On adaptive estimation using the *sup*-norm losses

G. Golubev, O. Lepski and B. Levit

Abstract

We consider the problem of recovering smooth functions from noisy data, using the *sup*-norm as the quality criterion. Starting with a natural projection estimator, we show a data-driven procedure to be adaptive asymptotically minimax.

1 Introduction

In this paper we study the problem of recovering an unknown function $f(t)$, $t \in [0,1]$, from noisy data

$$dY(t) = f(t)dt + \sigma db(t), \tag{1}$$

where $b(t)$ is a standard Wiener process. As the quality criterion of an estimator we will use the sup-norm

$$\|g\|_\infty = \sup_{t\in[0,1]} |g(t)|.$$

The optimality of an estimator can be assessed in several ways. Three main approaches have greatly influenced the development of modern theoretical statistics. With the minimax approach, one assumes that the unknown function $f(t)$ belongs to a given functional class $\mathcal{F}$ and that the performance of an estimator $\tilde{f}(t,Y)$ is characterized by its maximum risk

$$R(\tilde{f},\mathcal{F}) = \sup_{f\in\mathcal{F}} \mathbf{E}_f \|\tilde{f}(\cdot,Y) - f\|,$$

where $\|\cdot\|$ is a norm. The goal is to evaluate the minimax risk $R(\mathcal{F}) = \inf_{\tilde{f}} R(\tilde{f},\mathcal{F})$ and to construct an estimator $f^*(t,Y)$ that approaches this quantity. Usually one tries to find the so-called asymptotically minimax estimator f^* such that

$$\lim_{\sigma\to 0} \frac{R(f^*,\mathcal{F})}{R(\mathcal{F})} = 1.$$

The solutions are known for a selection of functional classes $\mathcal{F}$. Pinsker (1980) solved the problem using $\mathbf{L}_2$-losses and ellipsoidal restrictions. In the case of the Hölder classes

$\mathcal{F} = H(\beta, L)$, the solution for the *sup*-norm was obtained by Korostelev (1993) and Donoho (1994) who demonstrated that an asymptotically minimax estimator f^* can be found among the kernel smoothers. Another example is the class $\mathcal{F} = A(\tau, L)$ of functions analytic inside the strip around the real axis in the complex plain of size 2τ (Golubev, Tsybakov and Levit (1996)). The asymptotically minimax estimator was again a (spatial) kernel smoother which at the same time could be viewed as a projection estimator in the frequency domain.

In general, the asymptotically minimax estimators depend on the functional classes $\mathcal{F}$ at hand. Thus in the case of the Hölder classes both the optimal bandwidth and the shape of the optimal kernel depend on the parameters (β, L). In the case of analytic functions the asymptotically minimax estimator depends on the parameter τ. Since in practice precise information about the functional classes $\mathcal{F}$ is hardly ever available, applications of the "purely" minimax approach are very restricted.

To overcome this difficulty, a popular model selection approach is often used, associated with so-called "oracles". In contrast to the minimax approach, where the functional class is assumed to be given and no restrictions are imposed on the family of estimators, now the class of estimators, say $\mathcal{E} = \{f_h(x, Y), h \in H\}$, is chosen beforehand. For example, $\mathcal{E}$ may be the family of kernel estimators

$$f_h(x, Y) = \frac{1}{h} \int_0^1 K\left(\frac{x-t}{h}\right) dY(t),$$

where the bandwidth parameter h describes the family. The objective is to choose the "best" estimator within the family $\mathcal{E}$, which is of the same quality as the "oracle estimator" that achieves $\inf_{\tilde{f} \in \mathcal{E}} \mathbf{E}_f \|\tilde{f} - f\|$. Substantial progress in this area has been achieved recently in the case of the $\mathbf{L}_2$-losses, following the pioneering papers by Akaike (1973) and Mallows (1973). For the state of the art in this area see Nemirovskii (1998) and Barron, Birgé and Massart (1999).

The third approach can be seen as intermediate between the model selection and the minimax approach. It is often called the adaptive or functional scale approach. Here we are dealing with an appropriate family of functional classes $\mathcal{F}_\alpha$, $\alpha \in \Sigma$. We will call an estimator f^* *adaptive asymptotically minimax* if for any $\alpha \in \Sigma$

$$\lim_{\sigma \to 0} \frac{R(f^*, \mathcal{F}_\alpha)}{R(\mathcal{F}_\alpha)} = 1.$$

The adaptive approach has an obvious advantage over the minimax approach: the adaptive minimax estimator no longer depends on a particular functional class $\mathcal{F}_\alpha$.

It is perhaps natural that in constructing such estimators, one concentrates first of all on the functional scales for which the exact asymptotics of the minimax risk $R(\mathcal{F}_\alpha)$ are known. Even for such scales, finding the adaptive asymptotically minimax estimators may be quite challenging. Efroimovich and Pinsker (1984) obtained such estimators using $\mathbf{L}_2$-losses and Sobolev classes. A different approach to this problem was proposed by Golubev and Nussbaum (1990). A typical trait of such models is that the loss function $\|\tilde{f} - f\|$ has an asymptotically degenerate distribution.

However, adaptive minimax estimators do not always exist. For instance, they do not exist when the unknown function belongs to the Hölder scale $H(\beta, L)$, $\beta \in (0,1)$, $L > 0$, and the *sup*-norm losses are used (Lepski (1992)). This can be explained intuitively from the fact that for any given (β, L) both the kernel and the bandwidth of the asymptotically minimax estimator are derived from a strict balance between the bias and the variance (Korostelev (1993), Donoho (1997)). One therefore has little freedom in constructing such estimators, so an asymptotically minimax estimator for one Hölder class cannot at the same time be optimal for any other Hölder class.

In situations where the adaptive asymptotically minimax estimators do not exist, one can study both the optimal rates of convergence and the exact asymptotics of the best adaptive estimators. A great deal of work has been done on establishing the optimal rates of convergence of adaptive estimators (Lepski (1991)). In some cases asymptotically optimal adaptive estimators have been found, even in situations where explicit minimax estimators are still not known (see for instance Lepski and Spokoiny (1997), Tsybakov (1998)). Optimal adaptive pointwise estimators for analytic functions have been studied by Lepski and Levit (1998, 1999). A particular feature of the functional classes studied in the last two papers is that in the case of known functional classes the asymptotically minimax estimators were asymptotically unbiased.

