# Weierstraß–Institut
# für Angewandte Analysis und Stochastik

im Forschungsverbund Berlin e.V.


# Bootstrap confidence bands in nonparametric regression


Michael H. Neumann



submitted: 14th July 1994



Weierstrass Institute
for Applied Analysis
and Stochastics
Mohrenstrasse 39
D – 10117 Berlin
Germany



Preprint No. 107
Berlin 1994



*1991 Mathematics Subject Classification.* Primary 62G07, secondary 62G09, 62G15.
*Key words and phrases.* Nonparametric regression, confidence bands, bootstrap, local linear estimator.

Edited by
Weierstraß–Institut für Angewandte Analysis und Stochastik (WIAS)
Mohrenstraße 39
D — 10117 Berlin
Germany

Fax: + 49 30 2004975
e–mail (X.400): c=de;a=d400;p=iaas-berlin;s=preprint
e–mail (Internet): preprint@iaas-berlin.d400.de

# Bootstrap confidence bands in nonparametric regression

M. H. Neumann
Weierstrass Institute for Applied
Analysis and Stochastics
Mohrenstr. 39
Berlin, 10117, Germany

ABSTRACT. In the present paper we construct asymptotic confidence bands in nonparametric regression. Our assumptions admit unequal variances of the observations and nonuniform, possibly considerably clustered design. The confidence band is based on an undersmoothed local linear estimator, and an appropriate quantile is obtained via the wild bootstrap proposed by Härdle and Mammen (1990). We derive certain rates (in the sample size $n$) for the error in coverage probability, which is an improvement of existing results for methods that rely on the asymptotic distribution of the maximum of some Gaussian process. We propose a practicable rule for a data-dependent choice of the bandwidth.

*1991 Mathematics Subject Classification.* Primary 62G07; secondary 62G09, 62G15.
*Key words and phrases.* Nonparametric regression, confidence bands, bootstrap, local linear estimator.

## 1. Introduction

Whenever we have a nonparametric curve estimate, confidence bands are an important means to get an impression about the accuracy that can be expected for the particular estimator. Such bands seem to be much more informative than pointwise confidence intervals, which are also a major direction of research, when one has to decide if some feature of the estimated curve should be considered as structure of the unknown function or should be explained due to random fluctuations of the estimate. There already exists a long list on previous attempts on this subject, most of them are mentioned in the bibliography in Eubank and Speckman (1993). Much work was stimulated due to a paper by Bickel and Rosenblatt (1973), who primarily derived confidence bands for kernel density estimators, but provided additionally a useful technical result on the distribution of the maximum of certain Gaussian processes, which are stationary after centering, and serve as limit processes of the deviation process of kernel estimators if the sample size tends to infinity.

In the random design model, Liero (1982) for the Nadaraya-Watson kernel estimator, Johnston (1982) for the Yang estimator and Härdle (1989) for $M$-smoothers established confidence bands based on the limiting distribution of the deviation process. There exist similar results by Major (1973) for histogram estimators, Révész (1979) and Bjerve, Doksum, Yandell (1985) for nearest neighbor estimators. All of these authors used undersmoothing to make the effect of bias negligible.

A different approach was used in Knafl, Sacks, Ylvisaker (1982) and Hall, Titterington (1988), who constructed conservative confidence bands without undersmoothing, but on the basis of the prior knowledge of upper bounds for the roughness of the regression $m$.

Bootstrap methods were used in this context by Härdle, Bowman (1988) for pointwise confidence intervals and Härdle, Marron (1991) for the construction of a fixed number of simultaneous error bars. Bootstrap techniques were also proposed by Faraway and Jhun (1990) in density estimation and by Faraway (1990) in regression with i.i.d. errors for bandwidth choice and construction of confidence intervals. However, there was no rigorous result proved for the performance of confidence bands. An interesting comparison of the small sample behaviour of various methods was made by Loader (1993).

The latest development in this area that came to our attention is the paper by Eubank and Speckman (1993). These authors argued that methods which rely on undersmoothing are difficult to apply in practice, since there does not exist any natural guideline how to define an asymptotically undersmoothed bandwidth in a reasonable way for a fixed sample size $n$. Instead of pure undersmoothing they produced an estimator with asymptotically negligible bias by a two-step method due to adding a bias corrector to the initial estimator. It turns out that the estimators at both stages can be furnished with natural, MSE-optimal bandwidths, which makes the application of usual bandwidth selectors possible.

In the present paper we start with a fixed design model as Eubank and Speckman (1993) did, and we improve some of the shortcomings of that paper that were already mentioned by these authors. In particular, we admit heteroscedastic errors and nonuniform design, which result in a considerably nonstationary process as limit of the deviation process of our estimator. In view of the possibly considerably irregular design we apply the local linear estimator proposed by Fan (1992). It was shown in that paper that local linear estimators share the advantages of the Nadaraya-Watson estimator and the Gasser-Müller estimator both for random and regular nonuniform design. Another important improvement of the method of Eubank and Speckman is, that we also include the boundary region of the estimator, which can be quite large in practical applications with finite sample size.

We do not know if we can appropriately modify our equally sized confidence band to apply exact asymptotic results as given in Bickel, Rosenblatt (1973) or Qualls, Watanabe (1972) for essentially stationary Gaussian processes to determine a proper quantile in our situation. Instead, we apply the wild bootstrap proposed by Härdle, Mammen (1990), which was already implicitly contained in Wu (1986), to find an appropriate quantile for the error process. In distinction to all of the abovementioned papers we are able to derive certain rates (in $n$) for the decay of the error in coverage probability, which seems to be a strong argument in favor of our new method. Moreover, we conjecture that one could state for the approach based on the limiting process only an asymptotically vanishing error in coverage probability, but without any algebraic rate.

The treatment of the bias problem is essentially by undersmoothing, but we propose a practicable rule to determine the bandwidth also in a completely data-driven way. Even if we use formally undersmoothing, this method is not far from the approach in Eubank, Speckman (1993).

## 2. The method and the main result

Throughout this paper we consider the model

$$Y_i = m(x_i) + \varepsilon_i, \quad i = 1, \ldots, n, \tag{2.1}$$

where the errors $\varepsilon_i$ are independent, but not necessarily identically distributed with $E\varepsilon_i = 0$, $E\varepsilon_i^2 = v_i$, obeying

$(A_E)$ $\quad 0 < v_{\inf} \leq v_i \leq v_{\sup} < \infty, \quad E|\varepsilon_i|^M \leq C(M) < \infty \quad$ for all $\quad i, M$.

For the design points $x_i = x_i(n)$ we assume that there exist constants $0 < C_1 \leq C_2 < \infty$ with

$(A_D)$ $\quad C_1(n(b-a) - \log n) \leq \#\{i \mid x_i \in [a, b)\} \leq C_2(n(b-a) + \log n)$
for all $\quad 0 \leq a < b \leq 1$.

We adopt $(A_D)$ in our fixed design model rather than the frequently assumed "regular design", i.e. $\int_0^{x_i} f(t)\,dt = i/n$ for some probability density $f$, because it also includes cases with considerably more irregular, clustered designs. The following remark shows that also the often considered case of "random design" is covered by our assumption.

*Remark* 1. Assume that the design points $x_i$ are realizations of i.i.d. random variables with density $f$ supported on $[0,1]$, $0 < \inf_{x \in [0,1]} f(x) \leq \sup_{x \in [0,1]} f(x) < \infty$. Then $(A_D)$ is satisfied with probability exceeding $1 - n^{-\lambda}$ for arbitrary $\lambda$ and appropriately chosen $C_1$, $C_2$.

To treat a wide variety of possible designs appropriately, we apply a local linear estimator proposed by Fan (1992). It is known that it shares all positive properties of the Nadaraya-Watson as well as the Gasser-Müller kernel estimator. An additional advantage is, that it provides a simple solution to the usual boundary problem. Fan considered in his paper only a second order local linear estimator, i.e. an estimator which uses the presence of two derivatives of the regression function, but he claimed that it is possible to extend this idea to higher regularity. For greater generality, but also for some practical points with bandwidth selection described in Section 3 we consider higher order local linear estimators, too.
In the following we assume

$(A_S)\quad m \in C^k[0,1]$.

