The Annals of Statistics
1997, Vol. 25, No. 6, 2512–2546

OPTIMAL POINTWISE ADAPTIVE METHODS IN NONPARAMETRIC ESTIMATION[1]

BY O. V. LEPSKI AND V. G. SPOKOINY

Humboldt University and Weierstrass Institute for Applied Analysis and Stochastics

The problem of optimal adaptive estimation of a function at a given point from noisy data is considered. Two procedures are proved to be asymptotically optimal for different settings.

First we study the problem of bandwidth selection for nonparametric pointwise kernel estimation with a given kernel. We propose a bandwidth selection procedure and prove its optimality in the asymptotic sense. Moreover, this optimality is stated not only among kernel estimators with a variable bandwidth. The resulting estimator is asymptotically optimal among all feasible estimators. The important feature of this procedure is that it is fully adaptive and it "works" for a very wide class of functions obeying a mild regularity restriction. With it the attainable accuracy of estimation depends on the function itself and is expressed in terms of the "ideal adaptive bandwidth" corresponding to this function and a given kernel.

The second procedure can be considered as a specialization of the first one under the qualitative assumption that the function to be estimated belongs to some Hölder class $\Sigma(\beta, L)$ with unknown parameters β, L. This assumption allows us to choose a family of kernels in an optimal way and the resulting procedure appears to be asymptotically optimal in the adaptive sense in any range of adaptation with $\beta \leq 2$.

1. Introduction. The present paper is devoted to studying the problem of adaptive estimation of a function at a given point. In the context of function estimation, this problem can be treated as the problem of "local" or pointwise data-driven selection of smoothing parameters.

The inspection of the existing literature in this direction shows two different possible asymptotic approaches. First we mention the results with fixed-parameter asymptotics when an estimated function is kept fixed and the number of observations goes to infinity. In this setup the corresponding smoothing parameter (typically bandwidth) can be chosen root-n consistently; see Härdle and Marron (1985), Müller and Stadtmüller (1987), Staniswalis (1989), Jones, Marron and Park (1991), Vieu (1991), Hall and Johnstone (1992) and Brockmann, Gasser and Herrmann (1993) among others.

We also discuss another approach, usually called minimax, which is based on the assumption that the function to be estimated from noisy data belongs to some function (smoothness) class, for example, Hölder, Sobolev, Besov, and so on [see Ibragimov and Khasminskii (1980, 1981), Bretagnolle and Huber

Received February 1995; revised April 1996.
[1]Supported in part by INTAS Grant N93/1585.
AMS 1991 *subject classifications*. Primary 62G07; secondary 62G20.
Key words and phrases. Bandwidth selection, Hölder-type constraints, pointwise adaptive estimation.

(1979), Stone (1982), Nemirovskii (1985), Donoho and Johnstone (1992) and Kerkyacharian and Picard (1993)]. Such an assumption is very important because the rate (accuracy) of estimation and the corresponding optimal estimation rule depend on the structure and parameters of this function class. But, at the same time, this creates a serious problem for application of the nonparametric approach because typically we do not have a priori any information about the smoothness properties of the estimated function. To bypass this trouble, one or another kind of adaptive procedure is applied. It is assumed again that the function belongs to some function class but with unknown parameter values. After that the corresponding smoothness parameters (e.g., a bandwidth) of the estimation procedure are chosen automatically by data. We refer to Marron (1988) and Donoho and Johnstone (1995) for an overview on this topic. Note only that typical results on adaptive estimation deal with the problem of estimation of a function.

In the following discussion we shall focus on the problem of estimation of a function at a given point. Study of the problem of pointwise adaptive estimation in the minimax framework was initiated by Lepski (1990, 1992). In those papers a significant difference was shown between the problem of estimation of the whole function and of a value of a function at one point. More precisely, for the problem of pointwise estimation we encounter the phenomenon of lack of adaptability: if we knew that a function to be estimated belongs to a given Hölder class $\Sigma(\beta, L)$, then we would estimate this function at a given point with the accuracy $\varphi(\varepsilon) = \varepsilon^{2\beta/(2\beta+1)}$, ε being the noise level. But if the parameter β is unknown, then this accuracy is impossible to attain. The optimal adaptive rate was also calculated in Lepski (1990). It turned out to be $(\varepsilon\sqrt{\ln \varepsilon^{-1}})^{2\beta/(2\beta+1)}$ which differs from the nonadaptive one by the extra log-factor [see also Brown and Low (1992)].

Recently the problem of pointwise adaptive estimation has received a new impetus in connection with the problem of global function estimation for Besov classes. It was shown in Lepski, Mammen and Spokoiny (1997) that a kernel estimator with a variable data-driven bandwidth based on pointwise adaptation achieves the minimax rate of estimation over a wide scale of Besov classes and hence this kind of estimator shares rate optimality properties with wavelet estimators; see, for example, Donoho and Johnstone (1994, 1992), Kerkyacharian and Picard (1993) and Donoho et al. (1995).

In the present paper we continue studying the problem of pointwise adaptive estimation. The aim is to describe an asymptotically optimal (at the level of a constant) pointwise adaptive procedure.

Below we consider two settings for which an optimal (in the asymptotic sense) pointwise adaptive procedure can be shown explicitly. The first approach can be described as follows. Let a function $f(\cdot)$ be observed with noise and let us estimate the value of this function at a point t_0. We study the problem of an adaptive bandwidth selection in kernel estimation with a given kernel K. For pointwise adaptation we use the adaptive procedure from Lepski, Mammen and Spokoiny (1997) with a more accurate choice of its parameters. We prove that this estimation procedure is sharp optimal in the adaptive sense

over the class of all feasible estimators not only of kernel type. This kind of result is a little bit surprising since we know from Sacks and Strawderman (1982) that for nonadaptive pointwise estimation linear (and, in particular, kernel) methods are not sharp optimal.

The first setting assumes that the kernel K is given and only the bandwidth h is to be selected. The other approach is in a simultaneous selection of a kernel and a bandwidth. We consider this problem under the qualitative assumption that the function f belongs to some Hölder class with unknown parameters.

The paper is organized as follows. In the next section we formulate the problem of optimal bandwidth selection and present the related results. We also briefly discuss possible applications of these results to the problem of global function estimation. In Section 3 we consider the problem of optimal pointwise adaptive estimation under Hölder-type constraints on the estimated function. Some possible developments of the presented results are discussed in Section 4. The proofs are mostly deferred to Section 5.

2. Optimal bandwidth selection. In this section we consider the problem of data-driven bandwidth selection for a given kernel K. We propose a pointwise selection rule and show that the resulting estimator is optimal (asymptotically when the noise level goes to zero) among the class of all feasible estimators not only of kernel type.

2.1. *Model and kernel smoothers.* We consider the simplest "signal + white noise" model when an observed process $X(t)$, $t \in [0,1]$, obeys the following stochastic differential equation:

$$dX(t) = f(t)\,dt + \varepsilon\,dW(t). \tag{2.1}$$

Here ε is the level of noise and we assume that this level is "small"; that is, we consider the asymptotics as $\varepsilon \to 0$. The process $W = (W(t),\ t \geq 0)$ is a standard Wiener process. The function $f(\cdot)$ in (2.1) is to be estimated at a point $t_0 \in (0,1)$.

Let $K(\cdot)$, be a kernel, that is, a function satisfying the usual assumptions [see the following conditions $(K1)$–$(K5)$]. Consider the family of the kernel estimators $\tilde{f}_h(t_0)$ of the value $f(t_0)$:

$$\tilde{f}_h(t) = \frac{1}{h}\int K\left(\frac{t-t_0}{h}\right) dX(t), \tag{2.2}$$

with a positive bandwidth h. Furthermore, we assume that h is small enough and the support of the function $K((t-t_0)/h)$ is contained in $[0,1]$. This assumption allows us to avoid the boundary problem and to change integration over $[0,1]$ to integration over the whole real line. That is why we omit the integration limits here in the definition (2.2) and in what follows.

The problem is to select by the data X some bandwidth $\hat{h}$ to minimize the corresponding risk

$$E_f\left|\tilde{f}_{\hat{h}}(t_0) - f(t_0)\right|^r, \tag{2.3}$$

where $r \geq 1$ is a given power. The exact statement of the problem will be given later on. We start with a preliminary discussion.

2.2. *Preliminaries.* Denote, for $h > 0$,

$$\mathscr{K}_h f(t_0) = \frac{1}{h} \int K\left(\frac{t - t_0}{h}\right) f(t)\, dt.$$

We use the usual decomposition of the loss for the kernel estimators $\tilde{f}_h(t_0)$,

$$\tilde{f}_h(t_0) - f(t_0) = \mathscr{K}_h f(t_0) - f(t_0) + \xi(h) = B(h) + \xi(h), \tag{2.4}$$

where the stochastic term $\xi(h)$ with

$$\xi(h) = \frac{\varepsilon}{h} \int K\left(\frac{t - t_0}{h}\right) dW(t)$$

is obviously a Gaussian zero-mean random variable with variance

$$\sigma^2(h, \varepsilon) = \frac{\varepsilon^2 \|K\|^2}{h}.$$

The standard bandwidth choice in nonparametric estimation is motivated by the balance relation between the bias and stochastic terms in the decomposition (2.4). The bias term $B(h) = \mathscr{K}_h f(t_0) - f(t_0)$ for a bandwidth h is nonrandom but it depends on the function f, $B(h) = B_f(h)$, and it characterizes the accuracy of approximation of an estimated function by the applied method of approximation (in the present context by kernel smoothers). The stochastic term is a Gaussian random variable with zero mean and variance $\sigma^2(h, \varepsilon)$ and it depends typically on the error level ε, the kernel K and the bandwidth h but not on the function f.

Minimization of the losses leads to a balance equation of the form $B_f(h) \asymp \sigma(h, \varepsilon)$, where the symbol "$\asymp$" means the equivalence in order. Indeed, a decrease on the order of $\Delta(h)$ usually results in an increase on the order of $\sigma(h, \varepsilon)$ and vice versa, and such a balance relation is necessary for obtaining the optimal pointwise rate. But the function f is unknown and hence the bias function $B_f(h)$ is also unknown. One standard approach used here is based on the smoothness assumption that the function f belongs to some function class, for instance, to Hölder or Sobolev ball $\Sigma(\beta, L)$ with smoothness parameters β, L; see, for example, Ibragimov and Khasminskii (1981). Under such a constraint, one may estimate the bias $B_f(h)$ by Const. Lh^β and we arrive at the standard balance equation

$$Lh^\beta \asymp \sigma(h, \varepsilon) = \varepsilon \|K\| h^{-1/2}. \tag{2.5}$$

The approach proposed later develops this idea in the following sense. We try to adapt the estimation procedure and particularly the bandwidth selector rule not to some function class but to the function f itself. Of course, the quality of estimation still depends on some smoothness (or regularity) properties of this function. As soon as the kernel K is fixed (this means that the method of approximation of the function by its kernel smoothers is fixed), we measure

these regularity properties of the function f at the point t_0 by the bias function $B_f(h)$. [Note that the method of describing the smoothness properties of a function in terms of the rate of approximation by kernel smoothers is one of the standard approaches in approximation theory. For instance, if a kernel is of a proper regularity, then Sobolev or Besov smoothness function classes can be defined in these terms; see, e.g., Triebel (1992).]

Denote, for $h > 0$,

$$\Delta(h) = \Delta_f(h) = \sup_{0<\eta<h} \left|\mathscr{K}_\eta f(t_0) - f(t_0)\right|. \tag{2.6}$$

An adaptive procedure which would realize the bandwidth selection rule due to the following balance equation:

$$\Delta(h) \asymp \sigma(h, \varepsilon) \tag{2.7}$$

could be called "ideal" since such a procedure would adapt directly to local (pointwise) smoothness properties of the unknown function.

Unfortunately, such a balance equation [and even the classical balance equation (2.5)] cannot be realized for the problem of pointwise adaptive estimation. This phenomenon was discovered by Lepski (1990); see also Brown and Low (1992). The idea behind this is that the loss of a minimax estimator, being normalized, will not be asymptotically degenerate. [Recall that for the problem of global function estimation losses are typically degenerate; see, e.g., Lepski (1991).]

It turned out that, in order to handle an adaptive procedure in the case of an estimation at a point, one has to take some majorant for the stochastic term to control stochastic fluctuations. Namely, the balance relation

$$\Delta_f(h) \asymp \sigma(h, \varepsilon)\sqrt{\ln \varepsilon^{-1}} = \frac{\varepsilon\sqrt{\ln \varepsilon^{-1}}}{\sqrt{h}}$$

allows one to estimate adaptively but the corresponding rate also includes such a log-factor. One can say that this extra log-factor is an unavoidable payment for pointwise adaptation which can be neither removed nor improved (in the sense of rate of convergence).

