf{k} \in \mathbf{N}^d)$ provides an orthonormal basis of $L_2([0,1]^d)$. For a given $f \in L_2([0,1]^d)$, the following expansion holds in L_2:

$$f = \sum_{\mathbf{k} \in \mathbf{N}^d} \theta_{\mathbf{k}} \phi_{\mathbf{k}},$$

where

$$\theta_{\mathbf{k}} = \int_{[0,1]^d} f(\mathbf{x}) \phi_{\mathbf{k}}(\mathbf{x})\,d\mathbf{x}. \tag{22}$$

In particular (Parseval identity), $\|f\| = \|\theta\|_{l^2}$, where $\|\theta\|_{l^2}^2 = \sum_{\mathbf{k} \in \mathbf{N}^d} \theta_{\mathbf{k}}^2$. For simplicity, we will omit any further reference to the subscript l^2. Set

$$\Sigma(\boldsymbol{\beta}, L) := \left\{ \theta : \sum_{\mathbf{k} \in \mathbf{N}^d} \theta_{\mathbf{k}}^2 \left(1 + \sum_{i=1}^d k_i^{2\beta_i} \right) \leq L^2 \right\}. \tag{23}$$

REMARK 6. This definition of $\Sigma(\boldsymbol{\beta}, L)$ corresponds to the class (2) considered in the Introduction if we identify f and its expansion $\theta = (\theta_{\mathbf{k}})_{\mathbf{k} \in \mathbf{N}^d}$ in the basis $(\phi_{\mathbf{k}})_{\mathbf{k} \in \mathbf{N}^d}$ and assume that f is periodic. However, as we will see in (25) below, only a sequence space model is considered. Any further reference to $\Sigma(\boldsymbol{\beta}, L)$ will refer to the definition (23).

For a given $s < d$ and a direction $\mathbf{i}_s = (i_1, \ldots, i_s)$, where $i_1 < \cdots < i_s$, define

$$I(\mathbf{i}_s) = \left\{ (k_1, \ldots, k_d) \in \mathbf{N}^d : k_j = 0,\ \forall j \notin \{i_1, \ldots, i_s\} \right\}$$

and

$$\Sigma(\mathbf{i}_s) = \left\{ \theta : \sum_{\mathbf{k} \in I(\mathbf{i}_s)} \theta_{\mathbf{k}}^2 \left(1 + \sum_{i=1}^d k_i^{2\beta_i} \right) \leq L^2,\ \text{and } \theta_{\mathbf{k}} = 0,\ \mathbf{k} \notin I(\mathbf{i}_s) \right\}.$$

Note that, using property (21), if

$$f(x_1, \dots, x_d) = F(x_{i_1}, \dots, x_{i_s}),$$

then

$$\theta_{\mathbf{k}} = 0 \quad \text{for all } \mathbf{k} \in \mathbf{N}^d \setminus I(\mathbf{i}_s). \tag{24}$$

Thus, the functions which only depend on the direction $\mathbf{i}_s$ are contained in $\Sigma(\mathbf{i}_s)$.

Next, consider the observation X_ε given by (20). For $\mathbf{k} \in \mathbf{N}^d$, set

$$Y_{\mathbf{k}} = \int_{[0,1]^d} \phi_{\mathbf{k}}(\mathbf{x}) X_\varepsilon(d\mathbf{x}).$$

The following decomposition holds:

$$Y_{\mathbf{k}} = \theta_{\mathbf{k}} + \varepsilon \xi_{\mathbf{k}}, \qquad \mathbf{k} \in \mathbf{N}^d, \tag{25}$$

where $\theta_{\mathbf{k}}$ is defined by (22) and

$$\xi_{\mathbf{k}} = \int_{[0,1]^d} \phi_{\mathbf{k}}(\mathbf{x}) W(d\mathbf{x})$$

is a standard Gaussian random variable with zero mean and unit variance. Moreover, from the orthogonality of the $\phi_{\mathbf{k}}$, the $\xi_{\mathbf{k}}$ are independent. Thus, the white noise model given by (20) can be reformulated in terms of the sequence model (25). Because of the equivalence of (20) and (25), we can identify f and its expansion $\theta = (\theta_{\mathbf{k}})_{\mathbf{k} \in \mathbf{N}^d}$ in the basis $(\phi_{\mathbf{k}})_{\mathbf{k} \in \mathbf{N}^d}$. We also write P_θ^ε, E_θ^ε for P_f^ε, E_f^ε, respectively. Likewise, we identify any estimator $\tilde{f}_\varepsilon$ of f by its expansion $\tilde{\theta}_\varepsilon = (\tilde{\theta}_{\mathbf{k}})_{\mathbf{k} \in \mathbf{N}^d}$ in the basis $(\phi_{\mathbf{k}})_{\mathbf{k} \in \mathbf{N}^d}$. We can then define, for any $\rho_\varepsilon \in \Omega_\varepsilon$,

$$R_\varepsilon^{(r)}(\tilde{\theta}_\varepsilon, \Sigma, \rho_\varepsilon) = \sup_{\theta \in \Sigma} E_\theta^\varepsilon \{\rho_\varepsilon^{-p} \|\tilde{\theta}_\varepsilon - \theta\|^p\}.$$

We first consider the problem of constructing a random normalizing factor and an α-adaptive estimator w.r.t. $\Sigma(\mathbf{i}_s)$, for a given direction $\mathbf{i}_s$. We then treat the general case (i.e., considering all the directions $\mathbf{i}_s$ simultaneously) by means of the construction given in (16).

3.1. *The case of a given direction* $\mathbf{i}_s$.

3.1.1. *Construction of a random normalizing factor and* α*-adaptive estimator.* We begin with some notation. If $\mathbf{i}_s = (i_1, \dots, i_s)$, we abuse notation slightly by writing "$i \in \mathbf{i}_s$" for an index $i \in \mathbf{N}$ when we should actually write "$i \in \{i_1, \dots, i_s\}$."

Define β and $\beta(\mathbf{i}_s)$ by

$$\frac{1}{\beta} = \sum_{i=1}^{d} \frac{1}{\beta_i}, \qquad \frac{1}{\beta(\mathbf{i}_s)} = \sum_{i \in \mathbf{i}_s} \frac{1}{\beta_i}$$

and

$$C = \prod_{i=1}^{d} \beta_i^{1/2\beta_i}, \qquad C(\mathbf{i}_s) = \prod_{i \in \mathbf{i}_s} \beta_i^{1/2\beta_i},$$

$$Z_1 = L^{1/2\beta+1} C^{\beta/2\beta+1} \sqrt{\frac{2\beta+1}{2\beta}} 2^{1/4\beta+2}.$$

$Z_1(\mathbf{i}_s)$ is defined analogously, replacing β and C by $\beta(\mathbf{i}_s)$ and $C(\mathbf{i}_s)$, respectively;

$$Z_2 = L^{1/4\beta+1} C^{\beta/4\beta+1} \sqrt{\frac{4\beta+1}{4\beta}} 2^{1/4\beta+1}.$$

Many of the quantities involved hereafter depend on $\mathbf{i}_s$. We will sometimes omit the reference in $\mathbf{i}_s$, for notational simplicity. Set $K_\varepsilon = (K_1(\varepsilon), \ldots, K_s(\varepsilon))$, where, for any $j = 1, \ldots, s$ and $i_j \in \mathbf{i}_s$, we set

$$K_j(\varepsilon) = \beta_{i_j}^{1/2\beta_{i_j}} \left(\frac{\varepsilon^2 C(\mathbf{i}_s)}{2L^2} \right)^{-(1/\beta_{i_j})/[\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)]}.$$

Set $N_\varepsilon = (N_1(\varepsilon), \ldots, N_d(\varepsilon))$, where, for $i = 1, \ldots, d$,

$$N_i(\varepsilon) = \beta_i^{1/2\beta_i} \left(\frac{\varepsilon^2 \sqrt{C \ln(1/\alpha_\varepsilon)}}{4L^2} \right)^{-(1/\beta_i)/[2\beta/(4\beta+1)]}.$$

Set $M_\varepsilon = (M_1(\varepsilon), \ldots, M_d(\varepsilon))$, where, for $i = 1, \ldots, d$,

$$M_i(\varepsilon) = \beta_i^{1/2\beta_i} \left(\frac{\varepsilon^2 C}{2L^2} \right)^{-(1/\beta_i)/[\beta/(2\beta+1)]}.$$

We need to introduce the following five multiindex sets:

$$I = \{(k_1, \ldots, k_d) \in \mathbf{N}^d : k_j = 0,\ \forall j \notin \mathbf{i}_s\};$$

$$I_\varepsilon = \{(k_1, \ldots, k_d) \in I : k_{i_j} \le K_j(\varepsilon),\ j = 1, \ldots, s\};$$

$$J = \mathbf{N}^d \setminus I;$$

$$J_\varepsilon = \{(k_1, \ldots, k_d) \in J : k_i \le N_i(\varepsilon),\ i = 1, \ldots, d\};$$

$$Q_\varepsilon = \{(k_1, \ldots, k_d) \in \mathbf{N}^d : k_i \le M_i(\varepsilon),\ i = 1, \ldots, d\}.$$

Note that $I_\varepsilon = I_\varepsilon(\mathbf{i}_s)$ and $J_\varepsilon = J_\varepsilon(\mathbf{i}_s)$ and that the sets I_ε, J_ε and Q_ε are finite.

We consider the following estimators: $\hat{\theta}_\varepsilon = (\hat{\theta}_{\varepsilon,\mathbf{k}})_{\mathbf{k} \in \mathbf{N}^d}$ with

$$\hat{\theta}_{\varepsilon,\mathbf{k}} = \begin{cases} Y_{\mathbf{k}}, & \mathbf{k} \in Q_\varepsilon, \\ 0, & \mathbf{k} \in \mathbf{N}^d \setminus Q_\varepsilon, \end{cases}$$

and $\hat{\theta}_\varepsilon^{(0)}(\mathbf{i}_s) = (\hat{\theta}_{\varepsilon,\mathbf{k}}^{(0)}(\mathbf{i}_s))_{\mathbf{k}\in\mathbf{N}^d}$, with

$$\hat{\theta}_{\varepsilon,\mathbf{k}}^{(0)}(\mathbf{i}_s) = \begin{cases} Y_{\mathbf{k}}, & \mathbf{k} \in I_\varepsilon, \\ 0, & \mathbf{k} \in \mathbf{N}^d \setminus I_\varepsilon. \end{cases}$$

Note that $\hat{\theta}_\varepsilon = \hat{\theta}_\varepsilon(M_\varepsilon)$ and that $\hat{\theta}_\varepsilon^{(0)}(\mathbf{i}_s) = \hat{\theta}_\varepsilon^{(0)}(\mathbf{i}_s, K_\varepsilon)$. Note also that $\hat{\theta}_\varepsilon$ and $\hat{\theta}_\varepsilon^{(0)}$ are the usual projection estimators corresponding to Σ and $\Sigma(\mathbf{i}_s)$ respectively.

We also introduce the random variable $T_\varepsilon(\mathbf{i}_s) = T_\varepsilon(\mathbf{i}_s, N_\varepsilon)$, which will be used to construct a decision rule:

$$T_\varepsilon(\mathbf{i}_s) = \sum_{\mathbf{k}\in J_\varepsilon} (Y_{\mathbf{k}}^2 - \varepsilon^2).$$

Let

$$\varphi_\varepsilon(\Sigma) = \varepsilon^{2\beta/(2\beta+1)}$$

and

$$\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s) = \sup\left\{ \left(\varepsilon^4 \ln\frac{1}{\alpha_\varepsilon}\right)^{\beta/(4\beta+1)}, \varepsilon^{2\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)} \right\}.$$

We are now ready to define a RNF $\rho_\varepsilon^*(\mathbf{i}_s)$ and a corresponding estimator $\theta_\varepsilon^*(\mathbf{i}_s)$. Set

$$\lambda = \sqrt{2}C^{\beta/(4\beta+1)}(2L)^{1/(4\beta+1)}$$

and

$$\mathcal{A}_\varepsilon(\mathbf{i}_s) = \{T_\varepsilon(\mathbf{i}_s) \le \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s)^2\}.$$

Finally, our estimator $\theta_\varepsilon^*(\mathbf{i}_s)$ and random normalizing factor $\rho_\varepsilon^*(\mathbf{i}_s)$ are defined as follows:

$$\rho_\varepsilon^*(\mathbf{i}_s) = \begin{cases} \varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s), & \text{on } \mathcal{A}_\varepsilon(\mathbf{i}_s), \\ \varphi_\varepsilon(\Sigma), & \text{on } \mathcal{A}_\varepsilon^c(\mathbf{i}_s), \end{cases}$$

$$\theta_\varepsilon^*(\mathbf{i}_s) = \begin{cases} \hat{\theta}_\varepsilon^{(0)}(\mathbf{i}_s), & \text{on } \mathcal{A}_\varepsilon(\mathbf{i}_s), \\ \hat{\theta}_\varepsilon, & \text{on } \mathcal{A}_\varepsilon^c(\mathbf{i}_s). \end{cases}$$

3.1.2. *Main result for a given direction.*

THEOREM 1. *Let $p \ge 2$, $1/20 \ge \alpha_\varepsilon \ge \varepsilon^a$ for $\varepsilon \in (0,1)$ and some $a > 0$. Then $\rho_\varepsilon^*(\mathbf{i}_s)$ is an α-optimal random normalizing factor w.r.t. $\Sigma(\mathbf{i}_s)$ and $\theta_\varepsilon^*(\mathbf{i}_s)$ is α-adaptive. In particular*

$$\limsup_{\varepsilon\to 0} R_\varepsilon^{(r)}\big(\theta_\varepsilon^*(\mathbf{i}_s), \Sigma, \rho_\varepsilon^*(\mathbf{i}_s)\big) \le M^*(\mathbf{i}_s) \tag{26}$$

with

$$M^*(\mathbf{i}_s) = \begin{cases} Z_1^p + \sup\limits_{x \geq 1} \bigl(\lambda^2 x + Z_1^2(\mathbf{i}_s) + Z_2^2\bigr)^{p/2} e^{-(x-1)^2 \ln(1/\alpha)}, & \alpha \in (0,1), \\ Z_1^p, & \alpha = 0, \end{cases}$$

where $\alpha := \liminf_{\varepsilon \to 0} \alpha_\varepsilon$.

3.2. *The general case.* We now come to the construction of an optimal RNF ρ_ε^* and an α-adaptive estimator θ_ε^* w.r.t. the whole family of sets $\Sigma(\mathbf{i}_s)$ for all $\mathbf{i}_s$ and $1 \leq s \leq d-1$. Introduce

$$\mathcal{T}_d = \bigl\{\mathbf{i},\ \mathbf{i} = (i_1, \ldots, i_s),\ 1 \leq i_1 < \cdots < i_s \leq d,\ 1 \leq s \leq d-1\bigr\}.$$

Consider the family of pairs $\{(\rho_\varepsilon^*(\mathbf{i}), \theta_\varepsilon^*(\mathbf{i})),\ \mathbf{i} \in \mathcal{T}_d\}$. According to Section 2, we construct a new pair $(\rho_\varepsilon^*, \theta_\varepsilon^*)$ as follows. Let $\mathbf{i}^* \in \mathcal{T}_d$ be defined by

$$\rho_\varepsilon^*(\mathbf{i}^*) = \inf_{\mathbf{i} \in \mathcal{T}_d} \rho_\varepsilon^*(\mathbf{i}).$$

Put

$$\rho_\varepsilon^* = \rho_\varepsilon^*(\mathbf{i}^*)$$

and

$$\theta_\varepsilon^* = \theta_\varepsilon^*(\mathbf{i}^*).$$

THEOREM 2. *Let* $p \geq 2$, $1/20 \geq \alpha_\varepsilon \geq \varepsilon^a$ *for* $\varepsilon \in (0,1)$ *and some* $a > 0$. *Then* ρ_ε^* *is an* α*-optimal random normalizing factor w.r.t. the family* $\{\Sigma(\mathbf{i}),\ \mathbf{i} \in \mathcal{T}_d\}$ *and* θ_ε^* *is* α*-adaptive. In particular*

$$\limsup_{\varepsilon \to 0} R_\varepsilon^{(r)}(\theta_\varepsilon^*, \Sigma, \rho_\varepsilon^*) = M^*, \tag{27}$$

where

$$M^* \leq \sum_{\mathbf{i} \in \mathcal{T}_d} M^*(\mathbf{i}). \tag{28}$$

If in addition $\alpha := \liminf_{\varepsilon \to 0} \alpha_\varepsilon = 0$, *then*

$$M^* \leq \sup_{\mathbf{i} \in \mathcal{T}_d} M^*(\mathbf{i}). \tag{29}$$

REMARK 7. As we see, the improvement given by the optimal random normalizing factor ρ_ε^* does exist: all its values are essentially better (by some polynomial order) than the minimax rate of convergence over Σ (except of course the one corresponding to the MRC over Σ). In some cases, they even coincide with the MRC over some hypothesis sets $\Sigma(\mathbf{i})$.

REMARK 8. In general, the constant M^* grows in d approximately as 2^d, except for the case $\alpha = 0$. However, if $\alpha > 0$, it is possible to prove that M^* grows no faster than d. The proof is obtained similarly to the case $\alpha = 0$, and only involves more technicalities so we omit it.

3.3. *Adaptive estimation.* Proposition 3 in Section 2.3 gives us a sufficient condition for an α-adaptive estimator to be adaptive w.r.t. the family $\{\Sigma(\mathbf{i}),\ \mathbf{i} \in \mathcal{T}_d\}$ in the classical sense (18). We thus have the following corollary.

COROLLARY 2. *Let* $\alpha_\varepsilon = \mathcal{O}(\{\varepsilon^{2\beta^*/2\beta^*+d}\}^p)$ *as* $\varepsilon \to 0$, *where*

$$\beta^* = \sup_{1 \le i \le d} \beta_i.$$

Then θ_ε^* *is adaptive w.r.t. the family of sets* $\{\Sigma(\mathbf{i}),\ \mathbf{i} \in \mathcal{T}_d\}$, *that is*, $\forall \mathbf{i} \in \mathcal{T}_d$,

$$\limsup_{\varepsilon \to 0} \sup_{\theta \in \Sigma(\mathbf{i})} E_\theta^\varepsilon \{(\varphi_\varepsilon(\Sigma(\mathbf{i}))^{-1} \|\theta_\varepsilon^* - \theta\|)^p\} < \infty,$$

where $\varphi_\varepsilon(\Sigma(\mathbf{i})) = \varepsilon^{2\beta(\mathbf{i})/(2\beta(\mathbf{i})+1)}$.

4. Proofs.

4.1. *Proof of Theorem* 1. Let us begin with some technical lemmas

LEMMA 1. *For* $p \ge 2$, *the following upper bound holds*:

$$\limsup_{\varepsilon \to 0} \sup_{\theta \in \Sigma} E_\theta^\varepsilon \{\varphi_\varepsilon^{-p}(\Sigma) \|\hat{\theta}_\varepsilon - \theta\|^p\} \le Z_1^p.$$

LEMMA 2. *Let* $(\xi_i, 1 \le i \le N)$ *be a sequence of i.i.d. random variables*, *with* $\xi_1 \sim \mathcal{N}(0,1)$. *Put*

$$S_N = \sum_{i=1}^N (\xi_i^2 - 1).$$

Then, *for all* $N \ge 1$, $C_N \in (0, \infty)$ *and* $x \in [0, NC_N]$,

$$P\{S_N \ge x\} \le \exp\left\{-\frac{x^2}{4N(1+C_N)}\right\},$$

$$P\{S_N \le -x\} \le \exp\left\{-\frac{x^2}{4N(1+C_N)}\right\}.$$

If moreover $C_N = o(N^{-1/3})$ *as* $N \to \infty$ *the statement remains true if we replace the right-hand side of the inequality by*

$$\exp\left\{-\frac{x^2}{4N}\right\}(1 + o(1))$$

as $N \to \infty$, *uniformly in* $x \in [0, NC_N]$.

Lemma 1 is known (see for instance [40]). The proof of Lemma 2 is based on some results of [43] and can be found in [29].

We divide the proof of Theorem 1 into two steps. We prove the upper bound (26). Next, we prove a lower bound, namely that ρ_ε^* cannot be improved in the sense of Definition 1. Since $\mathbf{i}_s$ is fixed in this sequel, we will omit explicit reference to $\mathbf{i}_s$ when no confusion is possible. Likewise, we will omit the superscript ε in the definition of P_θ^ε and E_θ^ε.

