OPTIMAL POINTWISE ADAPTIVE METHODS IN NONPARAMETRIC ESTIMATION

LEPSKI, O.V. AND SPOKOINY, V.G.

Humboldt University, SFB 373, Spandauer Str. 1, 10178 Berlin

Weierstrass Institute for Applied Analysis and Stochastics,
Mohrenstr. 39, 10117 Berlin

October 1994

1991 *Mathematics Subject Classification.* 62G07; Secondary 62G20.

Key words and phrases. pointwise adaptive estimation, bandwidth selection, Hölder type constraints.

Authors thanks E.Mammen, A.Nemirovskii, A.Tsybakov and M.Neumann for the helpful remarks and discussion.

Abstract. 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 kernel. The resulting estimator is optimal among all feasible estimators.

The important feature of this procedure is that no prior information is used about smoothness properties of the estimated function i.e. the procedure is completely adaptive and "works" for the class of all functions. With it the attainable accuracy of estimation depends on the function itself and it is expressed in terms of "ideal" bandwidth corresponding to this function.

The second procedure can be considered as a specification 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 to choose a family of kernels in an optimal way and the resulting procedure appears to be asymptotically optimal in the adaptive sense.

1. Introduction

The standard minimax nonparametric approach is based on the assumption that the function to be estimated from noisy data belongs to some function (smoothness) class e.g. Hölder, Sobolev, Besov etc. (see Ibragimov and Khasminskii, 1980, 1981, Bretagnolle and Huber, 1976, Stone, 1982, Donoho and Johnstone, 1992b, Kerkyacharian and Picard, 1993). Such kind of assumption is of great importance 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 is the main drawback of the nonparametric approach because typically we don't have any prior information about 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 values of parameters. After that these parameters and the corresponding procedure are chosen automatically by data. We refer to Marron (1988) and Donoho and Johnstone (1992c) for an overview on this topic.

A number of authors considered local selection of a smoothing parameter which seems to be more reasonable for curves with inhomogeneous smoothness properties, see e.g. Marron (1988) and Vieu (1991). We mention the results from Lepski and Spokoiny (1994) where exact asymptotic bounds for the choice of "locality" parameter were established.

Principally another approach was proposed recently in the papers of Donoho, Johnstone, Kerkyacharian and Picard. The idea is to apply the nonlinear wavelet estimation procedure which does not use any prior information about smoothness properties of the function to be estimated and which is near minimax (up to log–factor) for the whole scale of Besov classes (including Hölder and Sobolev ones).

Moreover, for the case if smoothness of an estimated function is measured in weaker norm than the loss, no linear methods provide optimal or near optimal rate, Nemirovski (1985). This case corresponds to estimation of functions with inhomogeneous smoothness properties.

The power of the nonlinear wavelet procedure in such situations can be explained informally that it adapts automatically to the inhomogeneous structure of an estimated function.

This property of the wavelet procedure is called spatial adaptivity and it appeared to be extremely important as from theoretical point of view as for practical applications.

But one important question remains open. Why does the nonlinear wavelet procedure have such nice properties and what is most essential in this method?

An interesting step answering this question was made recently in Lepski, Mammen and Spokoiny(1994) where another procedure, namely, a kernel estimator with a variable data-driven bandwidth was considered and it was shown that this procedure is also spatially adaptive.

The idea of construction was to use the pointwise adaptive procedure i.e. the function is estimated at each point independently and adaptively. For pointwise adaptation a slightly modified version of Lepski (1992) was applied.

The key point of that paper can be stressed as follows. If some estimation procedure is pointwise adaptive, then it is spatially adaptive in the sense of rate optimality over the whole scale of Besov classes.

Studying the problem of pointwise adaptive estimation was initiated in Lepski (1990). In that paper significant difference was shown to be for estimation of the whole function and of a value of a function at one point . More precisely, for the problem of pointwise estimation we meet 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 occured to be $\left(\varepsilon\sqrt{\ln 1/\varepsilon}\right)^{2\beta/(2\beta+1)}$ that differs from the nonadaptive one by the extra log-factor (see also Brown and Low, 1992).

The above mentioned connection of pointwise adaptive estimation problem with the notion of spatial adaptivity motivates further studying the problem of adaptive estimation at a point.

Below we consider two settings for which 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 start from the situation that no assumptions were made a'priori on the smoothness properties of the function f. That means that we do not assume any smoothness conditions on f. We would like to apply a linear kernel estimation procedure. Moreover, we assume a kernel K to be given and only a bandwidth h for kernel estimation is chosen adaptively.

For pointwise adaptation we use the adaptive procedure from Lepski, Mammen and Spokoiny(1994) with a more accurate choice of its parameters.

We prove that this estimation procedure is exact optimal in adaptive sense within 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 non-adaptive pointwise estimation linear (and in particular kernel) methods do not provide asymptotic minimaxity.

The next natural step is to try to optimize a kernel to be used. But any optimal choice of kernel seems to be impossible without some prior assumptions on the structure of the estimated function.

One typical example of such kind is as follows: the function f is assumed to be in some Hölder class but with unknown parameters. The second studied in this paper problem corresponds just to this case. We show that under such a qualitative assumption the optimal choice of kernels can be made which provides optimal accuracy of estimation.

The paper is organizied as follows.

In the next section we formulate the problem of optimal bandwidth selection and present the related results.

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 noise level goes to zero) among the class of all feasible estimators not only of linear kernel type.

2.1. **Preliminaries.** First we precise the problem of adaptive estimation at a point.

Let a function $f(\cdot)$ be estimated from noisy data at a point t_0. To formulate the main results we introduce some characteristic of the function f at the point t_0 which can be treated as an "ideal" bandwidth for the kernel estimation with a given kernel.

The standard bandwidth choice is motivated by the balance relation between the bias and stochastic terms in the decomposition of losses for a kernel estimator. The bias term $b(h)$ 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 used method (in the present context by kernel smoothers). The stochastic term is random and it depends typically on the error level ε (or the number of observations n) and the bandwidth h but not on the function f. For usual kernel estimation procedure the stochastic term is a normal (or asymptotically normal) random variable with zero mean and the variance of order $\sigma^2(h) = \frac{\varepsilon^2}{h}$ (or $\frac{1}{nh}$).

The classical balance equation looks like

$$b_f(h) \asymp \sigma(h). \tag{2.1}$$

But the function f is unknown and hence the bias function $b(h)$ is also unknown. To choose a bandwidth one has to apply some kind of adaptation.

It is of interest to note that the balance rule (2.1) does not work in the pointwise adaptive estimation. This phenomenon was discovered by Lepski (1990), see also Brown and Low (1992). One has to take some majorant for stochastic term to control stochastic fluctuations. Namely, the balance relation

$$b_f(h) \asymp \psi(h)$$

with

$$\psi(h) = \sigma(h)\sqrt{\ln \frac{1}{\varepsilon}} = \frac{\varepsilon\sqrt{\ln 1/\varepsilon}}{\sqrt{h}}$$

allows to estimate adaptively but the corresponding rate includes also such a log-factor.

One can say that this extra log-factor is unavoidable payment for pointwise adaptation. We stress once again that such a property is of great importance: we lose a log at a point but we arrive to rate near optimality in the global estimation.

Moreover, presence of this log-factor allows to neglect the stochastic term and to select the parameters of the estimation procedure in an optimal way. This is just the subject of this section.

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

$$dX(t) = f(t)dt + \varepsilon dW(t). \tag{2.2}$$

Here ε is the level of noise and we assume below that this level is "small", i.e. 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 (2.2) is estimated at a point $t_0 \in (0,1)$.

2.3. **Kernel.** Let now a kernel $K(\cdot)$ be fixed satisfying the following conditions:

$(K1)$ the function $K(u)$ is symmetric i.e. $K(u) = K(-u)$, $u \geq 0$;

$(K2)$ the function $K(\cdot)$ is compactly supported i.e. $K(u) = 0$ for all u outside some compact set C on the real line;

$(K3)$

$$\int K(u)du = 1;$$

$(K4)$

$$\|K\|^2 = \int K^2(u)du < \infty;$$

$(K5)$

$$K(0) > \|K\|^2.$$

Note that no assumptions were made about smoothness properties of the kernel K i.e. it can be even discontinuous.

2.4. **Bandwidth Selection Problem. "Ideal" Bandwidth.** We consider below the family of the kernel estimators $\tilde{f}_h(t_0)$ of the value $f(t_0)$

$$\tilde{f}_h(t) = \frac{1}{h}\int K\left(\frac{t-t_0}{h}\right) dX(t) \tag{2.3}$$

with a positive bandwidth h.

Further we use the following standard decomposition of the loss for the kernel estimators $\tilde{f}_h(t_0)$

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

with the stochastic term $\xi(h)$

$$\xi(h) = \frac{\varepsilon}{h} \int K\left(\frac{t-t_0}{h}\right) dW(t) \tag{2.5}$$

which is obviously a Gaussian zero mean random variable with the variance

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

In what follows we assume that h is small enough and the support of the function $K(\frac{t-t_0}{h})$ contains in $[0,1]$. This assumption allows to neglect the boundary effects and to change integration over $[0,1]$ by integration over the whole real line. That is why we omit the integration limits here in the definition (2.3) and further.

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

$$E\left|\tilde{f}_{\hat{h}}(t_0) - f(t_0)\right|^p$$

where $p \geq 1$ is a given power.

Now we define the notion of an "ideal" bandwidth for the function f.

Denote for $h > 0$

$$\begin{aligned} \mathcal{K}_h f(t_0) &= \frac{1}{h} \int K\left(\frac{t-t_0}{h}\right) f(t) dt, \\ \Delta(h) &= \Delta_f(h) = \sup_{0<\eta<h} \left|\mathcal{K}_h f(t_0) - f(t_0)\right|. \end{aligned} \tag{2.6}$$

The value $\Delta(h)$ characterizes the accuracy of approximation of the function $f(\cdot)$ at the point t_0 by the kernel smoothers $\mathcal{K}_\eta f$ with $\eta \leq h$.

Before to give the definitions we precise the problem of bandwidth selection. We assume that besides the kernel K for each ε two values h_ε and h_ε^* are given such that $h_\varepsilon > \varepsilon^{-2}$, $h_\varepsilon^* \leq 1/2$ and

$$h_\varepsilon / h_\varepsilon^* \to 0, \qquad \varepsilon \to 0. \tag{2.7}$$

We will select a bandwidth h in the interval $h \in [h_\varepsilon, h_\varepsilon^*]$ that is h_ε is the smallest and h_ε^* is the largest admissible values of the bandwidth.

