The Annals of Statistics
1998, Vol. 26, No. 4, 1570–1613

REGRESSION-TYPE INFERENCE IN NONPARAMETRIC AUTOREGRESSION

BY MICHAEL H. NEUMANN AND JENS-PETER KREISS

Humboldt-Universität zu Berlin and Technische Universität Braunschweig

We derive a strong approximation of a local polynomial estimator (LPE) in nonparametric autoregression by an LPE in a corresponding nonparametric regression model. This generally suggests the application of regression-typical tools for statistical inference in nonparametric autoregressive models. It provides an important simplification for the bootstrap method to be used: It is enough to mimic the structure of a nonparametric regression model rather than to imitate the more complicated process structure in the autoregressive case. As an example we consider a simple wild bootstrap, which is used for the construction of simultaneous confidence bands and nonparametric supremum-type tests.

1. Introduction. Autoregressive models form an important class of processes in time series analysis. A nonparametric version of these models was introduced by Jones (1978). To allow for heteroscedastic modelling of the innovations, people often consider the model

$$X_t = m(X_{t-1}, \ldots, X_{t-p}) + v(X_{t-1}, \ldots, X_{t-q})\varepsilon_t, \tag{1.1}$$

where the ε_t are assumed to be i.i.d. with mean 0 and variance 1. Several authors dealt with the interesting statistical problem of estimating the autoregression function m nonparametrically. Robinson (1983), Tjøstheim (1994) and Masry and Tjøstheim (1995) considered usual Nadaraya–Watson estimators. Recently Masry (1996) and Härdle and Tsybakov (1997) investigated local polynomial estimators in this context. For some particular purposes of statistical inference like the construction of confidence sets and tests of hypotheses, it is also important to get knowledge about the statistical properties of the underlying estimator. Franke, Kreiss and Mammen (1997) consider time-series specific as well as regression-type bootstrap methods for model (1.1), and showed their consistency for the pointwise behavior of kernel smoothers of m. One of our goals is to show the validity of one of these bootstrap methods for statistics which concern the joint distribution of nonparametric estimators. This is motivated by potential applications to simultaneous confidence bands and nonparametric tests. In this paper, we also try to consider the situation from a more general point of view. We show first the closeness, in an appropriate sense, of a model like (1.1) to a corresponding regression model. To simplify notation, we restrict ourselves to the case of one

Received August 1996; revised January 1998.

AMS 1991 *subject classifications*. Primary 62G07, 62M05; secondary 62G09, 62G15.

Key words and phrases. Nonparametric autoregression, nonparametric regression, strong approximation, bootstrap, wild bootstrap, confidence bands.

lag; that is, $p = q = 1$. Without additional effort, we may allow the whole distribution of ε_t to depend on X_{t-1}. Accordingly, our basic assumption is that $X_0, \ldots, X_T$ is a realization from a strictly stationary time-homogeneous Markov chain.

The validity of our regression-type bootstrap is based on a strong approximation of a certain nonparametric estimator $\{\hat{m}_h(x)\}_{x \in I}$, construed as a process in x, by a corresponding estimator $\{\tilde{m}_h(x)\}_{x \in I}$ in a regression model, where I is a certain interval of interest. As a typical nonparametric estimator, we consider a local polynomial estimator. To this end, we construct first a strong approximation of partial sums w.r.t. dyadic intervals, $Z_{j,k} = \sum_{t:\, X_{t-1} \in I_{j,k}} \varepsilon_t$, $I_{j,k} = [F_X^{-1}((k-1)2^{-j}), F_X^{-1}(k2^{-j}))$, where F_X is the common cumulative distribution function of the X_t, by respective partial sums in a regression model. Such a partial sum approximation w.r.t. dyadic intervals is known to be a powerful tool to prove strong approximations in several instances; the best-known example is the seminal paper by Komlós, Major and Tusnády (1975). We achieve an approximate pairing of the random variables in the autoregressive model with the random variables in the regression model by Skorokhod embedding of the innovations/errors in a common set of Wiener processes assigned to the intervals $I_{j,k}$. Note that, in contrast to recent work of Brown and Low (1996) and Nussbaum (1996) who provide approximations of nonparametric *experiments* by simpler ones, we derive more practically oriented approximations of nonparametric *estimators* by their counterparts in observation models of simpler structure.

There will be several interesting consequences of our strong approximation result. Besides our particular use as a first step for proving the validity of a certain regression-type bootstrap, one can immediately derive similarities between properties of estimators in both models. For example, the pointwise or the integrated risks of corresponding nonparametric estimators are asymptotically the same. Moreover, one may also consider this result as a characterization of some kind of asymptotic equivalence of nonparametric autoregression and nonparametric regression concerning statistical inference about the autoregression–regression function. Accordingly, this provides a justification for the use of regression-type methods in the context of nonparametric autoregression, which has already been done for a long time.

The second step in proving the validity of the bootstrap proposal consists of constructing a strong approximation of the stochastic part of the LPE in the regression model and the bootstrap counterpart. Such an approximation has already been derived in a similar context in Neumann and Polzehl (1995), and we will borrow the corresponding result from there. Both strong approximations together yield the desired strong approximation of the stochastic part of the LPE $\hat{m}_h(x)$ by the bootstrap process. We apply this result to the construction of nonparametric confidence bands and supremum-type tests. The good rate for the approximation error suggests that the wild bootstrap is valid for several other purposes, too. Kreiss, Neumann and Yao (1998) prove this for L_2-type tests similar to that developed by Härdle and Mammen (1993) in the regression case.

The paper is organized as follows. In Section 2 we present the strong approximation result for partial sums on dyadic intervals (Theorem 2.1). This implies a strong approximation of an LPE in nonparametric autoregression by an LPE in nonparametric regression. Section 3 contains the wild bootstrap proposal and the corresponding strong approximation. We describe the application of the results to simultaneous bootstrap confidence bands and nonparametric tests. Further, we briefly discuss the higher-dimensional case and robustness properties against possible deviations from our structural model assumption. In Section 4 we present some simulation results in order to demonstrate the finite sample behavior of our proposal. All proofs are deferred to a final Section 5.

2. Approximation of nonparametric autoregression by nonparametric regression. Assume we observe a stretch $X_0, \ldots, X_T$ of a strictly stationary time-homogeneous Markov chain. We are interested in estimating the autoregression function $m(x) = E(X_t \mid X_{t-1} = x)$. First, we write the data generating process in the form of a nonparametric autoregressive model,

$$X_t = m(X_{t-1}) + \varepsilon_t, \qquad t = 1, \ldots, T, \tag{2.1}$$

where the distribution of ε_t is allowed to depend on X_{t-1} with

$$E(\varepsilon_t \mid X_{t-1}) = 0,$$
$$E(\varepsilon_t^2 \mid X_{t-1}) = v(X_{t-1}).$$

The conditional variance $v(X_{t-1})$ is assumed to be bounded away from zero and infinity on compact intervals. Note that, in contrast to the frequently used assumption of errors of the form $\sigma(X_{t-1})e_t$ with i.i.d. e_t's, the errors here can follow completely different distributions and are not necessarily independent. Such a dependence may arise because the distribution of ε_t depends, through X_{t-1}, on X_0 and $\varepsilon_1, \ldots, \varepsilon_{t-1}$. To ensure recurrence, we assume that

(A1) $\{X_t: t \geq 0\}$ is a (strictly) stationary time-homogeneous Markov chain. We denote by F_X the common cumulative distribution function of the X_t, which is assumed to be continuous. Furthermore, we assume absolute regularity (i.e., β-mixing) for $\{X_t\}$ and that the β-mixing coefficients decay at a geometric rate.

REMARK 1. For the definition of mixing we refer to the monograph by Doukhan (1994), Chapter 1. Assumption (A1) is, for example, fulfilled if we assume the following explicit structure of the data-generating process:

$$X_t = m(X_{t-1}) + \sigma(X_{t-1})e_t', \tag{2.2}$$

where $\sigma: \mathbb{R} \to (0, \infty)$ and (e_t') denote i.i.d. innovations with zero mean and unit variance. If one assumes further that

$$\limsup_{|x| \to \infty} \frac{E|m(x) + \sigma(x)e_1'|}{|x|} < 1$$

and that the distribution of e_1' possesses a nowhere vanishing Lebesgue density, then one may conclude that $\{X_t\}$ defined according to (2.2) is geometrically ergodic [cf. Doukhan (1994), pages 106, 107], which implies geometrical β-mixing if the chain is stationary, that is, $X_0 \sim F_X$. Moreover, the assumption of an everywhere positive density of the innovations can be relaxed. Besides a usual drift condition to a certain compact set K, it is enough that $\mathscr{L}(X_t \mid X_{t-1} = x)$ has a conditional density $p_{X_t \mid X_{t-1}}(y \mid x)$ bounded away from zero for $x, y \in K$ and $|x - y| \leq \varepsilon$ for some $\varepsilon > 0$; for details see Franke, Kreiss, Mammen and Neumann (1998).

The assumption that the chain is stationary may be avoided, since, for any initial distribution, we have geometric convergence to the unique stationary distribution by geometric ergodicity. Nevertheless, we assume throughout the whole paper that the underlying Markov chain is stationary.

Assumption (A1) will be used to show auxiliary results such as that sums of random variables derived from the X_t behave similarly to the independent case; see Lemma 2.1 below. Hence, it becomes clear that other mixing conditions are possible as well; geometric absolute regularity is merely assumed for convenience.

Although it is perhaps more natural to approximate nonparametric autoregression by nonparametric regression with *random* design, we establish here an approximation by nonparametric regression with nonrandom design. This is done in view of the proposed bootstrap method, which mimics just nonparametric regression with nonrandom design. Let $\{x_0, \ldots, x_{T-1}\}$ be a fixed realization of $\{X_0, \ldots, X_{T-1}\}$. As a counterpart to (2.1) we consider the nonparametric regression model

$$Y_t = m(x_{t-1}) + \eta_t, \qquad t = 1, \ldots, T, \tag{2.3}$$

where the η_t's are independent with $\eta_t \sim \mathscr{L}(\varepsilon_t \mid X_{t-1} = x_{t-1})$. Here we denote the independent variables by small letters to underline the fact that we consider the distribution of the Y_t's conditioned on a fixed realization of $\{X_0, \ldots, X_{T-1}\}$.

Before we turn to the main result of this subsection, we introduce some more notation and provide a useful auxiliary lemma. 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 2.1. Let $\{Z_T\}$ be a sequence of random variables and let $\{\alpha_T\}$ and $\{\beta_T\}$ be sequences of positive reals. We write

$$Z_T = \tilde{O}(\alpha_T, \beta_T),$$

if

$$P(|Z_T| > C\alpha_T) \le C\beta_T \tag{2.4}$$

holds for $T \ge 1$ and some $C < \infty$.

This definition is obviously stronger than the usual O_P and it is well suited for our particular purpose of constructing confidence bands and critical values for tests; see the applications in Section 4.

Whenever we claim that $\tilde{O}$ holds uniformly over a certain set, we mean that (2.4) is true for the same constant C. Here and in the following we make the convention that δ denotes a positive but arbitrarily small, and λ an arbitrarily large constant.

Before we turn to the strong approximation for the partial sums, we state a quite useful lemma about the stochastic behavior of sums of geometrically β-mixing random variables.

LEMMA 2.1. *Suppose that* $(Z_t)_{t=1,\dots,T}$ *is geometrically* β*-mixing and* $EZ_t = 0$.

(i) *If* $|Z_t| \le 1$ *almost surely, then*

$$\sum_{t=1}^{T} Z_t = \tilde{O}\left(\min\left\{\sqrt{T \log T}, \sqrt{\sum_t \operatorname{var}(Z_t)}\, \log T + (\log T)^2\right\}, T^{-\lambda}\right).$$

(ii) *Under the weaker assumption that* $\forall\ M < \infty\ \exists\ C_M < \infty$ *such that* $E|Z_t|^M \le C_M$, *we have*

$$\sum_{t=1}^{T} Z_t = \tilde{O}\left(\sqrt{\sum_t \operatorname{var}(Z_t)}\, \log T + T^{\delta}, T^{-\lambda}\right).$$

REMARK 2. If the Z_t were independent, we would obtain similar bounds where some of the logarithmic factors were improved. In the first case, Bernstein's inequality would immediately yield a stochastic upper bound of order $\tilde{O}(\sqrt{\sum \operatorname{var}(Z_t)}\, \sqrt{\log T} + \log T, T^{-\lambda})$; see, for example, Neumann (1996). In the second case, since $\sup\{|Z_t|\} = \tilde{O}(T^{\delta}, T^{-\lambda})$, we would obtain that $\sum Z_t = \tilde{O}(\sqrt{\sum \operatorname{var}(Z_t)}\, \sqrt{\log T} + T^{\delta}, T^{-\lambda})$. The additional logarithmic terms arise because of the blocking technique used to handle the weak dependence.

To ensure the desired behavior of weighted sums of the ε_t's and η_t's, respectively, we impose the following condition:

(A2) For all $M < \infty$ there exist finite constants C_M such that $\sup_{x \in \mathbb{R}}\{E(|\varepsilon_t|^M \mid X_{t-1} = x)\} \le C_M$.

Actually, it can be seen from the proofs that a certain finite number M of uniformly bounded moments would suffice. However, it seems to be difficult to get a minimal value for M, and therefore we do not make the attempt to give a particular value for it.

2.1. *Approximation of partial sums w.r.t. dyadic intervals.* The ultimate goal in the present paper is to show the validity of the wild bootstrap for statistics occuring with nonparametric estimators of the autoregression function m beyond the pointwise behavior of these estimators. To this end, we show in this section that the joint (in x) distribution of an LPE $\hat{m}_h(x)$ can be approximated by the joint distribution of an analogous estimator $\tilde{m}_h(x)$ defined in a corresponding nonparametric regression model.

It will be shown in the next subsection that $\hat{m}_h(x)$ can be well approximated by $\overline{m}(x) = \sum_t \overline{w}_h(x, X_{t-1})X_t$, where the weight function $\overline{w}_h(x, X_{t-1})$ depends only on the spatial position x and a single observation X_{t-1}. Then $\overline{m}(x)$ can be decomposed into a purely stochastic part $\sum \overline{w}_h(x, X_{t-1})\varepsilon_t$, and an essentially nonrandom part $\sum \overline{w}_h(x, X_{t-1})m(X_{t-1})$. Whereas the treatment of the latter part will not cause any substantial problems, the approximation of $\sum \overline{w}_h(x, X_{t-1})\varepsilon_t$ by its counterpart $\sum \overline{w}_h(x, x_{t-1})\eta_t$ requires more work.

To formalize such an approximation, we construct, on a sufficiently rich probability space, a pairing of the random vector $(X_0', \varepsilon_1', \ldots, \varepsilon_T')$ with another vector $(\eta_1', \ldots, \eta_T')$ such that:

1. $(X_0', \varepsilon_1', \ldots, \varepsilon_T') =_d (X_0, \varepsilon_1, \ldots, \varepsilon_T)$;
2. $(\eta_1', \ldots, \eta_T') =_d (\eta_1, \ldots, \eta_T)$; and
3. $\sup_{x \in I}\{|\sum[\overline{w}_h(x, X_{t-1}')\varepsilon_t' - \overline{w}_h(x, x_{t-1})\eta_t']|\}$ is small with high probability, where I is a certain interval of interest.

To facilitate notation, we will not distinguish between the original random variables X_t, ε_t, η_t and their artificial counterparts X_t', ε_t', η_t', and simply use the notation without prime. A first step toward an approximation as in (3) is an approximation of partial sums $\sum_{t:\, X_{t-1} \in I_l} \varepsilon_t$ by $\sum_{t:\, x_{t-1} \in I_l} \eta_t$, where $\{I_l\}$ is an appropriate set of intervals. Using Skorokhod embedding techniques as in the proof of Theorem 2.1, it is rather straightforward to establish such an approximation for nonoverlapping intervals $I_l = [(l-1)\rho_T, l\rho_T)$. This can imply a significant result for the difference in (3) if ρ_T tends to zero at a faster rate than the bandwidth of the LPE. However, as will be explained below, one often gets a smaller approximation error by specific constructions that yield simultaneously good approximations of partial sums w.r.t. intervals of different lengths. The perhaps best-known example is the strong approximation for empirical processes of i.i.d. random variables by Komlós, Major and Tusnády (1975). We establish first a multiscale approximation for partial sums w.r.t. dyadic intervals $I_{j,k} = [F_X^{-1}((k-1)2^{-j}), F_X^{-1}(k2^{-j}))$, $(j,k) \in \mathscr{I}_T$, where $\mathscr{I}_T = \{(j,k) \mid 0 \le j \le j^*, 1 \le k \le 2^j\}$, $2^{j^*} \le T < 2^{j^*+1}$. [We set $F_X^{-1}(0) = -\infty$ and $F_X^{-1}(1) = \infty$.] This implies in particular a useful strong approximation w.r.t. the "natural" dyadic intervals $[(k-1)2^{-j}, k2^{-j})$. Let

$$Z_{j,k} = \sum_{t:\, X_{t-1} \in I_{j,k}} \varepsilon_t$$

and

$$Z_{j,k}' = \sum_{t:\, x_{t-1} \in I_{j,k}} \eta_t$$

be partial sums of the errors according to the autoregressive model (2.1) and the regression model (2.3), respectively.