The problem studied in the present paper bears both the hallmarks of the functional scales for which adaptive minimax estimators have been constructed so far. The variance of the estimators we are studying dominates their bias and in addition the *sup*-norm losses $\|\tilde{f} - f\|_\infty$ exhibit degenerate asymptotic behavior.

Our functional scales will be introduced in terms of the Fourier coefficients, with regard to orthonormal bases in $\mathbf{L}_2(0,1)$ that possess some additional properties. Examples of such functional scales include harmonic polynomials and periodic analytic functions. The goal of the present paper is to construct adaptive asymptotically minimax estimators for the scales of such functional classes, with respect to the *sup*-norm. To achieve this objective we start with the projection-type estimators. By obvious analogy with band limitation, we will refer to the number of terms in a projection estimator as its *bandwidth.*

2 Main results

In this paper we study projection estimators with data-driven bandwidths. To specify such estimators one needs, first of all, a basis. Let $\varphi_k(x)$ be an orthonormal basis in the

Hilbert space $\mathbf{L}_2(0,1)$ equipped with the inner product

$$<f,g> = \int_0^1 f(t)g(t)\,dt.$$

Any function $f \in \mathbf{L}_2(0,1)$ can then be expanded into the Fourier series

$$f(t) = \sum_{k=0}^{\infty} <f,\varphi_k>\varphi_k(t).$$

A natural way to recover an unknown function $f(t)$ from the noisy data (1) is to use the projection estimator

$$\widehat{f}_W(t,Y) = \sum_{k=1}^{W} <\dot{Y},\varphi_k>\varphi_k(t), \tag{2}$$

where

$$<\dot{Y},\varphi_k>= \int_0^1 \varphi_k(t)\,dY(t).$$

The number W of terms in (2) will be called the *bandwidth* of the projection estimator. If W is a non-integer, we will adopt the following convention: $\sum_{W_1}^{W_2} = \sum_{\lceil W_1 \rceil}^{\lfloor W_2 \rfloor}$.

Since the performance of $\widehat{f}_W(t,Y)$ strongly depends on W, choosing the bandwidth is one of the primary statistical problems. A projection estimator can easily be split into the stochastic or variance term and the non-random part or bias. Accordingly, the risk of $\widehat{f}_W(t,Y)$ is controlled by

$$\mathbf{E}\|\widehat{f}_W - f\|_\infty \le \sigma \mathbf{E} \sup_{t\in[0,1]} \left| \sum_{k=1}^{W} \xi_k \varphi_k(t) \right| + \sup_{t\in[0,1]} \left| \sum_{k=W+1}^{\infty} <f,\varphi_k>\varphi_k(t) \right|, \tag{3}$$

where ξ_k are i.i.d. standard normal $\mathcal{N}(0,1)$.

To evaluate the variance term, we will impose some additional conditions on the basis $\varphi_k(x)$. First, we will assume that the functions $\varphi_k(x)$ are bounded uniformly in k and x: $\sup_{x\in[0,1],k} |\varphi_k(x)| < \infty$. In addition, some properties of $\varphi_k(t)$ will be formulated in terms of the 'incomplete reproducing kernel'

$$K_r^N(t,s) = \frac{1}{N} \sum_{l\in[rN,N]} \varphi_l(t)\varphi_l(s), \qquad r,t,s \in [0,1].$$

We will assume that for any $r \in [0,1]$ and $N \to \infty$ the kernel $K_r^N(t,s)$ satisfies the following conditions:

- for any $\gamma > 0$ uniformly in $t \in [N^{-1+\gamma}, 1 - N^{-1+\gamma}]$

$$K_r^N(t,t) = 1 - r + o(1); \tag{4}$$

- for any $\gamma > 0$ and some positive constant C uniformly in $t, s \in [N^{-1+\gamma}, 1 - N^{-1+\gamma}]$
$$\left|K_r^N(t,s)\right| \le \frac{C}{N|(t-s)|}; \tag{5}$$

- uniformly in $s \in [0,1]$
$$\left|\frac{\partial K_r^N(t,s)}{\partial t}\right|_{t=s} = o(N); \tag{6}$$

- uniformly in $s \in [0,1]$
$$\left|\frac{\partial^2 K_r^N(t,s)}{\partial s \partial t}\right|_{t=s} = O(N^2). \tag{7}$$

Later on we will need the following result, which can be derived e.g. from Cramér and Leadbetter (1967), Ch. 13.5, p. 294 or Lifshits (1995), Ch. 14, p. 203.

Lemma 1 *Let the assumptions (4–6) be satisfied and ξ_k be i.i.d. $\mathcal{N}(0,1)$. Then there exists a constant C such that all N and $r \in [0,1)$*

$$\mathbf{P}\left\{\frac{1}{\sqrt{N(1-r)}} \sup_{t\in[0,1]} \left|\sum_{k=Nr}^{N} \xi_k \varphi_k(t)\right| > x\right\} \le CN(1-r)\exp(-x^2/2).$$

Let $D(x) = \sqrt{2x \log x}$. Using the lemma, the variance term of the projection estimator can be bounded by

$$\mathbf{E} \sup_{t\in[0,1]} \left|\sum_{k=1}^{W} \xi_k \varphi_k(t)\right| \le (1+o(1))D(W), \quad W \to \infty. \tag{8}$$

In order to control the bias, we assume that the unknown function $f(x)$ belongs to the scale of functional classes

$$\mathcal{F}_W = \left\{f : \quad \sum_{k=W+1}^{\infty} |\langle f, \varphi_k \rangle| \le \sigma\rho(W)D(W)\right\}. \tag{9}$$

Here $\rho(x) \to 0$ as $x \to \infty$ is an arbitrary fixed function, which however is typically unknown to the statistician. The parameter W, also unknown, effectively determines the dimensionality of the statistical problem. One of the typical features of non-parametric problems is increasing dimensionality. Therefore we will assume that $W \ge W_{\min}$ for some $W_{\min} \to \infty$.

