Bootstrap Confidence Bands for the Autoregression Function

Article · February 1970

Source: CiteSeer

Michael H. Neumann
Friedrich Schiller University Jena

J.-P. Kreiss
Technische Universität Braunschweig

Bootstrap confidence bands for the autoregression function

Michael H. Neumann
Weierstraß-Institut für Angewandte
Analysis und Stochastik
Mohrenstraße 39
D – 10117 Berlin
Germany

Jens-Peter Kreiss
Technische Universität Braunschweig
Pockelsstraße 14
D – 38106 Braunschweig
Germany

1991 Mathematics Subject Classification. Primary 62G07, 62M05; secondary 62G09, 62G15.
Keywords and Phrases. Nonparametric autoregression, nonparametric regression, strong approximation, bootstrap, wild bootstrap, confidence bands.

The second author gratefully acknowledges the hospitality and support of the SFB 373 at Humboldt University, Berlin. We thank R. Tschernig for helpful comments on this paper.
The research was carried out within Sonderforschungsbereich 373 at Humboldt University Berlin and was printed using funds made available by the Deutsche Forschungsgemeinschaft.

ABSTRACT. 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. Besides our particular application to simultaneous confidence bands, this suggests the validity of wild bootstrap for several other statistical purposes.

1. INTRODUCTION

In this paper we deal with a nonparametric autoregressive model

$$X_t = m(X_{t-1}) + \varepsilon_t.$$

Such processes generalize well-known linear first order autoregressive models. Several authors dealt with the interesting statistical problem of estimating m nonparametrically. Robinson (1983), Tjøstheim (1994) and Masry and Tjøstheim (1995) dealt with usual Nadaraya-Watson type estimators. Recently (Härdle and Tsybakov (1995)) the interest was directed to local polynomial estimators for this setup. Of course, it is important to get knowledge about the statistical properties of particular nonparametric estimates. Besides asymptotic results the bootstrap offers a powerful tool for this purpose. Franke, Kreiss and Mammen (1996) consider a time series specific bootstrap as well as a wild bootstrap proposal in order to obtain pointwise confidence intervals for kernel smoothers in nonparametric autoregression with conditional heteroscedasticity. Successful application of the bootstrap for time series models can be found for example in Tjøstheim and Auestad (1994).

In this paper we consider the situation from a more general point of view. As a typical nonparametric estimator we consider local polynomials. We derive a strong approximation of a local polynomial estimator (LPE) in the autoregressive setup by an LPE in a corresponding nonparametric regression model. Besides the application of this main result to our particular example of simultaneous confidence bands, it contains the general message that nonparametric autoregression and nonparametric regression are asymptotically equivalent in a certain sense concerning statistical inference about the autoregression/regression function. Of course, this suggests and justifies to use regression-type methods for statistical inference in the context of nonparametric autoregression, too.

Further, from Neumann and Polzehl (1995) it is essentially known that one can find a strong approximation of an LPE in nonparametric regression by a random process generated by an appropriate bootstrap technique. Together with the strong approximation result in the present paper we are able to present a strong approximation of an LPE in nonparametric autoregression by a process generated according to the

wild bootstrap idea. Finally, we apply the strong approximation results to simultaneous confidence bands. On the basis of a result of Hall (1991), it can be shown that the proposed bootstrap approximation outperforms the approach using first-order asymptotic theory for the supremum of an appropriate Gaussian process.
But, quite general, the results suggest that the wild bootstrap is valid for several other purposes, too. In a forthcoming manuscript we discuss in detail bootstrap tests for the hypothesis of a parametric model for m. Such results have been developed in the regression model by Härdle and Mammen (1993).
The paper is organized as follows. In Section 2 we present the main ideas and results leading to a strong approximation of an LPE in nonparametric autoregression by an LPE in nonparametric regression (Theorem 2.1). Furthermore, we collect in this section the necessary assumptions and some auxiliary results. Section 3 contains the wild bootstrap proposal and the corresponding strong approximation result. The application of the results to simultaneous bootstrap confidence bands is given in Section 4. There we also 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 $m(x) = E(X_t | \; X_{t-1} = x)$. First, we write the data generating process in the form of a nonparametric autoregressive model,

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

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

$$\begin{aligned} E\left(\varepsilon_t \mid X_{t-1}\right) &= 0, \\ E\left(\varepsilon_t^2 \mid X_{t-1}\right) &= v(X_{t-1}). \end{aligned}$$

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 $\sigma(X_{t-1})\varepsilon_t$ with i.i.d. ε_t's, the errors here can follow completely different distributions and are not necessarily independent. Such a dependence arises because the distribution of ε_t depends 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 P_X the stationary distribution. 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 of Doukhan (1995, Chapter 1). Assumption (A1) is for example fulfilled if we assume the following explicit structure of the data-generating process:

$$X_t \;=\; m(X_{t-1}) \,+\, s(X_{t-1})\varepsilon_t', \tag{2.2}$$

where $s : \mathbb{R} \to (0,\infty)$ and (ε_t') denote i.i.d. innovations with zero mean and unit variance. We assume that

$$\limsup_{|x|\to\infty} \frac{E|m(x)+s(x)\varepsilon_1'|}{|x|} < 1$$

and that the distribution of ε_1' possesses a nowhere vanishing Lebesgue density. ¿From these conditions one may conclude that $\{X_t\}$ defined according to (2.2) is geometrically ergodic (cf. Doukhan (1995, p. 106/107)), which implies geometrical β-mixing if the chain is stationary, i.e. $X_0 \sim P_X$.
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.
Processes as defined in (2.2) play an important role in financial time series. Usually they are called ARCH-processes. Finally, we like to mention that we need assumption (A1), especially the geometric β-mixing, to give a not too complicated proof to Lemma 2.1. There we need more or less an exponential inequality.

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

(A2) The stationary density p_X of X_t satisfies $p_X(x) \geq C > 0$ for all $x \in [a,b]$

and construct a confidence band for this interval $[a,b]$. In this paper we focus our attention to so-called local polynomial estimators. These estimators are introduced in a paper by Stone (1977). Fan (1992, 1993) and Fan and Gijbels (1992, 1995) discuss the behavior of LPE for nonparametric regression in full detail. Recently Härdle and Tsybakov (1995) applied LPE to nonparametric autoregressive models.
A p-th order local polynomial estimator $\widehat{m}_h(x)$ of $m(x)$ is given as $\widehat{a}_0 = \widehat{a}_0(x, X_0, \ldots, X_T)$, where $\widehat{a} = (\widehat{a}_0, \ldots, \widehat{a}_{p-1})'$ minimizes

$$M_x = \sum_{t=1}^{T} K\left(\frac{x - X_{t-1}}{h}\right)\left(X_t - \sum_{q=0}^{p-1} a_q \left(\frac{x-X_{t-1}}{h}\right)^q\right)^2 . \qquad (2.3)$$

At the moment we only assume that the bandwidth h of the local polynomial estimator satisfies $h = O(T^{1-\delta})$ and $h^{-1} = O(T^{\delta})$ for some $\delta > 0$. We assume that the kernel K is a nonnegative function of bounded total variation with $supp(K) \subseteq [-1,1]$. 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 $\widehat{m}_h$ can be written as

$$\widehat{m}_h(x) = \sum_{t=1}^{T} w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) X_t = \left[(D_x' K_x D_x)^{-1} D_x' K_x \underline{X}\right]_1 , \qquad (2.4)$$

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

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

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

On first sight the analysis of $\widehat{m}_h$ seems to be quite involved, because the X_t's are dependent and enter into the right-hand side of (2.4) several times. To simplify the investigation of the deviation field $\{\widehat{m}_h(x) - m(x)\}_{x\in[a,b]}$ we approximate it by an analogous deviation field defined by observations according to a nonparametric regression model with independent errors.