4.1.1. *Upper bound.* Let us prove inequality (26). Put, for $\theta \in \Theta$,

$$\begin{aligned} R_\varepsilon^{(1)}(\theta) &= E_\theta\{(\rho_\varepsilon^*)^{-p}\|\theta_\varepsilon^* - \theta\|^p 1_{\mathcal{A}_\varepsilon}\} \\ &= E_\theta\{\varphi_\varepsilon^{-p}(\alpha_\varepsilon)\|\hat{\theta}_\varepsilon^{(0)} - \theta\|^p 1_{\mathcal{A}_\varepsilon}\} \end{aligned}$$

and

$$\begin{aligned} R_\varepsilon^{(2)}(\theta) &= E_\theta\{(\rho_\varepsilon^*)^{-p}\|\theta_\varepsilon^* - \theta\|^p 1_{\mathcal{A}_\varepsilon^c}\} \\ &= E_\theta\{\varphi_\varepsilon^{-p}(\Sigma)\|\hat{\theta}_\varepsilon - \theta\|^p 1_{\mathcal{A}_\varepsilon^c}\}. \end{aligned}$$

Obviously

$$R_\varepsilon^{(r)}(\theta_\varepsilon^*, \Sigma, \rho_\varepsilon^*) \le \sup_{\theta\in\Sigma} R_\varepsilon^{(1)}(\theta) + \sup_{\theta\in\Sigma} R_\varepsilon^{(2)}(\theta).$$

Therefore, it is sufficient to obtain bounds for $R_\varepsilon^{(1)}(\theta)$ and $R_\varepsilon^{(2)}(\theta)$, respectively. Let us first study $R_\varepsilon^{(2)}(\theta)$. Clearly

$$R_\varepsilon^{(2)} = \sup_{\theta\in\Sigma} R_\varepsilon^{(2)}(\theta) \le \sup_{\theta\in\Sigma} E_\theta\{\varphi_\varepsilon^{-p}(\Sigma)\|\hat{\theta}_\varepsilon - \theta\|^p\}.$$

From Lemma 1,

$$E_\theta\{\varphi_\varepsilon^{-p}(\Sigma)\|\hat{\theta}_\varepsilon - \theta\|^p\} \le Z_1^p(1 + o(1))$$

as $\varepsilon \to 0$, uniformly in $\theta \in \Sigma$. Finally

$$\limsup_{\varepsilon\to 0} R_\varepsilon^{(2)} \le Z_1^p. \tag{30}$$

Let us turn to $R_\varepsilon^{(1)}(\theta)$. The following notation will prove to be useful: for any multiindex set $A \subset \mathbf{N}^d$,

$$\tilde{S}(A) = \sum_{\mathbf{k}\in A} (\xi_{\mathbf{k}}^2 - 1).$$

From the definition of $\hat{\theta}_\varepsilon^{(0)}$, it is easily seen that

$$R_\varepsilon^{(1)}(\theta) = \varphi_\varepsilon^{-p}(\alpha_\varepsilon) E_\theta\left\{\left(\sum_{\mathbf{k}\in\mathbf{N}^d\setminus I_\varepsilon} \theta_{\mathbf{k}}^2 + \varepsilon^2\tilde{S}(I_\varepsilon) + \varepsilon^2|I_\varepsilon|\right)^{p/2} 1_{\mathcal{A}_\varepsilon}\right\}. \tag{31}$$

Note that $|I_\varepsilon| = \prod_{j=1}^s K_{i_j}(\varepsilon)$. From the decomposition

$$\mathbf{N}^d \setminus I_\varepsilon = J_\varepsilon \cup (I \setminus I_\varepsilon) \cup (J \setminus J_\varepsilon),$$

where the unions are disjoint, we obtain

$$\sum_{\mathbf{k}\in\mathbf{N}^d\setminus I_\varepsilon} \theta_{\mathbf{k}}^2 = H_\varepsilon(\theta) + S_\varepsilon^{(1)}(\theta) + S_\varepsilon^{(2)}(\theta),$$

where

$$H_\varepsilon(\theta) = \sum_{\mathbf{k}\in J_\varepsilon} \theta_{\mathbf{k}}^2,$$

$$S_\varepsilon^{(1)}(\theta) = \sum_{\mathbf{k}\in I\setminus I_\varepsilon} \theta_{\mathbf{k}}^2,$$

$$S_\varepsilon^{(2)}(\theta) = \sum_{\mathbf{k}\in J\setminus J_\varepsilon} \theta_{\mathbf{k}}^2.$$

Now, from the definition (23) of the ellipsoid Σ,

$$S_\varepsilon^{(1)}(\theta) \le L^2 \sum_{j=1}^s K_{i_j}^{-2\beta_{i_j}}(\varepsilon)\bigl(1+o(1)\bigr) \tag{32}$$

and

$$S_\varepsilon^{(2)}(\theta) \le L^2 \sum_{i=1}^d N_i^{-2\beta_i}(\varepsilon)\bigl(1+o(1)\bigr) \tag{33}$$

as $\varepsilon \to 0$, uniformly in $\theta \in \Sigma$. Note that the sum $\varepsilon^2 \tilde{S}(I_\varepsilon)$ satisfies the assumptions of Lemma 2; hence

$$\sup_{\theta\in\Sigma} P_\theta\bigl\{\varepsilon^2 \tilde{S}(I_\varepsilon) > \varepsilon^{\mu_1}\bigr\} \le \exp\{-\varepsilon^{-\mu_2}\}\bigl(1+o(1)\bigr) \tag{34}$$

as $\varepsilon \to 0$ for some positive μ_1 and μ_2 which only depend on L and B. Set

$$\mathcal{V}_\varepsilon = L^2\left(\sum_{j=1}^s K_{i_j}^{-2\beta_{i_j}}(\varepsilon) + \sum_{i=1}^d N_i^{-2\beta_i}(\varepsilon)\right) + \varepsilon^2 \prod_{j=1}^s K_{i_j}(\varepsilon)$$

and define

$$\bar{R}_\varepsilon^{(1)}(\theta) = \varphi_\varepsilon(\alpha_\varepsilon)^{-p}\bigl(H_\varepsilon(\theta) + \mathcal{V}_\varepsilon\bigr)^{p/2} P_\theta\bigl\{T_\varepsilon \le \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\bigr\}.$$

In view of (32), (33) and (34) we can state that

$$\limsup_{\varepsilon\to 0} \sup_{\theta\in\Sigma} R_\varepsilon^{(1)} \le \limsup_{\varepsilon\to 0} \sup_{\theta\in\Sigma} \bar{R}_\varepsilon^{(1)}.$$

Therefore, only an upper bound on $\bar{R}_\varepsilon^{(1)}$ is needed. Fix some $\delta > 0$, assumed to be small, and define

$$\Theta_{\delta,\varepsilon} = \left\{\theta \in \Sigma : H_\varepsilon(\theta) \le \frac{1+\delta}{1-\delta}\lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\right\}.$$

Set

$$\bar{R}_\varepsilon^{(1,1)} = \sup_{\theta\in\Theta_{\delta,\varepsilon}} \bar{R}_\varepsilon^{(1)}(\theta),$$

$$\bar{R}_\varepsilon^{(1,2)} = \sup_{\theta\in\Sigma\setminus\Theta_{\delta,\varepsilon}} \bar{R}_\varepsilon^{(1)}(\theta).$$

Obviously

$$\bar{R}_\varepsilon^{(1,1)} \le \left(\frac{1+\delta}{1-\delta}\lambda^2 + \varphi_\varepsilon(\alpha_\varepsilon)_\varepsilon^{-2}\mathcal{V}_\varepsilon\right)^{p/2}.$$

Letting $\varepsilon \to 0$ and using the definitions of $\varphi_\varepsilon(\alpha_\varepsilon)$ and $\mathcal{V}_\varepsilon$, we obtain, for all $\delta > 0$,

$$\limsup_{\varepsilon\to 0} \bar{R}_\varepsilon^{(1,1)} \le \left(\frac{1+\delta}{1-\delta}\lambda^2 + Z_1^2(\mathbf{i}_s) + Z_2^2\right)^{p/2}$$

and, letting $\delta \to 0$, we finally obtain

$$\limsup_{\varepsilon\to 0} \bar{R}_\varepsilon^{(1,1)} \le \left(\lambda^2 + Z_1^2(\mathbf{i}_s) + Z_2^2\right)^{p/2}. \tag{35}$$

Let us now turn to $\bar{R}_\varepsilon^{(1,2)}$. We plan to use the following decomposition:

$$T_\varepsilon = H_\varepsilon(\theta) + \varepsilon^2 \tilde{S}(J_\varepsilon) + \eta_\varepsilon(\theta),$$

where

$$\eta_\varepsilon(\theta) = 2\varepsilon \sum_{\mathbf{k}\in J_\varepsilon} \theta_{\mathbf{k}}\xi_{\mathbf{k}}.$$

We first show that the term $\eta_\varepsilon(\theta)$ can be neglected. We have

$$\eta_\varepsilon(\theta) \sim \mathcal{N}\left(0, 4\varepsilon^2 H_\varepsilon(\theta)\right).$$

It follows that

$$P_\theta\left\{|\eta_\varepsilon(\theta)| \ge \delta H_\varepsilon(\theta)\right\} \le 2\exp\left\{-\frac{1}{2}\frac{\delta^2}{\varepsilon^2}H_\varepsilon(\theta)\right\}, \tag{36}$$

where we have used the classical bound $P(|\zeta| \ge t) \le 2\exp(-t^2/2)$ if $\zeta \sim \mathcal{N}(0,1)$. Moreover, for $\theta \in \Sigma \setminus \Theta_{\delta,\varepsilon}$ we have $H_\varepsilon(\theta) > \frac{1+\delta}{1-\delta}\lambda^2\varphi_\varepsilon(\alpha_\varepsilon)^2$. This, together with (36) yields

$$P_\theta\left\{|\eta_\varepsilon(\theta)| \ge \delta H_\varepsilon(\theta)\right\} \le 2\exp\left\{-\varepsilon^{-\mu_3}\right\}\left(1 + o(1)\right)$$

as $\varepsilon \to 0$, uniformly in $\theta \in \Sigma \setminus \Theta_{\delta,\varepsilon}$, for some positive μ_3. Thus, for $\theta \in \Sigma \setminus \Theta_{\delta,\varepsilon}$,

$$P_\theta\{T_\varepsilon \le \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\} \le P_\theta\{H_\varepsilon(\theta)(1-\delta) + \varepsilon^2 \tilde{S}(J_\varepsilon) \le \lambda^2 \varphi_\varepsilon^2(\alpha_\varepsilon)\} + \exp\{-\varepsilon^{-\mu_3}\}(1+o(1)).$$

Hence

$$\bar{R}_\varepsilon^{(1,2)} \le F_\varepsilon(\theta)(1+o(1)),$$

where

$$F_\varepsilon(\theta) = \big(\varphi_\varepsilon(\alpha_\varepsilon)^{-2} H_\varepsilon(\theta) + Z_1(\mathbf{i}_s)^2 + Z_2^2\big)^{p/2} \times P_\theta\big(H_\varepsilon(\theta)(1-\delta) + \varepsilon^2 \tilde{S}(J_\varepsilon) \le \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\big).$$

Let $b > 0$ be some constant to be specified below. Let us introduce the set

$$\Theta_b = \left\{\theta \in \Sigma : H_\varepsilon(\theta) \ge \left(\sqrt{b \ln \frac{1}{\varepsilon}} + 1\right)^2 \lambda \varphi_\varepsilon(\alpha_\varepsilon)^2\right\}.$$

Then, in view of Lemma 2,

$$\begin{aligned} &P_\theta\{H_\varepsilon(\theta)(1_\delta) + \varepsilon^2 \tilde{S}_\varepsilon(J_\varepsilon) \le \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\} \\ &\le P_\theta\left\{\tilde{S}_\varepsilon(J_\varepsilon) \le -\frac{\lambda^2 \sqrt{\ln(1/\varepsilon)} \varphi_\varepsilon(\alpha_\varepsilon)}{\varepsilon^2}\right\} \\ &\le \exp\left\{-\frac{b \ln(1/\varepsilon) \varphi_\varepsilon(\alpha_\varepsilon)^4 \lambda^4}{\varepsilon^4 |J_\varepsilon|}\right\} \le \exp\left\{-b \ln \frac{1}{\varepsilon} \ln \frac{1}{\alpha_\varepsilon}\right\} \le \varepsilon^b \end{aligned} \tag{37}$$

as $\varepsilon \to 0$. Now, take b such that

$$\varphi_\varepsilon(\alpha_\varepsilon)^{-p} \varepsilon^b \to 0 \tag{38}$$

as $\varepsilon \to 0$, a choice which is obviously possible since $\varphi_\varepsilon(\alpha_\varepsilon)^{-p} \le \varepsilon^{-2\beta p/(1+2\beta)}$. Recall also that we consider parameters $\theta \in \Sigma$ satisfying

$$\sum_{(i_1,\dots,i_d) \in \mathbf{N}^d} \theta_{i_1 \cdots i_d}^2 \le L^2;$$

therefore $H_\varepsilon(\theta) \le L^2$. Keeping this in mind, we obtain the following from (37) and (38):

$$\sup_{\theta \in \Theta_b} F_\varepsilon(\theta) \le \big(L^2 + Z_1^2(\mathbf{i}_s) + Z_2^2\big)^{p/2} \varphi_\varepsilon(\alpha_\varepsilon)^{-p} \varepsilon^b \to 0 \tag{39}$$

as $\varepsilon \to 0$. For each x such that $1 + \delta \le x \le \sqrt{b \ln \frac{1}{\varepsilon}} + 1$, let us introduce the set $\Theta(x)$ as

$$\Theta(x) = \left\{\theta \in \Sigma : \frac{H_\varepsilon(\theta)(1-\delta)}{\lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2} = x\right\}.$$

Then

$$\Theta_{\delta,\varepsilon} \setminus \Theta_b = \bigcup_{x\in[1+\delta,\sqrt{b\ln(1/\varepsilon)}+1]} \Theta(x). \tag{40}$$

Put $R_\varepsilon(x) := \sup_{\theta\in\Theta(x)} F_\varepsilon(\theta)$. Successively

$$\begin{aligned} R_\varepsilon(x) &\le \left(\frac{\lambda^2}{1-\delta}x + Z_1^2(\mathbf{i}_s) + Z_2^2\right)^{p/2} P_\theta\left\{\tilde{S}(J_\varepsilon) \le -\frac{\lambda^2\varphi_\varepsilon(\alpha_\varepsilon)^2(x-1)^2}{\varepsilon^2}\right\} \\ &\le \left(\frac{\lambda^2}{1-\delta}x + Z_1^2(\mathbf{i}_s) + Z_2^2\right)^{p/2} \exp\left\{-\frac{\lambda^4\varphi_\varepsilon(\alpha_\varepsilon)^4(x-1)^2}{4\varepsilon^4|J_\varepsilon|}\right\} \qquad (41) \\ &\le \left(\frac{\lambda^2}{1-\delta}x + Z_1^2(\mathbf{i}_s) + Z_2^2\right)^{p/2} \exp\left\{-(x-1)^2\ln\frac{1}{\alpha_\varepsilon}\right\}. \end{aligned}$$

From (39), (40) and (41), we derive

$$\begin{aligned} \sup_{\theta\in\Theta_{\delta,\varepsilon}} F_\varepsilon(\theta) \le \sup_{x\in[1+\delta,\sqrt{b\ln(1/\varepsilon)}+1]} &\left\{\left(\frac{\lambda^2}{1-\delta}x + Z_1^2(\mathbf{i}_s) + Z_2^2\right)^{p/2}\right. \\ &\left.\times \exp\left(-(x-1)^2\ln\frac{1}{\alpha_\varepsilon}\right)\right\}. \end{aligned} \tag{42}$$

Therefore, from (35) and (42), we derive

$$\begin{aligned} &\limsup_{\varepsilon\to 0}\sup_{\theta\in\Sigma} F_\varepsilon(\theta) \\ &\quad\le \sup_{x\ge 1}\left(\lambda^2 x + Z_1^2(\mathbf{i}_s) + Z_2^2\right)^{p/2} \exp\left\{-(x-1)^2\ln\frac{1}{\alpha}\right\}, \end{aligned} \tag{43}$$

where $\alpha = \liminf_{\varepsilon\to 0}\alpha_\varepsilon \ge 0$. Finally putting together (43) and (30), we obtain the statement of Theorem 1.

4.1.2. *Lower bound.* First, let us note that for the directions $\mathbf{i}_s$ such that $\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s) = \varepsilon^{2\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)}$, in other words the improvement coincides with the minimax rate of convergence on $\Sigma(\mathbf{i}_s)$, the required lower bound follows from Proposition 1 in [29]. Therefore only the case of the directions such that $\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s) > \varepsilon^{2\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)}$ needs to be studied.

Let ρ_ε be an arbitrary RNF in Ω_ε for which

$$\frac{x(\rho_\varepsilon)}{x(\hat{\rho}_\varepsilon)} \to 0 \qquad \text{as } \varepsilon\to 0. \tag{44}$$

We need to prove that

$$\liminf_{\varepsilon\to 0}\inf_{\tilde{\theta}_\varepsilon} R_\varepsilon\left(\tilde{\theta}_\varepsilon, \Sigma, \rho_\varepsilon\right) = +\infty, \tag{45}$$

where the infimum is taken over all estimators. Define

$$\tilde{J}_\varepsilon = \left\{(k_1,\dots,k_d)\in\mathbf{N}^d : k_i \le N_i(\varepsilon),\ i=1,\dots,d\right\}.$$

Put $V = \{-1, +1\}^{|\tilde{J}_\varepsilon|}$. Thus, every $v \in V$ can be written as

$$v = (v_{\mathbf{k}})_{\mathbf{k} \in \tilde{J}_\varepsilon}, \qquad v_{\mathbf{k}} = \pm 1.$$

Let

$$\psi_\varepsilon = L\left(\sum_{i=1}^{d} N_i(\varepsilon)^{2\beta_i}\right)^{-1/2} \left(\prod_{i=1}^{d} N_i(\varepsilon)\right)^{-1/2}$$

and note that

$$\psi_\varepsilon^2 \asymp \left(\varepsilon^4 \ln \frac{1}{\alpha_\varepsilon}\right)^{(2\beta+1)/(4\beta+1)}.$$

We consider the family $\mathcal{U}_{\tilde{J}_\varepsilon}$ of size $2^{|\tilde{J}_\varepsilon|}$ of sequences $\theta(v)$, $v \in V$, indexed by $\mathbf{N}^d$ and defined as follows:

$$\theta(v)_{\mathbf{k}} = \begin{cases} \psi_\varepsilon v_{\mathbf{k}}, & \text{if } \mathbf{k} \in \tilde{J}_\varepsilon, \\ 0, & \text{otherwise.} \end{cases}$$

We claim that

$$\mathcal{U}_{\tilde{J}_\varepsilon} \subset \Sigma. \tag{46}$$

To show (46), it is enough to check, for any $\theta = \theta(v) \in \mathcal{U}_{\tilde{J}_\varepsilon}$, the following inequality:

$$A(\theta) = \sum_{j=1}^{d} \sum_{(i_1,\dots,i_d) \in \mathbf{N}^d} \theta^2_{(i_1,\dots,i_d)} i_j^{2\beta_j} \le L^2.$$

Let $\theta \in \mathcal{U}_{\tilde{J}_\varepsilon}$. Then

$$\begin{aligned} A(\theta) &= \psi_\varepsilon^2 \sum_{j=1}^{d} \sum_{(i_1,\dots,i_d) \in \tilde{J}_\varepsilon} i_j^{2\beta_j} \\ &= \psi_\varepsilon^2 \sum_{j=1}^{d} \prod_{i=1, i \ne j}^{d} N_i(\varepsilon) N_j(\varepsilon)^{2\beta_j+1} \\ &= \psi_\varepsilon^2 \left(\sum_{i=1}^{d} N_i(\varepsilon)^{2\beta_i}\right)\left(\prod_{i=1}^{d} N_i(\varepsilon)\right) \le L^2. \end{aligned}$$

In the following, we will denote by P_0 the probability law of the observation X_ε if the true parameter is $\theta = 0$. If $\theta = \theta(v) \in \mathcal{U}_{\tilde{J}_\varepsilon}$, we will use the abbreviation P_v for $P_{\theta(v)}$.

Define

$$V_{\mathbf{k}}^{(1)} = \{v \in V : v_{\mathbf{k}} = +1\}, \qquad V_{\mathbf{k}}^{(-1)} = \{v \in V : v_{\mathbf{k}} = -1\},$$
$$V_{\mathbf{k}}^{(0)} = \big\{v = (v_{\mathbf{l}}),\ \mathbf{l} \in \tilde{J}_\varepsilon : v_{\mathbf{l}} = \pm 1 \text{ if } \mathbf{l} \ne \mathbf{k} \text{ and } v_{\mathbf{k}} = 0\big\}.$$

Let

$$\mathcal{B}_\varepsilon = \{\rho_\varepsilon = x_\varepsilon(\rho_\varepsilon)\}.$$

We will prove that there exists some absolute constant $p_0 > 0$ such that, for all $\mathbf{k} \in \tilde{J}_\varepsilon$,

$$\frac{1}{2^{|\tilde{J}_\varepsilon|-1}} \sum_{v \in V_{\mathbf{k}}^{(0)}} P_v\{\mathcal{B}_\varepsilon\} \geq p_0. \tag{47}$$

Let us first show that (45) follows from (47). Let $\tilde{\theta}_\varepsilon$ be an arbitrary estimator. From the definition of $\mathcal{B}_\varepsilon$,

$$\begin{aligned} R_\varepsilon(\tilde{\theta}_\varepsilon, \Sigma, \rho_\varepsilon) &\geq \sup_{\theta \in \Sigma} E_\theta\{\rho_\varepsilon^{-p} \|\tilde{\theta}_\varepsilon - \theta\|^p 1_{\mathcal{B}_\varepsilon}\} \\ &\geq \sup_{\theta(v),\ v \in V} E_\theta\Big\{\big(x_\varepsilon(\rho_\varepsilon)^{-1} \|\tilde{\theta}_\varepsilon - \theta\|\big)^p 1_{\mathcal{B}_\varepsilon}\Big\} \\ &\geq \left(\frac{1}{2^{|\tilde{J}_\varepsilon|}} \sum_{v \in V} E_v\{x_\varepsilon(\rho_\varepsilon)^{-1} \|\tilde{\theta}_\varepsilon - \theta\|^2 1_{\mathcal{B}_\varepsilon}\}\right)^{p/2} =: \big(R_\varepsilon(\tilde{\theta}_\varepsilon)\big)^{p/2}. \end{aligned}$$

Here we used Jensen's inequality: $E\{|Z|^q\} \geq (E\{|Z|\})^q$ if $q \geq 1$. Note that from the definition of $\tilde{J}_\varepsilon$, $\theta_{\mathbf{k}} = 0$ if $\mathbf{k} \in \mathbf{N}^d \setminus \tilde{J}_\varepsilon$ for any $\theta \in \mathcal{U}_{\tilde{J}_\varepsilon}$. Using that

$$\|\tilde{\theta}_\varepsilon - \theta\|^2 \geq \sum_{\mathbf{k} \in \tilde{J}_\varepsilon} \big(\tilde{\theta}_{\varepsilon,\mathbf{k}} - \theta_{\mathbf{k}}\big)^2$$

we have

$$\begin{aligned} R_\varepsilon(\tilde{\theta}_\varepsilon) &\geq \frac{x_\varepsilon(\rho_\varepsilon)^{-2}}{2^{|\tilde{J}_\varepsilon|}} \sum_{v \in V} \sum_{\mathbf{k} \in \tilde{J}_\varepsilon} E_v\Big\{\big(\tilde{\theta}_{\mathbf{k}} - \theta_{\mathbf{k}}(v)\big)^2 1_{\mathcal{B}_\varepsilon}\Big\} \\ &= \frac{x_\varepsilon(\rho_\varepsilon)^{-2}}{2^{|J_\varepsilon|}} \sum_{\mathbf{k} \in \tilde{J}_\varepsilon} \Bigg(\sum_{v \in V_{\mathbf{k}}^{(+1)}} E_v\Big\{\big(\tilde{\theta}_{\varepsilon,\mathbf{k}} - \psi_\varepsilon\big)^2 1_{\mathcal{B}_\varepsilon}\Big\} \\ &\qquad + \sum_{v \in V_{\mathbf{k}}^{(-1)}} E_v\Big\{\big(\tilde{\theta}_{\varepsilon,\mathbf{k}} + \psi_\varepsilon\big)^2 1_{\mathcal{B}_\varepsilon}\Big\}\Bigg). \end{aligned}$$

For $v \in V$, let $\check{v}^{\mathbf{k}} = (\check{v}^{\mathbf{k}})_{\mathbf{l}}$ be defined, for $\mathbf{l} \in \mathbf{N}^d$, by

$$\check{v}^{\mathbf{k}} = \begin{cases} v_{\mathbf{l}}, & \text{if } \mathbf{l} \neq \mathbf{k}, \\ 0, & \text{otherwise.} \end{cases}$$