The link between the two sampling schemes (2.1) and (2.3) will be reached by Skorokhod embeddings of the ε_t's and η_t's, respectively, in a common set of Wiener processes. Such an embedding was introduced by Skorokhod (1965) for independent random variables and is known as a possible tool to derive strong approximations for partial sums of independent random variables [cf. Csörgő and Révész (1981), Chapter 2]. Later the technique has been extended to martingales by several authors; a convenient description of the main ideas can be found in Hall and Heyde (1980), Appendix A.1.

THEOREM 2.1. *Suppose that* (A1) *and* (A2) *are fulfilled. On an appropriate probability space, there exists a pairing of the random variables from* (2.1) *with those from* (2.3) *such that*

$$P\Big(|Z_{j,k} - Z'_{j,k}| > C_\lambda\big\{(T2^{-j})^{1/4}\log T + T^\delta\big\} \text{ for any } (j,k) \in \mathscr{I}_T\Big) = O(T^{-\lambda})$$

holds uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$ *for an appropriate set* Ω_T *with* $P((X_0, \ldots, X_{T-1}) \notin \Omega_T) = O(T^{-\lambda})$.

From this approximation w.r.t. the intervals $I_{j,k}$, we can immediately derive an analogous approximation for arbitrary intervals. This includes as special cases the natural dyadic intervals $[(k-1)2^{-j}, k2^{-j})$.

COROLLARY 2.1. *Suppose that* (A1) *and* (A2) *are fulfilled, and let* Ω_T *be as in Theorem* 2.1. *Then there exists a pairing of the random variables from* (2.1) *and* (2.3) *such that*

$$P\left(\sup_{-\infty \le c < d \le \infty}\left\{\frac{|\sum_{t:\, X_{t-1}\in[c,d)} \varepsilon_t - \sum_{t:\, x_{t-1}\in[c,d)} \eta_t|}{\big[TP\big(X_0 \in [c,d)\big)\big]^{1/4}\log T + T^\delta}\right\} > C_\lambda\right) = O(T^{-\lambda})$$

holds uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$.

Using a Haar wavelet expansion of an arbitrary weighting function w we can now establish a link between $\sum_t w(X_{t-1})\varepsilon_t$ and $\sum_t w(x_{t-1})\eta_t$. Such an approximation will hold in a uniform manner and simultaneously in a whole class $\mathscr{W}$ of such weighting functions. We use $TV(w)$ to denote the total variation of w.

COROLLARY 2.2. *Suppose that* (A1) *and* (A2) *are fulfilled, and let* Ω_T *be as in Theorem* 2.1. *Moreover, we assume that the stationary density* p_X *is bounded. Then there exists a pairing of the random variables from* (2.1) *and* (2.3) *such that*

$$P\left(\sup_{w\in\mathscr{W}}\left\{\frac{|\sum_t w(X_{t-1})\varepsilon_t - \sum_t w(x_{t-1})\eta_t|}{T^{1/4}\big[TV(w)\big]^{3/4}\|w\|_1^{1/4}\log T + TV(w)T^\delta}\right\} > C_\lambda\right) = O(T^{-\lambda})$$

holds uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$.

2.2. *Strong approximation for local polynomial estimators.* We intend to construct an asymptotic confidence band for the conditional mean function m. This makes sense for a region where we have enough information about m. To facilitate the technical calculations, we assume

(A3) The stationary density p_X of X_t is bounded and satisfies $p_X(x) \geq C > 0$ for all $x \in [a, b]$

and construct a confidence band for this interval $[a, b]$.

We focus our attention to so-called local polynomial estimators. These estimators were introduced in a paper by Stone (1977). Tsybakov (1986), Korostelev and Tsybakov [(1993), Chapter 1], Fan (1992, 1993) and Fan and Gijbels (1992, 1995) discussed the behavior of LPE for nonparametric regression in full detail. Recently Masry (1996) and Härdle and Tsybakov (1997) applied LPE to nonparametric autoregressive models. A pth order local polynomial estimator $\hat{m}_h(x)$ of $m(x)$ is given as $\hat{a}_0 = \hat{a}_0(x, X_0, \ldots, X_T)$, where $\hat{a} = (\hat{a}_0, \ldots, \hat{a}_{p-1})'$ minimizes

$$M_x = \sum_{t=1}^{T} K\left(\frac{x - X_{t-1}}{h}\right)\left(X_t - \sum_{q=0}^{p-1} a_q \left(\frac{x - X_{t-1}}{h}\right)^q\right)^2. \tag{2.5}$$

At the moment we only assume that the bandwidth h of the local polynomial estimator satisfies $h = O(T^{-\kappa})$ and $h^{-1} = O(T^{1-\kappa})$ for some $\kappa > 0$. We assume that the kernel K is a nonnegative function of bounded total variation with $\operatorname{supp}(K) \subseteq [-1, 1]$, which is bounded away from zero on a set of positive Lebesgue measure. We do not impose any further smoothness condition on K, because only a particular choice of p, which makes a certain rate of convergence possible, can be motivated from the estimation point of view. From least-squares theory it is clear that $\hat{m}_h$ can be written as

$$\begin{aligned} \hat{m}_h(x) &= \sum_{t=1}^{T} w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) X_t \\ &= \left[(D_x' K_x D_x)^{-1} D_x' K_x \underline{X}\right]_1, \end{aligned} \tag{2.6}$$

where $\underline{X} = (X_1, \ldots, X_T)'$,

$$D_x = \begin{pmatrix} 1 & \frac{x - X_0}{h} & \cdots & \left(\frac{x - X_0}{h}\right)^{p-1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & \frac{x - X_{T-1}}{h} & \cdots & \left(\frac{x - X_{T-1}}{h}\right)^{p-1} \end{pmatrix},$$

$$K_x = \operatorname{Diag}\left[K\left(\frac{x - X_0}{h}\right), \ldots, K\left(\frac{x - X_{T-1}}{h}\right)\right]$$

and $[\cdot]_1$ denotes the first entry of a vector. It is shown at the end of the proof of Lemma 2.2 that $(D_x' K_x D_x)^{-1}$ exists with a probability exceeding $1 - O(T^{-\lambda})$.

On first sight the analysis of $\hat{m}_h$ seems to be quite involved, because the X_t's are dependent and enter into the right-hand side of (2.6) several times. To simplify the investigation of the deviation process (in x), $\{\hat{m}_h(x) - m(x)\}_{x \in [a,b]}$, we approximate it by an analogous deviation process defined by observations according to the nonparametric regression model (2.3).

In analogy to (2.6) we define a local polynomial estimator as

$$\tilde{m}_h(x) = \sum_{t=1}^{T} w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\}) Y_t. \tag{2.7}$$

Before we turn to the main approximation step, we derive first some approximations to $\hat{m}_h$ and $\tilde{m}_h$, which allow us to replace the local polynomial estimators by quantities of a simpler structure.

According to (2.6), the weights of the local polynomial estimator can be written as

$$\begin{aligned} &w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) \\ &\quad = \sum_{q=0}^{p-1} d_q(x, \{X_0, \ldots, X_{T-1}\}) K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q, \end{aligned} \tag{2.8}$$

where $d_q(x, \{X_0, \ldots, X_{T-1}\}) = ((D_x' K_x D_x)^{-1})_{1, q+1}$. [$M_{ij}$ denotes the (i,j)th entry of a matrix M.] The functions d_q depend on $\{X_0, \ldots, X_{T-1}\}$ in a smooth manner ("smooth" is meant in the sense of bounded total variation, which leads to appropriately decaying coefficients in a Haar series expansion) and yields the following nonrandom approximation.

LEMMA 2.2. *Assume* (A1) *and* (A3). *Then there exist nonrandom functions* $d_q^{(\infty)}(x)$, $d_q^{(\infty)}(x) = ((E D_x' K_x D_x)^{-1})_{1, q+1} = O((Th)^{-1})$, *such that*

$$\sup_{x \in [a,b]} \left\{ |d_q(x, \{X_0, \ldots, X_{T-1}\}) - d_q^{(\infty)}(x)| \right\} = \tilde{O}\left((Th)^{-3/2} \log T, T^{-\lambda}\right).$$

This lemma allows us to introduce weights $\overline{w}_h(x, X_{t-1})$, which depend only on a single value X_{t-1}, namely

$$\overline{w}_h(x, X_{t-1}) = \sum_{q=0}^{p-1} d_q^{(\infty)}(x) K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q. \tag{2.9}$$

Now we obtain the following assertions, which finally allow us to consider the difference between $\sum_t \overline{w}_h(x, X_{t-1}) \varepsilon_t$ and $\sum_t \overline{w}_h(x, x_{t-1}) \eta_t$ rather than that between the more involved quantities $\hat{m}_h(x)$ and $\tilde{m}_h(x)$.

PROPOSITION 2.1. *Suppose that* (A1) *to* (A3) *are fulfilled. Then*

$$\sup_{x\in[a,b]}\left\{\left|\sum_t\left[w_h(x,X_{t-1},\{X_0,\dots,X_{T-1}\})-\overline{w}_h(x,X_{t-1})\right]\varepsilon_t\right|\right\}$$
$$=\tilde{O}\big((Th)^{-1}(\log T)^{3/2},T^{-\lambda}\big).$$

Analogously,

$$\sup_{x\in[a,b]}\left\{\left|\sum_t\left[w_h(x,x_{t-1},\{x_0,\dots,x_{T-1}\})-\overline{w}_h(x,x_{t-1})\right]\eta_t\right|\right\}$$
$$=\tilde{O}\big((Th)^{-1}(\log T)^{3/2},T^{-\lambda}\big)$$

holds uniformly in $(x_0,\dots,x_{T-1})\in\Omega_T$, *where* Ω_T *is an appropriate set with* $P((X_0,\dots,X_{T-1})\notin\Omega_T)=O(T^{-\lambda})$.

For the next assertion concerning a term, which plays a role similar to the usual bias term in nonparametric regression, we need the following assumption:

(A4) Here m is p-times differentiable with $\sup_{x\in[a-\delta,b+\delta]}\{|m^{(p)}(x)|\}<\infty$, for some $\delta>0$.

PROPOSITION 2.2. *Suppose that* (A1), (A3) *and* (A4) *are fulfilled. As an approximation to the bias-type term we consider the nonrandom quantity*

$$b_\infty(x)=\sum_{q=0}^{p-1}d_q^{(\infty)}(x)\sum_t E\left\{K\left(\frac{x-X_{t-1}}{h}\right)\left(\frac{x-X_{t-1}}{h}\right)^q\right.$$
$$\left.\times\int_x^{X_{t-1}}\frac{(X_{t-1}-s)^{p-1}}{(p-1)!}m^{(p)}(s)\,ds\right\}.$$

Then

$$\sup_{x\in[a,b]}\{|b_\infty(x)|\}=O(h^p)$$

and

$$\sup_{x\in[a,b]}\left\{\left|\sum_t w_h(x,X_{t-1},\{X_0,\dots,X_{T-1}\})m(X_{t-1})-m(x)-b_\infty(x)\right|\right\}$$
$$=\tilde{O}\big(h^p(Th)^{-1/2}\log T,T^{-\lambda}\big).$$

To establish now the desired approximation of $\sum_t\overline{w}_h(x,X_{t-1})\varepsilon_t$ by $\sum_t\overline{w}_h(x,x_{t-1})\eta_t$, we only have to find upper bounds to the total variation and the L_1-norm of $\overline{w}_h(x,\cdot)$. This leads to the following assertion.

PROPOSITION 2.3. *Suppose that* (A1) *to* (A3) *are fulfilled, and let* Ω_T *be as in Theorem* 2.1. *Then there exists a pairing of the random variables from* (2.1)

and (2.3) *such that*

$$\sup_{x\in[a,b]}\left\{\left|\sum_t \overline{w}_h(x, X_{t-1})\varepsilon_t - \sum_t \overline{w}_h(x, x_{t-1})\eta_t\right|\right\} = \tilde{O}\big((Th)^{-3/4}\log T, T^{-\lambda}\big)$$

holds uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$.

The approximations given in the Propositions 2.1, 2.2 and 2.3 lead also to the approximation of an LPE in nonparametric autoregression by an LPE in nonparametric regression.

THEOREM 2.2. *Suppose that* (A1) *to* (A4) *are fulfilled, and let* Ω_T *be as in Theorem* 2.1. *Then there exists a pairing of the random variables from* (2.1) *and* (2.3) *such that*

$$\sup_{x\in[a,b]}\big\{|\hat{m}_h(x) - \tilde{m}_h(x)|\big\} = \tilde{O}\big(h^p(Th)^{-1/2}\log T + (Th)^{-3/4}\log T, T^{-\lambda}\big)$$

holds uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$.

Besides the technical quantification of a certain upper bound of the rate of approximation of $\hat{m}_h(x)$ by $\tilde{m}_h(x)$, the more important fact is that the difference between $\hat{m}_h(x)$ and $\tilde{m}_h(x)$ is of smaller order than the stochastic fluctuations of $\hat{m}_h(x)$, which are $O_P((Th)^{-1/2})$. This can be interpreted as some kind of asymptotic equivalence of nonparametric autoregression and nonparametric regression. It is the first step in proving the validity of a regression-type bootstrap, the so-called wild bootstrap, in nonparametric autoregression. With a simple extra argument, it can also be used for proving asymptotic equivalence of the mean squared error of nonparametric estimators in models (2.1) and (2.3).

REMARK 3. As was already mentioned, it perhaps would have been more natural to approximate nonparametric autoregression by nonparametric regression with *random* design. That is, instead of (2.3) we consider the nonparametric regression model

$$Z_t = m(Y_t) + \eta_t, \qquad t = 1, \ldots, T, \tag{2.10}$$

where the pairs (Y_t, Z_t) are i.i.d. according to the stationary distribution of the vector (X_{t-1}, X_t) in model (2.1). Let $\check{m}_h(x)$ be the local polynomial estimator in model (2.10), which is defined analogously to (2.7). It is easily seen that the statement in Theorem 2.1 implies the asymptotic equivalence of LPE's in models (2.1) and (2.10). Strictly speaking, under (A1) to (A4) there exists a pairing of the random variables from (2.1) with those of (2.10) such that

$$\sup_{x\in[a,b]}\big\{|\hat{m}_h(x) - \check{m}_h(x)|\big\} = \tilde{O}\big(h^p(Th)^{-1/2}\log T + (Th)^{-3/4}\log T, T^{-\lambda}\big).$$

3. The bootstrap. To motivate the particular resampling scheme proposed here, first note the different nature of the stochastic and the "bias-type" term. Even if the current value of the stochastic term is unknown, its distribution can be consistently mimicked by the bootstrap. In contrast, the bias can only be handled if some degrees of smoothness of m are not used by $\hat{m}_h(x)$. In nonparametric regression and density estimation, there exist two main approaches to handle the bias problem: undersmoothing and explicit bias correction.

We mimic only the stochastic term of the LPE in the bootstrap world, and we will use separate adjustments for the bias. In view of the possibly inhomogeneous conditional variances, we use here the wild bootstrap technique, which has been introduced by Wu (1986). A detailed description of this resampling scheme can be found in the monograph by Mammen (1992). It has successfully been used in nonparametric regression in the already mentioned paper by Härdle and Mammen (1993). Let $(x_0, \ldots, x_T)$ be the realization of $(X_0, \ldots, X_T)$ at hand. We generate independent bootstrap innovations $\varepsilon_1^*, \ldots, \varepsilon_T^*$ with

$$E^*\varepsilon_t^* = 0, \qquad E^*(\varepsilon_t^*)^2 = \hat{\varepsilon}_t^2 = \bigl(x_t - \hat{m}_h(x_{t-1})\bigr)^2.$$

The notation E^* is used to underline the conditional character of the distribution $\mathcal{L}(\varepsilon_1^*, \ldots, \varepsilon_T^* \mid X_0, \ldots, X_T)$. An appropriate counterpart to model (2.3) in the bootstrap world is given by

$$X_t^* = \hat{m}_h(x_{t-1}) + \varepsilon_t^*, \qquad t = 1, \ldots, T.$$

Since we mimic the stochastic term $\sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\})\varepsilon_t$ of the local polynomial estimator only, we do not use the X_t^*'s explicitly.

3.1. *A strong approximation for the bootstrap process.* In order to be able to apply devices such as Lemma 2.1, we have to ensure that for all integers M there exists a finite constant $C_M > 0$ such that

$$E^*|\varepsilon_t^*|^M \le C_M|\hat{\varepsilon}_t|^M.$$

This can be ensured if we assume that $\varepsilon_t^* = \hat{\varepsilon}_t\eta_t^*$ for a sequence of i.i.d. random variables $\eta_1^*, \ldots, \eta_T^*$ with $E^*\eta_1^* = 0$, $E^*(\eta_1^*)^2 = 1$, and $E^*|\eta_1^*|^M < \infty$, for all integers M.

For getting the desired strong approximation of $\sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\})\varepsilon_t$ by $\sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\varepsilon_t^*$, it remains to establish a connection between the latter process (in x) and $\sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\eta_t$. This can be achieved by a construction along the lines of the proof of Theorem 2.1 in Neumann and Polzehl (1995). In contrast to the strong approximation of nonparametric autoregression by nonparametric regression, where we had to devise an embedding scheme which takes the consecutive order of the observations into account, we have to combine two models with independent observations. Based on Sakhanenko's (1991) strong approximation result for partial sums of independent random variables, we can derive a strong approximation with a smaller error than in our Theorem 2.1.