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

$$Y_t = m(x_{t-1}) + \eta_t, \quad t = 1, \ldots, T, \tag{2.5}$$

where the η_t's are independent with $\eta_t \sim \mathcal{L}(\varepsilon_t \mid X_{t-1} = x_{t-1})$. 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 $\{X_0, \ldots, X_{T-1}\}$. In analogy to (2.4) we define a local polynomial estimator as

$$\widetilde{m}_h(x) = \sum_{t=1}^{T} w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})Y_t. \tag{2.6}$$

In this section we show that on a sufficiently rich probability space there exists a pairing of $(X_0, \varepsilon_1, \ldots, \varepsilon_T)$, having a joint distribution according to model (2.1), with $(\eta_1, \ldots, \eta_T)$, having a joint distribution according to (2.5), such that $\widehat{m}_h$ and $\widetilde{m}_h$ are close to each other in the supremum norm on $[a, b]$. Before we turn to the main approximation step, we derive first some approximations to $\widehat{m}_h$ and $\widetilde{m}_h$, which allow to replace the local polynomial estimators by quantities of a simpler structure.

2.1. Simplification of the problem by approximating the local polynomial estimators. 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

$$P(|Z_T| > C\alpha_T) \le C\beta_T$$

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; see the application in Section 4 where we obtain in conjunction with Lemma 4.1 upper estimates for the error in coverage probability of the confidence bands.
In the following we have to deal with random functions of X_{t-1}, which also depend on the whole set $\{X_0, \ldots, X_{T-1}\}$. For example, the weights $w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\})$ of the local polynomial estimator are of this structure. To get nonrandom approximations of them we show that the number of X_t's that fall into some fixed interval converges to the expected number at a certain rate; cf. Lemma 2.1. Then we expand the functions of interest into a Haar wavelet series and show that this series converges in the supremum norm to a nonstochastic limit.
Here and in the following λ denotes an arbitrarily large constant.

Lemma 2.1. *Assume (A1). Then*

$$\left|\sum_{t=1}^{T} \{I(X_{t-1} \in [c_1, c_2]) - P_X[c_1, c_2]\}\right| = \tilde{O}\left(\min\{\sqrt{TP_X[c_1, c_2]T^{2\delta}} + (\log T)^2, \sqrt{T \log T}\}, T^{-\lambda}\right)$$

holds uniformly in $-\infty \le c_1 < c_2 \le \infty$.

In the following we specify this and other approximations to intervals of the form

$$I_{j,k} = \left[k2^{-j}, (k+1)2^{-j}\right). \tag{2.7}$$

We define $\mathcal{I}_T = \{(j,k) \mid \ 0 \le j \le j^*, (a-\delta)2^j < k \le (b+\delta)2^j\}$, where $2^{j^*} = O(T)$. Here large values of j refer to small intervals, whereas $I_{0,k} = [k, k+1)$.
As an immediate consequence of Lemma 2.1 we obtain that

$$\begin{aligned} &\#\{t: \ X_{t-1} \in I_{j,k}\} - TP(X_{t-1} \in I_{j,k}) \\ &\quad = \tilde{O}\left(\min\{\sqrt{TP(X_{t-1} \in I_{j,k})T^{2\delta}} + (\log T)^2, \sqrt{T \log T}\}, T^{-\lambda}\right) \end{aligned} \tag{2.8}$$

holds uniformly in $(j,k) \in \mathcal{I}_T$.
According to (2.4), the weights of the local polynomial estimator can be written as

$$w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) = \sum_{q=0}^{p-1} d_q(x, \{X_0, \ldots, X_{T-1}\}) K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q, \tag{2.9}$$

where $d_q(x, \{X_0, \ldots, X_{T-1}\}) = ((D_x' K_x D_x)^{-1})_{1,q+1}$. 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 (A2). Then there exist nonrandom functions* $d_q^{(\infty)}(x)$, $d_q^{(\infty)}(x) = \left((ED_x' K_x D_x)^{-1}\right)_{1,q+1} = O((Th)^{-1})$ *, such that*

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

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

$$\overline{w}_h(x, X_{t-1}) \;=\; \sum_{q=0}^{p-1} d_q^{(\infty)}(x) K\left(\frac{x - X_{t-1}}{h}\right)\left(\frac{x - X_{t-1}}{h}\right)^q. \tag{2.10}$$

Now we obtain the following assertions, which finally allow 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 between the more involved quantities $\widehat{m}_h(x)$ and $\widetilde{m}_h(x)$. To ensure the desired behaviour of weighted sums of the ε_t's and η_t's, respectively, we impose the following condition.

(A3) For all $M < \infty$ and arbitrary $\delta > 0$ there exist finite constants C_M such that $\sup_{x\in[a-\delta,b+\delta]}\left\{E\left(|\varepsilon_t|^M \mid \quad X_{t-1} = x\right)\right\} \le C_M$

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.

Proposition 2.1. *Assume (A1) to (A3). Then*

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

Analogously,

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

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

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

Proposition 2.2. *Assume (A1), (A2) and (A4).*
As an approximation to the bias-type term we consider the nonrandom quantity

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

Then

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

and

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

2.2. Approximation of autoregression by regression via Skorokhod embedding. In the previous subsection we derived some helpful technical approximations to reduce the problem of finding a close connection between the processes $\{\widehat{m}_h(x)\}_{x\in[a,b]}$ and $\{\widetilde{m}_h(x)\}_{x\in[a,b]}$ to the simpler task of finding a link between $\{\sum_t \overline{w}_h(x, X_{t-1})\varepsilon_t\}_{x\in[a,b]}$ and $\{\sum_t \overline{w}_h(x, x_{t-1})\eta_t\}_{x\in[a,b]}$. Now we construct such a pairing of the observations in (2.1) and (2.5), which provides a good approximation of partial sums of the ε_t's by partial sums of the η_t's corresponding to certain subintervals of $[a-\delta, b+\delta]$. Using a Haar wavelet expansion we then obtain the desired connection between $\{\sum_t \overline{w}_h(x, X_{t-1})\varepsilon_t\}_{x\in[a,b]}$ and $\{\sum_t \overline{w}_h(x, x_{t-1})\eta_t\}_{x\in[a,b]}$.
The link between the two sampling schemes (2.1) and (2.5) will be reached by Skorokhod embeddings of the ε_t's and η_t's, respectively, in the same 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). As in the previous subsection, we consider partial sums of the ε_t's and η_t's, respectively, according to subintervals $I_{j,k}$, $(j,k) \in \mathcal{I}_T$. 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 (2.5), respectively. Using Skorokhod embedding techniques we can establish the following fundamental lemma. Here and in the following $\delta > 0$ denotes an arbitrarily small, but fixed constant.

Lemma 2.3. *Assume (A1) to (A3). There exist sets of events Ω_T, $P((X_0, \dots, X_{T-1}) \notin \Omega_T) = O(T^{-\lambda})$, such that there exists on an appropriate probability space a pairing of the random variables from (2.1) and (2.5) with*

$$P\left(|Z_{j,k} - Z'_{j,k}| > [TP(X_{t-1} \in I_{j,k})]^{1/4} T^\delta + T^\delta \quad \textit{for any } (j,k) \in \mathcal{I}_T \right) = O(T^{-\lambda})$$

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

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 $\mathcal{W} = \{w \mid supp(w) \subseteq [c,d]\}$ of such weighting functions, where $c < d$ are any fixed constants.

