

**ON ORACLE INEQUALITIES RELATED
TO SMOOTHING SPLINES**

Y. CAO¹ AND Y. GOLUBEV²

¹Université de Provence (Aix-Marseille 1) CMI
39 rue F. Joliot-Curie, 13453 Marseille, France
E-mail: cao@cmi.univ-mrs.fr

²CNRS, Université de Provence (Aix-Marseille 1) CMI
39 rue F. Joliot-Curie, 13453 Marseille, France, and
Inst. for Problems of Information Transmission, Moscow, Russia
E-mail: golubev.yuri@googlemail.com

Smoothing splines provide very efficient algorithms for univariate regression estimation. When an unknown regression function is estimated with the help of smoothing splines, the principal problem is related to statistical properties of data-driven methods for choosing a smoothing spline parameter. This paper focuses on the unbiased risk estimation and the generalized cross validation and for these techniques we derive the so-called oracle inequalities controlling the performance of splines with the data-driven smoothing parameter.

Key words: splines, mean square risk, oracle inequality, unbiased risk estimation, generalized cross validation.

2000 Mathematics Subject Classification: Primary 62G05, 62G20; secondary 62C20.

1. Introduction and Main Results

The classical “spline”, a wooden beam was probably invented to draw ship hulls. The earliest available mention of a spline (see Figure 1) seems to be [4]. Nowadays, computerized splines are widely used in various engineering applications. In statistics, splines are often viewed as a basic tool since they provide a very efficient smoothing technique. We refer the interested readers to [5], [7], and [17], where numerous examples of splines applications can be found. This paper focuses on some statistical problems related to the so-called data-driven methods for choosing the spline smoothing parameter. For simplicity, we will deal with the Gaussian regression assuming that we have at our disposal the noisy data

$$(1) \quad Y_i = f(X_i) + \sigma\varepsilon_i, \quad i = 1, \dots, n,$$

©2006 by Allerton Press, Inc. Authorization to photocopy individual items for internal or personal use, or the internal or personal use of specific clients, is granted by Allerton Press, Inc. for libraries and other users registered with the Copyright Clearance Center (CCC) Transactional Reporting Service, provided that the base fee of \$50.00 per copy is paid directly to CCC, 222 Rosewood Drive, Danvers, MA 01923.

FIGURE 1. A mechanical spline from 1700s

where $X_1 < X_2 < \dots < X_n$ are known design points in $[0, 1]$, and the ε_i are i.i.d. $\mathcal{N}(0, 1)$. The function $f(x)$ is assumed to be unknown but smooth and our goal is to estimate this function on the basis of Y_1, \dots, Y_n . In this paper, $f(\cdot)$ is recovered with the help of a smoothing spline defined as a solution of the following optimization problem:

$$(2) \quad \hat{f}_\alpha(x) = \arg \min_f \left\{ \frac{1}{n} \sum_{i=1}^n [Y_i - f(X_i)]^2 + \alpha \int_0^1 [f^{(m)}(u)]^2 du \right\},$$

where $f^{(m)}(\cdot)$ denotes the derivative of order m and $\alpha > 0$ is called the *smoothing spline parameter*. In typical statistical applications [7], $m = 2$, since this choice provides an excellent compromise between numerical and statistical efficiencies of the spline method. The most remarkable numerical properties of $\hat{f}_\alpha(x)$ are due to the Reinsch algorithm, which allows for computing the spline in $O(n)$ operations (see [7] for details). In order to describe statistical properties of smoothing splines, we measure the quality of $\hat{f}_\alpha(x)$ by its mean square risk

$$R(\hat{f}_\alpha, f) = \mathbf{E} \frac{1}{n} \sum_{i=1}^n [f(X_i) - \hat{f}_\alpha(X_i)]^2,$$

where \mathbf{E} stands for the expectation with respect to the measure generated by Y_1, \dots, Y_n . Obviously this risk depends on α and the main goal of this paper is to study certain data-driven methods for choosing this parameter. We restrict ourselves to the principle of unbiased risk estimation [14], which is very popular in statistics. The motivation of this method, going back to [1], [9], and [12], is rather transparent. Since the optimization problem (2) is quadratic in f , $\hat{f}_\alpha(x)$ is a linear functional of $\mathbf{Y} = (Y_1, \dots, Y_n)^\top$, and thus

$$\hat{\mathbf{f}}_\alpha = S_\alpha \mathbf{Y},$$

where $\hat{\mathbf{f}}_\alpha = (\hat{f}_\alpha(X_1), \dots, \hat{f}_\alpha(X_n))^\top$, and S_α is an $n \times n$ matrix. Therefore we obviously have

$$(3) \quad R(\hat{f}_\alpha, f) = \frac{1}{n} \mathbf{E} \|\mathbf{f} - S_\alpha \mathbf{Y}\|^2 = \frac{1}{n} \|(I - S_\alpha)\mathbf{f}\|^2 + \frac{1}{n} \sigma^2 \text{tr}(S_\alpha S_\alpha^\top).$$

The main idea of the unbiased risk estimation is to minimize the right-hand side of (3) on the basis of the observations. Replacing $\|(I - S_\alpha)\mathbf{f}\|^2$ by its unbiased estimate, which can be easily computed since

$$\mathbf{E}\|(I - S_\alpha)\mathbf{Y}\|^2 = \|(I - S_\alpha)\mathbf{f}\|^2 + \sigma^2 n - 2\sigma^2 \text{tr}(S_\alpha) + \sigma^2 \text{tr}(S_\alpha S_\alpha^\top),$$

we arrive with (3) at the following data-driven choice:

$$(4) \quad \hat{\alpha} = \arg \min_{\alpha > 0} \left\{ \|\mathbf{Y} - S_\alpha \mathbf{Y}\|^2 + 2\sigma^2 \text{tr}(S_\alpha) \right\}.$$

From the practical viewpoint, this method has a drawback because $\hat{\alpha}$ depends on σ^2 , which is hardly known in practice. Fortunately, there is a simple trick to overcome this difficulty. Let us estimate σ^2 by

$$\hat{\sigma}^2 = \frac{1}{n} \|\mathbf{Y} - S_\alpha \mathbf{Y}\|^2$$

and plug-in this estimate in (4). Thus we get

$$\hat{\alpha} \approx \arg \min_{\alpha > 0} \left\{ \|\mathbf{Y} - S_\alpha \mathbf{Y}\|^2 \left[1 + \frac{2}{n} \text{tr}(S_\alpha) \right] \right\}.$$

In practice, a slightly modified criterion is typically used. To motivate this method let us note that $\text{tr}(S_\alpha)$ can be viewed as an effective dimension of the smoothing spline. So, the spline method provides a visible data smoothing if $\text{tr}(S_\alpha) \ll n$. In this case, the set of admissible α can be reduced to $\text{tr}(S_\alpha)/n \ll 1$, and therefore by the Taylor formula

$$\left[1 + \frac{2}{n} \text{tr}(S_\alpha) \right] \approx \left[1 - \frac{1}{n} \text{tr}(S_\alpha) \right]^{-2}.$$

Thus, we get the following method

$$(5) \quad \hat{\alpha}_{GCV} = \arg \min_{\alpha > 0} \frac{\|\mathbf{Y} - S_\alpha \mathbf{Y}\|^2}{[1 - \text{tr}(S_\alpha)/n]^2},$$

which is called *generalized cross validation* (GCV). Notice here that there is a very effective numerical algorithm for computing $\hat{\alpha}_{GCV}$ (see [7] for details).