We need to introduce the following likelihood ratios:

$$Z_{\mathbf{k}}^{(1)} = \frac{dP_v}{dP_{\check{v}^{\mathbf{k}}}}(Y) \quad \text{for } v \in V_{\mathbf{k}}^{(1)},$$
$$Z_{\mathbf{k}}^{(-1)} = \frac{dP_v}{dP_{\check{v}^{\mathbf{k}}}}(Y) \quad \text{for } v \in V_{\mathbf{k}}^{(-1)},$$

where $Y = (Y_{\mathbf{k}})$ is the observation process. Note that, under $P_{\check{v}^{\mathbf{k}}}$,

$$Z_{\mathbf{k}}^{(1)} = \exp\Big\{\frac{\psi_\varepsilon}{\varepsilon}\xi_{\mathbf{k}}\theta_{\mathbf{k}} - \frac{\psi_\varepsilon^2}{2\varepsilon^2}\theta_{\mathbf{k}}^2\Big\},$$
$$Z_{\mathbf{k}}^{(-1)} = \exp\Big\{-\frac{\psi_\varepsilon}{\varepsilon}\xi_{\mathbf{k}}\theta_{\mathbf{k}} - \frac{\psi_\varepsilon^2}{2\varepsilon^2}\theta_{\mathbf{k}}^2\Big\}$$

and that the distribution of $Z_{\mathbf{k}}^{(1)}$ and $Z_{\mathbf{k}}^{(1)}$ does not depend on $v_{\mathbf{l}}$, $\mathbf{l} \neq \mathbf{k}$, and ε. Moreover $\varepsilon^{-1}\psi_\varepsilon \to 0$ as $\varepsilon \to 0$. Hence, for all $\check{v}^{\mathbf{k}}$ and small enough ε and $\delta > 0$, the following inequality holds:

$$P_{\check{v}^{\mathbf{k}}}\big\{\big(Z_{\mathbf{k}}^{(1)} < 1-\delta\big) \cup \big(Z_{\mathbf{k}}^{(-1)} < 1-\delta\big)\big\} \le \delta. \tag{48}$$

Define

$$D_{\mathbf{k}} = \big\{Z_{\mathbf{k}}^{(1)} \ge 1-\delta\big\} \cap \big\{Z_{\mathbf{k}}^{(-1)} \ge 1-\delta\big\}.$$

It follows that

$$R_\varepsilon(\tilde\theta_\varepsilon) \ge \frac{x_\varepsilon(\rho_\varepsilon)^{-2}}{2^{|\tilde J_\varepsilon|}} \sum_{\mathbf{k}\in\tilde J_\varepsilon} \sum_{v\in V_{\mathbf{k}}^{(0)}} E_{\check{v}^{\mathbf{k}}}\big\{\big(Z_{\mathbf{k}}^{(1)}(\tilde\theta_{\varepsilon,\mathbf{k}} - \psi_\varepsilon)^2 + Z_{\mathbf{k}}^{(-1)}(\tilde\theta_{\varepsilon,\mathbf{k}} + \psi_\varepsilon)^2\big)1_{\mathcal{B}_\varepsilon}\big\}$$
$$\ge (1-\delta)\frac{x_\varepsilon(\rho_\varepsilon)^{-2}}{2^{|\tilde J_\varepsilon|}} \sum_{v\in V_{\mathbf{k}}^{(0)}} \psi_\varepsilon^2 P_{\check{v}^{\mathbf{k}}}\{\mathcal{B}_\varepsilon \cap D_{\mathbf{k}}\}$$

where we used that

$$(\tilde\theta_{\varepsilon,\mathbf{k}} - \psi_\varepsilon)^2 + (\tilde\theta_{\varepsilon,\mathbf{k}} + \psi_\varepsilon)^2 \ge 2\psi_\varepsilon^2.$$

Applying inequality (48) yields, for small enough ε and δ,

$$P_{\check{v}^{\mathbf{k}}}\{\mathcal{B}_\varepsilon \cap D_{\mathbf{k}}\} \ge P_{\check{v}^{\mathbf{k}}}\{\mathcal{B}_\varepsilon\} - \delta.$$

Hence, choosing $\delta < p_0/2$ we finally obtain

$$R_\varepsilon(\tilde\theta_\varepsilon) \ge (1-\delta)x_\varepsilon^{-2}(\rho_\varepsilon)\, p_0 \psi_\varepsilon^2 |J_\varepsilon|.$$

From the choice of ψ_ε and $|\tilde J_\varepsilon|$ we conclude that

$$\inf_{\tilde\theta_\varepsilon} R_\varepsilon(\tilde\theta_\varepsilon) \ge (1-\delta)\, p_0 \Big(\frac{x_\varepsilon(\rho_\varepsilon)}{\varphi_\varepsilon(\alpha_\varepsilon)}\Big)^{-2}.$$

It remains to show that

$$x_\varepsilon(\rho_\varepsilon)/\varphi_\varepsilon(\alpha_\varepsilon) \to 0 \quad \text{as } \varepsilon \to 0. \tag{49}$$

This will follow from (44) and

$$x_\varepsilon(\hat{\rho}_\varepsilon) \le \varphi_\varepsilon(\alpha_\varepsilon), \tag{50}$$

where

$$x_\varepsilon(\hat{\rho}_\varepsilon) = \inf\Big\{x \in (0, \varphi_\varepsilon(\Sigma)] : \inf_{\theta \in \Sigma(\mathbf{i}_s)} P_\theta\{\hat{\rho}_\varepsilon \le x\} \ge 1 - \alpha_\varepsilon\Big\}.$$

We will show that

$$\limsup_{\varepsilon \to 0} \alpha_\varepsilon^{-1} \sup_{\theta \in \Sigma(\mathbf{i}_s)} P_\theta\big\{\hat{\rho}_\varepsilon = \varphi_\varepsilon(\alpha_\varepsilon)\big\} \le 1.$$

From the definition of $\hat{\rho}_\varepsilon$, it is enough to prove that

$$\sup_{\theta \in \Sigma(\mathbf{i}_s)} P_\theta\big\{T_\varepsilon > \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\big\} \le \alpha_\varepsilon. \tag{51}$$

Under P_θ, $\theta \in \Sigma(\mathbf{i}_s)$, we have

$$T_\varepsilon = \varepsilon^2 \sum_{\mathbf{k} \in J_\varepsilon} (\xi_{\mathbf{k}}^2 - 1).$$

On the other side, from the definition of λ, N_ε and $\varphi_\varepsilon(\alpha_\varepsilon)$, we readily check that

$$\varepsilon^{-2} \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2 \ge 2\left(\ln \frac{1}{\alpha_\varepsilon}\right)^{1/2} \prod_{i=1}^d N_i^{1/2}(\varepsilon) =: z_\varepsilon.$$

It follows that

$$P_\theta\big\{T_\varepsilon > \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\big\} \le P_\theta\left\{\sum_{\mathbf{k} \in J_\varepsilon} (\xi_{\mathbf{k}}^2 - 1) > z_\varepsilon\right\} \tag{52}$$

uniformly in $\theta \in \Sigma(\mathbf{i}_s)$. Moreover

$$z_\varepsilon \le 2\left(a \ln \frac{1}{\varepsilon}\right)^{1/2} \prod_{i=1}^d N_i^{1/2}(\varepsilon)$$

from the assumption $\alpha_\varepsilon \ge \varepsilon^a$. Note that

$$|J_\varepsilon| = \prod_{i=1}^d N_i(\varepsilon)\big(1 + o(1)\big)$$

as $\varepsilon \to 0$. Therefore, we can choose C_ε such that $C_\varepsilon = o(|J_\varepsilon|^{-1/3})$ as $\varepsilon \to 0$ and $0 < z_\varepsilon \le C_\varepsilon \prod_{i=1}^d N_i(\varepsilon)$. Applying Lemma 2 together with (52) yields

$$P_\theta\big\{T_\varepsilon > \lambda^2 \varphi_\varepsilon(\alpha_\varepsilon)^2\big\} \le \alpha_\varepsilon\big(1 + o(1)\big)$$

as $\varepsilon \to 0$, uniformly in $\theta \in \Sigma(\mathbf{i}_s)$. This proves (50).

We complete the proof by proving (47). Since $\rho_\varepsilon \in \Omega_\varepsilon$ and $0 \in \Sigma(\mathbf{i}_s)$, for all $\mathbf{i}_s$, $s = 1, \dots, d-1$, we have, for all $\delta > 0$ and ε sufficiently small,

$$\alpha_\varepsilon^{-1} P_0\{\mathcal{B}_\varepsilon^c\} \le 1 + \delta. \tag{53}$$

Define

$$Z_\varepsilon = \frac{1}{2^{|\tilde{J}_\varepsilon|-1}} \sum_{v \in V_{\mathbf{k}}^{(0)}} \frac{dP_v}{dP_0}(Y)$$

and

$$\mathcal{P}_\varepsilon = \frac{1}{2^{|\tilde{J}_\varepsilon|-1}} \sum_{v \in V_{\mathbf{k}}^{(0)}} P_v\{\mathcal{B}_\varepsilon\}.$$

From (53)

$$\mathcal{P}_\varepsilon = E_0\{Z_\varepsilon 1_{\mathcal{B}_\varepsilon}\} \ge E_0\{Z_\varepsilon 1_{\mathcal{B}_\varepsilon} + c\alpha_\varepsilon^{-1} 1_{\mathcal{B}_\varepsilon^c}\} - c(1+\delta),$$

where $c > 0$ is a constant to be specified below. It is clear that

$$\inf_{\mathcal{F}_\varepsilon}(Z_\varepsilon 1_{\mathcal{F}_\varepsilon} + c\alpha_\varepsilon^{-1} 1_{\mathcal{F}_\varepsilon^c}) = Z_\varepsilon 1_{Z_\varepsilon < c\alpha_\varepsilon^{-1}} + c\alpha_\varepsilon^{-1} 1_{Z_\varepsilon \ge c\alpha_\varepsilon^{-1}}.$$

It follows that

$$\begin{aligned}
\mathcal{P}_\varepsilon &\ge E_0\{Z_\varepsilon 1_{Z_\varepsilon < c\alpha^{-1}}\} - c(1+\delta) \\
&= \frac{1}{2^{|\tilde{J}_\varepsilon|-1}} \sum_{v \in V_{\mathbf{k}}^{(0)}} P_v\{Z_\varepsilon < c\alpha^{-1}\} - c(1+\delta) \\
&= 1 - c(1+\delta) - \frac{1}{2^{|J_\varepsilon|-1}} \sum_{v \in V_{\mathbf{k}}^{(0)}} P_v\{Z_\varepsilon \ge c\alpha^{-1}\}.
\end{aligned}$$

Applying Chebyshev's inequality yields

$$\begin{aligned}
\mathcal{P}_\varepsilon &\ge 1 - c(1+\delta) - \frac{\alpha_\varepsilon}{c 2^{|\tilde{J}_\varepsilon|-1}} \sum_{v \in V_{\mathbf{k}}^{(0)}} E_v\{Z_\varepsilon\} \\
&= 1 - c(1+\delta) - \alpha_\varepsilon c^{-1} E_0\{Z_\varepsilon^2\}.
\end{aligned}$$

We claim that

$$E_0\{Z_\varepsilon^2\} = \alpha_\varepsilon^{-1/2}\big(1 + o(1)\big) \tag{54}$$

as $\varepsilon \to 0$. This shows that for sufficiently small ε and δ, and since $\alpha_\varepsilon \le 1/20$, we have

$$\mathcal{P}_\varepsilon \ge \left(1 - c - \frac{1}{2\sqrt{5}c}\right)\big(1 + o(1)\big) = 1 - \frac{2}{\sqrt{2\sqrt{5}}}\big(1 + o(1)\big) > 0$$

for the choice $c = 1/\sqrt{2\sqrt{5}}$, which proves (47). It remains to show (54). From

$$Z_\varepsilon = \prod_{\mathbf{k}\in\tilde{J}_\varepsilon} \left(\tfrac{1}{2}Z_{\mathbf{k}}^{(1)} + \tfrac{1}{2}Z_{\mathbf{k}}^{(-1)}\right)$$

and since the random variables $\xi_{\mathbf{k}}$ are independent, elementary computation shows that

$$\begin{aligned} E_0\{Z_\varepsilon^2\} &= \left(E_0\left\{\frac{1}{2}\exp\left(\frac{\psi_\varepsilon}{\varepsilon}\xi_{\mathbf{k}} - \frac{1}{2}\frac{\psi_\varepsilon^2}{\varepsilon^2}\right) + \frac{1}{2}\exp\left(-\frac{\psi_\varepsilon}{\varepsilon}\xi_{\mathbf{k}} - \frac{1}{2}\frac{\psi_\varepsilon^2}{\varepsilon^2}\right)\right\}\right)^{|\tilde{J}_\varepsilon|-1} \\ &= \left(\frac{1}{2}\exp\psi_\varepsilon^2 + \frac{1}{2}\exp -\psi_\varepsilon^2\right)^{|\tilde{J}_\varepsilon|-1} \\ &= \left(1 + \psi_\varepsilon^4 + \mathcal{O}\left(\psi_\varepsilon^6\right)\right)^{|\tilde{J}_\varepsilon|-1} \\ &= \exp\left\{|\tilde{J}_\varepsilon|\frac{\psi_\varepsilon^4}{\varepsilon^4}\right\}(1+o(1)) \end{aligned}$$

since $\psi_\varepsilon^6|\tilde{J}_\varepsilon| \to 0$ as $\varepsilon \to 0$. From $\psi_\varepsilon^4|\tilde{J}_\varepsilon| = -\frac{1}{2}\ln\alpha_\varepsilon$, the conclusion follows. This ends the proof of Theorem 1.

4.2. *Proof of Theorem* 2. Since the pair $(\rho_\varepsilon^*, \theta_\varepsilon^*)$ is obtained through the canonical construction (16), Proposition 1 shows the existence of an upper bound M^* described by (28) for the minimax risk and the optimality of ρ_ε^*. Consequently, only (29) needs to be proved. For this, we use Corollary 1. Therefore, it is enough to show that

$$\lim_{\varepsilon\to 0} \sum_{\mathbf{i}\in\mathcal{T}_d} \sup_{\theta\in\Sigma} E_\theta\left\{\zeta_\varepsilon^p(\mathbf{i}) 1_{\zeta_\varepsilon(\mathbf{i}) > M^*(\mathbf{i})}\right\} = 0, \tag{55}$$

where

$$\zeta_\varepsilon(\mathbf{i}) = \hat{\rho}(\mathbf{i})^{-1}\|\theta_\varepsilon^*(\mathbf{i}) - \theta\|.$$

From the Cauchy–Schwarz inequality

$$\begin{aligned} \sum_{\mathbf{i}\in\mathcal{T}_d} E_\theta\left\{\zeta_\varepsilon^p(\mathbf{i}) 1_{\zeta_\varepsilon(\mathbf{i}) > M^*(\mathbf{i})}\right\} &\le \sum_{\mathbf{i}\in\mathcal{T}_d} E_\theta\left\{\zeta_\varepsilon^{2p}(\mathbf{i})\right\}^{1/2} P_\theta\left\{\zeta_\varepsilon(\mathbf{i}) > M^*(\mathbf{i})\right\}^{1/2} \\ &\le M^*|\mathcal{T}_d| \sup_{\mathbf{i}\in\mathcal{T}_d} P_\theta\left\{\zeta_\varepsilon(\mathbf{i}) > M^*(\mathbf{i})\right\}^{1/2}, \end{aligned}$$

where

$$M^* = M_{2p}^* = \sup_{\mathbf{i}\in\mathcal{T}_d} M^*(\mathbf{i})^{1/2}. \tag{56}$$

Here $M^*_{2p}(\mathbf{i})$ means that we take the constant $M^*(\mathbf{i})$ associated with the power $2p$, which is finite since the choice of p is free in Theorem 1. Therefore, it is enough to show that

$$\lim_{\varepsilon\to 0}\sup_{\theta\in\Sigma(\mathbf{i})} P_\theta\{\zeta_\varepsilon(\mathbf{i}) > M^*(\mathbf{i})\} = 0 \qquad \forall \mathbf{i}\in\mathcal{T}_d. \tag{57}$$

Note that

$$P_\theta\{\zeta_\varepsilon(\mathbf{i}) > M^*(\mathbf{i})\} \le P_\theta\{\varphi_\varepsilon^{-1}(\Sigma)\|\hat\theta_\varepsilon - \theta\| \ge Z_1\}$$

so that (57) follows by classical arguments using Lemma 2.

The same arguments show that

$$\sup_{\theta\in\Theta_{\delta,\varepsilon}} P_\theta\{\zeta_\varepsilon(\mathbf{i}) > M^*(\mathbf{i}), \mathcal{A}_\varepsilon\} \le \sup_{\theta\in\Theta_{\delta,\varepsilon}} P_\theta\{\varphi_\varepsilon^{-1}(\alpha_\varepsilon)\|\hat\theta^{(0)} - \theta\| \ge M^*(\mathbf{i})\} \to 0$$

as $\varepsilon\to 0$. From the assumption $\liminf \alpha_\varepsilon = 0$ and in view of (43), we obtain

$$\lim_{\varepsilon\to 0}\sup_{\Theta\setminus\Theta_{\delta,\varepsilon}} P_\theta\{\varphi_\varepsilon(\alpha_\varepsilon)^{-1}\|\hat\theta^{(0)} - \theta\| \ge M^*(\mathbf{i})\} = 0,$$

which completes the proof of Theorem 2.

APPENDIX

A.1. Proof of Proposition 1. For $i = 1,\dots,N$, we set $x^*_\varepsilon(i) = x_\varepsilon(\rho^*_\varepsilon, i)$. Let us first show that, for $i = 1,\dots,N$,

$$x^*_\varepsilon(i) \le \varphi_{\varepsilon,i}(\alpha_\varepsilon). \tag{58}$$

Indeed

$$\begin{aligned}
\inf_{f\in\Sigma_i} P^\varepsilon_f\{\rho^*_\varepsilon \le \varphi_{\varepsilon,i}(\alpha_\varepsilon)\} &= \inf_{f\in\Sigma_i} P^\varepsilon_f\Big\{\inf_{j=1,\dots,N}\rho^*_{\varepsilon,j} \le \varphi_{\varepsilon,i}(\alpha_\varepsilon)\Big\}\\
&\ge \inf_{f\in\Sigma_i} P^\varepsilon_f\{\rho^*_{\varepsilon,i} \le \varphi_{\varepsilon,i}(\alpha_\varepsilon)\}\\
&= \inf_{f\in\Sigma_i} P^\varepsilon_f\{\rho^*_{\varepsilon,i} = \varphi_{\varepsilon,i}(\alpha_\varepsilon)\} \ge 1-\alpha_\varepsilon,
\end{aligned}$$

where we used property $P_1(i)$. From the definition of $x^*_\varepsilon(i)$, we derive (58). Note that, for $f\in\Sigma$,

$$\begin{aligned}
E^\varepsilon_f\{(\rho^*_\varepsilon)^{-p}\|f^*_\varepsilon - f\|^p\} &= \sum_{i=1}^N E^\varepsilon_f\{(\rho^*_{\varepsilon,i})^{-p}\|f^*_{\varepsilon,i} - f\|^p 1_{\{i^*=i\}}\}\\
&\le \sum_{i=1}^N E^\varepsilon_f\{(\rho^*_{\varepsilon,i})^{-p}\|f^*_{\varepsilon,i} - f\|^p\}.
\end{aligned}$$

Therefore

$$\limsup_{\varepsilon\to 0}\sup_{f\in\Sigma} E_f^{\varepsilon}\{(\rho_\varepsilon^*)^{-p}\|f_\varepsilon^* - f\|^p\} \le \sum_{i=1}^{N}(M_i^*)^p < \infty, \tag{59}$$

where we used property $P_2(i)$. Hence the bound (10) of Definition 1 is satisfied.

We now prove (13). Assume on the contrary that ρ_ε^* is not optimal. Then, there exist $\bar\rho_\varepsilon \in \Omega_\varepsilon$, $\bar j \in \{1,\dots,N\}$ and a constant $0 < \bar M < \infty$ such that

$$\frac{x_\varepsilon(\bar\rho_\varepsilon, \bar j)}{x_\varepsilon^*(\bar j)} \to 0 \qquad \text{as } \varepsilon \to 0 \tag{60}$$

and

$$\limsup_{\varepsilon\to 0}\sup_{f\in\Sigma} E_f^{\varepsilon}\{(\bar\rho_\varepsilon)^{-p}\|\bar f_\varepsilon - f\|^p\} \le \bar M \tag{61}$$

for some estimator $\bar f_\varepsilon$. Set

$$\bar\rho_{\varepsilon,\bar j} = \begin{cases} x_\varepsilon(\bar\rho_\varepsilon, \bar j), & \text{if } \bar\rho_\varepsilon \le x_\varepsilon(\bar\rho_\varepsilon, \bar j), \\ \varphi_\varepsilon(\Sigma), & \text{if } \bar\rho_\varepsilon > x_\varepsilon(\bar\rho_\varepsilon, \bar j). \end{cases}$$

First, note that $\bar\rho_{\varepsilon,\bar j} \ge \bar\rho_\varepsilon$ and therefore we have from (61)

$$\limsup_{\varepsilon\to 0}\sup_{f\in\Sigma} E_f^{\varepsilon}\{(\bar\rho_{\varepsilon,\bar j})^{-p}\|\bar f_\varepsilon - f\|^p\} \le \bar M. \tag{62}$$

Moreover

$$\inf_{f\in\Sigma_{\bar j}} P_f^{\varepsilon}\{\bar\rho_{\varepsilon,\bar j} = x_\varepsilon(\bar\rho_\varepsilon, \bar j)\} = \inf_{f\in\Sigma_{\bar j}} P_f^{\varepsilon}\{\bar\rho_\varepsilon \le x_\varepsilon(\bar\rho_\varepsilon, \bar j)\} \ge 1 - \alpha_\varepsilon. \tag{63}$$

From (58) and (60)

$$\frac{x_\varepsilon(\bar\rho_\varepsilon, \bar j)}{\varphi_{\varepsilon,\bar j}(\alpha_\varepsilon)} \le \frac{x_\varepsilon(\bar\rho_\varepsilon, \bar j)}{x_\varepsilon^*(\bar j)} \to 0 \qquad \text{as } \varepsilon \to 0.$$

This, together with (62) and (63), contradicts $P_3(\bar j)$ and shows the optimality of ρ_ε^*. From this point, it remains to note that (59) shows that f_ε^* is α-adaptive, which completes the proof of Proposition 1.

A.2. Proof of Corollary 1. We have, for $f \in \Sigma$,

$$\begin{aligned} &E_f^{\varepsilon}\{(\rho_\varepsilon^*)^{-p}\|f_\varepsilon^* - f\|^p\} \\ &\quad= \sum_{i=1}^{N} E_f^{\varepsilon}\{[\xi_{\varepsilon,i}(f)]^p 1_{i^*=i}\} \\ &\quad\le \sum_{i=1}^{N} M_i^p P_f^{\varepsilon}\{i^* = i\} + \sum_{i=1}^{N} E_f^{\varepsilon}\{[\xi_{\varepsilon,i}(f)]^p 1_{\{\xi_{\varepsilon,i}(f) > M_i, i^*=i\}}\} \end{aligned}$$

$$\leq \Big(\sup_{i=1,\dots,N} M_i \Big)^p + \sum_{i=1}^{N} \sup_{f\in\Sigma} E_f^\varepsilon \big\{ [\xi_{\varepsilon,i}(f)]^p 1_{\{\xi_{\varepsilon,i}(f) > M_i\}} \big\}.$$

Applying (17) completes the proof of Corollary 1.

