The Annals of Statistics
1997, Vol. 25, No. 3, 929–947

OPTIMAL SPATIAL ADAPTATION TO INHOMOGENEOUS SMOOTHNESS: AN APPROACH BASED ON KERNEL ESTIMATES WITH VARIABLE BANDWIDTH SELECTORS[1]

BY O. V. LEPSKI,[2] E. MAMMEN AND V. G. SPOKOINY

Institute for System Analysis and Humboldt-Universität; Universität Heidelberg; and Institute for Information Transmission Problems and Institute of Applied Analysis and Stochastics

A new variable bandwidth selector for kernel estimation is proposed. The application of this bandwidth selector leads to kernel estimates that achieve optimal rates of convergence over Besov classes. This implies that the procedure adapts to spatially inhomogeneous smoothness. In particular, the estimates share optimality properties with wavelet estimates based on thresholding of empirical wavelet coefficients.

1. Introduction. In nonparametric curve estimation the statistical analysis may focus on the inference of the qualitative structure of the analyzed curve. Often, interesting features of the curve are connected with spatially inhomogeneous smoothness. In this case, curve estimates that are spatially adaptive are appropriate.

A variety of such procedures have been proposed in the literature. In Breiman, Friedman, Olshen and Stone (1983) piecewise constant least squares estimates are considered with a data adaptive choice of the pieces (CART). More generally, Friedman and Silverman (1989), Friedman (1991) and Luo and Wahba (1995) use variable knot splines (MARS). Knot points are added, removed and allocated recursively using cross-validation techniques. These methods have shown good performance in simulations and real data examples. However, no asymptotic theory is available.

Mammen and van de Geer (1997) discuss penalized least squares curve estimation for spatially inhomogeneous curves. They propose penalty terms which allow more spatial inhomogeneity than the usual L^2-norms of derivatives of the curve. The estimates turn out to be variable knot splines [see also Mammen (1991)]. Results on rates of convergence and a pointwise asymptotic distribution theory are given.

Müller and Stadtmüller (1987), Staniswalis (1989) and Brockmann, Gasser and Hermann (1993) propose kernel estimation with locally variable bandwidth selectors. The calculation of local bandwidths is based on pilot estimation of local smoothness characteristics. An asymptotic analysis is available

Received July 1994; revised August 1996.

[1]This work was supported by the Deutsche Forschungsgemeinschaft, Sonderforschungsbereich 373 "Quantifikation und Simulation ökonomischer Prozesse," Berlin, Germany.

[2]Research supported by the Deutsche Forschungsgemeinschaft under personal Grant 436 RUS 17/241/93.

AMS 1991 *subject classification*. 62G07.

Key words and phrases. Kernel estimate, bandwidth choice, Besov spaces, spatial adaptation, minimax rate of convergence.

here, however, only under additional smoothness conditions on the curve [for a discussion of this point see also Gijbels and Mammen (1994)]. Spatially adaptive local polynomial regression estimates were introduced and discussed in Fan and Gijbels (1995). For a comparison of wavelet estimates and local polynomial regression estimates with variable bandwidth selector see Fan, Hall, Martin and Patil (1993). In a series of papers Donoho, Johnstone, Kerkyacharian and Picard have shown that wavelet analysis offers a powerful technology for spatially adaptive curve estimation. Curve estimates based on thresholding empirical wavelet coefficients are optimal for a wide range of loss functions and smoothness classes [see Donoho, Johnstone, Kerkyacharian and Picard, (1995), Kerkyacharian and Picard (1993) and Delyon and Juditsky (1994)]. In particular, they achieve optimal minmax rates over balls in smoothness classes (e.g., Sobolev or Besov classes) with a norm that is weaker than the loss function used (e.g., $L_{p'}$-loss for functions with assumed bounded $\int |f^{(k)}|^p$, where $p' > p$). This shows spatial adaptivity of wavelet estimates because such classes contain functions with spatially inhomogeneous smoothness. In such classes optimal rates cannot be achieved by estimates that are linear in the observations [e.g., kernel estimates with deterministic bandwidth, orthogonal series estimates, regression splines with equidistant knot points, see Nemirovski (1985)]. The reason is that linear estimates cannot adapt to spatial inhomogeneity. Spatial adaptivity of an estimate can also be characterized by pointwise properties. For wavelet estimates it has been shown that up to a logarithmic factor they achieve the same risk as a variable knot spline with optimally placed (deterministic) knot points. This holds for every function [see Donoho and Johnstone (1992, 1994)] and has been called ideal spatial adaptation.

In this paper, a new variable bandwidth kernel estimate is proposed. The bandwidth selector is based on a modification of a procedure for adaptive estimation due to Lepski (1990). This estimate is a reasonable alternative to wavelet estimates. Another curve estimate using the approach of Lepski has been proposed in Goldenshluger and Nemirovski (1994); see also Goldenshluger and Nemirovski (1996). Our estimate shares some decision theoretical optimality properties with wavelet estimates. In particular, we prove that it achieves optimal rates over the whole scale of Besov classes and L_p-losses (see Theorems 3.1 and 3.2). This shows that this estimate adapts to spatial inhomogeneity (like wavelet estimates). Furthermore, a result on ideal spatial adaptation is given (see Theorem 3.3). All results are based on a general bound for the pointwise risk of our estimate (see Proposition 3.4). This result can be used to get the rates of our estimate also in other setups.

Our model and our procedure will be described in the next section. Section 3 contains our results. Performance of the estimate is illustrated by simulated data sets in Section 4. The proofs are postponed to Section 5.

2. A data adaptive local bandwidth selector. In this paper we consider the white noise model

$$dY(t) = f(t)\,dt + \sigma\,dW(t), \qquad 0 \le t \le 1, \tag{2.1}$$

where $W(t)$, $0 \le t \le 1$, is a Brownian motion and f is an unknown (regression) function. Performance of estimates of f is studied for $\sigma \to 0$. Model (2.1) gives an asymptotic description for density estimation with i.i.d. observations and for nonparametric regression with i.i.d. Gaussian errors and sample size of order σ^{-2} [see Brown and Low (1996), Low (1992) and Nussbaum (1996)]. In particular, our results on rates of convergence can be shown to hold for regression models under conditions on the tails of the error distribution.

We will study kernel estimates $\hat{f}_h$ with kernel K and bandwidth h:

$$\hat{f}_h(x) = \int K_h(x-t)\,dY(t), \tag{2.2}$$

where $K_h(x) = h^{-1}K(x/h)$. We also write

$$f_h(x) = E_f \hat{f}_h(t) = \int K_h(x-t)f(t)\,dt, \tag{2.3}$$

$$v^2(h) = \operatorname{Var} \hat{f}_h(t) = \sigma^2 h^{-1} \int K^2(u)\,du. \tag{2.4}$$

We propose a local bandwidth selector $\hat{h}(t)$. It takes values in the geometrical grid

$$\mathscr{H}_\sigma = \{h \in [\sigma^2, h_\sigma^*]\colon h = a^{-j}h_\sigma^*,\ \ j = 0, 1, 2, \ldots\}.$$

Here $a > 1$ is an arbitrary constant. The upper bound h_σ^* will be specified below. We write L_σ for the number of elements of $\mathscr{H}_\sigma$.