Denote

$$d_\varepsilon = \sqrt{p \ln \frac{h_\varepsilon^*}{h_\varepsilon}}.$$

By (2.7) one has $d_\varepsilon \to \infty$ as $\varepsilon \to 0$.

This factor d_ε enters in the expression of the minimax rate of convergence. In some sense this factor is our payment for adaptation and the less is the range $[h_\varepsilon, h_\varepsilon^*]$ of adaptive bandwidth choice the less is this payment. In any case d_ε is not larger (in order) than $\sqrt{\ln 1/\varepsilon}$ and this is the typical order.

Introduce some more notation. Put

$$\begin{aligned} \psi(h) &= \sigma(h)d_\varepsilon = \frac{\|K\|\varepsilon}{\sqrt{h}}\sqrt{p\ln\frac{h_\varepsilon^*}{h_\varepsilon}}, \qquad (2.8)\\ C(K) &= \frac{K(0)}{\|K\|^2}-1,\\ b(h) &= C(K)\psi(h) = \frac{K(0)-\|K\|^2}{\|K\|}\frac{\varepsilon}{\sqrt{h}}\sqrt{p\ln\frac{h_\varepsilon^*}{h_\varepsilon}}. \qquad (2.9)\end{aligned}$$

Definition 2.1. Let the function $b(h)$ be defined by (2.9) and let given $f(\cdot)$ the function $\Delta(h)=\Delta_f(h)$ be defined by (2.6). The value h_f with

$$h_f = \sup\{h \le h_\varepsilon^* : \Delta(h) \le b(h)\} \qquad (2.10)$$

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

The function $\Delta(h)$ is by definition monotonously increasing and the function $b(h)$ is in the contrary monotonously decreasing with $b(h)\to\infty$ as $h\downarrow 0$. This provides correctness of the definition (2.10).

Now we are ready to formulate the main results.

2.5. **Main Results. 1. Uniform optimality.** Denote by $\mathcal{F}_\varepsilon$ the class of functions $f(\cdot)$ with $h_f \ge h_\varepsilon$,

$$\mathcal{F}_\varepsilon = \{f(\cdot) : h_f \ge h_\varepsilon\}.$$

Put

$$r(h) = b(h)+\psi(h) = \frac{K(0)}{\|K\|}\frac{\varepsilon}{\sqrt{h}}\sqrt{p\ln\frac{h_\varepsilon^*}{h_\varepsilon}}.$$

The first result describes the rate which is attained by the proposed estimator $\hat{f}_\varepsilon$ (for the explicit construction see below).

Theorem 2.1. *Let $K(\cdot)$ be a kernel satisfying the conditions $(K1)-(K5)$ and also the following condition*

$(K6)$

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

Then there exists an estimator $\hat{f}_\varepsilon(t_0)$ which is a kernel estimator with an adaptive bandwidth $\hat{h}$ such that

$$\sup_{f\in\mathcal{F}_\varepsilon} E\left|\frac{\hat{f}_\varepsilon(t_0)-f(t_0)}{r(h_f)}\right|^p = 1+o_\varepsilon(1).$$

Remark 2.1. Here and in what follows we denote by $o_\varepsilon(1)$ some absolute values depending possibly on ε, p 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 i.e. this estimator is asymptotically efficient.

Theorem 2.2. *Let a kernel $K(\cdot)$ satisfy the conditions $(K1)-(K5)$ and also the condition*

$(K7)$

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

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

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

The scope of results of Theorems 2.1 and 2.2 claims asymptotic optimality of the estimator $\hat f_\varepsilon(t_0)$ if the kernel K satisfies the conditions $(K1)-(K7)$.

The question for which kernels these conditions are fulfilled, is discussed in the next section.

Remark 2.2. The first result states the properties of the estimator $\hat h_\varepsilon(t_0)$ which are uniform on the very wide function class $\mathcal{F}_\varepsilon$ whereas the lower bounds result from Theorem 2.2 is stated on the class consisting of two functions. Moreover, we will use $f_0(t)\equiv 0$ and only f_1 depends on ε.

Remark 2.3. It is of interest to observe which accuracy of estimation provides the estimator $\hat f_\varepsilon(t_0)$ from Theorem 2.1 in the usual sense.

Let the function f to be estimated belong to some Hölder class $\Sigma(\beta,L)$ i.e., with $m=\lfloor\beta\rfloor$,

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

If the kernel K has the regularity m that is K is orthogonal to polynomials of degree from 1 till m, then one has easily

$$\Delta(h)=\Delta_f(h)=\sup_{\eta\le h}|\mathcal{K}_\eta f(t_0)-f(t_0)|\le Mh^\beta$$

with some absolute constant M depending only on β,L and $K(\cdot)$. Now the balance equation $\Delta(h_f)=b(h_f)=C(K)\varepsilon d_\varepsilon/\sqrt{h_f}$ arrives to

$$h_f\asymp(\varepsilon d_\varepsilon)^{\frac{2}{2\beta+1}}$$

and

$$r(h_f)\asymp(\varepsilon d_\varepsilon)^{\frac{2\beta}{2\beta+1}}\approx\left(\varepsilon\sqrt{\ln\varepsilon^{-1}}\right)^{\frac{2\beta}{2\beta+1}}$$

where the symbol "$\asymp$" means the equivalence in order. Therefore, and it was expected, the result of Theorem 1 guarantees the suboptimal rate for the pointwise adaptive estimation over the Hölder classes.

Moreover, an optimal kernel choice provides optimal pointwise-adaptive estimation over the Hölder classes but the discussion of this topic is the subject of the second part of this paper.

2.6. **Main Results. 2. Adaptive optimality.** In Lepski (1991, 1992) a new approach to define optimal adaptive rate was proposed for the situation without adaptation, in particular, for the problem of estimation at a point.

Below we present a result in this spirit which is, however, not only of theoretical interest. Some features of this result appear to be extremely important, for instance, for the problem of spatially adaptive estimation.

First we explain in which sense the results of Theorems 2.1 and 2.2 are not completely satisfactory. Due to Theorem 2.2 if we have only two functions, one with the smoothness parameter h_ε, another with h_ε^*, then we have to pay for the adaptation the factor $d_\varepsilon = (p\ln(h_\varepsilon^*/h_\varepsilon))^{1/2}$. Therefore, in the uniform sense in the whole range of adaptation the results of Theorems 2.1 and 2.2 cannot be improved.

But a function f to be estimated may have a smoothness parameter h_f inside the interval $[h_\varepsilon, h_\varepsilon^*]$, and perhaps, an improvement is possible for each particular function. Theorem 2.2 prompts to take a factor $d_\varepsilon' = (p\ln(h_\varepsilon^*/h_f))^{1/2}$ instead of d_ε as in the case if h_f were just the smallest value of bandwidth.

Put for $h \in [h_\varepsilon, h_\varepsilon^*]$

$$\psi'(h) = \sigma(h)\sqrt{1 \vee [p\ln(h_\varepsilon^*/h)]} \tag{2.11}$$

where $a \vee b$ means $\max\{a, b\}$.

Set also

$$\begin{aligned} b'(h) &= C(K)\psi'(h) = \frac{K(0) - \|K\|^2}{\|K\|}\frac{\varepsilon}{\sqrt{h}}\sqrt{1 \vee [p\ln(h_\varepsilon^*/h)]}, \\ r'(h) &= \psi'(h) + b'(h) = \frac{K(0)}{\|K\|}\frac{\varepsilon}{\sqrt{h}}\sqrt{1 \vee [p\ln(h_\varepsilon^*/h)]}. \end{aligned}$$

Define the "ideal bandwidth" for a function f by Definition 2.1 changing in it $b(h)$ by $b'(h)$.

Then we modify correspondingly the bandwidth selector. The properties of the resulting estimator are described by the following result.

Theorem 2.3. *Let a kernel K satisfy the conditions $(K1)$–$(K6)$. Then the estimator $\check{f}_\varepsilon(t_0)$ corresponding to the modified adaptive bandwidth $\check{h}$ provides*

$$\sup_{f\in\mathcal{F}_\varepsilon} E\left|\frac{\check{f}_\varepsilon(t_0) - f(t_0)}{r'(h_f)}\right|^p \le C$$

where C is some absolute constant.

Moreover, for each $f \in \mathcal{F}_\varepsilon$ with $h_f = o_\varepsilon(1)h_\varepsilon^*$ one has

$$E\left|\frac{\check{f}_\varepsilon(t_0) - f(t_0)}{r'(h_f)}\right|^p \le 1 + o_\varepsilon(1).$$

Remark 2.4. First we compare the accuracy provided by the modified adaptive procedure with that of the original one. Obviously $\psi'(h) \le \psi(h)$ and similarly for $b'(h)$. Hence the second statement of the last theorem claims better performance of the modified procedure in each range of adaptation separated away from the upper value h_ε^*.

The first statement is also of great importance. Due to this result, in particular, the modified estimator provides for functions f with $h_f = h_\varepsilon^*$ the non-adaptive rate $\varepsilon/\sqrt{h_\varepsilon^*}$ (since $b'(h_\varepsilon^*) = \text{Const.}\varepsilon/\sqrt{h_\varepsilon^*}$).

This phenomenon was used in Lepski, Mammen and Spokoiny (94) to construct a spatially adaptive procedure which gives the optimal rate for Besov function classes in an integral norm. The crucial point of that paper is that for functions from such classes a pointwise characteristic $h_f(t)$ coincides for "almost all" t with the upper value h_ε^* (because of truncation at this level) where the extra log-factor disappears.

Remark 2.5. The result of Theorem 2 with h_f instead of h_ε guarantees that the properties of modified estimator $\hat{f}_\varepsilon$ cannot be improved, at least, in the asymptotic sense.

To emphasize this fact explicitely, we formulate it as a separate result.

Theorem 2.4. *Let a kernel K satisfy the condition $(K1)$ through $(K5)$ and $(K7)$. If an estimator T_ε of $f(t_0)$ is such that*

$$E_0\left|\frac{T_\varepsilon - f(t_0)}{\varepsilon/\sqrt{h_\varepsilon^*}}\right|^p \le C$$

where C is some absolute constant and E_0 means the expectation for the model (2.2) with the zero function $f \equiv 0$.

Then for any $h = o_\varepsilon(1)h_\varepsilon^*$ there is f with $h_f \ge h$ such that

$$E\left|\frac{T_\varepsilon - f(t_0)}{r'(h_f)}\right|^p \ge 1 - o_\varepsilon(1).$$

Now we give the construction of the adaptive bandwidth $\hat{h}$ and $\check{h}$ and the corresponding estimators.