Corollary 2.1. *Assume (A1) to (A3) and let Ω_T be as in Lemma 2.3. Then there exists a pairing of the random variables from (2.1) and (2.5) such that*

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

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

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 $\overline{w}_h(x,.)$. This leads to the following assertion.

Proposition 2.3. *Assume (A1) to (A3) and let Ω_T be as in Lemma 2.3. Then there exists a pairing of the random variables from (2.1) and (2.5) such that*

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

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 now to the desired approximation of nonparametric autoregression by nonparametric regression.

Theorem 2.1. *Assume (A1) to (A4) and let Ω_T be as in Lemma 2.3. Then there exists a pairing of the random variables from (2.1) and (2.5) such that*

$$\sup_{x\in[a,b]}\{|\widehat{m}_h(x) \;-\; \widetilde{m}_h(x)|\} \;=\; \tilde{O}\left(h^p(Th)^{-1/2}T^\delta \;+\; (Th)^{-3/4}T^\delta, T^{-\lambda}\right)$$

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 $\widehat{m}_h(x)$ by $\widetilde{m}_h(x)$, the more important fact is that the difference between $\widehat{m}_h(x)$ and $\widetilde{m}_h(x)$ is of smaller order than the stochastic fluctuations of $\widehat{m}_h(x)$, which are $O_P((Th)^{-1/2})$. Although we use this result here only for the particular purpose of constructing simultaneous confidence bands for the autoregression function m, it seems to be of much greater importance. It provides the fundamental message that nonparametric autoregression and nonparametric regression are asymptotically equivalent. This explains in particular why methods, which were first developed in the regression context can be literally applied to autoregression. The approximation given in Theorem 2.1 can also be used to transfer a testing method developed in Härdle and Mammen (1993) for the case of nonparametric autoregression. This will be done in a forthcoming paper.

Remark 2. 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 (2.4) we consider the nonparametric regression model

$$Z_t = m(Y_t) + \eta_t, \quad t = 1, \ldots, T, \tag{2.11}$$

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 $\overline{m}_h(x)$ be the local polynomial estimator in model (2.11), which is defined analogously to (2.6). It is easily seen that the statement in Theorem 2.1 implies the asymptotic equivalence of the experiments (2.1) and (2.11). Strictly speaking, under (A1) to (A4) there exists a pairing of the random variables from (2.1) with those of (2.11) such that

$$\sup_{x \in [a,b]} \left\{ |\widehat{m}_h(x) - \overline{m}_h(x)| \right\} = \tilde{O}\left(h^p (Th)^{-1/2} T^{\delta} + (Th)^{-3/4} T^{\delta}, T^{-\lambda} \right).$$

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 explicitly estimated, if some degrees of smoothness of m are not used by $\widehat{m}_h(x)$. In nonparametric regression and density estimation there exist two main approaches to handle the bias problem: undersmoothing and explicit bias correction.

Here we take an undersmoothed estimator $\widehat{m}_h(x)$, that is the bandwidth h is chosen such that the order of the bias is smaller than the order of the standard deviation. Then it is not necessary to model $m(x)$ in the bootstrap world, because the deviation process $\widehat{m}_h(x) - m(x)$ is dominated by the stochastic term $\sum_t w_h(x, X_{t-1}, \{X_0, \ldots, X_{T-1}\}) \varepsilon_t$. 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

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

An appropriate counterpart to model (2.5) in the bootstrap world is given by

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

As argued above, 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. From this it is clear that we do not use the X_t^*'s explicitly.

We have to ensure that for all integers M there exists a finite constant $C_M > 0$ such that

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

This can be ensured if we assume that $\varepsilon_t^* = \widehat{\varepsilon}_t \eta_t^*$ for a sequence of i.i.d. random variables $\eta_1^*, \dots, \eta_T^*$ with $E^*\eta_1^* = 0$, $E^*(\eta_1^*)^2 = 1$, and $E^*|\eta_1^*|^M < \infty$, for all integers M. Exactly along the lines of Section 2 we obtain the following results.

Lemma 3.1. *On a sufficiently rich probability space there exists a pairing of* $\eta_1, \dots, \eta_T$ *with* $\varepsilon_1^*, \dots, \varepsilon_T^*$ *such that*

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

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

In conjunction with Theorem 2.1 we get

Theorem 3.1. *On a sufficiently rich probability space there exists a pairing of* $X_0, \varepsilon_1, \dots, \varepsilon_T$ *with* $\varepsilon_1^*, \dots, \varepsilon_T^*$ *such that*

$$\sup_{x\in[a,b]} \left\{ \left| \widehat{m}_h(x) - m(x) - \sum_t w_h(x, x_{t-1}, \{x_0, \dots, x_{T-1}\})\varepsilon_t^* \right| \right\} = \tilde{O}\left(h^p + (Th)^{-3/4}T^{\delta}, T^{-\lambda}\right)$$

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

4. An application of the bootstrap: confidence bands

We consider two possibilities for asymptotic confidence bands for m to a prescribed level $1-\alpha$. We develop simultaneous bands as opposed to confidence bands which attain pointwise a certain coverage probability. First we can construct a confidence band of a uniform size. To get the appropriate width for such a band, we consider the quantity

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

which is introduced to mimic

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

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

$$I_\alpha^*(x) = [\widehat{m}_h(x) - t_\alpha^*, \widehat{m}_h(x) + t_\alpha^*] \tag{4.1}$$

is supposed to form an asymptotic confidence band of 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 $\widehat{m}_h(x)$. Whereas the size of I_α^* is essentially driven by the worst case, that is by the supremum of $v(x) = var(\widehat{m}_h(x))$, a variable confidence band follows in size the local variability of $\widehat{m}_h(x)$. 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 $\widehat{\varepsilon}_t$ can also be used to estimate $v(x)$ by

$$\widehat{v}(x) = \sum_t w_h^2(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\widehat{\varepsilon}_t^2. \tag{4.2}$$

Let t_α^{**} be the (1-α)-quantile of the distribution of

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

which mimics

$$V_T = \sup_{x\in[a,b]} \left\{ |\widehat{m}_h(x) - m(x)| / \sqrt{\widehat{v}(x)} \right\}.$$

This leads to a confidence band of the form

$$I_\alpha^{**}(x) = \left[\widehat{m}_h(x) - \sqrt{\widehat{v}(x)} t_\alpha^{**}, \widehat{m}_h(x) + \sqrt{\widehat{v}(x)} t_\alpha^{**} \right]. \tag{4.3}$$

We already know from Theorem 3.1 that the process $\widehat{m}_h(x) - m(x)$ is pathwise close to the conditional (conditioned on $X_0, \varepsilon_1, \ldots, \varepsilon_T$) process $\sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\varepsilon_t^*$ on an appropriate probability space. The following lemma provides a lower bound for probabilities that $\sup_{x\in[a,b]}\{|\sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\varepsilon_t^*|\}$ falls into small intervals. Finally, these two results will lead to an estimate of the error in coverage probability of the proposed confidence bands.