According to this assumption, we apply a $k$-th order local linear estimator $\widehat{m}(x)$ of $m(x)$, which is given as $a_1(x, Y_1, \ldots, Y_n)$, where $a = (a_1, \ldots, a_k)'$ minimizes

$$M_x = \sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right)\left(Y_i - a_1 - a_2(x - x_i) - \cdots - a_k(x - x_i)^{k-1}\right)^2. \quad (2.2)$$

We assume that $K$ is a continuous nonnegative function with $K(x) > 0$ iff $|x| < 1$. It is clear that

$$\widehat{m}(x) = \sum w_j(x) Y_j = W_x' \underline{Y} = [(D_x' K_x D_x)^{-1} D_x' K_x \underline{Y}]_1 , \quad (2.3)$$

where $\underline{Y} = (Y_1, \ldots, Y_n)'$,

$$D_x = \begin{pmatrix} 1 & \frac{x-x_1}{h} & \cdots & (\frac{x-x_1}{h})^{k-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{x-x_n}{h} & \cdots & (\frac{x-x_n}{h})^{k-1} \end{pmatrix},$$

$$K_x = Diag\left[K(\frac{x - x_1}{h}), \ldots, K(\frac{x - x_n}{h})\right].$$

To give a first impression about the performance of this estimator, we state the following lemma.

**Lemma 2.1.** *Assume $(A_E)$, $(A_D)$, $(A_S)$. Then*
(i) $var(\widehat{m}(x)) = O((nh)^{-1})$,
(ii) $E\widehat{m}(x) - m(x) = O(h^k)$
*hold uniformly in $x \in [0,1]$.*

In the present paper we consider confidence bands of the form

$$I_x = [\widehat{m}(x) - t, \widehat{m}(x) + t] , \quad (2.4)$$

and we intend to determine such a value of $t$ that the property

$$P\left(m(x) \in I_x \quad \text{for all} \quad x \in [0,1]\right) \longrightarrow 1 - \alpha \quad (2.5)$$

is satisfied for some prescribed $\alpha$, $0 < \alpha < 1$.
In the special case of i.i.d. errors $\varepsilon_i$ Eubank and Speckman (1993) approximated the process $\{\widehat{m}(x)/\sqrt{var(\widehat{m}(x))}\}_{x\in[0,1]}$, via a strong approximation for partial sums of i.i.d. random variables, by some stationary Gaussian process and determined the

asymptotic $(1-\alpha)$-quantile of the maximum of the absolute value of the latter process by a result of Bickel and Rosenblatt (1973). This yields a uniform confidence band with an error in coverage probability of order $o(1)$.
In our considerable inhomogeneous situation due to unequal variances, nonequidistant design and the inclusion of the boundary region we do not know if one can use any available result on the maximum of the limiting process to get an analytic expression for an asymptotically correct $t$. Therefore we use the simple idea of bootstrap, which is usually applied whenever we do not know what to do with analytic methods. On the other hand, in avoiding the approximation step for the distribution of the maximal deviation of some Gaussian process we hope to get a better coverage accuracy for the confidence band. Because of the heteroscedastic errors, we apply the wild bootstrap proposed by Härdle and Mammen (1990). Starting from the residuals

$$\widehat{\varepsilon}_i = Y_i - \widehat{m}(x_i),$$

we draw independent random variables $\varepsilon_i^*$ with zero mean, variances $\widehat{\varepsilon}_i^2$ and appropriately bounded higher order moments. For simplicity we restrict ourselves to either

(i) $\quad \varepsilon_i^* \sim N(0, \widehat{\varepsilon}_i^2)$

or

(ii) $\quad P(\varepsilon_i^* = -\widehat{\varepsilon}_i) = P(\varepsilon_i^* = +\widehat{\varepsilon}_i) = 1/2.$

Now we attempt to mimic the stochastic part $\widehat{m}_0(x) = \sum w_j(x)\varepsilon_j$ of the process $(\widehat{m}(x))_{x\in[0,1]}$ by

$$\widehat{m}_0^*(x) = \sum w_j(x)\varepsilon_j^*.$$

Let $t_\alpha^*$ be the $(1-\alpha)$-quantile of the (random) distribution of the quantity

$$T_{n0}^* = \sup_{x\in[0,1]} \{|\widehat{m}_0^*(x)|\},$$

which is introduced to mimic

$$T_n = \sup_{x\in[0,1]} \{|\widehat{m}(x) - m(x)|\}.$$

Throughout the paper let $\delta > 0$ be an arbitrarily small and $\lambda < \infty$ an arbitrarily large constant. The following theorem, which is proved in Section 4, establishes an upper bound for the error in coverage probability of the confidence band of size $t_\alpha^*$ around $\widehat{m}(x)$.

**Theorem 2.1.** *Assume $(A_D)$, $(A_E)$, $(A_S)$. Then*

$$\begin{aligned} &P\left(m(x) \in [\widehat{m}(x) - t_\alpha^*, \widehat{m}(x) - t_\alpha^*] \quad \textit{for all} \quad x \in [0,1]\right) \\ &= 1 - \alpha + O\left(n^\delta (nh)^{-1/2} + (nh)^{1/2}(\log n)^{1/2} h^k\right). \end{aligned}$$

It follows that the rate for the coverage probability is nearly optimized by the choice

$$h \asymp n^{-1/(k+1)}.$$

On the other hand, it is known for kernel estimators that all commonly used bandwidth selectors are designed to minimize the risk, usually the mean square error, of the estimator. Such a bandwidth would be of order $n^{-1/(2k+1)}$ in our case, and their use would lead to a nonvanishing error in coverage probability. A practicable and heuristically motivated method to determine an appropriate bandwidth is discussed in the next section.

In view of Remark 1, for random design the assertion $(A_D)$ of the theorem holds conditioned on $\underline{X} = (X_1, \ldots, X_n)'$ with probability exceeding $1 - n^{-\lambda}$. Hence, the unconditioned error in coverage probability will be of the same order as given in the above theorem.

## 3. A PRACTICABLE RULE FOR THE BANDWIDTH CHOICE

In the literature on pointwise confidence intervals one can find two main approaches to tackle the bias problem, "undersmoothing" and "bias correction." The essential difference between them is, that for the first one the quantile $t_\alpha$ is chosen according to the stochastic part of that estimator, which defines the center of the confidence interval, whereas bias correction usually means that one takes the quantile in accordance to the stochastic part of some initial estimator, which is then corrected by an explicit bias estimator. If both approaches exploit the same amount of smoothness of the curve, undersmoothing is shown to be potentially better than explicit bias correction, which was rigorously proved in Hall (1991) for confidence intervals for a density, Hall (1992) for intervals in regression with i.i.d. errors and Neumann (1992) for regression with heteroscedastic errors.

In principle it is possible to define an appropriate explicit bias estimator also for local linear estimators, but rather than spending too much time for the consideration of this presumably worse method, we restrict our considerations in the present paper to undersmoothing.

The usual difficulty with undersmoothing in applications is, that all commonly used bandwidth selection techniques are closely connected to the optimization of the mean square error of the estimator. It turns out that these methods balance bias and standard deviation in such a way that they decrease to zero at the same rate. Hence, they are not immediately applicable for confidence bands.

To provide some motivation for our following proposal, we urge the reader, to compare first local linear estimators of different regularity. Every inclusion of an additional term in the local polynomials to be fitted could also be interpreted as a refinement of the former local linear estimator. Keeping this idea in mind, we can choose $h$ mean square error-optimal for some local linear estimator of lower regularity. For example, we could apply cross-validation to determine $h$. Because we think that someone could object that the use of another estimator for the bandwidth choice is quite arbitrary and unnatural, we hasten to point out that the same is done by Eubank and Speckman (1993) for confidence bands based on bias correction. Although the confidence band is centered around a bias corrected estimator of higher regularity than the initial estimator, the authors proposed to choose the bandwidth MSE-optimally for the latter one.