Now, by using (3), (8) and (9) the risk of the estimator $\widehat{f}_W(t, Y)$, is bounded, uniformly over $\mathcal{F}_W$, by

$$\sup_{f\in\mathcal{F}_W} \mathbf{E}\|\widehat{f}_W - f\|_\infty \le (1+o(1))\sigma D(W), \quad W_{\min} \to \infty. \tag{10}$$

It is not difficult to show (cf. section 5) that the projection estimator (2) is asymptotically minimax with respect to the corresponding class $\mathcal{F}_W$, i.e.

$$\inf_{\widehat{f}} \sup_{f \in \mathcal{F}_W} \mathbf{E}\|\widehat{f} - f\|_\infty = (1 + o(1))\sigma D(W), \quad W_{\min} \to \infty, \tag{11}$$

where *inf* is taken over all possible estimators.

Remark 1. Many different orthonormal systems satisfy the above conditions (4) — (7). In a sense, they are all close to the classical trigonometric bases, for which these conditions can be verified by simple algebra. One such system that is often used consists of the eigenfunctions of the following boundary value problem (see e.g. Dunford and Schwartz (1971), Sect. XIX.4)

$$\begin{aligned} (-1)^m \frac{d^{2m}}{dt^{2m}} \varphi_k(t) - \lambda_k \varphi_k(t) = 0, \\ \varphi_k^{(l)}(0) = \varphi_k^{(l)}(1), \quad l = m, \dots, 2m-1. \end{aligned}$$

In particular, one obtains for $m = 1$ the well-known cosine basis $\varphi_0(t) = 1$, $\varphi_k(t) = \sqrt{2}\cos(\pi k t)$.

Remark 2. At first sight, our definition of the functional classes $\mathcal{F}_W$ may seem somewhat artificial. To shed some additional light on this definition, consider the following two examples. Assume first that $\rho(x) \equiv 0$. In this case $\mathcal{F}_W$ is the linear space spanned by functions $\{\varphi_0(t), \dots, \varphi_W(t)\}$. For a less trivial example, assume that $f \in A(\tau, L)$, where

$$A(\tau, L) = \left\{ f : \sum_{k=0}^{\infty} <f, \varphi_k>^2 \exp(2\tau k) \le L^2 \right\},$$

and $\varphi_k(t)$ is the ordinary trigonometric basis. By the Cauchy-Schwartz inequality

$$\sum_{k=W}^{\infty} |<f, \varphi_k>| \le L \exp(-\tau W)(1 - \exp(-2\tau))^{-1/2}$$

and, therefore, $A(\tau, L) \subseteq \mathcal{F}_W$, with $W = \tau^{-1}\log(L/\sigma)$ and

$$\rho(W) = D^{-1}(W)(1 - \exp(-2\tau))^{1/2}.$$

Note that our assumption $W \ge W_{\min}$ is then equivalent to the following a priori restrictions $\tau \le \tau_{\max}$, $L \ge L_{\min}$, with some $\tau_{\max} < \infty$ and $L_{\min} > 0$. In this case $W_{\min} = \tau_{\max}^{-1} \log(L_{\min}/\sigma) \to \infty$ when $\sigma \to 0$.

The above-mentioned optimality property of the estimator $\hat{f}_W$ referred to the situation where the classes $\mathcal{F}_W$ were known. We now turn to the situation where W is unknown except for $W \ge W_{\min}$. Our adaptive estimator in this case will be constructed as follows. Let

$$w_k = W_{\min}(1 + \delta)^k, \quad k = 0, ..., \log(W_{\min})$$

be an exponential grid, where $\delta > 0$ is a sufficiently small number. Consider the family of projection estimators

$$\widehat{f}_{w_k}(t,Y) = \sum_{l=0}^{w_k} <\dot{Y}, \varphi_l> \varphi_l(t), \quad k = 0, ..., \log(W_{\min}).$$

Our goal is to select the "best" estimator within this family. To achieve this objective we use the method proposed by Lepski (1991), which consists of comparing the differences $\|\widehat{f}_{w_k} - \widehat{f}_{w_l}\|_\infty$. When $f \in \mathcal{F}_W$ and $w_l > w_k \geq W$, both the biases of $\widehat{f}_{w_k}$ and $\widehat{f}_{w_l}$ will be much smaller than $\sigma D(w_l - w_k)$ (cf. (9)). Therefore by Lemma 1 the random variables $\|\widehat{f}_{w_k} - \widehat{f}_{w_l}\|_\infty$ with a high probability do not exceed $(1+\delta)\sigma D(w_l - w_k)$. We therefore arrive at the following "estimator" of W

$$\widehat{W} = \min\left\{w_k : \|\widehat{f}_{w_k} - \widehat{f}_{w_l}\|_\infty \leq (1+\delta)\sigma D(w_l - w_k) \text{ for all } l > k\right\}.$$

The adaptive estimator is then simply

$$f^*(t,Y) = \widehat{f}_{\widehat{W}}(t,Y).$$

The main difficulty in analyzing this estimator is connected with evaluating part of its risk corresponding to the event $\widehat{W} < W$. This difficulty is caused by, first, the use of the *sup*-norm and, second, the need to demonstrate that no extra losses are involved in having to estimate the unknown bandwidth W or, in other words, that one doesn't have to pay a price for the adaptation. The following theorem represents the main result of the paper.

Theorem 1 *For any $C_0 \geq 1$ uniformly in $W \in [W_{\min}, C_0 W_{\min}]$ and σ as $W_{\min} \to \infty$*

$$R(f^*, \mathcal{F}_W)/[\sigma D(W)] \leq 1 + O(\delta).$$

Since δ is arbitrary and can be chosen slowly converging to zero as $W_{\min} \to \infty$, one can conclude that, according to (11), the estimator f^* is adaptive asymptotically minimax. Note that this estimator does not depend on the constant C_0 appearing in the Theorem.

3 Auxiliary results

The following result will play a key role in the proof of Theorem 1. Consider two independent Gaussian random processes

$$\eta_1(t) = \frac{1}{\sqrt{N}} \sum_{i=N+1}^{\alpha N} \xi_i \varphi_i(t), \quad \eta_2(t) = \frac{1}{\sqrt{N}} \sum_{i=0}^{N} \xi_i \varphi_i(t),$$

where $\alpha > 1$ and ξ_i are i.i.d. $\mathcal{N}(0,1)$.

Lemma 2 *For any sufficiently small $\varepsilon > 0$, uniformly in $g \in \mathbf{R}^1$ and $T \subseteq [0,1]$ as $N \to \infty$*

$$\mathbf{P}\Big\{\sup_{t\in T} |g+\eta_1(t)| \le A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{\sup_{t\in T} |g+\eta_2(t)| \ge A_\varepsilon^-(N)\Big\} = o(1), \tag{12}$$

where $A_\varepsilon^+(N) = (1+\varepsilon)\sqrt{2(\alpha-1)\log N}$, $A_\varepsilon^-(N) = (1-\varepsilon)\sqrt{2\alpha\log N}$.