LEMMA 3.1. *Suppose that* (A2) *is fulfilled. On a sufficiently rich probability space, there exists a pairing of* $\eta_1, \ldots, \eta_T$ *with* $\varepsilon_1^*, \ldots, \varepsilon_T^*$ *such that*

$$\sup_{x \in [a,b]} \left\{ \left| \sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\}) \eta_t \right. \right.$$
$$\left. \left. - \sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\}) \varepsilon_t^* \right| \right\}$$
$$= \tilde{O}\left((Th)^{-1} T^{\delta}, T^{-\lambda}\right)$$

holds uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$.

In conjunction with Proposition 2.1 and 2.3, we obtain the following theorem.

THEOREM 3.1. *Suppose that* (A1) *to* (A3) *are fulfilled. On a sufficiently rich probability space, there exists a pairing of* $X_0, \varepsilon_1, \ldots, \varepsilon_T$ *with* $\varepsilon_1^*, \ldots, \varepsilon_T^*$ *such that*

$$\sup_{x \in [a,b]} \left\{ \left| \sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) \varepsilon_t \right. \right.$$
$$\left. \left. - \sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\}) \varepsilon_t^* \right| \right\}$$
$$= \tilde{O}\left((Th)^{-3/4} \log T, T^{-\lambda}\right)$$

holds uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$.

Theorem 3.1 basically says that the stochastic part of $\hat{m}_h(x)$ can be successfully mimicked by its wild bootstrap analogue. Hence, one can apply the bootstrap to approximate, besides the pointwise distribution of nonparametric estimators which was already shown in Franke, Kreiss and Mammen (1997), supremum-type functionals of the LPE. The perhaps most often studied problem in this context are confidence bands for a certain function to be estimated nonparametrically. In addition, we may also employ the wild bootstrap for determining critical values for nonparametric supremum-type tests. We study both problems in the following.

3.2. *Bootstrap confidence bands.* The construction of nonparametric confidence bands is a classical field of application for bootstrap methods. It is well known that first-order asymptotic theory for the supremum of an approximating Gaussian process leads to an error in coverage probability of order $(\log T)^{-1}$; see Hall (1991). In contrast, we can obtain an algebraic rate of convergence by using the bootstrap.

In contrast to confidence intervals for the global mean, in nonparametric statistics one always encounters a serious problem due to the bias. The point

is that an optimal tuning of nonparametric estimators is achieved by forcing the magnitude of the stochastic term and the bias term to be of the same order. Although one can sufficiently well estimate the behavior of the stochastic term, there remains the uncertainty about the bias term. Hence, without some extra information, it is indeed impossible to construct confidence intervals or bands with an asymptotically correct coverage probability whose length decreases with the rate of optimal estimators. This is quite obvious in the case of pointwise confidence intervals where the stochastic fluctuations of the estimator are just of the same order as its standard deviation. In the case of simultaneous confidence bands, one has to consider the supremum deviation, which is known to be degenerate. The size of the stochastic part is actually by a factor of order $\sqrt{\log T}$ larger than the pointwise standard deviation; however, the stochastic fluctuations of the supremum deviation is nevertheless of smaller order of magnitude than the pointwise standard deviation. There are several options for handling this problem. In the regression context, some authors constructed conservative confidence bands for the mean function m on the basis of some prior information about the maximal roughness of m; see Knafl, Sacks and Ylvisaker (1985), Hall and Titterington (1988) and Sun and Loader (1994). As Hall and Titterington (1988) mention, such prior knowledge may sometimes arise from physical considerations or previous empirical evidence. However, this approach is clearly restricted to the case where such prior information is indeed available.

If this is not the case, there does not exist an entirely satisfactory strategy for the construction of confidence bands. The perhaps cleanest solution is to consider bands for a smoothed version of m,

$$(K_h m)(x) = \sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) m(X_{t-1}),$$

rather than for m itself. Now the problem is much easier to deal with, and with bands for $K_h m$, we have also more freedom to choose h. Note that $K_h m$ itself is random; however, it can be interpreted as some local average of m.

Let $1 - \alpha$ be the nominal coverage probability. We develop simultaneous bands as opposed to confidence bands which attain pointwise a certain coverage probability. To construct a confidence band of uniform size, we consider the quantity

$$U_T^* = \sup_{x \in [a,b]} \left\{ \left| \sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\}) \varepsilon_t^* \right| \right\},$$

which is introduced to mimic

$$U_T = \sup_{x \in [a,b]} \{ |\hat{m}_h(x) - (K_h m)(x)| \}.$$

Let t_α^* be the [random, because it depends on the sample $X_0, \ldots, X_T$ in model (2.1)] $(1 - \alpha)$-quantile of U_T^*. Then

$$I_\alpha^*(x) = [\hat{m}_h(x) - t_\alpha^*, \hat{m}_h(x) + t_\alpha^*] \tag{3.1}$$

is supposed to form an asymptotic confidence band for $K_h m$ to the prescribed level $1 - \alpha$.

A more reasonable and perhaps more natural alternative are simultaneous confidence bands whose size is proportional to an estimate of the standard deviation of $\hat{m}_h(x)$. Whereas the size of I_α^* is essentially driven by the worst case, that is, by the supremum of $V(x) = \mathrm{var}(\hat{m}_h(x))$, a variable confidence band follows in size the local variability of $\hat{m}_h(x)$. It can be expected that the area of such a confidence band is smaller than that of a band of uniform size. Moreover, it can serve as a visual diagnostic tool to detect regions where there are difficulties for the estimator—either because of large variances of the ε_t's or because of too sparse a design.

Now we describe the construction of a confidence band of variable size in detail. The residuals $\hat{\varepsilon}_t$ can also be used to estimate the variance of $\hat{m}_h(x)$, $V(x)$, by

$$\hat{V}(x) = \sum_t w_h^2(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\hat{\varepsilon}_t^2. \tag{3.2}$$

Let t_α^{**} be the $(1-\alpha)$-quantile of the distribution of

$$V_T^* = \sup_{x \in [a,b]} \left\{ \left| \sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\varepsilon_t^* / \sqrt{\hat{V}(x)} \right| \right\},$$

which mimics

$$V_T = \sup_{x \in [a,b]} \left\{ |\hat{m}_h(x) - (K_h m)(x)| / \sqrt{\hat{V}(x)} \right\}. \tag{3.3}$$

This leads to a confidence band of the form

$$I_\alpha^{**}(x) = \left[\hat{m}_h(x) - \sqrt{\hat{V}(x)}\, t_\alpha^{**}, \hat{m}_h(x) + \sqrt{\hat{V}(x)}\, t_\alpha^{**} \right]. \tag{3.4}$$

We already know from Theorem 3.1 that the process $\hat{m}_h(x) - (K_h m)(x)$ is pathwise close to the conditional process $\sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\varepsilon_t^*$ (conditioned on $X_0, \varepsilon_1, \ldots, \varepsilon_T$) on an appropriate probability space. In order to get statistically relevant results, we have to show that the approximation error is in magnitude below the size of the fluctuations of the supremum functional. If, for example, V_T had a density p_{V_T}, then the consistency of the bootstrap confidence bands would follow from a relation like

$$\| p_{V_T} \|_\infty (Th)^{-3/4} \log T = o_P(1).$$

Since the existence of such a density is not guaranteed, we formulate the following assertion, which provides a lower bound for probabilities that U_T^* and V_T^* fall into small intervals.

LEMMA 3.2. *Suppose that* (A1) *to* (A3) *are fulfilled. Then*:

(i) $P(U_T^* \in [c_1, c_2]) = O((c_2 - c_1)(Th)^{1/2}(\log T)^{1/2} + (Th)^{-1/2}T^\delta)$;
(ii) $P(V_T^* \in [c_1, c_2]) = O((c_2 - c_1)(\log T)^{1/2} + (Th)^{-1}T^\delta)$

hold uniformly in $(x_0, \ldots, x_T) \in \Omega_T$.

Part (i) of this lemma follows immediately from Lemma 2.2 in Neumann and Polzehl (1995), and part (ii) is an immediate consequence of (i) and

$$\sup_{x\in[a,b]}\{|\hat{V}(x)-V(x)|\}=\hat{O}\big((Th)^{-3/2}\log T, T^{-\lambda}\big). \tag{3.5}$$

In conjunction with Theorem 3.1, we now obtain an upper bound to the error in coverage probability for I_α^* and I_α^{**}, respectively.

THEOREM 3.2. *Suppose that* (A1) *to* (A3) *are fulfilled. Then*:

(i) $$P\big((K_h m)(x)\in[\hat{m}_h(x)-t_\alpha^*,\hat{m}_h(x)+t_\alpha^*]\ \text{for all } x\in[a,b]\big)$$
$$=1-\alpha+O\big((Th)^{-1/4}(\log T)^{3/2}\big);$$

(ii) $$P\Big((K_h m)(x)\in\Big[\hat{m}_h(x)-\sqrt{\hat{V}(x)}\,t_\alpha^{**},\hat{m}_h(x)+\sqrt{\hat{V}(x)}\,t_\alpha^{**}\Big]\ \text{for all } x\in[a,b]\Big)$$
$$=1-\alpha+O\big((Th)^{-1/4}(\log T)^{3/2}\big).$$

Although the construction of confidence intervals or bands for the unsmoothed function m is more problematic, there is nevertheless great interest in such methods. This is reflected by a large amount of literature, mostly in the context of nonparametric density estimation and regression; for a recent survey see Neumann and Polzehl (1995). There are two main routes to deal with the notorious bias problem: undersmoothing and a subsequent bias correction. Undersmoothing means that we choose h such that the bias-type term $\sum w_h(x, X_{t-1}, \{X_0,\dots,X_{T-1}\})m(X_{t-1})-m(x)$, which is $O(h^p)$, is of smaller order of magnitude than the stochastic fluctuations of U_T^*. Accordingly, the additional condition

$$h^p=o\big((Th)^{-1/2}(\log T)^{-1/2}\big) \tag{3.6}$$

would imply that I_α^* and I_α^{**} are confidence bands for m with an asymptotically correct simultaneous coverage probability; for details see Neumann and Kreiss (1996). Alternatively, we may employ an explicit bias correction. Let $\hat{B}_T(x)$ be any estimate of $B_T(x)=\sum w_h(x, X_{t-1}, \{X_0,\dots,X_{T-1}\})m(X_{t-1})-m(x)$ with

$$\sup_{x\in[a,b]}\{|\hat{B}_T(x)-B_T(x)|\}=o_P\big((Th)^{-1/2}(\log T)^{-1/2}\big). \tag{3.7}$$

Then the bias-corrected bands

$$I_{\alpha,c}^*=\big[\hat{m}_h(x)-\hat{B}_T(x)-t_\alpha^*,\hat{m}_h(x)-\hat{B}_T(x)+t_\alpha^*\big]$$

and

$$I_{\alpha,c}^{**}=\Big[\hat{m}_h(x)-\hat{B}_T(x)-\sqrt{\hat{V}(x)}\,t_\alpha^{**},\hat{m}_h(x)-\hat{B}_T(x)+\sqrt{\hat{V}(x)}\,t_\alpha^{**}\Big]$$

are asymptotic confidence bands for m with an asymptotic coverage probability $1 - \alpha$. Another possibility for handling the bias problem for bootstrap confidence bands is to use an oversmoothed estimator $\hat{m}_g$, $g \gg h$, as the underlying conditional mean function in the bootstrap world. That is, we define an appropriate counterpart to model (2.3) by

$$X_t^* = \hat{m}_g(x_{t-1}) + \varepsilon_t^*, \qquad t = 1, \ldots, T.$$

Such a proposal has been used in Härdle and Mammen (1993) for regression and in Franke, Kreiss and Mammen (1997) for autoregression. In view of an expansion like the one given in Proposition 2.2 for the bias-term in the bootstrap situation it has to be ensured that the pth derivative $\hat{m}_g^{(p)}$ estimates $m^{(p)}$ consistently.

If we denote the $(1 - \alpha)$-quantiles of $U_T^{*,\prime} = \sup_{x \in [a,b]}\{|\hat{m}_h^*(x) - \hat{m}_g(x)|\}$ and $V_T^{*,\prime} = \sup_{x \in [a,b]}\{|\hat{m}_h(x) - \hat{m}_g(x)|/\sqrt{\hat{V}(x)}\}$ by $t_\alpha^{*,\prime}$, $t_\alpha^{**,\prime}$, respectively, then we again obtain bias-corrected confidence bands for m with an asymptotic coverage probability $1 - \alpha$. These bands take exactly the form of I_α^* and I_α^{**} [cf. (3.1) and (3.4)], where t_α^*, t_α^{**} have to be replaced by $t_\alpha^{*,\prime}$, $t_\alpha^{**,\prime}$, respectively. For the simulations in Section 4 we make use of this last proposal; that is, we report on simulation results for

$$I_\alpha^{**,\prime} = \left[\hat{m}_h(x) - \sqrt{\hat{V}(x)}\, t_\alpha^{**,\prime}, \hat{m}_h(x) + \sqrt{\hat{V}(x)}\, t_\alpha^{**,\prime}\right]. \tag{3.8}$$

Both of the latter methods have the practical advantage that one may use a bandwidth h of optimal order, which in particular allows choosing it automatically by any of the popular criteria. In all cases, however, we have to admit that one has to give up optimality considerations for the confidence bands. In the case of undersmoothing, one has to use a suboptimal bandwidth, whereas we have to reserve some additional degrees of smoothness for the bias-correction step of the second and third method.

We do not dwell on the effect of a data-driven bandwidth choice which is important for a real application of this method. Usually data-driven bandwidths $\hat{h}$ are intended to approximate a certain nonrandom bandwidth h_T. If $(\hat{h} - h_T)/h_T$ converges at an appropriate rate, then the estimators $\hat{m}_{\hat{h}}$ and $\hat{m}_{h_T}$ are sufficiently close to each other, such that the results obtained in this paper remain valid; see Neumann (1995) for a detailed investigation of these effects for pointwise confidence intervals in nonparametric regression.

3.3. *A supremum-type test.* Another classical field of application of bootstrap methods is the determination of critical values for tests. In our nonparametric context, it might be of interest to test the appropriateness of a certain parametric model for m. The theory developed in this paper can be readily applied to certain nonparametric tests, as will be shown in the following.

We allow a composite hypothesis, that is,

$$H_0\colon m \in \mathcal{M},$$

where the only requirement is that the function class $\mathcal{M}$ allows a faster rate of convergence than the nonparametric model. We will assume that

(A5) There exists an estimator $\hat{\hat{m}}$ of m such that

$$\sup_{x\in\mathbb{R}}\left\{\left|\sum_{t=1}^{T}K\left(\frac{x-X_{t-1}}{h}\right)\left[\hat{\hat{m}}(X_{t-1})-m(X_{t-1})\right]\right|\right\} = o_P\left((Th)^{1/2}(\log T)^{-1/2}\right).$$

A sufficient condition for (A5) is obviously that $\hat{\hat{m}}$ itself converges in the supremum norm to m with a faster rate than $(Th)^{-1/2}(\log T)^{-1/2}$, which can be expected to hold in certain parametric models, $\mathcal{M}=\{m_\theta \mid \theta\in\Theta\}$. For the particular purpose of testing, there is no reason to use an LPE which automatically adapts to irregularities in the design. Rather, one may choose the test statistic under the aspect of convenience; for example,

$$W_T = \sup_{x\in\mathbb{R}}\left\{\left|\sum_{t=1}^{T}K\left(\frac{x-X_{t-1}}{h}\right)\left[X_t-\hat{\hat{m}}(X_{t-1})\right]\right|\right\}. \tag{3.9}$$

This roughly corresponds to a contrast function which weights the difference between $m(x)$ and $m_\theta(x)$ with a factor proportional to the stationary density $\pi(x)$.

In principle, it is also possible to look at the difference of a nonparametric estimator to (a smoothed version of) $\hat{\hat{m}}$ directly. This would lead to a test statistic like

$$W_T' = \sup_{x\in\mathbb{R}}\left\{\left|\sum_{t=1}^{T}w_h(x,X_{t-1},\{X_0,\ldots,X_{T-1}\})\left[X_t-\hat{\hat{m}}(X_{t-1})\right]g(X_{t-1})\right|\right\},$$

where g is an appropriate function which downweights contributions from regions where the stationary density is low, and where, therefore, any nonparametric estimator necessarily deteriorates. In the sequel, we restrict our considerations to the technically simpler, and not less natural, test statistic W_T. Let t_α^W be the $(1-\alpha)$-quantile of the (random) distribution of

$$W_T^* = \sup_{x\in\mathbb{R}}\left\{\left|\sum_{t=1}^{T}K\left(\frac{x-X_{t-1}}{h}\right)\varepsilon_t^*\right|\right\}. \tag{3.10}$$

By analogous considerations as above, one may prove the following theorem.

THEOREM 3.3. *Suppose that* (A1) *and* (A3) *to* (A5) *are fulfilled. Then, for arbitrary* $m\in\mathcal{M}$,

$$P_m\left(W_T > t_\alpha^W\right) = \alpha + o(1).$$

REMARK 4. It seems that L_2-tests such as those proposed by Härdle and Mammen (1993) in the regression set-up, are the most popular ones among nonparametric statisticians. Such tests can be optimal for testing against smooth alternatives, where supremum-type tests have less power in such a

situation. On the other hand, supremum-type tests can also outperform L_2-tests for testing against local alternatives having the form of sharp peaks; see Konakov, Läuter and Liero (1995) and Spokoiny (1996) for more details. Theory for L_2-tests in nonparametric autoregression is developed in Kreiss, Neumann and Yao (1998).