2.7. **Bandwidth Selector.** Put

$$\delta_\varepsilon = (d_\varepsilon)^{-1} = (p\ln h_\varepsilon^*/h_\varepsilon)^{-1/2} \to 0 \quad \varepsilon \to 0.$$

and define the grid $\mathcal{H}$

$$\mathcal{H} = \{h \in [h_\varepsilon, h_\varepsilon^*] : h = h_\varepsilon(1+\delta_\varepsilon)^k,\ k = 0,1,2,\dots\}. \tag{2.12}$$

Set now

$$\hat{h} = \max\left\{h \in \mathcal{H} : |\tilde{f}_h(t_0) - \tilde{f}_\eta(t_0)| \le (1+2\alpha_\varepsilon)\psi(\eta) \quad \forall \eta < h, \eta \in \mathcal{H}\right\}. \tag{2.13}$$

Here $\tilde{f}_h(t_0)$ is defined by (2.3) and

$$\alpha_\varepsilon = (d_\varepsilon)^{-1/3} = (p\ln h_\varepsilon^*/h_\varepsilon)^{-1/6} \to 0, \quad \varepsilon \to 0.$$

Finally put

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

For the modified selectors $\check{h}$ we use the same procedure with $\psi(\eta)$ changed by $\psi'(\eta)$,

$$\check{h} = \max\left\{h \in \mathcal{H} : |\tilde{f}_h(t_0) - \tilde{f}_\eta(t_0)| \le (1+2\alpha_\varepsilon)\psi'(\eta) \quad \forall \eta < h, \eta \in \mathcal{H}\right\}. \tag{2.14}$$

where δ_ε and α_ε are as before.

The modified estimator uses just this new bandwidth selector,

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

Notice that the number of elements in the grid $\mathcal{H}$ is of order $\ln 1/\varepsilon$ and the definition of the adaptive bandwidth $\hat{h}$ is based on comparison of the corresponding number of kernels estimators.

2.8. **Kernel Choice. Conditions on kernels.** Here we discuss briefly some aspects of the choice of the kernel $K(\cdot)$ and the range $[h_\varepsilon, h_\varepsilon^*]$.

Strictly speaking, Theorems 2.1 – 2.3 can be applied only for the kernels satisfying the conditions $(K1)$ – $(K7)$. Note, however, that the procedure makes sense for any kernel under $(K1)$–$(K4)$ and due to Lepski, Mammen and Spokoiny (1994) it will be spatially adaptive. It is of interest to see what can be obtained by the methods from Theorems 2.1 and 2.2 if the conditions $K(5)$ – $(K7)$ are not fulfilled.

Analysis of the proofs allows to extract the following. Denote

$$\begin{aligned} s_1^2 &= s_1^2(K) = \sup_{0<c\le 1} \int |K(u) - cK(uc)|^2 du/\|K\|^2, \\ C_1(K) &= \sup_{0<c\le 1} \left| K(0) - \int K(u)K(uc)du \right|. \end{aligned}$$

Put now

$$\begin{aligned} \psi_1(h) &= s_1\psi(h), \\ b_1(h) &= C_1(K)\psi(h). \end{aligned}$$

and define the "ideal" bandwidth h_f and the adaptive bandwidth $\hat{h}_1$ similarly to above but with $\psi_1(h), b_1(h)$ instead of $\psi(h), b(h)$. Now Theorem 2.1 remains valid with $r_1(h_f) = \psi_1(h_f) + b_1(h_f)$ instead of $r(h)$.

But Theorem 2.2 gives the same lower bound, namely with rate $r(h_f)$. This means that the corresponding estimator $\tilde{f}_{\hat{h}_1}$ has the efficiency ρ with

$$\rho = \frac{r_1(h_f)}{r(h_f)} = \frac{1 + C(K)}{s_1 + C_1(K)}. \tag{2.15}$$

Another interesting question is the optimization of the kernel K. The first insight prompts to take a kernel which minimizes the ratio $K(0)/\|K\|^2$ because this expression enters in the value of the minimax risk. But it is not true. This value enters also in the definition of the "ideal" bandwidth and smaller values of this ratio lead to smaller values of the corresponding "ideal" bandwidth and hence a smaller accuracy.

More detail analysis shows the inverse conclusion. The optimal kernel is produced by the optimization subproblem: to maximize the value $K(0)/\|K\|^2$ in the 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^* of the mentioned optimization subproblem satisfies automatically the condition $(K6)$ and $(K7)$.

One more important question for the kernel choice is a proper kernel regularity.

We see from Theorems 2.1–2.3 that the attainable accuracy of estimation is closely related to the accuracy of approximation of a given function by its kernel smoothers. To provide good rate of approximation for large bandwidth values one has to take kernels of high regularity.

More precise, the following conclusion is motivated by the theoretical results from the previous section: If we take a kernel of regularity $m \ge 1$ i.e. the kernel K is orthogonal to polynomials of lower degree, then the upper bound h_ε^* of the adaptation range can be taken about $\varepsilon^{\frac{2}{2m+1}}$ (or $n^{-\frac{1}{2m+1}}$). Indeed, the regularity m of the kernel $K(\cdot)$ provides good rate of approximation just for bandwidth up to this value.

Nevertheless, for practical calculation consideration of larger bandwidth values seems to be also reasonable.

The lower bound h_ε is recommended to take as small as possible. For the considered abstract "white noise" model this bound is of order ε^2.

For more realistic models (see below Section 4) this choice is restricted by reasons of the experiments equivalence. The mentioned there results of Brown and Low (1990) and Nussbaum (1993) prompt to take for h_ε the value of order $h_\varepsilon \asymp \varepsilon^{\frac{2}{1+1/2}} = \varepsilon^{4/3}$ corresponding to the smoothness parameter $1/2$.

2.9. **Nested Kernels.** Now we consider one generalization of the considered above problem. Namely we study the situation if one takes different kernels for different bandwidth values.

This idea is quite natural since small values of bandwidth correspond to low regularity functions and no necessity to take high-order kernels.

More precise, we assume that a system (net) of kernels $\mathbf{K} = (K_h, h > 0)$ depending possible on ε is given. The considered above case of a fixed kernel corresponds to $K_h(\cdot) = K(\cdot)$. Similarly to above we impose some conditions on these kernels:

(**K1**) the functions $K_h(u)$ are symmetric i.e. $K_h(u) = K_h(-u)$, $u \ge 0$;

(**K2**) the system of functions $K_h(\cdot)$ is compactly supported i.e. $K_h(u) = 0$ for all h and all u outside some compact set C on the real line;

(**K3**)

$$\int K_h(u)du = 1;$$

(**K4**)

$$\begin{aligned} \sup_h \|K_h\|^2 &= \sup_h \int K_h^2(u)du < \infty, \\ \inf_h \|K_h\|^2 &> 0; \end{aligned}$$

(**K5**) Set

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

Then there exist two positive constants C_1, C_2 such that

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

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

(**K6**) Uniformly in h

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

(**K7**) Uniformly in h

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

Similarly to above, consider the family of kernel estimators

$$\tilde{f}_h(t_0) = \frac{1}{h} \int K_h \left(\frac{t - t_0}{h} \right) dX(t). \tag{2.16}$$

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

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

Let again an interval $[h_\varepsilon, h_\varepsilon^*]$ be given with

$$h_\varepsilon^*/h_\varepsilon \to \infty.$$

and we choose a bandwidth in this range.

Denote similarly to above

$$\begin{aligned}
\mathcal{K}_h f(t_0) &= \frac{1}{h} \int K_h \left(\frac{t - t_0}{h} \right) f(t) dt, \\
\Delta(h) &= \Delta_f(h) = \sup_{0<\eta<h} \left| \mathcal{K}_h f(t_0) - f(t_0) \right|. \\
d_\varepsilon &= \sqrt{2p \ln \frac{\sigma(h_\varepsilon)}{\sigma(h_\varepsilon^*)}}, && (2.17) \\
\psi(h) &= \sigma(h) d_\varepsilon, && (2.18) \\
b(h) &= C(h)\psi(h) = (K_h(0) - \|K_h\|^2)\|K_h\|^{-2}\sigma(h) d_\varepsilon, && (2.19) \\
r(h) &= \psi(h) + b(h) = K_h(0)\|K_h\|^{-2}\sigma(h) d_\varepsilon. && (2.20)
\end{aligned}$$

Keep now the definitions of h_f from the previous section i.e.

$$h_f = \sup\{h \le h_\varepsilon^* : \Delta(h) \le b(h)\}. \tag{2.21}$$

Finally define the adaptive bandwidth $\hat{h}$ as above with modifications made.

The method of the proofs of Theorems 2.1 and 2.2 can be extended without any changes on the considered situation. We formulate the corresponding results for the reference convenience taking into account further applications for the problem of estimation over Hölder classes.

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

$$\sup_{f \in \mathcal{F}_\varepsilon} E \left| \frac{\hat{f}_\varepsilon(t_0) - f(t_0)}{r(h_f)} \right|^p \le 1 + o_\varepsilon(1).$$

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

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

In conclusion we present the version of Theorems 2.3 and 2.4 for the case of nested kernels $\mathbf{K}$.

We use again the notation $\sigma(h) = \varepsilon \|K_h\|/\sqrt{h}$.

Put for $h \in [h_\varepsilon, h_\varepsilon^*]$

$$\psi'(h) = \sigma(h)\sqrt{1 \vee \left(2p \ln \frac{\sigma(h)}{\sigma(h_\varepsilon^*)}\right)}. \tag{2.22}$$

Keep further the definition (2.16) and change in (2.19)–(2.20) and in the definition of the adaptive bandwidth $\hat{h}$ and the estimator $\hat{f}_\varepsilon(t_0)$ the value $\psi(h)$ with $\psi'(h)$ from (2.22).

The properties of the resulting bandwidth $\check{h}$ and the estimator $\check{f}_\varepsilon(t_0)$ are described by the following results.