Lemma 4.1. *Assume (A1) to (A3). Then*

$$\begin{aligned} &P\left(\sup_{x\in[a,b]} \left\{ |\sum_t w_h(x, x_{t-1}, \{x_0, \ldots, x_{T-1}\})\varepsilon_t^*| \right\} \in [c_1, c_2] \right) \\ &\quad = O\left((c_2 - c_1)(Th)^{1/2}(\log T)^{1/2} + (Th)^{-1/2}T^\delta \right). \end{aligned}$$

This lemma follows immediately from Lemma 2.2 in Neumann and Polzehl (1995). In conjunction with Theorem 3.1, we now obtain an upper bound of the error in coverage probability for I_α^*.

Theorem 4.1. *Assume (A1) to (A4). Then*

$$\begin{aligned} &P\left(m(x) \in [\widehat{m}_h(x) - t_\alpha^*, \widehat{m}_h(x) + t_\alpha^*] \quad \textit{for all} \quad x \in [a,b]\right) \\ &\qquad = 1 - \alpha + O\left((h^p + (Th)^{-3/4}T^\delta)(Th)^{1/2}\sqrt{\log T} \right). \end{aligned}$$

Analogously, we are able to give an upper bound for the error in coverage probability for I_α^{**} .

Theorem 4.2. *Assume (A1) to (A4). Then*

$$P\left(m(x) \in [\widehat{m}_h(x) - \sqrt{\widehat{v}(x)}t_\alpha^{**}, \widehat{m}_h(x) + \sqrt{\widehat{v}(x)}t_\alpha^{**}] \quad \textit{for all} \quad x \in [a,b]\right)$$
$$= 1 - \alpha + O\left((h^p + (Th)^{-3/4}T^\delta)(Th)^{1/2}\sqrt{\log T}\right).$$

As already mentioned, we propose to use undersmoothing to handle the bias problem. This means that the bandwidth $h = h(T)$ has to be chosen in such a way that $h^p \ll (Th)^{-1/2}$. If we do this appropriately, the error terms in Theorems 4.1 and 4.2 vanish, that is the confidence bands have asymptotically the prescribed coverage probability $1 - \alpha$.

We conclude this section with some simulation results. For this purpose let us consider the following two models:

$$X_t = 4 \cdot \sin(X_{t-1}) + \varepsilon_t \tag{4.4}$$

and

$$X_t = 0.8 \cdot X_{t-1} + \sqrt{1 + 0.2\ X_{t-1}^2} \cdot \varepsilon_t \tag{4.5}$$

The latter model is an 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 model (4.4) we assume a double exponential distribution, while model (4.5) 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 for $m_1(x) = 4 \cdot \sin(x)$ and $m_2(x) = 0.8 \cdot x$. This is done by simulating the 90% -quantile of V_T from 1000 Monte Carlo replications. The results are reported in Figures 1 and 2. m_1 is estimated by a local linear estimator $\widehat{m}_h$, $h = 0.4$, while for m_2 we make use of an usual Nadaraya-Watson type kernel estimator, i.e. a local constant smoother, with bandwidth $h = 1.0$. The thick lines show m_1 and m_2, respectively, whereas the thin lines represent confidence bands of the form $\widehat{m}_h(x) \pm \sqrt{\widehat{v}(x)}t_{\alpha,i}$, where $t_{\alpha,i}$ is chosen such that $m_i(x)$ is covered in 900 cases by the above band.

[Please insert Figures 1 and 2 about here.]

Finally, we choose at random three time series realizations from each model (4.4) and (4.5) in order to carry through the bootstrap. All three resulting bootstrap confidence bands $I_{0.10}^{**}(x)$, cf. (4.3), are given in Figures 3a-3c and 4a-4c, respectively. In order to obtain an impression of the stochastic fluctuation of these simultaneous confidence bands, we additionally report a plot (cf. Figures 5 and 6) which contains all bootstrap confidence bands together with the simulated true confidence bands from Figures 1 and 2 (thick lines).

[Please insert Figures 3, 4, 5, and 6 about here.]

Although this is only a small simulation study, the results demonstrate that the bootstrap offers a powerful tool in order to construct not only pointwise but also simultaneous confidence bands for nonparametric estimators in nonlinear autoregression.

The authors are very grateful to D. Seidel, Technical University of Braunschweig, for doing the programming.

5. Proofs

Proof of Lemma 2.1. The assertion can be concluded from a kind of Bernstein inequality, which, for example, is given in Doukhan (1995, Theorem 4, p. 36). Define

$$Z_t \;=\; I(X_{t-1} \in [c_1, c_2]) \;-\; P_X([c_1, c_2])$$

and abbreviate $EZ_t^2 \;=\; P_X[c_1, c_2](1 - P_X[c_1, c_2]) \;=\; \sigma^2$. Since geometric β-mixing implies geometric strong mixing (i.e. α-mixing) we obtain from a covariance inequality (cf. Doukhan (1995), Theorem 3, p. 9) for all $\delta > 0$ that

$$E\left(\sum_{t=1}^{T} Z_t\right)^2 \;\leq\; 2T\sigma^{2(1-\delta)}.$$

¿From the above mentioned Theorem 4 of Doukhan (1995) we obtain for $\kappa, M > 0$ large enough and all $\varepsilon > 0$, uniformly in $-\infty < c_1 < c_2 < \infty$, that

$$\begin{aligned}
&P\left(\left|\sum_{t=1}^{T} Z_t\right| \;\geq\; M \min\left\{\sqrt{TP_X[c_1,c_2]T^{2\delta}} + (\log T)^2, \sqrt{T\log T}\right\}\right) \\
&\leq\; P\left(\left|\sum_{t=1}^{T} Z_t\right| \;\geq\; M\min\left\{\sqrt{T\sigma^2 T^{2\delta}} + (\log T)^2, \sqrt{T\log T}\right\}\right) \\
&\leq\; 4\exp\left\{-M^2 \frac{(1-\varepsilon)\min^2\{\sqrt{T\sigma^2T^{2\delta}} + (\log T)^2, \sqrt{T\log T}\}}{2\left(2T\sigma^{2(1-\delta)} + \kappa \log T \min\{\sqrt{T\sigma^2 T^{2\delta}} + (\log T)^2, \sqrt{T\log T}\}\right)}\right\} \;+\; O(T^{-\lambda}).
\end{aligned}$$

In the case that the minimum in the denominator represents the dominating term it is easy to see that the exponent is at least of magnitude *const* $\cdot \log T$, i.e. the whole expression is of order $O(T^{-\lambda})$.