Now we define

$$\hat{h}(t) = \sup\{h \in \mathscr{H}_\sigma\colon |\hat{f}_h(t) - \hat{f}_\eta(t)| \le \psi(h, \eta) \text{ for all } \eta < h,\ \ \eta \in \mathscr{H}_\sigma\},$$

where, for $h > \eta$,

$$\psi(h, \eta) = D_1 v(h)\lambda(h) + v(h, \eta)\lambda(\eta), \tag{2.5}$$

$$\lambda(h) = \max\left\{1, \sqrt{D_2 \log(h_\sigma^*/h)}\right\}. \tag{2.6}$$

Here D_1 and D_2 are positive constants, $v(h)$ is the standard deviation of $\hat{f}_h(t)$ [see (2.4)] and $v(h, \eta)$, is the standard deviation of the difference $\hat{f}_h(t) - \hat{f}_\eta(t)$,

$$\begin{aligned} v^2(h, \eta) &= \sigma^2 \int [K_h(u) - K_\eta(u)]^2\,du \\ &= \sigma^2 \eta^{-1} \int [K(u) - (\eta/h)K(uh/\eta)]^2\,du. \end{aligned} \tag{2.7}$$

We propose the estimate

$$\hat{f}(t) = \hat{f}_{\hat{h}(t)}(t).$$

A modification of $\hat{f}$ based on piecewise constant choices of $\hat{h}$ is discussed in Lepski and Spokoiny (1994). The bandwidth $\hat{h}(t)$ has a nice statistical interpretation. It is the largest bandwidth h such that $\hat{f}_h(t)$ does not differ

"significantly" from kernel estimates with smaller bandwidth: one chooses a resolution level such that no significant features are visible on a finer resolution level. For $\eta < h$ the difference $\hat{f}_h(t) - \hat{f}_\eta(t)$ is of stochastic order $v(h, \eta)$. The additional logarithmic factor $\lambda(\eta)$ has been added because the definition of $\hat{h}_\sigma(t)$ is based on an increasing number of comparisons of different bandwidths. This additional factor is essential for our results (see the remark at the beginning of the proof of Proposition 3.4). Without this factor too small bandwidths would be chosen.

Our approach has an essential difference from wavelet estimation techniques based on thresholding of empicial wavelet coefficients. Empirical wavelet coefficients are related to the values

$$Z_{j,\sigma}(t) = \hat{f}_{2^{-j}h^*_\sigma}(t) - \hat{f}_{2^{-j-1}h^*_\sigma}(t).$$

A kernel estimate analogue of the wavelet threshold estimates would look like

$$\tilde{f}(t) = \hat{f}_{h^*_\sigma}(t) + \sum_{j \geq 0} Z_{j,\sigma}(t) \mathbb{I}(|Z_{j,\sigma}(t)| \geq C_{j,\sigma})$$

with appropriate threshold values $C_{j,\sigma}$; $\mathbb{I}$ denotes the indicator function. In particular, in contrast to $\hat{f}$, this method is based on comparison of neighboring resolution levels and it may find that "significant" features are present in the data for arbitrarily many resolution levels.

3. Near minimaxity and ideal spatial adaptation. We will study the rate of convergence of the estimate $\hat{f}$, defined in the last section, over balls $B^s_{p,q}(M)$ in Besov spaces $B^s_{p,q}$ and show that our curve estimate achieves optimal rates of convergence over these function classes. The following characterization of a Besov ball will be used in the proofs of Theorems 3.1 and 3.2:

$$B^s_{p,q}(M) = \{f\colon \|f\|_{B^s_{p,q}} \leq M\},$$

where

$$\|f\|_{B^s_{p,q}} = \begin{cases} \|f\|_p + \left[\int_0^1 h^{-sq} \|\operatorname{osc} f(\cdot, h)\|_p^q \frac{dh}{h}\right]^{1/q}, & \text{if } q < \infty, \\ \|f\|_p + \sup\limits_{0 \leq h \leq 1} h^{-s} \|\operatorname{osc} f(\cdot, h)\|_p, & \text{if } q = +\infty. \end{cases}$$

Here $\|f\|_p$ is the L_p-norm $\|f\|_p^p = \int_0^1 |f|^p$. Furthermore, for the definition of the local oscillation $\operatorname{osc} f(x, h)$ of the function f, an arbitrary $r \in \mathbb{N}$ with $r \geq s$ and a real u have to be chosen. The constant u has to fulfill

$$\begin{aligned} &1 \leq u \leq +\infty && \text{if } sp > 1, \\ &1 \leq u < +\infty && \text{if } sp = 1, \\ &1 \leq u < p(1 - sp)^{-1} && \text{if } sp < 1. \end{aligned}$$

With this choice of r and u the local oscillation $\operatorname{osc} f(x,h)$ of f is defined as

$$\operatorname{osc} f(x,h) = \begin{cases} \inf \sup_{|y-x|\le h} |f(y)-P(y)|, & \text{if } u=+\infty, \\ \inf\left[\frac{1}{2h}\int_{|y-x|\le h} |f(y)-P(y)|^u\,dy\right]^{1/u}, & \text{if } u<+\infty. \end{cases} \tag{3.1}$$

The infimum in (3.1) is taken over all polynomials of order r.

A proof that $\|\cdot\|_{B^s_{p,q}}$ is a norm of $B^s_{p,q}$ can be found in Triebel [(1992), Section 3.5.1]. Other equivalent norms are discussed there, too.

For the kernel K we make the following assumptions for an integer $k \ge 1$.

(K1). The kernel K has a compact support (say, $[-1,1]$); K is continuous and fulfills $\int K(u)\,du = 1$ and it has k vanishing moments: $\int u^i K(u)\,du = 0$, for $1 \le i \le k$. For $t < h$ and $t > 1-h$ the kernel K_h is replaced by boundary kernels K^t_h (kernels with support $[-t,h]$ or $[-h,1-t]$, respectively). We assume that the functions $hK^t_h(\cdot)$ are uniformly bounded, and that the kernels K^t_h fulfill $\int K^t_h(u)\,du = 0$ and have k vanishing moments.

For simplicity, our notation will not take into account the modifications at the boundary, in particular we will skip the superscript t in K^t_h. For the case that the inequalities $s \le 1/p$ and $q < +\infty$ hold, we need the following additional condition.

(K2). The kernel K can be decomposed as

$$K(u) = 2M(u) - \frac{1}{2}M\left(\frac{u}{2}\right),$$

where M is a bounded function with compact support (say, $[-1/2,+1/2]$) and with $\int M(u)\,du = 1$. Without any indication in the notation, modifications of M are used again at the boundary. They are assumed to be uniformly bounded.

Note that (K2) implies $\int K(u)\,du = 1$ and $\int uK(u)\,du = 0$.

We will study maximal $L_{p'}$-risks of $\hat f$ over $B^s_{p,q}$ balls of bounded functions. For fixed $M > 0$ and $L > 0$ we put

$$R_\sigma(\hat f, B^s_{p,q}, p') = \sup_{f \in B(M,L)} E_f \|\hat f - f\|^{p'}_{p'},$$

where

$$B(M,L) = \begin{cases} B^s_{p,q}(M), & \text{if } sp > 1, \\ B^s_{p,q}(M) \cap \{f\colon |f| \le L\}, & \text{if } sp \le 1. \end{cases} \tag{3.2}$$

For $sp \le 1$ functions in $B^s_{p,q}(M)$ are not uniformly bounded. This is the reason why the restriction $|f| \le L$ has been added for this case.

For simplicity, our notation does not always indicate every dependence. For instance, remember that $\hat f$ depends on σ and the choice of D_1, D_2, a and

h_σ^*. Furthermore, it depends on the kernel K (and its number k of vanishing moments).

We are now ready to state our main result.

THEOREM 3.1. *Suppose that for the parameters of the Besov class it holds that* $1 \le p, q \le +\infty$, $1 \le p' < +\infty$, $s > (1/p - 1/p')_+$. *The kernel* K *is assumed to fulfill* (K1) *with* $k > [s]$. *Additionally, in the case that* $s \le 1/p$ *and* $q < +\infty$ *hold,* K *is supposed to satisfy* (K2). *Furthermore, assume that* $\hat{h}(t)$ *is calculated with* $D_1 > 0$, $D_2 \ge 2p'$ *and with* $h_\sigma^* = \sigma^{2/(2s+1)}$. *Then, for* σ *small enough, the risks of* $\hat{f}$ *satisfy*

$$R_\sigma(\hat{f}, B^s_{p,q}, p') \le \text{const.} \begin{cases} \sigma^{p'r}, & \text{if } sp > \dfrac{p'-p}{2}, \\ \left[\sigma\sqrt{\log\left(\dfrac{1}{\sigma}\right)}\right]^{p'r'} \log\left(\dfrac{1}{\sigma}\right), & \text{if } sp = \dfrac{p'-p}{2}, \\ \left[\sigma\sqrt{\log\left(\dfrac{1}{\sigma}\right)}\right]^{p'r'}, & \text{if } sp < \dfrac{p'-p}{2}. \end{cases} \tag{3.3}$$

where

$$r = \frac{2s}{2s+1},$$
$$r' = \frac{s - 1/p + 1/p'}{s - 1/p + 1/2}$$

and const. *depends only on the kernel* K, *the parameters* s, p, L, M *of the function class, the norm power* p' *and the parameters* D_1, D_2, a *of the bandwidth selector.*

The exponent of σ in (3.3) gives the optimal rate. For $sp \ne (p'-p)/2$ this holds also for the logarithmic factor. Small choices of the class parameter p correspond to Besov classes that contain functions with spatially inhomogeneous smoothness. Because our estimates achieve optimal (or nearly optimal) rates in all Besov classes this shows that the estimates adapt well to spatially inhomogeneous smoothness. For a discussion of minimax rates in Besov spaces we refer to Donoho et al. (1995) and Delyon and Juditsky (1994).