This phenomenon also admits the following interpretation. An adaptive estimation means that we have to estimate not only the unknown value $f(t_0)$ of the function f at t_0 but also the underlying smoothing parameter which leads to some loss of efficiency. One may characterize this loss of efficiency as a noise magnification with some factor d_ε which is treated as a payment for adaptation. Now we denote $\tilde{\varepsilon} = \varepsilon d_\varepsilon$ and we are trying to realize the same balance equation (2.5) or (2.7) with $\tilde{\varepsilon}$ in place of ε, that is,

$$\Delta(h) \asymp \sigma(h, \tilde{\varepsilon}). \tag{2.8}$$

The minimal (in order) value of d_ε for which this can be done could be called "an adaptive factor." In the context of estimation over Hölder classes, due to Lepski (1990), this factor was found to be $\sqrt{\ln \varepsilon^{-1}}$.

We shall now describe the similar result in the context of kernel estimation with a fixed kernel. First we indicate the corresponding adaptive factor which appeared to be $\sqrt{r \ln(h_{\max}/h_{\min})}$, where $[h_{\min}, h_{\max}]$ is the range of adaptation. Then we show that the presence of this log-factor leads to degenerate behavior of the corresponding normalized losses that allows us to optimize the balance equation (2.8) in the following sense: we rewrite it in the form

$$\Delta(h) = C\sigma(h, \tilde{\varepsilon}) \tag{2.9}$$

and try to select the constant C in an optimal way in order to get the minimal value of the risk in (2.3). After the constant C has been specified, the bandwidth h_f which is the solution of this balance equation might be called an "ideal adaptive bandwidth" corresponding to the given kernel and to local regularity properties of an estimated function expressed by the bias function $B_f(h)$. In the following discussion we propose an expression for the constant $C = C(K)$, present an adaptive bandwidth selector and show that the resulting estimator performs in a way as if the value h_f were known. Moreover, we show that this estimator is optimal among the class of all feasible estimators, not only of kernel type.

Of course, the resulting notion of optimality, particularly the notion of the "ideal adaptive bandwidth," the corresponding accuracy of estimation and the corresponding adaptive procedure depend on the kernel K. This dependence is natural and in some sense unavoidable. Taking another kernel, we will get another procedure and another accuracy of estimation. But this dependence is not crucial from the point of view of rate of estimation. If the kernel K is of proper regularity, then the proposed procedure achieves the usual minimax rate for all standard smoothness classes; see Remark 2.7. Moreover, considering the problem of adaptive estimation at a point under Hölder-type constraints, a kernel (more precisely, a family of kernels) can be chosen in such a way that the resulting procedure becomes asymptotically optimal in the classical minimax sense over Hölder function classes; see Section 3.

2.3. *Kernel.* Now let a fixed kernel $K(\cdot)$ satisfy the following conditions:

($K1$) The function $K(u)$ is symmetric, that is, $K(u) = K(-u)$, $u \geq 0$.
($K2$) The function $K(\cdot)$ is compactly supported, that is, $K(u) = 0$ for all u outside some compact set C on the real line.
($K3$) $\int K(u)\,du = 1$.
($K4$) $\|K\|^2 = \int K^2(u)\,du < \infty$.
($K5$) The function $K(\cdot)$ is continuous at the point $t = 0$ and $K(0) > \|K\|^2$.

Note that no assumptions were made about the smoothness properties of the kernel K; that is, it can be even discontinuous.

2.4. *Bandwidth selection problem: "ideal adaptive bandwidth."* Now we make precise the problem of bandwidth selection and define the notion of an "ideal adaptive bandwidth" for the function f.

We assume that, besides the kernel K, two values $h_{\min,\varepsilon}$ and $h_{\max,\varepsilon}$ are given for each ε such that $h_{\min,\varepsilon} > \varepsilon^2$, $h_{\max,\varepsilon} \le 1$ and

$$h_{\min,\varepsilon}/h_{\max,\varepsilon} \to 0, \qquad \varepsilon \to 0. \tag{2.10}$$

For notational simplicity, we will omit the subindex ε in $h_{\min,\varepsilon}$ and $h_{\max,\varepsilon}$.

We will select a bandwidth h in the interval $h \in [h_{\min}, h_{\max}]$; that is, $h_{\min}$ is the smallest and $h_{\max}$ is the largest admissible value of the bandwidth.

Denote

$$d_\varepsilon = \sqrt{r \ln \frac{h_{\max}}{h_{\min}}}. \tag{2.11}$$

By (2.10) one has $d_\varepsilon \to \infty$ as $\varepsilon \to 0$.

First we show that "the payment for adaptation" or "the adaptive factor" could not be less in order than d_ε.

THEOREM 2.1. *Let d'_ε be such that $d'_\varepsilon/d_\varepsilon \to 0$ as $\varepsilon \to 0$ and let $\varepsilon' = \varepsilon d'_\varepsilon$. Then for any $C, C' > 0$ and ε small enough there exist two functions f_0 and f_1 such that*

$$\Delta_{f_0}(h_{\max}) \le C\sigma(h_{\max}, \varepsilon'),$$

$$\Delta_{f_1}(h_{\min}) \le C\sigma(h_{\min}, \varepsilon')$$

and, for any estimator T_ε of $f(t_0)$,

$$\max\left\{ E_{f_0}\left|\frac{T_\varepsilon - f_0(t_0)}{\sigma(h_{\max}, \varepsilon')}\right|^r, E_{f_1}\left|\frac{T_\varepsilon - f_1(t_0)}{\sigma(h_{\min}, \varepsilon')}\right|^r \right\} > C'.$$

The assertion of this theorem claims that if d'_ε is less in order than d_ε, then the balance rule (2.8) with ε' in place of $\tilde\varepsilon$ "does not work" in the sense that the corresponding normalized risk tends to infinity.

Furthermore, we explore the case with the adaptive factor d_ε from (2.11). Note that this factor enters automatically in the expression of the minimax rate of convergence and the less is the range $[h_{\min}, h_{\max}]$ the less is the payment for adaptation. In any case d_ε is not larger (in order) than $\sqrt{\ln \varepsilon^{-1}}$ and this is the typical order.

Now we define the notion of an "ideal adaptive bandwidth" which refers to the balance equation (2.9) with the optimal choice of a constant. Set

$$C(K) = \frac{K(0)}{\|K\|^2} - 1. \tag{2.12}$$

DEFINITION 2.1. Let d_ε be defined by (2.11), $\tilde\varepsilon = \varepsilon d_\varepsilon$ and $\sigma(h, \tilde\varepsilon) = \|K\|\tilde\varepsilon h^{-1/2}$. Let also, given $f(\cdot)$, the function $\Delta(h) = \Delta_f(h)$ be defined by (2.6). The value h_f^* with

$$h_f^* = \sup\{h \le h_{\max}: \Delta(h) \le C(K)\sigma(h, \tilde\varepsilon)\}$$

is called the "ideal adaptive bandwidth" for the function f.

REMARK 2.1. The function $\Delta(h)$ is, by definition, monotonely increasing and the function $\sigma(h, \tilde{\varepsilon})$ is, on the contrary, monotonely decreasing with $\sigma(h, \tilde{\varepsilon}) \to \infty$ as $h \downarrow 0$. This provides that the set $\{h\colon \Delta(h) \le C(K)\sigma(h, \tilde{\varepsilon})\}$ is nonempty and h_f^* is uniquely defined. Moreover, the function $\sigma(h, \varepsilon)$ is continuous and if the function $\Delta(h)$ is also continuous, then we arrive at the equation of the form (2.9) for the "ideal adaptive bandwidth." But, generally speaking, the function $\Delta(h)$ is right-continuous and hence one cannot guarantee that the equation (2.9) has a root.

The choice of a constant $C(K)$ is motivated by the following optimality results. In fact, we will propose an estimator which is optimal (in the adaptive sense) among all feasible estimators and which provides just the accuracy corresponding to this balance equation.

REMARK 2.2. The definition of the "ideal adaptive bandwidth" given here depends on the kernel K, the noise level ε and the function f which is quite natural. But also this notion depends essentially on the range of adaptation $[h_{\min}, h_{\max}]$. This is one more reason why we prefer to speak about an "ideal adaptive bandwidth" rather than about an "ideal bandwidth."

Now we are ready to describe an adaptive bandwidth selector and then formulate the main results.

2.5. *A bandwidth selector.* Denote

$$d(h) = \sqrt{1 \vee [p \ln(h_{\max}/h)]},$$
$$\alpha(h) = \frac{1}{\sqrt{d(h)}} \tag{2.13}$$

and define the grid $\mathscr{H}$ inductively by

$$\mathscr{H} = \left\{ h \in [h_{\min}, h_{\max}]\colon h_0 = h_{\max}, h_{k+1} = \frac{h_k}{1 + \alpha(h_k)},\ k = 0, 1, 2, \ldots \right\}. \tag{2.14}$$

Now set

$$\hat{h} = \max\{h \in \mathscr{H}\colon |\tilde{f}_h(t_0) - \tilde{f}_\eta(t_0)| \le (1 + \alpha(\eta))\sigma(\eta, \varepsilon)d(\eta)\ \forall\ \eta < h,\ \eta \in \mathscr{H}\}.$$

Here $\tilde{f}_h(t_0)$ is the usual kernel estimator defined by (2.2). Finally, set

$$\hat{f}_\varepsilon(t_0) = \tilde{f}_{\hat{h}}(t_0).$$

REMARK 2.3. The proposed adaptive procedure is based on a comparison of kernel estimators with different bandwidths from the grid $\mathscr{H}$ which is of geometric-type structure. We will see that the total number of elements in $\mathscr{H}$ and hence the total number of compared bandwidths is of logarithmic order.

The bandwidth $\hat{h}$ can be treated in the following way. It is the largest bandwidth h such that $\tilde{f}_h$ does not differ "significantly" from kernel estimators with smaller bandwidth.

REMARK 2.4. It is easy to observe that the whole construction of the proposed bandwidth selector requires knowledge of the noise level ε, the kernel K, the norm degree r and the range of adaptation $[h_{\min}, h_{\max}]$. For practical applications only the exact knowledge of the noise level ε is somewhat problematic. But this difficulty can be handled in the usual way by using a pilot estimator of the noise level. Note also that the value $h_{\min}$ is used only for the definition of the grid $\mathscr{H}$. This fact is rather important since it allows us to apply this adaptive rule for any adaptive range with the fixed upper value $h_{\max}$ (and with the smallest $h_{\min}$). Particularly, if one admits $h_{\max} = 1$, then one gets a uniform adaptive procedure. The question of a reasonable choice of the parameter $h_{\max}$ is also discussed later in this section in the context of global function estimation.

The following results claim asymptotic optimality of this bandwidth selector $\hat{h}$.

2.6. *Main results.* Let, given a function f, the corresponding smoothness characteristic (the "optimal adaptive bandwidth") h_f^* be defined by Definition 2.1. Denote by $\mathscr{F}_\varepsilon$ the class of functions $f(\cdot)$ with $h_f^* \geq h_{\min}$:

$$\mathscr{F}_\varepsilon = \{f(\cdot)\colon h_f^* \geq h_{\min}\}.$$

We shall assume that the estimated function f belongs to $\mathscr{F}_\varepsilon$. The meaning of this assumption is clear. If we start the adaptive procedure from the bandwidth $h_{\min}$, then we have to be sure that the regularity of the function f described by the value h_f^* is not less than $h_{\min}$. Note also that the constraint of the sort $f \in \mathscr{F}_\varepsilon$ is rather mild. If, for instance, $h_{\min} \asymp \varepsilon^2$, then any function with locally (around t_0) bounded variation belongs to $\mathscr{F}_\varepsilon$.

Now we formulate the main results. The first result describes the accuracy which is attained by the proposed estimator $\hat{f}_\varepsilon$.

THEOREM 2.2. *Let $K(\cdot)$ be a kernel satisfying conditions $(K1)$–$(K5)$ and also the following condition*:

$$(K6) \qquad \sup_{0<c\leq 1} \int \left|K(u) - cK(cu)\right|^2 du = \int K^2(u)\, du.$$

Then the estimator $\hat{f}_\varepsilon(t_0)$ which is a kernel estimator with an adaptive bandwidth $\hat{h}$ provides

$$\sup_{f\in\mathscr{F}_\varepsilon} E_f \left| \frac{\hat{f}_\varepsilon(t_0) - f(t_0)}{\sigma(h_f^*, \tilde{\varepsilon})} \right|^r = (C(K)+1)^r (1 + o_\varepsilon(1)).$$

REMARK 2.5. Here and in what follows we denote by $o_\varepsilon(1)$ any sequence depending possibly on ε, r and the kernel K but not on a function f and such that

$$o_\varepsilon(1) \to 0, \qquad \varepsilon \to 0.$$

The next result shows that the performance of the estimator $\hat{f}_\varepsilon$ cannot be improved in the minimax sense, that is, this estimator is asymptotically efficient.