3.4. *Some additional remarks. Generalization to higher dimensions.* It is quite straightforward to generalize our results to the higher-dimensional case. Suppose that we observe $X_{1-d}, \ldots, X_T$ which obey the model

$$X_t = m(X_{t-1}, \ldots, X_{t-d}) + \varepsilon_t, \tag{3.11}$$

where $\mathscr{L}(\varepsilon_t \mid X_{t-1}, \ldots, X_{1-d}) = \mathscr{L}(\varepsilon_t \mid X_{t-1}, \ldots, X_{t-d})$ and $E(\varepsilon_t \mid X_{t-1}, \ldots, X_{t-d}) \equiv 0$. An appropriate counterpart to (3.11) is given by

$$Y_t = m(x_{t-1}, \ldots, x_{t-d}) + \eta_t, \tag{3.12}$$

where $x_{1-d}, \ldots, x_T$ is a fixed realization of (3.11), and $\eta_t \sim \mathscr{L}(\varepsilon_t \mid (X_{t-1}, \ldots, X_{t-d}) = (x_{t-1}, \ldots, x_{t-d}))$ are independent.

We define a hierarchical set of intervals

$$I_{j;k_1,\ldots,k_d} = \big[(k_1 - 1)2^{-j}, k_1 2^{-j}\big) \times \cdots \times \big[(k_d - 1)2^{-j}, k_d 2^{-j}\big),$$

and corresponding partial sums

$$Z_{j;k_1,\ldots,k_d} = \sum_{t:(X_{t-1},\ldots,X_{t-d}) \in I_{j;k_1,\ldots,k_d}} \varepsilon_t,$$

$$Z^*_{j;k_1,\ldots,k_d} = \sum_{t:(x_{t-1},\ldots,x_{t-d}) \in I_{j;k_1,\ldots,k_d}} \eta_t.$$

Following the lines of the proof of Theorem 2.1, we can construct a pairing of the random variables from (3.11) with those from (3.12), such that

$$\begin{aligned} &|Z_{j;k_1,\ldots,k_d} - Z'_{j;k_1,\ldots,k_d}| \\ &\quad = \tilde{O}\Big(\big[TP\big((X_{t-1},\ldots,X_{t-d}) \in I_{j;k_1,\ldots,k_d}\big)\big]^{1/4} \log T + T^{\delta}, T^{-\lambda}\Big). \end{aligned} \tag{3.13}$$

If we intend, for example, to devise a test as described in the previous subsection, we have to analyze the difference between

$$\sum_{t=1}^{T} K\left(\frac{y_1 - X_{t-1}}{h}\right) \cdot \ldots \cdot K\left(\frac{y_d - X_{t-d}}{h}\right) \varepsilon_t$$

and

$$\sum_{t=1}^{T} K\left(\frac{y_1 - x_{t-1}}{h}\right) \cdot \ldots \cdot K\left(\frac{y_d - x_{t-d}}{h}\right) \eta_t,$$

where we choose h such that $h = O(T^{-\kappa})$ and $h^{-d} = O(T^{1-\kappa})$ for some $\kappa > 0$. We approximate these sums by corresponding truncated Haar series expansions. Let $\phi_{j,k}(x) = I(x \in [(k-1)2^{-j}, k2^{-j}))$ and $\psi_{j,k}(x) = I(x \in$

$[(k-1)2^{-j},(k-1/2)2^{-j})) - I(x \in [(k-1/2)2^{-j}, k2^{-j}))$. Define

$$\alpha_{k_1,\dots,k_d} = \int \cdots \int K\left(\frac{y_1 - z_1}{h}\right) \cdot \cdots \cdot K\left(\frac{y_d - z_d}{h}\right) \phi_{0,k_1}(z_1) \cdot \cdots \cdot \phi_{0,k_d}(z_d)\, dz_1 \cdots dz_d,$$

and

$$\alpha_{j;k_1,\dots,k_d}^{(i_1,\dots,i_d)} = \int \cdots \int K\left(\frac{y_1 - z_1}{h}\right) \cdot \cdots \cdot K\left(\frac{y_d - z_d}{h}\right) \psi_{j,k_1}^{(i_1)}(z_1) \cdot \cdots \cdot \psi_{j,k_d}^{(i_d)}(z_d)\, dz_1 \cdots dz_d,$$

where $\psi_{j,k}^{(1)} = \psi_{j,k}$ and $\psi_{j,k}^{(0)} = \phi_{j,k}$. If we assume that K is compactly supported and Lipschitz, then

$$\sum_{k_1,\dots,k_d} |\alpha_{k_1,\dots,k_d}| = O(h^d)$$

and

$$\sum_{k_1,\dots,k_d} |\alpha_{j;k_1,\dots,k_d}^{(i_1,\dots,i_d)}| = O\left(2^{jd/2} h^d \left(\frac{1}{h2^j} \wedge 1\right)\right).$$

By (3.13), we obtain, for the difference at the scale j, that

$$\begin{aligned} &\sum_{(i_1,\dots,i_d)\in\{0,1\}^d\setminus(0,\dots,0)} \sum_{k_1,\dots,k_d} \alpha_{j;k_1,\dots,k_d}^{(i_1,\dots,i_d)} \\ &\qquad \times \sum_t \left[\psi_{j,k_1}^{(i_1)}(X_{t-1}) \cdot \cdots \cdot \psi_{j,k_d}^{(i_d)}(X_{t-d})\,\varepsilon_t \right. \\ &\qquad \left. - \psi_{j,k_1}^{(i_1)}(x_{t-1}) \cdot \cdots \cdot \psi_{j,k_d}^{(i_d)}(x_{t-d})\,\eta_t\right] \\ &= \tilde{O}\left(2^{jd} h^d \left(\frac{1}{h2^j} \wedge 1\right)\left[(T2^{-jd})^{1/4} \log T + T^{\delta}\right], T^{-\lambda}\right). \end{aligned} \tag{3.14}$$

In contrast to the one-dimensional case, this upper estimate diverges as $j \to \infty$. Hence, we have to choose the finest scale of our truncated Haar series expansion more carefully. Let j^* be chosen such that both

$$2^{-j^*} = O(T^{-\nu} h) \tag{3.15}$$

and

$$2^{j^* d} h^d \left(\frac{1}{h2^{j^*}} \wedge 1\right)\left[(T2^{-j^* d})^{1/4} \log T + T^{\delta}\right] = O\left(T^{-\nu}(Th^d)^{1/2}\right) \tag{3.16}$$

hold for some $\nu > 0$.

According to (3.14) and (3.16), the difference between the truncated Haar series expansions is

$$\begin{aligned}
&\sum_{k_1,\dots,k_d} \alpha_{k_1,\dots,k_d}\bigl(Z_{0;k_1,\dots,k_d} - Z'_{0;k_1,\dots,k_d}\bigr) \\
&\quad + \sum_{0\le j\le j^*} \sum_{(i_1,\dots,i_d)\in\{0,1\}^d\setminus(0,\dots,0)} \sum_{k_1,\dots,k_d} \alpha_{j;k_1,\dots,k_d}^{(i_1,\dots,i_d)} \\
&\qquad\qquad \times \sum_t \Bigl[\psi_{j,k_1}^{(i_1)}(X_{t-1})\cdot\cdots\cdot\psi_{j,k_d}^{(i_d)}(X_{t-d})\,\varepsilon_t \\
&\qquad\qquad\qquad - \psi_{j,k_1}^{(i_1)}(x_{t-1})\cdot\cdots\cdot\psi_{j,k_d}^{(i_d)}(x_{t-d})\,\eta_t\Bigr] \\
&\quad = \tilde{O}\bigl(T^{-\nu}(Th^d)^{1/2}, T^{-\lambda}\bigr),
\end{aligned} \tag{3.17}$$

which is below the magnitude of pointwise fluctuations of $\sum_{t-1}^{T} K((y_1 - X_{t-1})/h)\cdot\cdots\cdot K((y_d - X_{t-d})/h)\varepsilon_t$, which are $O((Th^d)^{1/2})$. Because of the additional factor $T^{-\nu}$, it is also below the magnitude of fluctuations of the supremum functional.

The truncation errors, that is, the differences between the statistics of interest and the corresponding truncated Haar series expansion, can be treated as follows. Because of

$$\begin{aligned}
&\Delta(y_1,\dots,y_d;z_1,\dots,z_d) \\
&\quad = \Biggl\{K\Bigl(\frac{y_1 - z_1}{h}\Bigr)\cdot\cdots\cdot K\Bigl(\frac{y_d - z_d}{h}\Bigr) \\
&\qquad - \Biggl[\sum_{k_1,\dots,k_d} \alpha_{k_1,\dots,k_d}\phi_{0,k_1}(z_1)\cdot\cdots\cdot\phi_{0,k_d}(z_d) \\
&\qquad\quad + \sum_{0\le j\le j^*} \sum_{(i_1,\dots,i_d)\in\{0,1\}^d\setminus(0,\dots,0)} \sum_{k_1,\dots,k_d} \alpha_{j;k_1,\dots,k_d}^{(i_1,\dots,i_d)} \\
&\qquad\qquad\qquad \times \sum_t \psi_{j,k_1}^{(i_1)}(z_1)\cdot\cdots\cdot\psi_{j,k_d}^{(i_d)}(z_d)\Biggr]\Biggr\} \\
&\quad = O(2^{-j^*}h^{-1}),
\end{aligned}$$

we obtain, for any *fixed* $(y_1,\dots,y_d)$, that

$$\begin{aligned}
&\sum_{t=1}^{T} \Delta(y_1,\dots,y_d;X_{t-1},\dots,X_{t-d})\,\varepsilon_t \\
&\qquad = \tilde{O}\bigl(T^{-\nu}(Th^d)^{1/2}\log T + T^{\delta}, T^{-\lambda}\bigr).
\end{aligned} \tag{3.18}$$

Analogously, we can show for the pointwise sums corresponding to the regression model (3.12) that

$$\sum_{t=1}^{T} \Delta(y_1, \ldots, y_d; x_{t-1}, \ldots, x_{t-d})\eta_t = \tilde{O}\left(T^{-\nu}(Th^d)^{1/2}\log T + T^{\delta}, T^{-\lambda}\right). \tag{3.19}$$

By establishing (3.18) and (3.19) on a sufficiently fine grid, we obtain finally, in conjunction with (3.16), the desired approximation for the supremum deviation:

$$\sup_{y_1,\ldots,y_d} \left\{ \left| \sum_{t=1}^{T} K\left(\frac{y_1 - X_{t-1}}{h}\right) \cdot \cdots \cdot K\left(\frac{y_d - X_{t-d}}{h}\right)\varepsilon_t - \sum_{t=1}^{T} K\left(\frac{y_1 - x_{t-1}}{h}\right) \cdot \cdots \cdot K\left(\frac{y_d - x_{t-d}}{h}\right)\eta_t \right| \right\} = \tilde{O}\left(T^{-\nu}(Th^d)^{1/2}\log T + T^{\delta}, T^{-\lambda}\right). \tag{3.20}$$

Note that the approximation error is below the magnitude of the fluctuations of the supremum deviation, which are $O_P((Th^d)^{1/2}(\log T)^{-1/2})$.

Robustness against deviations from the model assumptions. Although authors often avoid, for obvious reasons, such a discussion, robustness against any kind of deviations from structural model assumptions is an important issue for the reliability of a method in practical applications. Our proofs are clearly based on the Markov property of the time series, since the application of the Skorokhod embedding requires that $E(\varepsilon_t \mid X_0, \ldots, X_{t-1}) \equiv 0$. On the other hand, even if the data generating process does not obey a structural model like (2.1), it makes sense nevertheless to fit such a nonparametric autoregressive model. It would be interesting to know whether the wild bootstrap remains valid in such a case of an inadequate model. Under mixing and some extra condition on the joint densities, Robinson (1983) showed that the effect of weak dependence vanishes asymptotically for nonparametric estimators. Hart (1995) coined the term "whitening by windowing" for this effect. It is generally connected with rare events as, for example, the event that a certain X_t falls into the range of a compactly supported kernel, scaled with a bandwidth h tending to zero. In our context of supremum-type statistics, we need an appropriate version of the whitening by windowing principle beyond the pointwise properties of nonparametric estimators. Using techniques completely different from those applied here, Neumann (1997, 1998) derived such results in the context of nonparametric density estimation and nonparametric estimation of the autoregression function, respectively, from weakly dependent random variables. The rate for the approximation in this general context is of course worse than that obtained in the present paper.

4. Simulations. In this section we present the results of a simulation study. The first part deals with simulated simultaneous confidence bands for the conditional mean function $m(x)$ (cf. Section 3.2). For this purpose let us consider the following two models:

$$X_t = 4\sin(X_{t-1}) + \varepsilon_t \tag{4.1}$$

and

$$X_t = 0.8X_{t-1} + \sqrt{1 + 0.2X_{t-1}^2}\,\varepsilon_t \tag{4.2}$$

The latter model is a usual linear first order autoregression with so-called ARCH-errors.

The innovations ε_t are assumed to be i.i.d. with zero mean and unit variance. For the innovations in model (4.1) we assume a double exponential distribution, while model (4.2) is assumed to have normally distributed errors. Based upon $T = 500$ observations $X_1, \ldots, X_T$ we simulate simultaneous confidence bands of variable size [cf. (3.4)] for $m_1(x) = 4\sin(x)$ and $m_2(x) = 0.8x$. This is done by simulating the 90%-quantile $t_{0.90}$ of $\sup_{x \in [a,b]}\{|\hat{m}_h(x) - m(x)|/\sqrt{\hat{V}(x)}\}$ from 1000 Monte Carlo replications. The actual 90%-confidence bands of the form $I_{0.10} = [\hat{m}_h(x) - \sqrt{\hat{V}(x)}\,t_{0.90}, \hat{m}_h(x) + \sqrt{\hat{V}(x)}\,t_{0.90}]$ [cf. (2.6), (3.2) for the definition of $\hat{m}_h$, $\hat{V}$, respectively] together with corresponding bootstrap confidence bands [cf. (3.8)] are shown in Figures 1a–d and 2a–d. Figure 1a–d corresponds to model (4.1) while Figure 2a–d shows results for model (4.2). In both situations we report on four replications, that is, four different underlying sets of data, which result in different estimators $\hat{m}_h$ and different $(1 - \alpha)$-quantiles in the bootstrap world.

The thick lines in all figures represent bootstrap confidence bands, while thin lines are used for actual confidence bands. Broken lines in all figures indicate the LPE estimates $\hat{m}_h$ for each underlying data set of the corresponding figure.

Then m_1 is estimated by a local linear estimator $\hat{m}_h$, while for m_2 we make use of a usual Nadaraya–Watson type kernel estimator, that is, a local constant smoother. The bandwidth h in all cases is chosen according to a cross validation criterion.

Although the above simulation results are qualitative only, it is indicated that the bootstrap offers a practicable tool in order to construct not only pointwise but also simultaneous confidence bands for nonparametric estimators in nonlinear autoregression.

The following part is devoted to simulation results for the supremum-type test proposed in Section 3.3. As underlying true models we choose an ordinary first order linear autoregression

$$X_t = 0.9X_{t-1} + \varepsilon_t, \qquad \varepsilon_t \sim \text{normal} \tag{4.3}$$

and

$$X_t = 0.9\sin(X_{t-1}) + \varepsilon_t, \qquad \varepsilon_t \sim \text{double exponential}. \tag{4.4}$$

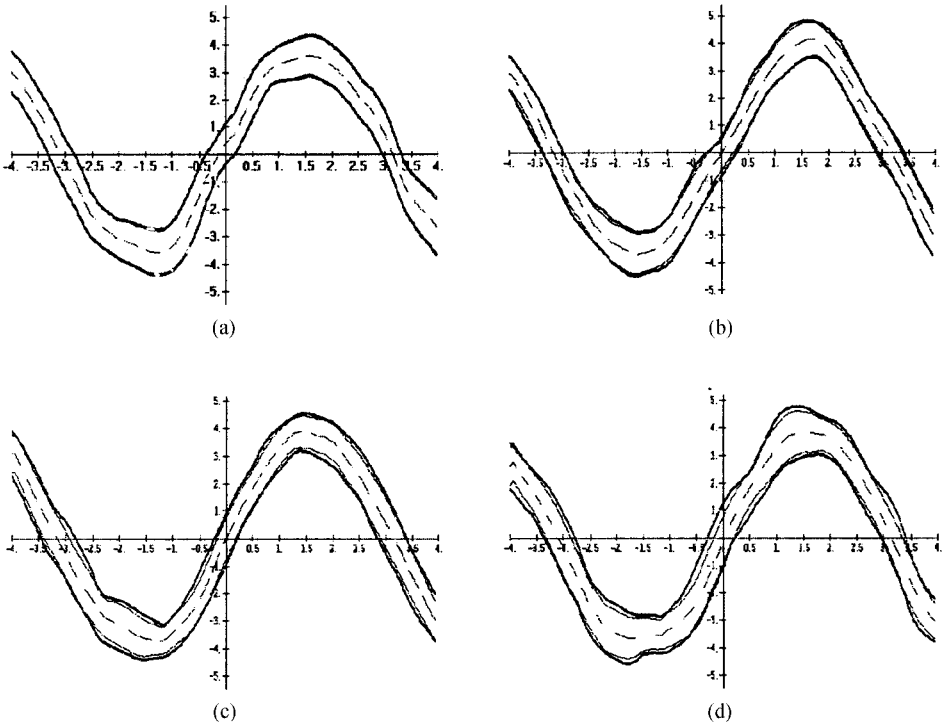

FIG. 1. *Bootstrap confidence bands* (*thick*) *based on LPE estimator* (*broken*) *and actual band* (*thin*).