The main goal of this paper is to provide the so-called oracle inequalities related to $\hat{\alpha}$ and $\hat{\alpha}_{GCV}$. In other words, we are looking for uniform upper bounds for $R(\hat{f}_{\hat{\alpha}}, f)$ and $R(\hat{f}_{\hat{\alpha}_{GCV}}, f)$ in terms of the risk of the oracle $\inf_{\alpha} R(\hat{f}_{\alpha}, f)$. Our first result concerns the method of the unbiased risk estimation.

Theorem 1. *Uniformly in f and $\gamma \in (0, 1)$,*

$$(6) \quad R(\hat{f}_{\hat{\alpha}}, f) \leq \frac{1}{1 - \gamma} \left\{ \inf_{\alpha} R(\hat{f}_{\alpha}, f) + \frac{C\sigma^2}{n\gamma} \right\}.$$

Here and in the rest of the paper, C denotes a generic constant.

The case of the unknown noise variance is more complicated and we could not prove a similar fact for the original GCV without restrictions on possible values of α . Therefore we consider the following truncated GCV criterion:

$$(7) \quad \hat{\alpha}_G = \arg \min_{\alpha: \text{tr}(S_\alpha) \leq \sqrt{n}} \frac{\|\mathbf{Y} - S_\alpha \mathbf{Y}\|^2}{[1 - \text{tr}(S_\alpha)/n]^2}.$$

Theorem 2. *Uniformly in f and $\gamma \in (0, 1)$,*

$$(8) \quad R(\hat{f}_{\hat{\alpha}_G}, f) \leq \frac{1}{1 - \gamma} \left\{ \left(1 - \frac{1}{\sqrt{n}}\right)^{-2} \inf_{\alpha: \text{tr}(S_\alpha) \leq \sqrt{n}} R(\hat{f}_\alpha, f) + \frac{C\sigma^2}{n\gamma} \right\}.$$

Unfortunately, the precise evaluation of the constant C in Theorems 1 and 3 is a very delicate problem. Our estimate of C by the Monte-Carlo method gives $C \leq 3$.

Remark 1. It seems at the first glance that the difference between $\hat{\alpha}_{GCV}$ and $\hat{\alpha}_G$ may result in different oracle inequalities. However, if the underlying regression function is sufficiently smooth, then

$$\min_{\alpha: \text{tr}(S_\alpha) \leq \sqrt{n}} R(\hat{f}_\alpha, f) = \min_{\alpha} R(\hat{f}_\alpha, f).$$

A heuristic explanation of this equality is based on the fact that $\text{tr}(S_\alpha)$ is related to the effective dimension of the linear space used by the spline method and therefore, $\text{tr}(S_\alpha) = O(\text{tr}(S_\alpha S_\alpha^\top))$. Obviously, smooth functions can be well approximated by linear spaces with low dimensions thus resulting in $\min_{\alpha} R(\hat{f}_\alpha, f) \leq O(\sigma^2 n^{-1/2})$. This means, in view of (3), that

$$\arg \min_{\alpha} R(\hat{f}_\alpha, f) \in \{\alpha: \text{tr}(S_\alpha) \leq \sqrt{n}\}.$$

Notice also that the rate of convergence $O(\sigma^2 n^{-1/2})$ means that the underlying regression function has Sobolev's smoothness $1/2$.

Remark 2. In fact, the condition $\{\alpha: \text{tr}(S_\alpha) \leq \sqrt{n}\}$ can be relaxed to $\{\alpha: \text{tr}(S_\alpha) \leq n/5\}$. However, in the latter case, the proof of Theorem 2 becomes cumbersome and therefore we preferred, in this paper, to present a less general but more transparent proof.

Standard mathematical applications of Theorems 1 and 2 are related to asymptotically minimax ($n \rightarrow \infty$) convergence rates on Sobolev balls. Let $\mathbb{W}_2^m(L)$ be the Sobolev ball

$$\mathbb{W}_2^m(L) = \left\{ f: \int_0^1 [f^{(m)}(u)]^2 du \leq L \right\}$$

and suppose that the performance of an estimate $\hat{f}(x)$ is measured by its minimax risk

$$r_n(\hat{f}) = \sup_{f \in \mathbb{W}_2^m(L)} \frac{1}{n} \mathbf{E} \sum_{i=1}^n [f(X_i) - \hat{f}(X_i)]^2.$$

Then, under mild assumptions on the design \mathbf{X} , combining Theorems 1 and 2 with the lower bound from [11], one can show that as $n \rightarrow \infty$

$$(9) \quad \inf_{\hat{f}} r_n(\hat{f}) = (1 + o(1))\kappa(m)r_n(\hat{f}_{\hat{\alpha}_G}) = (1 + o(1))\kappa(m)r_n(\hat{f}_{\hat{\alpha}}),$$

where \inf is taken over all estimators and the constant $\kappa(m) < 1$ represents the efficiency of the standard smoothing spline (2) with respect to the asymptotically best estimator. For mathematical details, we refer the reader to [13] and [10]. Equation (9) yields, in particular,

$$\inf_{\hat{f}} r_n(\hat{f}) \asymp r_n(\hat{f}_{\hat{\alpha}_G}) \asymp r_n(\hat{f}_{\hat{\alpha}}) \asymp L^{1/(2m+1)} \left(\frac{\sigma^2}{n} \right)^{2m/(2m+1)}.$$

Notice also that $\kappa(m)$ is close to 1, $\kappa(m) \geq 0.91$ for all $m \geq 2$ (see, e.g., Fig. 1 in [6]).

From the probabilistic viewpoint, the present paper can be viewed as a natural generalization of [2], where we extensively exploited martingale properties of the Wiener process $W(t)$. In particular, the following well-known fact

$$\mathbf{E} \max_{t \geq 0} [W(t) - \mu t]_+^p \leq \frac{C(p)}{\mu^p}, \quad p, \mu > 0,$$

was at the very core of our approach. For smoothing splines, we need similar facts but for essentially more general random processes. Therefore following [8], we introduce a simple notion of zero mean *ordered process* $\xi(t)$ characterized by the property $\mathbf{E}\xi(t_1)\xi(t_2) \geq \min\{\mathbf{E}\xi^2(t_1), \mathbf{E}\xi^2(t_2)\}$, and show with the standard chaining argument (see, e.g., [16]) that

$$\mathbf{E} \max_{t \geq 0} [\xi(t) - \mu \mathbf{E}\xi^2(t)]_+^p \leq \frac{C(p)}{\mu^p}.$$

We will see that this inequality almost immediately results in Theorems 1 and 2.