Now we turn to the effect of the randomness of such a data-driven bandwidth to the error in coverage probability. To get some feeling for this effect, we state first a simple lemma.

**Lemma 3.1.** *Assume $(A_D)$, $(A_E)$, $(A_S)$, $h \asymp n^{-\gamma}$ and $\widehat{h} - h = O_P(n^{-\mu})$. Then*

$$\widehat{m}_{\widehat{h}}(x) - \widehat{m}_h(x) = O_P\left(n^{\gamma-\mu}(n^{\delta}(nh)^{-1/2} + h^k)\right).$$

The more important question however is, whether our procedure remains consistent in the case of a randomly selected bandwidth. Of course, we could try to mimic this randomness also by the bootstrap, but this seems to make the method even more involved, and the effect is also not immediately clear. The following proposition provides an upper bound for the coverage accuracy with random bandwidth $\widehat{h}$.

**Proposition 3.1.** *Assume $(A_D)$, $(A_E)$, $(A_S)$, $h \asymp n^{-\gamma}$ and $P\left(|\widehat{h} - h| \geq Cn^{-\mu}\right) \leq Cn^{-\lambda}$. Then*

$$P\left(m(x) \in [\widehat{m}_{\widehat{h}}(x) - t^*_\alpha, \widehat{m}_{\widehat{h}}(x) + t^*_\alpha] \quad \text{for all} \quad x \in [0,1]\right)$$
$$= 1 - \alpha + O\left(n^{\delta}(nh)^{-1/2} + (nh)^{1/2}(\log n)^{1/2}h^k + n^{\gamma-\mu}(n^{\delta} + (nh)^{1/2}(\log n)^{1/2}h^k) + n^{-\lambda}\right).$$

In view of this result, each randomly chosen bandwidth $\widehat{h}$ with $\widehat{h} - h = \widetilde{O}(n^{-\delta}h, n^{-\lambda})$ for some nonrandom bandwidth $h$ leads to a confidence band with asymptotically correct coverage probability.

## 4. Proof of the main theorem

Before we turn to the proof of Theorem 2.1, we begin with some preparatory considerations and establish several lemmas on approximations to the deviation process $(\widehat{m}(x) - m(x))_{x \in [0,1]}$.
If we compare the cumulative distribution functions of two random variables, then we can expect that they are close to each other, if the difference between the random variables is small with high probability. Because of the frequent use of this fact we formalize it by introducing the following notion.

**Definition 4.1.** Let $\{Y_n\}$ and $\{Z_n\}$ $(Z_n \geq 0$ *a.s.*$)$ be sequences of random variables, and let $\{\gamma_n\}$ be a sequence of positive reals. We write

$$Y_n = \widetilde{O}(Z_n, \gamma_n),$$

if

$$P(|Y_n| > CZ_n) \leq C\gamma_n$$

holds for $n \geq 1$ and some $C < \infty$.

This notion differs obviously from the usual $O_p$, which would provide a similar property for $n \geq n_0$ and an arbitrary constant $\gamma$ instead of $C\gamma_n$ on the right-hand side. As a rule, for arbitrary $\delta$, $\lambda > 0$ we can conclude under sufficiently strong moment conditions on the distributions of the errors by Markov's and Whittle's inequalities that

$$(a_n)'\underline{\varepsilon} = \widetilde{O}(n^{\delta}\,\|a_n\|,\, n^{-\lambda}) \tag{4.1}$$

and

$$\underline{\varepsilon}' A_n \underline{\varepsilon} - E \underline{\varepsilon}' A_n \underline{\varepsilon} \; = \; \tilde{O}\left( n^{\delta} \sqrt{tr(A_n A_n')}\, , n^{-\lambda} \right) \tag{4.2}$$

hold uniformly over $a_n \in \mathbb{R}^n$ and arbitrary $(n \times n)$-matrices $A_n$, where $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$. Furthermore, we obtain similar assertions for random quantities $a_n$ and $A_n$, which is made rigorous by Lemma 5.3 in the next section.
The following lemma shows how $\tilde{O}$ can be used to prove the closeness of two random variables.

**Lemma 4.1.** *Let $\{X_n\}$ be a sequence of random variables with densities $p_n$, $\sup_t\{p_n(t)\} \leq c_n$. Further, we assume $Y_n = \tilde{O}(\gamma_{n1}, \gamma_{n2})$. Then*

$$P(X_n + Y_n < t) = P(X_n < t) + O(c_n \gamma_{n1} + \gamma_{n2})$$

*holds uniformly in $t \in (-\infty, \infty)$.*

The proof of this lemma follows immediately from the inequalities

$$P(X_n < t - C\gamma_{n1}) - P(|Y_n| > C\gamma_{n1}) \leq P(X_n + Y_n < t) \leq P(X_n < t + C\gamma_{n1}) + P(|Y_n| > C\gamma_{n1}).$$

Now we begin with our series of approximations. First we approximate

$$T_{n0} = \sup_{x \in [0,1]} \{ | \sum w_j(x) \varepsilon_j | \} \tag{4.3}$$

on an appropriate probability space by some version of

$$U_{n0} \; = \; \sup_{x \in [0,1]} \{ | \sum w_j(x) \xi_j | \}, \tag{4.4}$$

where $\xi_j \sim N(0, v_j)$ are independent.

**Lemma 4.2.** *Assume $(A_D)$, $(A_E)$. Then there exist versions of $T_{n0}$, $U_{n0}$ on a joint probability space such that*

$$T_{n0} \; - \; U_{n0} \; = \; \tilde{O}\left( n^{\delta} (nh)^{-1}, n^{-\lambda} \right).$$

*Proof.* Let

$$S_j \; = \; \sum_{i \leq j} \varepsilon_i$$

be the partial sum process and let

$$t_j \; = \; \sum_{i \leq j} v_i.$$

Then we have by Corollary 4 in Sakhanenko (1989, p. 54), that there exists a probability space such that

$$\max_{1 \leq j \leq n} \{ |S_j \; - \; W(t_j)| \} \; = \; \tilde{O}\left( n^{\delta}, n^{-\lambda} \right), \tag{4.5}$$

which implies by Lemma 5.2 that

$$\begin{aligned}
|T_{n0} - U_{n0}| &\leq \sup_x \left\{ |\sum w_j(x)(\varepsilon_j - \xi_j)| \right\} \\
&\leq \sup_x \left\{ \sum_{j=1}^{n-1} |w_j(x) - w_{j+1}(x)||S_j - W(t_j)| + |w_n(x)||S_n - W(t_n)| \right\} \\
&= \tilde{O}\left(n^{\delta}(nh)^{-1}, n^{-\lambda}\right).
\end{aligned}$$

□

In the same way we can prove the analog in the bootstrap world. Let

$$T_{n0}^* = \sup_x \left\{ |\sum_j w_j(x)\varepsilon_j^*| \right\}$$

and

$$U_{n0}^* = \sup_x \left\{ |\sum_j w_j(x)\xi_j^*| \right\},$$

where $\xi_j^* \sim N(0, v_j^*)$.

**Lemma 4.3.** *Assume ($A_D$), ($A_E$). Then, conditioned on $\underline{Y}$,*

$$T_{n0}^* - U_{n0}^* = \tilde{O}\left(n^{\delta}(nh)^{-1}, n^{-\lambda}\right)$$

*holds on an appropriate probability space with probability exceeding* $1 - n^{-\lambda}$.