Theorem 2.7. *Let a system of kernels* $\mathbf{K} = (K_h)$ *satisfy the conditions* $(\mathbf{K1})$–$(\mathbf{K6})$. *Then the estimator* $\check{f}_\varepsilon(t_0)$ *corresponding to the adaptive bandwidth* $\check{h}$ *with the modification (2.22) provides*

$$\sup_{f \in \mathcal{F}_\varepsilon} E \left| \frac{\check{f}_\varepsilon(t_0) - f(t_0)}{r'(h_f)} \right|^p \le C$$

where $r'(h) = K(0)\|K\|^{-2}\psi'(h)$ *and* C *is some absolute constant.*

For each $f \in \mathcal{F}_\varepsilon$ *with* $h_f = o_\varepsilon(1)h_\varepsilon^*$

$$E \left| \frac{\check{f}_\varepsilon(t_0) - f(t_0)}{r'(h_f)} \right|^p \le 1 + o_\varepsilon(1).$$

Theorem 2.8. *Let the conditions* $(\mathbf{K1})$–$(\mathbf{K5})$ *and* $(\mathbf{K7})$ *be fulfilled.*

Let estimators T_ε *of* $f(t_0)$ *satisfy the condition*

$$\lim_{\varepsilon \to 0} E_0 \left| \frac{T_\varepsilon}{\varepsilon / \sqrt{h_\varepsilon^*}} \right|^p \le C$$

where E_0 *means the expectation for the model (2.2) with the zero signal* $f(\cdot) \equiv 0$. *Then for any* $h = o_\varepsilon(1)h_\varepsilon^*$ *there exists a function* $f(\cdot)$ *with* $h_f \ge h$ *such that*

$$E \left| \frac{T_\varepsilon - f(t_0)}{r(h_f)} \right|^p \ge 1 - o_\varepsilon(1).$$

3. Estimation under Hölder type constraints

In the present section we consider the problem of pointwise adaptive estimation for the model (2.2) under the qualitative assumption that the function f belongs to some Hölder class $\Sigma(\beta, L)$. Given β, L this is the set of functions f such that, with $m = \lfloor \beta \rfloor$,

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

We deal with the situation when the parameters β, L are unknown.

Surprisingly, this adaptation can be performed in an optimal way and the result presented below describe the optimal adaptive procedure and the optimal attainable accuracy.

First we 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 parameters β is of the most importance for us. It can be qualified as an expected number of bounded derivatives of the function f. For the Lipschitz constant L we need actually only the qualitative assumption that it is separated away from zero and infinity. Apparently the results and the procedure can be formulated in 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 (1994), Donoho and Low (1992), Donoho (1994a,1994b):

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

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

Remark 3.1. The explicit solution g_β and the value $\operatorname{val}(P_\beta) = g_\beta(0)$ are known only for $\beta \le 1$ from Korostelev(1994). Set

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

Then

$$g_\beta(t) = a f_\beta(bt)$$

where the constants a, b are defined by

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

In particular,

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

The case $\beta > 1$ is much more difficult and in our knowledge only for $\beta = 2$ the solution g_β admits some explicit description.

Some more qualitative properties of the functions g_β are discussed below in this section.

As in the case of the bandwidth selection problem we present two kinds of results and two procedures. The first ones are uniform on the whole range of adaptation and the second ones describe optimal result for each particular β, L and relate to the notion of adaptive optimality.

First we fomulate the results, the procedures are described below.

In our results we assume that $\beta^* \le 2$. Possible extentions on the case of arbitrary β^* are discussed in the next section.

Theorem 3.1. *Let estimators $\hat{T}_\varepsilon$ of $f(t_0)$ be described below. Uniformly in $\beta \in [\beta_*, \beta^*]$ and $L \in [L_*, L^*]$*

$$\sup_{f \in \Sigma(\beta, L)} E \left| \frac{\hat{T}_\varepsilon - f(t_0)}{r(\beta, L)} \right|^p \le 1 + o_\varepsilon(1) \tag{3.1}$$

where

$$r(\beta, L) = g_\beta(0)\, L^{\frac{1}{2\beta+1}} \left[p \left(\frac{2}{2\beta_* + 1} - \frac{2}{2\beta^* + 1} \right) \right]^{\frac{\beta}{2\beta+1}} \left(\varepsilon \sqrt{\ln \frac{1}{\varepsilon}} \right)^{\frac{2\beta}{2\beta+1}} \tag{3.2}$$

and $g_\beta(0) = \operatorname{val}(P_\beta)$ is the solution to (P_β).

Remark 3.2. The value $r(\beta, L)$ can be treated as the accuracy attainable by the estimator $\hat{T}_\varepsilon$ on the Hölder class $\Sigma(\beta, L)$. We see that this accuracy has the near optimal rate $\left(\varepsilon\sqrt{\ln\frac{1}{\varepsilon}}\right)^{\frac{2\beta}{2\beta+1}}$ which is worse than the non-adaptive one with the extra log factor.

Notice also that the parameters β, L have different influence on the accuracy of estimation. Growing β means smoothness increasing and the corresponding accuracy becomes better. Growing L means in the contrary worser smoothness properties and worser accuracy.

In particular, the whole range of adaptation is described as follows: the lower point is (β_*, L^*) and the upper point is (β^*, L_*).

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 exists a function $f_\varepsilon \in \Sigma(\beta_*, L^*)$ such that for any estimator T_ε of $f(t_0)$*

$$\max\left\{E_0\left|\frac{T_\varepsilon}{r(\beta^*, L_*)}\right|^p, E_{f_\varepsilon}\left|\frac{T_\varepsilon - f_\varepsilon(t_0)}{r(\beta_*, L^*)}\right|^p, \right\} \geq 1 - o_\varepsilon(1)$$

where E_0 means the expectation for the model (2.2) with the signal f equal to zero, and E_{f_ε} corresponds to the case $f = f_\varepsilon$.

The next two results describe the optimal attainable accuracy of adaptive estimation if we use one universal estimator for all values of β, L in the interval of adaptation but the perfomance is studied on a fixed class $\Sigma(\beta, L)$.

First we explaim the properties of the estimators $\check{T}_\varepsilon$, see below.

Theorem 3.3. *There exists a constant C depending possibly only on β_*, β^* and L_*, L^* such that uniformly in $\beta \in [\beta_*, \beta^*]$ and $L \in [L_*, L^*]$*

$$\sup_{f\in\Sigma(\beta,L)} E\left|\frac{\hat{T}_\varepsilon - f(t_0)}{r'(\beta, L)}\right|^p \leq C. \tag{3.3}$$

Moreover, for each $\delta > 0$ uniformly in $\beta \in [\beta_*, \beta^* - \delta]$ and $L \in [L_*, L^*]$

$$\sup_{f\in\Sigma(\beta,L)} E\left|\frac{\check{T}_\varepsilon - f(t_0)}{r'(\beta, L)}\right|^p \leq 1 + o_\varepsilon(1) \tag{3.4}$$

where

$$r'(\beta, L) = g_\beta(0)\, L^{\frac{1}{2\beta+1}} \max\left\{1, p\left(\frac{2}{2\beta+1} - \frac{2}{2\beta^*+1}\right)\ln\frac{1}{\varepsilon}\right\}^{\frac{\beta}{2\beta+1}} \varepsilon^{\frac{2\beta}{2\beta+1}}. \tag{3.5}$$

The assertion of Theorem 3.2 with β instead of β^* claims optimality of the result 3.4 for each particular β, L. But we formulate this statement as a separate result because of its importance in the considered context.

Theorem 3.4. *If an estimator T_ε provides*

$$E_0\left|\frac{T_\varepsilon}{\varepsilon^{2\beta^*/(2\beta^*+1)}}\right|^p \leq C$$

with some absolute constant C *, then for any* $\beta < \beta^*$ *and any* $L > 0$

$$\sup_{f \in \Sigma(\beta,L)} E \left| \frac{T_\varepsilon - f(t_0)}{r'(\beta, L)} \right|^p \geq 1 - o_\varepsilon(1) \tag{3.6}$$

Now we describe the estimation rule.

3.1. **Estimation Procedures.** We present two procedures which differs slightly from each other. The first one corresponds to Theorem 3.1 and it performs uniformly on the whole interval of adaptation.

The properties of the second procedure are described by Theorem 3.3.

The both procedures are specifications of the procedures from the previous section with special choice of kernels K.

The construction of these kernels is closely related to the solutions g_β to the problems (P_β) from above. Roughly speaking, kernels K_β are obtained by normalization from g_β to provide $\int K_\beta = 1$.

Unfortunately the functions g_β were not stated generally to be compactly supported and, in particular, from nothing follows that the integral $\int g_\beta$ is finite.

Apparently this values do not enter in the answer and the desirable kernels are defined by a proper truncation.

Define the modification of the problem (P_β) under support constraints, Donoho (1994a): Given $A > 0$,

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

One has easily $\mathrm{val}(P_\beta) \leq \mathrm{val}(P_\beta[-A, A])$ and we use also the property (Donoho, 94a, Lemma 6.1)

$$\mathrm{val}(P_\beta[-A, A]) \to \mathrm{val}(P_\beta), \qquad A \to \infty. \tag{3.7}$$

Moreover, by methods of 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 large enough to provide the asymptotics in (3.7) for all $\beta \in [\beta_*, \beta^*]$.

Denote by $g_{\beta,A}$ the solution to $(P_\beta[-A, A])$. For more information about behavior of the functions $g_{\beta,A}$ see Lemma 3.1 below.

To apply the procedure from above section we have to state correspondence between bandwidth h and the smoothness parameters β, L.

Denote

$$\kappa = \sqrt{p \left(\frac{2}{2\beta_* + 1} - \frac{2}{2\beta^* + 1} \right)}$$

and given β, L set

$$h(\beta, L) = \left(\frac{\kappa \varepsilon \sqrt{\ln 1/\varepsilon}}{L} \right)^{\frac{2}{2\beta+1}} \tag{3.8}$$

and put

$$\begin{aligned} h_\varepsilon^* &= h(\beta^*, L_*), \\ h_\varepsilon &= h(\beta_*, L^*). \end{aligned}$$

Finally define the function $h(\beta)$ as the solution in β of the equation

$$h^{\beta} = \frac{\kappa\varepsilon\sqrt{\ln 1/\varepsilon}}{\sqrt{h}},$$

that is,

$$h(\beta) = \frac{\ln(\kappa\varepsilon\sqrt{\ln 1/\varepsilon})}{\ln h} - 1/2. \tag{3.9}$$

Finally we define $\hat{T}_\varepsilon$ as the estimator $\hat{f}_\varepsilon(t_0)$ from the previous section with the system of kernels $\mathbf{K} = (K_h)$ where we use for $h \in [h_\varepsilon, h_\varepsilon^*]$ and $\beta = \beta(h)$

$$K_h = \lambda_\beta^{-1} g_{\beta,A} \mathbf{1}_{[-A,A]} \tag{3.10}$$

with $\lambda_\beta = \int_{-A}^{A} g_{\beta,A}(t)dt$.