THEOREM 2.3. *Let a kernel $K(\cdot)$ satisfy conditions $(K1)$–$(K5)$ and also the condition*

$$(K7) \qquad \inf_{0<c\le 1} \int K(u)K(cu)\,du = \int K^2(u)\,du.$$

Then for each $\varepsilon > 0$ there exist two functions $f_0(\cdot)$ and $f_1(\cdot)$ (depending on ε) such that $h^*_{f_0} = h_{\max}$, $h^*_{f_1} \ge h_{\min}$ and, for any estimator T_ε,

$$\max_{f\in\{f_0,f_1\}} E_f \left| \frac{T_\varepsilon - f(t_0)}{\sigma(h^*_f, \tilde{\varepsilon})} \right|^r \ge (C(K)+1)^r(1-o_\varepsilon(1)).$$

The scope of the results of Theorems 2.2 and 2.3 claims the asymptotic optimality of the estimator $\hat{f}_\varepsilon(t_0)$ if the kernel K satisfies conditions $(K1)$–$(K7)$. The question for which kernels these conditions are fulfilled is discussed later in this section.

REMARK 2.6. The first result states the properties of the estimator $\hat{f}_\varepsilon(t_0)$ which are uniform over the very wide function class $\mathscr{F}_\varepsilon$, whereas the lower bounds result from Theorem 2.3 is stated for the class consisting of two functions. Moreover, we will use $f_0(t) \equiv 0$ and only f_1 depends on ε.

REMARK 2.7. It is of interest to observe which accuracy of estimation provides the estimator $\hat{f}_\varepsilon(t_0)$ from Theorem 2.2 in the usual sense.

Let the function f to be estimated belong to some Hölder class $\Sigma(\beta, L)$ (for the precise definition, see Section 3). Let m be the maximal integer which is less than β. If the kernel K has the regularity m, that is, K is orthogonal to polynomials $t, \dots, t^m$, then one easily has

$$\Delta(h) = \Delta_f(h) = \sup_{\eta\le h} \left| \mathscr{K}_\eta f(t_0) - f(t_0) \right| \le CLh^\beta,$$

with some constant C depending only on β and $K(\cdot)$. Now the balance equation $\Delta(h^*_f) = C(K)\sigma(h^*_f, \tilde{\varepsilon}) \le C\tilde{\varepsilon}(h^*_f)^{-1/2}$ leads to the asymptotic relation $h^*_f \ge C\tilde{\varepsilon}^{2/(2\beta+1)}$ and hence

$$\sigma(h^*_f, \tilde{\varepsilon}) \le C\tilde{\varepsilon}^{2\beta/(2\beta+1)} \asymp \left(\varepsilon\sqrt{\ln \varepsilon^{-1}}\right)^{2\beta/(2\beta+1)}.$$

Therefore, and as expected, the result of Theorem 2.1 guarantees the near optimal rate of estimation over Hölder classes. Moreover, an optimal kernel choice provides asymptotically optimal (up to a constant) pointwise-adaptive estimation over the Hölder classes. The discussion of this problem is the subject of the second part of the paper.

Due to Theorem 2.3, the result of Theorem 2.2 cannot be improved in the uniform sense over the class $\mathscr{F}_\varepsilon$; that is, we have to pay for adaptation at least

$d_\varepsilon = (r \ln(h_{\max}/h_{\min}))^{1/2}$. But, considering a single function f, this payment could be possibly brought down. In fact, the lower bound from Theorem 2.3 can be applied to any range of adaptation containing this function and the best result will be obtained for the case of the maximal possible value of $h_{\min}$, that is, as if the characteristic h_f^* of the function f due to Definition 2.1 were exactly $h_{\min}$.

Now we modify Definition 2.1 in this spirit and present the result.

DEFINITION 2.2. Let $C(K)$ be due to (2.12) and let, for a given function f, the functions $\Delta_f(h)$ and $d(h)$ be defined by (2.6) and (2.13), respectively. Define

$$h_f = \sup\bigl\{h \le h_{\max}\colon \Delta_f(h) \le C(K)\sigma(h,\varepsilon)\, d(h)\bigr\}. \tag{2.15}$$

Compared with Definition 2.1, one may observe that the last definition depends only on the upper value $h_{\max}$ of the range of adaptation. Note also that for any $f \in \mathscr{F}_\varepsilon$ one has $h_f \ge h_f^*$ since $d(h) \le d_\varepsilon$ for $h \in [h_{\min}, h_{\max}]$. We will see from the next result that the value $d(h_f)$ can be viewed as the individual payment for adaptation for a particular function f from the range of adaptation.

Before we state the assertion, let us point out one more important question. Theorems 2.2 and 2.3 assume for the range of adaptation $[h_{\min}, h_{\max}]$ that $h_{\max}/h_{\min} \to \infty$ as $\varepsilon \to 0$. Now, speaking about the individual payment for adaptation, we change $h_{\min}$ to h_f^*. But what happens if h_f^* is about $h_{\max}$? The answer to this question is of special importance in view of applications of the pointwise adaptive method to the problem of global estimation; see Section 2.7. The next result shows that the proposed procedure "works" for $h_f^* \asymp h_{\max}$, too, but we are able to state only rate optimality there.

THEOREM 2.4. *Let $K(\cdot)$ be a kernel satisfying conditions $(K1)$–$(K6)$. Then the estimator $\hat f_\varepsilon(t_0)$ provides for some constant C depending on K and r only and ε small enough*

$$\sup_{f\in\mathscr{F}_\varepsilon} E_f \left| \frac{\hat f_\varepsilon(t_0) - f(t_0)}{\sigma(h_f,\varepsilon)\, d(h_f)} \right|^r \le C.$$

Moreover, if $f \in \mathscr{F}_\varepsilon$ is such that $h_f/h_{\max} = o_\varepsilon(1)$, then

$$E_f \left| \frac{\hat f_\varepsilon(t_0) - f(t_0)}{\sigma(h_f,\varepsilon)\, d(h_f)} \right|^r = (C(K)+1)^r(1+o_\varepsilon(1)).$$

2.7. *Applications to global function estimation.* Now we briefly discuss possible applications of Theorem 2.2 or 2.4 to the problem of (global) function estimation. We consider the estimator of a function f which is an implementation of the pointwise adaptive procedure at each point $t \in [0,1]$. As already mentioned, the resulting estimator $\hat f(\cdot)$ can be viewed as a kernel estimator with the variable bandwidth $\hat h(\cdot)$, that is, $\hat f(t) = \tilde f_{\hat h(t)}(t)$.

In the following discussion we explore the properties of this estimator in the standard minimax sense. This means that the quality of the estimator $\hat{f}(t)$ is measured by the mean integrated error of the form

$$R_f(\hat{f}) = E\int_0^1 |\hat{f}(t) - f(t)|^r\, dt = \int_0^1 E|\hat{f}(t) - f(t)|^r\, dt,$$

and we are interested in the limit behavior (rate of decay to zero as $\varepsilon \to 0$) of the maximal value

$$\sup_{f\in\mathscr{F}} R_f(\hat{f}),$$

where $\mathscr{F}$ is a prescribed function class, for example, Hölder, Sobolev or Besov ball.

The results presented previously allow us to split the analysis of the properties of $\hat{f}$ into two different parts: for the first statistical part, everything is done by Theorem 2.4:

$$R_f(\hat{f}) \le C\int_0^1 |\sigma(h_f(t), \varepsilon)\, d(h_f(t))|^r\, dt,$$

where $h_f(t)$ is the pointwise smoothness characteristic of a function f due to Definition 2.2.

The second part relates to the approximation theory: given $\mathscr{F}$, we have to calculate or estimate the value

$$\sup_{f\in\mathscr{F}} \int_0^1 |\sigma(h_f(t), \varepsilon)\, d(h_f(t))|^r\, dt. \tag{2.16}$$

One example of such a calculation can be found in Lepski, Mammen and Spokoiny (1997) for the case of $\mathscr{F}$ being an arbitrary Besov ball. The resulting conclusion is as follows. If the kernel K is of proper regularity and if the value $h_{\max}$ is about 1, then the procedure achieves near minimax rate of convergence for this function class (within a log-factor).

Moreover, the proper choice of $h_{\max}$ leads to the exact minimax rate (without a log-factor). Of course, the corresponding optimal choice of $h_{\max}$ depends on the parameters of the Besov ball and hence requires some information a priori. For a data-driven choice of this parameter, one more global adaptive procedure is to be done; cf. Juditsky (1995).

The idea why the proper choice of the parameter $h_{\max}$ allows us to remove the extra log-factor and to get the minimax rate can be explained in the following way. This factor obviously comes from the multiplier $d(h_f(t))$ in (2.16) which is typically of the logarithmic order. The only exception is, due to Theorem 2.4, for the case when $h_f(t)$ is near the upper value $h_{\max}$ of the range of adaptation. But, if this level is taken properly, then it appears that for all functions f from a prescribed Besov ball and for "almost" all points $t \in [0, 1]$ the pointwise characteristic $h_f(t)$ is about $h_{\max}$ (because of the trimming from above at this level). For the points with $h_f(t)/h_{\max} = o_\varepsilon(1)$ we still have some loss of accuracy of logarithmic order but their contribution into the integral

(2.16) is relatively small. This explains how this extra log-factor can be removed from the global rate of estimation.

2.8. *Choice of a kernel and a range of adaptation.* Here we briefly discuss some aspects of the choice of the kernel $K(\cdot)$ and the range $[h_{\min}, h_{\max}]$.

Strictly speaking, Theorems 2.2–2.4 can be applied only for kernels satisfying conditions $(K1)$–$(K7)$. Note, however, that the procedure makes sense for any kernel under $(K1)$–$(K4)$, see, for example, Lepski, Mammen and Spokoiny (1997).

Another interesting question is the optimization of the kernel K. It turns out that the optimal kernel is produced by the optimization subproblem: to maximize the value $K(0)/\|K\|^2$ over a given function class. The discussion of this problem for the case of Hölder function classes is the subject of the next section. Note only that the solution K^* to the aforementioned optimization subproblem automatically satisfies conditions $(K6)$ and $(K7)$.

The lower bound $h_{\min}$ is recommended to be taken as small as possible. For the abstract "white noise" model under consideration, this bound is of order ε^2. For more realistic models (see Section 4), this choice is restricted by reasons of the experiment's equivalence. In particular, the relevant results of Brown and Low (1996) and Nussbaum (1996) suggest taking for $h_{\min}$ the value of order $h_{\min} \asymp \varepsilon^{2/(1+1/2)} = \varepsilon^{4/3}$ corresponding to the smoothness parameter $1/2$.

2.9. *Nested kernels.* Now we consider one generalization of the problem considered previously. Namely, we study the situation if one takes different kernels for different bandwidth values. This idea is quite natural since small bandwidth values correspond to functions of low regularity and there is no necessity to take high-order kernels. The last hint is justified by the results in the next section on optimal estimation over Hölder classes. We will see that the optimal procedure takes different kernels for different bandwidths and the kernel regularity increases as the bandwidth becomes larger; see Section 3.

Keeping in mind this application and for reference convenience, we formulate a general result on optimal bandwidth selection for a given set of kernels. More precisely, we assume that a system (net) of kernels $\mathbf{K} = (K_h,\ h > 0)$ depending possibly on ε is given. The case considered previously of a fixed kernel corresponds to $K_h(\cdot) = K(\cdot)$. As before, we impose some conditions on these kernels:

(**K1**) The functions $K_h(u)$ are symmetric, that is, $K_h(u) = K_h(-u)$, $u \geq 0$.
(**K2**) The system of functions $K_h(\cdot)$ is compactly supported, that is, $K_h(u) = 0$ for all h and all u outside some compact set C on the real line.
(**K3**) $\int K_h(u)\,du = 1$.
(**K4**) $\sup_h \|K_h\|^2 = \sup_h \int K_h^2(u)\,du < \infty$ and $\inf_h \|K_h\|^2 > 0$.
(**K5**) Set

$$C(h) = K_h(0)\|K_h\|^{-2} - 1.$$

Then there exist two positive constants C_1, C_2 such that

$$C_1 \le C(h) \le C_2.$$

We also introduce two conditions which are natural generalizations of $(K6)$ and $(K7)$.