In the case that

$$2T\sigma^{2(1-\delta)} \;\geq\; \kappa\log T\min\{\sqrt{T\sigma^2T^{2\delta}} + (\log T)^2, \sqrt{T\log T}\} \;=\; \kappa\log T\sqrt{T\log T}$$

we obtain from $\sigma^{2(1-\delta)} \leq 1$ an upper bound to the exponential term by

$$\exp\left(-\frac{M^2(1-\varepsilon)}{8}\log T\right).$$

Finally, for the remaining case

$$2T\sigma^{2(1-\delta)} \;\geq\; \kappa\log T\min\{\sqrt{T\sigma^2T^{2\delta}}+(\log T)^2, \sqrt{T\log T}\} \;=\; \kappa\log T\left(\sqrt{T\sigma^2T^{2\delta}} + (\log T)^2\right)$$

we have, since (from the above inequality) $T\sigma^2 \geq 1$, the following upper bound to the exponential term:

$$\exp\left(-\frac{M^2(1-\varepsilon)}{8}\frac{(\sqrt{T\sigma^2T^{2\delta}} + (\log T)^2)^2}{T\sigma^{2(1-\delta)}}\right) \;\leq\; \exp\left(-\frac{M^2(1-\varepsilon)}{8}T^\delta\right) \;=\; O(T^{-\lambda}).$$

□

Proof of Lemma 2.2. First we investigate how good the random quantity $(D_x' K_x D_x)_{ij}$ is approximated by its expectation. Let $g(z) = K\left(\frac{x-z}{h}\right)\left(\frac{x-z}{h}\right)^{i+j-2}$. Note that, for T large enough and $x \in [a,b]$, g is supported on $[a-h, b+h]$. Hence, we can apply the estimate given by (2.8). We approximate g by a truncated Haar wavelet series expansion

$$\tilde{g}(z) \;=\; \sum_k \alpha_k \phi_k(z) \;+\; \sum_{0 \le j < j^*} \sum_k \alpha_{j,k} \psi_{j,k}(z), \tag{5.1}$$

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

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

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

$$\sum_k |\alpha_k| \;\le\; \|g\|_{L_1} \;=\; O(h) \tag{5.2}$$

and

$$\sum_k |\alpha_{j,k}| \;=\; O\left(\min\{\|\psi_{j,k}\|_\infty \|g\|_1, \|\psi_{j,k}\|_1 TV(g)\}\right) \;=\; O(\min\{2^{j/2}h, 2^{-j/2}\}). \tag{5.3}$$

Define $F_T(z) = \sum_{t=1}^T I(X_{t-1} < z)$ and $F_T^{(\infty)} = \sum_{t=1}^T P(X_{t-1} < z)$. Then, by (5.2), (5.3) and (2.8),

$$\begin{aligned} & \left| \sum_{t=1}^T \tilde{g}(X_{t-1}) \;-\; \sum_{t=1}^T E\tilde{g}(X_{t-1}) \right| \\ =\;& \left| \int \tilde{g}(z)\, dF_T(z) \;-\; \int \tilde{g}(z)\, dF_T^{(\infty)}(z) \right| \\ \le\;& \sum_k |\alpha_k| \left| (F_T(k+1) - F_T(k)) \;-\; \left(F_T^{(\infty)}(k+1) - F_T^{(\infty)}(k)\right) \right| \\ & + \sum_{0 \le j < j^*} \sum_k |\alpha_{j,k}| \left| \int \psi_{j,k}(z) \left[dF_T(z) - dF_T^{(\infty)}(z) \right] \right| \\ =\;& \tilde{O}\left(h\sqrt{T \log T}, T^{-\lambda} \right) \\ & + \sum_{j:\, 2^j \le h^{-1}} O(2^{j/2}h) O(2^{j/2}) \tilde{O}(\sqrt{T2^{-j}}T^\delta, T^{-\lambda}) \\ & + \sum_{j:\, j<j^*,\, 2^j > h^{-1}} O(2^{-j/2}) O(2^{j/2}) \tilde{O}(\sqrt{T2^{-j}}T^\delta, T^{-\lambda}) \\ =\;& \tilde{O}\left(\sqrt{Th} T^\delta, T^{-\lambda} \right). \end{aligned} \tag{5.4}$$

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

$$\sum_k \|g - \widetilde{g}\|_{L_\infty(I_{j^*,k})} \leq TV(g) = O(1).$$

This implies that

$$\begin{aligned}
&\sum_t g(X_{t-1}) - \widetilde{g}(X_{t-1}) \\
&\quad = \sum_k \sum_{t:\, X_{t-1} \in I_{j^*,k}} g(X_{t-1}) - \widetilde{g}(X_{t-1}) \\
&\quad \leq \sum_k \|g - \widetilde{g}\|_{L_\infty(I_{j^*,k})} \# \{t : \; X_{t-1} \in I_{j^*,k}\} \\
&\quad = \widetilde{O}\left(T2^{-j^*} + \sqrt{T2^{-j^*}}T^\delta, T^{-\lambda}\right), \qquad (5.5)
\end{aligned}$$

and, analogously,

$$\sum_t Eg(X_{t-1}) - E\widetilde{g}(X_{t-1}) = O(T2^{-j^*}). \qquad (5.6)$$

¿From (5.4) to (5.6) we obtain that

$$|(D_x' K_x D_x)_{ij} - E(D_x' K_x D_x)_{ij}| = \widetilde{O}(\sqrt{Th}T^\delta, T^{-\lambda}),$$

which implies

$$\|D_x' K_x D_x - ED_x' K_x D_x\| = \widetilde{O}(\sqrt{Th}T^\delta, T^{-\lambda}). \qquad (5.7)$$

Recall that p_X denotes the stationary density of $\{X_t\}$. Because of

$$E(D_x' K_x D_x)_{ij} = T \int K\left(\frac{x-z}{h}\right)\left(\frac{x-z}{h}\right)^{i+j-2} p_X(z)\,dz = Th \int_{-1}^{1} K(z) z^{i+j-2} p_X(x-hz)\,dz$$

we obtain that

$$ED_x' K_x D_x \geq C\,Th \left(\left(\int_{-1}^{1} K(z) z^{i+j-2}\,dz\right)\right)_{i,j=1,\dots,p}, \qquad (5.8)$$

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

$$\begin{aligned}
&\left\|(D_x' K_x D_x)^{-1} - (ED_x' K_x D_x)^{-1}\right\| \\
&\quad \leq \left\|(D_x' K_x D_x)^{-1}\right\| \|D_x' K_x D_x - ED_x' K_x D_x\| \left\|(ED_x' K_x D_x)^{-1}\right\| \\
&\quad = \widetilde{O}\left((Th)^{-3/2}T^\delta, T^{-\lambda}\right). \qquad (5.9)
\end{aligned}$$

With the definition

$$d_q^{(\infty)}(x) = \left((ED_x' K_x D_x)^{-1}\right)_{1,q+1}$$

we obtain the assertion. □

Proof of Proposition 2.1. By $\|K((x-.)/h)((x-.)/h)^q\|_1 = O(h)$ and $TV\left(K((x-.)/h)((x-.)/h)^q\right) = O(1)$ we conclude from Corollary 2.1 that

$$\sum_t K\left(\frac{x-X_{t-1}}{h}\right)\left(\frac{x-X_{t-1}}{h}\right)^q \varepsilon_t$$
$$= \sum_t K\left(\frac{x-x_{t-1}}{h}\right)\left(\frac{x-x_{t-1}}{h}\right)^q \eta_t + \tilde{O}\left((Th)^{1/4}T^{\delta}, T^{-\lambda}\right) \qquad (5.10)$$

holds for $(x_0,\ldots,x_{T-1}) \in \Omega_T$, Ω_T according to Lemma 2.3.
For $(x_0,\ldots,x_{T-1}) \in \Omega_T$ we obtain by Theorem 4 in Amosova (1972) that

$$Z'_{j,k} = \sum_{t: x_{t-1}\in I_{j,k}} \eta_t = \tilde{O}\left(\sqrt{T2^{-j}}\sqrt{\log T}, T^{-\lambda}\right),$$

which implies, by calculations similar to those in the proof of Corollary 2.1 below, that

$$\sup_{x\in[a,b]}\left\{\left|\sum_t K\left(\frac{x-x_{t-1}}{h}\right)\left(\frac{x-x_{t-1}}{h}\right)^q \eta_t\right|\right\} = \tilde{O}\left(\sqrt{Th}\sqrt{\log T}, T^{-\lambda}\right). \quad (5.11)$$