2. Proofs

2.1. PRELIMINARIES. We begin with a standard linear transformation of the observations, which enables us to simplify numerous technical details. Recall that the design points X_i , $i = 1, \dots, n$, are assumed to be different and belonging to $[0, 1]$. Let $\{\varphi_k(x)$, $k = 1, \dots, n\}$ be the Reinsch–Demmler [3] basis defined by

$$(10) \quad \frac{1}{n} \sum_{i=1}^n \varphi_k(X_i) \varphi_j(X_i) = \delta_{kj}, \quad \int_0^1 \varphi_i^{(m)}(x) \varphi_j^{(m)}(x) dx = \lambda_i \delta_{ij},$$

$$0 = \dots = 0 < \lambda_m < \dots < \lambda_n,$$

where δ_{ij} is Kronecker's delta. The eigenvalues λ_k , $k = 1, \dots, n$, are related to Kolmogorov's diameter (see [15] for details) of the Sobolev ball

$$\mathbb{W}_2^m = \left\{ f: \int_0^1 [f^{(m)}(x)]^2 dx \leq 1 \right\}.$$

To explain this fact, suppose we want to approximate functions in \mathbb{W}_2^m with the help of linear methods. It means that for given functions ϕ_i , $i = 1, \dots, N$, we are interested in the maximal approximation error

$$D^N(\phi) = \sup_{f \in \mathbb{W}_2^m} \inf_{c_1, \dots, c_N} \frac{1}{n} \sum_{i=1}^n \left[f(X_i) - \sum_{k=1}^N c_k \phi_k(X_i) \right]^2.$$

Then Kolmogorov's diameter d^N of \mathbb{W}_2^m is defined as the minimal approximation error

$$d^N = \inf_{\phi} D^N(\phi),$$

where \inf is computed over all functions ϕ_k , $k = 1, \dots, N$. With a relatively simple algebra one can show that the Reinsch–Demmler basis provides the Kolmogorov diameter of \mathbb{W}_2^m , i.e.,

$$d^N = D^N(\varphi) = \lambda_{N+1}^{-1}.$$

Using this basis, we make the following linear transformation of the observations:

$$(11) \quad Z_k = \frac{1}{n} \sum_{i=1}^n Y_i \varphi_k(X_i) = \theta_k + \varepsilon \xi_k,$$

where

$$\theta_k = \frac{1}{n} \sum_{i=1}^n f(X_i) \varphi_k(X_i), \quad \xi_k = \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i \varphi_k(X_i), \quad \varepsilon = \frac{\sigma}{\sqrt{n}}.$$

Obviously, the sequence space model (11) is equivalent to (1) and therefore from now on we deal with (11). Notice that the risk of an estimator \hat{f} can be rewritten in terms of (11) as

$$R(\hat{f}, f) = \mathbf{E} \sum_{k=1}^n [\hat{\theta}_k - \theta_k]^2, \quad \text{where} \quad \hat{\theta}_k = \frac{1}{n} \sum_{i=1}^n \hat{f}(X_i) \varphi_k(X_i).$$

It is also easy to check, in view of (10), that the smoothing spline admits the following spectral representation:

$$\hat{f}_\alpha(x) = \sum_{k=1}^n h_k(\alpha) Z_k \varphi_k(x) \quad \text{with} \quad h_k(\alpha) = \frac{1}{1 + \alpha \lambda_k}$$

and its risk is computed as follows:

$$R(\hat{f}_\alpha, f) = R(\alpha, \theta) = \sum_{k=1}^n [1 - h_k(\alpha)]^2 \theta_k^2 + \varepsilon^2 \sum_{k=1}^n h_k^2(\alpha).$$

Finally notice that $\hat{\alpha}$ and $\hat{\alpha}_G$ can be expressed in terms of Z_k , $k = 1, \dots, n$, as

$$(12) \quad \hat{\alpha} = \arg \min_{\alpha \geq 0} \left\{ \sum_{k=1}^n [1 - h_k(\alpha)]^2 Z_k^2 + 2\varepsilon^2 \|h(\alpha)\|_1 \right\},$$

$$(13) \quad \hat{\alpha}_G = \arg \min_{\alpha: \|h(\alpha)\| \leq \sqrt{n}} \sum_{k=1}^n \frac{[1 - h_k(\alpha)]^2 Z_k^2}{[1 - \|h(\alpha)\|_1/n]^2},$$

where $\|h(\alpha)\|_1 = \sum_{k=1}^n h_k(\alpha)$.

2.2. ORDERED PROCESSES AND THEIR PROPERTIES. Our method of deriving oracle inequalities is related to a special class of random processes. Let $\xi(t)$, $t \geq 0$, be a random process with $\mathbf{E}\xi(t) = 0$ and a finite variance $\mathbf{E}\xi^2(t) = \sigma^2(t)$, which is assumed to be continuous and monotone

$$\sigma^2(t_2) \geq \sigma^2(t_1), \quad t_2 \geq t_1.$$

The process $\xi(t)$, $t \geq 0$, is called *ordered* if it is separable and for all $t_2 \geq t_1$

$$(14) \quad \mathbf{E}[\xi(t_2) - \xi(t_1)]^2 \leq \sigma^2(t_2) - \sigma^2(t_1).$$

Obviously this inequality admits the following equivalent form

$$\mathbf{E}\xi(t_2)\xi(t_1) \geq \min \{ \mathbf{E}\xi^2(t_2), \mathbf{E}\xi^2(t_1) \}.$$

Thus, an ordered process can be viewed as a natural generalization of the Wiener process $W(t)$ for which $\mathbf{E}W(t_1)W(t_2) = \min\{\mathbf{E}W(t_1), \mathbf{E}W(t_2)\}$. We will see that the class of ordered processes is sufficiently broad.

Denote for brevity

$$\Delta_\xi(t_1, t_2) = \frac{\xi(t_1) - \xi(t_2)}{\sqrt{\mathbf{E}[\xi(t_1) - \xi(t_2)]^2}}.$$

The main property of ordered processes is given by the following lemma.

Lemma 1. *Suppose that there exists $\lambda > 0$ such that*

$$(15) \quad \varphi(\lambda) \triangleq \sup_{t_1, t_2} \mathbf{E} \cosh \{ \lambda \Delta_\xi(t_1, t_2) \} < \infty.$$

Then there exists a constant C depending on λ such that for all $T > 0$ and all $p \geq 1$

$$(16) \quad \left[\mathbf{E} \sup_{t, s \in [0, T]} |\xi(t) - \xi(s)|^p \right]^{1/p} \leq Cp\sigma(T).$$

Proof. Notice that for $p \geq 1$ the function $L(x) = \log^p(x + e^{p-1})$ is concave on $(0, \infty)$ since

$$L''(x) = \frac{p \log^{p-2}(x + e^{p-1})}{(x + e^{p-1})^2} [p - 1 - \log(x + e^{p-1})] \leq 0.$$

In order to prove (16), we use the standard chaining argument [16]. For a given integer $s \geq 0$ define the points t_k^s on $[0, T]$ by

$$\sigma^2(t_k^s) = 2^{-s} k \sigma^2(T), \quad k = 0, \dots, 2^s - 1,$$

and denote by \mathcal{T}^s the set of these points. Let u be an arbitrary point in \mathcal{T}^s . Then we can find a chain, i.e., the points $\tau_k(u) \in \mathcal{T}^k$, $k = 0, \dots, s$, such that

- $u = \tau_s(u)$, $0 = \tau_0(u)$,
- $|\sigma^2(\tau_k(u)) - \sigma^2(\tau_{k-1}(u))| \leq 2^{-k+1} \sigma^2(T)$.