A.3. Proof of Proposition 3. By definition, Σ is bounded by Q. We may thus assume without loss of generality that $\|f_\varepsilon^*\| \leq 3Q$. Let $i \in \{1,\dots,N\}$. Proposition 3 is equivalent to

$$\limsup_{\varepsilon\to 0} R_\varepsilon\big(f_\varepsilon^*, \Sigma_i, \varphi_\varepsilon(\Sigma_i)\big) < \infty. \tag{64}$$

We have

$$\begin{aligned} R_\varepsilon\big(f_\varepsilon^*, \Sigma_i, \varphi_\varepsilon(\Sigma_i)\big) \leq{}& \sup_{f\in\Sigma_i} E_f^\varepsilon \big\{ \varphi_\varepsilon^{-p}(\Sigma_i) \|f_\varepsilon^* - f\|^p 1_{\{\rho_\varepsilon^* \leq x_\varepsilon(\rho_\varepsilon^*, i)\}} \big\} \\ &+ \sup_{f\in\Sigma_i} E_f^\varepsilon \big\{ \varphi_\varepsilon^{-p}(\Sigma_i) \|f_\varepsilon^* - f\|^p 1_{\{\rho_\varepsilon^* > x_\varepsilon(\rho_\varepsilon^*, i)\}} \big\}. \end{aligned}$$

Hence from the definition of $\hat{f}_{\varepsilon,i}$, we have

$$\begin{aligned} R_\varepsilon\big(f_\varepsilon^*, \Sigma_i, \varphi_\varepsilon(\Sigma_i)\big) \leq{}& R_\varepsilon\big(\hat{f}_{\varepsilon,i}, \Sigma_i, \varphi_\varepsilon(\Sigma_i)\big) \\ &+ (4Q)^p \Big(\inf_{j=1,\dots,N} \varphi_\varepsilon^{-p}(\Sigma_j) \Big) \sup_{f\in\Sigma_i} P_f^\varepsilon \big\{ \rho_\varepsilon^* > x_\varepsilon(\rho_\varepsilon, i) \big\}. \end{aligned}$$

Since $\hat{f}_{\varepsilon,i}$ is asymptotically optimal on Σ_i, the first term on the right-hand side of the last inequality is bounded. Moreover, since ρ_ε^* is α-optimal, we have

$$\limsup_{\varepsilon\to 0} \alpha_\varepsilon^{-1} \sup_{f\in\Sigma_i} P_f^\varepsilon \big\{ \rho_\varepsilon^* > x_\varepsilon(\rho_\varepsilon, i) \big\} \leq 1.$$

Finally, we have that $\alpha_\varepsilon \varphi_\varepsilon^{-p}(\Sigma_i)$ is bounded for all i by assumption. The conclusion (64) follows.

Acknowledgments. The first author is grateful to M. Nussbaum for supporting a two-month visit at WIAS, where this work was initiated. The careful remarks and comments of three referees were valuable to us in improving considerably a former version of the manuscript.

REFERENCES

[1] AKAIKE, H. (1974). A new look at the statistical model identification. *IEEE Trans. Automat. Control* **19** 716–723.

[2] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.

[3] BREIMAN, L. and FREEDMAN, D. (1983). How many variables should be entered in a regression model? *J. Amer. Statist. Assoc.* **78** 131–136.

[4] BROWN, L. D. and LOW, M. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.
[5] CHERNOFF, H. (1956). Large sample theory: parametric case. *Ann. Math. Statist.* **27** 1–22.
[6] CSISZÁR, I. and KÖRNER, J. (1981). *Information Theory: Coding Theorems for Discrete Memoryless Systems*. Akademic, New York.
[7] DELYON, B. and JUDITSKY, A. (1996). On minimax wavelet estimators. *Appl. Comput. Harmon. Anal.* **3** 215–228.
[8] DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.
[9] DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: Asymptopia? *J. Roy. Statist. Soc. Ser. B* **57** 301–369.
[10] DYCHAKOV, A. G. (1971). On a search model of false coins. In *Colloquia Mathematica Societatis Janos Bolyai: Topics in Information Theory*. North-Holland, Amsterdam.
[11] EFROMOVICH, S. YU. (1985). Nonparametric estimation of a density of unknown smoothness. *Theory Probab. Appl.* **30** 524–534.
[12] EFROMOVICH, S. YU. and PINSKER, M. S. (1984). Adaptive algorithms for nonparametric filtering. *Automat. Remote Control* **11** 54–60.
[13] FREIDLINA, V. L. (1975). On one problem of screening experimental design. *Theory Probab. Appl.* **20** 100–114.
[14] GOLDENSHLUGER, A. and NEMIROVSKI, A. (1997). On spatially adaptive estimation of nonparametric regression. *Math. Methods Statist.* **6** 135–170.
[15] GOLUBEV, G. K. (1990). Quasilinear estimates of a signal in L_2. *Problems Inform. Transmission* **26** 15–20.
[16] GRAMA, I. and NUSSBAUM, M. (1998). Asymptotic equivalence for nonparametric generalized linear models. *Probab. Theory Related Fields* **111** 167–214.
[17] HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **3** 1465–1481.
[18] HALL, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* **22** 215–232.
[19] HALL, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694.
[20] HALL, P., KERKYACHARIAN, G. and PICARD, D. (1998). Block threshold rules for curve estimation using kernel and wavelet methods. *Ann. Statist.* **26** 922–942.
[21] HALL, P., KERKYACHARIAN, G. and PICARD, D. (1999). A note on the wavelet oracle. *Statist. Probab. Lett.* **43** 415–420.
[22] HALL, P., KERKYACHARIAN, G. and PICARD, D. (1999). On the minimax optimality of block thresholding wavelet estimators. *Statist. Sinica* **9** 33–49.
[23] IBRAGIMOV, I. A. and KHASMINSKI, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.
[24] IBRAGIMOV, I. A. and KHASMINSKI, R. Z. (1984). More on estimation of the density of a distribution. *J. Soviet. Math.* **25** 1155–1165.
[25] LEPSKI, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.
[26] LEPSKI, O. V. (1991). Asymptotic minimax adaptive estimation. 1. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.
[27] LEPSKI, O. V. (1992). Asymptotic minimax adaptive estimation. 2. Statistical models without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37** 433–448.
[28] LEPSKI, O. V. (1992). On problems of adaptive estimation in white Gaussian noise. In *Advances in Soviet Mathematics* (R. Z. Khasminskii, ed.) **12** 87–106. Amer. Math. Soc., Providence, RI.

[29] LEPSKI, O. V. (1999). How to improve the accuracy of estimation. *Math. Methods Statist.* **8** 441–486.
[30] LEPSKI, O. V. and SPOKOINY, V. G. (1995). Local adaptation to inhomogeneous smoothness: resolution level. *Math. Methods Statist.* **4** 239–258.
[31] LEPSKI, O. V. and SPOKOINY, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546.
[32] 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.
[33] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554.
[34] MALYUTOV, M. B. and TSITOVICH, I. I. (1996). On sequential search for significant variables of unknown function. In *Proceedings of 6th Lukacs Symposium* 155–178. VSP, Utrecht.
[35] MESHALKIN, P. S. (1970). To the justification of random balance method. *Industrial Laboratory* **36**.
[36] NEUMANN, M. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* **23** 1937–1959.
[37] NEUMANN, M. and VON SACHS, R. (1995). Wavelet thresholding in anisotropic function classes and application to the adaptive estimation of evolutionary spectra. Preprint, WIAS, Berlin.
[38] NIKOLSKII, S. M. (1975). *Approximation of Functions of Several Variables and Imbedding Theorems*. Springer, Berlin.
[39] NUSSBAUM, M. (1983). Optimal filtration of a function of many variables in white Gaussian noise. *Problems Inform. Transmission* **19** 23–29.
[40] NUSSBAUM, M. (1986). On nonparametric estimation of a regression function, being smooth on a domain in $\mathbb{R}^k$. *Theory Probab. Appl.* **31** 118–125.
[41] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.
[42] PATEL, M. S., ed. (1987). *Experiments in Factor Screening. Comm. Statist. Theory Methods* **16**(10). (Special issue.)
[43] PETROV, V. V. (1975). *Sums of Independent Random Variables*. Springer, Berlin.
[44] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence intervals for pointwise curve estimation. *Ann. Statist.* **28** 298–335.
[45] POLYAK, B. T. and TSYBAKOV, A. B. (1990). Asymptotic optimality of the C_p test in the projection estimation of a regression. *Theory Probab. Appl.* **35** 293–306.
[46] RÉNYI, A. (1965). On the theory of random search. *Bull. Amer. Math. Soc.* **71** 809–828.
[47] SCHWARZ, G. (1978). Estimating the dimension of a model. *Ann. Statist.* **6** 461–464.
[48] STONE, C. J. (1986). The dimensionality reduction principle for generalized additive models. *Ann. Statist.* **14** 590–606.

LABORATOIRE DE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UFR DE MATHÉMATIQUES, CASE 7012
CNRS-UMR 7599 ET UNIVERSITÉ PARIS VII
2 PLACE JUSSIEU, 75251 PARIS CEDEX 05
FRANCE
E-MAIL: hoffmann@math.jussieu.fr

LABORATOIRE D'ANALYSE,
TOPOLOGIE, PROBABILITÉS
CENTRE DE MATHÉMATIQUES
ET D'INFORMATIQUE
CNRS-UMR 6632
UNIVERSITÉ DE PROVENCE
39, RUE JOLIOT CURIE
13453 MARSEILLE CEDEX 13
FRANCE
E-MAIL: lepski@cmi.univ-mrs.fr

DISCUSSION

LUCIEN BIRGÉ

Université Paris VI and UMR CNRS 7599

1. Introduction: Lepski's initial paper. In the last 10 years, an enormous amount of work has been produced in order to build so-called *adaptive* estimators and, in particular, one of the authors of this paper designed a special method, now known as *Lepski's method*, to build such estimators. Adaptive estimators have the advantage that they estimate the parameter better when it is easier to estimate. In some sense, the estimator does the best that is possible in view of the properties (smoothness or complexity) of the underlying function to be estimated. For instance it will choose the right bin width for a histogram estimator or the right bandwidth for a kernel estimator. This is quite satisfactory but, even if you know that this is the case, the estimator does not tell you how well it does, because its performance depends on the unknown properties of the parameter. In particular, you have no idea about the order of magnitude of the distance between your estimator and the truth and cannot build nontrivial confidence sets from it.

This has been the main justification for the introduction of the *random normalizing factors* (RNFs) and α_ε-adaptive estimators in Lepski [3]. In this fundamental paper he dealt with the simplest situation where one assumes that either f belongs to some (typically large) set Σ or that it is easier to estimate because it belongs to some much smaller set $\Sigma_0 \subset \Sigma$. In this case, an adaptive estimator estimates f better (and indeed in an optimal way with respect to Σ_0) if f actually belongs to Σ_0 but one will not know it because the estimator does not indicate that $f \in \Sigma_0$. Something is missing here, namely a test procedure that tells you whether f belongs to Σ_0 or not. Lepski's construction essentially involves two estimators $\hat{f}$ and $\hat{f}_0$, especially designed to be optimal when f belongs to Σ and Σ_0 respectively, and a test procedure that f belongs to Σ_0. Of course this is an oversimplified presentation and not all such triplets would do. For the construction to work, special combinations of those three elements have to be chosen in a quite sophisticated way. The close connection between RNFs, α_ε-adaptive estimators and testing whether Σ_0 is true or not is made quite explicit in Proposition 3 of Lepski (1999).

In Section 5 of Lepski [3], the author gives two concrete applications of his ideas after having listed in Section 4 a number of potentially interesting problems to be solved, concluding that "the treatment of minimax risk with RNFs for multidimensional models is the subject of a series of forthcoming papers." The paper by Marc Hoffmann and Oleg Lepski (hereafter H&L) to be discussed is one of these and, in view of the numerous connections between both papers, my discussion will deal with the two of them simultaneously.

2. The new extended framework. In view of handling a problem of variable selection in multidimensional regression, it is necessary to consider the case where there is more than one alternative to Σ. Here the authors deal with N subsets Σ_i, $1 \leq i \leq N$, of Σ, which, as they say, requires a nontrivial extension of the results of Lepski [3]. From the conceptual point of view (definition and general properties of RNFs and α_ε-adaptive estimators), this extension is quite natural and hopefully does not involve serious additional technicalities. The extension essentially preserves the properties of RNFs and α_ε-adaptive estimators as described in Sections 2 and 3 of Lepski [3]. Moreover, and this can be viewed as the main result concerning the new framework, the solution for the case $N > 1$ is essentially equivalent to the solution of the N problems Σ versus Σ_i separately. Proposition 1 actually provides a complete solution for the construction of optimal RNFs and α_ε-adaptive estimators for the general situation starting from optimal RNFs and α_ε-adaptive estimators for each of the N corresponding binary problems, and Proposition 2 offers a reciprocal. This also clearly emphasizes the importance of the initial construction of Lepski [3].

This general fact being established once and for all, it follows that, in view of this equivalence, the hard work now lies in finding optimal RNFs and α_ε-adaptive estimators when $N = 1$. An inspection of the proofs of Theorems 1 and 2 in Lepski [3] or Theorem 1 in H&L immediately confirms the impression that it is not easy. In each case, the authors actually provide a specific construction tailored for the problem at hand. This leads to a very natural question: can one find some more or less generic method to solve the case $N = 1$, at least for some classes of sets Σ and Σ_0? It may be difficult, but certainly quite exciting, to design some general basic methods to solve the problem, even if they have to be tuned in particular situations.

3. The problem of adaptive confidence sets. As mentioned before, adaptive estimators do not tell us anything about their real performance and it is a merit of RNFs to provide the statistician with a rough idea of the distance between the estimator being used and the true underlying parameter. Why should one need to have an evaluation of this distance, if not for approximately locating the true parameter? And approximately locating the true parameter essentially means building confidence sets. The problem of building adaptive confidence sets has not often been considered in the literature up to now, with a few exceptions mentioned in Section 2.4 of H&L, although it is an important and delicate problem which is far from being solved. From my point of view the main merit of the theory of RNFs and α_ε-adaptive estimators developed by H&L is to give a general method for constructing adaptive confidence sets. I only regret that the authors put the main emphasis on normalized versions of the risk and then derive their confidence sets from them through the very rough Markov inequality, rather than directly looking at adaptive confidence sets. In many situations, risks are computed by integrating deviation inequalities and it would be more natural to start from such deviation

inequalities to derive adaptive confidence sets. I am not sure that there is a need for a single object, namely an α_ε-adaptive estimator, for solving both problems of adaptation and problems of adaptive confidence intervals. Since the problem of adaptation has been widely investigated, my point of view would be to concentrate directly on the construction of adaptive confidence sets, not necessarily based on previous risk evaluations.

There is another reason for separating the two problems of estimation and getting confidence sets. In the estimation case, we know how to build adaptive estimators over all Hölder classes simultaneously without any restriction on the parameters. The construction of RNFs requires the assumption that the true unknown f belongs to some *known* set Σ of functions. This may be technically and theoretically unavoidable but obviously leads to some difficulties in order to apply the theory. How should we choose this space Σ which appears to be an essential tool in the construction? In some sense, this is opposite to the philosophy of adaptation and many adaptive methods do not require such knowledge. One can of course make the conservative choice of a very large set Σ but this will be at the price of a very slow rate when the assumption that f belongs to $\bigcup_{1\le i\le N}\Sigma_i$ is rejected.

4. Asymptotics versus nonasymptotics. My main concern about RNFs and related concepts is about their fundamentally asymptotic nature. They are primarily based on rates and comparison of rates. I do not believe in the concept of rates for any practical purpose and, even from a theoretical point of view, I find it terribly misleading. Here is an elementary illustration. If we denote by $H(\beta, L)$ the space of Hölder densities on $[0, 1]$ with smoothness β and constant L, that is, assuming that $0 < \beta \le 1$, the set of densities that satisfy

$$|f(x) - f(y)| \le L|x - y|^\beta \qquad \text{for all } x, y \in [0, 1],$$

the rate of convergence of good estimators based on n i.i.d. observations is known to be bounded by $C(Ln^{-\beta})^{1/(2\beta+1)}$ and this is optimal from a minimax point of view, apart from the constant C. As a consequence, if f belongs to $H(1, L_1)$ it can be estimated at a better rate, namely $n^{-1/3}$, than if f belongs to $H(1/2, L_2)$, the rate being only $n^{-1/4}$. Nevertheless, for all reasonable sample sizes, one can get a more accurate estimator for the second case if $L_1 = 125$ and $L_2 = 1$.

Another illustration of the difficulties connected with the purely asymptotic point of view considered by H&L is as follows. Assume that we have at hand n i.i.d. observations from some unknown density f on $[0, 1]$ belonging to $H(1, L)$ for some unknown value of L. This is actually a simple but quite realistic situation because most usual densities are indeed Lipschitz for a large enough value of L. Here n may be large (a few thousands) but is definitely finite. This is a situation where we know how to estimate f adaptively using a histogram based on a regular partition of $[0, 1]$ with $\hat{D}$ pieces, where $\hat{D}$ is determined by the data only. Both theoretical (Castellan [2]) and practical (Birgé and Rozenholc [1]) nonasymptotic

results are available for this situation and they do not require an a priori upper bound on L although, for practical purposes, $L \leq 10^6$ would probably do. It is also easily seen from simulations that the cases $L = 1$ and $L = 50$ lead to quite different results. Since the relevant estimators are adaptive in a nonasymptotic sense they lead to good estimation procedures in both cases but do not provide confidence sets for f, which is a serious drawback. On the other hand, the theory of RNFs, as presented in Lepski [3] or H&L, provides no solution at all to this problem because it is based on comparison of convergence rates and the rate is the same, namely $n^{-1/3}$, whatever the value of L. I do not mean here that the idea of RNFs is uninteresting or that the estimators which are constructed in Lepski [3] and H&L are bad, but rather that the theory should be modified to take this fact into account and that the quality of the resulting estimators, or rather of the resulting confidence sets, should be evaluated using some nonasymptotic criteria.

To go on with this apology of the nonasymptotic approach, let us consider the effect of N. Of course, as shown by Proposition 1, the effect of N is negligible from an asymptotic point of view. But let us look at the proof carefully. The fact that one can reduce the general problem to the simpler case $N = 1$ is based on the finiteness of $R = \sum_{i=1}^{N} (M_i^*)^p$. Nevertheless, for large values of N this quantity R can be quite large and its presence can completely hide the effect of the rates (generally of order ε^δ for some small value of δ). For all realistic values of ε, the only visible effect will be connected with the size of R. Moreover, interesting applications typically involve a large number of possible alternatives to the larger space Σ. It is even more true that Σ has to be known and therefore it will be natural, for safety reasons, to take it quite large. In the variable selection problem with d variables considered in H&L, the number of possible subsets of significant variables is 2^d. Thinking of $d = 10$ gives an idea of the size of N and of the asymptotic nature of the results. Looking at the residual term in the risk bound for Corollary 1, which is considered negligible in the proof, also casts some doubt about the relevance of such a result when N is large.

To conclude, I would say that the theory of RNFs is a quite interesting attempt to provide some general solution to the delicate problem of finding adaptive confidence intervals for nonparametric problems. Nevertheless, I believe that this theory, in its present form, suffers from its purely asymptotic nature and should be modified in a suitable way. In particular the concept should be more directly oriented toward confidence sets and should have more relevance from a nonasymptotic point of view.

REFERENCES

[1] BIRGÉ, L. and ROZENHOLC, Y. (2002). How many bins should be put in a regular histogram? Technical report, Univ. Paris VI.

[2] CASTELLAN, G. (1999). Modified Akaike's criterion for histogram density estimation. Technical Report, Univ. Paris-Sud, Orsay.
[3] LEPSKI, O. (1999). How to improve the accuracy of estimation. *Math. Methods Statist.* **8** 441–486.

UMR 7599 "PROBABILITÉS ET MODÈLES ALÉATOIRES"
LABORATOIRE DE PROBABILITÉS, BOÎTE 188
UNIVERSITÉ PARIS VI, 4 PLACE JUSSIEU
F-75252 PARIS CEDEX 05
FRANCE
E-MAIL: lb@ccr.jussieu.fr

DISCUSSION

LAWRENCE D. BROWN AND YI LIN

University of Pennsylvania

1. Introduction. We congratulate the authors for a stimulating paper (referred to as HL in the following). As the authors correctly stated, the number of variables does not affect the optimal rate of convergence in a regular parametric model, but it does affect the optimal rate of convergence in nonparametric models. To be more precise, the optimal rate of convergence in a nonparametric function estimation problem depends on the "effective" nonparametric dimension of the model. For example, the "effective" dimension of a nonparametric additive model is 1, no matter how many variables are in the model. Therefore, variable selection in the additive model does not affect the optimal rate of convergence of the model. In light of this, we would like to consider the problem of adaptive estimation in the nonparametric functional ANOVA setup. It is readily seen that the variable selection problem considered in HL is a special case of the problem we consider. Following the example of HL, we will concentrate on the white noise model setting. This can be motivated by the results in Brown and Low [1] and Nussbaum [3].

2. Rates of convergence. Before turning to our main topic we would like to point out a feature of the result in HL that we found surprising. We hope that they will be able to comment on this, and perhaps provide some additional background and a heuristic explanation. The feature that concerns us first appears in the formula for $z_n(\mathbf{i}_s)$ near the end of Section 1, and is repeated in various forms later on, including in Theorem 1. To focus on this feature, let us consider the case of one direction, $\mathbf{i}_s$, as in Theorem 1 and, for simplicity, we consider only isotropic regression. Thus, assume a fixed smoothness, say m, throughout the model. Let the full dimension of the model be d, and let the dimension of the "direction" $\mathbf{i}_s$ of interest be s, say. Then $\beta = m/d$ and $\beta(\mathbf{i}_s) = m/s$.

As HL note, if $s > d/2$ one may choose $\alpha = \alpha_n = n^{-a}$ for suitable $a > 0$ and then $z_n(\mathbf{i}_s)$ coincides with the rate of convergence on the set $\Sigma(\mathbf{i}_s)$. This situation seems entirely satisfactory; we are concerned with the opposite situation where $s \le d/2$. In that case, for any α_n converging to 0,

$$z_n(\mathbf{i}_s) = \left(\frac{\sqrt{\ln(1/\alpha_n)}}{n}\right)^{2\beta/(4\beta+1)} > \left(\frac{1}{n}\right)^{\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)}$$
$$= \text{rate of convergence on } \Sigma(\mathbf{i}_s).$$

Hence, no fully adaptive estimation is possible in the RNF sense of HL. What about the usual sense of adaptive estimation? Is fully adaptive estimation possible here in the usual sense? Or, is the rate $z_n(\mathbf{i}_s)$ the best possible ordinary rate of convergence over $\Sigma(\mathbf{i}_s)$ for an estimator in this situation if the optimal convergence rate is also desired over the full space?

The ordinary sense of adaptation does not involve a choice of α. So if $z_n(\mathbf{i}_s)$ is the best possible ordinary rate of convergence over $\Sigma(\mathbf{i}_s)$, then it is important in constructing an ordinary adaptive estimator to choose α_n in an optimal way. For ordinary adaptive estimation, what is the optimal choice of α_n, and what is the corresponding optimal adaptive rate result?

In this connection we note for the construction in Section 3.1 it appears that any choice $\alpha_n \to 0$ will yield an estimator that converges at the rate with respect to *asymptotic risk*, defined as

$$\lim_{B\to\infty} \limsup_{n\to\infty} \sup_{f\in\Sigma(\mathbf{i}_s)} E_f^n \left\{\varphi_n(\alpha_n, \mathbf{i}_s)^{-2}\big(\min(\|\hat{f}_n - f\|, B)\big)^2\right\} \le \infty.$$

On the other hand it appears one must choose $\alpha_n = n^{-a}$ for suitable $a > 0$ in order to attain the appropriately normed limiting rate, as defined via (3) of HL. If this is so, it is of additional technical interest as an instance where the asymptotic and the limiting risks can differ, and also where no optimal asymptotic risk is attained.