Table 1 reports upon simulated values of the power function of a statistical test based on the test statistic W_T [cf. (3.10)], for the hypothesis of a first-order linear autoregression with unknown parameter θ. That is, we expect for model (4.3) values around the level α, while for model (4.4) we obtain an impression of the power of the testing procedure.

As values for the sample size T we use 100 and 200. Again, a cross-validation technique is used for the selection of the bandwidth h, which is in accordance with the theoretical results of the paper (cf. the discussion at the end of Section 3.2). The number of Monte Carlo replications both for the bootstrap and the testing procedure is chosen equal to 500. Table 2 contains similar results for a statistical test of the hypothesis H: $m \in \mathcal{M} = \{\theta \sin(\cdot) \mid \theta \in \mathbb{R}\}$, where θ is not known. The results presented in Table 2 are for just the opposite situation to Table 1. Now the hypothesis is the nonlinear model (4.4) with unknown coefficient θ. The bootstrap distribution mimics the stochastic term $\sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\})\varepsilon_t$, only, which ensures that even under the alternative a reasonable approximation of the distribution of the test statistic under the hypothesis is achieved. This guarantees reasonable power values of the proposed sup-type test, as can be seen from Tables 1 and 2. It

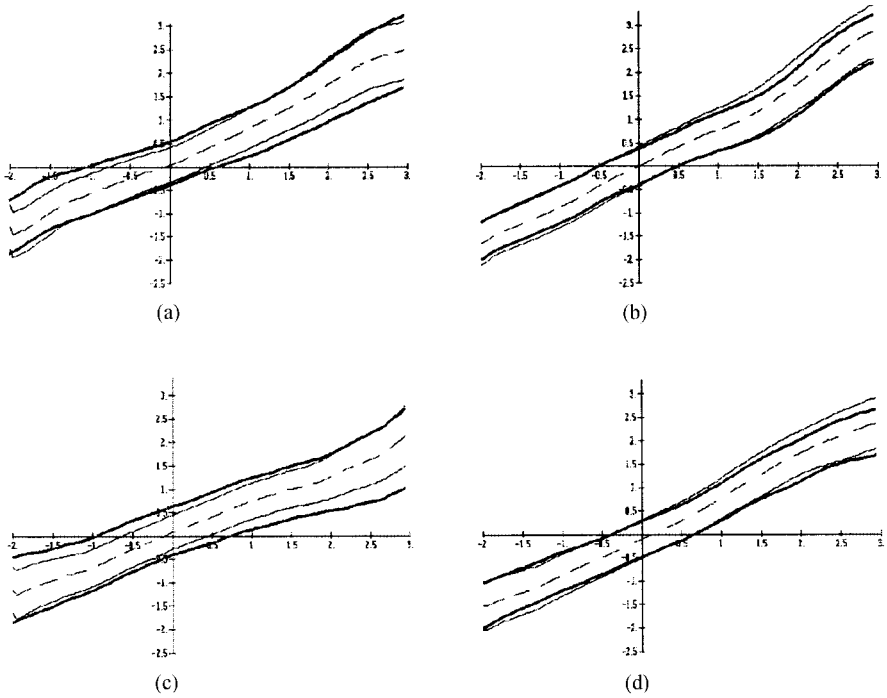

FIG. 2. *Bootstrap confidence bands* (*thick*) *based on LPE estimator* (*broken*) *and actual band* (*thin*).

can also be seen from these tables that the test on H: $m \in \{\theta \sin(\cdot) \mid \theta \in \mathbb{R}\}$ (θ unknown) for model (4.3) has much more power than the test on linearity for model (4.4). This can be explained through the much more widely spread stationary distribution of model (4.3). This implies essentially that deviations from the underlying conditional mean function over a larger interval will be taken into account by the sup-distance (cf. Section 3.3).

TABLE 1
Simulated power values for testing on first-order linear autoregression

Model	Theoretical level	Sample size	Simulated power
(4.3)	0.05	100	0.046
		200	0.044
	0.10	100	0.091
		200	0.088
(4.4)	0.05	100	0.422
		200	0.764
	0.10	100	0.600
		200	0.872

TABLE 2
Simulated power values for testing on first-order nonlinear autoregression

Model	Theoretical level	Sample size	Simulated power
(4.4)	0.05	100	0.034
		200	0.036
	0.10	100	0.068
		200	0.058
(4.3)	0.05	100	0.986
		200	1.000
	0.10	100	0.984
		200	1.000

To obtain an impression of the quality of the approximation of the bootstrap distribution $\mathscr{L}(W_T^*)$, cf. (3.11), for the distribution of the test statistic W_T, [cf. (3.10)], we include in Figures 3 and 4 slightly smoothed plots for the density of $\mathscr{L}(W_T)$ (thick line) and five randomly chosen bootstrap approximations of this distribution (thin lines). The bootstrap approximation of $\mathscr{L}(W_T)$ is used to compute critical values for the testing procedure. From Figures 3 and 4 it can be seen that the distribution of the test statistic is rather skewed.

All presented plots are based on 1000 Monte Carlo replications.

5. Proofs.

PROOF OF LEMMA 2.1. To handle the dependence, we consider blocks of consecutive observations $\{Z_t, t \in \mathscr{T}_i\}$, where $\mathscr{T}_i = \{(i-1)\rho_T + 1, \ldots, i\rho_T \wedge T\}$ and $\rho_T = [C_\lambda \log T]$. Without loss of generality, we consider the blocks with

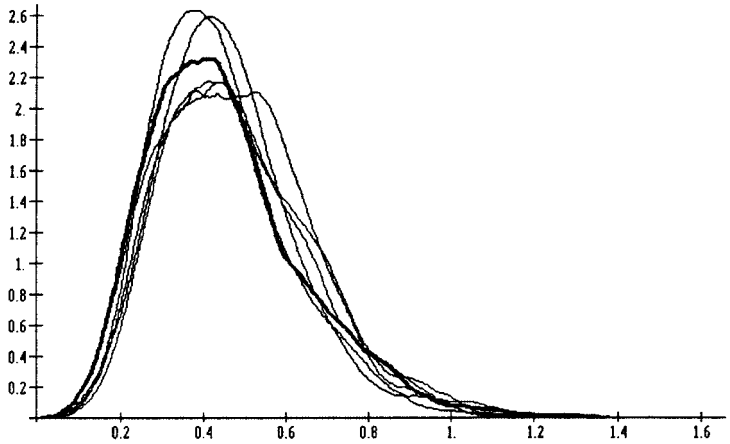

FIG. 3. $T = 100$, *model* (4.3). *Distribution of test statistic* (*thick*) *with five bootstrap approximations* (*thin*).

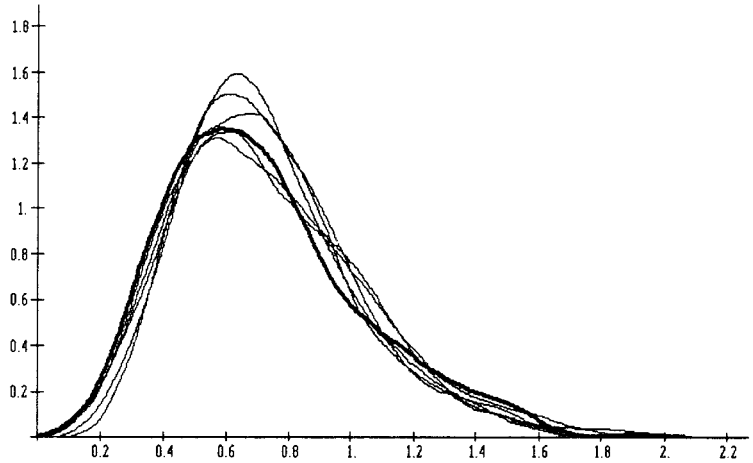

FIG. 4. $T = 200$, *model* (4.4). *Distribution of test statistic* (*thick*) *with five bootstrap approximations* (*thin*).

odd numbers. By Proposition 2 in Doukhan, Massart and Rio (1995), there exists a sequence of independent blocks $\{Z_t', t \in \mathcal{T}\}$, i odd, with the property

$$P\big((Z_t', t \in \mathcal{T}_i) \neq (Z_t, t \in \mathcal{T}_i) \text{ for any odd } i\big) = O(T^{-\lambda}), \tag{5.1}$$

where we have to choose the value of C_λ in dependence of λ.

After this reduction to the independent case, we will obtain the assertion from Bernstein's inequality, which we quote for the reader's convenience from Shorack and Wellner [(1986), page 855]: Let $U_1, \ldots, U_n$ be independent random variables with $EU_i = 0$ and $|U_i| \leq K_n$ almost surely. Then, for $U = \sum U_i$,

$$\begin{aligned} P(U > c) &\leq \exp\left(-\frac{c^2/2}{\operatorname{var}(U) + (K_n c)/3}\right) \\ &\leq \exp\left(-\frac{c^2}{4\operatorname{var}(U)}\right) + \exp\left(-\frac{3c}{4K_n}\right) \end{aligned}$$

holds for arbitrary $c > 0$.

Setting

$$c_\lambda = \sqrt{\operatorname{var}(U)}\sqrt{4\lambda \log(n)} + (4/3) K_n \lambda \log(n)$$

we get

$$P\big(|U| > c_\lambda\big) \leq 4\exp(-\lambda \log(n)).$$

In other words, we have that

$$U = \tilde{O}\Big(\sqrt{\operatorname{var}(U)}\sqrt{\log(n)} + K_n \log(n), n^{-\lambda}\Big). \tag{5.2}$$

From Davydov (1970) we get, for $p, q, r \geq 1$ and $1/p + 1/q + 1/r = 1$, that

$$\begin{aligned} \operatorname{var}\Big(\sum_{t \in \mathscr{T}_i} Z'_t\Big) &= \sum_{s,\, t \in \mathscr{T}_i} \operatorname{cov}(Z'_s, Z'_t) \\ &\leq \sum_{s,\, t \in \mathscr{T}_i} 8\beta(|s-t|)^{1/r} \|Z_s\|_p \|Z_t\|_q = O(\log T). \end{aligned} \tag{5.3}$$

On the other hand, we obtain by Jensen's inequality that

$$\operatorname{var}\Big(\sum_{t \in \mathscr{T}_i} Z'_t\Big) \leq \rho_T \sum_{t \in \mathscr{T}_i} \operatorname{var}(Z_t). \tag{5.4}$$

Assertion (i) now follows from (5.1) to (5.4) as well as $\|\sum_{t \in \mathscr{T}_i} Z'_t\|_\infty = O(\log T)$. Under the hypothesis of (ii), we have in particular that

$$\sup_{t=1,\ldots,T} \{|Z_t|\} = \tilde{O}(T^{\delta'}, T^{-\lambda}) \tag{5.5}$$

holds for any $\delta' \in (0, \delta)$. Using the truncated random variables $Z''_t = Z_t I(|Z_t| < T^{\delta'})$ instead of Z_t, we obtain (ii) from (5.1), (5.2), (5.4) and (5.5). □

PROOF OF THEOREM 2.1.

(i) *General idea*. The pairing of the observations in the autoregression model (2.1) with those in the regression model (2.3), which provides a close connection between $Z_{j,k}$ and $Z'_{j,k}$, is made via a Skorokhod embedding of the ε_t's and η_t's, respectively, in a common set of Wiener processes. This technique makes use of the well-known fact that any random variable Y with $EY = 0$ and $EY^2 < \infty$ can be represented as the value of a Wiener process stopped at an appropriate random time. Moreover, such a representation is also possible for the partial sum process of independent random variables as well as for a discrete time martingale; see, for example, Hall and Heyde (1980), Appendix A.1 for a convenient description. In particular, one can show asymptotic normality for a martingale with this approach.

However, here we have a different task. We are not interested in a close connection of the two *global* partial sum processes $S_n = \sum_{t=1}^n \varepsilon_t$ and $S'_n = \sum_{t=1}^n \eta_t$, but we are interested in a close connection of the sums of those ε_t's and η_t's which correspond to X_{t-1}'s and x_{t-1}'s, respectively, that fall into a particular interval. A quite obvious modification of the usual Skorokhod embedding in one Wiener process would be to relate the sets of random variables $\{\varepsilon_1, \ldots, \varepsilon_T\}$ and $\{\eta_1, \ldots, \eta_T\}$ to independent Wiener processes W_k, which correspond to the intervals $I_{j^*,k}$ on the finest resolution scale under consideration. This would lead to such a pairing of $\{\varepsilon_1, \ldots, \varepsilon_T\}$ with $\{\eta_1, \ldots, \eta_T\}$, which provides a close connection between $Z_{j^*,k}$ and $Z'_{j^*,k}$. If j^* is chosen fine enough, that is if $2^{-j^*} \ll h$, then we also get $\hat{m}_h(x) - \tilde{m}_h(x) = o_P((Th)^{-1/2})$. However, although this *monoscale* approximation is quite good for the differences between $Z_{j,k}$ and $Z'_{j,k}$ for j close to j^*, it is not optimal at coarser scales $j \ll j^*$. In view of this inefficiency we apply here a refined, truly *multiscale* approximation scheme. Accordingly we will relate the ε_t's and η_t's to Wiener processes $W_{j,k}$ for $(j,k) \in \mathscr{I}_T$.

In the following we describe this construction in detail for the autoregressive process (2.1). The construction in the regression setting (2.3) is completely analogous, and will only be mentioned briefly. Then we draw conclusions for the rate of approximation of $Z_{j,k}$ by $Z'_{j,k}$, which will complete the proof.

(ii) *Embedding of* ε_1. Let $W_{j,k}$, $(j,k) \in \mathcal{I}_T$, be independent Wiener processes. We will use each of these processes only on a certain time interval $[0, T_{j,k}]$, where the values of the $T_{j,k}$'s will be specified in part (v) below. At the moment it is only important to know that $T_{0,k} = \infty$.

Let k_1 be that random number with $X_0 \in I_{j^*,k_1}$. Now we represent ε_1 by the Wiener process W_{j^*,k_1}. This should be done with the aid of a stopping time $\tau^{(1)}$, which is constructed according to Lemma A.2 in Hall and Heyde (1980), Appendix A.1. However, since we want to use W_{j^*,k_1} up to some time T_{j^*,k_1} only, it might happen that this is not enough for representing ε_1. In this case we use additionally a certain stretch of the process $W_{j^*-1,[k_1/2]}$, and so on.

To formalize this construction, let $k^{(j)}$ be such that

$$I_{j^*,k} \subseteq I_{j^*-1,k^{(j^*-1)}} \subseteq \cdots \subseteq I_{0,k^{(0)}},$$

that is, $k^{(j)} = [k2^{j-j^*}]$, where $[a]$ denotes the largest integer not greater than a. According to the above description we represent ε_1 by the following Wiener process:

$$W^{(1)}(s) = \begin{cases} W_{j^*,k_1}(s), & \text{if } 0 \le s \le T_{j^*,k_1}, \\ W_{j^*,k_1}\bigl(T_{j^*,k_1}\bigr) + \cdots + W^{(j+1)}_{j+1,k_1}\bigl(T^{(j+1)}_{j+1,k_1}\bigr) \\ \quad + W_{j,k_1^{(j)}}\bigl(s - T_{j^*,k_1} - \cdots - T^{(j+1)}_{j+1,k_1}\bigr), \\ \qquad \text{if } T_{j^*,k_1} + \cdots + T_{j+1,k_1^{(j+1)}} < s \le T_{j^*,k_1} + \cdots + T^{(j)}_{j,k_1}. \end{cases}$$

($W^{(1)}$ is indeed a Wiener process on $[0,\infty)$, since $T_{0,k} = \infty$.)

According to Lemma A.2 in Hall and Heyde (1980), we have

$$\mathcal{L}(\varepsilon_1 \mid X_0 = x_0) = W^{(1)}(\tau^{(1)})$$

for an appropriate stopping time $\tau^{(1)}$.

To explain the following steps in a formally correct way, we introduce stopping times $\tau^{(t)}_{j,k}$, $t = 0, \ldots, T$, assigned to the corresponding Wiener processes $W_{j,k}$. Define

$$\tau^{(0)}_{j,k} = 0 \quad \text{for all } (j,k) \in \mathcal{I}_T.$$

To get $\tau^{(1)}_{j,k}$ we redefine all those $\tau^{(0)}_{j,k}$'s, which are assigned to Wiener processes $W_{j,k}$ that were needed to represent ε_1. According to the above construction we set

$$\tau^{(1)}_{j^*,k_1} = \tau^{(1)} \wedge T_{j^*,k_1}.$$

We redefine further

$$\tau_{j,k_1}^{(1)(j)} = \begin{cases} \left[\tau^{(1)} - T_{j^*,k_1} - \cdots - T_{j+1,k_1}^{(j+1)}\right] \wedge T_{j,k_1}^{(j)}, \\ \qquad \text{if } T_{j^*,k_1} + \cdots + T_{j-1,k_1}^{(j-1)} < \tau^{(1)}, \\ 0, \quad \text{otherwise.} \end{cases}$$

The remaining stopping times $\tau_{j,l}^{(1)}$ with $l \neq k_1^{(j)}$ keep their preceding value $\tau_{j,l}^{(0)} = 0$. This procedure will be repeated for all other ε_t's, with the modification that we use only stretches of the Wiener processes, which are still untouched by the previous construction steps.