*Proof.* All we have to prove is some analog to ($A_E$) for the bootstrap random variables $\varepsilon_1^*, \ldots, \varepsilon_n^*$. It is clear that the complete analog of ($A_E$) is not guaranteed for each individual random variable, since it is not excluded that the $\hat{\varepsilon}_j$'s take on quite large values. However, it is easy to see that

$$\frac{1}{n} \sum_{j=1}^{n} E^*|\varepsilon_j^*|^M \leq \tilde{C}(M) \tag{4.6}$$

holds for appropriate $\tilde{C}(M) < \infty$ with probability exceeding $1 - n^{-\lambda}$. Hence, we can again apply Corollary 4 of Sakhanenko (1989) to show that the analog of (4.5) is true on an appropriate probability space for $S_j^* = \sum_{i \leq j} \varepsilon_i^*$ and $t_j^* = \sum_{i \leq j} v_i^*$ instead of $S_j$ and $t_j$, respectively. The rest of this proof goes in complete analogy to that of Lemma 4.2. □

**Lemma 4.4.** *Assume ($A_D$), ($A_E$). Then there exist versions of $U_{n0}$ and $U_{n0}^*$ on a joint probability space with*

$$U_{n0} - U_{n0}^* = \tilde{O}\left(n^{\delta}(nh)^{-1}, n^{-\lambda}\right).$$

*Proof.* First we remark that, if we follow the pattern of the proofs of the Lemmas 4.2 and 4.3, then we would get a weaker estimate. Proceeding in this way, one could easily show that

$$\sup_{1\le j\le n}\left\{|t_j - t_j^*|\right\} = \tilde{O}\left(n^{\delta+1/2} + h^{-1}, n^{-\lambda}\right),$$

which implies

$$|W(t_j) - W(t_j^*)| = \tilde{O}\left(n^{\delta}(n^{1/4} + h^{-1/2}), n^{-\lambda}\right)$$

and now, along the lines of the abovementioned proofs,

$$U_{n0} - U_{n0}^* = \tilde{O}\left(n^{\delta}(n^{1/4} + h^{-1/2})(nh)^{-1}, n^{-\lambda}\right).$$

On the other hand, it is easy to see that we can get for a simple histogram estimator with block length of order $h$ an approximation of the order given in Lemma 4.4. To prove the assertion of our lemma, we must improve the naive approach sketched above in two directions. On the one hand, since we have not two sequences of different distributions with coinciding variances, but sequences of distributions with unequal variances, we must *localize* our partial sum approach to packages of each $O(nh)$ consecutive random variables. On the other hand, for two random variables $Z_1 \sim N(0,\sigma_1^2)$ and $Z_2 \sim N(0,\sigma_2^2)$, $\sigma_1 < \sigma_2$, we observe that $\widetilde{Z}_2 = \frac{\sigma_2}{\sigma_1}X_1 \sim N(0,\sigma_2^2)$ is closer to $Z_1$ than $\widehat{Z}_2 = Z_1 + Z_3$ with $Z_3 \sim N(0,\sigma_2^2 - \sigma_1^2)$ independent of $Z_1$. In other words, a *multiplicative* reconstruction is more powerful than an *additive* one, and hence we will use the same stretches of a Wiener Process to get appropriate versions of $\{\varepsilon_i\}$ and $\{\varepsilon_i^*\}$, respectively.
First, we split up the error vectors $\underline{\varepsilon} = (\varepsilon_1, \dots, \varepsilon_n)'$ and $\underline{\varepsilon}^* = (\varepsilon_1^*, \dots, \varepsilon_n^*)'$ in $\Delta \asymp h^{-1}$ packages of length $d_j \asymp nh$,

$$\underline{\varepsilon} = (\varepsilon_{11}, \dots, \varepsilon_{1d_1}, \dots, \varepsilon_{\Delta 1}, \dots, \varepsilon_{\Delta d_\Delta})',$$

$\underline{\varepsilon}^*$ is defined analogously. Let $v_{jk} = E\varepsilon_{jk}^2$, $v_{jk}^* = E{\varepsilon_{jk}^*}^2$ and $w_{jk}(x) = w_l(x)$, if $l$ corresponds to $(j,k)$. Further, let $V_j = \sum_{k=1}^{d_j} v_{jk}$, $V_j^* = \sum_{k=1}^{d_j} v_{jk}^*$ $(j = 1, \dots, \Delta)$.
We define

$$t_{jk} = \sum_{l\le k} v_{jk} \quad , \quad t_{jk}^* = \sum_{l\le k} v_{jk}^*,$$
$$s_{jk} = (j-1) + t_{jk}/V_j \quad , \quad s_{jk}^* = (j-1) + t_{jk}^*/V_j^*.$$

Let $W(t)$ be a Wiener Process. We define the following versions of $\underline{\varepsilon}$ and $\underline{\varepsilon}^*$ on a joint probability space:

$$\begin{aligned} \varepsilon_{jk} &= V_j^{1/2}\left(W(s_{jk}) - W(s_{j,k-1})\right), \\ \varepsilon_{jk}^* &= V_j^{*1/2}\left(W(s_{jk}^*) - W(s_{j,k-1}^*)\right). \end{aligned}$$

Obviously, the $\varepsilon_{jk}$'s as well as the $\varepsilon_{jk}^*$'s are independent, $var(\varepsilon_{jk}) = v_{jk}$, $var(\varepsilon_{jk}^*) = v_{jk}^*$.
As indicated above, we have certain averaged versions of the error processes, $\sum_{k=1}^{d_j} \varepsilon_{jk}$ and $\sum_{k=1}^{d_j} \varepsilon_{jk}^*$, which are *multiplicatively* connected.

We decompose

$$\sum_{j,k} w_{jk}(x)[\varepsilon_{jk} - \varepsilon^*_{jk}] = \Delta_1(x) + \Delta_2(x)$$

in a "coarse structure" term

$$\Delta_1(x) = \sum_j \left(V_j^{1/2} - V_j^{*1/2}\right) \sum_k w_{jk}(x) \left(W(s^*_{jk}) - W(s^*_{j,k-1})\right)$$

and a "fine structure" term

$$\Delta_2(x) = \sum_j V_j^{1/2} \sum_k \left[(W(s_{jk}) - W(s_{j,k-1})) - (W(s^*_{jk}) - W(s^*_{j,k-1}))\right].$$

In the next section we show that

$$\max_{j,k}\left\{|t_{jk} - t^*_{jk}|\right\} = \tilde{O}\left(n^\delta (nh)^{1/2}, n^{-\lambda}\right), \tag{4.7}$$

which implies $V_j \asymp V_j^* \asymp nh$ and

$$\max_j\left\{|V_j^{1/2} - V_j^{*1/2}|\right\} = \tilde{O}\left(n^\delta, n^{-\lambda}\right).$$

Therefore we have

$$\sup_x\{\Delta_1(x)\} = \tilde{O}\left(n^\delta (nh)^{-1}, n^{-\lambda}\right).$$

We rewrite

$$\begin{aligned}
\Delta_2(x) &= V_j^{1/2} \sum_{j,k} \left[\int_{s_{j,k-1}}^{s_{jk}} w_{j,k}(x)\, dW(t) - \int_{s^*_{j,k-1}}^{s^*_{jk}} w_{j,k}(x)\, dW(t)\right] \\
&= V_j^{1/2} \int_{k-1}^{k} [w_t - w_t^*]\, dW(t),
\end{aligned}$$

where

$$\begin{aligned}
w_t &= w_{j,k}(x) \quad \text{if} \quad t \in (s_{j,k-1}, s_{jk}], \\
w_t^* &= w_{j,k}(x) \quad \text{if} \quad t \in (s^*_{j,k-1}, s^*_{jk}].
\end{aligned}$$

By (4.7) and Lemma 5.2 we obtain

$$\Delta_2(x) = \tilde{O}\left(n^\delta (nh)^{-1}, n^{-\lambda}\right).$$

□

**Lemma 4.5.** *Assume $(A_D)$, $(A_E)$. Let $p_n$ denote the density of $U_{n0}$. Then*

$$\sup_t\{|p_n(t)|\} = O((nh)^{1/2}(\log n)^{1/2}).$$

*Proof.* First, we split the interval $[0,1]$ into $\Delta$ subintervals, $\Delta$ even, $1/(4h) \leq \Delta < 1/(2h)$. Define

$$Z_i = \sup_{x \in \Delta_i} \left\{ \sum w_j(x) \varepsilon_j \right\},$$

where $\Delta_i = [(i-1)/\Delta, i/\Delta)$.