The second procedure is defined in the similar way with a few modifications. Namely we set

$$\begin{aligned} d(\beta) &= \max\left\{1, p\left(\frac{2}{2\beta+1} - \frac{2}{2\beta^*+1}\right)\ln 1/\varepsilon\right\}^{1/2} \\ h'(\beta, L) &= \left(\frac{\varepsilon d(\beta)}{L}\right)^{\frac{2}{2\beta+1}}, \\ h_\varepsilon^* &= h'(\beta^*, L_*), \\ h_\varepsilon &= h'(\beta_*, L^*), \end{aligned}$$

and let the function $\beta'(h)$ be the solution in β of the equation

$$h^{\beta+1/2} = \varepsilon d(\beta).$$

With these corrections the estimator $\check{T}_\varepsilon$ is just $\check{f}_\varepsilon(t_0)$ from above.

3.2. **Proof of Theorems 3.1, 3.3.** We deduce Theorems 3.1, 3.3 as corollaries of Theorems 2.5 – 2.7. For this we have to check the conditions (**K1**) through (**K6**) for the kernels (K_h) and to verify that the results of Theorems 2.5, 2.7 provide just the accuracy necessary for Theorems 3.1, 3.3.

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

Lemma 3.1. *Let A be arbitrary positive and $\beta^* \le 2$. The following stastements are fulfilled for each $\beta \le \beta^*$:*

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

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

(ii) *the function $g_{\beta,A}$ is symmetric, i.e. $g_{\beta,A}(t) = g_{\beta,A}(-t)$, $t \in R^1$;*

(iii) *the function $g_{\beta,A}$ has maximum at $t = 0$ and for $\beta > 1$*

$$g'_{\beta,A}(0) = 0;$$

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

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

and particulary

$$\int_{-A}^{A} g_{\beta,A} \ge g_{\beta,A}(0)^{-1}; \tag{3.11}$$

(v) *Functions* $g_{\beta,A}$ *are continuous in* $\beta \le \beta^*$ *and* $u \in [-A, A]$.

Proof. The first four statements follow immediately from the general results of convex analysis. Show, for instance, (iv).

Let $f \in \Sigma(\beta, 1)$ and $f(0) = g_{\beta,A}(0)$. Then for each $\alpha \in [0, 1]$ one has $(1-\alpha)g_{\beta,A} + \alpha f \in \Sigma(\beta, 1)$. Now by definition of $g_{\beta,A}$

$$\int_{-A}^{A} [(1-\alpha)g_{\beta,A} + \alpha f]^2 = \int_{-A}^{A} g_{\beta,A}^2 + 2\alpha \int_{-A}^{A} g_{\beta,A}(f - g_{\beta,A}) + \alpha^2 \int_{-A}^{A} f^2 \ge \int_{-A}^{A} g_{\beta,A}^2.$$

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

The relation (3.11) is the specification of the proved above with $f \equiv g_{\beta,A}(0)$.

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

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

that again depends on β in a continuous way. □

Now we check the properties of the kernels (K_h) from (3.10). The conditions (**K1**)–(**K5**) follows directly from Lemma 3.1. To state (**K6**) we use the following simple fact.

Lemma 3.2. *Let the kernels* (K_h) *be defined by (3.9), (3.10). 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 (3.9) of $\beta(h)$ that for each $c \in (0, 1)$

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

This yields the assertion through (v) of Lemma 3.1. □

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

$$\int (g_{\beta,A}\mathbf{1}_{[-A,A]}(u) - c g_{\beta,A}\mathbf{1}_{[-A,A]}(cu))^2 du \le \int_{-A}^{A} g_{\beta,A}^2(u) du. \tag{3.12}$$

Now evidently $g_{\beta,A}(cu) \in \Sigma(\beta,1)$ for $c \le 1$ and by (iv) of Lemma 3.1

$$\int (g_{\beta,A}\mathbf{1}_{[-A,A]}(u) - c g_{\beta,A}\mathbf{1}_{[-A,A]}(cu))^2 du = (1+c)\int_{-A}^{A} g_{\beta,A}^2(u)du +$$
$$+ 2c\int_{-A}^{A} g_{\beta,A}(u) g_{\beta,A}(cu)du \le (1-c)\int_{-A}^{A} g_{\beta,A}^2(u)du$$

and the assertion (3.12) follows.

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

$$r(h_f) = \frac{K_{h_f}(0)}{\|K_{h_f}\|}\frac{\varepsilon\sqrt{p\ln h_\varepsilon^*/h_\varepsilon}}{\sqrt{h_f}}$$

where h_f is defined by Definition 2.1. Theorem 3.1 will be proved if we show that for each β, L and any $f \in \Sigma(\beta, L)$

$$\begin{aligned} h_f &\ge h(\beta,L)(1+o_\varepsilon(1)) & (3.13)\\ r(\beta,L) &= r(h(\beta,L))(1+o_\varepsilon(1)). & (3.14)\end{aligned}$$

Here $h(\beta,L)$ is defined by (3.8) and $r(\beta,L)$ by (3.2), that is

$$\begin{aligned} r(\beta,L) &= g_\beta(0)L^{1/(2\beta+1)}\left(\kappa\varepsilon\sqrt{\ln 1/\varepsilon}\right)^{2\beta/(2\beta+1)} = g_\beta(0)\frac{\kappa\varepsilon\sqrt{\ln 1/\varepsilon}}{h(\beta,L)}, & (3.15)\\ r(h(\beta,L)) &= \frac{K_{h(\beta,L)}}{\|K_{h(\beta,L)}\|}\frac{\varepsilon\sqrt{p\ln(h_\varepsilon^*/h_\varepsilon)}}{\sqrt{h(\beta,L)}}. & (3.16)\end{aligned}$$

By direct calculation

$$\sqrt{p\ln\frac{h_\varepsilon^*}{h_\varepsilon}} = \sqrt{p\left(\frac{2}{2\beta_*+1} - \frac{2}{2\beta^*+1}\right)\ln\frac{1}{\varepsilon}}\,(1+o_\varepsilon(1)) = \kappa\sqrt{\ln 1/\varepsilon}\,(1+o_\varepsilon(1)). \tag{3.17}$$

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

$$|h(\beta,L)|^\beta = \frac{\kappa\varepsilon\sqrt{\ln 1/\varepsilon}}{h(\beta,L)}. \tag{3.18}$$

Easily

$$\beta' = \beta(1+o_\varepsilon(1)). \tag{3.19}$$

Now using the definition of the kernels (K_h), (3.7), (i) and (v) of Lemma 3.1 and Lemma 3.2 we conclude

$$\frac{K_{h(\beta,L)}(0)}{\|K_{h(\beta,L)}\|} = g_{\beta',A}(0) = g_{\beta,A}(0)(1+o_\varepsilon(1)) = g_\beta(0)(1+o_{\varepsilon,A}(1)) \tag{3.20}$$

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

Putting together (3.15) – (3.20) we get (3.14).

It remains to prove (3.13). For this due to Definition 2.1 we have to check that, given β, L and $f \in \Sigma(\beta, L)$, one has with $h = h(\beta,L)$ and $\eta < h$

$$|\mathcal{K}_\eta f(t_0) - f(t_0)| \le \frac{K_h(0) - \|K_h\|^2}{\|K_h\|}\frac{\varepsilon\sqrt{p\ln(h_\varepsilon^*/h_\varepsilon)}}{h}(1+o_\varepsilon(1)). \tag{3.21}$$

As above, we get for $h = h(\beta, L)$

$$(K_h(0) - \|K_h\|^2)/\|K_h\| = g_{\beta,A}(0) - \lambda_\beta^{-1}. \tag{3.22}$$

Next

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

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

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

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 (3.21) follows from (3.17), (3.18), (3.22) and the next technical statement.

Lemma 3.3. *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

$$G(\beta', \beta; A) \to G(\beta, \beta; A) = g_{\beta,A}(0) - \lambda_\beta^{-1}, \qquad \beta' \to \beta. \tag{3.23}$$

Proof. The first statement follows for $\beta', \beta \le 1$ and for $\beta', \beta \in (1, 2]$ from (v) of Lemma 3.1. 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 (3.23). 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 (i) and (iv) of Lemma 3.1 since without loss of generality one may assume $g(0) = g_{\beta,A}(0)$. □

3.3. **Proof of Theorem 3.2.** We cannot apply directly Theorem 2.6 since the function f_1 from the latter theorem does not belong necessary to $\Sigma(\beta, L)$ (with $\beta = \beta_*$ and $L = L^*$.

But the idea of the proof remains valid and the choice $f_0 \equiv 0$, $f_1(t) = (1-\alpha_\varepsilon)\frac{r(h_\varepsilon)}{g_\beta(0)} g_\beta((t-t_0)/h_\varepsilon)$ provides the assertion of the theorem. We omit the details since they are literally the same that of in the proof of Theorem 2.2, see Section 5 below.

4. Further Developments

4.1. **Other Nonparametric Statistical Models.** In this paper we concentrate ourselves on the simplest "white noise" model (2.2). This type of models allows to emphasize 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 nongaussian 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(1990), Low (1992) and Nussbaum (1993) 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 motivate applicability of Theorems 2.1 – 2.3 for these models.

4.2. **Other type of constraints.** In Section 4 we considered the Hölder type constraints. But, of course, other kinds can be considered as well. We mention only the case of Sobolev type constraints as a reasonable one in the present context. This type of constraints means that β-th generalized derivative of the estimated function is bounded by a constant L, β, L being unknown.

All the consideration from above remains valid with the obvious modifications. Only the family of optimization problems (P_β) should be considered with the corresponding type of constraints.

4.3. **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 (1992), Donoho (1994b), Efroimovich and Low (1994). The corresponding result show close relation between the particular problem of pointwise estimation and a general problem for an arbitrary linear functional.

We conjeucture that all consideration from above can be extended in a similar way on the general case.

4.4. **The case $\beta^* > 2$.** The fact $\beta^* \le 2$ was used essentially in the proof of Theorems 3.1 through 3.4, in particular, for the proof of important Lemma 3.1.

For the case $\beta^* > 2$ the statements of Theorems 3.1 or 3.3 cannot be extended directly from the considered case $\beta^* \le 2$ because the structure of Hölder classes is not embedded: $\Sigma(\beta', 1)$ does not belongs to $\Sigma(\beta, 1)$ for $\beta' < \beta$.