Proof. Without loss of generality one can assume that $g > 0$ and that the set T is closed since the functions $\eta_1(t)$, $\eta_2(t)$ are continuous. It is clear that for any $s \in T$

$$\mathbf{P}\Big\{\sup_{t\in T} |g+\eta_1(t)| \le A_\varepsilon^+(N)\Big\} \le \mathbf{P}\Big\{g+\eta_1(s) \le A_\varepsilon^+(N)\Big\}.$$

Since the functions $\varphi_l^2(t)$ are uniformly bounded, $\mathbf{E}\eta_1^2(t) \le C$ for some $C > 0$. Therefore the right-hand side of the last inequality tends to 0 if

$$g > (1+2\varepsilon)\sqrt{2(\alpha-1)\log N}.$$

The case

$$0 \le g \le (1+2\varepsilon)\sqrt{2(\alpha-1)\log N}. \tag{13}$$

has to be considered next. Define the points $\{t_0, t_1, t_2, \dots, t_M\}$ in T as follows

$$t_0 = \min\{t \in T : t \ge 0\},\ \dots,\ t_{k+1} = \min\{t \in T : t > t_k + N^{\gamma-1}\},\ \dots,$$

where $\gamma > 0$ is a sufficiently small number. Obviously, $\{t_0, t_1, t_2, \dots, t_M\}$ is an $N^{\gamma-1}$-net in T. Let $T_k = \{t \in [0,1] : |t-t_k| \le N^{\gamma-1}/2\}$ and

$$\mathcal{T} = \bigcup_{k=1}^{M-1} T_k.$$

We have

$$\mathbf{P}\Big\{\sup_{t\in T} |g+\eta_1(t)| \le A_\varepsilon^+(N)\Big\} \le \mathbf{P}\Big\{g + \sup_{k=1,\dots M-1} \eta_1(t_k) \le A_\varepsilon^+(N)\Big\}$$
$$+\ \mathbf{P}\Big\{g+\eta_1(t_0) \le A_\varepsilon^+(N)\Big\}\ +\ \mathbf{P}\Big\{g+\eta_1(t_M) \le A_\varepsilon^+(N)\Big\}.$$

Therefore

$$\begin{aligned}
&\mathbf{P}\Big\{\sup_{t\in T} |g+\eta_1(t)| \le A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{\sup_{t\in T} |g+\eta_2(t)| \ge A_\varepsilon^-(N)\Big\} \\
&\le\ \mathbf{P}\Big\{g + \sup_{k=1,\dots M-1} \eta_1(t_k) \le A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{g + \sup_{t\in\mathcal{T}} |\eta_2(t)| \ge A_\varepsilon^-(N)\Big\} \\
&\quad + \mathbf{P}\Big\{g+\eta_1(t_0) \le A_\varepsilon^+(N)\Big\}\ \mathbf{P}\ \Big\{g + \sup_{t\in T_0} |\eta_2(t)| \ge A_\varepsilon^-(N)\Big\} \\
&\quad + \mathbf{P}\Big\{g+\eta_1(t_M) \le A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{g + \sup_{t\in T_M} |\eta_2(t)| \ge A_\varepsilon^-(N)\Big\}.
\end{aligned} \tag{14}$$

The last two terms in the above inequality are small. Indeed using (13) and Lemma 1 one obtains

$$\begin{aligned}
&\mathbf{P}\Big\{g+\eta_1(t_M)\le A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{g+\sup_{t\in T_M}|\eta_2(t)|\ge A_\varepsilon^-(N)\Big\} \qquad (15)\\
&\le\ \mathbf{P}\Big\{\sup_{t\in T_M}|\eta_2(t)|\ge\sqrt{2\log N}[\sqrt{\alpha}(1-\varepsilon)-\sqrt{\alpha-1}(1+2\varepsilon)]\Big\}=o(1)
\end{aligned}$$

and similarly

$$\mathbf{P}\Big\{g+\eta_1(t_0)\le A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{g+\sup_{t\in T_0}|\eta_2(t)|\ge A_\varepsilon^-(N)\Big\}=o(1). \qquad (16)$$

To bound the first term on the right side of (14), we will approximate the process $\eta_1(t)$ on the set $\{t_1,t_2,\dots,t_{M-1}\}$ by a sequence of independent Gaussian random variables. Consider Gaussian vector η, where $\eta_k=\eta_1(t_k),\ k=1,\dots,M-1$. Let B be the covariance matrix of η. One can represent η as $\eta=\sqrt{B}\xi$, where ξ is $\mathcal{N}(0,E)$, the matrix $\sqrt{B}$ is positive semidefinite and $\sqrt{B}\sqrt{B}=B$. Therefore

$$\eta=\sqrt{\operatorname{diag}B}\,\xi+\zeta,$$

where $\zeta=(\sqrt{B}-\sqrt{\operatorname{diag}B})\xi$ and the diagonal matrix $\operatorname{diag}B$ has the entries $(\operatorname{diag}B)_{kk}=B_{kk}$. To evaluate $\mathbf{E}\,\zeta_k^2$, the following inequality can easily be checked

$$\begin{aligned}
&(\sqrt{B}-\sqrt{\operatorname{diag}B})(\sqrt{B}-\sqrt{\operatorname{diag}B})^T\\
&\quad\le\ (\operatorname{diag}B)^{-1/2}(B-\operatorname{diag}B)(B-\operatorname{diag}B)^T(\operatorname{diag}B)^{-1/2}.
\end{aligned}$$

It now follows from (5) that

$$\begin{aligned}
\mathbf{E}\,\zeta_k^2\ &\le\ C\sum_{j=1,j\ne k}^{M}B_{jk}^2\le\frac{C}{N^2}\sum_{j=1,j\ne k}^{M}\frac{1}{\sin^2\pi(t_j-t_k)}\\
&\le\ CN^{-2\gamma}\sum_{j=1,j\ne k}^{N^{1-\gamma}}\frac{1}{(j-k)^2}\le CN^{-2\gamma}.
\end{aligned}$$

Therefore any $x>0$

$$\mathbf{P}\Big\{\max_k|\zeta_k|>x\Big\}\le\sum_k\mathbf{P}\Big\{|\zeta_k|>x\Big\}\le N^{1-\gamma}\exp(-Cx^2N^{2\gamma})=o(1).$$

This together with (4) implies

$$\begin{aligned}
&\mathbf{P}\Big\{g+\sup_{k=1,\dots,M-1}\eta_1(t_k)\le A_\varepsilon^+(N)\Big\} \qquad (17)\\
&\le\ \Phi^{M-1}\left((1+\varepsilon)\sqrt{2\log N}-\frac{g}{\sqrt{\alpha-1}}\right)+o(1),
\end{aligned}$$

where $\Phi(x)$ is the standard normal cdf.