Let $p_{n1}$, $p_{n1}^-$, $p_{n2}$ and $p_{n2}^-$ denote the densities of $\max_{i\,odd}\{Z_i\}$, $\min_{i\,odd}\{Z_i\}$, $\max_{i\,even}\{Z_i\}$ and $\min_{i\,even}\{Z_i\}$, respectively. (Their existence follows by Theorem 1 in Tsirel'son (1975).) Because of $p_{nj}(t) = p_{nj}^-(-t)$, $j = 1,2$, we have

$$p_n(t) \leq 2p_{n1}(t) + 2p_{n2}(t).$$

W.l.o.g. we derive an upper estimate for $p_{n1}(t)$.

Let $j \leq \Delta$ be any odd number and let $\underline{\varepsilon}_j = (\varepsilon_{j1}, \ldots, \varepsilon_{jd_j})'$ be the vector of those random variables from $\underline{\varepsilon}$, which are necessary to compute $\widehat{m}_0(x) = \sum_l w_l(x)\varepsilon_l$ on the interval $\Delta_j$. (The numeration here need not coincide with those from the proof of Lemma 4.4.)

Let $e_j = \tilde{e}_j / \|\tilde{e}_j\|$, $\tilde{e}_j = (v_{j1}^{-1/2}, \ldots, v_{jd_j}^{-1/2})'$ and $\Sigma_j = cov(\underline{\varepsilon}_j)$. It is easy to see that $\underline{\varepsilon}_j$ can be decomposed into the independent summands $\Sigma_j^{1/2} e_j e_j' \Sigma_j^{-1/2} \underline{\varepsilon}_j = 1\|\tilde{e}_j\|^{-1} e_j' \Sigma_j^{-1/2} \underline{\varepsilon}_j$ and $\Sigma_j^{1/2}(I - e_j e_j')\Sigma_j^{-1/2}\underline{\varepsilon}_j$. We decompose $\widehat{m}_0(x)$ correspondingly as

$$\widehat{m}_0(x) = \widehat{m}_{01}(x) + \widehat{m}_{02}(x),$$

where, because of $\sum_k w_{jk}(x) = 1$ for all $x \in \Delta_j$,

$$\widehat{m}_{01}(x) = \sum_k w_{jk}(x) e_j' \Sigma_j^{-1/2} \underline{\varepsilon}_j = \|\tilde{e}_j\|^{-1} e_j' \Sigma_j^{-1/2} \underline{\varepsilon}_j \sim N\left(0, \|\tilde{e}_j\|^{-2}\right)$$

and

$$\widehat{m}_{02}(x) = \widehat{m}_0(x) - \widehat{m}_{01}(x).$$

Let $m_{j1} = \widehat{m}_{01}(x)$ for any $x \in \Delta_j$ and $m_{j2} = \sup_{x \in \Delta_j}\{\widehat{m}_{02}(x)\}$. Since $\widehat{m}(x)$ uses only observations $Y_j$ with $|x - x_j| \leq h$, we get that $Z_1, \ldots, Z_{\Delta-1}$ are independent. It is clear that $(m_{11}, \ldots, m_{\Delta-1,1})$ is independent of $m_2^{odd} = (m_{12}, m_{32}, \ldots, m_{\Delta-1,2})$. Hence, we have for the conditional distribution of $Z = \max_{j\,odd}\{Z_j\}$ that

$$\begin{aligned} & P\left(Z \geq t \,\middle|\, m_2^{odd}\right) \\ &= P\left(Z_1 \geq t \,\middle|\, m_2^{odd}\right) + P\left(Z_1 < t,\, Z_3 \geq t \,\middle|\, m_2^{odd}\right) \\ &\quad + \ldots + P\left(Z_1 < t, \ldots, Z_{\Delta-3} < t, Z_{\Delta-1} \geq t \,\middle|\, m_2^{odd}\right) \\ &= P(m_{11} \geq t - m_{12}) + P(m_{11} < t - m_{12})\, P(m_{31} \geq t - m_{32}) \\ &\quad + \ldots + P(m_{11} < t - m_{12}) \cdots P(m_{\Delta-3,1} < t - m_{\Delta-3,2})\, P(m_{\Delta-1,1} \geq t - m_{\Delta-1,2}), \end{aligned} \tag{4.8}$$

which implies for the conditional density of $Z$

$$\begin{aligned} p_{Z|m_2^{odd}}(t) &= \frac{d}{dt}\left\{-P\left(Z \geq t \,\middle|\, m_2^{odd}\right)\right\} \\ &\leq p_{m_{11}}(t - m_{12}) + p_{m_{31}}(t - m_{32}) P(m_{11} < t - m_{12}) \\ &\quad + \ldots + p_{m_{\Delta-1,1}}(t - m_{\Delta-1,2}) P(m_{11} < t - m_{12}, \ldots, m_{\Delta-3,1} < t - m_{\Delta-3,2}). \end{aligned} \tag{4.9}$$

Since $m_{j1} \sim N(0, \|\tilde{e}_j\|^{-2})$, it is easy to see that

$$p_{m_{j1}}(s) \leq P(m_{j1} \geq s) \|\tilde{e}_j\| \left(C + \sqrt{c \log n}\right) + Cn^{-c/2},$$

which implies by (4.8) and (4.9)

$$\begin{aligned} p_{Z|m_2^{odd}}(t) &\leq P\left(Z \geq t \,\middle|\, m_2^{odd}\right) \max_j\{\|\tilde{e}_j\|\} \left(C + \sqrt{c \log n}\right) + C\Delta n^{-c/2} \\ &= O\left((nh)^{1/2}\sqrt{\log n}\right). \end{aligned}$$

Integration over all possible realizations of $m_2^{odd}$ finishes the proof. □

Now we turn to the proof of the main theorem.

*Proof of Theorem* 2.1. By (ii) of Lemma 2.1 and Lemma 4.2 we obtain

$$T_n - U_{n0} = \tilde{O}\left(n^{\delta}(nh)^{-1} + h^k, n^{-\lambda-1}\right),$$

which yields due to Lemma 4.1 and Lemma 4.5

$$P(T_n < t) = P(U_{n0} < t) + O\left(n^{\delta}(nh)^{-1/2} + h^k(nh)^{1/2}(\log n)^{1/2}\right) \tag{4.10}$$

for each *nonrandom* $t$.
By the Lemmas 4.1, 4.3 through 4.5 we conclude

$$P(U_{n0} < t) = P(T_{n0}^* < t \mid \underline{Y}) + \tilde{O}\left(n^{\delta}(nh)^{-1/2}, n^{-\lambda-1}\right) \tag{4.11}$$

for each *nonrandom* $t$, which yields in conjunction with (4.10)

$$P(T_n < t) = P(T_{n0}^* < t \mid \underline{Y}) + \tilde{O}\left(n^{\delta}(nh)^{-1/2} + h^k(nh)^{1/2}(\log n)^{1/2}, n^{-\lambda-1}\right) \tag{4.12}$$

for each *nonrandom* $t$.
Now it is easy to show that

$$\sup_t \{|P(T_n < t) - P(T_{n0}^* < t \mid \underline{Y})|\} = \tilde{O}\left(n^{\delta}(nh)^{-1/2} + h^k(nh)^{1/2}(\log n)^{1/2}, n^{-\lambda}\right), \tag{4.13}$$

which implies in particular

$$\begin{aligned} P(T_n < t)|_{t=t_\alpha^*} &= P(T_{n0}^* < t_\alpha^* \mid \underline{Y}) + \tilde{O}\left(n^{\delta}(nh)^{-1/2} + h^k(nh)^{1/2}(\log n)^{1/2}, n^{-\lambda}\right) \\ &= 1 - \alpha + \tilde{O}\left(n^{\delta}(nh)^{-1/2} + h^k(nh)^{1/2}(\log n)^{1/2}, n^{-\lambda}\right). \end{aligned} \tag{4.14}$$