To verify that such points exist, one can imagine the standard binary tree. The nodes of this tree at the level k are associated with the points t_j^k , $j = 0, \dots, 2^k - 1$. It is clear that there exists a unique way connecting $u \in \mathcal{T}^s$ and 0 (top of the tree). This way passes via nodes, which are denoted by $\tau_k(u)$. So we can write

$$u = \sum_{k=0}^{s-1} [\tau_{k+1}(u) - \tau_k(u)]$$

and for two arbitrary points u, v in \mathcal{T}^s we get

$$u - v = \sum_{k=0}^{s-1} [\tau_{k+1}(u) - \tau_k(u)] - \sum_{k=0}^{s-1} [\tau_{k+1}(v) - \tau_k(v)].$$

Therefore by (15) we have

$$\begin{aligned} (17) \quad & \left[\mathbf{E} \sup_{u, v \in \mathcal{T}^s} |\xi(u) - \xi(v)|^p \right]^{1/p} \leq 2 \sum_{k=0}^{s-1} \left[\mathbf{E} \sup_{u \in \mathcal{T}^{k+1}} |\xi(\tau_{k+1}(u)) - \xi(\tau_k(u))|^p \right]^{1/p} \\ & = 2 \sum_{k=0}^{s-1} \left[\mathbf{E} \sup_{u \in \mathcal{T}^{k+1}} |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|^p \left[\mathbf{E} [\xi(\tau_{k+1}(u)) - \xi(\tau_k(u))]^2 \right]^{p/2} \right]^{1/p} \\ & \leq 2\sigma(T) \sum_{k=0}^{s-1} 2^{-k} \left[\mathbf{E} \sup_{u \in \mathcal{T}^{k+1}} |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|^p \right]^{1/p}. \end{aligned}$$

Next, using concavity of $L(x)$, $x \geq 0$, and (15), we obtain

$$\begin{aligned} & \left[\mathbf{E} \sup_{u \in \mathcal{T}^{k+1}} |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|^p \right]^{1/p} \\ & \leq \frac{1}{\lambda} \log \left[\sum_{u \in \mathcal{T}^{k+1}} \mathbf{E} \exp[\lambda |\Delta_\xi(\tau_{k+1}(u), \tau_k(u))|] + e^{p-1} \right] \\ & \leq \frac{1}{\lambda} \log [2^{k+2} \varphi(\lambda) + e^{p-1}] = \frac{(k+2) \log(2) + p - 1}{\lambda} + \frac{\log[\varphi(\lambda)]}{\lambda}. \end{aligned}$$

Substituting this in (17), we arrive at the inequality

$$\left[\mathbf{E} \sup_{u, v \in \mathcal{T}^s} |\xi(u) - \xi(v)|^p \right]^{1/p} \leq C\sigma(T),$$

which proves the lemma, since by separability of $\xi(t)$

$$\left[\mathbf{E} \sup_{u,v \in [0,T]} |\xi(u) - \xi(v)|^p \right]^{1/p} = \limsup_{s \rightarrow \infty} \left[\mathbf{E} \sup_{u,v \in T^s} |\xi(u) - \xi(v)|^p \right]^{1/p}. \quad \square$$

Lemma 1 almost immediately results in the following fact.

Lemma 2. *Let $\xi(t)$ be an ordered process satisfying (15) and such that $\xi(0) = 0$. Then there exists a constant C depending on λ such that for all $\alpha > 0$*

$$(18) \quad \mathbf{E} \sup_{t \geq 0} [\xi(t) - \alpha \sigma^2(t)]_+^p \leq \frac{C(4p+4)^{2p+2}}{\alpha^p},$$

where $[x]_+ = \max(0, x)$.

Proof. Without loss of generality, we can assume that $\lim_{t \rightarrow \infty} \sigma^2(t) = \infty$. Then for any integer $k \geq 0$ we define $t_k(\alpha)$ by

$$\sigma(t_k(\alpha)) = \frac{k}{\alpha}.$$

In what follows, we will use the following form of the Markov inequality

$$(19) \quad \mathbf{E} \eta^p \mathbf{1}\{\eta > x\} \leq \frac{\mathbf{E} |\eta|^{p+q}}{x^q},$$

which follows immediately from the inequality $\eta^p \mathbf{1}\{\eta > x\} \leq |\eta|^p |\eta/x|^q$.

Using that $f(x) = x^p \mathbf{1}\{x > x_0\}$ is monotone in $x > 0$, we have

$$(20) \quad \begin{aligned} \mathbf{E} \sup_{t \geq 0} [\xi(t) - \alpha \sigma^2(t)]_+^p &\leq \sum_{k=0}^{\infty} \mathbf{E} \sup_{t \in [t_k(\alpha), t_{k+1}(\alpha)]} \xi^p(t) \mathbf{1}\{\xi(t) \geq \alpha \sigma^2(t)\} \\ &\leq \sum_{k=0}^{\infty} \mathbf{E} \sup_{t \in [t_k(\alpha), t_{k+1}(\alpha)]} \xi^p(t) \mathbf{1}\{\xi(t) \geq \alpha \sigma^2(t_k(\alpha))\} \\ &\leq \sum_{k=0}^{\infty} \mathbf{E} \sup_{t \in [t_k(\alpha), t_{k+1}(\alpha)]} \xi^p(t) \mathbf{1}\left\{ \sup_{t \in [t_k(\alpha), t_{k+1}(\alpha)]} \xi(t) \geq \alpha \sigma^2(t_k(\alpha)) \right\} \\ &\leq \mathbf{E} \sup_{0 \leq t \leq t_1(\alpha)} |\xi(t)|^p \\ &\quad + \sum_{k=1}^{\infty} \mathbf{E} \sup_{0 \leq t \leq t_{k+1}(\alpha)} \xi^p(t) \mathbf{1}\left\{ \sup_{0 \leq t \leq t_{k+1}(\alpha)} \xi(t) \geq \alpha \sigma^2(t_k(\alpha)) \right\}. \end{aligned}$$