Next using Lemma 1 we get

$$\mathbf{P}\Big\{g + \sup_{t\in\mathcal{T}} |\eta_2(t)| \geq A_\varepsilon^-(N)\Big\} \leq C(M-1)N^\gamma \exp\left[-\frac{1}{2}(A_\varepsilon^-(N)-g)^2\right]. \tag{18}$$

We still have to check that the product of the terms appearing on the right-hand sides of (17) and (18) correspondingly tends to 0 as $N \to \infty$. Let us assume that the right-hand of (17) is bounded away from 0. Using the asymptotics of $\Phi(x)$ for large x and (13) we obtain

$$M-1 \leq C\sqrt{\log N}\exp\left[\frac{1}{2}\left((1+\varepsilon)\sqrt{2\log N} - \frac{g}{\sqrt{\alpha-1}}\right)^2\right].$$

With this the inequality (18) becomes

$$\begin{aligned}
&\mathbf{P}\Big\{g + \sup_{t\in\mathcal{T}^\gamma} |\eta_2(t)| \geq A_\varepsilon^-(N)\Big\} \\
&\leq\ CN^\gamma\sqrt{\log N}\exp\left[\frac{1}{2}\left((1+2\varepsilon)\sqrt{2\log N} - \frac{g}{\sqrt{\alpha-1}}\right)^2\right] \\
&\quad\times\ \exp\left[-\frac{1}{2}\left((1-\varepsilon)\sqrt{2\alpha\log N} - g\right)^2\right] \leq CN^\gamma\sqrt{\log N} \\
&\quad\times \exp\left\{2\log N\left[(1+2\varepsilon)^{1/2}(1-x)^2 - (1-\varepsilon)^{1/2}\left(\sqrt{\alpha} - x\sqrt{\alpha-1}\frac{1+2\varepsilon}{1-\varepsilon}\right)^2\right]\right\},
\end{aligned}$$

where $x = g[2\log N(\alpha-1)]^{-1/2}(1+2\varepsilon)^{-1} \leq 1$, cf. (13)). Thus to prove that the right-hand side of the last inequality is small, we only have to check that

$$\max_{x\in[0,1]} \Lambda(x) < 0,$$

where $\Lambda(x) = (1-x)^2 - (\sqrt{\alpha} - x\sqrt{\alpha-1})^2$. Obviously $\Lambda(x) = (2-\alpha)x^2 + \dots$ attains its maximum over $[0,1]$ at the boundary points when $\alpha \leq 2$. Note that $\Lambda(0) = 1-\alpha < 0$, and $\Lambda(1) = -(\sqrt{\alpha} - \sqrt{\alpha-1})^2 < 0$. When $\alpha > 2$, both roots of $\Lambda(x)$

$$x_{1,2} = \frac{\sqrt{\alpha}\pm 1}{\sqrt{\alpha-1}\pm 1}.$$

satisfy $x_{1,2} > 1$. Therefore $\Lambda(x) < 0$ for $x \in [0,1]$. This together with (14–16) proves Lemma 3.

Lemma 3 *For any $\varepsilon > 0$ uniformly in $g(t)$ as $N \to \infty$*

$$\mathbf{P}\Big\{\sup_{t\in[0,1]} |g(t)+\eta_1(t)| \leq A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{\sup_{t\in[0,1]} |g(t)+\eta_2(t)| \geq A_\varepsilon^-(N)\Big\} = o(1).$$

Proof. One may assume that

$$\sup_{t\in[0,1]} |g(t)| \le (1+2\varepsilon)\sqrt{2(\alpha-1)\log N}. \tag{19}$$

Indeed if

$$|g(t_0)| > (1+2\varepsilon)\sqrt{2(\alpha-1)\log N}$$

for some $t_0 \in [0,1]$, then

$$\begin{aligned}\mathbf{P}\Big\{\sup_{t\in[0,1]} |g(t)+\eta_1(t)| \le A_\varepsilon^+(N)\Big\} &\le \mathbf{P}\Big\{|g(t_0)+\eta_1(t_0)| \le A_\varepsilon^+(N)\Big\} \\ &\le \mathbf{P}\Big\{\eta_1(t_0) \le -\varepsilon\sqrt{2(\alpha-1)\log N}\Big\} \to 0.\end{aligned}$$

Now consider for a sufficiently small $\delta > 0$ the sequence $g_k = \sqrt{2\log N}k\delta$, with $|k| \le (1+2\varepsilon)\sqrt{(\alpha-1)}/\delta$ (cf. (19)). Let

$$T_k = \Big\{t \in [0,1] :\ |g_k - g(t)| \le \delta\sqrt{2\log N}/2\Big\}.$$

Then

$$\begin{aligned}&\mathbf{P}\Big\{\sup_{t\in[0,1]} |g(t)+\eta_1(t)| \le A_\varepsilon^+(N)\Big\}\mathbf{P}\Big\{\sup_{t\in[0,1]} |g(t)+\eta_2(t)| \ge A_\varepsilon^-(N)\Big\} \\ &\quad\le \sum_{|k|<\sqrt{\alpha}/\delta} \mathbf{P}\Big\{\sup_{t\in T_k} |g_k+\eta_1(t)| \le A_\varepsilon^+(N)\Big\} \\ &\qquad\quad \times \mathbf{P}\Big\{\sup_{t\in[0,1]} |g_k+\eta_2(t)| \ge A_\varepsilon^-(N)(1-\delta)\Big\} \\ &\quad\le \sum_{|k|<\sqrt{\alpha}/\delta} \mathbf{P}\Big\{\sup_{t\in T_k} |g_k+\eta_1(t)| \le A_\varepsilon^+(N)(1+\delta)\Big\} \\ &\qquad\quad \times \mathbf{P}\Big\{\sup_{t\in T_k} |g_k+\eta_2(t)| \ge A_\varepsilon^-(N)(1-\delta)\Big\}.\end{aligned}$$

This together with Lemma 2 proves Lemma 3.