3. Functional ANOVA formulations. A general d-dimensional nonparametric function estimation problem has an optimal rate of convergence that depends on the magnitude of d. For even moderately large d, the rate of convergence is very slow compared to that of one-dimensional problems. This is one aspect of the so-called curse of dimensionality. To circumvent the curse of dimensionality, we often consider the functional ANOVA decomposition. Consider a d-dimensional nonparametric function with the following decomposition:

$$(1) \qquad f(x_1, x_2, \ldots, x_d) = \text{constant} + \sum_{i=1}^{d} f_i(x_i) + \sum_{i<j} f_{ij}(x_i, x_j) + \cdots,$$

where the components satisfy side conditions which generalize the usual side conditions for parametric ANOVA to function spaces, and the series may be truncated in some manner.

There are two different types of functional ANOVA models commonly considered. They differ in the function spaces considered. Let $H^m([0,1]^d)$ be the mth-order Sobolev Hilbert space of functions on $[0,1]^d$. Stone [4] assumes smoothness conditions like $f_i \in H^m([0,1])$, $f_{ij} \in H^m([0,1]^2)$ and so on (Stone actually assumed Hölder spaces, which are similar to the Sobolev Hilbert spaces) and showed that the optimal rate of the model is $n^{-2m/(2m+s)}$, the same as that of s-dimensional full function problems, where s is the highest order of interactions considered. Therefore the effective dimension of such functional ANOVA models is s. In the following we will refer to such models as partial derivative ANOVA (PD-ANOVA) models. The smoothing spline ANOVA models introduced in Wahba [5] and discussed in detail in Wahba et al. [6] make a different type of assumption on the component functions in the functional ANOVA decomposition. That is, after determining the function space of each main effect, the function space in which an interaction lies is assumed to be the tensor product space of the function spaces of the interacting main effects. Therefore, in the tensor product space ANOVA model, if we assume the main effects are in $H^m([0,1])$, the kth-order interactions lie in $\bigotimes^k H^m([0,1])$. We will refer to them as tensor product space ANOVA (TPS-ANOVA) models.

For a Hilbert space E_1 of functions of x_1 and a Hilbert space E_2 of functions of x_2, the tensor product space of E_1 and E_2 is defined as the completion of the class of functions $\{\sum_{i=1}^k f_i(x_1)g_i(x_2),\ f_i \in E_1,\ g_i \in E_2\}$ under a norm induced by the norms in E_1 and E_2. It is known (Lin [2]) that the tensor product space of d Sobolev spaces $H^m([0,1])$ is equivalent to

$$\Omega_m = \left\{ f : \frac{\partial^{|\mathbf{i}|} f(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{i}}} \in L_2([0,1]^d),\ \forall \mathbf{i} = \{i_1, i_2, \ldots, i_d\} \in \mathbf{R}^d \right.$$
$$\left. \text{such that } \max_j i_j \le m \right\},$$

where $\mathbf{x} = \{x_1, x_2, \ldots, x_d\}$, $|\mathbf{i}| = \sum_j i_j$. The norm for any $f \in \Omega_m$ is

$$\|f\|^2_{\Omega_m} = \sum_{\{\mathbf{i}:\ \max_j i_j \le m\}} \int \left[\frac{\partial^{|\mathbf{i}|} f(\mathbf{x})}{\partial \mathbf{x}^{\mathbf{i}}} \right]^2 d\mathbf{x}.$$

Since any function in $\bigotimes^k H^m([0,1])$ has one derivative of order km (order m in each direction), and some other derivatives of orders higher than m, we can see that TPS-ANOVA models put higher order smoothness conditions on interactions than on main effects, and the order of the smoothness condition imposed on an interaction increases with the order of the interaction. This reveals an intuitively appealing aspect of the tensor product ANOVA model: starting from an additive model, when we make the model more and more complex by throwing in higher and higher order interaction terms, we assume stronger and stronger smoothness conditions on the new terms thrown in to keep the model manageable. This is

consistent with the philosophy of the ANOVA modeling strategy of throwing away higher order interaction terms.

Lin [2] showed that the optimal rate of convergence for the tensor product space ANOVA model is $[n(\log n)^{1-s}]^{-2m/(2m+1)}$, where s is the highest order of interactions considered. Notice this implies that the optimal rate of the saturated tensor product space model is $[n(\log n)^{1-d}]^{-2m/(2m+1)}$. This is only a log factor away from the optimal rate of the one-dimensional nonparametric problems. Therefore the optimal rate of convergence of TPS-ANOVA models depends on the number of variables only through a log term.

HL consider a particular family of submodels related to the problem of variable selection, but their formulation can be applied to other subfamilies. For example, we can apply the framework to the model selection problem in the functional ANOVA framework. In principle we can take the function space Σ of HL to be the space corresponding to any functional ANOVA model and consider adaptive estimation with respect to smaller ANOVA submodels. For simplicity, we assume Σ to be a function space corresponding to the saturated ANOVA model, and we consider adaptive estimation with respect to the functional ANOVA submodels.

3.1. *PD-ANOVA model.* The saturated function space considered in PD-ANOVA is the same as the one considered in HL with $\beta_1 = \beta_2 = \cdots = \beta_d = m$. Here Σ is the Sobolev space $H^m([0,1]^d)$. It is anticipated that the same results on rate of convergence in HL should also be valid PD-ANOVA. This can actually be proved by the same line of proof in HL. We only provide a brief description of how HL's proofs can be modified to give results in PD-ANOVA setting. First consider a given functional ANOVA submodel M_1 of the form (1), with the highest order of interaction to be s. Let S be the set of index sets corresponding to the generators of M_1. For example, for the model

$$f(x_1, x_2, x_3) = f_0 + f_1 + f_2 + f_3 + f_{12} + f_{13} + f_{23},$$

we have $S = \{\{1,2\}, \{1,3\}, \{2,3\}\}$. The index set I in HL corresponds to the set of $\mathbf{k}$ such that $\theta_{\mathbf{k}}$ [defined in (22) in HL] are not zero for the functions in the submodel. In the ANOVA setting, the index set I can be defined correspondingly as

$$I = \bigcup_{\mathbf{i}_l \in S} \{(k_1, k_2, \ldots, k_d) \in \mathbf{N}^d : k_j = 0,\ \forall j \notin \mathbf{i}_l\}. \tag{2}$$

To avoid technicalities, we concentrate on the rate of convergence and ignore the constants that do not depend on ε. We now show that, ignoring the constants, Theorem 1 of HL applies to PD-ANOVA. The β in HL should now be m/d, and $\beta(\mathbf{i}_s)$ should now be replaced by m/s. All the other quantities in HL can then be defined similarly ignoring the constants. The proof of the upper bound part of Theorem 1 follows the same line of argument as that of the proof of upper

bound in HL. The lower bound for the PD-ANOVA model is a corollary of the lower bound part of Theorem 1 of HL, since the function space in the PD-ANOVA model contains the function space in the variable selection model. Once Theorem 1 is established, the generalization to adaptation to multiple subspaces follows the development in HL.

3.2. *Tensor product space model.* Tensor product space framework is relevant to both variable selection and TPS-ANOVA model choice. In such a framework the saturated space Σ is $\bigotimes^d H^m([0,1])$, which is different from the function space $H^m([0,1]^d)$ considered in HL. However, the framework in HL can still apply. For simplicity in this discussion we will consider only the TPS variable selection model. That is equivalent to a TPS-ANOVA model with only one term with dimension s. It should be clear that the rates we obtain are valid for the general TPS-ANOVA model.

We basically follow the notation of HL. For notational simplicity, we concentrate on the case $p = 2$. We concentrate on the rate of convergence and ignore all the constants that do not affect rates. These include C's, Z's and λ in HL. Also, we do not need β and $\beta(\mathbf{i}_s)$ in our tensor product space case. We also assume α_ε to be a small fixed number independent of ε, though we keep α_ε in the notation just so that it is easy to see to which term in HL it corresponds. The full space Σ corresponds to

$$\Sigma(m, L) = \left\{\theta : \sum_{\mathbf{k}\in\mathbf{N}^d} \left(\theta_{\mathbf{k}}^2 \prod_{i=1}^d (1+k_i)^{2m}\right) \le L^2\right\}. \tag{3}$$

I and J are defined the same way as in HL. That is,

$$I = \{(k_1, k_2, \ldots, k_d) \in \mathbf{N}^d : k_j = 0,\ \forall j \notin \mathbf{i}_s\},$$

$$J = \mathbf{N}^d \setminus I.$$

We define

$$\Sigma(\mathbf{i}_s) = \left\{\theta : \sum_{\mathbf{k}\in I} \left(\theta_{\mathbf{k}}^2 \prod_{i=1}^d (1+k_i)^{2m}\right) \le L^2\right\}.$$

We further define

$$I_\varepsilon = \left\{(k_1, k_2, \ldots, k_d) \in I : \prod_{j=1}^s (1+k_{i_j})^{2m} \le K(\varepsilon)\right\},$$

$$J_\varepsilon = \left\{(k_1, k_2, \ldots, k_d) \in J : \prod_{i=1}^d (1+k_i)^{2m} \le N(\varepsilon)\right\},$$

$$Q_\varepsilon = \left\{(k_1, k_2, \ldots, k_d) \in \mathbf{N}^d : \prod_{i=1}^d (1+k_i)^{2m} \le M(\varepsilon)\right\},$$

with

$$K(\varepsilon) = \left[\varepsilon^2 \left(\log \frac{1}{\varepsilon}\right)^{s-1}\right]^{-2m/(2m+1)},$$

$$M(\varepsilon) = \left[\varepsilon^2 \left(\log \frac{1}{\varepsilon}\right)^{d-1}\right]^{-2m/(2m+1)},$$

$$N(\varepsilon) = \left[\varepsilon^4 \left(\log \frac{1}{\varepsilon}\right)^{d-1}\right]^{-2m/(4m+1)}.$$

Note that the way the above six quantities are defined is different from that in HL. Notice also for our definition it is always true that $N(\varepsilon) > K(\varepsilon)$. This is a consequence of the fact that the rates of convergence in TPS models differ by only a log term.

Notations $\hat{\theta}_\varepsilon$, $\hat{\theta}_\varepsilon^{(0)}(\mathbf{i}_s)$ and $T_\varepsilon(\mathbf{i}_s)$ have the same definitions as those in HL, but with our new I_ε, J_ε, Q_ε. We define

$$\varphi_\varepsilon(\Sigma) = \left[\varepsilon^2 \left(\log \frac{1}{\varepsilon}\right)^{d-1}\right]^{m/(2m+1)},$$

$$\varphi_\varepsilon\left(\alpha_\varepsilon, \mathbf{i}_s\right) = \left[\varepsilon^2 \left(\log \frac{1}{\varepsilon}\right)^{s-1}\right]^{m/(2m+1)}.$$

Notations $A_\varepsilon(\mathbf{i}_s)$, $\rho_\varepsilon^*(\mathbf{i}_s)$ and $\theta_\varepsilon^*(\mathbf{i}_s)$ have the same definitions as those in HL, but with our new $T_\varepsilon(\mathbf{i}_s)$, $\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s)$, $\varphi_\varepsilon(\Sigma)$, $\hat{\theta}_\varepsilon$ and $\hat{\theta}_\varepsilon^{(0)}(\mathbf{i}_s)$. We will show that Theorem 1 in HL holds in our case with our definitions (ignoring the constants). To do that, we first prove Lemma 1. This is a different lemma since all the quantities involved are defined differently now. Therefore we provide a detailed proof:

$$E_\theta \|\hat{\theta}_\varepsilon - \theta\|^2 = \sum_{Q_\varepsilon} \varepsilon^2 + \sum_{Q_\varepsilon^c} \theta_{\mathbf{k}}^2 = \varepsilon^2 |Q_\varepsilon| + \sum_{Q_\varepsilon^c} \theta_{\mathbf{k}}^2.$$

We have the following approximation:

$$|Q_\varepsilon| \sim \int_{\prod_{i=1}^d (1+x_i)^{2m} \le M(\varepsilon)} 1\, dx_1\, dx_2 \cdots dx_d$$

$$= \int_{\prod_{i=1}^d (1+x_i) \le M^{1/2m}(\varepsilon)} 1\, dx_1\, dx_2 \cdots dx_d.$$

Changing the variable in the integral, $z_i = \prod_{j \le i}(1 + x_j)$, $i = 1, 2, \ldots, d$, the above quantity becomes

$$\int_1^{M^{1/2m}(\varepsilon)} \left[\int_1^{z_d} \cdots \int_1^{z_2} z_1^{-1} \cdots z_{d-1}^{-1} \, dz_1 \cdots dz_{d-1} \right] dz_d$$
$$= \int_1^{M^{1/2m}(\varepsilon)} \left[(\log z_d)^{d-1} \right] dz_d$$
$$= z_d (\log z_d)^{d-1} \Big|_1^{M^{1/2m}(\varepsilon)} - (d-1) \int_1^{M^{1/2m}(\varepsilon)} \left[(\log z_d)^{d-2} \right] dz_d$$
$$\sim M^{1/2m}(\varepsilon) (\log M)^{d-1}.$$

On the other hand, by the definition of Q_ε and Σ, we have

$$\sum_{Q_\varepsilon^c} \theta_{\mathbf{k}}^2 M(\varepsilon) \le \sum_{Q_\varepsilon^c} \theta_{\mathbf{k}}^2 \prod_{i=1}^{d} (1 + k_i)^{2m} \le L^2.$$

So we have

$$E_\theta \|\hat{\theta}_\varepsilon - \theta\|^2 \sim \varepsilon^2 M^{1/2m}(\varepsilon) (\log M)^{d-1} + M^{-1}(\varepsilon) L^2 \sim \varphi_\varepsilon^2(\Sigma).$$

Hence Lemma 1 is proved. Lemma 2 stays the same since it does not depend on our notation.

Similar to the derivation of $|Q_\varepsilon|$, we have

$$|I_\varepsilon| \sim K^{1/2m}(\varepsilon) (\log K)^{s-1}.$$

With these in mind, the proof of the upper bound in HL goes through with our definition of the notation. In particular, the rate $\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s)$ is $[\varepsilon^2 (\log(\frac{1}{\varepsilon}))^{s-1}]^{m/(2m+1)}$. The proof of the lower bound is not needed. It follows directly from the fact that $\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s) = [\varepsilon^2 (\log(\frac{1}{\varepsilon}))^{s-1}]^{m/(2m+1)}$ coincides with the minimax rate of convergence on $\Sigma(\mathbf{i}_s)$.

It is interesting to note that since the rate of convergence $\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s)$ in the tensor product space framework is $[\varepsilon^2 (\log \frac{1}{\varepsilon})^{s-1}]^{m/(2m+1)}$, the same as the minimax rate of convergence on $\Sigma(\mathbf{i}_s)$, the estimation is fully adaptive in the RNF sense of HL. The generalization to adaptation to multiple subspaces can be made following the development in HL. We do not pursue that here.

REFERENCES

[1] BROWN, L. D. and LOW, M. G. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

[2] LIN, Y. (2000). Tensor product space ANOVA models. *Ann. Statist.* **28** 734–755.

[3] NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.

[4] STONE, C. J. (1994). The use of polynomial splines and their tensor products in multivariate function estimation (with discussion). *Ann. Statist.* **22** 118–184.

[5] WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

[6] WAHBA, G., WANG, Y., GU, C., KLEIN, R. and KLEIN, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. *Ann. Statist.* **23** 1865–1895.

DEPARTMENT OF STATISTICS
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104-6302
E-MAIL: lbrown@wharton.upenn.edu

DISCUSSION

SAM EFROMOVICH

University of New Mexico

This is a nice and very stimulating article. The article suggests estimating an underlying function together with its normalization factor. An estimate of the normalization factor is referred to as a random rate.

The article is also thought-provoking. Actually after reading this article I realized that I had many unanswered questions so I decided to use this discussion primarily for clarification of some of the issues. In what follows I denote by HL a procedure or a notion introduced in the article.

1. Can an HL estimator be data-driven? As I understand it, the HL setting is defined as follows. There are given an underlying (principal) function space Σ and its N subspaces $\Sigma_1, \ldots, \Sigma_N$. A statistician (and thus an estimator) knows everything about these spaces. The space Σ is the one traditionally studied in the nonparametric minimax literature, and the subspaces describe a belief that an estimated signal may have a simpler structure and thus can be estimated more accurately. Then, according to the HL paradigm, an HL estimator must be minimax over the largest space Σ but it is allowed to be not minimax over the subspaces where, instead of being necessarily minimax, an HL random rate (a new notion introduced by the authors) should satisfy some restrictions.

As a result, there are $N + 1$ candidates (the underlying space and its N subspaces) to choose from and this is the reason the authors refer to an HL procedure of estimation as an adaptive one. However, let us stress that all the HL procedures suggested require complete information about these $N + 1$ candidates.

Can this requirement about knowing the candidates be dropped? In other words, can an HL procedure be data-driven? (To avoid possible confusion with notions of adaptive estimates used in the article, I use the notion of a data-driven estimation to stress that an estimator does not a priori depend on the spaces, in particular on the underlying Σ.)

At first glance the answer is "yes" because a similar setting was considered in the original articles by Lepski [3, 4] where a Lepski (L) adaptive procedure

was suggested that later was successfully developed into a data-driven estimator. In those articles a univariate problem was considered and it was assumed that an unknown underlying space is an element of a given net $\Sigma_1^*, \dots, \Sigma_m^*$ of spaces. The goal was to find an adaptive estimator that was minimax over an underlying space.

The HL setting looks very similar to the L setting but the differences, in my opinion, are dramatic. For instance, in the L setting all spaces in the net are treated equally, whereas in the HL setting there is an underlying set Σ for which and only for which the minimaxity of an HL estimator is required. As a result, while for the L approach a data-driven estimator is developed by considering a fine net of spaces that "approximates" an underlying space, it is not clear that such a possibility exists in the HL setting. Indeed, if a data-driven estimator makes a mistake and considers a larger space $\Sigma' \supset \Sigma$ in place of the underlying Σ, then this estimator is no longer necessarily minimax over the underlying Σ and thus it does not satisfy the HL paradigm. The outcome is similar when $\Sigma' \subset \Sigma$.

Thus, how do we construct a data-driven estimator that preserves the HL paradigm?

2. Adaptation versus HL estimation. I found this line of the article very interesting and (naturally) biased toward HL estimation. The authors are very categorical in their conjecture that it is impossible to compute the accuracy (meaning the random rate) of an adaptive estimator; see the paragraph below (7) and note that this assertion is softened a bit in Remark 5 of Section 2.3.

To shed light on this conjecture, let us consider an example in which finding HL random rates is trivial, but the construction of an adaptive estimator is a serious problem. I use a classical example of estimation of analytic functions; see Efromovich [2, Chapter 7]. In this example the minimax rate is proportional to $(\ln(n)n^{-1})^{1/2}$ and, because it does not depend on an underlying analytic space, this is the HL random rate as well. Thus, finding a random rate (normalization factor) is trivial. On the other hand, adaptive (data-driven) minimax estimation is not an elementary problem but an Efromovich–Pinsker estimator $\hat{f}_{\mathrm{EP}}$ will do it (see Efromovich [2]). As a result, in the HL terminology, $\{(\ln(n)n^{-1})^{1/2}, \hat{f}_{\mathrm{EP}}\}$ is a 0-adaptive HL estimator. Note that this is also an example of an HL estimator that is adaptive in the HL sense but has a different structure than the hypothetical adaptive estimator discussed in Proposition 3.

To be fair to the HL approach, I would like to present an example in which adaptive estimation is trivial but finding random rates is not. I use a familiar linear inverse problem where the available signal is $\int h(t-x)f(x)\,dx$; it is observed in additive noise and the kernel h is supersmooth (see Efromovich [2], pages 299–301). Here the minimax rate crucially depends on the smoothness of f [it is proportional to $\ln(n)^{-\beta/\nu}$, where ν describes the kernel] and thus finding HL random rates is a problem. On the other hand, an elementary (not adaptive) estimator is simultaneously minimax over a wide set of function spaces; see again Efromovich [2].

I like the latter example because, in my opinion, it sheds new light on the HL approach and this light is not shaded by complicated estimates of f.

3. Does an HL estimator improve accuracy of estimation? This is another interesting line in the article where using the notion "accuracy" confuses me. I accept the HL position that knowing an underlying rate gives extra information to a statistician, but does this knowledge improve the accuracy of estimation of an underlying function f?

To discuss this issue, I would like to present an example of a minimax estimator that always outperforms the HL estimator in terms of mean integrated squared error (MISE) convergence.

The example is very simple. I again consider an Efromovich–Pinsker estimator and compare it with the canonical HL estimator suggested for the filtering model. Let us begin with the HL estimator. For each $f \in \Sigma^*$, where Σ^* is one of $N+1$ possible spaces, the fastest rate of its MISE convergence is the minimax rate for Σ^* (I explain this in more detail in the next section). On the other hand, the Efromovich–Pinsker estimator is superefficient over Σ^* (using the terminology of Brown, Low and Zhao [1]) and thus

$$E\|\hat{f}_{EP} - f\|^2_{L_2} / E\|\hat{f}_{HL} - f\|^2_{L_2} = o(1). \tag{1}$$

This assertion does not contradict the optimal minimax results about the canonical HL estimator because here a pointwise approach is used. Also, the Efromovich–Pinsker estimator does not estimate rates. However, the result presented raises the following questions. If there exists an estimator A whose MISE for every underlying function decreases faster than the MISE of an estimator B, then why should a statistician be interested in the random rate (normalization factor) of estimator B? Also, if a statistician knows the normalization factor of the estimator B but does not know the normalization factor of estimator A, then is it right to call estimator B more accurate?

4. Minimax and estimation. The minimax approach considered in the article is called global and it assumes that an underlying function can be any point in an underlying space Σ. For instance, if $\Sigma = \Sigma(\beta, L)$ is a univariate Sobolev body (23), then the minimax rate is $n^{-\beta/(2\beta+1)}$ and it is attained by a projection estimator $\hat{f}_{\mathrm{HL}}(x) = \sum_{k=0}^{n^{1/(2\beta+1)}} \hat{\theta}_k \phi_k(x)$ used in the article. This estimate is very simple; it is globally minimax and it is also very convenient for both L and HL estimation procedures. However, it also has a statistical drawback—for each $f \in \Sigma(\beta, L)$ its MISE is equal to $O(1)n^{-2\beta/(2\beta+1)}$ regardless of smoothness of the function f.

Also note that the global minimax approach allows us to consider different underlying functions for different n. This is clearly not the case in applications, and thus other types of minimax approaches have been developed.

Probably the most studied one is a minimax approach where, for a given function f_0, it is assumed that the difference $f - f_0$ belongs to Σ and that f approximates f_0, for instance, $\|f - f_0\|_{L_\infty} \to 0$. For known examples this minimax approach does not change rates but it has been instrumental in finding sharp constants. It also implies more sophisticated adaptive procedures; see the discussion in Chapter 7 of my book.