Using now Lemma 2.2, (2.9), and (2.10) we obtain the assertions. □

Proof of Proposition 2.2. Because of $\sum_t w_h(x, X_{t-1}, \{X_0,\ldots,X_{T-1}\}) = 1$ and $\sum_t w_h(x, X_{t-1}, \{X_0,\ldots,X_{T-1}\})(X_{t-1}-x)^q = 0$ for $q = 1,\ldots,p-1$ we get from a Taylor series expansion with integral remainder that

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

Since $g(z) = K\left(\frac{x-z}{h}\right)\left(\frac{x-z}{h}\right)^q \int_x^z \frac{(z-s)^{p-1}}{(p-1)!} m^{(p)}(s)\, ds$ satisfies $\|g\|_1 = O(h^{p+1})$ and $TV(g) = O(h^p)$, we obtain analogously to (5.7) that

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

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

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

□

Proof of Lemma 2.3. (i) *General idea*
The pairing of the observations in the autoregression model (2.1) with those in the regression model (2.5), 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 certain 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 e.g. 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 $\widehat{m}_h(x)-\widetilde{m}_h(x)=o_P((Th)^{-1/2})$. 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, truely *multiscale* approximation scheme. Accordingly we will relate the ε_t's and η_t's to Wiener processes $W_{j,k}$ for $(j,k)\in\mathcal{I}_T$.
In the following we describe this construction in detail for the autoregressive process (2.1). The construction in the regression setting (2.5) 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 defined 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 by means 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 to represent ε_1. In this case we additionally use 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

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

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:

$$W^{(1)}(s) = \begin{cases} W_{j^*,k_1}(s), & \text{if } 0 \le s \le T_{j^*,k_1} \\ W_{j^*,k_1}(T_{j^*,k_1}) + \ldots + W_{j+1,k_1^{(j+1)}}(T_{j+1,k_1^{(j+1)}}) + W_{j,k_1^{(j)}}(s - T_{j^*,k_1} - \ldots - T_{j+1,k_1^{(j+1)}}), \\ \qquad \text{if } T_{j^*,k_1} + \ldots + T_{j+1,k_1^{(j+1)}} < s \le T_{j^*,k_1} + \ldots + T_{j,k_1^{(j)}} \end{cases}$$

($W^{(1)}$ 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

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

for an appropriate stopping time $\tau^{(1)}$.
To explain the following steps in a formally correct way we introduce stopping times $\tau_{j,k}^{(t)}$, $t = 0, \ldots, T$, assigned to the corresponding Wiener processes $W_{j,k}$. Define

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

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

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

We redefine further

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

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 so far.

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

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

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

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

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

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

For all (j,l) with $l \ne k_t^{(j)}$ 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 $W_{j,k}$, $(j,k) \in \mathcal{I}_T$. 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 $I_{j,m} \supset I_{j^*,k}$. Then we will 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

$$P\left(\sum_t \tau^{(t)} I(X_{t-1} \in I_{j,k}) < \sum_{(l,m): I_{l,m} \subseteq I_{j,k}} T_{l,m} \quad \text{for any } (j,k) \in \mathcal{I}_T \setminus \{(0,k)\}\right) = O(T^{-\lambda}) \tag{5.12}$$

and

$$P\left(\sum_t \tilde{\tau}^{(t)} I(x_{t-1} \in I_{j,k}) < \sum_{(l,m): I_{l,m} \subseteq I_{j,k}} T_{l,m} \quad \text{for any } (j,k) \in \mathcal{I}_T \setminus \{(0,k)\}\right) = O(T^{-\lambda}). \tag{5.13}$$

To achieve this, we study first the behaviour of the above sums of the stopping times assigned to the interval $I_{j,k}$.
Define the σ-field $\mathcal{F}_t = \sigma\left(X_0, \varepsilon_1, \ldots, \varepsilon_t, \{W_{j,k}(s),\ 0 \leq s \leq \tau_{j,k}^{(t)}\}_{(j,k)\in\mathcal{I}_T}\right)$. According to Theorem A.1 in Hall and Heyde (1980, Appendix A.1), the stopping time $\tau^{(t)}$ is $\mathcal{F}_t$-measurable with

$$E\left(\tau^{(t)} \mid \mathcal{F}_{t-1}\right) = E\left(\varepsilon_t^2 \mid \mathcal{F}_{t-1}\right) = v(X_{t-1}) \quad \text{a.s.}$$

and

$$E\left((\tau^{(t)})^M \mid \mathcal{F}_{t-1}\right) \leq C_M E\left(\varepsilon_t^{2M} \mid \mathcal{F}_{t-1}\right) = C_M E\left(\varepsilon_t^{2M} \mid X_{t-1}\right) \quad \text{a.s.}$$

Further, $\left\{\sum_{s=1}^t [\tau^{(s)} - v(X_{s-1})] I(X_{s-1} \in I_{j,k}), \mathcal{F}_t, t \geq 1\right\}$ is a martingale.
Let $\varepsilon > 0$ be chosen such that $\delta > \varepsilon/(4 + 2\varepsilon)$. Further, define $g(X_{t-1}) = v(X_{t-1}) I(X_{t-1} \in I_{j,k}) - E v(X_0) I(X_0 \in I_{j,k})$ and $p_{j,k} = P(X_0 \in I_{j,k})$. Since $\{X_t\}$ is geometrically β-mixing, we obtain by Rosenthal's inequality (see Doukhan (1995), Theorem 2, p. 26)) that

$$\begin{aligned}
& E\left|\sum_{t=1}^T g(X_t)\right|^M \\
\leq\ & C(M,\varepsilon) \max\left\{\sum_{t=1}^T \left(E|g(X_t)|^{M+\varepsilon}\right)^{M/(M+\varepsilon)}, \left[\sum_{t=1}^T \left(E|g(X_t)|^{2+\varepsilon}\right)^{2/(2+\varepsilon)}\right]^{M/2}\right\} \\
=\ & O\left(T p_{j,k}^{M/(M+\varepsilon)} + [T p_{j,k}^{2/(2+\varepsilon)}]^{M/2}\right) \\
=\ & O\left(T(p_{j,k} + 1/T) T^{\varepsilon/(M+\varepsilon)} + [T(p_{j,k} + 1/T) T^{\varepsilon/(2+\varepsilon)}]^{M/2}\right). \qquad (5.14)
\end{aligned}$$

Define $f(X_{t-1}) = E\left([\tau^{(t)} - v(X_{t-1})]^2 I(X_{t-1} \in I_{j,k}) \mid \ \mathcal{F}_{t-1}\right)$. Then we obtain that

$$E\left|\sum_{t=1}^{T} f(X_{t-1})\right|^{M/2} = O\left([T(p_{j,k} + 1/T)]^{M/2}\right),$$

which implies by Rosenthal's inequality for martingales (cf. Hall and Heyde (1980, Theorem 2.12, p. 23/24)) that

$$\begin{aligned}
& E\left|\sum_{t=1}^{T}[\tau^{(t)} - v(X_{t-1})] I(X_{t-1} \in I_{j,k})\right|^{M} \\
&= O\left(E\left(\sum_{t=1}^{T} f(X_{t-1})\right)^{M/2} + \sum_{t=1}^{T} E|\tau^{(t)} - v(X_{t-1})|^M I(X_{t-1} \in I_{j,k})\right) \\
&= O\left([T(p_{j,k} + 1/T)]^{M/2}\right). \qquad (5.15)
\end{aligned}$$