Integrating over $t_\alpha^*$ we obtain the assertion. □

## 5. Proofs of the subordinate assertions and some technical lemmas

### 5.1. Some additional lemmas.

**Lemma 5.1.** *Assume $(A_D)$. Then*

$$\left\|(D'_x K_x D_x)^{-1}\right\| = O\left((nh)^{-1}\right)$$

*holds uniformly in $x \in [0,1]$.*

*Proof.* First, observe that

$$D'_x K_x D_x = (Q'_x)^{-1} Q_x{}^{-1},$$

where $Q_x$ is such that

$$Q'_x D'_x K_x D_x Q_x = I_k.$$

Keeping the Gram-Schmidt orthogonalization algorithm in mind, it is easy to see that one possible choice for $Q_x$ is the following one:

$$Q_x = \begin{pmatrix} 1 & -\frac{(D_{x1},D_{x2})_K}{(D_{x1},D_{x1})_K} & \cdots & -\frac{(D_{x1},D_{xk})_K}{(D_{x1},D_{x1})_K} \\ 0 & 1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & -\frac{(D_{x,k-1},D_{xk})_K}{(D_{x,k-1},D_{x,k-1})_K} \\ 0 & \cdots & 0 & 1 \end{pmatrix} *$$

$$* \begin{pmatrix} \frac{1}{\|D_{x1}\|_K} & 0 & \cdots & 0 \\ 0 & \frac{1}{\|D_{x2}-\frac{(D_{x1},D_{x2})_K}{(D_{x1},D_{x1})_K}D_{x1}\|_K} & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & \frac{1}{\|D_{xk}-\frac{(D_{x1},D_{xk})_K}{(D_{x1},D_{x1})_K}D_{x1}-\cdots-\frac{(D_{x,k-1},D_{xk})_K}{(D_{x,k-1},D_{x,k-1})_K}D_{x,k-1}\|_K} \end{pmatrix}$$

It is easy to see that

$$(D_{xl}, D_{xl})_K = \sum K\left(\frac{x-x_i}{h}\right)\left(\frac{x-x_i}{h}\right)^{2l-2} \asymp nh.$$

If we prove

$$\left\|D_{xl} - \frac{(D_{x1},D_{xl})_K}{(D_{x1},D_{x1})_K}D_{x1} - \cdots - \frac{(D_{x,l-1},D_{xl})_K}{(D_{x,l-1},D_{x,l-1})_K}D_{x,l-1}\right\|_K \geq C\,nh, \qquad (5.1)$$

then we immediately obtain that

$$\|Q_x\| = O\left((nh)^{-1/2}\right)$$

holds uniformly in $x \in [0,1]$, which yields the assertion.

For simplicity we sketch the proof of (5.1) only for the simplest case $l = 2$. By $K(x) \geq C > 0$ for $|x| \leq 1/2$ we get

$$\left\| D_{x2} - \frac{(D_{x1}, D_{x2})_K}{(D_{x1}, D_{x1})_K} D_{x1} \right\|_K$$

$$\geq \; Cnh \int_{(x-h/2)\vee 0}^{(x+h/2)\wedge 1} K(z) \left[ z - \frac{(D_{x1}, D_{x2})_K}{(D_{x1}, D_{x1})_K} \right]^2 dz \; + \; o(nh) \; \geq \; C\, nh.$$

The proof of (5.1) for $l > 2$ is analogous. □

**Lemma 5.2.** *Assume $(A_D)$. Then*

(i) $w_j(x) = O\left((nh)^{-1}\right)$,
(ii) $w_j(x) - w_{j+1}(x) = O\left((nh)^{-2}\right)$,
(iii) $\frac{d}{dx}\{w_j(x)\} = O\left(h^{-1}(nh)^{-1}\right)$

*hold uniformly in $j$ and $x \in [0,1]$.*

*Proof.* Observe that

$$w_j(x) = \left[ (D_x' K_x D_x)^{-1} \left( K(\frac{x - x_j}{h}), K(\frac{x - x_j}{h})(\frac{x - x_j}{h}), \ldots, K(\frac{x - x_j}{h})(\frac{x - x_j}{h})^{k-1} \right)' \right]_1,$$

which immediately yields (i) and (ii) due to Lemma 5.1.
Further, we have

$$\frac{d}{dx}\{w_j(x)\} = \left[ (D_x' K_x D_x)^{-1} \frac{d}{dx}\{(\ldots)'\} - (D_x' K_x D_x)^{-1} \frac{d}{dx}\{D_x' K_x D_x\} (D_x' K_x D_x)^{-1} (\ldots)' \right]_1,$$

which implies (iii). □

**Lemma 5.3.** *(uniform $\tilde{O}$-approximation) Let $\mathcal{A}^n = \{a_\theta^{(n)}\}_{\theta \in \Theta}$ and $\mathcal{A}^{n\times n} = \{A_\theta^{(n)}\}_{\theta\in\Theta}$ be families of $n$-vectors and $(n \times n)$-matrices, respectively. Further, define the $\epsilon$-entropy $E_\epsilon(\mathcal{A}^{n\times n})$ of $\mathcal{A}^{n\times n}$, as the minimal number of $(n \times n)$-matrices $A_i$ with the property that each $A \in \mathcal{A}^{n\times n}$ can be approximated by some $A_i$ with $\|A - A_i\| \leq \epsilon$. Analogously we define the $\epsilon$-entropy $E_\epsilon(\mathcal{A}^n)$ of $\mathcal{A}^n$.*
*Assume $(A_E)$, $E_{n^{-1/2-\beta}}(\mathcal{A}^n) = O(n^\gamma)$ and $E_{n^{-1-\beta}}(\mathcal{A}^{n\times n}) = O(n^\gamma)$ for some $\beta > 0, \gamma < \infty$. Then*
*(i)* $\sup_{\theta\in\Theta}\{(\|a_\theta^{(n)}\| + n^{-\beta})^{-1} |a_\theta^{(n)\prime} \underline{\varepsilon}|\} = \tilde{O}(n^\delta, n^{-\lambda})$,
*(ii)* $\sup_{\theta\in\Theta}\{(\sqrt{tr(A_\theta^{(n)} A_\theta^{(n)\prime})} + n^{-\beta})^{-1} |\underline{\varepsilon}' A_\theta^{(n)} \underline{\varepsilon} - E \underline{\varepsilon}' A_\theta^{(n)} \underline{\varepsilon}|\} = \tilde{O}(n^\delta, n^{-\lambda})$
*holds for appropriate $\delta > 0$ and $\lambda < \infty$, which can be chosen arbitrarily small and large, respectively, if all moments of the $\varepsilon_i$'s are uniformly bounded.*

*Proof.* For a one-element set $\Theta = \{\theta_0\}$ we obtain (i) and (ii) by Markov's and Whittle's inequalities, see Whittle (1960). For general $\Theta$ we derive (i) and (ii) on the basis of that set of vectors and matrices, just given by the definition of the $n^{-1/2-\beta}$-entropy and $n^{-1-\beta}$- entropy, respectively. Let $\widehat{\theta}$ denote this parameter from the approximating grid with $\|a_\theta^{(n)} - a_{\widehat{\theta}}^{(n)}\| \leq n^{-1/2-\beta}$. By Markov's, Whittle's and Bonferroni's inequalities we obtain that, for appropriate positive $\delta$ and $\lambda$,

$$\begin{aligned}
\|(a_\theta^{(n)})'\underline{\varepsilon}\| &\leq \|(a_{\widehat{\theta}}^{(n)})'\underline{\varepsilon}\| + \|a_\theta^{(n)} - a_{\widehat{\theta}}^{(n)}\| \, \|\underline{\varepsilon}\| \\
&= O\left(n^\delta \|a_{\widehat{\theta}}^{(n)}\| + n^{-1/2-\beta} n^{1/2+\delta}\right) \\
&= O\left(n^\delta \|a_\theta^{(n)}\| + n^\delta \, n^{-\beta}\right)
\end{aligned}$$

holds uniformly over $\theta \in \Theta$ with a probability exceeding $1 - O(n^{-\lambda})$, which implies (i). (ii) can be proved analogously. □

## 5.2. Proofs of the subordinate assertions.

*Proof of Remark* 1. W.l.o.g. we prove this assertion for the simplest case $X_i \sim U[0,1]$, i.e. $f \equiv 1$. The general case follows then immediately by the transformation $X_i = F^{-1}(U_i)$, where $F$ is the c.d.f. of $X_i$ and $U_i \sim U[0,1]$ are independent random variables. Because of our assumption $0 < \inf f(x) \leq \sup f(x) < \infty$, we have $0 < \inf\{\frac{d}{dx}F^{-1}(x)\} \leq \sup\{\frac{d}{dx}F^{-1}(x)\} < \infty$, which provides the assertion in the general case.