(**K6**) Uniformly in h,

$$\sup_{0<c\le 1} \frac{\int |K_h(u) - cK_{h/c}(cu)|^2\,du}{\int K_h^2(u)\,du} = 1 + o_\varepsilon(1).$$

(**K7**) Uniformly in h,

$$\sup_{0<c\le 1} \frac{K_h(0) - \int K_{ch}(u)K_h(cu)\,du}{K_h(0) - \int K_h^2(u)\,du} = 1 + o_\varepsilon(1).$$

Now we consider the family of kernel estimators

$$\tilde f_h(t_0) = \frac{1}{h}\int K_h\left(\frac{t - t_0}{h}\right) dX(t). \tag{2.17}$$

The stochastic term for such an estimator has variance $\sigma^2(h, \varepsilon)$ with

$$\sigma^2(h, \varepsilon) = \frac{\varepsilon^2 \|K_h\|^2}{h}.$$

Let again an interval $[h_{\min}, h_{\max}]$ be given with

$$h_{\max}/h_{\min} \to \infty$$

and we choose a bandwidth h in this range. Denote similarly to before

$$\begin{aligned}
\mathscr{K}_h f(t_0) &= \frac{1}{h}\int K_h\left(\frac{t - t_0}{h}\right) f(t)\,dt,\\
\Delta(h) = \Delta_f(h) &= \sup_{0<\eta<h} \left|\mathscr{K}_h f(t_0) - f(t_0)\right|,\\
d_\varepsilon &= \sqrt{2r \ln \frac{\sigma(h_{\min}, \varepsilon)}{\sigma(h_{\max}, \varepsilon)}},\\
\tilde\varepsilon &= \varepsilon d_\varepsilon,\\
d(h) &= \sqrt{2r \ln \frac{\sigma(h, \varepsilon)}{\sigma(h_{\max}, \varepsilon)}},\\
\psi(h, \varepsilon) &= (C(h) + 1)\sigma(h, \varepsilon) = K_h(0)\|K_h\|^{-2}\sigma(h, \varepsilon).
\end{aligned} \tag{2.18}$$

Now the definitions of the "ideal adaptive bandwidth" h_f^* or h_f and of the bandwidth selector $\hat h$ are kept fixed with the modifications indicated previously.

The method of the proofs of Theorems 2.2–2.4 can be extended without any changes to the situation under consideration.

THEOREM 2.5. *Let a system of kernels* $\mathbf{K} = (K_h)$ *satisfy conditions* (**K1**)–(**K6**). *Then the estimator* $\hat{f}_\varepsilon(t_0)$ *corresponding to the adaptive bandwidth* $\hat{h}$ *provides*

$$\sup_{f\in\mathscr{F}_\varepsilon} E_f \left| \frac{\hat{f}_\varepsilon(t_0) - f(t_0)}{\psi(h_f^*, \tilde{\varepsilon})} \right|^r \le 1 + o_\varepsilon(1).$$

THEOREM 2.6. *Let conditions* (**K1**)–(**K5**) *and* (**K7**) *be fulfilled. Then for each* $\varepsilon > 0$ *there exist two functions* $f_0(\cdot)$ *and* $f_1(\cdot)$ *such that* $h_{f_0} = h_{\max}$, $h_{f_1} \ge h_{\min}$ *and, for an arbitrary estimator* T_ε,

$$\max_{f\in\{f_0,f_1\}} E_f \left| \frac{T_\varepsilon - f(t_0)}{\psi(h_f^*, \tilde{\varepsilon})} \right|^r \ge 1 - o_\varepsilon(1).$$

THEOREM 2.7. *Under* (**K1**)–(**K6**) *there is a constant* C *such that for* ε *small enough*

$$\sup_{f\in\mathscr{F}_\varepsilon} E_f \left| \frac{\hat{f}_\varepsilon(t_0) - f(t_0)}{\psi(h_f, \varepsilon\, d(h_f))} \right|^r \le C.$$

If $f \in \mathscr{F}_\varepsilon$ *is such that* $h_f/h_{\max} = o_\varepsilon(1)$, *then*

$$E_f \left| \frac{\hat{f}_\varepsilon(t_0) - f(t_0)}{\psi(h_f, \varepsilon\, d(h_f))} \right|^r \le 1 + o_\varepsilon(1).$$

3. Adaptive pointwise estimation over Hölder classes. In this section we consider the problem of pointwise adaptive estimation for the model (2.1) under the qualitative assumption that the function f belongs to some Hölder class $\Sigma(\beta, L)$. Given β, L, define m as the maximal integer number less than β. Then $\Sigma(\beta, L)$ is the set of functions f such that

$$|f^{(m)}(t) - f^{(m)}(s)| \le L|t - s|^{\beta-m}, \qquad t, s \in R^1.$$

Here $f^{(m)}$ means the mth derivative of f.

We explore the case of adaptive estimation of $f(t_0)$ when the parameters β, L are unknown. Surprisingly, this adaptation can be performed in an optimal way and the following results describe the optimal adaptive procedure and the optimal attainable accuracy.

First we make more precise the problem of adaptive estimation. We assume that the parameters β, L lie in given intervals $\beta \in [\beta_*, \beta^*]$, $L \in [L_*, L^*]$ with some positive $\beta_* < \beta^*$ and $L_* \le L^*$. These parameters characterize the range of adaptation in the case under consideration. Note that the smoothness parameter β is of the most importance to us. For the Lipschitz constant L, we only need the qualitative assumption that it be separated away from zero and infinity. Apparently the results and the procedure can be stated in such a way when the values L_* and L^* are not used.

To formulate the main results, we introduce the following optimization problem which is an optimal recovery problem; see Korostelev (1993), Donoho and Liu (1991), Donoho and Low (1992) and Donoho (1994a, b):

$$(P_\beta)\colon \sup g(0) \quad \text{subject to} \quad \begin{cases} \int g^2 \le 1, \\ g \in \Sigma(\beta, 1). \end{cases}$$

Let g_β solve this problem and let $\mathrm{val}(P_\beta)$ mean $g_\beta(0)$.

REMARK 3.1. The explicit solution g_β and the value $\mathrm{val}(P_\beta) = g_\beta(0)$ are known only for $\beta \le 1$; see, for example, Korostelev (1993). Set

$$f_\beta(t) = (1 - |t|^\beta)_+.$$

Then $g_\beta(t) = af_\beta(bt)$, where the constants a, b are defined by

$$ab^\beta = 1, \qquad ab^{-1/2} = \|f_\beta\|_2 = 1.$$

In particular,

$$\mathrm{val}(P_\beta) = g_\beta(0) = ((2\beta+1)(\beta+1)/4\beta^2)^{\beta/(2\beta+1)}.$$

The case $\beta > 1$ is much more difficult and, to our knowledge, the solution g_β admits an explicit description only for $\beta = 2$. Some more qualitative properties of the functions g_β are discussed later in this section.

It is useful to introduce the notation

$$\tau = (\beta, L),$$

$$\mathscr{T} = \{\tau = (\beta, L)\colon \beta \in [\beta_*, \beta^*],\ L \in [L_*, L^*]\}.$$

For each $\tau \in \mathscr{T}$ denote also

$$\varphi(\tau, \varepsilon) = g_\beta(0) L^{1/(2\beta+1)} \varepsilon^{2\beta/(2\beta+1)}. \tag{3.1}$$

Note that $\varphi(\tau, \varepsilon\sqrt{\ln \varepsilon^{-1}})$ is the optimal asymptotic accuracy of estimation over Hölder class $\Sigma(\tau) = \Sigma(\beta, L)$ in sup-norm; see Korostelev (1993) and Donoho (1994b).

For the pointwise estimation, the rate is $\varepsilon^{2\beta/(2\beta+1)}$ but it cannot be attained adaptively for an unknown β; see Lepski (1990) and Brown and Low (1992). The optimal adaptive rate turned out to be again $(\varepsilon\sqrt{\ln \varepsilon^{-1}})^{2\beta/(2\beta+1)}$; see Lepski (1990). Now we describe the optimal pointwise adaptive procedure and evaluate the corresponding asymptotic risk. Denote

$$\varkappa_\varepsilon = \left[r\left(\frac{1}{2\beta_* + 1} - \frac{1}{2\beta^* + 1} \right) \ln \varepsilon^{-1} \right]^{1/2},$$

$$\bar{\varepsilon} = \varepsilon \varkappa_\varepsilon,$$

$$\varkappa(\beta) = \left[r\left(\frac{1}{2\beta + 1} - \frac{1}{2\beta^* + 1} \right) \ln \varepsilon^{-1} \right]^{1/2},$$

with $\beta < \beta^*$.

The proposed adaptive estimator will be described later in this section. Now we formulate the results where we assume that $\beta^* \le 2$. Possible extensions to the case of arbitrary β^* are discussed in the next section.

THEOREM 3.1. *Let the estimators $\hat{T}_\varepsilon$ of $f(t_0)$ be as defined in Section* 3.1. *Then*

$$\sup_{\tau\in\mathscr{T}} \sup_{f\in\Sigma(\beta,L)} E_f \left| \frac{\hat{T}_\varepsilon - f(t_0)}{\varphi(\tau,\bar{\varepsilon})} \right|^r \le 1 + o_\varepsilon(1).$$

To formulate the lower bound, we have to describe first the extreme points of the range of adaptation $\mathscr{T}$. Note that the parameters β, L have different influence on the accuracy of estimation. For larger β, the value $\varphi(\tau,\varepsilon)$ is smaller. But, if L increases, then $\varphi(\tau,\varepsilon)$ also increases. Denote $\tau_* = (\beta_*, L^*)$ and $\tau^* = (\beta^*, L_*)$. Obviously,

$$\varphi(\tau^*,\varepsilon) \le \varphi(\tau,\varepsilon) \le \varphi(\tau_*,\varepsilon).$$

The next result claims optimality of the estimator $\hat{T}_\varepsilon$ in the uniform sense on the whole interval of adaptation.

THEOREM 3.2. *For each $\varepsilon > 0$ there exist functions $f_0 \in \Sigma(\tau^*)$ and $f_1 \in \Sigma(\tau_*)$ such that, for any estimator T_ε of $f(t_0)$,*

$$\max\left\{ E_{f_0}\left| \frac{T_\varepsilon - f_0(t_0)}{\varphi(\tau^*,\bar{\varepsilon})} \right|^r, E_{f_1}\left| \frac{T_\varepsilon - f_1(t_0)}{\varphi(\tau_*,\bar{\varepsilon})} \right|^r \right\} \ge 1 - o_\varepsilon(1).$$

Finally, we describe the performance of the estimator $\hat{T}_\varepsilon$ on each particular class $\Sigma(\tau)$, $\tau \in \mathscr{T}$.

THEOREM 3.3. *Let $\tau = (\beta, L)$ with $\beta < \beta^*$. Then*

$$\sup_{f\in\Sigma(\tau)} E_f \left| \frac{\hat{T}_\varepsilon - f(t_0)}{\varphi(\tau, \varepsilon\varkappa(\beta))} \right|^r \le 1 + o_\varepsilon(1).$$

Now we present the estimation rule.

3.1. *Estimation procedures.* The proposed procedure is a specialization of the procedure from the previous section for a set of nested kernels with a special choice of kernels K. The construction of these kernels is closely related to the solutions g_β to the problems (P_β) given previously. Roughly speaking, the kernels K_β are obtained by normalization from g_β to provide $\int K_\beta = 1$.

Unfortunately, it is unknown whether the functions g_β are compactly supported and, in particular, it is not clear whether the integrals $\int g_\beta$ are finite. Apparently these values do not enter into the answer and the desirable kernels can be defined using a proper truncation.

Define the modification of the problem (P_β) under support constraints [Donoho (1994b)]. Given $A > 0$,

$$(P_\beta[-A, A]):\ \sup g(0) \quad \text{subject to} \quad \begin{cases} \int_{-A}^{A} g^2 \le 1, \\ g \in \Sigma(\beta, 1). \end{cases}$$

One has easily $\mathrm{val}(P_\beta) \le \mathrm{val}(P_\beta[-A, A])$ and we also use the property [Donoho (1994b), Lemma 6.1]

$$\mathrm{val}(P_\beta[-A, A]) \to \mathrm{val}(P_\beta), \qquad A \to \infty. \tag{3.2}$$

Moreover, using the method from Donoho and Low (1992), Theorem 3, one may state this assertion uniformly in β. In what follows we assume that a number A is taken depending on ε and growing as $\varepsilon \to 0$, for instance, $A = A_\varepsilon = \log \varepsilon^{-1}$.

Denote by $g_{\beta, A}$ the solution to $(P_\beta[-A, A])$. For more information about the behavior of the functions $g_{\beta, A}$, see Lemma 5.5.

To apply the procedure from the preceding section, we have to state the correspondence between the bandwidth h and the smoothness parameters $\tau = (\beta, L)$. Denote

$$\begin{aligned} h(\tau) &= (\varepsilon \varkappa(\beta)/L)^{2/(2\beta+1)}, \\ h_{\min} &= h(\tau_*), \\ h_{\max} &= h(\tau^*). \end{aligned} \tag{3.3}$$

Next, introduce a function $\beta(h)$ as a solution in β of the equation

$$h^\beta = \sigma(h, \varepsilon \varkappa(\beta)) = \varepsilon \varkappa(\beta) h^{-1/2},$$

that is,

$$\beta(h) = \frac{\ln(\varepsilon \varkappa(\beta))}{\ln h} - \frac{1}{2}. \tag{3.4}$$

Denote also

$$K_h = \lambda_\beta^{-1} g_{\beta, A} \mathbf{1}_{[-A, A]}, \tag{3.5}$$

with $\lambda_\beta = \int_{-A}^{A} g_{\beta,A}(t)\, dt$. Now the data-driven bandwidth $\hat{h}$ is defined in Sections 2.5 and 2.9 and we let $\hat{T}_\varepsilon = \tilde{f}_{\hat{h}}(t_0)$.