It can be illustrate explicitly on the first statement of Lemma 3.3 where one has easily $G(\beta', \beta; A) = \infty$, for instance, if $\beta' = 1$ and $\beta = 3$ since $\Sigma(3, 1)$ contains all linear functions.

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

In this case we need only a generalization of Lemma 3.1 on the considered situation. But further discussion of this matter beyond the scope of this article.

5. Proofs

In the present section we prove Theorems 2.1 and 2.2. Necessary corrections for proofs of Theorems 2.3 through 2.8 are obvious and omitted.

5.1. **Proof of Theorem 2.1.** Denote by N_ε the number of elements in the grid $\mathcal{H}$ (see (2.12)) and show that this value is of logarithmic order i.e. we choose between logarithmic number of possible values for the bandwidth h.

Lemma 5.1. *One has for the grid* $\mathcal{H}$

$$N_\varepsilon = \#\mathcal{H} \le 2d_\varepsilon^3 = 2\left(p \ln h_\varepsilon^*/h_\varepsilon\right)^{3/2}.$$

Proof. The definition of $\mathcal{H}$ provides

$$N_\varepsilon \le \frac{\ln h_\varepsilon^*/h_\varepsilon}{\ln(1+\delta_\varepsilon)} \le \frac{2d_\varepsilon^2}{\delta_\varepsilon} \le 2d_\varepsilon^3.$$

Here we used that δ_ε is small for ε small and hence $\ln(1+\delta_\varepsilon) \ge \delta_\varepsilon/2$. □

Let us fix some function f from $\mathcal{F}_\varepsilon$ and let $h_f, \hat{h}$ be defined as above. By definition of $\mathcal{F}_\varepsilon$ we have $h_f \ge h_\varepsilon$.

Without loss of generality we assume that $h_f \in \mathcal{H}$. Otherwise we can change h_f by the closest from below point of $\mathcal{H}$ and the result of Theorem 2.1 remains valid.

Recall that the definition of h_f provides for each $h \le h_f$ the inequality

$$|\mathcal{K}_h f(t_0) - f(t_0)| \le b(h_f) = C(K)\psi(h_f). \tag{5.1}$$

The following notation will be helpful below. Put for each $h \in \mathcal{H}$

$$\mathcal{H}(h) = \{h' \in \mathcal{H} : h' \le h\}, \qquad \mathcal{H}^-(h) = \{h' \in \mathcal{H} : h' < h\}.$$

One has

$$\begin{aligned} E\left|\hat{f}_\varepsilon(t_0) - f(t_0)\right|^p &= E\left|\hat{f}_\varepsilon(t_0) - f(t_0)\right|^p \mathbf{1}(\hat{h} \ge h_f) + \\ &+ E\left|\hat{f}_\varepsilon(t_0) - f(t_0)\right|^p \mathbf{1}(\hat{h} < h_f) = \\ &= R_\varepsilon^+ + R_\varepsilon^-. \end{aligned}$$

Now through the definition of $\hat{h}$, the decomposition (2.4) and (5.1) we find

$$\begin{aligned} R_\varepsilon^+ &\le E\left(|\tilde{f}_{\hat{h}}(t_0) - \tilde{f}_{h_f}(t_0)| + |\tilde{f}_{h_f}(t_0) - f(t_0)|\right)^p \mathbf{1}(\hat{h} \ge h_f) \le \\ &\le E\left((1+2\alpha_\varepsilon)\psi(h_f) + b(h_f) + |\xi(h_f)|\right)^p \mathbf{1}(\hat{h} \ge h_f). \end{aligned}$$

If now

$$r'(h) = (1+2\alpha_\varepsilon)\psi(h) + b(h) = r(h) + 2\alpha_\varepsilon\psi(h)$$

and $q(h)$ is an arbitrary positive function, then

$$\begin{aligned} R_\varepsilon^+ \le\ & (r'(h_f) + q(h_f))^p\, P(\hat{h} \ge h_f) + \\ & + E\left(r'(h_f) + |\xi(h_f)|\right)^p \mathbf{1}(\xi(h_f) \ge q(h_f)). \end{aligned}$$

Similarly for R_ε^-

$$\begin{aligned} R_\varepsilon^- &= \sum_{h\in\mathcal{H}^-(h_f)} E\left|\hat{f}_\varepsilon(t_0)-f(t_0)\right|^p \mathbf{1}(\hat{h}=h) \le \\ &\le \sum_{h\in\mathcal{H}^-(h_f)} E\left(b(h_f)+|\xi(h)|\right)^p \mathbf{1}(\hat{h}=h) \le \\ &\le \sum_{h\in\mathcal{H}^-(h_f)} \left(b(h_f)+q(h)\right)^p P(\hat{h}=h) + \\ &\quad + \sum_{h\in\mathcal{H}^-(h_f)} E\left(b(h_f)+|\xi(h)|\right)^p \mathbf{1}(|\xi(h)|>q(h)). \end{aligned}$$

Getting together the estimates for R_ε^+ and for R_ε^- we obtain

$$\begin{aligned} R_\varepsilon &\le (r'(h_f)+q(h_f))^p P(\hat{h}\ge h_f) + \\ &\quad + \sum_{h\in\mathcal{H}^-(h_f)} \left(b(h_f)+q(h)\right)^p P(\hat{h}=h) + \\ &\quad + |r'(h_f)|^p \sum_{h\in\mathcal{H}(h_f)} E\left(1+\left|\frac{\xi(h)}{r'(h_f)}\right|\right)^p \mathbf{1}(|\xi(h)|>q(h)). \end{aligned} \tag{5.2}$$

First we consider the last sum in this expression.

Lemma 5.2.

$$S_3 = \sum_{h\in\mathcal{H}(h_f)} E\left(1+\left|\frac{\xi(h)}{r'(h_f)}\right|\right)^p \mathbf{1}(|\xi(h)|>q(h)) = o_\varepsilon(1). \tag{5.3}$$

Proof. By (2.5) $\xi(h)\asymp\mathcal{N}(0,\sigma^2(h))$. Thus

$$\begin{aligned} &E\left(1+\left|\frac{\xi(h)}{r(h_f)}\right|\right)^p \mathbf{1}(|\xi(h)|>q(h)) \le \\ &\le 2^{p-1}E\left(1+\left|\frac{\sigma(h)}{r(h_f)}\right|^p|\zeta|^p\right)\mathbf{1}(|\zeta|>d(h)) \le \\ &\le 2^{p-1}\left(1+C(p)\left|\frac{\sigma(h)}{r(h_f)}\right|^p\right)\exp\{-\frac{1}{2}d^2(h)\}. \end{aligned}$$

Here ζ means a standard normal random variable, $d(h)=q(h)/\sigma(h)$ and we used that for a large enough

$$\begin{aligned} P(|\zeta|>a) &\le \exp\{-a^2/2\}, \\ E|\zeta|^p\mathbf{1}(|\zeta|>a) &\le C(p)\exp\{-a^2/2\} \end{aligned}$$

with some absolute constant $C(p)$ depending only on p.

Now we choose the values $q(h)$ and hence $d(h)$ in a proper way. Set for $h\le h_f$

$$d(h) = \sqrt{p\ln\frac{h_f}{h}+2d_\varepsilon^{1/3}} = \sqrt{2p\ln\frac{\sigma(h)}{\sigma(h_f)}+2d_\varepsilon^{1/3}}, \tag{5.4}$$

$$q(h) = \sigma(h)d(h), \tag{5.5}$$

where recall $d_\varepsilon=\sqrt{p\ln h_\varepsilon^*/h_\varepsilon}$.

Now $\#\mathcal{H}(h_f) \le \#\mathcal{H} = N_\varepsilon$ and

$$\begin{aligned} S_3 &\le \sum_{h\in\mathcal{H}(h_f)} 2^{p-1}\left(1 + C(p)\left|\frac{\sigma(h)}{r(h_f)}\right|^p\right)\exp\{-\frac{1}{2}2p\ln\frac{\sigma(h)}{\sigma(h_f)} - d_\varepsilon^{1/3}\} \le \\ &\le N_\varepsilon 2^{p-1}\exp\{-d_\varepsilon^{1/3}\}\left(1 + C(p)\left|\frac{\sigma(h_f)}{r(h_f)}\right|^p\right). \end{aligned}$$

By definition

$$\sigma(h_f)/r(h_f) = \frac{1+C(K)}{d_\varepsilon} = o_\varepsilon(1)$$

and Lemma 5.1 provides

$$N_\varepsilon \le 2d_\varepsilon^3.$$

The assertion of the lemma follows now from

$$d_\varepsilon^3\exp\{-d_\varepsilon^{1/3}\} = o_\varepsilon(1).$$

□

Next we estimate the second term in (5.2).

One has by the definitions (5.4) – (5.5) that $q(h_f) = \sigma(h_f)d(h_f) = \sigma(h_f)\sqrt{2d_\varepsilon^{1/3}}$ and

$$\begin{aligned} S_1 &= (r'(h_f) + q(h_f))^p\, P(\hat{h} \ge h_f) = \\ &= r^p(h_f)\left(1 + 2\alpha_\varepsilon + \frac{\sqrt{2d_\varepsilon^{1/3}}}{d_\varepsilon}\right) P(\hat{h} \ge h_f) = \\ &= r^p(h_f)\,(1 + o_\varepsilon(1))\, P(\hat{h} \ge h_f). \end{aligned} \tag{5.6}$$

It remains to estimate S_2 with

$$\sum_{h\in\mathcal{H}^-(h_f)} (b(h_f) + q(h))^p\, P(\hat{h} = h). \tag{5.7}$$

Define the value h_1 by the equality

$$2b(h_f) = \alpha_\varepsilon\psi(h_1) = d_\varepsilon^{1/3}\psi(h_1). \tag{5.8}$$

Of course, $h_1 < h_f$. Denote

$$\begin{aligned} \mathcal{H}^-(h_1) &= \{h \in \mathcal{H} : h < h_1\}, \\ \mathcal{H}^+(h_1) &= \{h \in \mathcal{H} : h_1 \le h < h_f\}. \end{aligned}$$

We split the summation in (5.7) into two parts over $\mathcal{H}^-(h_1)$ and $\mathcal{H}^+(h_1)$. Before to estimate these two sums we state some simple properties of the functions $q(h)$ and $d(h)$ from (5.4) – (5.5).