For the interpretation of the exponents in (3.3) let us briefly remark that for the case of $sp \le (p'-p)/2$ we have $r' > 0$. This follows from $s - 1/p + 1/p' > 0$ and $s - 1/p + 1/2 > 0$. The first of these two inequalities follows from $p' > p$. For the proof of the second inequality we apply $sp \le (p'-p)/2$ and our condition $s > (1/p - 1/p')_+ = (1/p - 1/p')$ to obtain

$$\frac{p'}{2} - 1 \ge sp + \frac{p}{2} - 1 > \left(\frac{1}{p} - \frac{1}{p'}\right)p + \frac{p}{2} - 1 = \frac{p}{p'}\left(\frac{p'}{2} - 1\right).$$

Because of $p' > p$ this implies $p' > 2$ and $sp - 1 + p/2 > 0$.

The choice of h_σ^* in Theorem 3.1 requires explicit knowledge of s. The next theorem helps to understand the performance of $\hat{f}$ in case of unknown degree s of smoothness.

THEOREM 3.2. *Under the assumptions of Theorem* 3.1 *for a choice of* h_σ^* *with* $\sigma^{2/(2s+1)} \le h_\sigma^* \le 1$ *one gets for* σ *small enough*

$$R_\sigma(\hat{f}, B^s_{p,q}, p') \le \text{const.} \begin{cases} \left[\sigma\sqrt{\log\left(\frac{1}{\sigma}\right)}\right]^{p'r}, & \text{if } sp > \frac{p'-p}{2}, \\ \left[\sigma\sqrt{\log\left(\frac{1}{\sigma}\right)}\right]^{p'r'} \log\left(\frac{1}{\sigma}\right), & \text{if } sp = \frac{p'-p}{2}, \\ \left[\sigma\sqrt{\log\left(\frac{1}{\sigma}\right)}\right]^{p'r'}, & \text{if } sp < \frac{p'-p}{2}. \end{cases} \tag{3.4}$$

Here r *and* r' *are defined as in Theorem* 3.1.

By comparing the results of Theorems 3.1 and 3.2 we find that using $h_\sigma^* = 1$ gives the optimal rate for $sp \le (p'-p)/2$ and an additional logarithmic factor for $sp > (p'-p)/2$.

Now we state a property of $\hat{f}$ that was called ideal spatial adaptation in Donoho and Johnstone (1994). We fix some point $t \in [0,1]$ and study the pointwise risk $r_\sigma(t,f) = E_f|\hat{f}(t) - f(t)|^2$ (here $p' = 2$). We would like to compare this risk with $\inf E|\hat{f}_h(t) - f(t)|^2$, where the infimum runs over all (deterministic) bandwidths h with $\sigma^2 < h < 1$. Note that

$$E|\hat{f}_h(t) - f(t)|^2 = (f_h(t) - f(t))^2 + \operatorname{Var}\hat{f}_h(t) = (f_h(t) - f(t))^2 + \sigma^2\|K\|_2^2 h^{-1}.$$

Denote

$$r_{\text{ideal}}(t,f) = \inf_{a\sigma^2 \le h \le 1}\Big[\sup_{0\le\eta\le h}(f_\eta(t) - f(t))^2 + \sigma^2\|K\|_2^2 h^{-1}\Big].$$

The first term $\sup_{0\le\eta\le h}(f_\eta(t)-f(t))^2$ reflects the local smoothness of f in the interval $[t-h, t+h]$. The second term is the variance of $\hat{f}_h(t)$. The minimizing "ideal" bandwidth h_{ideal} provides a trade-off between these two terms, but it depends on unknown characteristics of the function f. It is known from Lepski (1990) and Brown and Low (1992) that in pointwise estimation one has to pay an additional logarithmic factor for not knowing smoothness properties of f. In the present context this means that no estimate achieves the risk of order $r_{\text{ideal}}(t,f)$ adaptively (uniformly over large enough function classes). This loss of efficiency can be viewed as payment for estimation of unknown smoothness parameters of f. The loss can be quantified by an increased noise level:

$$r_{\text{adapt}}(t,f) = \inf_{a\sigma^2 \le h \le 1}\Big[\sup_{0\le\eta\le h}(f_\eta(t) - f(t))^2 + \sigma^2\log(1/\sigma)\|K\|_2^2 h^{-1}\Big].$$

Obviously, $r_{\text{adapt}}(t,f) \le r_{\text{ideal}}(t,f)\log(1/\sigma)$.

THEOREM 3.3. *Suppose that the kernel K fulfills* (K1) *with $k \geq 1$, and $\hat{h}(t)$ is calculated with $D_1 > 0$, $D_2 \geq 4$ and $h_\sigma^* = 1$. Then one has, for each $L > 0$ uniformly in $f \in \mathscr{F}(L) = \{f\colon \sup_{x\in[0,1]} f(x) - \inf_{x\in[0,1]} f(x) \leq L\}$ and $t \in (0,1)$ for σ small enough,*

$$E_f(\hat{f}(t) - f(t))^2 \leq C r_{\text{adapt}}(t, f), \tag{3.5}$$

where the constant C depends only on p', a, D_1, D_2 and the kernel K.

It can be shown that for every estimate that fulfills (3.5) [where $r_{\text{adapt}}(t, f)$ is defined with a kernel K satisfying the conditions of Theorem 3.2], the statement (3.4) of Theorem 3.2 also holds (with $p' = 2$). This means that every estimate which is locally adaptive at each point t in the sense of (3.5) is automatically globally (spatially) adaptive and nearly minimax over the wide scale of Besov classes in the sense of Theorem 3.2. In the context of adaptive estimation the rate r_{adapt} is optimal [see Lepski and Spokoiny (1996)]. For further discussions of adaptive estimation see also Lepski (1991, 1992).

For the proof of our three theorems we will make use of the following proposition. There an upper bound is given for pointwise risks $r_\sigma(t, f) = E_f|\hat{f}(t) - f(t)|^{p'}$. Like Theorem 3.3 the proposition relates the pointwise risk of $\hat{f}$ to the pointwise risk of a kernel estimate with deterministic bandwidth $h_\sigma(t, f)$. By application of this proposition the proofs of Theorems 3.1 and 3.2 are reduced to approximation theoretical considerations. The proposition can be used to treat function classes other than Besov classes.

PROPOSITION 3.4. *Let the kernel K obey* (K1) *with $k \geq 1$. For an $L > 0$, let $\hat{h}(t)$ be calculated with $D_1 > 0$, $D_2 \geq 2p'$ and $H\sigma^2 \leq h_\sigma^* \leq 1$, where $H = \exp\{2L^2/(D_1^2 D_2)\}$. Then there exists a constant C (depending on p', a, D_1, D_2 and the kernel K) such that the following inequality holds uniformly in $f \in \mathscr{F}(L)$ and $t \in (0,1)$:*

$$E_f|\hat{f}(t) - f(t)|^{p'} \leq C\{v(h_\sigma(t, f))\lambda(h_\sigma(t, f))\}^{p'}. \tag{3.6}$$

Here $h_\sigma(t, f)$ is the following local (deterministic) bandwidth:

$$\begin{aligned} &h_\sigma(t, f) \\ &\quad = \sup\{h \in \mathscr{H}_\sigma\colon |f_\eta(t) - f(t)| \leq D_1 v(h)\lambda(h) \text{ for all } \eta \in \mathscr{H}_\sigma,\ \eta \leq h\}. \end{aligned} \tag{3.7}$$

The quantities $v^2(h)$, $\psi(h, \eta)$, $\lambda(h)$ and $v^2(h, \eta)$ have been defined in (2.4), (2.5), (2.6) *or* (2.7), *respectively. The set $\mathscr{F}(L)$ has been introduced in Theorem* 3.3.