4. Further developments.

4.1. *Other nonparametric statistical models.* In this paper we concentrate on the simplest "white noise" model (2.1). This type of model allows us to em-

phasize more clearly the main ideas, avoiding a lot of technical details which correspond to more realistic models. However, we believe that other kinds of nonparametric statistical models (discrete-time regression models with Gaussian and non-Gaussian errors, density or spectral density function models, etc.) can be considered in the same manner, perhaps under some technical assumptions. The results of Brown and Low (1996), Low (1992) and Nussbaum (1996) can be mentioned in this context. These results guarantee equivalence in some sense between the regression or density function models and a proper white noise model if the smoothness parameter is more than $1/2$. This motivates the applicability of Theorems 2.2–2.4 for these models.

4.2. *Estimation of linear functionals.* The problem of estimation at a point can be considered as the particular case of the problem of estimation of a linear functional. The problem of estimation of linear functionals was studied intensively in the present context in Donoho and Low (1992), Donoho and Liu (1991), Donoho (1994b) and Efroimovich and Low (1994). The corresponding results show a close relation between the particular problem of pointwise estimation and a general problem for an arbitrary linear functional. We conjecture that all of the results given previously can be extended in a similar way to the general case.

4.3. *The case* $\beta^* > 2$. The fact $\beta^* \le 2$ was used essentially in the proof of Theorems 3.1–3.3, in particular, for the proof of Lemma 5.5.

For the case $\beta^* > 2$, the statements of Theorem 3.1 or 3.3 cannot be extended directly from the case $\beta^* \le 2$ because the structure of Hölder classes is not embedded: $\Sigma(\beta', 1)$ does not belong to $\Sigma(\beta, 1)$ for $\beta' < \beta$. It can be illustrated explicitly by the first statement of Lemma 5.7 where one easily has $G(\beta', \beta; A) = \infty$, for instance, if $\beta' = 1$ and $\beta = 3$ since $\Sigma(3, 1)$ contains all linear functions.

Nevertheless, we conjecture that all the results stated previously can be extended to the case of an arbitrary β under some additional constraints on the Hölder classes $\Sigma(\beta, L)$ for $\beta > 2$ type of boundedness of all derivatives of order $1, \dots, \lfloor \beta \rfloor$.

5. Proofs. In this section we prove Theorems 2.2 and 2.3. The result of Theorem 2.1 follows from Theorem 2.3. The necessary corrections for the proofs of Theorems 2.4–2.7 are obvious and omitted. Finally, we show how one can derive the result of Theorems 3.1–3.3 from the general results of Theorems 2.5–2.7.

5.1. *Proof of Theorem* 2.2. Let us fix some function f from $\mathscr{F}_\varepsilon$ and let $h_f, \hat{h}$ be due to Definition 2.2. Without loss of generality we assume that $h_f \in \mathscr{H}$. Otherwise we can replace h_f by the closest from below point of $\mathscr{H}$ and the result of Theorem 2.2 remains valid.

The definition of h_f yields for each $h \le h_f$ the inequality

$$\left| \mathscr{K}_h f(t_0) - f(t_0) \right| \le C(K) \sigma(h_f, \varepsilon) d(h_f), \tag{5.1}$$

where

$$d(h) = \bigl(r \ln(h_{\max}/h)\bigr)^{1/2}.$$

We consider the case when h_f is separated away from $h_{\max}$, that is,

$$|d(h_f)|^{-1} = o_\varepsilon(1). \tag{5.2}$$

The case with $d(h_f) \asymp h_{\max}$ can be considered in a similar way; see also Lepski, Mammen and Spokoiny (1997).

Recall the notation

$$\psi(h, \varepsilon) = (C(K) + 1)\sigma(h, \varepsilon) = (C(K) + 1)\, \varepsilon \|K\| h^{-1/2},$$
$$\alpha(h) = d^{-1/2}(h)$$

and define the value $h_{f,1}$ as a solution of the equation

$$\alpha(h_{f,1})\sigma(h_{f,1}, \varepsilon)\, d(h_{f,1}) = 4C(K)\sigma(h_f, \varepsilon)\, d(h_f). \tag{5.3}$$

One has obviously $h_{f,1} < h_f$ for ε small enough in view of (5.2).

Now we split the grid $\mathscr{H}$ into three parts

$$\mathscr{H}^{(1)} = \{h \in \mathscr{H}\colon h \geq h_f\},$$
$$\mathscr{H}^{(2)} = \{h \in \mathscr{H}\colon h_{f,1} < h < h_f\},$$
$$\mathscr{H}^{(3)} = \{h \in \mathscr{H}\colon h \leq h_{f,1}\},$$

and decompose the risk of the estimator $\hat{f}_\varepsilon(t_0)$ in a similar way:

$$\begin{aligned} E_f\bigl|\hat{f}_\varepsilon(t_0) - f(t_0)\bigr|^r &= E_f\bigl|\hat{f}_\varepsilon(t_0) - f(t_0)\bigr|^r \mathbf{1}(\hat{h} \geq h_f) \\ &\quad + E_f\bigl|\hat{f}_\varepsilon(t_0) - f(t_0)\bigr|^r \mathbf{1}(h_{f,1} < \hat{h} < h_f) \\ &\quad + E_f\bigl|\hat{f}_\varepsilon(t_0) - f(t_0)\bigr|^r \mathbf{1}(\hat{h} \leq h_{f,1}) \\ &= R_\varepsilon^{(1)} + R_\varepsilon^{(2)} + R_\varepsilon^{(3)}. \end{aligned}$$

We shall explore each term in this decomposition separately starting from $R_\varepsilon^{(1)}$.

Set, for $h > 0$,

$$\zeta(h) = \sigma^{-1}(h, \varepsilon)\xi(h).$$

Then the random variable $\zeta(h)$ is for any h standard normal. Using now the definition of $\hat{h}$, the decomposition (2.4) and (5.1), we find

$$\begin{aligned} R_\varepsilon^{(1)} &\leq E_f\bigl(\bigl|\tilde{f}_{\hat{h}}(t_0) - \tilde{f}_{h_f}(t_0)\bigr| + \bigl|\tilde{f}_{h_f}(t_0) - f(t_0)\bigr|\bigr)^r \mathbf{1}(\hat{h} \geq h_f) \\ &\leq E_f\bigl[(1 + \alpha(h_f))\sigma(h_f, \varepsilon)\, d(h_f) \\ &\qquad + C(K)\sigma(h_f, \varepsilon)\, d(h_f) + |\xi(h_f)|\bigr]^r \mathbf{1}(\hat{h} \geq h_f) \qquad (5.4) \\ &\leq (1 + \alpha(h_f))^r E_f\bigl[\psi(h_f, \varepsilon)\, d(h_f) + \sigma(h_f, \varepsilon)|\zeta(h_f)|\bigr]^r \mathbf{1}(\hat{h} \geq h_f) \\ &\leq (1 + \alpha(h_f))^r\, \psi^r(h_f, \varepsilon\, d(h_f))\, E_f\bigl(1 + d^{-1}(h_f)|\zeta(h_f)|\bigr)^r \mathbf{1}(\hat{h} \geq h_f). \end{aligned}$$

To estimate the last expression, we use the following technical assertion.

LEMMA 5.1. *Let ζ be a standard Gaussian random variable and let A be a random event on the same probability space. Then the following statements are satisfied*:

(i) *There exist positive constants C_r, γ_r such that, for any $\gamma < \gamma_r$,*

$$E(1+\gamma|\zeta|)^r \mathbf{1}(A) \le (1+2r\sqrt{\gamma})P(A) + C_r \exp\{-1/(2\gamma)\}.$$

(ii) *It holds for some constant C'_r and any positive numbers γ, Γ, with $\Gamma \ge 1$ and $\gamma\Gamma \ge 1$,*

$$E(1+\gamma|\zeta|)^r \mathbf{1}(A) \le C'_r(\gamma\Gamma)^r[P(A) + \exp\{-\Gamma^2/2\}].$$

PROOF. We get

$$E(1+\gamma|\zeta|)^r \mathbf{1}(A) \le E(1+\sqrt{\gamma})^r \mathbf{1}(A) + E(1+\gamma|\zeta|)^r \mathbf{1}(|\zeta| > \gamma^{-1/2}).$$

Now (i) follows from the fact that for $\gamma < \gamma_r = 2^{1/(r-1)} - 1$

$$(1+\sqrt{\gamma})^r - 1 \le 2r\sqrt{\gamma},$$

$$E(1+\gamma|\zeta|)^r \mathbf{1}(|\zeta| > \gamma^{-1/2}) \le C_r \exp\{-1/(2\gamma)\},$$

which can be readily verified.

Next, one obtains under $\Gamma \ge 1$ and $\gamma\Gamma \ge 1$ in a similar manner

$$\begin{aligned} E(1+\gamma|\zeta|)^r \mathbf{1}(A) &\le E(1+\gamma\Gamma)^r \mathbf{1}(A) + E(1+\gamma|\zeta|)^r \mathbf{1}(|\zeta| > \Gamma) \\ &\le (\gamma\Gamma)^r \big[(1+(\gamma\Gamma)^{-1})^r P(A) + ((\gamma\Gamma)^{-1} + |\zeta|/\Gamma)^r \mathbf{1}(|\zeta| > \Gamma)\big] \\ &\le C'_r(\gamma\Gamma)^r \big[P(A) + \exp\{-\Gamma^2/2\}\big]. \end{aligned}$$

□

Since $d^{-1}(h_f) = o_\varepsilon(1)$ and hence $\alpha(h_f) = o_\varepsilon(1)$, we get from (5.4) using Lemma 5.1(i), with $\gamma = d^{-1}(h_f)$,

$$R_\varepsilon^{(1)} \le \psi^r(h_f, \varepsilon\, d(h_f)) P(\hat{h} \le h_f)(1 + o_\varepsilon(1)).$$

Next, we estimate $R_\varepsilon^{(2)}$. One has, similar to before,

$$\begin{aligned} R_\varepsilon^{(2)} &= \sum_{h \in \mathscr{H}^{(2)}} E_f \big|\tilde{f}(h)(t_0) - f(t_0)\big|^r \mathbf{1}(\hat{h} = h) \\ &\le \sum_{h \in \mathscr{H}^{(2)}} E_f \big(C(K)\sigma(h_f, \varepsilon) d(h_f) + |\xi(h)|\big)^r \mathbf{1}(\hat{h} = h) \\ &\le \sum_{h \in \mathscr{H}^{(2)}} E_f \big(\psi(h_f, \varepsilon) d(h_f) + \sigma(h_f, \varepsilon)|\zeta(h)|\big)^r \mathbf{1}(\hat{h} = h) \\ &\le \psi^r(h_f, \varepsilon\, d(h_f)) \sum_{h \in \mathscr{H}^{(2)}} E_f \big(1 + \gamma(h)|\zeta(h)|\big)^r \mathbf{1}(\hat{h} = h), \end{aligned} \tag{5.5}$$

where

$$\gamma(h) = \frac{\sigma(h, \varepsilon)}{\psi(h_f, \varepsilon)\, d(h_f)} = \frac{\sigma(h, \varepsilon)}{(C(K)+1)\sigma(h_f, \varepsilon)\, d(h_f)}. \tag{5.6}$$

Evidently, $\gamma(h) \le \gamma(h_{f,1})$ for $h \in \mathscr{H}^{(2)}$ and using Lemma 5.1(i), we get

$$R_\varepsilon^{(2)} \le \psi^r(h_f, \varepsilon\, d(h_f)) \sum_{h\in\mathscr{H}^{(2)}} \bigl(1 + 2r\sqrt{\gamma(h)}\bigr) P(\hat h = h) + C_r \exp\{-1/(2\gamma(h))\}$$

$$\le \psi^r(h_f, \varepsilon\, d(h_f))\Bigl[\Bigl(1 + 2r\sqrt{\gamma(h_{f,1})}\Bigr) P(h_{f,1} < \hat h < h_f) + C_r \#(\mathscr{H}^{(2)}) \exp\{-1/(2\gamma(h_{f,1}))\}\Bigr].$$

Here $\#(\mathscr{H}^{(2)})$ means the number of elements in $\mathscr{H}^{(2)}$. Due to (5.3) and (5.6),

$$\gamma(h_{f,1}) = \frac{\sigma(h_{f,1}, \varepsilon)}{(C(K)+1)\sigma(h_f, \varepsilon)\, d(h_f)} \le \frac{C(K)}{4(C(K)+1)\, d(h_{f,1})\alpha(h_{f,1})} \le 4\alpha(h_{f,1})$$

and hence $\gamma(h_{f,1}) = o_\varepsilon(1)$. Next, it is easy to estimate

$$\#(\mathscr{H}^{(2)}) \le \frac{\ln(h_f/h_{f,1})}{1 - (1+\alpha(h_{f,1}))^{-1}} \le \frac{\ln(h_f/h_{f,1})}{\alpha(h_{f,1})}$$

and hence

$$\#(\mathscr{H}^{(2)}) \exp\{-1/(2\gamma(h_{f,1}))\} = o_\varepsilon(1).$$

Therefore,

$$R_\varepsilon^{(2)} \le \psi^r(h_f, \varepsilon\, d(h_f))\bigl[P(h_{f,1} < \hat h < h_f) + o_\varepsilon(1)\bigr].$$

Getting together the estimates for $R_\varepsilon^{(1)}$ and for $R_\varepsilon^{(2)}$, we obtain

$$R_\varepsilon^{(1)} + R_\varepsilon^{(2)} \le \psi^r(h_f, \varepsilon\, d(h_f))\bigl[P(\hat h > h_{f,1}) + o_\varepsilon(1)\bigr] \le \psi^r(h_f, \varepsilon\, d(h_f))\,(1 + o_\varepsilon(1)).$$

It remains to show that

$$R_\varepsilon^{(3)} = \psi^r(h_f, \varepsilon\, d(h_f))\, o_\varepsilon(1).$$

One gets similarly to (5.5)

$$R_\varepsilon^{(3)} \le \psi^r(h_f, \varepsilon\, d(h_f)) \sum_{h\in\mathscr{H}^{(3)}} E_f\bigl(1 + \gamma(h)|\zeta(h)|\bigr)^r \mathbf{1}(\hat h = h), \tag{5.7}$$

where $\gamma(h)$ is due to (5.6) and, for $h \in \mathscr{H}^{(3)}$,

$$\gamma(h) < \frac{\sigma(h, \varepsilon)}{\sigma(h_f, \varepsilon)} \le \sqrt{\frac{h_f}{h}}.$$

We shall estimate this sum using Lemma 5.1(i), but first we show that the probability $P_f(\hat h = h)$ is small enough.