(iii) *Embedding of* ε_t. Let k_t be that random number with $X_{t-1} \in I_{j^*,k_t}$. We represent ε_t by means of parts of W_{j^*,k_t}, $W_{j^*-1,k_t^{(j^*-1)}}, \ldots, W_{0,k_t^{(0)}}$, which have not been used to far.

First note that, because of the strong Markov property, these remaining parts $W_{j,k_t^{(j)}}(s + \tau_{j,k_t^{(j)}}^{(t-1)}) - W_{j,k_t^{(j)}}(\tau_{j,k_t^{(j)}}^{(t-1)})$ are again Wiener processes. Hence,

$$W^{(t)}(s) = \begin{cases} W_{j^*,k_t}\left(s + \tau_{j^*,k_t}^{(t-1)}\right) - W_{j^*,k_t}\left(\tau_{j^*,k_t}^{(t-1)}\right), \quad \text{if } 0 \le s \le T_{j^*,k_t} - \tau_{j^*,k_t}^{(t-1)}, \\ \left(W_{j^*,k_t}\left(T_{j^*,k_t}\right) - W_{j^*,k_t}\left(\tau_{j^*,k_t}^{(t-1)}\right)\right) \\ \quad + \cdots + \left(W_{j+1,k_t^{(j+1)}}\left(T_{j+1,k_t^{(j+1)}}\right) - W_{j+1,k_t^{(j+1)}}\left(\tau_{j+1,k_t^{(j+1)}}^{(t-1)}\right)\right) \\ \quad + \Big(W_{j,k_t^{(j)}}\Big(s - \left(T_{j^*,k_t} - \tau_{j^*,k_t}^{(t-1)}\right) \\ \qquad\qquad - \cdots - \left(T_{j+1,k_t^{(j+1)}} - \tau_{j+1,k_t^{(j+1)}}^{(t-1)}\right) + \tau_{j,k_t^{(j)}}^{(t-1)}\Big) \\ \qquad\qquad - W_{j,k_t^{(j)}}\left(\tau_{j,k_t^{(j)}}^{(t-1)}\right)\Big), \\ \quad \text{if } \left(T_{j^*,k_t} - \tau_{j^*,k_t}^{(t-1)}\right) + \cdots + \left(T_{j+1,k_t^{(j+1)}} - \tau_{j+1,k_t^{(j+1)}}^{(t-1)}\right) < s \\ \quad \le \left(T_{j^*,k_t} - \tau_{j^*,k_t}^{(t-1)}\right) + \cdots + \left(T_{j,k_t^{(j)}} - \tau_{j,k_t^{(j)}}^{(t-1)}\right) \end{cases}$$

is again a Wiener process on $[0, \infty)$.

Now we take, according to the construction in Lemma A.2 in Hall and Heyde (1980), a stopping time $\tau^{(t)}$ with

$$\mathscr{L}(\varepsilon_t \mid X_{t-1} = x_{t-1}) = W^{(t)}(\tau^{(t)}).$$

To get $\tau_{j,k}^{(t)}$, we redefine those stopping times $\tau_{j,k}^{(t-1)}$, which are assigned to Wiener processes $W_{j,k}$ that were used to represent ε_t. We set

$$\tau_{j,k_t^{(j)}}^{(t)} = \begin{cases} \left[\tau_{j,k_t^{(j)}}^{(t-1)} + \left(\tau^{(t)} - \left(T_{j^*,k_t} - \tau_{j^*,k_t}^{(t-1)}\right) - \cdots - \left(T_{j+1,k_t^{(j+1)}} - \tau_{j+1,k_t^{(j+1)}}^{(t-1)}\right)\right)\right] \\ \quad \wedge T_{j,k_t^{(j)}}, \\ \qquad \text{if } \left(T_{j^*,k_t} - \tau_{j^*,k_t}^{(t-1)}\right) + \cdots + \left(T_{j+1,k_t^{(j+1)}} - \tau_{j+1,k_t^{(j+1)}}^{(t-1)}\right) < \tau^{(t)}, \\ \tau_{j,k_t^{(j)}}^{(t-1)}, \quad \text{otherwise.} \end{cases}$$

For all (j, l) with $l \neq k_t^{(j)}$ we define

$$\tau_{j,l}^{(t)} = \tau_{j,l}^{(t-1)}.$$

After embedding $\varepsilon_1, \ldots, \varepsilon_T$ we arrive at stopping times $\tau_{j,k}^{(T)}$.

(iv) *Embedding of* $\eta_1, \ldots, \eta_T$. We embed $\eta_1, \ldots, \eta_T$ in complete analogy to the embedding of $\varepsilon_1, \ldots, \varepsilon_T$ in the same Wiener processes $W_{j,k}$, $(j,k) \in \mathscr{I}_T$. In this way, we arrive at stopping times $\tilde{\tau}_{j,k}^{(t)}$, which play the same role as the $\tau_{j,k}^{(t)}$'s.

(v) *Choice of the values for* $T_{j,k}$. To motivate our particular choice of the $T_{j,k}$'s we consider first two extreme cases. If $T_{j^*,k} = \infty$, then $Z_{j^*,k}$ and $Z'_{j^*,k}$ are both completely represented by $W_{j^*,k}$. This will lead to a close connection of $Z_{j^*,k}$ and $Z'_{j^*,k}$. However, this choice is not favorable for scales j with $j \ll j^*$. If, for simplicity, $T_{j^*,k} = \infty$ for all k, then the representations of $Z_{j,k}$ and $Z'_{j,k}$, for $j < j^*$, depend very much on the particular values of $\{X_0, \ldots, X_{T-1}\}$ and $\{x_0, \ldots, x_{T-1}\}$. In general, in the case of too large a $T_{j^*,k}$ there will be a tendency that for the representation of $Z_{j,k}$ and $Z'_{j,k}$ too many different stretches of the Wiener processes $W_{j^*,m}$ with $I_{j^*,m} \subset I_{j,k}$ are used, which leads to a suboptimal connection of $Z_{j,k}$ and $Z'_{j,k}$.

On the other hand, if $T_{j^*,k}$ is quite small, then $Z_{j^*,k}$ and $Z'_{j^*,k}$ will be represented in large parts by stretches of Wiener processes $W_{j,m}$, $j < j^*$, which correspond to intervals $I_{j,m} \supset I_{j^*,k}$. Moreover, these stretches used for $Z_{j^*,k}$ will be mostly different from those used for $Z'_{j^*,k}$, and therefore we would get a suboptimal connection of $Z_{j^*,k}$ and $Z'_{j^*,k}$.

To find a good compromise between these two conflicting aims, we choose the $T_{j,k}$'s as large as possible, but with the additional property that the stretches $[0, T_{j,k}]$, $j \neq 0$, are used up in the representation of $\{\varepsilon_1, \ldots, \varepsilon_T\}$ and $\{\eta_1, \ldots, \eta_T\}$ with high probability. Strictly speaking, we choose the $T_{j,k}$'s in such a way that

$$P\Bigg(\sum_t \tau^{(t)} I\big(X_{t-1} \in I_{j,k}\big) < \sum_{(l,m):\ I_{l,m} \subseteq I_{j,k}} T_{l,m} \text{ for any } (j,k) \in \mathscr{I}_T \setminus \{(0,k)\} \Bigg) = O(T^{-\lambda}) \tag{5.6}$$

and

$$P\Bigg(\sum_t \tilde{\tau}^{(t)} I\big(x_{t-1} \in I_{j,k}\big) < \sum_{(l,m):\ I_{l,m} \subseteq I_{j,k}} T_{l,m} \text{ for any } (j,k) \in \mathscr{I}_T \setminus \{(0,k)\} \Bigg) = O(T^{-\lambda}). \tag{5.7}$$

To achieve this, we study first the behavior of the above sums of the stopping times assigned to the interval $I_{j,k}$.

Recall that the conditional distribution of ε_t depends only on X_{t-1}. By taking a closer look at the construction of the Skorokhod embedding described in Hall and Heyde (1980), one can see $\tau^{(t)}$ depends only on ε_t and $\{W^{(t)}(s), 0 \leq s \leq \tau^{(t)}\}$. Since, for $t \neq t'$, $\{W^{(t)}(s), 0 \leq s \leq \tau^{(t)}\}$ and $\{W^{(t')}(s), 0 \leq s \leq \tau^{(t')}\}$ correspond to disjoint stretches of the Wiener processes $W_{j,k}$ separated by stopping times, the random variables $\tau^{(t)} I(X_{t-1} \in I_{j,k})$ are

geometrically β-mixing. Hence, we obtain by Lemma 2.1(ii) that

$$P\left(\left|\sum_{t=1}^{T}\left\{\tau^{(t)}I\left(X_{t-1}\in I_{j,k}\right)-E\left[\tau^{(t)}I\left(X_{t-1}\in I_{j,k}\right)\right]\right\}\right| > \left[C_\lambda\sqrt{T2^{-j}}\log T+T^\delta\right]\right)=O(T^{-\lambda}) \tag{5.8}$$

and, analogously,

$$P\left(\left|\sum_{t=1}^{T}\left\{\tilde\tau^{(t)}I\left(x_{t-1}\in I_{j,h}\right)-E\left[\tilde\tau^{(t)}I\left(x_{t-1}\in I_{j,k}\right)\right]\right\}\right| > \left[C_\lambda\sqrt{T2^{-j}}\log T+T^\delta\right]\right)=O(T^{-\lambda}) \tag{5.9}$$

holds uniformly in $(x_0,\dots,x_{T-1})\in\Omega_T$, where Ω_T is an appropriate set of "not too irregular" realizations of $(X_0,\dots,X_{T-1})$ with $P((X_0,\dots,X_{T-1})\notin\Omega_T)=O(T^{-\lambda})$. Define

$$S_{j,k}=\sum_{t=1}^{T}E\tau^{(t)}I\left(X_{t-1}\in I_{j,k}\right)-\left[C_\lambda\sqrt{T2^{-j}}\log T+T^\delta\right].$$

Further, we define

$$T_{j,k}=S_{j,k}-\sum_{(l,m):\ I_{l,m}\subset I_{j,k}}S_{l,m}.$$

(Then $S_{j,k}=\sum_{(l,m):\ I_{l,m}\subseteq I_{j,k}}T_{l,m}$; that is, $T_{j,k}$ is chosen such that $\{W_{j,k}(s),\ s\in[0,T_{j,k}]\}$ is actually used up with a high probability.)

By (5.8) and (5.9) we obtain (5.6) and (5.7).

(vi) *Conclusions for* $|Z_{j,k}-Z'_{j,k}|$. By (5.6) we obtain with a probability exceeding $1-O(T^{-\lambda})$ that

$$Z_{j,k}=\sum_{(l,m):\ I_{l,m}\subseteq I_{j,k}}W_{l,m}(T_{l,m})+\sum_{t:\ X_{t-1}\in I_{j,k}}\ \sum_{(l,m):\ I_{j,k}\subset I_{l,m}}W_{l,m}\left(\tau_{l,m}^{(t)}\right)-W_{l,m}\left(\tau_{l,m}^{(t-1)}\right), \tag{5.10}$$

and, by (5.7),

$$Z'_{j,k}=\sum_{(l,m):\ I_{l,m}\subseteq I_{j,k}}W_{l,m}(T_{l,m})+\sum_{t:\ x_{t-1}\in I_{j,k}}\ \sum_{(l,m):\ I_{j,k}\subset I_{l,m}}W_{l,m}\left(\tilde\tau_{l,m}^{(t)}\right)-W_{l,m}\left(\tilde\tau_{l,m}^{(t-1)}\right), \tag{5.11}$$

which holds again with a probability exceeding $1-O(T^{-\lambda})$, under the condition $(x_0,\dots,x_{T-1})\in\Omega_T$. At this point we see why our particular pairing of $\varepsilon_1,\dots,\varepsilon_T$ with $\eta_1,\dots,\eta_T$ provides a close connection between $Z_{j,k}$ and $Z'_{j,k}$: most of the randomness of $Z_{j,k}$ and $Z'_{j,k}$ is contained in the first terms on the right-hand side of (5.10) and (5.11), respectively. These terms are random, but

identical to each other. Assume now that both $\sum_{t=1}^{T} \tau^{(t)} I(X_{t-1} \in I_{j,k}) \geq S_{j,k}$ and $\sum_{t=1}^{T} \tilde{\tau}^{(t)} I(x_{t-1} \in I_{j,k})\} \geq S_{j,k}$ are satisfied. By (5.8) and (5.9) we have that

$$\sum_{t:\, X_{t-1} \in I_{j,k}} \sum_{(l,m):\, I_{j,k} \subset I_{l,m}} \tau_{l,m}^{(t)} - \tau_{l,m}^{(t-1)} = \sum_{t:\, X_{t-1} \in I_{j,k}} \tau^{(t)} - S_{j,k}$$
$$= \tilde{O}\left(\sqrt{T 2^{-j}} \log T + T^{\delta}, T^{-\lambda}\right)$$

and

$$\sum_{t:\, x_{t-1} \in I_{j,k}} \sum_{(l,m):\, I_{j,k} \subset I_{l,m}} \tilde{\tau}_{l,m}^{(t)} - \tilde{\tau}_{l,m}^{(t-1)} = \sum_{t:\, x_{t-1} \in I_{j,k}} \tilde{\tau}^{(t)} - S_{j,k}$$
$$= \tilde{O}\left(\sqrt{T 2^{-j}} \log T + T^{\delta}, T^{-\lambda}\right).$$

Note that, for fixed t and under $X_{t-1} \in I_{j,k}$, the pieces $\{W_{l,m}(s), \tau_{l,m}^{(t-1)} \leq s \leq \tau_{l,m}^{(t)}\}$ of the Wiener processes $W_{l,m}$ corresponding to intervals $I_{l,m} \supset I_{j,k}$ can be composed to a piece of a new Wiener process $W_{j,k}^{\mathrm{res},t}$ on the interval $[0, \tau_{j,k}^{\mathrm{res},t}]$, where $\tau_{j,k}^{\mathrm{res},t} = \sum_{(l,m):\, I_{j,k} \subset I_{l,m}} (\tau_{l,m}^{(t)} - \tau_{l,m}^{(t-1)})$. This is achieved by setting

$$W_{j,k}^{\mathrm{res},t}(s)$$
$$= \begin{cases} W_{j-1,[k/2]}\left(s + \tau_{j-1,[k/2]}^{(t-1)}\right) - W_{j-1,[k/2]}\left(\tau_{j-1,[k/2]}^{(t-1)}\right), \\ \qquad\qquad \text{if } 0 \leq s \leq \tau_{j-1,[k/2]}^{(t)} - \tau_{j-1,[k/2]}^{(t-1)}, \\ \left[W_{j-1,[k/2]}\left(\tau_{j-1,[k/2]}^{(t)}\right) - W_{j-1,[k/2]}\left(\tau_{j-1,[k/2]}^{(t-1)}\right)\right] \\ \qquad + \cdots + \left[W_{l+1,[k2^{l+1-j}]}\left(\tau_{l+1,[k2^{l+1-j}]}^{(t)}\right)\right. \\ \qquad\qquad \left. - W_{l+1,[k2^{l+1-j}]}\left(\tau_{l+1,[k2^{l+1-j}]}^{(t-1)}\right)\right] \\ + \left[W_{l,[k2^{l-j}]}(u) - W_{l,[k2^{l-j}]}\left(\tau_{l,[k2^{l-j}]}^{(t-1)}\right)\right], \text{if } s = \left(\tau_{j-1,[k/2]}^{(t)} - \tau_{j-1,[k/2]}^{(t-1)}\right) \\ \qquad + \cdots + \left(\tau_{l+1,[k2^{l+1-j}]}^{(t)} - \tau_{l+1,[k2^{l+1-j}]}^{(t-1)}\right) + \left(u - \tau_{l,[k2^{l-j}]}^{(t-1)}\right) \\ \qquad \text{and } u \leq \tau_{l,[k2^{l-j}]}^{(t)}. \end{cases}$$

(In the case of $X_{t-1} \notin I_{j,k}$ we simply set $\tau_{j,k}^{\mathrm{res},t} = 0$.)

Note that $\{W_{j,k}^{\mathrm{res},t}(s), 0 \leq s \leq \tau_{j,k}^{\mathrm{res},t}\}$ is $\mathscr{F}_t$-measurable. By the strong Markov property, the remaining parts of the Wiener processes $W_{j,k}$, that is, $\{W_{j,k}(s + \tau_{j,k}^{(t)}) - W_{j,k}(\tau_{j,k}^{(t)}), 0 \leq s < \infty\}$, form again independent Wiener processes, which are also independent of $\mathscr{F}_t$. Hence, we can compose all these parts of $W_{j,k}^{\mathrm{res},t}$ considered above to a Wiener process $W_{j,k}^{\mathrm{res}}$ by setting

$$W_{j,k}^{\mathrm{res}}(s) = \begin{cases} W_{j,k}^{\mathrm{res},1}(s), \quad \text{if } 0 \leq s \leq \tau_{j,k}^{\mathrm{res},1}, \\ W_{j,k}^{\mathrm{res},1}\left(\tau_{j,k}^{\mathrm{res},1}\right) + \cdots + W_{j,k}^{\mathrm{res},u-1}\left(\tau_{j,k}^{\mathrm{res},u-1}\right) \\ \quad + W_{j,k}^{\mathrm{res},u}\left(s - \tau_{j,k}^{\mathrm{res},1} - \cdots - \tau_{j,k}^{\mathrm{res},u-1}\right), \\ \qquad\qquad \text{if } \tau_{j,k}^{\mathrm{res},1} + \cdots + \tau_{j,k}^{\mathrm{res},u-1} < s \leq \tau_{j,k}^{\mathrm{res},1} + \cdots + \tau_{j,k}^{\mathrm{res},u}. \end{cases}$$

An analogous construction can be made for the $\tilde{\tau}_{l,m}^{(t)}$'s, leading to a Wiener process $\tilde{W}_{j,k}^{\mathrm{res}}$.