The bandwidth $h_\sigma(t, f)$ is well defined by (3.7). The supremum on the right-hand side of (3.7) is taken over a nonempty set. This follows from the following two inequalities (3.8) and (3.9). Using $f \in \mathscr{F}(L)$ and that K has support $[-1, 1]$, we get

$$|f_\eta(t) - f(t)| \leq L\|K\|_1 \leq \sqrt{2}L\|K\|_2. \tag{3.8}$$

From $h_\sigma^* \geq H\sigma^2$ and the definition of H it follows that

$$v(\sigma^2)\lambda(\sigma^2) = \sqrt{D_2 \log(h_\sigma^*/\sigma^2)}\|K\|_2 \geq \sqrt{D_2 \log H}\|K\|_2 = \sqrt{2}L\|K\|_2. \tag{3.9}$$

Our results hold for some modifications of $\hat{f}(t)$. In particular, one could replace $\psi(h, \eta)$ by

$$\psi(h, \eta) = \psi_{\mathrm{MOD}}(\eta) = D_3 \frac{\sigma}{\sqrt{\eta}} \sqrt{1 + \ln\left(\frac{h_\sigma^*}{\eta}\right)}. \tag{3.10}$$

It can be shown that Theorems 3.1, 3.2 and 3.3 continue to hold for this modification of $\hat{f}(t)$ if (in Theorems 3.1 and 3.2) D_3 has been chosen larger than $2\{1 + [4p' + 6]^{1/2}\|K\|\}$. For this modification of $\hat{f}$, simulations are presented in the next section.

4. Some simulated data sets. For the illustration of the performance of our estimate we have carried out some simulations in a regression setup. We have simulated the examples in Donoho and Johnstone (1994). Simulations of other estimates for these examples can be found in Goldenshluger and Nemirovski (1994), Luo and Wahba (1995) and Fan and Gijbels (1995). Figure 1 shows the regression functions with Gaussian noise (sample size 2048, error variance 1). In all four cases the signal-to-noise level is 7. In the estimation we have used the biweight kernel $K(u) = (1 - u^2)_+^2$. We have used no boundary kernels. As ψ we have chosen $\psi(h, \eta) = \psi_{\mathrm{MOD}}(\eta)$, see (3.10). In this setup the bandwidth $\hat{h}(t)$ is defined as

$$\hat{h}(t) = \sup\Bigg\{h \in \mathscr{H}\colon |\hat{f}_h(t) - \hat{f}_\eta(t)| \leq D_3 \frac{s}{\sqrt{n\eta}} \sqrt{1 + \ln\left(\frac{h_\sigma^*}{\eta}\right)} \quad \text{for all } \eta < h,\ \eta \in \mathscr{H}\Bigg\},$$

where n is the sample size, s^2 is the variance of the data and (with $a = 1.02$ and $h_\sigma^* = 0.4$)

$$\mathscr{H} = \{h \in [0.001, 0.4]\colon h = (1.02)^{-j}0.4,\ \ j = 0, 1, 2, \ldots\}.$$

The estimate $\hat{f}_h$ has been chosen as a Nadaraya–Watson estimate. Constants $D_3 = 1$, $D_3 = 1.5$ and $D_3 = 2$ have been used. The data sets have been generated 101 times. Figure 2 shows the kernel estimate $\hat{f}$ with median integrated squared error for $D_3 = 1.5$. The estimates look similiar to the curve estimates in Donoho and Johnstone (1994), Goldenshluger and Nemirovski (1994), Luo and Wahba (1995) and Fan and Gijbels (1995). For $D_3 = 1$ the estimates have a rougher shape, for $D_3 = 2$ they are smoother. Some further investigations are needed here. In particular this concerns the choice of constants, other modifications of $\hat{h}_\sigma(t)$ and applications to other smoothers (than Nadaraya–Watson estimates).

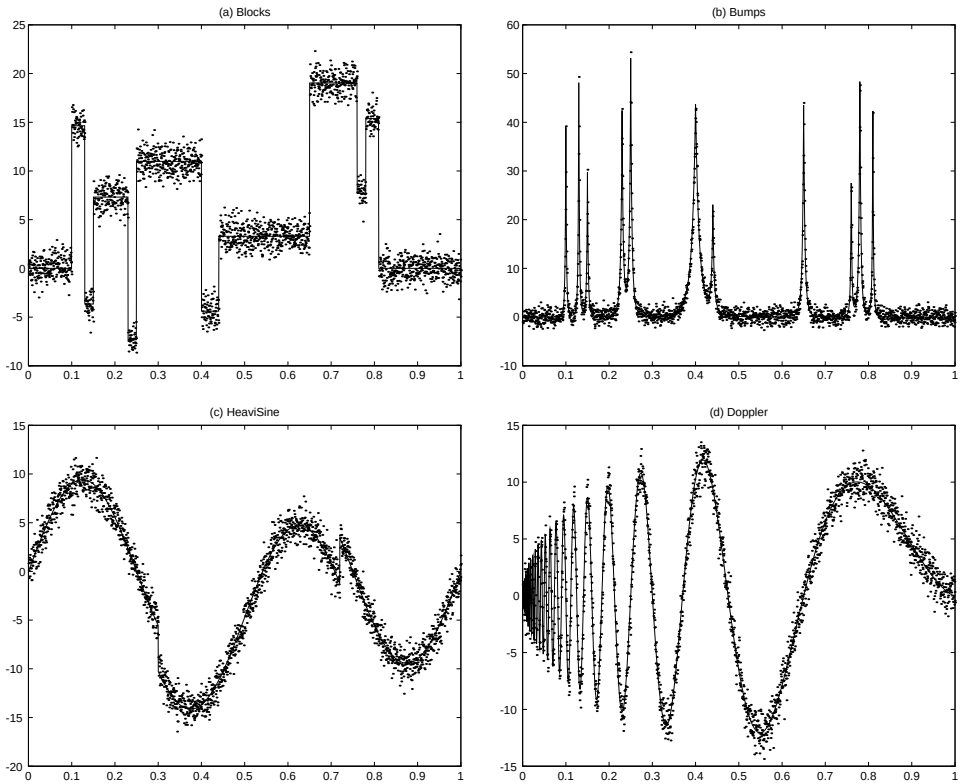

FIG. 1. *Four functions* [*see Donoho and Johnstone* (1992)] *with Gaussian white noise*, $\sigma = 1$, *with* f *rescaled to have signal-to-noise ratio* 7; *sample size* $n = 2048$.

5. Proofs. The proofs of Theorems 3.1, 3.2 and 3.3 rely on Proposition 3.4. We start with the proof of the proposition.

PROOF OF PROPOSITION 3.4. Let us fix $L > 0$, $f \in \mathcal{F}(L)$ and $t \in [0, 1]$ and let $h_\sigma(t, f)$ be defined by (3.7). For simplification, any dependence on f will not be indicated in the notation. For example, we write $h_\sigma(t) = h_\sigma(t, f)$ and $r_\sigma(t) = r_\sigma(t, f)$.