Lemma 5.3. *For each* $h \in \mathcal{H}(h_f)$

$$d(h) \le d_\varepsilon(1 + o_\varepsilon(1)). \tag{5.9}$$

For each $h \in \mathcal{H}^+(h_1)$ *and* ε *small enough*

$$q(h) \le b(h_f)d_\varepsilon^{-1/3}. \tag{5.10}$$

Proof. One has

$$
\begin{aligned}
d(h) &= \sqrt{p \ln \frac{h_f}{h} + 2d_\varepsilon^{1/3}} \le \\
&\le \sqrt{p \ln \frac{h_f}{h} + 2\sqrt{d_\varepsilon^{1/3}}} \le \\
&\le d_\varepsilon + \sqrt{2d_\varepsilon^{1/3}}
\end{aligned}
$$

and (5.9) follows.

The condition (5.8) implies for $h \in \mathcal{H}^+(h_1)$

$$\psi(h) \le 2b(h_f)d_\varepsilon^{1/3}.$$

Now

$$
\begin{aligned}
q(h) &= \sigma(h)d(h) = \\
&= \psi(h)\frac{d(h)}{d_\varepsilon} \le \\
&\le 2b(h_f)d_\varepsilon^{1/3}\frac{d(h)}{d_\varepsilon}
\end{aligned}
$$

and using the definitions $b(h) = C(K)\psi(h) = C(K)\sigma(h)d_\varepsilon$ we have

$$
\begin{aligned}
d(h) &= \sqrt{2p \ln \frac{\sigma(h)}{\sigma(h_f)} + 2d_\varepsilon^{1/3}} = \\
&= \sqrt{2p \ln \frac{\psi(h)C(K)}{b(h_f)} + 2d_\varepsilon^{1/3}} \le \\
&\le \sqrt{2p \ln (C(K)d_\varepsilon^{1/3}) + 2d_\varepsilon^{1/3}} \le \\
&\le d_\varepsilon^{1/3}/2
\end{aligned}
$$

if ε is small enough.

This yields

$$q(h) \le b(h_f)2d_\varepsilon^{1/3}\frac{d_\varepsilon^{1/3}}{2d_\varepsilon} = b(h_f)d_\varepsilon^{-1/3}.$$

□

The result (5.10) allows to get

$$
\begin{aligned}
S_2^+ &= \sum_{h \in \mathcal{H}^+(h_1)} (b(h_f) + q(h))^p \, P(\hat{h} = h) \le \\
&\le \left(b(h_f) + b(h_f)d_\varepsilon^{-1/3}\right)^p P\left(\hat{h} \in \mathcal{H}^+(h_1)\right) = \\
&= b^p(h_f)\,(1 + o_\varepsilon(1))\, P\left(\hat{h} \in \mathcal{H}^+(h_1)\right). \qquad (5.11)
\end{aligned}
$$

It remains to estimate the sum

$$S_2^- = \sum_{h\in\mathcal{H}^-(h_1)} (b(h_f) + q(h))^p \, P(\hat{h} = h).$$

Lemma 5.4. *For each* $h \in \mathcal{H}^-(h_1)$

$$P(\hat{h} = h) \le 2N_\varepsilon \exp\{-\frac{1+2\alpha_\varepsilon}{2} d_\varepsilon^2\}.$$

Proof. Let us fix some $h \in \mathcal{H}^-(h_1)$ and denote $h_+ = h(1+\delta_\varepsilon)$. Obviously h_+ is the next after h element of the grid $\mathcal{H}$. We use also the notation $\mathcal{H}^-(h) = \{\eta \in \mathcal{H} : \eta < h\}$.

The definition of $\hat{h}$ yields

$$P(\hat{h} = h) \le \sum_{\eta\in\mathcal{H}^-(h)} P\left(|\tilde{f}_\eta(t_0) - \tilde{f}_{h_+}(t_0)| > (1+2\alpha_\varepsilon)\psi(\eta)\right).$$

Since $h < h_f$, then $\eta, h_+ \le h_f$ and through (5.1)

$$\begin{aligned} |\mathcal{K}_\eta f(t_0) - f(t_0)| &\le b(h_f), \\ \left|\mathcal{K}_{h_+} f(t_0) - f(t_0)\right| &\le b(h_f). \end{aligned}$$

Hence

$$|\tilde{f}_\eta(t_0) - \tilde{f}_{h_+}(t_0)| \le 2b(h_f) + |\xi(\eta) - \xi(h_+)|.$$

But $h \in \mathcal{H}^-(h_1)$ and thus

$$2b(h_f) \le \alpha_\varepsilon \psi(h) \le \alpha_\varepsilon \psi(\eta).$$

Notice also that

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

i.e. this difference is normal $\mathcal{N}(0, \sigma^2(\eta, h_+))$ with

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

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

The condition $(K6)$ provides

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

and we arrive to the following estimate

$$\begin{aligned} P(\hat{h} = h) &\le \sum_{\eta\in\mathcal{H}^-(h)} P\left(|\xi(\eta) - \xi(h_+)| > (1+\alpha_\varepsilon)\psi(\eta)\right) = \\ &= \sum_{\eta\in\mathcal{H}^-(h)} P\left(|\zeta| > (1+\alpha_\varepsilon) d_\varepsilon \frac{\sigma(\eta)}{\sigma(\eta, h_+)}\right) \le \\ &\le \sum_{\eta\in\mathcal{H}^-(h)} P\left(|\zeta| > (1+\alpha_\varepsilon) d_\varepsilon\right) \le \\ &\le N_\varepsilon \exp\{-\frac{1}{2}(1+\alpha_\varepsilon)^2 d_\varepsilon^2\} \le \\ &\le N_\varepsilon \exp\{-\frac{1+2\alpha_\varepsilon}{2} d_\varepsilon^2\}. \end{aligned}$$

The lemma is proved. □

One has evidently $q(h) \le q(h_\varepsilon)$ for $h \in \mathcal{H}^-(h_1)$ and applying the result of the last lemma we obtain

$$\begin{aligned} S_2^- &\le \sum_{h\in\mathcal{H}^-(h_1)} (b(h_f)+q(h_\varepsilon))^p P(\hat h = h) \le \\ &\le 2N_\varepsilon^2 (b(h_f)+q(h_\varepsilon))^p \exp\{-\frac{1+2\alpha_\varepsilon}{2} d_\varepsilon^2\} \le \\ &\le 2^p N_\varepsilon^2 (b^p(h_f)+\sigma^p(h_\varepsilon)d^p(h_\varepsilon)) \exp\{-\frac{1+2\alpha_\varepsilon}{2} d_\varepsilon^2\} \le \\ &\le 2^p N_\varepsilon^2 (b^p(h_f)+\sigma^p(h_\varepsilon)d^p(h_\varepsilon)) \exp\left\{-p\ln\frac{\sigma(h_\varepsilon)}{\sigma(h_\varepsilon^*)} - \alpha_\varepsilon p \ln \frac{h_\varepsilon^*}{h_\varepsilon}\right\} \le \\ &\le 2^p N_\varepsilon^2 \left(\left| b(h_f)\frac{\sigma(h_\varepsilon^*)}{\sigma(h_\varepsilon)}\right|^p + |d(h_\varepsilon)\sigma(h_\varepsilon^*)|^p \right) \exp\{-p\alpha_\varepsilon d_\varepsilon^2\}. \end{aligned}$$

Using Lemma 5.1, the first statement of Lemma 5.3 and that $b(h_f) \le b(h_\varepsilon) \le C(K)\sigma(h_\varepsilon)d_\varepsilon$ we conclude

$$S_2^- \le 2^{p+2} d_\varepsilon^6 \, |\psi(h_\varepsilon^*)(1+C(K))|^p \exp\{-p d_\varepsilon^{2-1/3}\} = \psi^p(h_\varepsilon^*)\, o_\varepsilon(1).$$

Combining this estimate with (5.2), (5.3) (5.6) and (5.11) we obtain

$$\begin{aligned} R_\varepsilon &\le r^p(h_f)(1+o_\varepsilon(1)) \Big(P(\hat h \ge h_f) + P(h_1 \le \hat h < h_f) + o_\varepsilon(1) \Big) \le \\ &\le r^p(h_f)(1+o_\varepsilon(1)). \end{aligned}$$

Theorem 2.1 is proved.

5.2. **Proof of Theorem 2.2.** Define

$$f_0(t) \equiv 0$$

and

$$\begin{aligned} f_1(t) &= (1-\alpha_\varepsilon)\frac{\varepsilon}{\sqrt{h_\varepsilon}\|K\|}\sqrt{p\ln\frac{h_\varepsilon^*}{h_\varepsilon}}\, K\left(\frac{t-t_0}{h_\varepsilon}\right) = \\ &= (1-\alpha_\varepsilon)\frac{r(h_\varepsilon)}{K(0)}\, K\left(\frac{t-t_0}{h_\varepsilon}\right) \end{aligned} \tag{5.12}$$

where

$$\alpha_\varepsilon = d_\varepsilon^{-1/2} = (p\ln h_\varepsilon^*/h_\varepsilon)^{-1/4} = o_\varepsilon(1).$$

It is obvious that

$$h_{f_0} = h_\varepsilon^*.$$

Next we show that

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

In fact, for each $\eta < h_\varepsilon$ one has by $(K7)$

$$\begin{aligned}
|\mathcal{K}_\eta f_1(t_0) - f_1(t_0)| &= \left| \frac{1}{\eta} \int K\left(\frac{t-t_0}{\eta}\right) [f_1(t) - f_1(t_0)]\, dt \right| = \\
&= v_\varepsilon \left| \frac{1}{\eta} \int K\left(\frac{t-t_0}{\eta}\right) \left[K\left(\frac{t-t_0}{h_\varepsilon}\right) - K(0) \right] dt \right| = \\
&= v_\varepsilon \left| \frac{1}{\eta} \int K(u) \left[K(uc) - K(0) \right] dt \right| \le \\
&\le v_\varepsilon \left(K(0) - \|K\|^2 \right)
\end{aligned}$$

where

$$v_\varepsilon = (1-\alpha_\varepsilon) \frac{r(h_\varepsilon)}{K(0)}$$

and $c = \eta / h_\varepsilon \le 1$. This gives for f_1

$$\begin{aligned}
\Delta(h_\varepsilon) &= \Delta_{f_1}(h_\varepsilon) \le \\
&\le (1-\alpha_\varepsilon) r(h_\varepsilon) \frac{K(0) - \|K\|^2}{K(0)} = \\
&= (1-\alpha_\varepsilon) b(h_\varepsilon)
\end{aligned}$$

that means $h_{f_1} \ge h_\varepsilon$.

Let the measures $P_{0,\varepsilon}$ and $P_{1,\varepsilon}$ correspond to the model (2.2) with the functions f_0 and f_1 respectively.