4 Proof of the Theorem

Let $f \in \mathcal{F}_W$. Denote

$$\overline{W} = \min\{w_k : w_k \ge W\} := w_{\overline{k}}$$

an approximation of W by the the bandwidths nested in the exponential grid introduced in Section 2. Let $W_{\max} = w_{\lfloor \log(W_{\min}) \rfloor} \le W_{\min}^{1+\delta}$. We will split the risk of f^* into two parts:

$$R^- = \mathbf{E}_f \|f^* - f\|_\infty \mathbf{1}\Big\{\widehat{W} \le \overline{W}\Big\}, \quad \text{and} \quad R^+ = \mathbf{E}_f \|f^* - f\|_\infty \mathbf{1}\Big\{\widehat{W} > \overline{W}\Big\}. \tag{20}$$

First let us show that R^+ is small. Consider the following event

$$A = \{Y(t) : \ \|f^* - f\|_\infty \ge 2\sigma D(W_{\max})\}.$$

Now R^+ can be bounded as

$$R^+ \le \mathbf{E}_f \|f^* - f\|_\infty \mathbf{1}\Big\{\widehat{W} > \overline{W}\Big\}\mathbf{1}\Big\{A\Big\} + 2\sigma D(W_{\max})\mathbf{P}\Big\{\widehat{W} > \overline{W}\Big\}. \tag{21}$$

By the Cauchy-Schwartz inequality one obtains

$$\begin{aligned} &\mathbf{E}_f \|f^* - f\|_\infty \mathbf{1}\Big\{\widehat{W} > \overline{W}\Big\}\mathbf{1}\Big\{A\Big\} \le \mathbf{E}_f \|f^* - f\|_\infty \mathbf{1}\Big\{A\Big\} \\ &\le \sum_{\overline{k} < k \le \log(W_{\min})} \mathbf{E}_f^{1/2} \|\widehat{f}_{w_k} - f\|_\infty^2 \mathbf{P}^{1/2}\Big\{\|\widehat{f}_{w_k} - f\|_\infty \ge 2\sigma D(W_{\max})\Big\}. \end{aligned} \tag{22}$$

Next note that $\widehat{f}_{w_k}(t) - f(t) = b_{w_k}(t) + \sigma\eta_{w_k}(t)$, where

$$\eta_{w_k}(t) = \sum_{l=1}^{w_k} \xi_l \varphi_l(t), \quad b_{w_k}(t) = \sum_{l=w_k+1}^{\infty} \langle f, \varphi_l \rangle \varphi_l(t), \tag{23}$$

and ξ_l are i.i.d. $\mathcal{N}(0,1)$. By the definition of $\mathcal{F}_W$ (cf. (9)) we have $\|b_{w_k}\|_\infty \le \sigma D(W)/2$, since $w_k \ge \overline{W}$). Therefore

$$\|\widehat{f}_{w_k} - f\|_\infty \le \|\eta_{w_k}\|_\infty + \sigma D(W)/2. \tag{24}$$

Further by Lemma 1

$$\mathbf{P}\Big\{\|\eta_{w_k}\|_\infty \ge \sqrt{w_k}x\Big\} \le C w_k \exp(-x^2/2). \tag{25}$$

Combining (24), (25) we arrive at

$$\begin{aligned} &\mathbf{P}\Big\{\|\widehat{f}_{w_k} - f\|_\infty \ge 2\sigma D(C_0 W_{\min})\Big\} \le \mathbf{P}\Big\{\|\eta_{w_k}\|_\infty \ge 3\sigma D(w_k)/2\Big\} \\ &\le \ C w_k \exp(-9\log w_k/4) \le C w_k^{-5/4}. \end{aligned} \tag{26}$$

By similar reasoning we obtain for any $w_k > \overline{W}$

$$\mathbf{E}_f \|\widehat{f}_{w_k} - f\|_\infty^2 \le 4\sigma^2 w_k \log w_k.$$

Hence, by (22) and (26)

$$\mathbf{E}_f\|f^*-f\|_\infty \mathbf{1}\Big\{\widehat{W}>\overline{W}\Big\}\mathbf{1}\Big\{A\Big\} \le C\sigma W_{\max}^{1/4}\log^{1/2} W_{\max}. \tag{27}$$

The probability $\mathbf{P}\Big\{\widehat{W}>\overline{W}\Big\}$ has still to be evaluated. Note that by (9) for any $\delta>0$ and $w_k>\overline{W}$

$$\|b_{w_k}-b_{\overline{W}}\|_\infty \le 2\sigma\rho(W)D(W) < o(1)\delta\sigma D(w_k-\overline{W}).$$

Therefore using Lemma 1 one obtains

$$\begin{aligned}
\mathbf{P}\Big\{\widehat{W}>\overline{W}\Big\} &\le \sum_{k>i>\overline{k}} \mathbf{P}\Big\{\|\widehat{f}_{w_k}-\widehat{f}_{w_i}\|_\infty \ge \sigma(1+\delta)D(w_k-w_i)\Big\}\\
&\le \sum_{k>i>\overline{k}} \mathbf{P}\Big\{\|\eta_{w_k}-\eta_{w_i}\|_\infty \ge [1+\delta+o(1)]D(w_k-w_i)\Big\}\\
&\le \sum_{k>i>\overline{k}} (w_k-w_i)^{-\delta+o(1)} \le \sum_{k=1}^{\infty}[(1+\delta)^k-1]^{-\delta+o(1)}\sum_{i>\overline{k}} w_i^{-\delta+o(1)} \le CW^{-\delta+o(1)}\delta^{-4}.
\end{aligned}$$

Together with (21) and (27)) this finally gives the following bound

$$R^+ \le C\sigma W^{-\delta+o(1)}\delta^{-4}D(W_{\max}) = o(1)\sigma D(W). \tag{28}$$