If we choose $M \geq \lambda/(\delta - \varepsilon/(4 + 2\varepsilon))$, we obtain from (5.14) and (5.15) by Markov's inequality that

$$\begin{aligned}
& P\left(\left|\sum_{t=1}^{T} \tau^{(t)} I(X_{t-1} \in I_{j,k}) - T E v(X_0) I(X_0 \in I_{j,k})\right| > \sqrt{T P(X_0 \in I_{j,k})} T^{\delta} + T^{\delta}\right) \\
&= O\left(\frac{T(p_{j,k} + 1/T) T^{\varepsilon/(M+\varepsilon)} + [T(p_{j,k} + 1/T) T^{\varepsilon/(2+\varepsilon)}]^{M/2} + [T(p_{j,k} + 1/T)]^{M/2}}{[T^{1+2\delta}(p_{j,k} + 1/T)]^{M/2}}\right) \\
&= O(T^{-\lambda}). \qquad (5.16)
\end{aligned}$$

Accordingly, we have

$$P\left(\begin{array}{l} \left|\sum_{t:\ X_{t-1} \in I_{j,k}} \tau^{(t)} - T E v(X_0) I(X_0 \in I_{j,k})\right| \ > \ [\sqrt{T P(X_0 \in I_{j,k})} T^{\delta} + T^{\delta}] \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for any } (j,k) \in \mathcal{I}_T \setminus \{(0,k)\} \end{array} \right) = O(T^{-\lambda}). \qquad (5.17)$$

For the regression scheme (2.5) we have an analogous relation:

$$P\left(\begin{array}{l} \left|\sum_{t:\ x_{t-1} \in I_{j,k}} \tilde{\tau}^{(t)} - T E v(X_0) I(X_0 \in I_{j,k})\right| \ > \ [\sqrt{T P(X_0 \in I_{j,k})} T^{\delta} + T^{\delta}] \\ \qquad\qquad\qquad\qquad\qquad\qquad\qquad \text{for any } (j,k) \in \mathcal{I}_T \setminus \{(0,k)\} \end{array} \right) = O(T^{-\lambda}) \qquad (5.18)$$

uniformly in $(x_0, \ldots, x_{T-1}) \in \Omega_T$, where $P((X_0, \ldots, X_{T-1}) \notin \Omega_T) = O(T^{-\lambda})$. Here and in the following Ω_T denotes an appropriate set of "not too irregular" realizations of $(X_0, \ldots, X_{T-1})$.

Define

$$S_{j,k} = \sum_{t=1}^{T} E\tau^{(t)} I(X_{t-1} \in I_{j,k}) - [\sqrt{T P(X_0 \in I_{j,k})} T^{\delta} + T^{\delta}].$$

Further, we define

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

(Then $S_{j,k} = \sum_{(l,m):\ I_{l,m} \subseteq I_{j,k}} T_{l,m}$.)
By (5.17) and (5.18) we obtain (5.12) and (5.13).

(vi) *Conclusions for* $|Z_{j,k} - Z'_{j,k}|$

By (5.12) we obtain with a probability exceeding $1 - O(T^{-\lambda})$ that

$$Z_{j,k} = \sum_{(l,m):\ I_{l,m} \subseteq I_{j,k}} W_{l,m}(T_{l,m}) + \sum_{t:\ X_{t-1} \in I_{j,k}} \sum_{(l,m):\ I_{j,k} \subset I_{l,m}} W_{l,m}(\tau_{l,m}^{(t)}) - W_{l,m}(\tau_{l,m}^{(t-1)}), \tag{5.19}$$

and, by (5.13),

$$Z'_{j,k} = \sum_{(l,m):\ I_{l,m} \subseteq I_{j,k}} W_{l,m}(T_{l,m}) + \sum_{t:\ x_{t-1} \in I_{j,k}} \sum_{(l,m):\ I_{j,k} \subset I_{l,m}} W_{l,m}(\tilde{\tau}_{l,m}^{(t)}) - W_{l,m}(\tilde{\tau}_{l,m}^{(t-1)}), \tag{5.20}$$

which holds again with a probability exceeding $1 - O(T^{-\lambda})$ under the condition $(x_0, \ldots, x_{T-1}) \in \Omega_T$. At this point we see why our particular pairing of $\varepsilon_1, \ldots, \varepsilon_T$ with $\eta_1, \ldots, \eta_T$ 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 (5.19) and (5.20), respectively. These terms are random, but identical to each other. Assume now that $\min\{\sum_{t=1}^T \tau^{(t)} I(X_{t-1} \in I_{j,k}), \sum_{t=1}^T \tilde{\tau}^{(t)} I(x_{t-1} \in I_{j,k})\} \geq T_{j,k}$ is satisfied. By (5.17) and (5.18) we have that

$$\begin{aligned} \sum_{t:\ X_{t-1} \in I_{j,k}} \sum_{(l,m):\ I_{j,k} \subset I_{l,m}} \tau_{l,m}^{(t)} - \tau_{l,m}^{(t-1)} &= \sum_{t:\ X_{t-1} \in I_{j,k}} \tau^{(t)} - S_{j,k} \\ &= \tilde{O}\left(\sqrt{TP(X_0 \in I_{j,k})} T^{\delta} + T^{\delta}, T^{-\lambda}\right) \end{aligned}$$

and

$$\begin{aligned} \sum_{t:\ x_{t-1} \in I_{j,k}} \sum_{(l,m):\ I_{j,k} \subset I_{l,m}} \tilde{\tau}_{l,m}^{(t)} - \tilde{\tau}_{l,m}^{(t-1)} &= \sum_{t:\ x_{t-1} \in I_{j,k}} \tilde{\tau}^{(t)} - S_{j,k} \\ &= \tilde{O}\left(\sqrt{TP(X_0 \in I_{j,k})} T^{\delta} + T^{\delta}, T^{-\lambda}\right). \end{aligned}$$

Note that, for fixed t and under $X_{t-1} \in I_{j,k}$, the pieces $\left\{W_{l,m}(s),\ \tau_{l,m}^{(t-1)} \leq s \leq \tau_{l,m}^{(t)}\right\}$ 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 $W_{j,k}^{res,t}$ on the interval $[0, \tau_{j,k}^{res,t}]$, where $\tau_{j,k}^{res,t} = \sum_{(l,m):\ I_{j,k} \subset I_{l,m}} (\tau_{l,m}^{(t)} - \tau_{l,m}^{(t-1)})$. This is achieved by setting

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

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

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

$$W_{j,k}^{res}(s) \;=\; \begin{cases} W_{j,k}^{res,1}(s), \quad \text{if } 0 \le s \le \tau_{j,k}^{res,1} \\ W_{j,k}^{res,1}(\tau_{j,k}^{res,1}) + \ldots + W_{j,k}^{res,u-1}(\tau_{j,k}^{res,u-1}) + W_{j,k}^{res,u}(s - \tau_{j,k}^{res,1} - \ldots - \tau_{j,k}^{res,u-1}), \\ \qquad \text{if } \tau_{j,k}^{res,1} + \ldots + \tau_{j,k}^{res,u-1} \le s < \tau_{j,k}^{res,1} + \ldots + \tau_{j,k}^{res,u} \end{cases}$$