The proof of the proposition is based on the following idea. We distinguish the following two cases:

$$\hat{h}(t) \geq h_\sigma(t) \quad \text{and} \quad \hat{h}(t) < h_\sigma(t).$$

We denote the corresponding events by $A_\sigma(t) = \{\hat{h}(t) \geq h_\sigma(t)\}$ and $A_\sigma^c(t) = \{\hat{h}(t) < h_\sigma(t)\}$. For the pointwise risk we get

$$\begin{aligned} r_\sigma(t) &= E_f|\hat{f}(t) - f(t)|^{p'} \\ &\leq r_\sigma^+(t) + r_\sigma^-(t), \end{aligned}$$

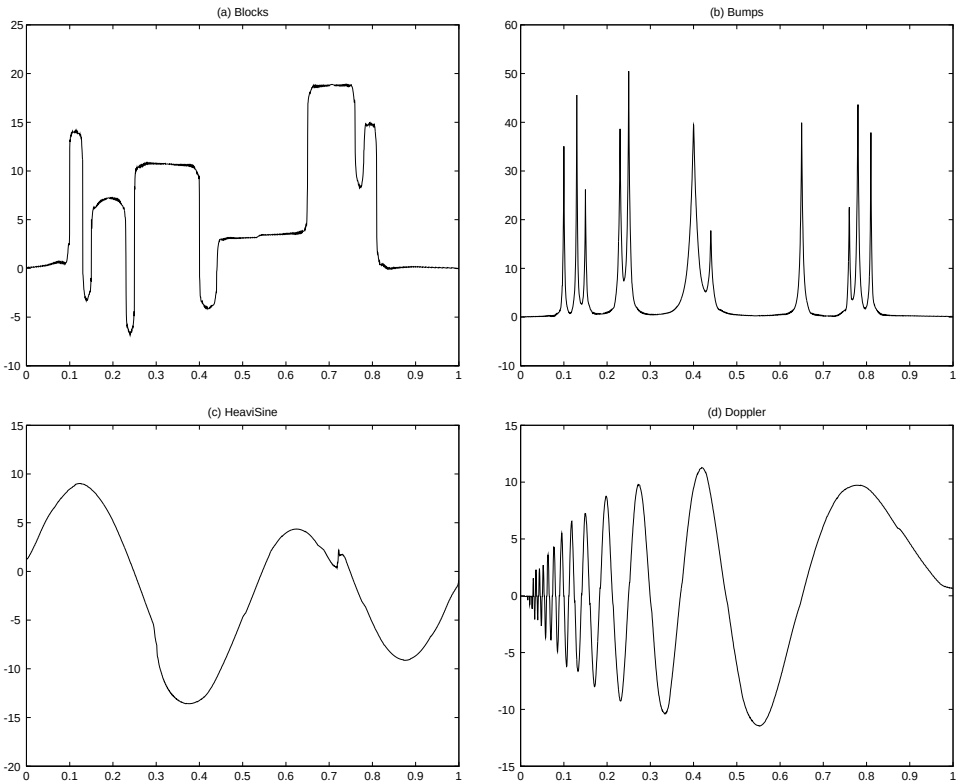

FIG. 2. *Kernel estimates of the four functions in Figure* 1; *local bandwidth* $\hat{h}$ *with* $\psi(h, \eta) = \psi_{\mathrm{MOD}}(\eta)$, $D_3 = 1.5$, $a = 1.02$ *and* $h_\sigma^* = 0.4$.

where

$$r_\sigma^+(t) = E_f|\hat{f}(t) - f(t)|^{p'}\mathbb{I}(A_\sigma(t)),$$

$$r_\sigma^-(t) = E_f|\hat{f}(t) - f(t)|^{p'}\mathbb{I}(A_\sigma^c(t)).$$

We will show

$$r_\sigma^+(t) \le \text{const.}[v(h_\sigma(t))\lambda(h_\sigma(t))]^{p'}, \tag{5.1}$$

$$r_\sigma^-(t) \le \text{const.}(\sigma^2/h_\sigma^*)^{p'/2}, \tag{5.2}$$

where const. depends on p', a, D_1, D_2 and the kernel K. The proof of (5.1) is rather simple. It uses that, by definition of $\hat{h}(t)$, on the event $A_\sigma(t)$ we can bound the difference between $\hat{f}(t)$ and $\hat{f}_{h_\sigma(t)}(t)$, [see (5.3)] [because of $\hat{h}(t) \ge h_\sigma(t)$]. The estimate $\hat{f}_{h_\sigma(t)}(t)$ is a kernel estimate with deterministic bandwidth. Its asympotic behavior is well understood. The proof of (5.2) is based on upper estimates of $E_f|\hat{f}(t) - f(t)|^{p'}\mathbb{I}(\hat{h}(t) = h)$ for $h < h_\sigma(t)$ in

$\mathscr{H}_\sigma$. The event $\{\hat{h}(t) = h\}$ can be bounded by unions of events $B_\sigma(t, h, \eta) = \{|\hat{f}_h(t) - \hat{f}_\eta(t)| > \psi(h, \eta)\}$ with $\eta < h$. At this point we need that the probability of the events $B_\sigma(t, h, \eta)$ decreases exponentially with $\eta \to 0$. This decrease is guaranteed by the logarithmic factor $\lambda(\eta)$ in the definition of $\hat{h}(t)$. It will allow us to bound the probability of $\{\hat{h}(t) = h\}$ by the sum of probabilities of the events $B_\sigma(t, h, \eta)$.

PROOF OF (5.1). First note that the definitions of $h_\sigma(t)$ and of $A_\sigma(t)$ imply

$$|\hat{f}(t) - \hat{f}_{h_\sigma(t)}(t)|\mathbb{I}(A_\sigma(t)) \le \psi^*(h_\sigma(t)), \tag{5.3}$$

$$|f(t) - f_{h_\sigma(t)}(t)| \le D_1 v(h_\sigma(t))\lambda(h_\sigma(t)), \tag{5.4}$$

where

$$\psi^*(h) = \sup\{\psi(h', h)\colon h' \in \mathscr{H}_\sigma, h' > h\}.$$

Using the bound $\psi^*(h) \le 2(D+1)v(h)\lambda(h)$ and the fact that $\hat{f}_{h_\sigma(t)}(t) - f_{h_\sigma(t)}(t)$ has a normal distribution with mean 0 and variance $v^2(h_\sigma(t))$ we get, from (5.3) and (5.4),

$$\begin{aligned} r_\sigma^+(t) &= E_f|\hat{f}(t) - f(t)|^{p'}\mathbb{I}(A_\sigma(t)) \\ &\le \text{const.}\big\{E_f|\hat{f}(t) - \hat{f}_{h_\sigma(t)}(t)|^{p'}\mathbb{I}(A_\sigma(t)) + E_f|\hat{f}_{h_\sigma(t)}(t) - f_{h_\sigma(t)}(t)|^{p'} \\ &\qquad + E_f|f_{h_\sigma(t)}(t) - f(t)|^{p'}\big\} \\ &\le \text{const.}'[v(h_\sigma(t))\lambda(h_\sigma(t))]^{p'}. \end{aligned}$$

The constants const. and const.$'$ depend only on D_1 and p'.

PROOF OF (5.2). Write for $h, \eta \in \mathscr{H}_\sigma$ with $h > \eta$,

$$B_\sigma(t, h, \eta) = \{|\hat{f}_h(t) - \hat{f}_\eta(t)| > \psi(h, \eta)\}.$$

With this notation we get from the definition of $\hat{h}(t)$ that, for each $h \in \mathscr{H}_\sigma$,

$$\{\hat{h}(t) = ha^{-1}\} \subseteq \bigcup_{\eta \in \mathscr{H}_\sigma(h)} B_\sigma(t, h, \eta),$$

where $\mathscr{H}_\sigma(h)$ is the set $\{\eta \in \mathscr{H}_\sigma\colon \eta < h\}$. Using

$$A_\sigma^c \subseteq \bigcup_{h \in \mathscr{H}_\sigma(ah_\sigma)} \bigcup_{\eta \in \mathscr{H}_\sigma(h)} B_\sigma(t, h, \eta),$$

we get

$$\begin{aligned} r_\sigma^- &\le \sum_{h \in \mathscr{H}_\sigma(ah_\sigma(t))} E_f|\hat{f}_h(t) - f(t)|^{p'}\mathbb{I}(\hat{h}(t) = h) \\ &\le \sum_{h \in \mathscr{H}_\sigma(ah_\sigma(t))} \sum_{\eta \in \mathscr{H}_\sigma(h)} E_f|\hat{f}_h(t) - f(t)|^{p'}\mathbb{I}(B_\sigma(t, h, \eta)). \end{aligned}$$