Let $U_n(t) = \sqrt{n}(G_n(t) - t)$, where $G_n(t) = n^{-1}\sum 1_{\xi_i \leq t}$, $\xi_1, \ldots, \xi_n \sim U[0,1]$ are independent. Applying Corollary 1 on p. 622 in Shorack and Wellner (1986) with $a = C(\log n)\, n^{-1}$, $b = \delta = 1/2$ and $\lambda = 3/\sqrt{2}\,\delta\sqrt{an}$, we obtain

$$P\left(\sup_{a \leq d-c \leq b} \frac{|U_n(d) - U_n(c)|}{\sqrt{d-c}} \geq \lambda\right) \leq \frac{24}{a\delta^3} \exp\left(-(1-\delta)^5 \frac{\lambda^2}{2}\right) = O(n^{-\lambda}), \quad (5.2)$$

if $C$ is chosen sufficiently large.

Let now

$$|U_n(d) - U_n(c)| \leq \lambda\sqrt{d-c}.$$

We distinguish two cases.

If $d - c \geq a$, then

$$\begin{aligned}
\left|\int_c^d dF_n - \int_c^d dF\right| &= n^{-1/2}\, |U_n(d) - U_n(c)| \\
&\leq n^{-1/2}\lambda\sqrt{d-c} \\
&= O\left(\sqrt{a}\sqrt{d-c}\right) = O(d-c).
\end{aligned}$$

If $d - c < a$, then

$$\int_c^d dF = a = O(n^{-1}\log n) \quad (5.3)$$

and

$$\begin{aligned}
\int_c^d dF_n &\leq \int_c^{c+a} dF_n = n^{-1/2}(U_n(c+a) - U_n(c)) + \int_c^d dF \\
&= O(n^{-1}\log n),
\end{aligned}$$

which completes the proof. □

*Proof of Lemma* 2.1. (i) follows immediately from Lemma 5.2.
First, note that $\sum_{l=1}^{k} \widehat{a}_l(x, Y_1, \ldots, Y_n)D_{xl}$ is just the projection in the norm $\|.\|_K$ of the vector $\underline{Y} = (Y_1, \ldots, Y_n)'$ into the linear subspace spanned by the vectors $D_{xl} = \left( (\frac{x-x_1}{h})^{l-1}, \ldots, (\frac{x-x_n}{h})^{l-1} \right)'$, $l = 1, \ldots, k$. Since the $D_{xl}$'s are linearly independent for large enough $n$, we have

$$\widehat{a}_i \left( x, (x - x_1)^{l-1}, \ldots, (x - x_n)^{l-1} \right) = h^{l-1} \delta_{il} \quad \text{for} \quad i, l \leq k.$$

This implies

$$\sum_j w_j(x)(x - x_j)^{l-1} = \widehat{a}_1 \left( x, (x - x_1)^{l-1}, \ldots, (x - x_n)^{l-1} \right) = \delta_{1l}.$$

Hence, we have by Taylor series expansion, for some $\xi_j$ between $x$ and $x_j$,

$$\begin{aligned} E\widehat{m}(x) - m(x) &= \sum_j w_j(x)(m(x_j) - m(x)) \\ &= \sum_{l=1}^{k-1} \sum_j w_j(x) \frac{m^{(l)}(x)}{l!}(x_j - x)^l + \sum_j w_j(x) \frac{m^{(k)}(\xi_j)}{k!}(x_j - x)^k \\ &= O(h^k). \end{aligned} \tag{5.4}$$

□

*Proof of Lemma* 3.1. From $w_j(x, \widehat{h}) - w_j(x, h) = O((\widehat{h} - h)h^{-1}(nh)^{-1})$ and Lemma 5.3 we get

$$\sum \left( w_j(x, \widehat{h}) - w_j(x, h) \right) \varepsilon_j = O_P \left( n^{\delta}(\widehat{h} - h)h^{-1}(nh)^{-1/2} \right) = O_P \left( n^{\delta} n^{\gamma-\mu}(nh)^{-1/2} \right).$$

Because of (5.4) and $\overline{w}_j(x) = \frac{d}{dh}\{w_j(x, h)\} = O(h^{-1}(nh)^{-1})$ we obtain

$$\sum \overline{w}_j(x)(m(x_j) - m(x)) = \sum \overline{w}_j(x) \frac{m^{(k)}(\xi_j)}{k!}(x_j - x)^k = O(h^{k-1}),$$

which yields

$$\sum \left( w_j(x, \widehat{h}) - w_j(x, h) \right) m(x_j) = O_P \left( (\widehat{h} - h)h^{k-1} \right) = O_P \left( n^{\gamma-\mu} h^k \right).$$

□

*Proof of Proposition* 3.1. According to the proof of Lemma 3.1 we get

$$|T_n(\widehat{h}) - T_n(h)| \geq \sup_x \{ |\widehat{m}_{\widehat{h}}(x) - \widehat{m}_h(x)| \} = \tilde{O} \left( n^{\gamma-\mu}(n^{\delta}(nh)^{-1/2} + h^k) \right)$$

and, by similar considerations,

$$t^*_{\alpha}(\widehat{h}) - t^*_{\alpha}(h) = \tilde{O} \left( n^{\gamma-\mu}(n^{\delta}(nh)^{-1/2} + h^k) \right);$$

which proves the assertion in conjunction with Theorem 2.1. □

*Proof of* (4.7). From $\hat{\varepsilon}_i^2 = \varepsilon_i^2 - 2\varepsilon_i(\widehat{m}(x_i) - m(x_i)) + (\widehat{m}(x_i) - m(x_i))^2$ we obtain the decomposition

$$\begin{aligned} |t_{jk} - t_{jk}^*| &\geq \left|\sum_{l\leq k} \varepsilon_{jl}^2 - v_{jl}^2\right| \\ &\quad + 2\left|\sum_{l\leq k}(\widehat{m}(x_{jl}) - E\widehat{m}(x_{jl}))\varepsilon_{jl}\right| \\ &\quad + 2\left|\sum_{l\leq k}(E\widehat{m}(x_{jl}) - m(x_{jl}))\varepsilon_{jl}\right| \\ &\quad + \sum_{l=1}^{d_j}(\widehat{m}(x_{jl}) - m(x_{jl}))^2 \\ &= R_{jk1} + R_{jk2} + R_{jk3} + R_{j4}. \end{aligned}$$

Now we have by (4.2)

$$R_{jk1} = \tilde{O}\left(n^\delta k^{1/2}, n^{-\lambda-1}\right).$$

Note that

$$\sum_{l\leq k}(\widehat{m}(x_{jl}) - E\widehat{m}(x_{jl}))\varepsilon_{jl} = \underline{\varepsilon}' A_{jl}\underline{\varepsilon}$$

holds for some matrix $A_{jl}$ with $(A_{jk})_{st} = O((nh)^{-1})$ and $(A_{jl})_{st} = 0$ for $|s-t| > Cnh$. This implies $tr(A_{jl}A_{jl}') = O(1)$, which yields by (4.2)

$$R_{jk2} = \tilde{O}\left(n^\delta, n^{-\lambda-1}\right).$$

Further, we have by $E\widehat{m}(x_{jl}) - m(x_{jl}) = O(h^2)$ and (4.1)

$$R_{jk3} = \tilde{O}\left(n^\delta h^2 k^{1/2}, n^{-\lambda-1}\right).$$

Finally, we get

$$R_{j4} = \tilde{O}\left(n^\delta, n^{-\lambda-1}\right),$$

which completes the proof. □

*Acknowledgment.* I thank Professor V. Konakov for reference to the papers by Sakhanenko and Tsirel'son.