The reason this minimax approach does not change rates is clear—functions studied are not sufficiently fixed and the considered local space is too large. The situation changes if we consider the following local space:

$$\Sigma(\beta, K, f_0) = \Bigg\{ f\colon\ f(x) = \sum_{k=0}^{K} \theta_{0k}\phi_k(x) + \sum_{k>K} \theta_k\phi_k(x), \\ \sum_{k>K} (\theta_k^2 - \theta_{0k}^2)(1 + k^{2\beta}) \le 0, \theta_{0k} = \int f_0(x)\phi_k(x)\,dx \Bigg\}.$$

The new element here is that the low frequency part of an estimated f is fixed and only the high frequency part is flexible. (Actually, a variety of projections can be used but this simple example clarifies the approach.) Note that all considered functions belong to the boundary of the space $\Sigma(\beta, L_0)$, $L_0 = \sum_{k\ge 0} \theta_{0,k}^2(1 + k^{2\beta})$; nevertheless under mild assumptions the minimax MISE converges faster than the global minimax $n^{-2\beta/(2\beta+1)}$, that is,

$$\inf_{\tilde f} \sup_{f\in\Sigma(\beta,K,f_0)} E\|\tilde f - f\|_{L_2}^2 = o(1)n^{-2\beta/(2\beta+1)}.$$

This rate is attained by the Efromovich–Pinsker estimator, and sharp constants are also known for this setting.

This local minimax approach sheds new light on (1) in Section 3 above. Namely, the HL canonical estimator is not locally minimax and this explains why other estimators can dominate it at each point. On the other hand, I conjecture that the HL global minimax setting can be extended to the local minimax setting.

5. Random rates in anisotropic regression? This was the most confusing question that I dealt with. Title, key words and Introduction promise the reader that a regression model is considered and some results for the regression model are presented. I may be wrong and missing something, but is there a single regression result?

I also would like to say that I am truly impressed by the regression model suggested for the study and the minimal assumptions made. To be specific, in HL's (1) and (2) the model $Y_i = f(X_i) + \varepsilon_i$, $i = 1, 2, \ldots, n$, is considered, where f belongs to the anisotropic class (2) with positive β_i's and the errors ε_i are uncorrelated zero-mean noise variables. Nothing more is assumed.

I am very interested in looking at proofs for this model. First, I would like to see how errors with infinite moments will be dealt with. The case of dependent errors

is very interesting. Also, typically even for a univariate setting some restrictions on β are required to obtain optimal adaptive results, and some restrictions are necessary for the equivalence between a regression and filtering models. How are they bypassed?

Thus, I conclude my discussion with the following question to the authors. Can you formulate a mathematical HL result for the model (1)–(2) and explain its proofs for the above-highlighted cases?

REFERENCES

[1] BROWN, L. D., LOW, M. G. and ZHAO, L. H. (1997). Superefficiency in nonparametric function estimation. *Ann. Statist.* **25** 2607–2625.

[2] EFROMOVICH, S. (1999). *Nonparametric Curve Estimation: Methods, Theory and Applications*. Springer, New York.

[3] LEPSKI, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

[4] LEPSKI, O. V. (1992). Asymptotic minimax adaptive estimation. 2. Statistical models without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37** 433–448.

DEPARTMENT OF MATHEMATICS
AND STATISTICS
UNIVERSITY OF NEW MEXICO
ALBUQUERQUE, NEW MEXICO 87131-1141
E-MAIL: efrom@math.unm.edu

DISCUSSION

DOMINIQUE PICARD AND KARINE TRIBOULEY

Université Paris VII—Denis Diderot

In their paper, the authors make a new interesting step toward "ideal" estimation in the nonparametric framework: they face at the same time the estimation problem and the control of accuracy.

Going back in history, one of the major steps has probably been to set up the problem in the minimax framework. Let us put it in the form mentioned by the authors. The question is to find $\hat{f}_n$ such that

$$\limsup_{n\to+\infty} \sup_{f\in\Sigma} \mathbb{E}_f \Phi_n(\Sigma) \|\hat{f}_n - f\| < \infty, \tag{1}$$

where Σ is a class of functions and $\Phi_n(\Sigma)$ is a rate of convergence that cannot be improved.

The following step in this path toward "ideal" estimation is the notion of adaptation: for Π denoting a collection of classes of functions,

$$\Psi_n(f) = \inf\{\Phi_n(\Sigma),\ f \in \Sigma,\ \Sigma \in \Pi\}$$

replaces $\Phi_n(\Sigma)$ in (1). It has been proved by various authors that some methods of estimation share this property of being adaptive: among them, penalized model selection methods, Lepski's method for selecting the bandwidth and wavelet thresholding.

However, in this context of adaptation, although we know that the method produces an estimator achieving the ideal accuracy $\Psi_n(f)$, this quantity is unknown and remains hidden. But in practice, this quantity is very important because it somehow tells us the confidence we can put in the estimation. Some methods like resampling methods allow us to estimate the expected error but they do not have minimax properties. One of the purposes of this paper is to give information about the quantity $\Psi_n(f)$ in the minimax context. The authors are able to produce a procedure providing, in addition to an optimal adaptive estimator, a "kind"of estimation of the accuracy $\Psi_n(f)$ called *the random normalizing factor* (RNF). The procedure seems to perform very well when Π is not too wide a class.

We comment on two aspects which seem to us especially interesting: the first concerns the link between the RNF and the adaptive confidence intervals; the second, the link between the RNF and the maxisets.

1. RNF and adaptive confidence intervals. Of course, another standard situation in statistical estimation where a normalizing factor plays an important role is the construction of confidence intervals, since it is obvious that, in such a context, the length of the confidence interval appears as a quantity relatively close to the notion of RNF.

There exists a very developed theory for confidence intervals based on kernel estimators with nonrandom bandwidth, showing their consistency and calculating the rates for the coverage probabilities (see, for instance, Hall [2, 3]). Recently, more sophisticated constructions have been able to provide confidence intervals using kernel estimators with data-driven selected bandwidth (Neumann [6]; see also Faraway [1]). However, at this stage, the length of the confidence intervals is not data driven. So they remain hidden except by assuming a priori that the function belongs to a prescribed class.

In Picard and Tribouley [10], using methods from wavelet thresholding, confidence intervals were provided adapting to functions with a very different kind of regularity (eventually having singularities or highly oscillating in small intervals) and without incorporating any knowledge of the regularity in the construction. In this case, the length of the interval is also data driven. Briefly, to fix the ideas, let us summarize roughly a procedure of Picard and Tribouley [10] for finding an adaptive confidence interval in an equispaced regression framework (which is very close to the white noise model investigated here).

Let ϕ and ψ be a pair of father and mother wavelets. First, let us construct the following crucial indicator:

$$\hat{j}(x_0) = \sup\{j \geq 0,\ \psi_{jk}(x_0) \neq 0,\ |\hat{\beta}_{jk}| \geq M t_n\}, \qquad t_n = \sqrt{\frac{\log n}{n}},$$

the maximum scale size where coefficients are significant in a neighborhood of x_0. Here $\hat{\beta}_{jk}$ is an estimate of the wavelet coefficient constructed in the standard way. Then we choose, for the center of the interval,

$$\hat{f}(x_0) = \sum_k \hat{\alpha}_{j_0k}\phi_{j_0k}(x_0) + \sum_{j=j_0}^{\hat{j}(x_0)} \sum_k \hat{\beta}_{jk}\psi_{jk}(x_0).$$

This central part is generally quite close to the standard thresholding estimator. It can be proved that the interval centered on $\hat{f}(x_0)$ with length $2u_{1-\alpha/2}v(n)$, where $u_{1-\alpha/2}$ is a Gaussian quantile and where the leading term is given by

$$v(n)^2 = \frac{1}{n}\left(\sum_k \phi_{j_0k}^2(x_0) + \sum_{j=j_0}^{\hat{j}(x_0)} \sum_k \psi_{jk}^2(x_0)\right),$$

has (up to a logarithm term) an optimal coverage up to the first order. Roughly, if we want to improve the order of accuracy, we have to reduce the bias by replacing $\hat{j}(x_0)$ with $\hat{j}_\eta(x_0)$, which is a quantity increasing with the order of accuracy η, and to correct for the higher order moments occurring in the Edgeworth expansions. We note here that $v(n)$ obviously plays the role of the RNF.

We have several comments:

1. An important difference of the work provided here is the fact that our central issue is to focus on the unknown quantity $f(x_0)$ at a fixed point, instead of a band around a norm. We think that, for applications, it is important to have a result with a visual significance. Hence, as for a band constructed on the $\mathbb{L}_2$-norm, it is difficult to be satisfied from this point of view. However, a band constructed using the $\mathbb{L}_\infty$-norm would be mostly satisfactory. In the authors' opinion, what can be done in this direction?
2. A central issue for providing confidence intervals (or confidence bands) is the problem of the bias. There are two commonly used methods to deal with the bias, undersmoothing and explicit bias correction. In every case, the bias part does contribute. This is completely different from the minimax situation where in general there is a trade-off between the stochastic term and the bias. Even for adaptive situations, the bias may be the leading term.
3. In the confidence interval problem, not only the size of the interval is important but also the coverage accuracy. A considerable effort has been made by the statistical community to improve the accuracy of confidence intervals and the construction in Picard and Tribouley [10] benefits from much previous work (see Hall [2, 3], Neumann [6]). Do the authors think that the inequality governing the order of confidence of the RNF can give more precise results about the coverage?
4. Nevertheless, the confidence interval presented above does not overcome the obstruction mentioned in Low [5]. It does require some assumptions (in

addition to the usual regularity conditions) which can be interpreted in the following way. We do not need to know the regularity of f, but the precision of the estimation is linked with the local regularity in x_0 and, somehow, this local regularity has to be estimated. This problem has no solution without extraneous assumptions except for tremendous rates of convergence. Hence, our assumptions have to be understood as the fact that the observations contain enough information relative to the complexity of the function around x_0.

5. In the context presented by the authors, they are able to chose a normalizing factor among a finite set of possible alternatives. What would be the price to pay to be able to choose among a much wider class of regularity, as is the case for the previous adaptive confidence intervals? This is obviously connected to the possibility of increasing the number N with n.

2. RNF and maxisets. Let us consider a remark of the authors:

> It may well happen that $\hat{\rho}_n$ is smaller than $\varphi_n(\Sigma)$, without having $f \in \bigcup_j \Sigma_j$. However, this suggests that f is somehow close to $\bigcup_j \Sigma_j$.

In fact, such a phenomenon can be analyzed in a slightly different way if we think in terms of "maxisets" instead of thinking in terms of functional classes.

It is common in nonparametric statistics to index a set of functions by smoothing parameters. However, when looking for minimax procedures over fixed functional sets, or adaptive procedures with respect to a range of sets indexed by a smoothing parameter (like Hölder spaces indexed by the smoothness parameter α), we are in fact influenced by functions in this set which are the most difficult to estimate with a general procedure. Actually, this set of "bad functions" strongly depends on the definition of smoothness. Most unfavorable a priori measures or sets of functions in Assouad's cube or Fano's pyramid do not look the same at all if we refer to Hölder classes or to Sobolev spaces, for instance. Moreover they usually do not reflect well what we expect to find in practical situations. This explains why many people find it a bit arbitrary to look for special properties over a specific class of regularity which does not necessarily have any connection with the practical situation.

In Kerkyacharian and Picard [4] the focus is on the concept of "maxiset," that is, the maximal set where a procedure has some given rate of convergence. The setting is not extremely different from the minimax context but it has the main advantage of providing a functional set which is genuinely connected to the procedure and the model.

If we come back to the present situation, both the procedure of estimation and the RNF may be chosen arbitrarily, but in practice they are generally constructed from a particular estimation procedure. If this is the case, it can be a serious advantage to start with a set of Σ_j's associated with the maxisets of the procedure. More precisely, in Kerkyacharian and Picard [4], the problems of finding maxisets

and oracle inequalities are connected. A typical situation is the following. Let us define

$$F(f)(j) := \sup_{j' \geq j} 2^{-j'/2} \|E_{j'} f - f\|, \tag{2}$$

where $E_j f$ is the mean of $\hat{f}_j$, a sequence of estimators of the function f indexed by the tuning parameter j (e.g., a bandwidth). We consider the following quantities:

$$j_\lambda^F(f) := \inf\{j \in \mathbb{N}, F(f)(j) \leq \lambda\}, \qquad \lambda_n = \frac{1}{\sqrt{n}}, \qquad j_n^F = j_{\lambda_n}^F(f).$$

Then, we say that the estimator $\hat{f}$ satisfies an oracle inequality on $\mathcal{V}$, associated with a sequence of estimators $\hat{f}_j$ and the functional F at the rate $c_n = 1 + \log n$ if the following two inequalities are true for all $n \geq 1$:

$$\mathbb{E}_n \|\hat{f} - f\|_p^p \leq C c_n (2^{j_n^F/2} \lambda_n)^p \qquad \forall f \in \mathcal{V},$$

$$\|E_{j_\lambda^F(f)} f - f\|_p^p \leq C' (2^{j_\lambda^F(f)/2} \lambda)^p \qquad \forall f \in \mathcal{V},\ \forall \lambda > 0.$$

It seems to us that if we take as starting point an estimate satisfying an oracle inequality of the above type, $\Pi = \{\Sigma_k,\ k = 1, \ldots, N\}$ with

$$\Sigma_k = \{f, F(f)(j) \leq 2^{-\alpha_k j},\ \forall j \geq 0\},$$

we can reasonably guess that a RNF can be produced. Do the authors agree with this guess? We also think that if N were large enough, then the RNF also would be a fairly good indicator of the class Σ_k.

REFERENCES

[1] FARAWAY, J. (1990). Bootstrap selection of bandwidth and confidence bands for nonparametric regression. *J. Statist. Comput. Simulation* **37** 37–44.

[2] HALL, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* **22** 215–232.

[3] HALL, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20** 675–694.

[4] KERKYACHARIAN, G. and PICARD, D. (2000). Minimax or maxisets? Technical report, Univ. Paris VI and Paris VII.

[5] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554.

[6] NEUMANN, M. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* **23** 1937–1959.

[7] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335.

UER MATHÉMATIQUES 45-55-5EME
UNIVERSITÉ PARIS VII—DENIS DIDEROT
2, PLASE JUSSIEU
75251 PARIS CEDEX 05
FRANCE
E-MAIL: picard@math.jussieu.fr

DISCUSSION

ALEXANDRE B. TSYBAKOV

Université Paris VI

The main and important contribution of this paper is in developing the concept of optimal random normalizing factors (RNFs). A special case (RNF taking two values) had been introduced earlier by Lepski [8]. The theory of optimal RNF can be viewed as an attempt to consider estimation, testing and confidence regions within a unique scheme. Although it can be applied to parametric problems as well, the paper focuses on nonparametrics, arguing in the rate optimality terms natural for nonparametric function estimation.

A simple motivation for the optimal RNF theory comes from the well-known practice of pretesting. One often does testing to choose a simple model and to estimate in this model afterwards. In such a procedure the statistician typically evaluates the errors of testing and estimation separately and does not try to specify the error of the compound procedure. Using RNF, Marc Hoffmann and Oleg Lepski suggest a hybrid test–estimation measure of the error of such a compound procedure, when a finite number of models are to be tested (RNF takes a finite number of values equal to the number of tested hypotheses). They also define the concept of optimality with respect to this new hybrid error measure. Two main applications of the optimal RNF approach suggested in the paper are (i) adaptive nonparametric confidence sets and (ii) adaptive estimators on nonstandard (nonnested) functional classes appearing in anisotropic regression (ANR).

1. Adaptive confidence sets. Most of the previous work was devoted to nonparametric confidence intervals at a fixed point. The methods either were based on undersmoothed estimators (and thus handled only the variance term, the bias being negligible) or used empirical estimates of the bias, assuming higher smoothness of the function f than the one for which the bias expression was obtained. In both cases valid confidence intervals were obtained, but they were applicable to suboptimal (undersmoothed) estimators. A limitation of this type turns out to be unavoidable for confidence intervals at a fixed point. As shown by Low [11], it is impossible to construct adaptive confidence intervals for nonparametric estimators at a fixed point conserving optimal rates of convergence. This means, in particular, that if one admits that f might perhaps have only β derivatives, the expected length of the pointwise confidence interval must be of the order $n^{-\beta/(2\beta+1)}$, even if f is infinitely differentiable. Marc Hoffmann and Oleg Lepski show that for confidence intervals in the L_2-norm the situation is more optimistic. In particular, an adaptive confidence set can be constructed as the L_2-ball around a rate optimal (and adaptive) estimator, and the data-dependent radius of this ball, expressed in terms of RNF, becomes of smaller order as the

smoothness of f increases [see (11)]. Thus, unlike the pointwise case, in the L_2-case we have a possibility to construct confidence sets that are "honestly" adaptive to the smoothness of f. However, this important result is not developed enough to get implementable recipes. Several points are left open:

1. The bound in (11) depends on M and $\hat{\rho}_\varepsilon$. How do we compute these values? If we turn to Theorem 2 for the concrete example of ANR, the value $\hat{\rho}_\varepsilon = \rho^*_\varepsilon$ is explicitly defined but it depends on L and $\boldsymbol{\beta}$, which are unknown in practice, while for M we have only an asymptotic expression $\tilde{M}^*$ depending on the same unknown parameters.
2. The derivation of (11) from (10) is based on Markov's inequality: this is a rough method; the bounds are certainly poor. Some exponential inequalities should be used instead, at least for the Gaussian case.
3. The values of the RNF $\hat{\rho}_\varepsilon$ in (11) are discretized: $\hat{\rho}_\varepsilon$ takes $N+1$ values, where N is the number of candidate classes Σ_j. The result also depends on the choice of the envelope class Σ. If the discretization is very rough, one gets poor confidence sets. How do we discretize and how do we choose the envelope set Σ? In the ANR example, the discretization is done w.r.t. the unknown directions $\mathbf{i}_s$, but not w.r.t. the unknown smoothness, which is certainly not fair in the confidence sets context. Ideally, one would like Σ to be as large as possible, for example, the entire L_2, while one would like the Σ_j's as small as posssible, up to $\Sigma_j = \{f_0\}$, where f_0 is an individual function. This means that the discretization given by Σ_j should be very fine. It is mentioned after (12) that $N = N_\varepsilon$ can grow as $\varepsilon \to 0$. What are the limitations?

A related idea is "honest confidence sets" in L_2 suggested by Li [10]. He constructs confidence sets based on Stein's unbiased risk estimator. The construction of Li is "honest" in the sense that his confidence bounds are valid for all functions f in L_2; there is no adaptation to the smoothness. A price for such a generality is that the intervals are wide; the L_2-radius of the ball is $\sim n^{-1/4}$. Li studied only the rates and did not provide the expressions for coverage probabilities in a ready-to-use form. Nevertheless, in his framework this might be possible with some work, since the constants appearing there are absolute. It is interesting to compare the size of confidence sets in [10] with those obtained by Marc Hoffmann and Oleg Lepski. Intuitively, one would think that the results should match when the envelope set Σ approaches L_2 (i.e., $\beta \to 0$). But this is not the case. Let, for example, $d = 1$, $N = 1$, and let Σ_1 be a class with smoothness $\beta' \gg \beta$. Then the RNF of Hoffmann and Lepski giving the size of the confidence set is of the order

$$z_n = \max\left\{\left(\frac{\sqrt{\ln(1/\alpha)}}{n}\right)^{2\beta/(4\beta+1)}, n^{-\beta'/(2\beta'+1)}\right\}$$

in the "best" case where Σ_1 is accepted. In the worst case the confidence sets are even wider. Since $\beta' \gg \beta$, we have $z_n \sim n^{-2\beta/(4\beta+1)}$ (up to a log-factor), and thus $z_n \gg n^{-1/4}$ whenever $\beta < 1/4$. This suggests that the confidence sets of [10] become preferable, as the envelope set Σ becomes large (at least, $\beta < 1/4$).

2. Adaptive estimation on general scales of classes. It is well known that the usual adaptation schemes (Mallows C_p, cross-validation, wavelet thresholding, Lepski's scheme etc.) achieve minimax adaptivity on various functional classes, such as one-dimensional or isotropic multidimensional Hölder, Sobolev or Besov classes. In this usual context, adaptation to the unknown smoothness is realized, and the functional classes are nested or ordered in a certain sense by the values of their smoothness parameters. For the ANR setting, Marc Hoffmann and Oleg Lepski do not consider adaptation to unknown smoothness (the smoothness parameters for all the directions are fixed and known), but rather adaptation to unknown direction $\mathbf{i}_s$. However, their general theory of Section 2 does not exclude a possibility of adaptation to both smoothness and direction. Another crucial point is that Hoffmann and Lepski do not aim to recover the "true" direction $\mathbf{i}_s$, but *a* direction having the same complexity $\frac{1}{\beta_{i_1}} + \cdots + \frac{1}{\beta_{i_s}}$ as the true one, and hence yielding the same convergence rate. So, more precisely, one should speak about adaptation to the rate of the unknown true direction $\mathbf{i}_s$ and not to the direction itself.

For each class $\Sigma(\mathbf{i}_s)$ there is a rate optimal projection estimator $\hat{\theta}_\varepsilon^{(0)}(\mathbf{i}_s)$. The Hoffmann–Lepski adaptation procedure selects a data-driven $\mathbf{i}^*$ and prescribes using the estimator $\hat{\theta}_\varepsilon^{(0)}(\mathbf{i}^*)$. The definition of $\mathbf{i}^*$ can be written in the form

$$\mathbf{i}^* = \arg\min_{\mathbf{i}_s \in \mathcal{N}} \varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s) \tag{1}$$

if the set $\mathcal{N} = \{\mathbf{i}_s : T_\varepsilon(\mathbf{i}_s) \le \lambda^2 \varphi_\varepsilon^2(\alpha_\varepsilon, \mathbf{i}_s)\}$ is not empty. If $\mathcal{N}$ is empty, the procedure prescribes using $\hat{\theta}_\varepsilon$, a rate optimal projection estimator on the envelope set Σ.

Writing $\mathbf{i}^*$ in the form (1) allows us to see that the Hoffmann–Lepski procedure is a member of the family of adaptation rules that can be called *pretesting aggregation*. These rules are defined as follows. Assume that our aim is to select an element $\mathbf{i}$ from a given set $\mathcal{T}$ of general nature. For each $\mathbf{i}$ associate a test T_i and say that $\mathbf{i}$ is accepted if the test statistic is smaller than a certain threshold. Introduce a partial ordering $\preceq$ on the set $\mathcal{T}$, and define the aggregated value

$$\mathbf{i}^* = \min\nolimits_{\preceq}\{\text{accepted } \mathbf{i}\},$$

where $\min_{\preceq}$ denotes the minimum w.r.t. the ordering $\preceq$. A general recipe is that: (i) the relation $\preceq$ should order in the sense of increasing variance terms or, more precisely, guaranteed upper bounds for stochastic error terms; (ii) the test T_i should reject if the appropriately chosen test statistic is larger than the corresponding guaranteed bound.