Now the definition of $h_\sigma(t)$ yields for any $h \le h_\sigma(t)$ that

$$|f_h(t) - f(t)| \le Dv(h_\sigma(t))\lambda(h_\sigma(t)) \le Dv(h)\lambda(h). \tag{5.5}$$

Using (5.5) we get, for $\eta < h \le h_\sigma$,

$$\begin{aligned} B_\sigma(t,h,\eta) &\subseteq \{2D_1 v(h)\lambda(h) + v(h,\eta)|\xi(t,h,\eta)| > \psi(h,\eta)\} \\ &\subseteq \{|\xi(t,h,\eta)| > \lambda(\eta)\}, \end{aligned}$$

where $\xi(t,h,\eta) = v^{-1}(h,\eta)\{[\hat{f}_\eta(t) - \hat{f}_h(t)] - [f_\eta(t) - f_h(t)]\}$ is a Gaussian variable with mean 0 and variance 1. This gives

$$r_\sigma^- \le \sum_{h\in\mathscr{H}_\sigma(ah_\sigma(t))} \sum_{\eta\in\mathscr{H}_\sigma(h)} E_f|\hat{f}_h(t) - f(t)|^{p'}\mathbb{I}\big(|\xi(t,h,\eta)| > \lambda(\eta)\big).$$

Again applying (5.5) we get

$$\begin{aligned} r_\sigma^- \le \sum_{h\in\mathscr{H}_\sigma(ah_\sigma(t))} \sum_{\eta\in\mathscr{H}_\sigma(h)} & E_f\big[Dv(h_\sigma)\lambda(h_\sigma) + v(h)|\xi(t,h)|\big]^{p'} \\ & \times \mathbb{I}\big(|\xi(t,h,\eta)| > \lambda(\eta)\big), \end{aligned} \tag{5.6}$$

where $\xi(t,h) = v^{-1}(h)\{\hat{f}_h(t) - f_h(t)\}$ is another Gaussian variable with mean 0 and variance 1. We now apply the following lemma.

LEMMA 5.1. *Let ξ and ξ' be standard Gaussian random variables. Then, for any constants $b \ge 0$ and $c \ge 1$,*

$$E(b + |\xi'|)^{p'}\mathbb{I}(|\xi| > c) \le \text{const.}|b + c|^{p'}\exp(-c^2/2),$$

where const. *depends only on p'.*

PROOF. Denote $\varrho = E\xi'\xi$. Then one can decompose ξ' into $\xi' = \varrho\xi + \sqrt{1-\varrho^2}\xi^\perp$, where $\xi^\perp$ has the standard normal distribution and is independent of ξ. The lemma follows from the following inequality:

$$\begin{aligned} E(b + |\xi'|)^{p'}\mathbb{I}(|\xi| > c) &\le E\big(b + |\varrho\xi| + \sqrt{1-\varrho^2}|\xi^\perp|\big)^{p'}\mathbb{I}(|\xi| > c) \\ &\le \text{const. } E\big[(1-\varrho^2)^{p'/2} + (b + |\varrho\xi|)^{p'}\big]\mathbb{I}(|\xi| > c) \\ &\le \text{const.}(b + c)^{p'}\exp(-c^2/2). \end{aligned}$$

□

By application of this lemma we get, from (5.6) (because of $D_2 \ge 2p'$),

$$\begin{aligned} r_\sigma^- &\le \text{const.} \sum_{h\in\mathscr{H}_\sigma(ah_\sigma(t))} \sum_{\eta\in\mathscr{H}_\sigma(h)} v(h)^{p'}[D_1\lambda(h_\sigma(t))v(h_\sigma(t))/v(h) + \lambda(\eta)]^{p'} \\ &\qquad \times \exp(-\lambda^2(\eta)/2) \\ &\le \text{const.}'\sigma^{p'} \sum_{h\in\mathscr{H}_\sigma(ah_\sigma(t))} h^{-p'/2} \sum_{\eta\in\mathscr{H}_\sigma(h)} (\eta/h_\sigma^*)^{p'}[\log(h_\sigma^*/\eta)]^{p'/2} \\ &\le \text{const.}''\sigma^{p'} \sum_{h\in\mathscr{H}_\sigma(ah_\sigma(t))} h^{-p'/2}(h/h_\sigma^*)^{p'}[\log(h_\sigma^*/h)]^{p'/2} \\ &\le \text{const.}'''\sigma^{p'} h_\sigma^{*\,-p'/2}, \end{aligned}$$

where const.$'''$ depends only on D_1, p', the grid factor a and the kernel K. This shows (5.2). $\square$

PROOF OF THEOREM 3.1. Choose $f \in B(M, L)$ [see (3.2)]. Because we consider risks over classes of uniformly bounded functions, $h_\sigma(t, f) > \sigma^2$ is well defined for σ small enough. Proposition 3.4 implies, for the risk $R_\sigma(f) = E_f \int_0^1 |\hat{f}(t) - f(t)|^{p'}\, dt$,

$$R_\sigma(f) = \int_0^1 r_\sigma(t, f)\, dt \le \text{const.} \int_0^1 [\phi_\sigma(h_\sigma(t, f))]^{p'}\, dt,$$

where

$$\phi_\sigma(h) = v(h)\lambda(h) = \sigma \|K\| h^{-1/2} \max\left\{1, \sqrt{D_2 \log(h_\sigma^*/h)}\right\}.$$

This can be written as

$$R_\sigma(f) \le \text{const.}\left[\int \phi_\sigma(h_\sigma^*)^{p'} \mathbb{I}(h_\sigma(t, f) = h_\sigma^*)\, dt + \sum_{h \in \mathscr{H}_\sigma} \int_{S_h} \phi_\sigma(h)^{p'}\, dt\right],$$

where $S_h = \{t\colon h_\sigma(t, f) = ha^{-1}\}$. On S_h it holds that

$$\Delta_h(t) \ge D\phi_\sigma(h), \tag{5.7}$$

where $\Delta_h(t) = \sup_{\eta \le h} |f_\eta(t) - f(t)|$. This follows from the definition (3.7) of $h_\sigma(t, f)$ and the monotonicity of $\Delta_h(t)$ and $\phi_\sigma(h)$ in h. Now, using (5.7), one gets, for each number $p(h) \in [0, p']$,

$$\begin{aligned} R_\sigma(f) &\le \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + \sum_{h \in \mathscr{H}_\sigma} \phi_\sigma(h)^{p'-p(h)} \int |\Delta_h(t)|^{p(h)}\, dt\right] \\ &= \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + \sum_{h \in \mathscr{H}_\sigma} \phi_\sigma(h)^{p'-p(h)} \|\Delta_h\|_{p(h)}^{p(h)}\right]. \end{aligned} \tag{5.8}$$

We will show that

$$\sup_{f \in B(M, L)} \sup_{0 \le h \le 1} h^{-s} \|\Delta_h\|_p < +\infty, \tag{5.9}$$

$$\sup_{f \in B(M, L)} \sup_{0 \le h \le 1} h^{-s'} \|\Delta_h\|_{p'} < +\infty \quad \text{if } sp \le \frac{p' - p}{2}, \tag{5.10}$$

where s' is defined as $s' = s - 1/p + 1/p'$.

Before we come to the proof of (5.9) and (5.10), let us show that both these statements imply Theorem 3.1.

We now define a function $p(h)$. We distinguish between the following four cases:

(i) $p \ge p'$;
(ii) $p' > p$ and $sp > (p' - p)/2$;
(iii) $sp = (p' - p)/2$;
(iv) $sp < (p' - p)/(2)$.

For the case of $p \geq p'$ we set $p(h) \equiv p'$. By (5.9) one gets $\|\Delta_h\|_{p'} \leq \|\Delta_h\|_p \leq$ const. h^s. Equation (5.8) gives

$$R_\sigma(f) \leq \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + \sum_{h\in\mathscr{H}_\sigma} h^{sp'}\right] \leq \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + h_\sigma^{*\,sp'}\right].$$

By substituting $h_\sigma^* = \sigma^{2/(2s+1)}$ one obtains the statement of the theorem for this case.