## References


1. Bickel, P. and Rosenblatt, M. (1973). On some global measures of the derivation of density function estimators. *Ann. Statist.* **1**, 1071-1095.
2. Bjerve, S., Doksum, K. A. and Yandell, B. S. (1985). Uniform confidence bounds for regression based on a simple moving average. *Scand. J. Statist.* **12**, 159-169.
3. Eubank, R. L. and Speckman, P. L. (1993). Confidence bands in nonparametric regression. *J. Amer. Statist. Assoc.* **88**, 1287-1301.
4. Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998-1004.
5. Faraway, J. (1990). Bootstrap selection of bandwidth and confidence bands for nonparametric regression. *J. Statist. Comput. Simul.* **37**, 37-44.

6. Faraway, J. and Jhun, M. (1990). Bootstrap choice of bandwidth for density estimation. *J. Amer. Statist. Assoc.* **85**, 1119-1122.
7. Hall, P. (1991). Edgeworth expansions for nonparametric density estimators, with applications. *Statistics* **22**, 215-232.
8. Hall, P. (1992). Effect of bias estimation on coverage accuracy of bootstrap confidence intervals for a probability density. *Ann. Statist.* **20**, 675-694.
9. Hall, P. and Titterington, D. M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multivariate Anal.* **27**, 228-254.
10. Härdle, W. (1989). Asymptotic maximal deviation of $M$-smoothers. *J. Multivariate Anal.* **29**, 163-179.
11. Härdle, W. and Bowman, A. (1988). Bootstrapping in nonparametric regression: Local adaptive smoothing and confidence bands. *J. Amer. Statist. Assoc.* **83**, 102-110.
12. Härdle, W. and Mammen, E. (1990) Bootstrap methods in nonparametric regression. CORE Discussion paper 9058, Université Catholique de Louvain, Louvain-la-Neuve, Belgium.
13. Härdle, W. and Marron, J. S. (1991). Bootstrap simultaneous error bars for nonparametric regression. *Ann. Statist.* **19**, 778-796.
14. Johnston, G. J. (1982). Probabilities of maximal deviations for nonparametric regression function estimates. *J. Multivariate Anal.* **12**, 402-414.
15. Knafl, J., Sacks, J. and Ylvisaker, D. (1985). Confidence bands for regression functions. *J. Amer. Statist. Assoc.* **80**, 683-691.
16. Liero, H. (1982). On the maximal deviation of the kernel regression function estimate. *Math. Operationsforsch. Statist., Ser. Statist.* **13**, 171-182.
17. Loader, C. R. (1993). Nonparametric regression, confidence bands and bias correction. *Proceedings of the Interface Between Statistics and Computer Science*, 131-136.
18. Major, P. (1973). On non-parametric estimation of the regression function. *Studia Sci. Math. Hungar.* **8**, 347-361.
19. Neumann, M. H. (1992). Pointwise confidence intervals in nonparametric regression with heteroscedastic error structure. Preprint No. 34, Institut für Angewandte Analysis und Stochastik, Berlin.
20. Qualls, C. and Watanabe, H. (1972). *Ann. Math. Statist.* **43**, 580-596.
21. Révész, P. (1979). On the nonparametric estimation of the regression function. *Problems of Control and Information Theory* **8**, 297-302.
22. Sakhanenko, A. I. (1989). On the accuracy of normal approximation in invariance principle. *Trudy Inst. Mat. (Novosibirsk)* **13**, 40-66. (in Russian)
23. Shorack, G. R. and Wellner, J. A. (1986). Empirical Processes with Applications to Statistics. Wiley, New York.
24. Tsirel'son, B. S. (1975). The density of the distribution of the maximum of a Gaussian process. *Theory Probab. Appl.* **20**, 847-856.
25. Whittle, P. (1960). Bounds for the moments of linear and quadratic forms in independent variables. *Theory Prob. Appl.* **5**, 302-305.
26. Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261-1343.

# Recent publications of the
# Weierstraß–Institut für Angewandte Analysis und Stochastik

## Preprints 1993

78. Grigori Milstein, Michael Nussbaum: Autoregression approximation of a nonparametric diffusion model.

## Preprints 1994

79. Anton Bovier, Véronique Gayrard, Pierre Picco: Gibbs states of the Hopfield model in the regime of perfect memory.

80. Roland Duduchava, Siegfried Prößdorf: On the approximation of singular integral equations by equations with smooth kernels.

81. Klaus Fleischmann, Jean–François Le Gall: A new approach to the single point catalytic super–Brownian motion.

82. Anton Bovier, Jean–Michel Ghez: Remarks on the spectral properties of tight binding and Kronig–Penney models with substitution sequences.

83. Klaus Matthes, Rainer Siegmund–Schultze, Anton Wakolbinger: Recurrence of ancestral lines and offspring trees in time stationary branching populations.

84. Karmeshu, Henri Schurz: Moment evolution of the outflow–rate from nonlinear conceptual reservoirs.

85. Wolfdietrich Müller, Klaus R. Schneider: Feedback stabilization of nonlinear discrete–time systems.

86. Gennadii A. Leonov: A method of constructing of dynamical systems with bounded nonperiodic trajectories.

87. Gennadii A. Leonov: Pendulum with positive and negative dry friction. Continuum of homoclinic orbits.

88. Reiner Lauterbach, Jan A. Sanders: Bifurcation analysis for spherically symmetric systems using invariant theory.

89. Milan Kučera: Stability of bifurcating periodic solutions of differential inequalities in $\mathbb{R}^3$.

90. Peter Knabner, Cornelius J. van Duijn, Sabine Hengst: An analysis of crystal dissolution fronts in flows through porous media Part I: Homogeneous charge distribution.

91. Werner Horn, Philippe Laurençot, Jürgen Sprekels: Global solutions to a Penrose–Fife phase–field model under flux boundary conditions for the inverse temperature.

92. Oleg V. Lepskii, Vladimir G. Spokoiny: Local adaptivity to inhomogeneous smoothness. 1. Resolution level.

93. Wolfgang Wagner: A functional law of large numbers for Boltzmann type stochastic particle systems.

94. Hermann Haaf: Existence of periodic travelling waves to reaction–diffusion equations with excitable–oscillatory kinetics.

95. Anton Bovier, Véronique Gayrard, Pierre Picco: Large deviation principles for the Hopfield model and the Kac–Hopfield model.

96. Wolfgang Wagner: Approximation of the Boltzmann equation by discrete velocity models.

97. Anton Bovier, Véronique Gayrard, Pierre Picco: Gibbs states of the Hopfield model with extensively many patterns.

98. Lev D. Pustyl'nikov, Jörg Schmeling: On some estimations of Weyl sums.

99. Michael H. Neumann: Spectral density estimation via nonlinear wavelet methods for stationary non–Gaussian time series.

100. Karmeshu, Henri Schurz: Effects of distributed delays on the stability of structures under seismic excitation and multiplicative noise.

101. Jörg Schmeling: Estimates of Weyl sums over subsequences of natural numbers.

102. Grigori N. Milstein, Michael V. Tret'yakov: Mean–square approximation for stochastic differential equations with small noises.

103. Valentin Konakov: On convergence rates of suprema in the presence of non–negligible trends.

104. Pierluigi Colli, Jürgen Sprekels: On a Penrose–Fife model with zero interfacial energy leading to a phase–field system of relaxed Stefan type.

105. Anton Bovier: Self–averaging in a class of generalized Hopfield models.

106. Andreas Rathsfeld: A wavelet algorithm for the solution of the double layer potential equation over polygonal boundaries.