LEMMA 5.2. *For each* $h \in \mathscr{H}^{(3)}$,

$$P(\hat{h} = h) \le 2\alpha^{-1}(h) \exp\{-d^2(h)/2 - d(h)\}.$$

PROOF. Let us fix some $h \in \mathscr{H}^{(3)}$ and let h_+ be the preceding element of $\mathscr{H}$, that is, $h = h_+/(1 + \alpha_{h_+})$. We also use the notation

$$\mathscr{H}^-(h) = \{\eta \in \mathscr{H}\colon \eta < h\}.$$

The definition of $\hat{h}$ yields

$$P(\hat{h} = h) \le \sum_{\eta \in \mathscr{H}^-(h)} P\big(\big|\tilde{f}_\eta(t_0) - \tilde{f}_{h_+}(t_0)\big| > (1 + \alpha_\eta)\sigma(\eta, \varepsilon)\, d(\eta)\big).$$

Since $h < h_{f,1}$, then $\eta, h_+ < h_f$ and by (5.1)

$$\begin{aligned}\big|\tilde{f}_\eta(t_0) - \tilde{f}_{h_+}(t_0)\big| &\le \big|\mathscr{K}_\eta f(t_0) - f(t_0)\big| + \big|\mathscr{K}_{h_+} f(t_0) - f(t_0)\big| + |\xi(\eta) - \xi(h_+)| \\ &\le 2C(K)\sigma(h_f, \varepsilon)\, d(h_f) + |\xi(\eta) - \xi(h_+)|.\end{aligned}$$

But $\eta \le h \le h_{f,1}$ and thus

$$2C(K)\sigma(h_f, \varepsilon)\, d(h_f) \le \alpha(h)\sigma(h, \varepsilon)\, d(h)/2 \le \alpha(\eta)\sigma(\eta, \varepsilon)\, d(\eta)/2.$$

Notice also that

$$\xi(\eta) - \xi(\eta_+) = \varepsilon \int \left(\frac{1}{\eta} K\left(\frac{t - t_0}{\eta}\right) - \frac{1}{h_+} K\left(\frac{t - t_0}{h_+}\right)\right) dW(t),$$

that is, this difference is normal $\mathscr{N}(0, \sigma^2(\eta, h_+; \varepsilon))$ with

$$\sigma^2(\eta, h_+; \varepsilon) = \frac{\varepsilon^2}{\eta} \int \big|K(u) - cK(uc)\big|^2\, du,$$

where $c = \eta/h_+ \le 1$.

The condition $(K6)$ provides

$$\sigma^2(\eta, h_+; \varepsilon) \le \sigma^2(\eta, \varepsilon).$$

We now obtain

$$\begin{aligned}P_f(\hat{h} = h) &\le \sum_{\eta \in \mathscr{H}^-(h)} P_f\big(|\xi(\eta) - \xi(h_+)| > (1 + \alpha(\eta)/2)\sigma(\eta, \varepsilon)\, d(\eta)\big) \\ &= \sum_{\eta \in \mathscr{H}^-(h)} P\left(|\zeta| > (1 + \alpha(\eta)/2)\, d(\eta) \frac{\sigma(\eta, \varepsilon)}{\sigma(\eta, h_+; \varepsilon)}\right) \\ &\le \sum_{\eta \in \mathscr{H}^-(h)} P\big(|\zeta| > (1 + \alpha(\eta)/2)\, d(\eta)\big) \\ &\le \sum_{\eta \in \mathscr{H}^-(h)} \exp\{-(1 + \alpha(\eta)/2)^2\, d^2(\eta)/2\}.\end{aligned}$$

To complete the proof of the lemma, we notice that

$$(1 + \alpha(\eta)/2)^2\, d^2(\eta) > d^2(\eta) + 2d(\eta)$$

and one derives easily, using the definition of the grid $\mathscr{H}$,

$$\sum_{\eta\in\mathscr{H}^-(h)} \exp\{-d^2(\eta)/2 - d(\eta)\} \le \exp\{-d(h)\} \sum_{\eta\in\mathscr{H}^-(h)} (h/h_{\max})^{r/2}$$
$$\le \exp\{-d(h)\}\alpha^{-1}(h)(h/h_{\max})^{r/2}$$

and the lemma follows.

Now we apply Lemma 5.1(ii) to each term in (5.7) with $\gamma = \gamma(h)$ and $\Gamma = \Gamma(h) = \sqrt{d^2(h) + 2d(h)}$. We obtain for each $h \in \mathscr{H}^{(3)}$, using Lemma 5.2,

$$\begin{aligned} &E_f\big(1 + \gamma(h)|\zeta(h)|\big)^r \mathbf{1}(\hat{h} = h) \\ &\quad \le C\big[d(h)\sqrt{h_f/h}\big]^r \big(P(\hat{h} = h) + \exp\{-d^2(h)/2 - d(h)\}\big) \\ &\quad \le C d^{r+1}(h)(h_f/h)^{r/2} \exp\{-d^2(h)/2 - d(h)\} \\ &\quad = C d^{r+1}(h) \exp\{-d(h)\}(h_f/h_{\max})^{r/2}. \end{aligned}$$

Therefore,

$$R_\varepsilon^{(3)} \le C\psi^r(h_f, \varepsilon\, d(h_f)) \exp\{-d^2(h_f)/2\} \sum_{h\in\mathscr{H}^{(3)}} d^{r+1}(h) \exp\{-d(h)\}.$$

The following lemma completes the proof of the theorem.

LEMMA 5.3. *There is a constant C such that*

$$\sum_{h\in\mathscr{H}^{(3)}} d^{r+1}(h) \exp\{-d(h)\} \le C.$$

PROOF. Let h_k be the kth element of the grid $\mathscr{H}$, that is,

$$h_k = h_{\max} \prod_{i=0}^{k-1} [1 + \alpha(h_i)]^{-1}.$$

Then obviously, for large k,

$$\begin{aligned} d^2(h_k) &= r \ln(h_{\max}/h) \\ &= r \sum_{i=0}^{k-1} (1 + \alpha(h_i)) \\ &\ge rk \ln(1 + \alpha(h_k)) \\ &\ge rk\alpha(h_k)/2 = rk d^{-1/2}(h_k)/2. \end{aligned}$$

This yields

$$d(h_k) \ge (rk/2)^{2/5}.$$

Similarly, one gets $d(h_k) \le \sqrt{rk}$.

Now let k_f be such that $h_{f,1} = h_{k_f}$. Then easily

$$\sum_{h\in\mathscr{H}^{(3)}} d^{r+1}(h)\exp\{-d(h)\} \le \sum_{k=k_f}^{\infty} (rk)^{(r+1)/2}\exp\{-(rk/2)^{2/5}\} \le C$$

and the assertion follows. □

Theorem 2.2 is proved.

5.2. *Proof of Theorem* 2.3. Define

$$f_0(t) \equiv 0$$

and

$$\begin{aligned} f_1(t) &= (1-\alpha_\varepsilon)\frac{\varepsilon}{\sqrt{h_{\min}}\|K\|}\sqrt{r\ln\frac{h_{\max}}{h_{\min}}}\,K\biggl(\frac{t-t_0}{h_{\min}}\biggr) \\ &= (1-\alpha_\varepsilon)\frac{\sigma(h_{\min},\varepsilon)d_\varepsilon}{\|K\|^2}K\biggl(\frac{t-t_0}{h_{\min}}\biggr) \\ &= v_\varepsilon K\biggl(\frac{t-t_0}{h_{\min}}\biggr), \end{aligned} \tag{5.8}$$

where

$$\alpha_\varepsilon = d_\varepsilon^{-1/2} = \bigl(p\ln h_{\max}/h_{\min}\bigr)^{-1/4} = o_\varepsilon(1)$$

and

$$v_\varepsilon = (1-\alpha_\varepsilon)\sigma(h_{\min},\varepsilon)d_\varepsilon\|K\|^{-2}.$$

It is obvious that

$$h_{f_0} = h_{\max}.$$

Next we show that

$$h_{f_1} \ge h_{\min}.$$

In fact, for each $\eta < h_{\min}$ one has by $(K7)$, for $c = \eta/h_{\min}$,

$$\begin{aligned} \bigl|\mathscr{K}_\eta f_1(t_0) - f_1(t_0)\bigr| &= \biggl|\frac{1}{\eta}\int K\biggl(\frac{t-t_0}{\eta}\biggr)\bigl[f_1(t)-f_1(t_0)\bigr]\,dt\biggr| \\ &= v_\varepsilon\biggl|\frac{1}{\eta}\int K\biggl(\frac{t-t_0}{\eta}\biggr)\biggl[K\biggl(\frac{t-t_0}{h_{\min}}\biggr)-K(0)\biggr]\,dt\biggr| \\ &= v_\varepsilon\biggl|\int K(u)[K(uc)-K(0)]\,dt\biggr| \\ &\le v_\varepsilon(K(0)-\|K\|^2). \end{aligned}$$

This gives, for f_1,

$$\Delta(h_{\min}) = \Delta_{f_1}(h_{\min}) \le (1-\alpha_\varepsilon)\frac{K(0)-\|K\|^2}{\|K\|^2}\sigma(h_{\min},\varepsilon)d_\varepsilon$$
$$= (1-\alpha_\varepsilon)C(K)\sigma(h_{\min},\varepsilon)d_\varepsilon,$$

which means that $h_{f_1} \ge h_{\min}$.

Let the measures $P_{0,\varepsilon}$ and $P_{1,\varepsilon}$ correspond to the model (2.1) with the functions f_0 and f_1, respectively. It is clear that these measures are Gaussian. Moreover, by Girsanov's theorem

$$\frac{dP_{1,\varepsilon}}{dP_{0,\varepsilon}} = \exp\left\{\varepsilon^{-1}\int f_1(t)\,dX(t) - \frac{1}{2}\varepsilon^{-2}\int f_1^2(t)\,dt\right\}$$
$$= \exp\left\{q_\varepsilon\zeta_\varepsilon - \frac{1}{2}q_\varepsilon^2\right\}$$

where

$$q_\varepsilon^2 = \varepsilon^{-2}\int f_1^2(t)\,dt,$$
$$\zeta_\varepsilon = \frac{\varepsilon^{-1}}{q_\varepsilon}\int f_1(t)\,dX(t) \tag{5.9}$$

and

$$\mathscr{L}(\zeta_\varepsilon \mid P_{0,\varepsilon}) = \mathscr{N}(0,1).$$

The theorem will follow if we show that, for any estimator T_ε,

$$\liminf_{\varepsilon\to 0} R_\varepsilon = 1, \tag{5.10}$$

where

$$R_\varepsilon = \max\left\{E_{0,\varepsilon}\left|\frac{T_\varepsilon}{\psi(h_{\max},\varepsilon)d_\varepsilon}\right|^r, E_{1,\varepsilon}\left|\frac{T_\varepsilon - f_1(t_0)}{\psi(h_{\min},\varepsilon)d_\varepsilon}\right|^r\right\}$$

and $E_{0,\varepsilon}$, $E_{1,\varepsilon}$ mean integration w.r.t. the measures $P_{0,\varepsilon}$, $P_{1,\varepsilon}$.