For the case of $p' > p$ and $sp > (p' - p)/2$ we put $p(h) \equiv p$. Then one gets

$$R_\sigma(f) \leq \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + \sum_{h\in\mathscr{H}_\sigma} h^{sp}\phi_\sigma(h)^{p'-p}\right].$$

The definition of the grid $\mathscr{H}_\sigma$ allows us to bound the last sum by

$$\begin{aligned}(\sigma\|K\|)^{p'-p} &\sum_{i=0}^{\infty}\left(h_\sigma^* a^{-i}\right)^{sp-(p'-p)/2}\left(2p'\max\{1, i\log a\}\right)^{(p'-p)/2}\\ &\leq \text{const.}\,\sigma^{p'-p}h_\sigma^{*\,sp-(p'-p)/2}\end{aligned}$$

and hence

$$R_\sigma(f) \leq \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + \sigma^{p'-p}h_\sigma^{*\,sp-(p'-p)/2}\right].$$

As above, the choice $h_\sigma^* = \sigma^{2/(2s+1)}$ leads to the bound const. $\sigma^{2sp'/(2s+1)}$ of Theorem 3.1.

Next we consider the case that $sp = (p' - p)/2$. Here we again take $p(h) \equiv p$ and estimate roughly $\phi_\sigma(h)$ by $\sigma h^{-1/2}\sqrt{\log(1/\sigma)}$. This gives, by (5.8) and (5.9),

$$\begin{aligned}R_\sigma(f) &\leq \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + \sum_{h\in\mathscr{H}_\sigma} h^{sp}\phi_\sigma(h)^{p'-p}\right]\\ &\leq \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + \left(\sigma\sqrt{\log(1/\sigma)}\right)^{p'-p}\sum_{h\in\mathscr{H}_\sigma} h^{sp-(p'-p)/2}\right]\\ &= \text{const.}\left[\phi_\sigma(h_\sigma^*)^{p'} + L_\sigma\left(\sigma\sqrt{\log(1/\sigma)}\right)^{p'-p}\right],\end{aligned}$$

where L_σ is the number of elements in $\mathscr{H}_\sigma$. This yields the assertion, since $L_\sigma \leq (\log\sigma^{-2})/\log a$ holds and since the equality $sp = (p' - p)/2$ implies $r'p' = p' - p$.

It remains to consider the case $sp < (p' - p)/2$. Here we set

$$p(h) = \begin{cases} p, & \text{if } h > h_1(\sigma),\\ p', & \text{if } h < h_1(\sigma),\end{cases}$$

where

$$h_1(\sigma) = \left[\sigma\sqrt{\log(1/\sigma)}\right]^{1/(s-1/p+1/2)}.$$

For $sp \le (p' - p)/2$ we have that $s - 1/p + 1/2 > 0$ (see the remark after Theorem 3.1). Therefore the definition of $h_1(\sigma)$ makes sense.

With this choice we get, from (5.8), (5.9) and (5.10),

$$R_\sigma(f) \le \text{const.}\,[\phi_\sigma(h_\sigma^*)^{p'} + R_1 + R_2],$$

where

$$R_1 = \sum_{\substack{h \ge h_1(\sigma) \\ h \in \mathcal{H}_\sigma}} \|\Delta_h\|_{p(h)}^{p(h)} |\phi_\sigma(h)|^{p'-p(h)} \le \text{const.} \sum_{\substack{h \ge h_1(\sigma) \\ h \in \mathcal{H}_\sigma}} h^{sp} |\phi_\sigma(h)|^{p'-p},$$

$$R_2 = \sum_{\substack{h < h_1(\sigma) \\ h \in \mathcal{H}_\sigma}} \|\Delta_h\|_{p(h)}^{p(h)} |\phi_\sigma(h)|^{p'-p(h)} \le \text{const.} \sum_{\substack{h < h_1(\sigma) \\ h \in \mathcal{H}_\sigma}} h^{s'p'}.$$

The sum R_2 is a geometric series and can be bounded by

$$\begin{aligned} \text{const.}\, h_1(\sigma)^{s'p'} &= \text{const.}\Big[\sigma\sqrt{\log(1/\sigma)}\Big]^{p'(s-1/p+1/p')/(s-1/p+1/2)} \\ &= \text{const.}\Big[\sigma\sqrt{\log(1/\sigma)}\Big]^{p'r'}, \end{aligned}$$

which is exactly of the same order as the right-hand side of (3.3) [for $sp < (p' - p)/2$].

It remains to consider the term R_1. Using the definition of $\phi_\sigma(h)$ we get

$$R_1 \le \text{const.}\Big[\sigma\sqrt{\log(1/\sigma)}\Big]^{p'-p} \sum_{\substack{h > h_1(\sigma) \\ h \in \mathcal{H}_\sigma}} h^{sp-(p'-p)/2}.$$

For the case of $sp - (p' - p)/2 < 0$ this gives

$$\begin{aligned} R_1 &\le \text{const.}\Big[\sigma\sqrt{\log(1/\sigma)}\Big]^{p'-p} h_1(\sigma)^{sp-(p'-p)/2} \\ &= \text{const.}\Big[\sigma\sqrt{\log(1/\sigma)}\Big]^{p'(s-1/p+1/p')/(s-1/p+1/2)} \\ &= \text{const.}\Big[\sigma\sqrt{\log(1/\sigma)}\Big]^{p'r'}. \end{aligned}$$

We come now to the proofs of (5.9) and (5.10).

PROOF OF (5.9). For $sp > 1$ the definition (3.1) of local oscillations with $u = +\infty$ implies that for $0 \le t \le 1$ and for each $\varepsilon > 0$ there exists a polynomial $P_{t,h}$ of degree k with

$$\sup_{|x-t| \le h} |f(x) - P_{t,h}(x)| \le \operatorname{osc} f(t, h) + \varepsilon.$$

This implies

$$\sup_{|x-t| \le h} |f(x) - f(t) - P_{t,h}(x) + P_{t,h}(t)| \le 2 \operatorname{osc} f(t, h) + 2\varepsilon.$$

Since K has k vanishing moments, we obtain $\Delta_h(t) \le \text{const.} \operatorname{osc} f(t, h)$. This shows (5.9).

For $sp \le 1$ and $q = +\infty$ we apply the definition (3.1) of local oscillations with $u = 1$. Arguing similarly to above we obtain

$$|f_h(t) - f_{h/2}(t)| \le \text{const.} \operatorname{osc} f(t, h).$$

Because of $\|f_\eta - f\|_p \to 0$ (for $\eta \to 0$) it holds that

$$\|f_h - f\|_p \le \sum_{i \ge 0} \|f_{2^{-i}h} - f_{2^{-i-1}h}\|_p.$$

Now $h^{-s}\| \operatorname{osc} f(t, h)\|_p \le \text{const.}$ provides

$$\|f_h - f\|_p \le h^s \sum_{i \ge 0} \text{const.}\, 2^{-is} \le \text{const.}\, h^s.$$

This shows (5.9).