By Lemma 1, the first term at the right-hand side of the above inequality is bounded by

$$(21) \quad \mathbf{E} \sup_{0 \leq t \leq t_1(\alpha)} |\xi(t)|^p \leq Cp^p \sigma^p(t_1(\alpha)) = \frac{Cp^p}{\alpha^p},$$

whereas for the second one, in view of (19), we get the following upper bound:

$$\begin{aligned} & \sum_{k=1}^{\infty} \mathbf{E} \sup_{0 \leq t \leq t_{k+1}(\alpha)} \xi^p(t) \mathbf{1} \left\{ \sup_{0 \leq t \leq t_{k+1}(\alpha)} \xi(t) \geq \alpha \sigma^2(t_k(\alpha)) \right\} \\ & \leq C(p+q)^{p+q} \sum_{k=1}^{\infty} \frac{\sigma^{p+q}(t_{k+1}(\alpha))}{[\alpha \sigma^2(t_k(\alpha))]^q} = \frac{C[2(p+q)]^{p+q}}{\alpha^p} \sum_{k=1}^{\infty} \frac{1}{k^{q-p}}. \end{aligned}$$

Setting $q = p + 2$ in the above inequality and using (20) together with (21), we prove (18). \square

2.3. EXAMPLES OF ORDERED PROCESSES. The simplest example of an ordered process is $\xi(t) = \xi t$, where ξ is a zero mean random variable with a finite exponential moment $\mathbf{E} \exp(\lambda|\xi|) < \infty$ for some $\lambda > 0$. As we have already mentioned, the Wiener process $W(t)$ is an ordered process. At the first glance, ξt and $W(t)$ are quite different, but from the viewpoint of Lemma 2 they are equivalent. Of course, the distribution of $\max_{t \geq 0} [W(t) - \alpha t]$ is well known

$$\mathbf{P} \left\{ \max_{t \geq 0} [W(t) - \alpha t] \geq x \right\} = \exp(-2\alpha x).$$

The following two examples play an essential role in adaptive estimation. Let $H_k(\mu) \in [0, 1]$ be a nonincreasing sequence in k and such that

$$(22) \quad \begin{aligned} H_k(\mu_2) &\geq H_k(\mu_1) \quad \text{for all } \mu_2 \geq \mu_1 \quad \text{and } k = 1, 2, \dots, \\ \lim_{\mu \rightarrow \infty} H_k(\mu) &= 1, \quad k = 1, 2, \dots \end{aligned}$$

Consider the following Gaussian process:

$$\xi_1(t) = \sum_{k=1}^{\infty} [1 - H_k(1/t)] \theta_k \xi_k,$$

where ξ_k are i.i.d. $\mathcal{N}(0, 1)$ and $\sum_{i=1}^{\infty} \theta_i^2 < \infty$. It is easy to see that $\xi_1(t)$ is an ordered process. Indeed, in view of (22) we have for $t_1 \leq t_2$

$$\mathbf{E} \xi_1^2(t_1) = \sum_{k=1}^{\infty} [1 - H_k(1/t_1)]^2 \theta_k^2 \leq \sum_{k=1}^{\infty} [1 - H_k(1/t_1)] [1 - H_k(1/t_2)] \theta_k^2 = \mathbf{E} \xi_1(t_1) \xi_1(t_2),$$

thus proving that $\xi_1(t)$ is an ordered process. For $\xi_1(t)$ Lemma 2 sounds as follows.

Lemma 3. *Let $H_k(\cdot)$ satisfies (22), then for all $\alpha > 0$*

$$(23) \quad \mathbf{E} \sup_{\mu \geq 0} \left[\sum_{k=1}^{\infty} [1 - H_k(\mu)] \theta_k \xi_k - \alpha \sum_{k=1}^{\infty} [1 - H_k(\mu)]^2 \theta_k^2 \right]_+^p \leq \frac{C(p)}{\alpha^p},$$

and

$$(24) \quad \mathbf{E} \sup_{\mu \geq 0} \left[\sum_{k=1}^{\infty} [1 - H_k(\mu)]^2 \theta_k \xi_k - \alpha \sum_{k=1}^{\infty} [1 - H_k(\mu)]^2 \theta_k^2 \right]_+^p \leq \frac{C(p)}{\alpha^p}.$$

In order to prove (24), it suffices to note that if $H_k(\cdot)$ satisfies (22), then the function $2H_k(\cdot) - H_k^2(\cdot)$ also satisfies this condition.

The next useful ordered process is defined by

$$\xi_2(t) = \sum_{k=1}^{\infty} H_k(t)(\xi_k^2 - 1),$$

where the ξ_k are i.i.d. $\mathcal{N}(0, 1)$ and $H_k(t)$ satisfies (22). It is easy to check that

$$\mathbf{E}\xi_2^2(t_1) \leq \mathbf{E}\xi_2(t_2)\xi_2(t_1), \quad t_1 \leq t_2.$$

So, in order to apply Lemma 2, it remains to check (15). Denoting for brevity

$$\|H(t_2) - H(t_1)\|^2 = \sum_{k=1}^{\infty} [H_k(t_2) - H_k(t_1)]^2,$$

we have

$$(25) \quad \mathbf{E} \exp[\lambda \Delta_{\xi}(t_2, t_1)] = \exp \left[-\frac{\lambda}{\sqrt{2}\|H(t_2) - H(t_1)\|} \sum_{k=1}^{\infty} [H_k(t_2) - H_k(t_1)] - \frac{1}{2} \sum_{k=1}^{\infty} \log \left(1 - \sqrt{2}\lambda \frac{H_k(t_2) - H_k(t_1)}{\|H(t_2) - H(t_1)\|} \right) \right].$$

Since obviously

$$\max_k [H_k(t_2) - H_k(t_1)] \leq \|H(t_2) - H(t_1)\|,$$

then using the Taylor expansion for $\log(1 - \cdot)$ in the right-hand side of (25), we get for $\lambda \leq 1/2$

$$\mathbf{E} \exp[\lambda \Delta_{\xi}(t_2, t_1)] \leq \exp(C\lambda^2),$$

thus proving (15). Therefore using Lemma 2, we obtain the following fact.

Lemma 4. *Let $H_k(\cdot)$ satisfies (22), then for all $\alpha > 0$*

$$(26) \quad \mathbf{E} \max_{\mu > 0} \left[\sum_{k=1}^{\infty} H_k(\mu)(\xi_k^2 - 1) - \alpha \sum_{k=1}^{\infty} H_k^2(\mu) \right]_+^p \leq \frac{C(p)}{\alpha^p},$$

and

$$(27) \quad \mathbf{E} \max_{\mu > 0} \left[\sum_{k=1}^{\infty} H_k(\mu)(1 - \xi_k^2) - \alpha \sum_{k=1}^{\infty} H_k^2(\mu) \right]_+^p \leq \frac{C(p)}{\alpha^p}.$$

The proof of (27) is quite similar to that of (26) and therefore omitted.