For example, the Hoffmann–Lepski procedure (if we discard the envelope class Σ which is certainly ballast in the adaptation context) is a special case of pretesting aggregation where "$\mathbf{i}_s$ is accepted" means $\mathbf{i}_s \in \mathcal{N}$, and $\mathbf{i}_s \preceq \mathbf{i}'_s$ means $\varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}_s) \le \varphi_\varepsilon(\alpha_\varepsilon, \mathbf{i}'_s)$. Another special case of pretesting aggregation is the Lepski adaptation scheme [7] where $\mathbf{i}$ is a smoothness parameter, the tests T_i check

whether pairwise distances between estimators exceed a threshold and the ordering is defined by usual inequalities between smoothness parameters. Examples of pretesting aggregation with nonstandard ordering, related to multivariate anisotropic settings, appear in [6], [9], [12]. These inherit the test structure close to the original Lepski scheme [7]: pairwise comparisons between estimators (or between estimators and pseudoestimators) are used to define the tests T_i. Note that for the Hoffmann–Lepski procedure the tests T_i are not of this type. The situation becomes even less standard in the statistical learning context where no explicit separation of risk into the bias and variance terms is available. Pretesting aggregation works here as well [14], however, with a quite different definition of tests: they should check nonemptiness of certain "insensitivity sets" of the empirical risk functional.

I guess the choice of the ANR model in the paper is mainly motivated by pedagogical reasons, as an example where the concretization of the optimal RNF theory can be done with minimal technicalities. From this point of view, the choice is perfect. However, the model is so nice and so particular that the adaptive estimators can be obtained in a better fashion without resorting to the RNF machinery. In fact, consider the blockwise Stein method. Define the values

$$T_1 = 5, \qquad \rho_\varepsilon = 1/\log(1/\varepsilon), \qquad N_{\max} = \lfloor \varepsilon^{-2} \rfloor,$$
$$T_m = \lfloor T_1(1+\rho_\varepsilon)^{m-1} \rfloor, \qquad m = 2, \dots, J-1, \qquad T_J = N_{\max} - \textstyle\sum_{j=1}^{J-1} T_j,$$
$$J = \min\left\{ m : T_1 + \sum_{j=2}^{m} \lfloor T_1(1+\rho_\varepsilon)^{j-1} \rfloor \ge N_{\max} \right\}$$

and introduce the weakly geometrically increasing sets of integers

$$A_1 = \{k \in \mathbf{N} : k \le T_1\}, \quad A_m = \left\{ k \in \mathbf{N} : \sum_{j=1}^{m-1} T_j < k \le \sum_{j=1}^{m} T_j \right\}, \quad m = 2, \dots, J.$$

For $\mathbf{m} = (m_1, \dots, m_d) \in \mathbf{N}^d$ such that $1 \le m_j \le J$ define the blocks

$$B_{\mathbf{m}} = \{\mathbf{k} \in \mathbf{N}^d : k_j \in A_{m_j},\ j = 1, \dots, d\}.$$

Consider the sequence model (25) of Hoffmann and Lepski. The blockwise Stein estimator of $\theta = (\theta_{\mathbf{k}},\ \mathbf{k} \in \mathbf{N}^d)$ is defined as $\tilde\theta_\varepsilon = (\tilde\theta_{\varepsilon,\mathbf{k}},\ \mathbf{k} \in \mathbf{N}^d)$, where

$$(2) \quad \tilde\theta_{\varepsilon,\mathbf{k}} = \lambda_{\mathbf{k}} Y_{\mathbf{k}}, \qquad \lambda_{\mathbf{k}} = \left(1 - \frac{\varepsilon^2 |B_{\mathbf{m}}|}{\sum_{\mathbf{k} \in B_{\mathbf{m}}} Y_{\mathbf{k}}^2}\right)_+, \qquad \mathbf{k} \in B_{\mathbf{m}},\ \mathbf{m} \in \{1, \dots, J\}^d,$$

and $\lambda_{\mathbf{k}} = 0$ if $\mathbf{k}$ does not belong to any of the $B_{\mathbf{m}}$'s. Let

$$\Sigma(\mathbf{i}_s) = \left\{ \theta : \sum_{\mathbf{k} \in I(\mathbf{i}_s)} \theta_{\mathbf{k}}^2 \left(1 + \sum_{k=1}^{d} k_i^{2\beta_i}\right) \le L^2,\ \theta_{\mathbf{k}} = 0,\ \mathbf{k} \notin I(\mathbf{i}_s) \right\},$$

and denote $\|\theta\|^2 = \sum_{\mathbf{k} \in \mathbf{N}^d} \theta_{\mathbf{k}}^2$, $\varphi_\varepsilon(\Sigma(\mathbf{i}_s)) = \varepsilon^{2\beta(\mathbf{i}_s)/(2\beta(\mathbf{i}_s)+1)}$. The following result holds.

PROPOSITION. *The estimator* $\tilde{\theta}_\varepsilon$ *defined by* (2) *satisfies*

$$\limsup_{\varepsilon\to 0}\sup_{\theta\in\Sigma(\mathbf{i}_s)}\mathbf{E}_\theta^\varepsilon\Big[\big(\varphi_\varepsilon^{-1}\big(\Sigma(\mathbf{i}_s)\big)\|\tilde{\theta}_\varepsilon-\theta\|\big)^2\Big]<\infty,$$

for every $1\le s\le d$, *every* $\boldsymbol{\beta}=(\beta_1,\dots,\beta_d)$ *with* $\beta_i>0$ *and every* $L>0$.

PROOF. Using the well-known inequality for blockwise linear oracles (see, e.g., [3], [5], [13]), we get

$$\mathbf{E}_\theta^\varepsilon\|\tilde{\theta}_\varepsilon-\theta\|^2\le S+4\varepsilon^2J^d+\sum_{\mathbf{k}\in\overline{B}}\theta_{\mathbf{k}}^2, \tag{3}$$

where $\overline{B}=\mathbf{N}^d\setminus\bigcup_{\mathbf{m}}B_{\mathbf{m}}$ and

$$S=\sum_{\mathbf{m}}\frac{\varepsilon^2|B_{\mathbf{m}}|\|\theta\|_{(\mathbf{m})}^2}{\varepsilon^2|B_{\mathbf{m}}|+\|\theta\|_{(\mathbf{m})}^2}.$$

If $\theta\in\Sigma(\mathbf{i}_s)$, the last sum on the RHS of (3) is of a smaller order than $\varphi_\varepsilon^2(\Sigma(\mathbf{i}_s))$, in view of the definition of $N_{\max}$ and $\Sigma(\mathbf{i}_s)$. The sum S contains only the terms for which $\|\theta\|_{(\mathbf{m})}\ne 0$, that is, such that $B_{\mathbf{m}}\cap I(\mathbf{i}_s)\ne\varnothing$. These correspond to the vectors $\mathbf{m}=(m_1,\dots,m_d)$ with components $m_j=1$ for $j\notin I(\mathbf{i}_s)$ [note that there are $d-s$ components $j\notin I(\mathbf{i}_s)$]. For such $\mathbf{m}$ we have $|B_{\mathbf{m}}|=T_1^{d-s}\prod_{j\in I(\mathbf{i}_s)}T_{m_j}=5^{d-s}|B_{s,\mathbf{m}}|$, where $B_{s,\mathbf{m}}$ denotes the trace of the block $B_{\mathbf{m}}$ on the space of indices $I(\mathbf{i}_s)$. Now, cut a hyperrectangle in the space of indices $I(\mathbf{i}_s)$ with sides $K_j=\varepsilon^{-2\beta(\mathbf{i}_s)/[\beta_{k_j}(2\beta(\mathbf{i}_s)+1)]}$, $j\in I(\mathbf{i}_s)$, and denote $M_V=\{\mathbf{m}:m_j=1,\,j\notin I(\mathbf{i}_s)$, and $\sum_{l=1}^{m_j}T_l\le K_j,\forall j\in I(\mathbf{i}_s)\}$, $M_B=\{\mathbf{m}:m_j=1,\,j\notin I(\mathbf{i}_s)$, and $\exists j\in I(\mathbf{i}_s):\sum_{l=1}^{m_j}T_l>K_j\}$. For $\theta\in\Sigma(\mathbf{i}_s)$ we get, for ε small enough,

$$\begin{aligned}S&\le\varepsilon^2 5^{d-s}\sum_{\mathbf{m}\in M_V}|B_{s,\mathbf{m}}|+\sum_{\mathbf{m}\in M_B}\|\theta\|_{(\mathbf{m})}^2\\&\le\varepsilon^2 5^{d-s}\prod_{j\in I(\mathbf{i}_s)}K_j+\sum_{\mathbf{k}\in I(\mathbf{i}_s):\ \exists k_j\ge K_j/2}\theta_{\mathbf{k}}^2,\end{aligned}$$

where for the last sum we used that $T_{m_j}\le(1/2)\sum_{l=1}^{m_j}T_l$ if ε is small enough. This is the usual bias–variance decomposition for the projection estimator with correct parameters for the class $\Sigma(\mathbf{i}_s)$. Using the definition of $\Sigma(\mathbf{i}_s)$ and of K_j we find

$$\sup_{\theta\in\Sigma(\mathbf{i}_s)}S=O\big(\varepsilon^{4\beta(\mathbf{i}_s)/[2\beta(\mathbf{i}_s)+1]}\big),\qquad\varepsilon\to 0.$$

To complete the proof, it remains to substitute this into (3) and to use that $J\le C\log^2 1/\varepsilon$ for some $C>0$ (see [13], Chapter 3). $\square$

An advantage of this result, as compared to Corollary 2 of Hoffmann and Lepski, is that $\boldsymbol{\beta}, s$ and L need not be known (the estimator $\tilde{\theta}_\varepsilon$ does not depend on these values); and that the adaptation is achieved simultaneously to the dimension s, to the direction $\mathbf{i}_s$ and to the smoothness $\boldsymbol{\beta}$. I believe that an even stronger result holds: not only the rate but also the exact minimax constant on every $\Sigma(\mathbf{i}_s)$ is attained with the above blockwise method (and/or with some other adaptation schemes); the techniques of [1–4] can be helpful to prove this. Note also that a special role of the envelope class Σ of "full dimension" is not needed here: it is just an example of $\Sigma(\mathbf{i}_s)$ with $s = d$, and the result holds for this class as well.

Of course, such luck is due to the fact that the problem is very particular. For instance, if we replace the Sobolev ellipsoids $\Sigma(\mathbf{i}_s)$ by the Hölder classes in the space of functions, the above construction is no longer valid, while the RNF approach is still applicable. Also, the RNF approach suggests (in principle) some confidence intervals, while for the blockwise estimators they are not known. It would be interesting to see other applications of the concept of optimal RNF, which is certainly a challenging issue.

REFERENCES

[1] CAVALIER, L., GOLUBEV, YU., PICARD, D. and TSYBAKOV, A. B. (2000). Oracle inequalities for inverse problems. *Ann. Statist.* To appear.

[2] CAVALIER, L. and TSYBAKOV, A. B. (2000). Sharp adaptation for inverse problems with random noise. *Probab. Theory Related Fields*. To appear. Available at www.proba.jussieu.fr.

[3] CAVALIER, L. and TSYBAKOV, A. B. (2001). Penalized blockwise Stein's method, monotone oracles and sharp adaptive estimation. *Math. Methods Statist.* **10** 247–282.

[4] GOLDENSHLUGER, A. and TSYBAKOV, A. (2001). Adaptive prediction and estimation in linear regression with infinitely many parameters. *Ann. Statist.* **29** 1601–1619.

[5] JOHNSTONE, I. M. (1998). Function estimation in Gaussian noise: sequence models. (Draft of a monograph; available at http://www-stat.stanford.edu/.)

[6] KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields*. **121** 137–170.

[7] LEPSKI, O. V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.

[8] LEPSKI, O. V. (1999). How to improve the accuracy of estimation. *Math. Methods Statist.* **8** 441–486.

[9] LEPSKI, O. V. and LEVIT, B. YA. (1999). Adaptive nonparametric estimation of smooth multivariate functions. *Math. Methods Statist.* **8** 344–370.

[10] LI, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008.

[11] LOW, M. G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554.

[12] POLZEHL, J. and SPOKOINY, V. G. (1999). Image denoising: pointwise adaptive approach. Preprint, Weierstrass Inst., Berlin.

[13] TSYBAKOV, A. B. (2001a). *Introduction à l'estimation non-paramétrique*. Unpublished manuscript (book, submitted for publication).

[14] TSYBAKOV, A. B. (2001b). Optimal aggregation of classifiers in statistical learning. Preprint 682, Laboratoire de Probabilités et Modèles Aléatoires, Univ. Paris VI and Paris VII. (Available at http://www.proba.jussieu.fr/mathdoc/preprints/index.html#2001.)

LABORATOIRE DE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UNIVERSITÉS PARIS 6
4 PL. JUSSIEU, BP 188
F-75252 PARIS CEDEX 05
FRANCE
E-MAIL: tsybakov@ccr.jussieu.fr

DISCUSSION

SARA VAN DE GEER

University of Leiden

1. What is an optimal rate? One of the most important issues in statistics is to go beyond point estimation, by estimating in addition the quality (error) of an estimator. This paper discusses this issue in the context of nonparametric curve estimation.

Rates of convergence of estimators in nonparametric models commonly depend not just on the model, but also on the quantity to be estimated. Therefore, the question arises how to define optimality of rates, as well as whether rates can be estimated.

The optimal rate of convergence in the minimax sense is defined as the best possible rate in the worst possible case. Thus, minimax rates are pessimistic and do not reward adaptive procedures. The challenge this paper takes up is to introduce a concept of optimality of random rates which extends the minimax approach, but does notice the possibility of faster rates at certain points (and lets this be known to the statistician).

Let us introduce some notation, adapted to the example we will consider later on, and for that reason slightly different from the notation used in the paper. Consider an observation $\mathbf{X}_\varepsilon$ from a probability distribution $\mathbf{P}_\theta$, where the unknown parameter θ is in a given parameter space Θ. Let $\hat{\theta}_\varepsilon$ be an estimator of θ. To settle things, let us suppose that $(\Theta, \|\cdot\|)$ is a normed vector space and that we are interested in the squared error $\|\hat{\theta}_\varepsilon - \theta\|^2$. The minimax mean squared error is

$$R_\varepsilon^2 = \inf_{\hat{\theta}_\varepsilon} \sup_{\theta \in \Theta} \mathbf{E}_\theta \|\hat{\theta}_\varepsilon - \theta\|^2.$$

The minimax rate of convergence ρ_ε is defined by

$$0 < \liminf_{\varepsilon \to 0} \frac{R_\varepsilon}{\rho_\varepsilon} \leq \limsup_{\varepsilon \to 0} \frac{R_\varepsilon}{\rho_\varepsilon} < \infty.$$

Clearly, the minimax rate depends on the model: $\rho_\varepsilon = \rho_\varepsilon(\Theta)$. Suppose now that we allow the rates to depend on the particular θ, say $\rho_{\theta,\varepsilon}$. It does not make sense to define optimal parameter dependent rates without referring to a particular model Θ. The reason is of course that the estimator $\hat{\theta}_\varepsilon \equiv \theta_0$ is optimal when θ happens to be equal to θ_0 (with rate $\rho_{\theta_0,\varepsilon} = 0$), but is useless for any other value of θ. In other words, one needs to have good behavior of an estimator for all possible values of θ, or even *uniformly* in θ. The authors have chosen the minimax approach: attention is restricted to estimators $\hat{\theta}_\varepsilon$ that attain the minimax rate ρ_ε, uniformly in $\theta \in \Theta$. There is, however, one drawback of this approach: the requirement of *uniformity* means that Θ has to be relatively small, because otherwise, the estimated rates will still be large.

Let $\{\hat{\theta}_\varepsilon\}$ be an estimator that attains the minimax rate uniformly on Θ. Suppose that, for some subset $\Theta_1 \subset \Theta$, the rate is actually faster:

$$\lim_{\varepsilon \to 0} \sup_{\theta \in \Theta_1} \frac{\mathbf{E}_\theta \|\hat{\theta}_\varepsilon - \theta\|^2}{\rho_\varepsilon^2} = 0.$$

For example, Θ_1 could be a single point $\{\theta\}$ (or even *all* singletons $\{\theta\}$). More interesting perhaps is the case where Θ_1 is an infinite subset. In that case, the best one can hope for is that $\hat{\theta}_\varepsilon$ converges with the minimax rate $\rho_\varepsilon(\Theta_1)$ uniformly on Θ_1. Two questions arise:

1. Does such an adaptive estimator exist? If not, how do we define the optimal rate on Θ_1?
2. Are lucky rates observable?

In the paper, both questions are addressed by considering so-called optimal random (in fact, estimated) normalizing factors (RNFs) $\hat{\rho}_\varepsilon$, which depend on $\mathbf{X}_\varepsilon$ and hence on θ via the distribution of $\mathbf{X}_\varepsilon$. A RNF $\hat{\rho}_\varepsilon$ is defined as optimal if it does not exceed the minimax rate and (roughly speaking) is as small as possible with large probability on Θ_1. Adaptivity occurs if the RNF does not exceed the minimax rate $\rho_\varepsilon(\Theta_1)$ with large probability. The paper extends the definition of optimal RNFs to the case where one has several (possibly nonnested) subsets $\Theta_1, \ldots, \Theta_N$. It provides optimal RNFs for general N by a canonical construction using optimal RNFs for each Θ_i.

The definition as given in the paper, of the optimal RNF, requires that

$$\limsup_{\varepsilon \to 0} \sup_{\theta \in \Theta} \mathbf{E}_\theta \left(\frac{\|\hat{\theta}_\varepsilon - \theta\|^2}{\hat{\rho}_\varepsilon^2} \right) < \infty.$$

Thus, the optimal RNF can be used to produce a confidence interval for θ. However, there is a conceptual problem here. The confidence interval consists of all θ for which $\|\hat{\theta}_\varepsilon - \theta\| \leq C\hat{\rho}_\varepsilon$, with C some constant, so this is a ball $B(R)$ in Θ, with radius $R = C\hat{\rho}_\varepsilon$. Clearly, the smaller R, the happier one is. On the other hand, Θ can be quite large, in which case $B(R)$ is also large, even if R is

small. This indicates that optimal RNFs are really different from optimal adaptive confidence sets, in the sense that small R can still mean large $B(R)$. In other words, the confidence set deduced from the optimal RNF seems to be conservative in testing $H\colon\ \theta \in \Theta_1$. (See also Sections 2 and 3 below, where the relation with testing is discussed further.) In curve estimation, constructing pointwise (adaptive) confidence intervals or (adaptive) confidence intervals in sup-norm is unfortunately very difficult. In any case, it appears that the confidence intervals considered in this paper have little in common with, for example, those in Picard and Tribouley [4] or Dümbgen and Spokoiny [2].

2. Anisotropic regression and other examples. The paper considers anisotropic regression as an example. Let me translate this problem in its most simple form. Suppose one has observations $\mathbf{X}_\varepsilon = \{X_{k,l,\varepsilon}\}$ with

$$X_{k,l,\varepsilon} = \theta_{k,l} + \varepsilon\xi_{k,l}, \qquad k, l \in \{1, 2, \ldots\},$$

where the $\xi_{k,l}$ are i.i.d. $\mathcal{N}(0, 1)$, and where

$$\{\theta_{k,l}\} \in \Theta = \left\{ \sum_{k=1}^{\infty} \sum_{l=1}^{\infty} \theta_{k,l}^2 (k^{2\beta_1} + l^{2\beta_2}) \le L^2 \right\}$$

and

$$\Theta_1 = \{\theta \in \Theta : \theta_{k,l} = 0 \text{ for all } l > 1\}.$$

Let $\|\cdot\|$ be the Euclidean norm. The minimax rate for this model is $\rho_\varepsilon(\Theta) = (\varepsilon^2)^{\beta/(2\beta+1)}$ on Θ, where β is defined by $1/\beta = 1/\beta_1 + 1/\beta_2$, and rate $\rho_\varepsilon(\Theta_1) = (\varepsilon^2)^{\beta_1/(2\beta_1+1)}$ on Θ_1. The authors show that adaptation occurs if $\beta_1 < 2\beta$ (i.e., $\beta_1 < \beta_2$). Otherwise, the optimal RNF is (apart from logarithmic factors) of order $(\varepsilon^2)^{2\beta/(4\beta+1)}$ on Θ_1. Thus, if Θ is large, optimal RNFs are also large. This seems to be the price to pay for requiring *uniformity* in Θ, that is, for the minimax approach.

Before briefly sketching the methodology used by the authors to arrive at this result, let me recall that most adaptive estimation methods, for example, the very general method with dimension penalties as considered in Barron, Birgé and Massart [1], tend to underestimate the complexity of a model. Thus, such estimation methods cannot be used to estimate the rate. It appears that one has to take a second order correction into account.

Consider the hypothesis $H\colon\ \theta \in \Theta_1$. Take as test statistic

$$T_\varepsilon = \sum_{k=1}^{N_1} \sum_{l=2}^{N_2} (X_{k,l,\varepsilon}^2 - \varepsilon^2),$$

where N_1 and N_2 are properly chosen. Let $\tilde{N} = N_1 N_2$, and let ϕ_ε^2 be the critical value for the test. Roughly speaking, and modulo logarithmic factors, T_ε behaves

like $\varepsilon^2\sqrt{\tilde{N}}$ under H, as a consequence of the central limit theorem. Therefore, we should also take ϕ_ε^2 of this order. Now, an optimal RNF will be

$$\hat{\rho}_\varepsilon = \begin{cases} \phi_\varepsilon \vee \rho_\varepsilon(\Theta_1), & \text{if } H \text{ is accepted,} \\ \rho_\varepsilon(\Theta), & \text{if } H \text{ is rejected.} \end{cases}$$

To have that this RNF is of the right order when H is not true but nevertheless accepted, one has to take care of the bias, that is, take $\tilde{N}$ large enough. Thus, the trade-off is now between $\varepsilon^2\sqrt{\tilde{N}}$ and squared bias, leading to choosing $\tilde{N}$ of order $(\varepsilon^2)^{2/(4\beta+1)}$ (again modulo logarithmic factors).

This procedure is as in Lepski [3]. Let me briefly describe it using a sequence space formulation. Consider the model

$$X_{k,\varepsilon} = \theta_k + \varepsilon\xi_k, \qquad k \in \{1, 2, \ldots\},$$

with $\theta \in \Theta$. Consider projection estimators

$$\hat{\theta}_k^{\mathcal{N}} = \begin{cases} X_{k,\varepsilon}, & \text{if } k \in \mathcal{N}, \\ 0, & \text{else.} \end{cases}$$

The variance–bias decomposition is now

$$\mathbf{E}_\theta \|\hat{\theta}^{\mathcal{N}} - \theta\|^2 = \varepsilon^2 |\mathcal{N}| + \sum_{k \notin \mathcal{N}} \theta_k^2. \tag{1}$$

Now, let us assume without loss of generality that choosing $\mathcal{N}$ as the first N_ε indexes k gives the optimal mean square error (for linear projection estimators) on Θ. Denote the estimator by $\hat{\theta}_{N_\varepsilon}$, where for general N, $\hat{\theta}_N = \hat{\theta}^{\{1,\ldots,N\}}$.