Note that $\tau_{j,k}^{\mathrm{res},1} + \cdots + \tau_{j,k}^{\mathrm{res},T} = \sum_{t:\, X_{t-1} \in I_{j,k}} \tau^{(t)} - S_{j,k}$. Now we obtain by Lemma 1.2.1 in Csörgő and Révész [(1981), page 29], that

$$\begin{aligned}
|Z_{j,k} - Z'_{j,k}| &\le \left| \sum_{t:\, X_{t-1} \in I_{j,k}} \sum_{(l,m):\, l>j,\, I_{j,k} \subset I_{l,m}} W_{l,m}\bigl(\tau_{l,m}^{(t)}\bigr) - W_{l,m}\bigl(\tau_{l,m}^{(t-1)}\bigr) \right| \\
&\quad + \left| \sum_{t:\, x_{t-1} \in I_{j,k}} \sum_{(l,m):\, l>j,\, I_{j,k} \subset I_{l,m}} W_{l,m}\bigl(\tilde{\tau}_{l,m}^{(t)}\bigr) - W_{l,m}\bigl(\tilde{\tau}_{l,m}^{(t-1)}\bigr) \right| \\
&= \left| W_{j,k}^{\mathrm{res}}\left(\sum_{t:\, X_{t-1} \in I_{j,k}} \tau^{(t)} - S_{j,k} \right) \right| + \left| \tilde{W}_{j,k}^{\mathrm{res}}\left(\sum_{t:\, x_{t-1} \in I_{j,k}} \tilde{\tau}^{(t)} - S_{j,k} \right) \right| \\
&= \tilde{O}\bigl([T2^{-j}]^{1/4} \log T + T^{\delta}, T^{-\lambda} \bigr),
\end{aligned}$$

which completes the proof. □

PROOF OF COROLLARY 2.1. Let $(x_0, \ldots, x_{T-1}) \in \Omega_T$. We assume throughout the proof that

$$|Z_{j,k} - Z'_{j,k}| \le C_{\lambda} \bigl\{ (T2^{-j})^{1/4} \log T + T^{\delta} \bigr\} \tag{5.12}$$

is fulfilled for all $(j,k) \in \mathscr{I}_T$, which holds true by Theorem 2.1 with a probability of $1 - O(T^{-\lambda})$.

Let $[c,d)$ be an arbitrary interval. Beginning from the coarsest scale, we approximate $[c,d)$ from below by a union of as large as possible intervals. There exist indices, $(j_1,k_1), \ldots, (j_m,k_m)$, such that

$$I_{j_1,k_1} \cup \cdots \cup I_{j_m,k_m} \subseteq [c,d),$$

where $j_1 < \cdots < j_{m-1} \le j_m \le j^*$. Here we choose j^* such that $(T2^{-j^*})^{1/4} \log T \asymp T^{\delta}$.

On the other hand, if we add two suitable intervals from the finest scale, say I_{j^*,l_1} and I_{j^*,l_2}, then we can approximate $[c,d)$ from above, that is,

$$[c,d) \subseteq I_{j_1,k_1} \cup \cdots \cup I_{j_m,k_m} \cup I_{j^*,l_1} \cup I_{j^*,l_2}.$$

Now we obtain, under (5.12), that

$$\begin{aligned}
&\left| \sum_{t:\, X_{t-1} \in [c,d)} \varepsilon_t - \sum_{t:\, x_{t-1} \in [c,d)} \eta_t \right| \\
&\quad \le \sum_{i=1}^{m} |Z_{j_i,k_i} - Z'_{j_i,k_i}| + |Z_{j^*,l_1}| + |Z_{j^*,l_2}| + |Z'_{j^*,l_1}| + |Z'_{j^*,l_2}| \\
&\quad = O\left(\sum_{i=1}^{m} \left[(T2^{-j_i})^{1/4} \log T + T^{\delta} \right] + T^{\delta} \right) \\
&\quad = O\bigl((T2^{-j_1})^{1/4} \log T + T^{\delta} \bigr) \\
&\quad = O\bigl([TP(X_0 \in [c,d))]^{1/4} \log T + T^{\delta} \bigr).
\end{aligned}$$

□

PROOF OF COROLLARY 2.2. We approximate w by a truncated Haar wavelet series expansion

$$\tilde{w}(z) = \sum_k \alpha_k \phi_k(z) + \sum_{0 \le j < j^*} \sum_k \alpha_{j,k} \psi_{j,k}(z), \tag{5.13}$$

where $\alpha_k = \int \phi_k(z) w(z)\, dz$, $\alpha_{j,k} = \int \psi_{j,k}(z) w(z)\, dz$ and $\phi_k(z) = I(k-1 \le z < k)$,

$$\psi_{j,k}(z) = \begin{cases} 2^{j/2}, & \text{if } (k-1)2^{-j} \le z < (k-1/2)2^{-j}, \\ -2^{j/2}, & \text{if } (k-1/2)2^{-j} \le z < k2^{-j}, \\ 0, & \text{otherwise.} \end{cases}$$

In view of the following calculations, we choose j^* such that $T2^{-j^*} \asymp T^{\delta}$. It holds that

$$\sum_k |\alpha_k| = O(\|w\|_1) \tag{5.14}$$

and

$$\begin{aligned} \sum_k |\alpha_{j,k}| &= O\big(\min\{\|\psi_{j,k}\|_\infty \|w\|_1, \|\psi_{j,k}\|_1 TV(w)\}\big) \\ &= O\big(\min\{2^{j/2}\|w\|_1, 2^{-j/2} TV(w)\}\big). \end{aligned} \tag{5.15}$$

This implies, with $\tilde{Z}_{j,k} = \sum_t I(X_{t-1} \in [(k-1)2^{-j}, k2^{-j}))\varepsilon_t$ and $\tilde{Z}'_{j,k} = \sum_t I(x_{t-1} \in [(k-1)2^{-j}, k2^{-j}))\eta_t$, that

$$\begin{aligned} &\left| \sum_{t=1}^{T} \tilde{w}(X_{t-1})\varepsilon_t - \sum_{t=1}^{T} \tilde{w}(x_{t-1})\eta_t \right| \\ &\quad \le \left| \sum_k \alpha_k \big[\tilde{Z}_{0,k} - \tilde{Z}'_{0,k}\big] \right| \\ &\qquad + \left| \sum_{0 \le j < j^*} \sum_k \alpha_{j,k} \sum_t \big[\psi_{j,k}(X_{t-1})\varepsilon_t - \psi_{j,k}(x_{t-1})\eta_t\big] \right| \\ &\quad \le \tilde{O}\big(\|w\|_1 T^{1/4} \log T, T^{-\lambda}\big) \\ &\qquad + \sum_{0 \le j < j^*} \sum_k |\alpha_{j,k}|\, \|\psi_{j,k}\|_\infty \max_l \big\{|\tilde{Z}_{j+1,l} - \tilde{Z}'_{j+1,l}|\big\} \\ &\quad = \tilde{O}\big(\|w\|_1 T^{1/4} \log T, T^{-\lambda}\big) \\ &\qquad + \sum_{0 \le j < j^*} \tilde{O}\big(\min\{2^{j/2}\|w\|_1, 2^{-j/2} TV(w)\} 2^{j/2} (T2^{-j})^{1/4} \log T, T^{-\lambda}\big) \\ &\quad = \tilde{O}\big(T^{1/4} (TV(w))^{3/4} \|w\|_1^{1/4} \log T, T^{-\lambda}\big). \end{aligned} \tag{5.16}$$

Further, we have

$$\sum_k \|w - \tilde{w}\|_{L_\infty([(k-1)2^{-j^*}, k2^{-j^*}))} \le TV(w),$$

which implies that

$$\begin{aligned}
&\sum_t (w(X_{t-1}) - \tilde{w}(X_{t-1}))\varepsilon_t \\
&\quad = \sum_k \sum_{t:\, X_{t-1} \in [(k-1)2^{-j^*}, k2^{-j^*})} (w(X_{t-1}) - \tilde{w}(X_{t-1}))\varepsilon_t \\
&\quad = \tilde{O}\big(T2^{-j^*}TV(w)T^{\delta/2}, T^{-\lambda}\big),
\end{aligned} \tag{5.17}$$

and, analogously,

$$\sum_t (w(x_{t-1}) - \tilde{w}(x_{t-1}))\eta_t = \tilde{O}\big(T2^{-j^*}TV(w)T^{\delta/2}, T^{-\lambda}\big). \tag{5.18}$$

The assertion follows now from (5.16) to (5.18).

PROOF OF LEMMA 2.2. First we investigate how well the random quantity $(D_x' K_x D_x)_{ij}$ is approximated by its expectation. We have $(D_x' K_x D_x)_{ij} = \sum g(X_{t-1})$, where $g(z) = K((x-z)/h)((x-z)/h)^{i+j-2}$. Note that, for $x \in [a, b]$, g is supported on $[a - h, b + h]$. We approximate g by a truncated Haar wavelet series expansion

$$\tilde{g}(z) = \sum_k \beta_k \phi_k(z) + \sum_{0 \le j < j^*} \sum_k \beta_{j,k} \psi_{j,k}(z), \tag{5.19}$$

where $\beta_k = \int \phi_k(z) g(z)\, dz$, $\beta_{j,k} = \int \psi_{j,k}(z) g(z)\, dz$. In view of the following calculations we choose j^* such that $T2^{-j^*} \asymp \sqrt{Th}$. It holds that

$$\sum_k |\beta_k| \le \|g\|_{L_1} = O(h) \tag{5.20}$$

and

$$\begin{aligned}
\sum_k |\beta_{j,k}| &= O\big(\min\{\|\psi_{j,k}\|_\infty \|g\|_1, \|\psi_{j,k}\|_1 TV(g)\}\big) \\
&= O\big(\min\{2^{j/2}h, 2^{-j/2}\}\big).
\end{aligned} \tag{5.21}$$

Define $F_T(z) = \sum_{t=1}^T I(X_{t-1} < z)$ and $F_T^{(\infty)} = TP(X_{t-1} < z)$. As an immediate consequence of Lemma 2.1 we obtain that

$$\begin{aligned}
&\#\{t:\, X_{t-1} \in [(k-1)2^{-j}, k2^{-j})\} - TP\big(X_{t-1} \in [(k-1)2^{-j}, k2^{-j})\big) \\
&\quad = \tilde{O}\Big(\min\Big\{\sqrt{T2^{-j}}\log T + (\log T)^2, \sqrt{T \log T}\Big\}, T^{-\lambda}\Big)
\end{aligned} \tag{5.22}$$

holds uniformly in $(j,k) \in \mathscr{I}_T$. Then, by (5.20), (5.21) and (5.22),

$$\begin{aligned}
&\left| \sum_{t=1}^{T} \tilde{g}(X_{t-1}) - \sum_{t=1}^{T} E\tilde{g}(X_{t-1}) \right| \\
&\quad = \left| \int \tilde{g}(z)\, dF_T(z) - \int \tilde{g}(z)\, dF_T^{(\infty)}(z) \right| \\
&\quad \le \sum_k |\beta_k| \left| \left(F_T(k) - F_T(k-1)\right) - \left(F_T^{(\infty)}(k) - F_T^{(\infty)}(k-1)\right) \right| \\
&\qquad + \sum_{0 \le j < j^*} \sum_k |\beta_{j,k}| \left| \int \psi_{j,k}(z) \left[dF_T(z) - dF_T^{(\infty)}(z) \right] \right| \\
&\quad = \tilde{O}\left(h\sqrt{T \log T}, T^{-\lambda}\right) \\
&\qquad + \sum_{j:\, 2^j \le h^{-1}} O(2^{j/2}h) O(2^{j/2}) \tilde{O}\left(\sqrt{T2^{-j}} \log T, T^{-\lambda}\right) \\
&\qquad + \sum_{j:\, j < j^*, 2^j > h^{-1}} O(2^{-j/2}) O(2^{j/2}) \tilde{O}\left(\sqrt{T2^{-j}} \log T, T^{-\lambda}\right) \\
&\quad = \tilde{O}\left(\sqrt{Th} \log T, T^{-\lambda}\right).
\end{aligned} \tag{5.23}$$

Since $\tilde{g}$ is the best piecewise constant approximation to g, that is $\tilde{g}(z) = 2^{-j^*} \int_{(k-1)2^{-j^*}}^{k2^{-j^*}} g(z)\, dz$ if $z \in [(k-1)2^{-j^*}, k2^{-j^*})$, we get

$$\sum_k \|g - \tilde{g}\|_{L_\infty([(k-1)2^{-j^*}, k2^{-j^*}))} \le TV(g) = O(1).$$

This implies that

$$\begin{aligned}
&\sum_t g(X_{t-1}) - \tilde{g}(X_{t-1}) \\
&\quad = \sum_k \sum_{t:\, X_{t-1} \in [(k-1)2^{-j^*}, k2^{-j^*})} g(X_{t-1}) - \tilde{g}(X_{t-1}) \\
&\quad \le \sum_k \|g - \tilde{g}\|_{L_\infty([(k-1)2^{-j^*}, k2^{-j^*}))} \#\left\{t:\, X_{t-1} \in I_{j^*,k}\right\} \\
&\quad = \tilde{O}\left(T2^{-j^*} + \sqrt{T2^{-j^*}} \log T, T^{-\lambda}\right),
\end{aligned} \tag{5.24}$$

and, analogously,

$$\sum_t Eg(X_{t-1}) - E\tilde{g}(X_{t-1}) = O(T2^{-j^*}). \tag{5.25}$$

From (5.23) to (5.25) we obtain that

$$|(D_x' K_x D_x)_{ij} - E(D_x' K_x D_x)_{ij}| = \tilde{O}\left(\sqrt{Th} \log T, T^{-\lambda}\right),$$

which implies

$$\|D_x' K_x D_x - ED_x' K_x D_x\| = \tilde{O}\left(\sqrt{Th} \log T, T^{-\lambda}\right). \tag{5.26}$$

According to Lemma 1 of Tsybakov (1986), we have

$$ED_x' K_x D_x \geq CTh\left(\left(\int_{-1}^{1} K(z) z^{i+j-2}\, dz\right)\right)_{i,j=1,\ldots,p}, \tag{5.27}$$

where $\lambda_{\min}(((\int_{-1}^{1} K(z) z^{i+j-2}\, dz))_{i,j=1,\ldots,p}) > 0$. Hence,

$$\begin{aligned} &\|(D_x' K_x D_x)^{-1} - (ED_x' K_x D_x)^{-1}\| \\ &\leq \|(D_x' K_x D_x)^{-1}\| \, \|D_x' K_x D_x - ED_x' K_x D_x\| \, \|(ED_x' K_x D_x)^{-1}\| \\ &= \tilde{O}\left((Th)^{-3/2} \log T, T^{-\lambda}\right), \end{aligned} \tag{5.28}$$

which proves the assertion.

PROOF OF PROPOSITION 2.1. By $\|K((x-\cdot)/h)((x-\cdot)/h)^q\|_1 = O(h)$ and $TV(K((x-\cdot)/h)((x-\cdot)/h)^q) = O(1)$, we conclude from Corollary 2.2 that

$$\begin{aligned} &\sum_t K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q \varepsilon_t \\ &= \sum_t K\left(\frac{x - x_{t-1}}{h}\right)\left(\frac{x - x_{t-1}}{h}\right)^q \eta_t + \tilde{O}\left((Th)^{1/4} T^{\delta}, T^{-\lambda}\right) \end{aligned} \tag{5.29}$$

holds for $(x_0, \ldots, x_{T-1}) \in \Omega_T$, Ω_T according to Theorem 2.1.