2.4. PROOF OF THEOREM 1. It follows the main lines of the proof of Theorem 1 in [2]. Let $\tilde{\alpha}$ be an arbitrary data-driven smoothing parameter. Then we can decompose the risk $\|h(\tilde{\alpha})Z - \theta\|^2$ as follows:

$$\begin{aligned}
(28) \quad \|h(\tilde{\alpha})Z - \theta\|^2 &= \sum_{k=1}^n [h_k(\tilde{\alpha})Z_k - Z_k + \varepsilon\xi_k]^2 \\
&= \sum_{k=1}^n [h_k(\tilde{\alpha}) - 1]^2 Z_k^2 + 2\varepsilon \sum_{k=1}^n [h_k(\tilde{\alpha}) - 1] Z_k \xi_k + \varepsilon^2 \sum_{k=1}^n \xi_k^2 \\
&= \left\{ \sum_{k=1}^n [1 - h_k(\tilde{\alpha})]^2 Z_k^2 + 2\varepsilon^2 \sum_{k=1}^n h_k(\tilde{\alpha}) \right\} - \varepsilon^2 \sum_{k=1}^n \xi_k^2 \\
&\quad + 2\varepsilon \sum_{k=1}^n [h_k(\tilde{\alpha}) - 1] \theta_k \xi_k + 2\varepsilon^2 \sum_{k=1}^n h_k(\tilde{\alpha}) (\xi_k^2 - 1).
\end{aligned}$$

The principal idea is to show that the last two terms of this equation are small. To implement this idea, notice that for any $\gamma \in (0, 1)$ the estimation error $\|h(\tilde{\alpha})Z - \theta\|^2$ admits the following representation:

$$\begin{aligned}
(29) \quad \|h(\tilde{\alpha})Z - \theta\|^2 &= (1 - \gamma) \|h(\tilde{\alpha})Z - \theta\|^2 + \gamma \|h(\tilde{\alpha})Z - \theta\|^2 \\
&= (1 - \gamma) \|h(\tilde{\alpha})Z - \theta\|^2 + \gamma \sum_{k=1}^n [1 - h_k(\tilde{\alpha})]^2 \theta_k^2 \\
&\quad + \gamma \varepsilon^2 \sum_{k=1}^n [h_k(\tilde{\alpha})]^2 \xi_k^2 - 2\gamma \varepsilon \sum_{k=1}^n h_k(\tilde{\alpha}) [1 - h_k(\tilde{\alpha})] \theta_k \xi_k \\
&= (1 - \gamma) \|h(\tilde{\alpha})Z - \theta\|^2 + \gamma \varepsilon^2 \sum_{k=1}^n [h_k(\tilde{\alpha})]^2 + \gamma \sum_{k=1}^n [1 - h_k(\tilde{\alpha})]^2 \theta_k^2 \\
&\quad + \gamma \varepsilon^2 \sum_{k=1}^n [h_k(\tilde{\alpha})]^2 (\xi_k^2 - 1) - 2\gamma \varepsilon \sum_{k=1}^n h_k(\tilde{\alpha}) [1 - h_k(\tilde{\alpha})] \theta_k \xi_k.
\end{aligned}$$

Our next step is to use the definition of $\hat{\alpha}$ (see (12)). We obviously have that for any given α

$$\begin{aligned}
&\left\{ \sum_{k=1}^n [1 - h_k(\hat{\alpha})]^2 Z_k^2 + 2\varepsilon^2 \sum_{k=1}^n h_k(\hat{\alpha}) \right\} \leq \left\{ \sum_{k=1}^n [1 - h_k(\alpha)]^2 Z_k^2 + 2\varepsilon^2 \sum_{k=1}^n h_k(\alpha) \right\} \\
&\leq R(\alpha, \theta) + \varepsilon^2 \sum_{k=1}^n \xi_k^2 + 2\varepsilon \sum_{k=1}^n [1 - h_k(\alpha)]^2 \theta_k \xi_k + \varepsilon^2 \sum_{k=1}^n [1 - h_k(\alpha)]^2 (\xi_k^2 - 1).
\end{aligned}$$

Therefore, combining this inequality with (28) and (29), we obtain

$$(30) \quad (1 - \gamma) \mathbf{E} \|h(\hat{\alpha})Z - \theta\|^2 \leq R(\alpha, \theta) + \mathbf{E} r_0[h(\alpha), \theta] + \mathbf{E} r_1[h(\hat{\alpha}), \theta] + \mathbf{E} r_2[h(\hat{\alpha})],$$

where

$$\begin{aligned} r_0[h, \theta] &= 2\varepsilon \sum_{k=1}^n (1 - h_k)^2 \theta_k \xi_k + \varepsilon^2 \sum_{k=1}^n (1 - h_k)^2 (\xi_k^2 - 1), \\ r_1[h, \theta] &= -2\varepsilon \sum_{k=1}^n \{1 - (1 + \gamma)h_k - \gamma h_k^2\} \theta_k \xi_k - \gamma \sum_{k=1}^n (1 - h_k)^2 \theta_k^2, \\ r_2[h] &= \varepsilon^2 \sum_{k=1}^n \{2h_k - \gamma [h_k]^2\} (\xi_k^2 - 1) - \gamma \varepsilon^2 \sum_{k=1}^n [h_k]^2. \end{aligned}$$

To bound from above these remainder terms, we apply Lemmas 3 and 4. It is easy to see that the sequence $(1 + \gamma)h_k(\alpha) - \gamma h_k^2(\alpha)$ with $\alpha = 1/\mu$ satisfies (22) for any $\gamma \in [0, 1]$. Therefore with (23) we obtain

$$\begin{aligned} (31) \quad \mathbf{E} r_1[h(\hat{\alpha}), \theta] &\leq \mathbf{E} \sup_{\alpha \geq 0} \left[-2\varepsilon \sum_{k=1}^n [1 - (1 + \gamma)h_k(\alpha) + \gamma h_k^2(\alpha)] \theta_k \xi_k \right. \\ &\quad \left. - \gamma \sum_{k=1}^n [1 - (1 + \gamma)h_k(\alpha) + \gamma h_k^2(\alpha)]^2 \theta_k^2 \right] \\ &\quad + \mathbf{E} \sup_{\alpha \geq 0} \left[\gamma \sum_{k=1}^n [1 - (1 + \gamma)h_k(\alpha) + \gamma h_k^2(\alpha)]^2 \theta_k^2 \right. \\ &\quad \left. - \gamma \sum_{k=1}^n [1 - h_k(\alpha)]^2 \theta_k^2 \right]_+ \leq \frac{C\varepsilon^2}{\gamma}. \end{aligned}$$