For $sp \le 1$ and $q < +\infty$ we recall that K can be decomposed as

$$K(x) = 2M(x) - \frac{1}{2} M\left(\frac{x}{2}\right).$$

Now

$$\begin{aligned} f_h(x) - f(x) &= \int M(t)[2f(x + ht) - f(x + 2ht) - f(x)]\, dt \\ &\le \text{const.} \int_{|t| \le 1} |2f(x + ht) - f(x + 2ht) - f(x)|\, dt. \end{aligned}$$

Equation (5.9) follows by application of Theorem 3.5.3 in Triebel (1992) and by using the embedding $B^s_{p,q} \subset B^s_{p,\infty}$. □

PROOF OF (5.10). For $p' \ge p$ the Besov space $B^{s'}_{p',q}$ can be embedded into $B^s_{p,q}$ for all $q \ge 1$ [see Triebel (1992)]. This means that

$$\sup_{f \in B^s_{p,q}(M)} \|f\|_{B^{s'}_{p',q}} < +\infty.$$

Note also that $s'p' < 1$, $s'p' = 1$ or $s'p' > 1$, if and only if $sp < 1$, $sp = 1$ and $sp > 1$, respectively. Thus, (5.10) can be shown by the same arguments as (5.9). □

PROOF OF THEOREM 3.2. We proceed similarly to the proof of Theorem 3.1. The first part of this proof is independent of the choice of h^*_σ and the bound (5.8) remains valid. If $sp \le (p' - p)/2$ holds, the risk $R_\sigma(f)$ can be treated as in the proof of Theorem 3.1. For the case of $sp > (p' - p)/2$ another definition of $p(h)$ will be used for estimating $R_\sigma(f)$. The following choice will do:

$$p(h) = \begin{cases} 0, & \text{if } h > \left[\sigma\sqrt{\log(1/\sigma)}\right]^{2/(2s+1)}, \\ \min\{p, p'\}, & \text{if } h \le \left[\sigma\sqrt{\log(1/\sigma)}\right]^{2/(2s+1)}. \end{cases}$$

□

PROOF OF THEOREM 3.3. Let $t \in [0,1]$, let f be fixed and let $h_\sigma = h_\sigma(t,f)$ be defined by (3.7). We get from Proposition 3.4 with $h_\sigma^* = 1$ that

$$r_\sigma(t,f) = E_f|\hat{f}(t) - f(t)|^2 \le \text{const.}\,\sigma^2\|K\|^2 h_\sigma^{-1}\log h_\sigma^{-1} \le \text{const.}\,\big(\sigma^2\log\sigma^{-1}\big)\|K\|^2 h_\sigma^{-1}. \tag{5.11}$$

Now we have to show that $r_{\text{adapt}}(t,f) \ge \text{const.}(\sigma^2\log\sigma^{-1})\|K\|^2 h_\sigma^{-1}$. Recall that

$$r_{\text{adapt}}(t,f) = \inf_{a\sigma^2 \le h \le 1}\{\Delta_h^2(t) + \big(\sigma^2\log\sigma^{-1}\big)\|K\|^2 h^{-1}\},$$

where $\Delta_h(t) = \sup_{\eta \le h}|f_\eta(t) - f(t)|$. Suppose that the infimum is attained at h_0. We now treat the cases of $h_0 \ge ah_\sigma$ and $h_0 < ah_\sigma$ separately. Note that $\Delta_h(t)$ is monotonely increasing in h. Consider first the case with $h_0 \ge ah_\sigma$. Then we have immediately from (5.11) that $r_\sigma(t,f) \le \text{const.}\, r_{\text{adapt}}(t,f)$.

If $h_0 < ah_\sigma$, then the definition (3.7) of h_σ gives $\Delta_{ah_\sigma}^2(t) > D_1^2\sigma^2(ah_\sigma)^{-1} \times \log(ah_\sigma)^{-1}$. Because of (5.11) this implies

$$r_{\text{adapt}}(t,f) \ge \Delta_{h_0}^2(t) \ge \Delta_{ah_\sigma}^2(t) > \frac{D_1^2\sigma^2}{ah_\sigma}\log(ah_\sigma)^{-1} \ge \text{const.}\, r_\sigma(t,f).$$

This is the statement of Theorem 3.3. □

REFERENCES

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1983). *Classification and Regression Trees*. Wadsworth, Belmont, CA.

BROCKMANN, M., GASSER, T. and HERRMANN, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *J. Amer. Statist. Assoc.* **88** 1302–1309.

BROWN, L. D. and LOW, M. G. (1992). Superefficiency and lack of adaptability in functional estimation. Preprint.

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

DELYON, B. and JUDITSKY, A. (1994). Wavelet estimators, global error measures revisited. Technical report, IRISA, Rennes.

DONOHO, D. L. and JOHNSTONE, I. M. (1996). Minimax estimation via wavelet shrinkage. Unpublished manuscript.

DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia? (with discussion). *J. Roy. Statist. Soc. Ser. B* **57** 301–369.

FAN, J. and GIJBELS, I. (1995). Data-driven bandwidth selection in local polynomial fitting: variable bandwidth and spatial adaptation. *J. Roy. Statist. Soc. Ser. B* **57** 371–394.

FAN, J., HALL, P., MARTIN, M. and PATIL, P. (1993). Adaptation to high spatial inhomogeneity based on wavelets and on local linear smoothing. Preprint.

FRIEDMAN, J. H. (1991). Multivariate adaptive regression splines (with discussion). *Ann. Statist.* **19** 1–41.

FRIEDMAN, J. H. and SILVERMAN, B. W. (1989). Flexible parsimonious smoothing and additive modelling (with discussion). *Technometrics* **31** 3–39.

GIJBELS, I. and MAMMEN, E. (1994). On local adaptivity of kernel estimates with plug-in local bandwidth selectors. Preprint, Sonderforschungsbereich 373, Humboldt Univ., Berlin.

GOLDENSHLUGER, A. and NEMIROVSKI, A. (1994). On spatial adaptive estimation of nonparametric regression. Research report, Technion–Israel Inst. Technology, Haifa, Israel.

GOLDENSHLUGER, A. and NEMIROVSKI, A. (1996). On spatial adaptive nonparametric estimation of functions satisfying differential inequalities. Research report, Technion–Israel Inst. Technology, Haifa, Israel.

KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by kernel and wavelet method, optimality in Besov space. *Statist. Probab. Lett.* **18** 327–336.

LEPSKI, O. V. (1990). One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 459–470

LEPSKI, O. V. (1991). Asymptotic minimax adaptive estimation. 1. Upper bounds. *Theory Probab. Appl.* **36** 654–659.

LEPSKI, O. V. (1992). Asymptotic minimax adaptive estimation. 2. Statistical model without optimal adaptation. Adaptive estimators. *Theory Probab. Appl.* **37** 468–481.

LEPSKI, O. V. and SPOKOINY, V. G. (1994). Local adaptivity to inhomogeneous smoothness. Resolution level. *Math. Methods Statist.* **4** 239–258.

LEPSKI, O. V. and SPOKOINY, V. G. (1996). Optimal pointwise adaptive methods in nonparametric estimation. Unpublished manuscript.

LOW, M. G. (1992). Renormalization and white noise approximation for nonparametric functional estimation. *Ann. Statist.* **20** 545–554.

LUO, Z. and WAHBA, G. (1995). Hybrid adaptive splines. Preprint.

MAMMEN, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.

MAMMEN, E. and VAN DE GEER, S. (1997). Locally adaptive regression splines. *Ann. Statist.* **25** 387–413.

MÜLLER, H.-G. and STADTMÜLLER, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182–201.

NEMIROVSKI, A. (1985). On nonparametric estimation of smooth regression functions. *Soviet J. Comput. Syst. Sci.* **23** 1–11.

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

STANISWALIS, J. G. S. (1989). Local bandwidth selection for kernel estimates. *J. Amer. Statist. Assoc.* **84** 284–288.

TRIEBEL, H. (1992). *Theory of Function Spaces II.* Birkhäuser, Basel.

O. V. LEPSKI
INSTITUTE FOR SYSTEM ANALYSIS
PROSPEKT 60-LET OKTJABRJA, 9
MOSCOW, 117312
RUSSIA
AND
SONDERFORSCHUNGSBEREICH 373
HUMBOLDT-UNIVERSITÄT
SPANDAUER STRASSE 1
10178 BERLIN
GERMANY

E. MAMMEN
INSTITUT FÜR ANGEWANDTE MATHEMATIK
UNIVERSITÄT HEIDELBERG
IM NEUENHEIMER FELD 294
69120 HEIDELBERG
GERMANY
E-MAIL: mammen@statlab.uni-heidelberg.de

V. G. SPOKOINY
INSTITUTE FOR INFORMATION TRANSMISSION PROBLEMS
ERMOLOVOY 19
MOSCOW, 101447
RUSSIA
AND
INSTITUTE OF APPLIED ANALYSIS AND STOCHASTICS
MOHRENSTRASSE 39
10117, BERLIN
GERMANY