Note that

$$\frac{f_1(t_0)}{\psi(h_{\min},\varepsilon)d_\varepsilon} = 1-\alpha_\varepsilon$$

and denote

$$\theta_\varepsilon = \frac{T_\varepsilon}{\psi(h_{\min},\varepsilon)d_\varepsilon(1-\alpha_\varepsilon)},$$
$$D_\varepsilon = \frac{\psi(h_{\min},\varepsilon)}{\psi(h_{\max},\varepsilon)} = \sqrt{\frac{h_{\max}}{h_{\min}}}.$$

With this notation,

$$R_\varepsilon = |1-\alpha_\varepsilon|^r \max\left\{D_\varepsilon^r E_{0,\varepsilon}|\theta_\varepsilon|^r, E_{1,\varepsilon}|1-\theta_\varepsilon|^r\right\}.$$

Now (5.10) is equivalent to

$$\liminf_{\varepsilon\to 0} \max\{D_\varepsilon^r E_{0,\varepsilon}|\theta_\varepsilon|^r, E_{1,\varepsilon}|1-\theta_\varepsilon|^r\} \geq 1.$$

Furthermore, due to (5.8) and (5.9),

$$\begin{aligned} q_\varepsilon^2 &= \int f_1^2(t)\,dt \\ &= \varepsilon^{-2} v_\varepsilon^2 \int K^2\left(\frac{t-t_0}{h_{\min}}\right) dt \\ &= \varepsilon^{-2} v_\varepsilon^2 h_{\min} \|K\|^2 \\ &= (1-\alpha_\varepsilon)^2 r \ln \frac{h_{\max}}{h_{\min}} \\ &= (1-\alpha_\varepsilon)^2 d_\varepsilon^2 \end{aligned}$$

and

$$\frac{1}{q_\varepsilon}\left(r\ln D_\varepsilon - \frac{1}{2}q_\varepsilon^2\right) = \frac{1}{(1-\alpha_\varepsilon)d_\varepsilon}\left(d_\varepsilon^2 - \frac{1}{2}(1-\alpha_\varepsilon)^2 d_\varepsilon^2\right) \geq \alpha_\varepsilon d_\varepsilon \to \infty$$

as $\varepsilon \to 0$.

Now the result of the theorem follows directly from the next lemma.

LEMMA 5.4. *Let for each $\varepsilon > 0$ two Gaussian measures $P_{0,\varepsilon}$ and $P_{1,\varepsilon}$ be given with*

$$\ln \frac{dP_{\varepsilon,1}}{dP_{\varepsilon,0}} = q_\varepsilon \zeta_\varepsilon - \frac{1}{2}q_\varepsilon^2,$$

where

$$\mathscr{L}(\zeta_\varepsilon \mid P_{0,\varepsilon}) = \mathscr{N}(0,1)$$

and $q_\varepsilon \to \infty$.

Then let the numbers D_ε be such that

$$\frac{1}{q_\varepsilon}\left(r\ln D_\varepsilon - \frac{1}{2}q_\varepsilon^2\right) \to \infty. \tag{5.11}$$

Then for any estimator θ_ε such that

$$\liminf_{\varepsilon\to 0} D_\varepsilon^r E_{0,\varepsilon}|\theta_\varepsilon|^r \leq C < \infty, \tag{5.12}$$

one has

$$\liminf_{\varepsilon\to 0} E_{1,\varepsilon}|\theta_\varepsilon - 1|^r \geq 1.$$

PROOF. Fix any estimators θ_ε satisfying (5.12). Then take an arbitrary $M > 0$ and denote

$$\pi = \frac{1}{2CM},$$

where C is from condition (5.12). This condition yields for ε small enough

$$D_\varepsilon^r E_{0,\varepsilon}|\theta_\varepsilon|^r \le 2C$$

and

$$R_\varepsilon = E_{1,\varepsilon}|\theta_\varepsilon - 1|^r \ge E_{1,\varepsilon}|1-\theta_\varepsilon|^r + \pi D_\varepsilon^r E_{0,\varepsilon}|\theta_\varepsilon|^r - 2C\pi.$$

Denote

$$Z_\varepsilon = \frac{dP_{0,\varepsilon}}{dP_{1,\varepsilon}}.$$

One has

$$Z_\varepsilon = \exp\{-q_\varepsilon\zeta_\varepsilon + \tfrac{1}{2}q_\varepsilon^2\} = \exp\{-q_\varepsilon(\zeta_\varepsilon - q_\varepsilon) - \tfrac{1}{2}q_\varepsilon^2\},$$

where by Girsanov's theorem

$$\mathscr{L}(\zeta_\varepsilon - q_\varepsilon \mid P_{1,\varepsilon}) = \mathscr{N}(0,1)$$

and hence

$$P_{1,\varepsilon}(\zeta_\varepsilon - q_\varepsilon \le M) = \Phi(M),$$

where $\Phi(\cdot)$ is the Laplace function.

Now set

$$\delta_\varepsilon = \exp\{-q_\varepsilon\}$$

and introduce events

$$A_\varepsilon = \{\theta_\varepsilon \le \delta_\varepsilon\},$$
$$B_\varepsilon = \{\zeta_\varepsilon - q_\varepsilon \le M\}.$$

Now one has, on $A_\varepsilon \cap B_\varepsilon$,

$$Z_\varepsilon \ge \exp\{-q_\varepsilon M - \tfrac{1}{2}q_\varepsilon^2\},$$
$$|1-\theta_\varepsilon| \ge 1 - \delta_\varepsilon,$$

and $|\theta_\varepsilon| \ge \delta_\varepsilon$ on the complement A_ε^c of A_ε. Therefore,

$$\begin{aligned} R_\varepsilon &\ge E_{1,\varepsilon}|1-\theta_\varepsilon|^r + \pi D_\varepsilon^r E_{0,\varepsilon}|\theta_\varepsilon|^r - \frac{1}{M} \\ &= E_{1,\varepsilon}\big(|1-\theta_\varepsilon|^r + \pi D_\varepsilon^r Z_\varepsilon|\theta_\varepsilon|^r\big) - \frac{1}{M} \\ &\ge P(A_\varepsilon)|1-\delta_\varepsilon|^r + \pi D_\varepsilon^r \delta_\varepsilon^r \exp\left\{-q_\varepsilon M - \frac{1}{2}q_\varepsilon^2\right\} P(A_\varepsilon^c \cap B_\varepsilon). \end{aligned}$$

Condition (5.11) implies

$$
\begin{aligned}
D_\varepsilon^r \delta_\varepsilon^r \exp\left\{-q_\varepsilon M - \frac{1}{2}q_\varepsilon^2\right\} &= \exp\left\{r \ln D_\varepsilon - \frac{1}{2}q_\varepsilon^2 - r q_\varepsilon - M q_\varepsilon\right\} \\
&= \exp\left\{q_\varepsilon\left[\frac{1}{q_\varepsilon}\left(r \ln D_\varepsilon - \frac{1}{2}q_\varepsilon^2\right) - r - M\right]\right\} \\
&\to \infty, \qquad \varepsilon \to 0.
\end{aligned} \tag{5.13}
$$

Now we use that $P(A_\varepsilon^c \cap B_\varepsilon) \geq P(A_\varepsilon^c) - P(B_\varepsilon^c)$ and $P(B_\varepsilon^c) = \bar{\Phi}(M) = 1 - \Phi(M)$. If $P(A_\varepsilon^c) \geq 2P(B_\varepsilon^c) = 2\bar{\Phi}(M)$, then R_ε is large in view of (5.13). But if $P(A_\varepsilon^c) \leq 2P(B_\varepsilon^c) = 2\bar{\Phi}(M)$, then

$$
\begin{aligned}
R_\varepsilon &\geq P(A_\varepsilon)(1-\delta_\varepsilon)^r - 1/M \\
&\geq (1 - 2\bar{\Phi}(M))|1-\delta_\varepsilon|^r - 1/M.
\end{aligned}
$$

This proves that

$$
\liminf_{\varepsilon \to 0} R_\varepsilon \geq (1 - 2\bar{\Phi}(M)) - 1/M
$$

for each finite $M > 0$, and the lemma follows. □

5.3. *Proof of Theorems* 3.1 *and* 3.3. We deduce Theorems 3.1 and 3.3 as corollaries of Theorems 2.5 and 2.7. For this we have to check conditions (**K1**)–(**K7**) for the kernels (K_h) and to verify that the results of Theorems 2.5 and 2.7 provide just the accuracy claimed in Theorems 3.1 and 3.3.

We start with a technical result describing some useful properties of the solution $g_{\beta, A}$ to the problem $(P_\beta[-A, A])$.

LEMMA 5.5. *Let* $A > 1$ *and* $\beta^* \leq 2$. *The following statements are fulfilled for each* $\beta \leq \beta^*$:

(i) *The solution* $g_{\beta, A}$ *to* $P_\beta[-A, A]$ *exists and is unique*.

(ii) *The function* $g_{\beta, A}$ *is symmetric, that is,* $g_{\beta, A}(t) = g_{\beta, A}(-t)$, $t \in R^1$.

(iii) *The function* $g_{\beta, A}$ *has maximum at* $t = 0$, *and for* $\beta > 1$ *one has* $g'_{\beta, A}(0) = 0$.

(iv) $\int_{-A}^{A} g_{\beta, A}^2 = 1$.

(v) *For any* $f \in \Sigma(\beta, 1)$ *with* $f(0) = g_{\beta, A}(0)$,

$$
\int_{-A}^{A} f g_{\beta, A} \geq \int_{-A}^{A} g_{\beta, A}^2 = 1
$$

and, in particular,

$$
\int_{-A}^{A} g_{\beta, A} \geq g_{\beta, A}(0)^{-1}. \tag{5.14}
$$

(vi) *The functions* $g_{\beta, A}$ *are continuous in* $\beta \leq \beta^*$ *and* $u \in [-A, A]$.

PROOF. For the case with $\beta \le 1$, all the statements can be checked directly using the explicit form of the solution $g_{\beta,A}$. Therefore, we are checking only the case with $\beta > 1$.

The first three statements follow immediately from the general results of convex analysis.

Statement (iv) can be proved by renormalization arguments; cf. Donoho and Low (1992). Indeed, let $f \in \Sigma(\beta, 1)$ be such that $\int_{-A}^{A} f^2(t)\,dt < 1$. For $a > 1$, the function $f_a = a^\beta f(a\,\cdot) \in \Sigma(\beta, 1)$ and

$$\int_{-A}^{A} f_a^2(t)\,dt = a^{2\beta-1} \int_{-Aa}^{Aa} f^2(t)\,dt.$$

Since $f \in \Sigma(\beta, 1)$, then, taking $a = 1 + \alpha$ with some positive α small enough, one gets $\int_{-A}^{A} f_a^2 < 1$ but $f_a(0) > f(0)$. This proves (iv).

To check (v), let us take any $f \in \Sigma(\beta, 1)$ with $f(0) = g_{\beta,A}(0)$. Then, for each $\alpha \in [0, 1]$, one has $f_\alpha = (1-\alpha)g_{\beta,A} + \alpha f \in \Sigma(\beta, 1)$. Obviously, $f_\alpha(0) = g_{\beta,A}(0)$ and, arguing as before and using the definition of $g_{\beta,A}$, one obtains $\int_{-A}^{A} f_\alpha^2 \ge 1$, that is,

$$\begin{aligned}
&\int_{-A}^{A} \bigl[(1-\alpha)g_{\beta,A} + \alpha f\bigr]^2 \\
&\quad = \int_{-A}^{A} g_{\beta,A}^2 + 2\alpha \int_{-A}^{A} g_{\beta,A}(f - g_{\beta,A}) + \alpha^2 \int_{-A}^{A} (f - g_{\beta,A})^2 \\
&\quad \ge \int_{-A}^{A} g_{\beta,A}^2.
\end{aligned}$$

This yields for α small that $\int_{-A}^{A} g_{\beta,A}(f - g_{\beta,A}) \ge 0$.

The relation (5.14) is the specialization of this inequality with $f \equiv g_{\beta,A}(0)$.

As claimed in (vi), the continuity of $g_{\beta,A}$ in β follows from the fact that the optimization criterion for the problem $P_\beta[-A, A]$ does not depend on β and the set of constraints is of the form

$$|g(s) - g(t)| \le |s - t|^\beta, \qquad \beta \le 1, \qquad s, t \in [-A, A],$$

$$|g'(s) - g'(t)| \le |s - t|^{\beta-1}, \qquad g'(0) = 0, \qquad \beta \in (1, 2], \qquad s, t \in [-A, A],$$

which again depends on β in a continuous way.

Now we shall check the properties of the kernels (K_h) from (3.5). Conditions (**K1**)–(**K5**) follow directly from Lemma 5.5. To verify (**K6**), we use the following simple fact.