In order to control $\mathbf{E} r_2(h(\hat{\alpha}))$, notice that $2x - \gamma x^2$ is a monotone function on $[0, 1]$ for any $\gamma \in [0, 1]$. Hence the sequence $[2h_k(\alpha) - \gamma h_k^2(\alpha)]/(2 - \gamma)$ with $\alpha = 1/\mu$ satisfies (22), and by Lemma 4 we obtain

$$\begin{aligned} (32) \quad (2 - \gamma) \mathbf{E} \sum_{k=1}^n \frac{2h_k(\hat{\alpha}) - \gamma [h_k(\hat{\alpha})]^2}{2 - \gamma} (\xi_k^2 - 1) &- \gamma \mathbf{E} \sum_{k=1}^n \left[\frac{2h_k(\hat{\alpha}) - \gamma [h_k(\hat{\alpha})]^2}{2 - \gamma} \right]^2 \\ &+ \gamma \mathbf{E} \sum_{k=1}^n \left[\frac{2h_k(\hat{\alpha}) - \gamma [h_k(\hat{\alpha})]^2}{2 - \gamma} \right]^2 - \gamma \mathbf{E} \sum_{k=1}^n [h_k(\hat{\alpha})]^2 \\ &\leq (2 - \gamma) \mathbf{E} \sup_{\alpha \geq 0} \left\{ \sum_{k=1}^n \frac{2h_k(\alpha) - \gamma [h_k(\alpha)]^2}{2 - \gamma} (\xi_k^2 - 1) \right. \\ &\quad \left. - \frac{\gamma}{2 - \gamma} \sum_{k=1}^n \left[\frac{2h_k(\alpha) - \gamma [h_k(\alpha)]^2}{2 - \gamma} \right]^2 \right\} \leq \frac{C(2 - \gamma)^2}{\gamma}, \end{aligned}$$

thus proving that

$$\mathbf{E} r_2[h(\hat{\alpha})] \leq \frac{C\varepsilon^2}{\gamma}.$$

To finish the proof, it remains to substitute this inequality and (31), (32) in (30) and to notice that obviously $\mathbf{E} r_0[h(\alpha), \theta] = 0$. \square

2.5. PROOF OF THEOREM 2. The main idea of the proof is related to the almost obvious fact that $\|h(\hat{\alpha}_G)Z - Z\|^2/n$ always provides a good *upper bound* for the unknown noise variance.

Lemma 5. *Let $\tilde{\alpha}$ be an arbitrary data-driven smoothing parameter. Then*

$$(33) \quad \mathbf{E} \left[\varepsilon^2 - \frac{\|h(\tilde{\alpha})Z - Z\|^2}{n} \right]_+ \leq \frac{\sqrt{2}\varepsilon^2}{\sqrt{n}} + \frac{2\varepsilon^2}{n} \mathbf{E} \sum_{k=1}^n h_k(\tilde{\alpha}) + \frac{C\varepsilon^2}{n}.$$

Proof. We start with the following decomposition:

$$(34) \quad \begin{aligned} \varepsilon^2 - \frac{\|h(\tilde{\alpha})Z - Z\|^2}{n} &= \frac{\varepsilon^2}{n} \sum_{k=1}^n (1 - \xi_k^2) \\ &+ \frac{1}{n} \left[-2\varepsilon \sum_{k=1}^n [1 - h_k(\tilde{\alpha})]^2 \theta_k \xi_k - \sum_{k=1}^n [1 - h_k(\tilde{\alpha})]^2 \theta_k^2 \right] \\ &+ \frac{\varepsilon^2}{n} \left[\sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})](1 - \xi_k^2) - \sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})] \right]. \end{aligned}$$

The first term in the right-hand side can be easily bounded since

$$(35) \quad \mathbf{E} \left[\frac{1}{n} \sum_{k=1}^n (\xi_k^2 - 1) \right]^2 = \frac{2}{n}.$$

To control the last two remainder terms, we apply Lemmas 3 and 4. By (24) we immediately get

$$(36) \quad \mathbf{E}_\theta \left[-2\varepsilon \sum_{k=1}^n [1 - h_k(\tilde{\alpha})]^2 \theta_k \xi_k - \sum_{k=1}^n [1 - h_k(\tilde{\alpha})]^2 \theta_k^2 \right]_+ \leq C\varepsilon^2.$$

Our final step is to control the last term, which can be rewritten as

$$\begin{aligned} &\mathbf{E} \left[\sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})](1 - \xi_k^2) - \sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})] \right] \\ &= \mathbf{E} \left[\sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})](1 - \xi_k^2) - \frac{1}{4} \sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})]^2 \right] \\ &\quad + \left[\frac{1}{4} \sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})]^2 - \sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})] \right]. \end{aligned}$$

Since $2h_k(\alpha) - h_k^2(\alpha)$ with $\alpha = 1/\mu$ satisfies (22), we get with the help of Lemma 4

$$\mathbf{E} \sup_{\alpha \geq 0} \left[\sum_{k=1}^n [h_k^2(\alpha) - 2h_k(\alpha)](1 - \xi_k^2) - \frac{1}{4} \sum_{k=1}^n [h_k^2(\alpha) - 2h_k(\alpha)]^2 \right]_+ \leq C$$

and therefore

$$\begin{aligned}
& \mathbf{E} \left[\sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})](1 - \xi_k^2) - \sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})] \right]_+ \\
& \leq \mathbf{E} \left[\sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})](1 - \xi_k^2) - \frac{1}{4} \sum_{k=1}^n [h_k^2(\tilde{\alpha}) - 2h_k(\tilde{\alpha})]^2 \right]_+ \\
& \quad + \mathbf{E} \left[\sum_{k=1}^n \left(h_k(\tilde{\alpha}) - \frac{h_k^2(\tilde{\alpha})}{2} \right)^2 + \sum_{k=1}^n [2h_k(\tilde{\alpha}) - h_k^2(\tilde{\alpha})] \right]_+ \\
& \leq \mathbf{E} \sup_{\alpha \geq 0} \left[\sum_{k=1}^n [h_k^2(\alpha) - 2h_k(\alpha)](1 - \xi_k^2) - \frac{1}{4} \sum_{k=1}^n [h_k^2(\alpha) - 2h_k(\alpha)]^2 \right]_+ \\
& \quad + 2\mathbf{E} \sum_{k=1}^n h_k(\tilde{\alpha}) \leq C + 2\mathbf{E} \sum_{k=1}^n h_k(\tilde{\alpha}).
\end{aligned}$$

Finally, combining this with (34)–(36), we finish the proof. \square

Proof of Theorem 3. We prove this theorem with the help of two basic inequalities. The first one follows immediately from Lemma 5. Namely,

$$(37) \quad \mathbf{E} \left[\varepsilon^2 - \frac{\|h(\hat{\alpha}_G)Z - Z\|^2}{n} \right]_+ \leq \frac{C\varepsilon^2}{\sqrt{n}},$$

which holds true since $\sum_{k=1}^n h_k(\hat{\alpha}_G) \leq \sqrt{n}$.