It is clear that these measures are Gaussian. Moreover, by Girsanov's theorem

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

where

$$\begin{aligned}
q_\varepsilon^2 &= \varepsilon^{-2} \int f_1^2(t) dt, \qquad (5.13) \\
\zeta_\varepsilon &= \frac{\varepsilon^{-1}}{q_\varepsilon} \int f_1(t) dX(t),
\end{aligned}$$

and

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

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

$$\liminf_{\varepsilon \to 0} R_\varepsilon = 1 \qquad (5.14)$$

where

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

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

Note that

$$\frac{f_1(t_0)}{r(h_\varepsilon)} = 1 - \alpha_\varepsilon$$

and denote

$$\begin{aligned} \theta_\varepsilon &= \frac{T_\varepsilon}{r(h_\varepsilon)(1-\alpha_\varepsilon)}, \\ D_\varepsilon &= \frac{r(h_\varepsilon)}{r(h_\varepsilon^*)} = \sqrt{h_\varepsilon^*/h_\varepsilon}. \end{aligned}$$

With this notation

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

Now (5.14) is equivalent to

$$\liminf_{\varepsilon\to 0} \max\left\{D_\varepsilon^p E_{0,\varepsilon}|\theta_\varepsilon|^p, E_{1,\varepsilon}|1-\theta_\varepsilon|^p\right\} \ge 1.$$

Further, due to (5.12) and (5.13)

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

and

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

as $\varepsilon \to 0$.

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

Lemma 5.5. *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

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

and $q_\varepsilon \to \infty$.

Let then numbers D_ε be such that

$$\frac{1}{q_\varepsilon}(p\ln D_\varepsilon - \frac{1}{2}q_\varepsilon^2) \to \infty. \tag{5.15}$$

Then for any estimator θ_ε such that

$$\liminf_{\varepsilon\to 0} D_\varepsilon^p E_{0,\varepsilon}|\theta_\varepsilon|^p \le C < \infty \tag{5.16}$$

one has

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

Proof. Fix any estimators θ_ε satisfying (5.16). Take then an arbitrary $M > 0$ and denote

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

where C is from the condition (5.16). This condition yields for ε small enough

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

and

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

Denote

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

One has

$$Z_\varepsilon = \exp\left\{-q_\varepsilon\zeta_\varepsilon + \frac{1}{2}q_\varepsilon^2\right\} = \exp\left\{-q_\varepsilon(\zeta_\varepsilon - q_\varepsilon) - \frac{1}{2}q_\varepsilon^2\right\}$$

where by Girsanov's theorem

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

and hence

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

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

Put now

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

and introduce events

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

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

$$\begin{aligned} Z_\varepsilon &\ge \exp\{-q_\varepsilon M - \frac{1}{2}q_\varepsilon^2\} \\ |1-\theta_\varepsilon| &\ge 1-\delta_\varepsilon, \end{aligned}$$

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

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

The condition (5.15) implies

$$\begin{aligned} D_\varepsilon^p \delta_\varepsilon^p \exp\{-q_\varepsilon M - \frac{1}{2}q_\varepsilon^2\} &= \exp\{p \ln D_\varepsilon - \frac{1}{2}q_\varepsilon^2 - pq_\varepsilon - Mq_\varepsilon\} = \\ &= \exp\left\{q_\varepsilon\left[\frac{1}{q_\varepsilon}(p\ln D_\varepsilon - \frac{1}{2}q_\varepsilon^2) - p - M\right]\right\} \\ &\to \infty, \qquad \varepsilon \to 0. \end{aligned} \tag{5.17}$$

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 through (5.17). But if $P(A_\varepsilon^c) \leq 2P(B_\varepsilon^c) = 2\bar{\Phi}(M)$, then

$$R_\varepsilon \geq P(A_\varepsilon)(1-\delta_\varepsilon)^p - 1/M \geq (1 - 2\bar{\Phi}(M))|1 - \delta_\varepsilon|^p - 1/M.$$

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. □

References

[1] Bretagnolle,J. and Huber,C. (1979). Estmation des densites: Risque minimax *Z. Wahrsch. Verw.Gebiete* **47** 119–137.

[2] Brockmann, M., Gasser, T. and Herrmann, E. (1993). Locally adaptive bandwidth choice for kernel regression estimators. *J. Amer. Statist. Assoc.* **88** 1302–1309.

[3] Brown, L.D. and Low, M.G. (1990). Asymptotic equivalence of nonparametric regression and white noise. *Technical Report*, Cornell University.

[4] Brown, L.D. and Low, M.G. (1992). Superefficiency and lack of adaptability in functional estimation. *Technical Report*, Cornell University.

[5] Delyon, B. and Juditsky, A. (1994). Wavelet estimators, global error measures, revisited. *Technical Report*, IRISA, Rennes.

[6] Donoho,D.L. and Johnstone,I.M. (1992a). Ideal spatial adaptation by wavelet shrinkage *Technical Report 400* Dep. of Statistics, Stanford University.

[7] Donoho,D.L. and Johnstone,I.M. (1992b). Minimax estimation via wavelet shrinkage *Technical Report 402* Dep. of Statistics, Stanford University.

[8] Donoho,D.L. and Johnstone,I.M. (1992c). Adapting to Unknown Smoothness via Wavelet Schrinkage. Dep. of Statistics, Stanford University.

[9] Donoho, D.L., Johnstone, I.M., Kerkyacharian, G. and Picard, D. (1994). Wavelet shrinkage: asymptopia? *J. Royal Statist. Soc.*, to appear.

[10] Donoho,D.L. and Liu,R.C. (1991). Geometrizing rate of convergence,III. *Ann. Statist.* **19** 668–701.

[11] Donoho,D.L. and Low,M.G. (1992). Renormalization exponents and optimal pointwise rates of convergence. *Ann. Statist.* **20** 944–970.

[12] Donoho,D.L. (1994a) Statistical Estimation and Optimal Recovery. *Ann. Statist.* **22** 238–270.

[13] Donoho,D.L. (1994b) Asymptotic minimax risk for sup-norm: Solution via optimal recovery. *Probab. Theory and Rel. Fields* **99** 145–170.

[14] Efroimovich, S.Y. and Low,M.G.(1994) Adaptive Estimates of Linear Functionals (unpublished manuscript)

[15] Gijbels, I. and Mammen, E. (1994). On local adaptivity of kernel estimates with plug-in local bandwidth selectors. Preprint, Sonderforschungsbereich 373, Humboldt Universität, Berlin.

[16] Hall,P. and Johnstone,I.(1992) Empirical functionals and efficient smoothing parameter selection. *J. Roy. Statist. Soc.* **B 54** 475–530.

[17] Härdle,W. and Marron,J.S.(1985). Optimal bandwidth selection in nonparametric regression function estimation *Ann. Statist.* **12** 1466–1481.

[18] Ibragimov,I.A. and Khasminskii,R.Z. (1980). Estimates of signal, its derivatives, and point of maximum for Gaussian observations *Theory Probab. Appl.* **25** N.4 703–716.

[19] Ibragimov,I.A. and Khasminskii,R.Z. (1981). *Statistical Estimation: Asymptotic Theory* Springer, New York.

[20] Jones,M.C., Marron,J.S. and Park,B. (1991) A simple root-n bandwidth selector *Ann. Statist.* **19**, 1919–1932.

[21] Kerkyacharian,G. and Picard,D.(1993). Density estimation by kernel and wavelet method, optimality in Besov space *Statistics and Probability Letters* **18** 327–336.

[22] Korostelev,A.P. (1994) Exact asymptotic minimax estimate for a nonparametric regsression in the uniform norm. *Theory Probab. Appl.* **39** ??

[23] Lepski,O.V.(1990). One problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** N.3 459–470.

[24] Lepski,O.V.(1991). Asymptotic minimax adaptive estimation. 1. Upper bounds. *Theory Probab. Appl.* **36** No.4, 645–659.

[25] Lepski,O.V.(1992). Asymptotic minimax adaptive estimation. 2. Statistical model without optimal adaptation. Adaptive estimators *Theory Probab. Appl.* **37** N.3, 468–481.

[26] Lepski, O.V. and Spokoiny, V.G. (1994). Local adaptivity to inhomogeneous smoothness. 1. Resolution level. Preprint, Institut für Angewandte Analysis und Stochastik, Berlin, Germany.

[27] Lepski, O.V., Mammen, E. and Spokoiny, V.G. (1994). Ideal spatial adaptation to inhomogeneous smoothness: an approach based on kernel estimates with variable bandwidth selection. (submitted in *Ann. Statist.*)

[28] Low,M.G.(1992). Renormalizing upper and lower bounds for integrated risk in the white noise model *Ann.Statist.* **21** 577–589.
[29] Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Ann. Statist.* **19** 741–759.
[30] Marron,J.S.(1988) Automatic smoothing parameter selection: a survey. *Empir. Econ.* **13**, 187–208.
[31] Müller, H.G. and Stadtmüller, U. (1987). Variable bandwidth kernel estimators of regression curves. *Ann. Statist.* **15** 182–201.
[32] Nemirovski, A. (1985) On nonparametric estimation of smooth regression function. *Sov. J. Comput. Syst. Sci* **23** (6): 1–11.
[33] Nussbaum, M. (1993). Asymptotic equivalence of density estimation and white noise. *Technical Report*, Institute of Applied Analysis and Stochastics, Berlin.
[34] Poljak,B.T. and Tsybakov,A.B. (1990). Asymptotic optimality of C_p-test for the orthogonal series estimation of regression *Theory Probab. Appl.* **35** N.2, 293–306
[35] 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, New York.
[36] Sacks,J. and Ylvisaker,D.(1981) Asymptotically optimum kernels for density estimation at a point. *Ann. Statist.* **9** 334–346.
[37] Staniswalis, J.G.S. (1989). Local bandwidth selection for kernel estimates. *J. Amer. Statist. Assoc.* **84** 284–288.
[38] Stone,C.J.(1982). Optimal global rates of convergence for nonparametric regression *Ann. Statist.* **10** N.4 1040–1053.
[39] Vieu,P.(1991) Nonparametric Regression: Optimal Local Bandwidth Choice. *J.R.Statist.Soc.* **B 53** No.2, 453–464.

Humboldt University, SFB 373, Spandauer Str. 1, 10178 Berlin

Weierstrass Institute for Applied Analysis and Stochastics, Mohrenstr. 39, 10117 Berlin, Germany, and Institute for Information Transmission Problems, Ermolovoy 19, Moscow 101447, Russia