Now let us consider the second term R^- of the risk. Let $B=\{Y(t):\|f^*-f\|_\infty \ge \sigma D(\overline{W})\}$. Now one can estimate R^- as follows

$$\begin{aligned}
R^- &= \mathbf{E}_f\|f^*-f\|_\infty \mathbf{1}\Big\{\widehat{W}\le\overline{W}\Big\}\mathbf{1}\Big\{B^c\Big\} + \mathbf{E}_f\|f^*-\|_\infty \mathbf{1}\Big\{\widehat{W}\le\overline{W}\Big\}\mathbf{1}\Big\{B\Big\} \qquad (29)\\
&\le \sigma D(\overline{W}) + \mathbf{E}_f\|f^*-f\|_\infty \mathbf{1}\{\widehat{W}\le\overline{W}\}\mathbf{1}\Big\{B\Big\}\mathbf{1}\Big\{\|f^*-f\|_\infty > 3\sigma D(W)\Big\}\\
&\quad + \mathbf{E}_f\|f^*-f\|_\infty \mathbf{1}\Big\{\widehat{W}\le\overline{W}\Big\}\mathbf{1}\Big\{B\Big\}\mathbf{1}\Big\{\|f^*-f\|_\infty \le 3\sigma D(W)\Big\}.
\end{aligned}$$

In the case of the event $\widehat{W}\le\overline{W}$ we have by the definition of f^*

$$\|f^*-\widehat{f}_{\overline{W}}\|_\infty \le (1+\delta)\sigma D(W).$$

Combined with the inequality $\|f^*-f\|_\infty \le \|f^*-\widehat{f}_{\overline{W}}\|_\infty + \|\widehat{f}_{\overline{W}}-f\|_\infty$ this gives a bound for the last term on the right-hand side of (29)

$$\begin{aligned}
&\mathbf{E}_f\|f^*-f\|_\infty \mathbf{1}\Big\{\widehat{W}\le\overline{W}\Big\}\mathbf{1}\Big\{B\Big\}\mathbf{1}\{\|f^*-f\|_\infty > 3\sigma D(W)\} \qquad (30)\\
&\le (1+\delta)\sigma D(\overline{W})\mathbf{P}_f\Big\{\|\widehat{f}_{\overline{W}}-f\|_\infty \ge (2-\delta)\sigma D(W)\Big\}\\
&\quad + \mathbf{E}_f\|\widehat{f}_{\overline{W}}-f\|_\infty \mathbf{1}\Big\{\|\widehat{f}_{\overline{W}}-f\|_\infty \ge (2-\delta)\sigma D(W)\Big\}.
\end{aligned}$$

Since on $\mathcal{F}_W$ the bias of $\widehat{f}_{\overline{W}}$ is bounded by $o(1)\sigma D(W)$, one can deduce that both right-hand terms are small. Indeed by Lemma 1

$$\mathbf{P}_f\Big\{\|\widehat{f}_{\overline{W}} - f\|_\infty \geq (2-\delta)\sigma D(W)\Big\} \leq CW^{-3+4\delta}$$

and

$$\begin{aligned}
&\mathbf{E}_f\|\widehat{f}_{\overline{W}} - f\|_\infty \mathbf{1}\Big\{\|\widehat{f}_{\overline{W}} - f\|_\infty \geq (2-\delta)\sigma D(W)\Big\} \\
&\leq \; C\sigma D(W) W \int_{2(1-\delta)\log^{1/2} W}^{\infty} \exp(-x^2/2)\, dx \leq C\sigma W^{-5/2+8\delta}.
\end{aligned}$$

Hence according to (30)

$$\mathbf{E}_f\|f^* - f\|_\infty \mathbf{1}\Big\{\widehat{W} \leq \overline{W}\Big\}\mathbf{1}\Big\{B\Big\}\mathbf{1}\Big\{\|f^* - f\|_\infty > 3\sigma D(W)\Big\} \leq C\sigma W^{-2}. \tag{31}$$

Consider now the last term in (29). We have

$$\begin{aligned}
&\mathbf{E}_f\|f^* - f\|_\infty \mathbf{1}\Big\{\widehat{W} \leq \overline{W}\Big\}\mathbf{1}\Big\{B\Big\}\mathbf{1}\Big\{\|f^* - f\|_\infty \leq 3\sigma D(W)\Big\} \\
&\leq \; 3\sigma D(W)\mathbf{E}_f \mathbf{1}\Big\{\widehat{W} \leq \overline{W}\Big\}\mathbf{1}\Big\{B\Big\} \\
&\leq \; 3\sigma D(W) \sum_{k<\overline{k}} \mathbf{E}_f \mathbf{1}\Big\{\|\widehat{f}_{w_k} - f\|_\infty \geq \sigma D(\overline{W})\Big\} \\
&\quad \times \mathbf{1}\Big\{\|\widehat{f}_{w_k} - f_{\overline{W}}\|_\infty \leq \sigma D(\overline{W} - w_k)\Big\}.
\end{aligned} \tag{32}$$

Note that by (23)

$$\Big\|\sum_{k=\overline{W}}^{\infty} <f,\varphi_k>\varphi_k(\cdot)\Big\|_\infty \leq o(1)D(W).$$

Let

$$g(t) = \sum_{k=w_k}^{\overline{W}} <f,\varphi_k>\varphi_k(t).$$

Since the processes $\widehat{f}_{w_k}$ and $\widehat{f}_{w_k} - f_{\overline{W}}$ are independent

$$\begin{aligned}
&\mathbf{E}_f \mathbf{1}\Big\{\|\widehat{f}_{w_k} - f\|_\infty \geq \sigma D(\overline{W})\Big\}\mathbf{1}\Big\{\|\widehat{f}_{w_k} - f_{\overline{W}}\|_\infty \leq \sigma D(\overline{W} - w_k)\Big\} \\
&\leq \; \mathbf{P}_f\Big\{\|\eta_{w_k} + g\|_\infty \geq (1-\delta)\sigma D(\overline{W})\Big\}\mathbf{P}_f\Big\{\|\eta_{w_k} - \eta_{\overline{W}} + g\|_\infty \leq \sigma D(\overline{W} - w_k)\Big\}.
\end{aligned}$$

Lemma 3 shows that right side in the above inequality is $o(1)$. Hence by (29–32) $R^- \leq (1+o(1))\sigma D(\overline{W})$. This inequality together with (20) and (28) proves the theorem.

5 Remark on the lower bound

In this section we will briefly discuss the minimax property of projection estimators. The next lemma provides easily verifiable conditions which ensure that projection estimators are asymptotically minimax over certain subsets $\Theta \subset \mathbf{L}_2[0,1]$.