The second inequality is quite banal

$$(38) \quad \frac{\|h(\hat{\alpha}_G)Z - Z\|^2}{[1 - \|h(\hat{\alpha}_G)\|_1/n]^2} \leq \frac{\|h(\alpha_G^*)Z - Z\|^2}{[1 - \|h(\alpha_G^*)\|_1/n]^2},$$

where

$$\alpha_G^* = \arg \min_{\alpha: \|h(\alpha)\|_1 \leq \sqrt{n}} R(\alpha, \theta).$$

To combine these inequalities, we bound from below the left-hand side of (38). Using that $(1 - x)^{-2} \geq 1 + 2x$, $x \in [0, 1]$, we immediately get

$$\begin{aligned}
& \|h(\hat{\alpha}_G)Z - Z\|^2 + \frac{2}{n} \|h(\hat{\alpha}_G)Z - Z\|^2 \|h(\hat{\alpha}_G)\|_1 - \varepsilon^2 \sum_{k=1}^n \xi_k^2 \\
& \leq \frac{\|h(\alpha_G^*)Z - Z\|^2}{[1 - \|h(\alpha_G^*)\|_1/n]^2} - \varepsilon^2 \sum_{k=1}^n \xi_k^2,
\end{aligned}$$

and therefore

$$\begin{aligned}
& \mathbf{E} \left[\|h(\hat{\alpha}_G)Z - Z\|^2 + \frac{2}{n} \|h(\hat{\alpha}_G)Z - Z\|^2 \|h(\hat{\alpha}_G)\|_1 - \varepsilon^2 \sum_{k=1}^n \xi_k^2 \right] \\
& \leq \frac{R(\alpha_G^*, \theta) - \varepsilon^2 \|h(\alpha_G^*)\|_1/n}{[1 - \|h(\alpha_G^*)\|_1/n]^2} \leq \frac{R(\alpha_G^*, \theta)}{[1 - \|h(\alpha_G^*)\|_1/n]^2}.
\end{aligned}$$

Next we rewrite this inequality in the following equivalent form:

$$(39) \quad \mathbf{E}\|h(\hat{\alpha}_G)Z - Z\|^2 - \varepsilon^2 \mathbf{E} \sum_{k=1}^n \xi_k^2 + 2\varepsilon^2 \mathbf{E}\|h(\hat{\alpha}_G)\|_1 \\ \leq \frac{R(\alpha_G^*, \theta)}{[1 - \|h(\alpha_G^*)\|_1/n]^2} + 2\mathbf{E}\|h(\hat{\alpha}_G)\|_1 \left[\varepsilon^2 - \frac{\|h(\hat{\alpha}_G)Z - Z\|^2}{n} \right].$$

Using simple algebra together with (31) and (32), one can bound from below the left-hand side of the above display as follows:

$$(40) \quad \mathbf{E}\|h(\hat{\alpha}_G)Z - Z\|^2 - \varepsilon^2 \mathbf{E} \sum_{k=1}^n \xi_k^2 + 2\varepsilon^2 \mathbf{E}\|h(\hat{\alpha}_G)\|_1 \\ = \mathbf{E}\|h(\hat{\alpha}_G)Z - \theta\|^2 \\ + 2\varepsilon \mathbf{E} \sum_{k=1}^n (1 - h_k(\hat{\alpha}_G))\theta_k \xi_k + 2\varepsilon^2 \mathbf{E} \sum_{k=1}^n h_k(\hat{\alpha}_G)(1 - \xi_k^2) \\ = (1 - \gamma) \mathbf{E}\|h(\hat{\alpha}_G)Z - \theta\|^2 + \gamma \mathbf{E} \sum_{k=1}^n [1 - h_k(\hat{\alpha}_G)]^2 \theta_k^2 \\ + 2\varepsilon \mathbf{E} \sum_{k=1}^n [1 - (1 + \gamma)h_k(\hat{\alpha}_G) + \gamma h_k^2(\hat{\alpha}_G)]\theta_k \xi_k \\ + \varepsilon^2 \mathbf{E} \sum_{k=1}^n [2h_k(\hat{\alpha}_G) - \gamma h_k^2(\hat{\alpha}_G)](1 - \xi_k^2) + \gamma \varepsilon^2 \mathbf{E} \sum_{k=1}^n [h_k(\hat{\alpha}_G)]^2 \\ \geq (1 - \gamma) \mathbf{E}\|h(\hat{\alpha}_G)Z - \theta\|^2 - \frac{C\varepsilon^2}{\gamma}.$$

The right-hand side of (39) can be easily controlled since $\|h(\hat{\alpha}_G)\|_1 \leq \sqrt{n}$ and it follows immediately from (37) that

$$\mathbf{E}\|h(\hat{\alpha}_G)\|_1 \left[\varepsilon^2 - \frac{\|h(\hat{\alpha}_G)Z - Z\|^2}{n} \right] \leq \sqrt{n} \mathbf{E} \left[\varepsilon^2 - \frac{\|h(\hat{\alpha}_G)Z - Z\|^2}{n} \right]_+ \leq C\varepsilon^2.$$

Combining this inequality with (39) and (40) we get

$$\mathbf{E}\|h(\hat{\alpha}_G)Z - \theta\|^2 \leq \frac{1}{1 - \gamma} \left[\frac{R(\alpha_G^*, \theta)}{(1 - 1/\sqrt{n})^2} + \frac{C\varepsilon^2}{\gamma} \right],$$

thus completing the proof of the theorem. \square

References

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

- [2] Y. Cao and Yu. Golubev, *On oracle inequalities related to polynomial fitting*, Math. Methods Statist., **14** (2005), 431–450.
- [3] A. Demmler and C. Reinsch, *Oscillation matrices with spline smoothing*, Numer. Math., **24** (1975), 375–382.
- [4] Duhamel du Monceau, *Eléments de l'Architecture Navale ou Traité de la Construction des Vaissaux*, Paris, 1752.
- [5] R. Eubank, *Nonparametric Regression and Spline Smoothing* (2nd ed.), Dekker, New York, 1999.
- [6] Yu. Golubev and W. Härdle, *On adaptive smoothing in partial linear models*, Math. Methods Statist., **11** (2002), 98–117.
- [7] P. J. Green and B. W. Silverman, *Nonparametric Regression and Generalized Linear Models. A roughness penalty approach*, Chapman and Hall, 1994.
- [8] A. Kneip, *Ordered linear smoothers*, Ann. Statist., **22** (1994), 835–866.
- [9] C. L. Mallows, *Some comments on C_p* , Technometrics, **15** (1973), 661–675.
- [10] M. Nussbaum, *Spline smoothing in regression models and asymptotic efficiency in L_2* , Ann. Statist., **13** (1985), 984–997.
- [11] M. S. Pinsker, *Optimal filtering of square integrable signals in Gaussian white noise*, Problems Inform. Transmission, **16** (1980), 120–133.
- [12] R. Shibata, *An optimal selection of regression variables*, Biometrika, **68** (1981), 45–54.
- [13] P. Speckman, *Spline smoothing and optimal rates of convergence in nonparametric regression*, Ann. Statist., **13** (1985), 970–983.
- [14] C. M. Stein, *Estimation of the mean of a multivariate normal distribution*, Ann. Statist., **9** (1981), 1135–1151.
- [15] V. Tikhomirov, *Fundamental Principles of the Theory of Extremal Problems*, Wiley, New York, 1986.
- [16] A. Van der Vaart and J. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York, 1996.
- [17] G. Wahba, *Spline Models for Observational Data*, SIAM, Philadelphia, 1990.

[Received May 2006; revised December 2006]