For $(x_0, \ldots, x_{T-1}) \in \Omega_T$ we obtain by Theorem 4 in Amosova (1972) that

$$\tilde{Z}'_{j,k} = \sum_{t:\, x_{t-1} \in [(k-1)2^{-j}, k2^{-j})} \eta_t = \tilde{O}\left(\sqrt{T 2^{-j}} \sqrt{\log T}, T^{-\lambda}\right),$$

which implies, by calculations similar to those in the proof of Corollary 2.2 below, that

$$\sup_{x \in [a,b]} \left\{\left| \sum_t K\left(\frac{x - x_{t-1}}{h}\right)\left(\frac{x - x_{t-1}}{h}\right)^q \eta_t \right|\right\} = \tilde{O}\left(\sqrt{Th} \sqrt{\log T}, T^{-\lambda}\right). \tag{5.30}$$

Using now Lemma 2.2, (2.8) and (2.9) we obtain the assertions. □

PROOF OF PROPOSITION 2.2. According to Lemma 1.4.2 of Korostelev and Tsybakov (1993), we have in the case that $D_x' K_x D_x$ is regular that $\sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) = 1$ and $\sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\})(X_{t-1} - x)^q = 0$ for $q = 1, \ldots, p-1$. Provided this is true (which is actually the case with a probability exceeding $1 - O(T^{-\lambda})$), we get from a Taylor series expan-

sion with integral remainder that

$$\begin{aligned}
&\sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) m(X_{t-1}) - m(x) \\
&\quad = \sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) \int_x^{X_{t-1}} \frac{(X_{t-1} - s)^{p-1}}{(p-1)!} m^{(p)}(s)\, ds \\
&\quad = \sum_{q=0}^{p-1} d_q(x, \{X_0, \ldots, X_{T-1}\}) \sum_t K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q \\
&\qquad\qquad \times \int_x^{X_{t-1}} \frac{(X_{t-1} - s)^{p-1}}{(p-1)!} m^{(p)}(s)\, ds.
\end{aligned}$$

Since $g(z) = K((x-z)/h)((x-z)/h)^q \int_x^z ((z-s)^{p-1}/(p-1)!) m^{(p)}(s)\, ds$ satisfies $\|g\|_q = O(h^{p+1})$ and $TV(g) = O(h^p)$, we obtain analogously to (5.26) that

$$\begin{aligned}
&\sum_t K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q \int_x^{X_{t-1}} \frac{(X_{t-1} - s)^{p-1}}{(p-1)!} m^{(p)}(s)\, ds \\
&\quad = E \sum_t K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q \int_x^{X_{t-1}} \frac{(X_{t-1} - s)^{p-1}}{(p-1)!} m^{(p)}(s)\, ds \\
&\qquad + \tilde{O}\left(h^p \sqrt{Th}\, \log T, T^{-\lambda}\right).
\end{aligned}$$

Since $E\sum_t K((x - X_{t-1})/h)((x - X_{t-1})/h)^q \int_x^{X_{t-1}} ((X_{t-1} - s)^{p-1}/(p-1)!) m^{(p)}(s)\, ds = O(Th^{p+1})$, we obtain, in conjunction with Lemma 2.2, that

$$\begin{aligned}
&\sup_{x \in [a,b]} \left\{ \left| \sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) m(X_{t-1}) - m(x) - b_\infty(x) \right| \right\} \\
&\quad = \tilde{O}\left(h^p (Th)^{-1/2} \log T, T^{-\lambda}\right). \qquad \square
\end{aligned}$$

PROOF OF PROPOSITION 2.3. By $\|\overline{w}_h\|_1 = O(T^{-1})$ and $TV(\overline{w}_h) = O((Th)^{-1})$, the assertion follows immediately from Corollary 2.2. $\square$

PROOF OF LEMMA 3.1. This proof is similar to that of Theorem 2.1 in Neumann and Polzehl (1995). In order to prove the assertion we introduce independent random variables $\xi_t \sim N(0, \mathrm{var}(\eta_t))$ as well as a second set of independent random variables in the bootstrap domain $\xi_t^* \sim N(0, \mathrm{var}(\varepsilon_t^*))$, whose relationship among each other as well as to the η_t's and the ε_t^*'s is described below.

We split up as follows:

$$\begin{aligned}\sum_t \overline{w}_h(x, x_{t-1})\eta_t - \sum_t \overline{w}_h(x, x_{t-1})\varepsilon_t^* &= \sum_t \overline{w}_h(x, x_{t-1})(\eta_t - \xi_t) + \sum_t \overline{w}_h(x, x_{t-1})(\xi_t - \xi_t^*) \\ &\quad + \sum_t \overline{w}_h(x, x_{t-1})(\xi_t^* - \varepsilon_t^*) \\ &= S_1(x) + S_2(x) + S_3(x).\end{aligned} \tag{5.31}$$

First we pair the random variables $\xi_1, \ldots, \xi_T$ with the random variables $\xi_1^*, \ldots, \xi_T^*$ in such a way that $S_2(x)$ is as small as possible. Some motivation for the particular construction used here is given in Neumann and Polzehl (1995).

We decompose the error vectors $\underline{\xi} = (\xi_1, \ldots, \xi_T)'$ and $\underline{\xi}^* = (\xi_1^*, \ldots, \xi_T^*)'$ into $\Delta \asymp h^{-1}$ packages of length $d_j \asymp Th$, respectively, that is,

$$\underline{\xi} = \left(\xi_{11}, \ldots, \xi_{1d_1}, \ldots, \xi_{\Delta 1}, \ldots, \xi_{\Delta d_\Delta}\right)'. \tag{5.32}$$

($\underline{\xi}^*$ is split up analogously.)

Let $v_{jk} = E\xi_{jk}^2$, $v_{jk}^* = E\xi_{jk}^{*2}$ and $w_{jk}(x) = \overline{w}_h(x, x_{t-1})$, if t corresponds to (j, k) in (5.32). Further, let $V_j = \sum_{k=1}^{d_j} v_{jk}$, $V_j^* = \sum_{k=1}^{d_j} v_{jk}^*$ $(j = 1, \ldots, \Delta)$. We define

$$\begin{aligned} t_{jk} &= \sum_{l \le k} v_{jl}, \qquad t_{jk}^* = \sum_{l \le k} v_{jl}^*, \\ s_{jk} &= (j-1) + t_{jk}/V_j, \qquad s_{jk}^* = (j-1) + t_{jk}^*/V_j^*. \end{aligned}$$

Now we represent the ξ_t's as well as the ξ_t^*'s by one and the same Wiener process $W(t)$; namely, we set

$$\xi_{jk} = V_j^{1/2}\big(W(s_{jk}) - W(s_{j,k-1})\big)$$

and

$$\xi_{jk}^* = V_j^{*1/2}\big(W(s_{jk}^*) - W(s_{j,k-1}^*)\big).$$

It is clear that the ξ_t's as well as the ξ_t^*'s are independent and have the desired distributions.

Now we decompose $S_2(x)$ in a "coarse structure" term

$$S_{21}(x) = \sum_j \left(V_j^{1/2} - V_j^{*1/2}\right)\sum_k w_{jk}(x)\big(W(s_{jk}^*) - W(s_{j,k-1}^*)\big)$$

and a "fine structure" term

$$S_{22}(x) = \sum_j V_j^{1/2}\sum_k w_{jk}(x)\big[\big(W(s_{jk}) - W(s_{j,k-1})\big) - \big(W(s_{jk}^*) - W(s_{j,k-1}^*)\big)\big].$$

We can easily show that

$$\begin{aligned}\max_{j,k}\left\{|t_{jk} - t_{jk}^*|\right\} &= \sum_{l \le k}\left(\varepsilon_{jl}^2 - v_{jl}\right) + \sum_{l \le k}\left(\hat{\varepsilon}_{jl}^2 - \varepsilon_{jl}^2\right) \\ &= \tilde{O}\big((Th)^{1/2}T^{\delta}, T^{-\lambda}\big),\end{aligned} \tag{5.33}$$

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

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

Therefore we have

$$\sup_x \{|S_{21}(x)|\} = \tilde{O}\big((Th)^{-1} T^\delta, T^{-\lambda}\big). \tag{5.34}$$

We rewrite

$$\begin{aligned} S_{22}(x) &= \sum_j V_j^{1/2} \sum_k w_{jk}(x) \left[\int_{s_{j,k-1}}^{s_{jk}} dW(t) - \int_{s_{j,k-1}^*}^{s_{jk}^*} dW(t) \right] \\ &= \sum_j V_j^{1/2} \int_{j-1}^{j} [\, w_t - w_t^* \,]\, dW(t), \end{aligned}$$

where $w_t = w_{j,k}(x)$, if $t \in (s_{j,k-1}, s_{jk}]$, and $w_t^* = w_{j,k}(x)$, if $t \in (s_{j,k-1}^*, s_{jk}^*]$.

By (5.33) and $w_{j,k}(x) - w_{j,k+1}(x) = O((Th)^{-2})$ we acquire $\sup_t\{|w_t - w_t^*|\} = \tilde{O}((Th)^{-3/2} T^\delta, T^{-\lambda})$, which implies that

$$S_{22}(x) = \tilde{O}\big((Th)^{-1} T^\delta, T^{-\lambda}\big). \tag{5.35}$$

To get a favorable pairing of the η_t's with the ξ_t's we consider the partial sum processes

$$P_t = \sum_{s \le t} \eta_s \quad \text{and} \quad \tilde{P}_t = \sum_{s \le t} \xi_s.$$

According to Corollary 4 in Sakhanenko [(1991), page 76], there exists a pairing of the ξ_j's and ξ_j^*'s, on a sufficiently rich probability space, such that

$$\max_{1 \le t \le T} \left\{ |P_t - \tilde{P}_t| \right\} = \tilde{O}(T^\delta, T^{-\lambda}),$$

which implies by $TV(\overline{w}_h(x, \cdot)) = O((Th)^{-1})$ that

$$\begin{aligned} \sup_{x \in [a,b]} \{|S_1(x)|\} &\le \sup_x \left\{ \sum_{t=1}^{T-1} |\overline{w}_h(x, x_{t-1}) - \overline{w}_h(x, x_t)|\, |P_t - \tilde{P}_t| \right. \\ &\qquad \left. + |\overline{w}_h(x, x_{T-1})|\, |P_T - \tilde{P}_T| \right\} \\ &= \tilde{O}\big((Th)^{-1} T^\delta, T^{-\lambda}\big). \end{aligned} \tag{5.36}$$

Analogously we can find a pairing of the ξ_t^*'s with the $\hat{\xi}_t$'s such that

$$\sup_{x \in [a,b]} \{|S_3(x)|\} = \tilde{O}\big((Th)^{-1} T^\delta, T^{-\lambda}\big). \tag{5.37}$$

The assertion follows now from (5.31) and (5.34) to (5.37). □

PROOF OF THEOREM 3.2. By Theorem 3.1, there exists a pairing of $X_0, \varepsilon_1, \ldots, \varepsilon_T$ with $\varepsilon_1^*, \ldots, \varepsilon_T^*$ such that

$$U_T - U_T^* = \tilde{O}\big((Th)^{-3/4} \log T, T^{-\lambda}\big)$$

holds, if $(X_0, \ldots, X_{T-1}) \in \Omega_T$. This implies, in conjunction with Lemma 3.2, that

$$\begin{aligned} &\sup_t \{|P(U_T < t) - P(U_T^* < t \mid X_0, \ldots, X_T)|\} \\ &\quad = O\big((Th)^{-1/4}(\log T)^{3/2} + (Th)^{-1/2}T^{\delta}\big) \end{aligned} \tag{5.38}$$

holds uniformly in $(X_0, \ldots, X_T) \in \Omega_T$. Hence, we obtain in particular that

$$\begin{aligned} P(U_T < t)\mid_{t=t_\alpha^*} &= P(U_T^* < t_\alpha^* \mid X_0, \ldots, X_T) + O\big((Th)^{-1/4}(\log T)^{3/2}\big) \\ &= 1 - \alpha + O\big((Th)^{-1/4}(\log T)^{3/2}\big), \end{aligned} \tag{5.39}$$

again for $(X_0, \ldots, X_T) \in \Omega_T$. In other words, for $(X_0, \ldots, X_T) \in \Omega_T$, t_α^* is between two *nonrandom* bounds $t_{\alpha,1}$ and $t_{\alpha,2}$ with $P(U_T < t_{\alpha,i}) = 1 - \alpha + O((Th)^{-1/4}(\log T)^{3/2})$, $i = 1, 2$. This implies (i).

The proof of (ii) follows from the same reasoning and the fact that

$$\sup_{x \in [a,b]} \{|\hat{V}(x) - V(x)|\} = \tilde{O}\big(T^{\delta}(Th)^{-3/2}, T^{-\lambda}\big). \qquad \square$$

Acknowledgment. The second author gratefully acknowledges the hospitality and support of the SFB 373 at Humboldt University and of the Weierstrass Institute, Berlin. We thank two anonymous referees and an Associate Editor for their suggestions that helped to improve an earlier version of this paper.

REFERENCES

AMOSOVA, N. N. (1972). On limit theorems for probabilities of moderate deviations. *Vestnik Leningrad. Univ.* **13** 5–14 (in Russian).

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

CSÖRGŐ, M. and RÉVÉSZ, P. (1981). *Strong Approximations in Probability and Statistics*. Academic, New York.

DAVYDOV, YU. A. (1970). The invariance principle for stationary processes. *Theory Probab. Appl.* **15** 487–498.

DOUKHAN, P. (1994). *Mixing*: *Properties and Examples*. *Lecture Notes in Statist.* **85**. Springer, New York.

DOUKHAN, P., MASSART, P. and RIO, E. (1995). Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.* **31** 393–427.

FAN, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87** 998–1004.

FAN, J. (1993). Local linear regression smoothers and their minimax efficiencies. *Ann. Statist.* **21** 196–216.

FAN, J. and GIJBELS, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20** 2008–2036.

FAN, J. and GIJBELS, I. (1995). *Local Polynomial Modeling and Its Application—Theory and Methodologies*. Chapman and Hall, New York.

FRANKE, J., KREISS, J.-P. and MAMMEN, E. (1997). Bootstrap of kernel smoothing in nonlinear time series. Discussion Paper 20/97, SFB 373, Humboldt Univ., Berlin.

FRANKE, J., KREISS, J.-P., MAMMEN, E. and NEUMANN, M. H. (1998). Properties of the nonparametric autoregressive bootstrap. Unpublished manuscript.

HALL, P. (1991). On the distribution of suprema. *Probab. Theory Related Fields* **89** 447–455.
HALL, P. and HEYDE, C. C. (1980). *Martingale Limit Theory and Its Application*. Academic Press, New York.
HALL, P. and TITTERINGTON, M. (1988). On confidence bands in nonparametric density estimation and regression. *J. Multvariate Anal.* **27** 228–254.
HÄRDLE, W. and MAMMEN, E. (1993). Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21** 1926–1947.
HÄRDLE, W. and TSYBAKOV, A. B. (1997). Local polynomial estimators of the volatility function in nonparametric autoregression. *J. Econometrics* **81** 223–242.
HART, J. D. (1995). Some automated methods of smoothing time-dependent data. *J. Nonparametr. Statist.* **6** 115–142.
JONES, D. A. (1978). Nonlinear regressive processes. *Proc. Roy. Soc. London Ser. A* **360** 71–95.
KNAFL, G., SACKS, J. and YLVISAKER, D. (1985). Confidence bands for regression functions. *J. Amer. Statist. Assoc.* **80** 683–691.
KOMLÓS, J., MAJOR, P. and TUSNÁDY, G. (1975). An approximation of partial sums of independent rv's and the sample distribution function. *Z. Wahrsch. Verw. Gebiete* **32** 111–131.
KONAKOV, V., LÄUTER, H. and LIERO, H. (1995). Comparison of the asymptotic power of tests based on L_2- and L_∞-norms under non-standard local alternatives. Discussion Paper 10/95, SFB 373, Humboldt Univ., Berlin.
KOROSTELEV, A. P. and TSYBAKOV, A. B. (1993). *Minimax Theory of Image Reconstruction*. *Lecture Notes in Statist.* **82**. Springer, New York.
KREISS, J.-P., NEUMANN, M. H. and YAO, Q. (1998). Bootstrap tests for simple structures in nonparametric time series regression. Preprint No. 98/07, TU Braunschweig.
MAMMEN, E. (1992). *When Does Bootstrap Work*? *Asymptotic Results and Simulations*. *Lecture Notes in Statist.* **77**. Springer, New York.
MASRY, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Ser. Anal.* **17** 571–599.
MASRY, E. and TJØSTHEIM, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory* **11** 258–289.
NEUMANN, M. H. (1995). Automatic bandwidth choice and confidence intervals in nonparametric regression. *Ann. Statist.* **23** 1937–1959.
NEUMANN, M. H. (1997). On robustness of model-based bootstrap schemes in nonparametric time series analysis. Discussion Paper 88/97, SFB 373, Humboldt Univ., Berlin.
NEUMANN, M. H. (1998). Strong approximation of density estimators from weakly dependent observations by density estimators from independent observations. *Ann. Statist.* **26** To appear.
NEUMANN, M. H. and KREISS, J.-P. (1996). Bootstrap confidence bands for the autoregression function. Preprint 263, Weierstrass Institute, Berlin.
NEUMANN, M. H. and POLZEHL, J. (1995). Simultaneous bootstrap confidence bands in nonparametric regression. *J. Nonparametr. Statist.* To appear.
NUSSBAUM, M. (1996). Asymptotic equivalence of density estimation and Gaussian white noise. *Ann. Statist.* **24** 2399–2430.
ROBINSON, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4** 185–207.
SAKHANENKO, A. I. (1991). On the accuracy of normal approximation in the invariance principle. *Siberian Adv. Math.* **1** 58–91.
SHORACK, G. R. and WELLNER, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
SKOROKHOD, A. V. (1965). *Studies in the Theory of Random Processes*. Addison-Wesley, Reading, MA.
SPOKOINY, V. G. (1996). Adaptive and spatially adaptive testing of a nonparametric hypothesis. *Math. Methods Statist.* To appear.
STONE, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5** 595–620.
SUN, J. and LOADER, C. R. (1994). Simultaneous confidence bands for linear regression and smoothing. *Ann. Statist.* **22** 1328–1345.

TJØSTHEIM, D. (1994). Non-linear time series: a selective review. *Scand. J. Statist.* **21**, 97–130.
TSYBAKOV, A. B. (1986). Robust reconstruction of functions by the local approximation method. *Problems Inform. Transmission* **22** 133–146.
WU, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14** 1261–1343.

HUMBOLDT-UNIVERSITÄT
SONDERFORSCHUNGSBEREICH 373
SPANDAUER STRASSE 1
D-10178 BERLIN
GERMANY
E-MAIL: neumann@wiwi.hu-berlin. de

TECHNISCHE UNIVERSITÄT BRAUNSCHWEIG
POCKELSSTRASSE 4
D-38106 BRAUNSCHWEIG
GERMANY
E-MAIL: j.kreiss@tu-bs.de