LEMMA 5.6. *Let the kernels (K_h) be defined by* (3.4) *and* (3.5). *Then there exists $c(\varepsilon) \to 0$ as $\varepsilon \to 0$ such that, uniformly in $c \in [c(\varepsilon), 1]$,*

$$\frac{\|K_{h/c}\|}{\|K_h\|} = 1 + o_\varepsilon(1).$$

PROOF. One has directly from the definition (3.4) of $\beta(h)$ that, for each $c \in (0,1)$,

$$\beta(h) - \beta(h/c) = o_\varepsilon(1).$$

This yields the assertion in view of Lemma 5.5(vi). □

The last result reduces (**K6**) to $(K6)$ for the kernels $K_\beta = \lambda_\beta^{-1} g_{\beta,A} \mathbf{1}_{[-A,A]}$, or, equivalently,

$$\int \bigl(g_{\beta,A}\mathbf{1}_{[-A,A]}(u) - c g_{\beta,A}\mathbf{1}_{[-A,A]}(cu)\bigr)^2 du \le \int_{-A}^{A} g_{\beta,A}^2(u)\,du. \tag{5.15}$$

Now evidently $g_{\beta,A}(cu) \in \Sigma(\beta,1)$ for $c \le 1$ and by Lemma 5.5(v)

$$\begin{aligned}
&\int \bigl(g_{\beta,A}\mathbf{1}_{[-A,A]}(u) - c g_{\beta,A}\mathbf{1}_{[-A,A]}(cu)\bigr)^2 du \\
&\quad = (1+c)\int_{-A}^{A} g_{\beta,A}^2(u)\,du - 2c\int_{-A}^{A} g_{\beta,A}(u) g_{\beta,A}(cu)\,du \\
&\quad \le (1-c)\int_{-A}^{A} g_{\beta,A}^2(u)\,du
\end{aligned}$$

and the assertion (5.15) follows.

Now we may apply Theorem 2.5 which guarantees for a function f the accuracy of estimation

$$\psi(h_f, \tilde\varepsilon) = \frac{K_{h_f}(0)}{\|K_{h_f}\|}\frac{\varepsilon d_\varepsilon}{\sqrt{h_f}},$$

where $d_\varepsilon = \sqrt{r\ln(h_{\max}/h_{\min})}$ and h_f is defined by Definition 2.1. Theorem 3.1 will be proved if we show that for each $\tau = (\beta, L)$ and any $f \in \Sigma(\beta, L)$,

$$h_f \ge h(\tau)(1 + o_\varepsilon(1)), \tag{5.16}$$

$$\varphi(\tau, \bar\varepsilon) = \psi(h(\tau), \tilde\varepsilon)(1 + o_\varepsilon(1)). \tag{5.17}$$

Here $h(\tau)$ is defined by (3.3) and $\varphi(\tau,\varepsilon)$ by (3.1), that is,

$$\varphi(\tau, \bar\varepsilon) = g_\beta(0) L^{1/(2\beta+1)} \bigl(\varepsilon\varkappa_\varepsilon\bigr)^{2\beta/(2\beta+1)} = g_\beta(0)\frac{\varepsilon\varkappa_\varepsilon}{\sqrt{h(\tau)}}, \tag{5.18}$$

$$\psi(h(\tau), \tilde\varepsilon) = \frac{K_{h(\tau)}}{\|K_{h(\tau)}\|}\frac{\varepsilon d_\varepsilon}{\sqrt{h(\tau)}}. \tag{5.19}$$

By straightforward calculation,

$$d_\varepsilon = \varkappa_\varepsilon(1 + o_\varepsilon(1)). \tag{5.20}$$

Let $\beta' = \beta(h(\tau))$ be the solution in β of the equation

$$|h(\tau)|^\beta = \frac{\varepsilon \varkappa_\varepsilon}{\sqrt{h(\tau)}}. \tag{5.21}$$

It easily follows that

$$\beta' = \beta(1 + o_\varepsilon(1)). \tag{5.22}$$

Now, using the definition of the kernels (K_h), (3.2), Lemma 5.5(iv) and (vi) and Lemma 5.6, we conclude

$$\frac{K_{h(\tau)}(0)}{\|K_{h(\tau)}\|} = g_{\beta', A}(0) = g_{\beta, A}(0)(1 + o_\varepsilon(1)) = g_\beta(0)(1 + o_{\varepsilon, A}(1)), \tag{5.23}$$

where $o_{\varepsilon, A}(1) \to 0$ as $\varepsilon \to 0$ and $A \to \infty$.

Putting together (5.18)–(5.23), we get (5.17).

It remains to prove (5.16). For this, due to Definition 2.1, we have to check that, given $\tau = (\beta, L)$ and $f \in \Sigma(\beta, L)$, one has, for $h = h(\tau)$ and $\eta < h$,

$$\left|\mathscr{K}_\eta f(t_0) - f(t_0)\right| \le C(h)\sigma(h, \varepsilon d_\varepsilon)(1 + o_\varepsilon(1)). \tag{5.24}$$

As before, we get, for $h = h(\tau)$,

$$C(h) = (K_h(0) - \|K_h\|^2)/\|K_h\| = g_{\beta, A}(0) - \lambda_\beta^{-1}. \tag{5.25}$$

Next

$$\begin{aligned} |\mathscr{K}_\eta f(t_0) - f(t_0)| &= \frac{1}{\eta}\int K_\eta\left(\frac{t - t_0}{\eta}\right)[f(t) - f(t_0)]\,dt \\ &\le \int K_\eta(u)[f(t_0 + u\eta) - f(t_0)]. \end{aligned}$$

Let $\beta' = \beta(\eta)$. One has $\beta' < \beta$ since $\eta < h$.

Note also that for $f \in \Sigma(\beta, L)$ one has $g(u) = (L\eta^\beta)^{-1}[f(t_0 + u\eta) - f(t_0)] \in \Sigma(\beta, 1)$. Hence

$$\begin{aligned} \left|\mathscr{K}_\eta f(t_0) - f(t_0)\right| &\le L\eta^\beta \lambda_{\beta'}^{-1} \int_{-A}^{A} g_{\beta', A}(u)[g(u) - g(0)]\,du \\ &= Lh^\beta(\eta/h)^\beta G(\beta', \beta; A), \end{aligned}$$

where

$$G(\beta', \beta; A) = \sup_{g \in \Sigma(\beta, 1)} \left| \int_{-A}^{A} \lambda_{\beta'}^{-1} g_{\beta', A}(u)[g(u) - g(0)]\,du \right|.$$

Now the required assertion (5.24) follows from (5.20), (5.21), (5.25) and the next technical statement.

LEMMA 5.7. *For any* $A > 0$ *and* $\beta^* \le 2$, *one has*, *uniformly in* $\beta', \beta \in [\beta_*, \beta^*]$, $\beta' < \beta$,

$$G(\beta', \beta; A) \le C < \infty$$

and

$$G(\beta', \beta; A) \to G(\beta, \beta; A) = g_{\beta, A}(0) - \lambda_\beta^{-1}, \qquad \beta' \to \beta. \tag{5.26}$$

PROOF. The first statement follows for $\beta', \beta \le 1$ and for $\beta', \beta \in (1, 2]$ from Lemma 5.5(vi). For $\beta' \le 1$ and $\beta > 1$ we use additionally the fact that for A large enough $g_{\beta', A} = g_{\beta'}$ and

$$\int u g_{\beta'}(u)\, du = 0.$$

Show the equality in (5.26). It can be rewritten as follows. For any $g \in \Sigma(\beta, 1)$,

$$\int_{-A}^{A} g_{\beta, A}(u)[g(0) - g(u)]\, du \le g_{\beta, A}(0) \int_{-A}^{A} g_{\beta, A}(u)\, du - 1.$$

But in this form the statement follows from Lemma 5.5(iv) and (v), since one may assume without loss of generality that $g(0) = g_{\beta, A}(0)$. □

5.4. *Proof of Theorem* 3.2. Theorem 2.6 cannot be applied directly since we are not sure that the function f_1 from this theorem belongs to $\Sigma(\tau_*)$ [with $\tau_* = (\beta_*, L^*)$]. But the idea of the proof remains valid. If the property of compactness of supports for g_β were proved, then we could take $f_0 \equiv 0$, $f_1(t) = (1 - \alpha_\varepsilon)\psi(h_{\min}, \tilde{\varepsilon}) g_\beta^{-1}(0)\, g_\beta((t - t_0)/h_{\min})$ with some small $\alpha_\varepsilon = o_\varepsilon(1)$ and proceed as in the proof of Theorem 2.6. Without the assumption of support compactness, one may use the method from Donoho (1994b) which relies on the solution of a special compactified optimization problem. We omit the details.

Acknowledgments. The authors thank E. Mammen, A. Nemirovskii, A. Tsybakov and M. Neumann for helpful remarks and discussion.

REFERENCES

BRETAGNOLLE, J. and HUBER, C. (1979). Estimation des densites: risque minimax. *Z. Wahrsch. Verw. Gebiete* **47** 119–137.

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. (1996). Asymptotic equivalence of nonparametric regression and white noise. *Ann. Statist.* **24** 2384–2398.

BROWN, L. D. and LOW, M. G. (1992). Superefficiency and lack of adaptability in functional estimation. Technical report, Cornell Univ.

DONOHO, D. L. (1994a). Statistical estimation and optimal recovery. *Ann. Statist.* **22** 238–270.
DONOHO, D. L. (1994b). Asymptotic minimax risk for sup-norm: solution via optimal recovery. *Probab. Theory Related Fields* **99** 145–170.
DONOHO, D. L. and JOHNSTONE, I. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455.
DONOHO, D. L. and JOHNSTONE, I. M. (1992). Minimax estimation via wavelet shrinkage. Unpublished manuscript.
DONOHO, D. L. and JOHNSTONE, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224.
DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1995). Wavelet shrinkage: asymptopia (with discussion)? *J. Royal Statist. Soc. Ser. B* **57** 301–369.
DONOHO, D. L. and LIU, R. C. (1991). Geometrizing rate of convergence. III. *Ann. Statist.* **19** 668–701.
DONOHO, D. L. and LOW, M. G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970.
EFROIMOVICH, S. Y. and LOW, M. G. (1994). Adaptive estimates of linear functionals. *Probab. Theory Related Fields* **98** 261–275.
HALL, P. and JOHNSTONE, I. (1992). Empirical functionals and efficient smoothing parameter selection. *J. Roy. Statist. Soc. Ser. B* **54** 475–530.
HÄRDLE, W. and MARRON, J. S. (1985). Optimal bandwidth selection in nonparametric regression function estimation. *Ann. Statist.* **13** 1466–1481.
IBRAGIMOV, I. A. and KHASMINSKII, R. Z. (1980). Estimates of signal, its derivatives, and point of maximum for Gaussian observations. *Theory Probab. Appl.* **25** 703–716.
IBRAGIMOV, I. A. and KHASMINSKII, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer, New York.
JONES, M. C., MARRON, J. S. and PARK, B. (1991). A simple root-n bandwidth selector. *Ann. Statist.* **19** 1919–1932.
JUDITSKY, A. (1995). Adaptive wavelet estimators. *Math. Methods Statist.* To appear.
KERKYACHARIAN, G. and PICARD, D. (1993). Density estimation by kernel and wavelet method, optimality in Besov space. *Statist. Probab. Lett.* **18** 327–336.
KOROSTELEV, A. P. (1993). Exact asymptotic minimax estimate for a nonparametric regression in the uniform norm. *Theory Probab. Appl.* **38** 737–743.
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** 645–659.
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., MAMMEN, E. and SPOKOINY, V. G. (1997). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. *Ann. Statist.* **25** 929–947.
LOW, M. G. (1992). Renormalizing upper and lower bounds for integrated risk in the white noise model. *Ann. Statist.* **20** 577–589.
MARRON, J. S. (1988). Automatic smoothing parameter selection: a survey. *Empir. Econom.* **13** 187–208.
MÜLLER, H. G. and STADTMÜLLER, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182–201.
NEMIROVSKII, A. (1985). On nonparametric estimation of smooth regression function. *Soviet J. Comput. System Sci.* **23** 1–11.
NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and white noise. *Ann. Statist.* **24** 2399–2430.
SACKS, J. and STRAWDERMAN, W. (1982). Improvements of linear minimax estimates. In *Statistical Decision Theory and Related Topics 3* (S. S. Gupta and J. O. Berger, eds.) **2** 287–304. Academic Press, New York.
STANISWALIS, J. G. S. (1989). Local bandwidth selection for kernel estimates. *J. Amer. Statist. Assoc.* **84** 284–288.

STONE, C. J. (1982). Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10** 1040–1053.
TRIEBEL, H. (1992). *Theory of Function Spaces* **2**. Birkhäuser, Basel.
VIEU, P. (1991). Nonparametric regression: optimal local bandwidth choice. *J. Roy. Statist. Soc. Ser. B* **53** 453–464.

HUMBOLDT UNIVERSITY
SFB 373
SPANDAUER STRASSE 1
10178 BERLIN
GERMANY
E-MAIL: lepski@iaas-berlin.d400.de

WEIERSTRASS INSTITUTE FOR APPLIED ANALYSIS AND STOCHASTICS
MOHRENSTRASSE 39
10117 BERLIN
GERMANY