Lemma 4 *Assume that conditions (4-6) are fulfilled. Let*

$$f_\theta(t) = \sum_{i=0}^{(1-\varepsilon)W} \theta_i \varphi_i(t),$$

where θ_i are i.i.d. $\mathcal{N}(0, \sigma^2 d_W)$. If $d_W \to \infty$ and $\mathbf{P}\{f_\theta \notin \Theta\} \to 0$ as $W \to \infty$ then

$$\lim_{W\to\infty} \inf_{\widehat{f}} \sup_{f\in\Theta} \mathbf{E}_f \|\widehat{f} - f\|_\infty / [\sigma D(W)] \geq (1 - O(\varepsilon)),$$

where inf is taken over all estimators.

We omit the proof, which is based on the Anderson lemma (see e.g. Ibragimov, Hasminskii (1981)) and the well-known arguments put forward by Pinsker (1980). Consider instead a simple example that shows how this lemma works. Assume that the Fourier coefficients of the periodic function $f(t)$ are decreasing exponentially. More precisely, we are dealing with the following functional class

$$\Theta = \{f : |{<}f, \varphi_k{>}| \leq L \exp(-\tau k)\}.$$

Suppose we want to recover $f(\cdot)$ from the noisy data (1), where $\sigma \to 0$. Using Lemma 1 it is not difficult to show that the projection estimator $\widehat{f}_{W_\sigma}(t, Y)$ in (2) with the bandwidth $W_\sigma = \tau^{-1} \log(L/\sigma)$ has the asymptotic risk (cf. Golubev, Levit and Tsybakov (1996))

$$R(\widehat{f}_{W_\sigma}, \Theta) = [1 + o(1)]\sigma \left(\frac{2}{\tau} \log \frac{L}{\sigma} \log\log \frac{L}{\sigma}\right)^{1/2}.$$

Here the estimator $\widehat{f}_{W_\sigma}(t, Y)$ is an asymptotically minimax estimator over Θ. Indeed according to Lemma 4 it is sufficient to exhibit a sequence $d_\sigma \to \infty$, such that for $\sigma \to 0$

$$\mathbf{P}\Big\{\sigma d_\sigma |\xi_i| \leq L \exp(-\tau i), \quad 0 \leq i \leq (1-\varepsilon)W_\sigma\Big\} \to 1.$$

Since $\exp(-\tau i) \geq \exp[-(1-\varepsilon)W_\sigma \tau] = L^{\varepsilon-1}\sigma^{1-\varepsilon}$ for $i \leq (1-\varepsilon)W_\sigma$ the above probability is bounded from below by

$$\begin{aligned} &\mathbf{P}\Big\{\sigma d_\sigma |\xi_i| \leq L^\varepsilon \sigma^{1-\varepsilon}, \quad 0 \leq i \leq (1-\varepsilon)W_\sigma\Big\} \qquad (33) \\ &= \Big[1 - \mathbf{P}\Big\{d_\sigma |\xi_1| > L^\varepsilon \sigma^{-\varepsilon}\Big\}\Big]^{W_\sigma(1-\varepsilon)} \geq 1 - (1-\varepsilon)W_\sigma \exp[-(L/\sigma)^{2\varepsilon}/(2d_\sigma)]. \end{aligned}$$

When d_σ is chosen equal to $\log(L/\sigma)$, the right-hand side of (33) tends to 1 as $\sigma \to 0$, proving that the projection estimator is indeed asymptotically minimax.

References

[1] Akaike, H. (1973). *Information theory and an extension of the maximum likelihood principle.* Proc. 2nd. Intern. Symp. Inf. Theory. Petrov P. N. and Csaki F. eds. Budapest, 267–281.

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

[3] Cramér, H. and Leadbetter M.R. (1967). Stationary and Related Stochastic Processes, Wiley, New York.

[4] Donoho, D. (1994). Asymptotic minimax risk for sup-norm loss: solution via optimal recovery. *Probab. Theory and Related Fields.* **99** 145–170.

[5] Dunford N. and Schwartz J.T. (1971). Linear Operators, v. III. Wiley, New York.

[6] Golubev, G., Levit, B. and Tsybakov, A. (1996). Asymptotically efficient estimation of analytic functions in Gaussian noise. *Bernoulli* **2** 167–181.

[7] Golubev, G. K. and Nussbaum, M. (1990). Adaptive spline estimates for nonparametric regression models. *Theory Probab. Appl.* **37** 521–529.

[8] Efroimovich, S.Y. and Pinsker M.S. (1984). Learning algorithm for nonparametric filtering. *Automation and Remote Control* **11** 1434–1440.

[9] Ibragimov, I.A. and Hasminskii, R.Z. (1981). *Statistical estimation. Asymptotic Theory,* Springer-Verlag.

[10] Korostelev, A. P. (1993). Asymptotically minimax regression estimator in the uniform norm up to exact constant. *Theory Probab. Appl.* **38** 737–743.

[11] Lepski, O. V. (1991). Asymptotically minimax adaptive estimation I: upper bounds. Optimal adaptive estimates. *Theory Probab. Appl.* **37** 682–697.

[12] Lepski, O. V. (1992). On problems of adaptive estimation in white Gaussian noise. In Topics in Nonparametric estimation. Advances in Soviet Math., vol. 12. Khasminskii R. Z. ed. Amer. Math. Soc., Providence, R. I. 87–106.

[13] Lepski, O. and Levit, B. (1998). Adaptive minimax estimation of infinitely differentiable functions, Math. Meth. Statistics, Vol. 7, 123-156.

[14] Lepski, O. and Levit, B. (1999). Adaptive non-parametric estimation of smooth multivariate functions, Math. Meth. Statistics, Vol. 8, 344-370

[15] Lepski, O. V. and Spokoiny, V. G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546.

[16] Lifshits, M.A. (1995). *Gaussian Random Functions.* Kluwer, Dordrecht.

[17] Mallows, C. L. (1973). Some comments on C_p. *Technometrics*, **15**, 661–675.

[18] Nemirovskii, A. (1998). *Topics in Non-parametric Estimation.* Ecole d'été de probabilités de St. Flour. Lecture notes.

[19] Pinsker, M. S. (1980). Optimal filtering of square integrable signals in Gaussian white noise. *Problems of Information Transmission*, 120–133.

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