Let $\Theta_1 \subset \Theta$, and suppose that on Θ_1 one can in fact do with the first K_ε indexes k, where $K_\varepsilon \leq N_\varepsilon$; that is, $\hat{\theta}_{K_\varepsilon}$ gives the smallest mean square error on Θ_1. Take the following as test statistic for $H : \theta \in \Theta_1$:

$$T_\varepsilon = \|\hat{\theta}_{\tilde{N}_\varepsilon} - \hat{\theta}_{K_\varepsilon}\|^2 - \varepsilon^2(\tilde{N}_\varepsilon - K_\varepsilon),$$

with critical value ϕ_ε^2 (modulo logarithmic factors) of order $\varepsilon^2\sqrt{\tilde{N}_\varepsilon}$. One has to trade off $\varepsilon^2\sqrt{\tilde{N}_\varepsilon}$ against the squared bias [instead of the usual variance–bias trade-off (1)]. Adaptation occurs if K_ε is of larger order than $\sqrt{\tilde{N}_\varepsilon}$.

For example, suppose

$$\Theta = \left\{ \sum_{k=1}^{\infty} \theta_k^2 k^{2\beta} \leq L^2 \right\}$$

and

$$\Theta_1 = \left\{ \sum_{k=1}^{\infty} \theta_k^2 k^{2\beta_1} \leq L^2 \right\},$$

with $\beta_1 > \beta$. Then N_ε is of order $(\varepsilon^2)^{1/(2\beta+1)}$ and K_ε is of order $(\varepsilon^2)^{1/(2\beta_1+1)}$, whereas $\sqrt{\tilde{N}_\varepsilon}$ is of order $(\varepsilon^2)^{1/(4\beta+1)}$ (modulo logarithmic factors).

3. How about applications? Constructing confidence intervals using the RNFs is as yet not practically feasible. One of the reasons is of course that the appropriate constants are not known (yet). But also, the temptation to use such a confidence interval to test the hypothesis $\theta \in \Theta_1$ (say) probably should be suppressed. It will be more natural to apply the test that was used to construct the confidence interval. In fact, the problem as stated in the paper, that "you know adaptive estimators converge very fast if the function is very smooth (or has prescribed complexity), but you can tell nothing about the estimated function itself" (Section 1) is not solved by using RNFs, because very unsmooth functions may still have very small RNFs. In fact, one wants to test the hypothesis

$$H\colon\ f \text{ is } \textit{not} \text{ smooth},$$

a notoriously hard problem.

A limitation of the approach used is that it assumes only a finite number of possibilities $\Theta_1, \ldots, \Theta_N$. From Corollary 1 in the paper, one may deduce that this can be generalized to the case where N grows as $\varepsilon \to 0$. However, I have no insight into the case when condition (17) of this corollary is met. The example also requires knowledge of the constant L, the radius of the Sobolev ball. This clearly also prohibits actual applications.

Thus, RNFs provide a new and natural generalization of the minimax approach, and they generate and recall many puzzling and persistent questions as well!

REFERENCES

[1] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.

[2] DÜMBGEN, L. and SPOKOINY, V. G. (2001). Multiscale testing of qualitative hypothesis. *Ann. Statist.* **29** 124–152.

[3] LEPSKI, O. V. (1999). How to improve the accuracy of estimation. *Math. Methods Statist.* **8** 441–486.

[4] PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence intervals for pointwise curve estimation. *Ann. Statist.* **28** 298–335.

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF LEIDEN
P.O. BOX 9512
2300 RA LEIDEN
THE NETHERLANDS
E-MAIL: geer@math.leidenuniv.nl

REJOINDER

M. HOFFMANN AND O. LEPSKI

Université Paris VII and Université de Provence

We are grateful to all the participants for their stimulating comments and insightful questions. Reading the notes of the contributers, we progressively came to a better understanding of the imperfections of our approach. It also has given an orientation for further efforts: trying to answer their questions, we found several interesting problems for future work. We also believe and hope that this discussion will be relevant for potential readers since it somehow represents the state of modern asymptotic theory in nonparametric statistics. In our rejoinder we mainly focus on the criticism that was raised throughout the discussion. We will develop hereafter some complementary results that can presumably be obtained in the same line as the proofs of the paper.

What do we mean by "improvement of accuracy"? The main concern which appeared in almost all the contributions—Birgé, Efromovich, Tsybakov, van de Geer—is that the initial space Σ [or $\Sigma(\boldsymbol{\beta}, L)$ for anisotropic regression] is known a priori, and this knowledge is used for all constructions presented in the paper. Before addressing this objection, let us separate our results into two groups:

1. ***Construction of RNFs and of $\boldsymbol{\alpha}$-adaptive estimators***—We really need to know Σ. We understand the improvement of the accuracy as a random normalizing factor (RNF) which is "better" than the minimax rate of convergence on Σ. We can interpret our results as follows: the optimal RNF and α-adaptive estimators are relevant if Σ is not "too large." For particular models, we show that if Σ is rather large, the possible improvement may exist, but is negligible for realistic sample sizes.
2. ***Lower bounds***—We can even be more precise with the last statement. Usually, one can prove that the characteristic $x_\varepsilon(\hat{\rho}_\varepsilon, i)$ of a RNF $\hat{\rho}_\varepsilon$ providing finiteness of the maximal risk over Σ does not exceed either the minimax rate of convergence on Σ_i or the minimax testing rate, say r_ε, corresponding to the following hypothesis testing problem. One needs to test the hypothesis $f \in \Sigma_i$ against the alternative $f \in \Sigma$ such that $\inf_{g\in\Sigma_i}\|f - g\| \geq r_\varepsilon$. Typically, this rate r_ε does not depend on Σ_i, but crucially depends on Σ: the larger Σ, the worse r_ε. This explains in particular why an adaptive version of our procedure ("w.r.t. Σ in all generality") is meaningless. For example, let us consider the white noise model with $\|\cdot\| = L_2$-norm, $N = 1$, and let $\Sigma = \Sigma(\beta, 1)$, $\beta > 0$, be a family of Sobolev balls in $L_2([0, 1])$ (we take unit balls for simplicity) and $\Sigma_1 =$ "space of constants." It is known (see [2]) that $r_\varepsilon = \varepsilon^{4\beta/(4\beta+1)}$. Suppose now that β is unknown. If one looks for $\hat{\rho}_\varepsilon$ and $\hat{f}_\varepsilon$ such that

$$\limsup_{\varepsilon\to 0} \sup_{f\in\Sigma(\beta,1)} E_f^\varepsilon\{(\hat{\rho}_\varepsilon)^{-2}\|\hat{f}_\varepsilon - f\|_2^2\} < \infty \qquad \forall \beta > 0,$$

then, necessarily,

$$\liminf_{\varepsilon\to 0} \frac{x_\varepsilon(\hat{\rho}_\varepsilon)}{\varepsilon^{4\beta/(4\beta+1)}} > 0 \qquad \forall \beta > 0.$$

This implies that

$$\frac{\text{“improved rate”}}{\text{“minimax rate of convergence”}} = \frac{x_\varepsilon(\hat{\rho}_\varepsilon)}{\varepsilon^{2\beta/(2\beta+1)}} \to \text{constant} \neq 0$$

when $\beta \to 0$.

In conclusion, a single procedure that reasonably improves in our sense the rate of convergence over an arbitrary Σ cannot exist in general! Even if we suppose that $\beta \geq \beta_{\min} > 0$ (here $\beta_{\min}$ is supposed to be known), then the improvement does exist but is ridiculous if $\beta_{\min}$ is small.

Thus, the knowledge of Σ is crucial, but even this is not sufficient to obtain satisfactory results. As already seen, Σ should not be “too huge.” A nice collection of such spaces in the context of multivariate estimation is developed by Brown and Lin in their Discussion. They elegantly show in a few lines how to extend our approach to the ANOVA setting.

Adaptive confidence sets. Birgé, Tsybakov, Picard and Tribouley and van de Geer have pointed out several interesting links between RNF and adaptive confidence sets. They also pose various thought-provoking questions. We think we have to try to shed some light, if we can, on some of their remarks. As we understand these remarks, the main concerns are as follows:

1. the fact that the proposed confidence sets explicitly depend on the knowledge of $\boldsymbol{\beta}$ and L for the case of anisotropic regression; in the abstract model, they depend on the initial space Σ;
2. the description of the confidence sets using L_2-norm;
3. the use of the Markov inequality to derive confidence sets.

Although our goal was not to provide adaptive confidence sets (ACSs), we realize that this issue is probably the most important aspect of the paper; by studying the links between RNF and ACS, we originally pursued the following objectives:

1. to “show that for confidence intervals in the L_2-norm the situation is more optimistic [*than for pointwise confidence intervals*]” (Tsybakov); in particular, to show that ACSs may exist in some cases;
2. to show that, even when the L_2-norm is considered, ACSs do not exist in a general setting;
3. to show that, if ACSs do not exist, the use of RNF may provide a smaller radius of the confidence set than the radius suggested by the minimax approach (see Section 1 in the discussion by Picard and Tribouley).

To be more specific, we again take the example we considered in the first section. Let $\Sigma(\beta, 1)$, $0 < \beta_{\min} \le \beta \le \beta_{\max} \le \infty$, be a family of Sobolev balls.

First, suppose that $\beta_{\max} < 2\beta_{\min}$. In this case, one can construct a RNF ρ_ε^* and an estimator f_ε^* such that the L_2 ball with center f_ε^* and a radius of the order ρ_ε^* is a fully ACS w.r.t. $\{\Sigma(\beta, 1),\ \beta_{\min} \le \beta \le \beta_{\max}\}$. It means that, for all $0 < \gamma < 1$ and some $C(\gamma, \beta_{\min})$, we have

$$\inf_{f \in \Sigma(\beta_{\min}, 1)} P_f^\varepsilon \{\| f_\varepsilon^* - f \|_2 \le C(\beta_{\min}, \gamma)\rho_\varepsilon^*\} \ge 1 - \gamma,$$

$$x(\rho_\varepsilon^*, \beta) \asymp \varepsilon^{2\beta/(2\beta+1)} \qquad \forall \beta \in [\beta_{\min}, \beta_{\max}].$$

To construct the pair $(\rho_\varepsilon^*, f_\varepsilon^*)$, we need to know how to handle the number N of hypotheses when $N = N_\varepsilon \to \infty$ as $\varepsilon \to 0$ if we want to use our canonical construction (this question was addressed simultaneously by Picard and Tribouley, by Tsybakov and by van de Geer). To do this with an abstract model, it is sufficient to replace our condition (17) in Section 2.2 by

$$(*) \qquad \lim_{\varepsilon \to 0} \sum_{i=1}^{N_\varepsilon} \sup_{f \in \Sigma} E_f^\varepsilon \{[\xi_{\varepsilon,i}(f)]^p 1_{\{\xi_{\varepsilon,i}(f) \ge M_i\}}\} = 0.$$

Assumption $(*)$ is very close to the sufficient condition in [4] for the existence of an optimal adaptive estimator. The sufficient condition in [4] is indeed verified in various nonparametric settings. (Compare the remarks in van de Geer's Section 3.)

In our example, the $(\rho_{\varepsilon,i}^*, f_{\varepsilon,i}^*)$ are the optimal RNF and α-adaptive estimator w.r.t. $\Sigma(\beta_i, 1)$, where $\beta_i = \beta_{\min} + \frac{i}{(\ln(1/\varepsilon))^2}$ and $\beta_i \in [\beta_{\min}, \beta_{\max}]$. In this case, it is readily seen that $N_\varepsilon \asymp (\ln(1/\varepsilon))^2$. The condition $(*)$ holds if one assumes that $\alpha_\varepsilon \to 0$ fast enough (see the assumption of Proposition 3).

Second, suppose that $\beta_{\max} > 2\beta_{\min}$. Repeating the proof of the lower bound given in the paper, one can show that an ACS does not exist anymore. To be more precise, if we understand an ACS as a ball $B(\hat{f}_\varepsilon, \hat{\rho}_\varepsilon)$ in L_2 with data-driven center $\hat{f}_\varepsilon$ and radius $\hat{\rho}_\varepsilon$, then for any $\gamma > 0$ small enough such that

$$\inf_{f \in \Sigma(\beta_{\min}, 1)} P_f^\varepsilon \{f \in B(\hat{f}_\varepsilon, \hat{\rho}_\varepsilon)\} \ge 1 - \gamma$$

we have

$$\sup_{f \in \Sigma(\beta, 1)} P_f^\varepsilon \{\hat{\rho}_\varepsilon > \text{const.}\ \varepsilon^{4\beta_{\min}/(4\beta_{\min}+1)}\} \ge \kappa \qquad \forall \beta \ge 2\beta_{\min},$$

for some $\kappa > 0$. It remains to note that $\varepsilon^{4\beta_{\min}/(4\beta_{\min}+1)} \gg \varepsilon^{2\beta/(2\beta+1)}$ for $\beta > 2\beta_{\min}$.

Third, if $\beta_{\max}$ is arbitrary, we still obtain an optimistic result: we can construct $(\rho_\varepsilon^*, f_\varepsilon^*)$ such that, for all $0 < \gamma < 1$ and some $C(\gamma, \beta_{\min})$, we have

$$\inf_{f \in \Sigma(\beta_{\min}, 1)} P_f^\varepsilon \{\| f_\varepsilon^* - f \|_2 \le C(\beta_{\min}, \gamma)\rho_\varepsilon^*\} \ge 1 - \gamma,$$

$$x(\rho_\varepsilon^*, \beta) \asymp \left\{ \begin{array}{ll} \left(\varepsilon\sqrt{\dfrac{1}{\ln \varepsilon}}\right)^{4\beta_{\min}/(4\beta_{\min}+1)}, & \beta \in (2\beta_{\min}, \beta_{\max}] \\ \varepsilon^{2\beta/(2\beta+1)}, & \beta \in [\beta_{\min}, 2\beta_{\min}] \end{array} \right\} \ll \varepsilon^{2\beta_{\min}/(2\beta_{\min}+1)}.$$

Keeping this in mind, we can now answer the first concern of the contributors: we really need to know $\beta_{\min}$ (and hence the initial space Σ) because if $\beta_{\min}$ is too small (if $\beta_{\min} = 0$, then Σ is "roughly" the L_2-space), the best possible improvement is negligible.

Let us now briefly discuss the use of the Markov inequality and the choice of the L_2-norm to construct confidence sets. First, we absolutely agree with the remarks of Birgé and Tsybakov that it is much better to use exponential bounds for probabilities to obtain more precise coverage of the confidence sets; see also the remarks of Picard and Tribouley. The Markov inequality is just the simplest way to relate probability deviations and squared expectation if only rates of convergence are sought. However, the reason we describe a confidence set as a ball in some normed space is a direct consequence of the application of the theory of RNFs. The question of how to apply confidence sets of this type to real problems is a very delicate issue. Here we present some of our conjectures where and how one can apply such constructions. From our point of view, one of the possible problems is a confidence interval for a value of a function at a *random* point. To be more specific, let us consider the nonparametric AR(d) model

$$Y_i = f(Y_{i-1}, \dots, Y_{i-d}) + \xi_i, \qquad i = 0, \dots, n,$$

with centered i.i.d. random variables ξ_i (innovations) and let $f \in \Sigma$, where Σ is a typical class considered in a nonparametric setting. Imagine that we are able to extend our theory for this model; that is, we can find an optimal RNF ρ_n^* and an α-adaptive estimator such that

$$(**) \qquad \sup_{f\in\Sigma} E_f^n\{(\rho_n^*)^{-2}\|f_n^* - f\|_2^2\} \le M$$

for large enough n. A problem of interest (scrutinized in many papers) is to predict the value Y_{n+1}. Clearly (because the ξ_{n+1} and $Y_n, Y_{n-1}, \dots$ are independent and moreover the law of ξ_1 does not depend on f) this problem is in fact equivalent to finding an estimator of

$$f(Y_n, \dots, Y_{n-d+1}).$$

A standard approach is then to use the following predictor:

$$\hat{Y}_{n+1} = \hat{f}_n(Y_n, \dots, Y_{n-d+1}),$$

where $\hat{f}_n(\cdot)$ is an estimator of f. We claim that using $(**)$ we can construct a data-driven *confidence interval* for $f(Y_n, \dots, Y_{n-d+1})$ as follows. Put, for $c > 0$,

$$\Gamma_c = \{t \in \mathbb{R}^d : |f_n^*(t) - f(t)| \ge c\rho_n^*\}.$$

Clearly

$$\text{measure}(\Gamma_c) \le c^{-2}(\rho_n^*)^{-2}\|f_n^* - f\|_2^2$$

and therefore

$$\text{(i)} \qquad \sup_{f\in\Sigma} E_f^n\{\text{measure}(\Gamma_c)\} \le c^{-2}M$$

for large enough n. In view of (i) and under some additional assumptions imposed on the law of ξ_1, one can prove that

$$\text{(ii)} \qquad \sup_{f\in\Sigma} P_f^n\{(Y_n,\dots,Y_{n-d+1}) \notin \Gamma_c\} \le a(c),$$

for some $a(c) \to 0$ as $c \to \infty$. We conclude from (ii) that, for all $0 < \gamma < 1$, there exists $C(\gamma)$ such that

$$\inf_{f\in\Sigma} P_f^n\{f(Y_n,\dots,Y_{n-d+1}) \in I_n\} \ge 1-\gamma,$$

where

$$I_n = [f_n^*(Y_n,\dots,Y_{n-d+1}) - C(\gamma)\rho_n^*,\ f_n^*(Y_n,\dots,Y_{n-d+1}) + C(\gamma)\rho_n^*].$$

We end the Rejoinder by addressing individually some remarks, comments and questions of the participants that are not covered above.

L. Birgé. Asymptotic versus nonasymptotic: the main concern of Birgé is the "fundamentally asymptotic nature" of our results. Certainly, we agree that any *nonasymptotic* mathematical result (if available) is much deeper than an asymptotic one and mathematical statistics is no exception. However, whenever one speaks about investigations in the area of mathematical statistics there are usually two types of difficulties to overcome. First, one needs to present a statistical procedure and to compute its characteristics (risk, power function or something else). At this stage, the possibility of obtaining a nonasymptotic result is based on the technical abilities of the researcher and mostly depends on the complexity of the problem to be solved. If we look at the upper bound obtained in our Theorem 1 we will see that it is in fact a nonasymptotic result. It is just because "the model is so nice and so particular" (Tsybakov). We also believe that, for some other statistical models which are not so "ideal," a nonasymptotic version of our upper bound could be established as well. Unfortunately, the main difficulty is not here. Whenever a statistical procedure is proposed, one usually wants to compare it with other procedures. In other words one needs an optimality criterion. Indeed, the criterion we propose (Definition 2) has an "asymptotic nature," and we do not have any idea about its nonasymptotic version. We also do not know any satisfactory "nonasymptotic" criteria allowing comparison of statistical procedures.

L. D. Brown and Y. Lin. Is fully adaptive estimation possible in multidimensional problems? For minimax risk described by L_2-losses, the estimators adaptive w.r.t. the scale of anisotropic Hölder spaces as well as the scale of anisotropic Besov spaces (in other words w.r.t. $\boldsymbol{\beta}$ and L) were found in [1] and in [6] respectively. In these papers $s = d$ and the adaptation to unknown direction $\mathbf{i}_s$ is not considered. The discussion by Tsybakov shows in an elegant and thorough way how to construct an estimator which is fully adaptive to $\boldsymbol{\beta}$, L, s and $\mathbf{i}_s$ using the blockwise Stein method. Some results for the case of L_p-losses, $1 \le p \le \infty$, have been obtained recently in [5].

S. Efromovich. Efromovich contests our opinion that it is impossible to compute the accuracy of an adaptive estimator and gives the example where the finding of a random rate is not needed (see also the example of functional space in the discussion by Birgé). We remind the reader that our "categorical conjecture" was done for the problems where different (in order!) rates of convergence are possible. This is not the case for the examples considered by Efromovich and Birgé.

Does the HL (our) estimator improve the accuracy of estimation? In the first section above we explained how we understand the notion of "improvement." With this in mind we can state that our procedure is optimal in view of Definitions 2 and 3. If so, how do we comment on the example of the estimator (say, estimator A) given by Efromovich which "outperforms the HL estimator" (estimator B)? Indeed, estimator A is better but in view of another criterion. To be mathematically correct one should compare procedures using one and the same criterion! Perhaps estimator A satisfies Definitions 2 and 3. In this case one can say that estimators A and B are equivalent in view of this criterion. It is also possible that estimator A is not optimal in view of Definitions 2 and 3. Clearly, it does not mean that estimator A is less "accurate" than estimator B. In general, the following question could be asked: for a given estimator, say $\tilde{f}$, does one need to estimate its accuracy $\|\tilde{f} - f\|$? Some results in this direction are obtained in [3] for the case $\|\cdot\| = \|\cdot\|_2$.

What about random rates in anisotropic regression? Indeed, we consider this model only in the Introduction in order to clarify the problems to be solved. The presentation of this less "theoretical" model in the Introduction of our paper was suggested by an Associate Editor and the referees simultaneously. We believe that our result can be extended to this model for the case of i.i.d. (or weakly dependent) noises satisfying some moment condition.

D. Picard and K. Tribouley. What can be done for L_∞-losses? In view of the relationship with hypothesis testing as discussed in the first section above, we are very pessimistic about the possibility of extending our results to L_∞-losses.

What about RNFs and maxisets? We did not investigate this direction. The concept of a maxiset seems indeed very natural to us in the context of RNFs. Possibly, this could be related to a notion of *random maxiset*.

REFERENCES

[1] BARRON, A., BIRGÉ, L. and MASSART, P. (1999). Risk bounds for model selection via penalization. *Probab. Theory Related Fields* **113** 301–413.

[2] INGSTER, Y. I. (1993). Asymptotically minimax hypothesis testing for nonparametric alternatives, I–III. *Math. Methods Statist.* **2** 85–114, 171–189, 249–268.

[3] IOUDITSKY, A. and LEPSKI, O. (2001). Evaluation of the accuracy of nonparametric estimators. *Math. Methods Statist.* To appear.

[4] LEPSKI O. V. (1991). Asymptotic minimax adaptive estimation. 1. Upper bounds. *Theory Probab. Appl.* **36** 682–697.

[5] KERKYACHARIAN, G., LEPSKI, O. and PICARD, D. (2001). Nonlinear estimation in anisotropic multi-index denoising. *Probab. Theory Related Fields* **121** 137–170.

[6] NEUMANN, M. H. and VON SACHS, R. (1997). Wavelet threshholding in anisotropic function classes and application to adaptive estimation of evolutionary spectra. *Ann. Statist.* **25** 38–76.

LABORATOIRE DE PROBABILITÉS
ET MODÈLES ALÉATOIRES
UFR DE MATHÉMATIQUES, CASE 7012
CNRS-UMR 7599 ET UNIVERSITÉ PARIS VII
2 PLACE JUSSIEU, 75251 PARIS CEDEX 05
FRANCE
E-MAIL: hoffmann@math.jussieu.fr

LABORATOIRE D'ANALYSE,
TOPOLOGIE, PROBABILITÉS
CENTRE DE MATHÉMATIQUES
ET D'INFORMATIQUE
CNRS-UMR 6632
UNIVERSITÉ DE PROVENCE
39, RUE JOLIOT CURIE,
13453 MARSEILLE CEDEX 13
FRANCE
E-MAIL: lepski@cmi.univ-mrs.fr