An analogous construction can be made for the $\widetilde{\tau}_{l,m}^{(t)}$'s, leading to a Wiener process $\widetilde{W}_{j,k}^{res}$.
Note that $\tau_{j,k}^{res,1} + \ldots + \tau_{j,k}^{res,T} \;=\; \sum_{t:\ X_{t-1}\in I_{j,k}} \tau^{(t)} \;-\; S_{j,k}$. Now we obtain by Lemma 1.2.1 in Csörgő and Révész (1981, p. 29) that

$$\begin{aligned} |Z_{j,k} \;-\; Z'_{j,k}| \;&\le\; \left| \sum_{t:\ X_{t-1}\in I_{j,k}} \sum_{(l,m):\ l>j, I_{j,k}\subset I_{l,m}} W_{l,m}(\tau_{l,m}^{(t)}) \;-\; W_{l,m}(\tau_{l,m}^{(t-1)}) \right| \\ &\quad + \left| \sum_{t:\ x_{t-1}\in I_{j,k}} \sum_{(l,m):\ l>j, I_{j,k}\subset I_{l,m}} W_{l,m}(\widetilde{\tau}_{l,m}^{(t)}) \;-\; W_{l,m}(\widetilde{\tau}_{l,m}^{(t-1)}) \right| \\ &= \left| W_{j,k}^{res}\Big(\sum_{t:\ X_{t-1}\in I_{j,k}} \tau^{(t)} \;-\; S_{j,k}\Big) \right| + \left| \widetilde{W}_{j,k}^{res}\Big(\sum_{t:\ x_{t-1}\in I_{j,k}} \widetilde{\tau}^{(t)} \;-\; S_{j,k}\Big) \right| \\ &= \widetilde{O}\left((TP(X_0 \in I_{j,k}))^{1/4} T^{\delta}, T^{-\lambda} \right), \end{aligned}$$

which finishes the proof. $\square$

Proof of Corollary 2.1. We choose j^* such that $T2^{-j^*} \asymp T^{\delta}$. Assume throughout this proof that

$$|Z_{j,k} \;-\; Z'_{j,k}| \;\le\; (T2^{-j})^{1/4} T^{\delta} \quad \text{for all } (j,k) \in \mathcal{I}_T, \tag{5.21}$$

which is satisfied with a probability exceeding $1 - O(T^{-\lambda})$. Further, assume that

$$\sum_{t:\ X_{t-1}\in I_{j^*,k}} |\varepsilon_t| \;+\; \sum_{t:\ x_{t-1}\in I_{j^*,k}} |\eta_t| \;\le\; C_\lambda T 2^{-j^*} \quad \text{for all } k, \tag{5.22}$$

which is also fulfilled with a probability exceeding $1 - O(T^{-\lambda})$ for an appropriate choice of C_λ. To prove the assertion we use an approach similar to the proof of Lemma 2.2. We approximate w again by a truncated Haar wavelet series expansion

$$\widetilde{w}(z) \;=\; \sum_k \beta_k \phi_k(z) \;+\; \sum_{0\le j<j^*} \sum_k \beta_{j,k}\psi_{j,k}(z), \tag{5.23}$$

where $\beta_k = \int \phi_k(z) w(z)\, dz$, $\beta_{j,k} = \int \psi_{j,k}(z) w(z)\, dz$. We have

$$\sum_k |\beta_k| \;=\; O(\|w\|_1) \tag{5.24}$$

and

$$\sum_k |\beta_{j,k}| \;=\; O\left(\min\{\|\psi_{j,k}\|_\infty \|w\|_1, \|\psi_{j,k}\|_1 TV(w)\}\right) \;=\; O(\min\{2^{j/2}\|w\|_1, 2^{-j/2}TV(w)\}). \tag{5.25}$$

This implies that

$$\begin{aligned}
&\left|\sum_{t=1}^T \widetilde{w}(X_{t-1})\varepsilon_t \;-\; \sum_{t=1}^T \widetilde{w}(x_{t-1})\eta_t\right| \\
&\quad\le\; \left|\sum_k \beta_k [Z_{0,k} \;-\; Z'_{0,k}]\right| \\
&\qquad +\left|\sum_{0\le j<j^*} \sum_k \beta_{j,k} \sum_t [\psi_{j,k}(X_{t-1})\varepsilon_t \;-\; \psi_{j,k}(x_{t-1})\eta_t]\right| \\
&\quad\le\; O\left(\|w\|_1 T^{1/4} T^\delta\right) \;+\; \sum_{0\le j<j^*} \sum_k |\beta_{j,k}| \|\psi_{j,k}\|_\infty \max_l \left\{|Z_{j+1,l} - Z'_{j+1,l}|\right\} \\
&\quad=\; O\left(\|w\|_1 T^{1/4} T^\delta\right) \;+\; \sum_{0\le j<j^*} O\left(\min\{2^{j/2}\|w\|_1, 2^{-j/2}TV(w)\} 2^{j/2}(T2^{-j})^{1/4} T^\delta\right) \\
&\quad=\; O\left(T^{1/4}(TV(w))^{3/4}\|w\|_1^{1/4} T^\delta\right).
\end{aligned} \tag{5.26}$$

Further we have

$$\sum_k \|w \;-\; \widetilde{w}\|_{L_\infty(I_{j^*,k})} \;\le\; TV(w),$$

which implies that

$$\begin{aligned}
&\sum_t \left(w(X_{t-1}) \;-\; \widetilde{w}(X_{t-1})\right)\varepsilon_t \\
&\quad=\; \sum_k \sum_{t:\; X_{t-1}\in I_{j^*,k}} \left(w(X_{t-1}) \;-\; \widetilde{w}(X_{t-1})\right)\varepsilon_t \\
&\quad=\; O\left(T2^{-j^*} TV(w)\right),
\end{aligned} \tag{5.27}$$

and, analogously,

$$\sum_t \left(w(x_{t-1}) \;-\; \widetilde{w}(x_{t-1})\right)\eta_t \;=\; O\left(T2^{-j^*} TV(w)\right). \tag{5.28}$$

The assertion follows now from (5.26) to (5.28). □

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 2.1. □

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 $\xi_t \sim N(0, var(\eta_t))$ as well as a second set of independent random variables in the bootstrap domain $\xi_t^* \sim N(0, var(\varepsilon_t^*))$, 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

$$\sum_t \overline{w}_h(x,x_{t-1})\eta_t - \sum_t \overline{w}_h(x,x_{t-1})\varepsilon_t^*$$

$$= \sum_t \overline{w}_h(x,x_{t-1})(\eta_t - \xi_t) + \sum_t \overline{w}_h(x,x_{t-1})(\xi_t - \xi_t^*) + \sum_t \overline{w}_h(x,x_{t-1})(\xi_t^* - \varepsilon_t^*)$$

$$= S_1(x) + S_2(x) + S_3(x). \tag{5.29}$$

First we pair the random variables $\xi_1,\dots,\xi_T$ with the random variables $\xi_1^*,\dots,\xi_T^*$ 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 $\underline{\xi} = (\xi_1,\dots,\xi_T)'$ and $\underline{\xi}^* = (\xi_1^*,\dots,\xi_T^*)'$ into $\Delta \asymp h^{-1}$ packages of length $d_j \asymp Th$, respectively, that is

$$\underline{\xi} = (\xi_{11},\dots,\xi_{1d_1},\dots,\xi_{\Delta 1},\dots,\xi_{\Delta d_\Delta})'. \tag{5.30}$$

($\underline{\xi}^*$ is splitted 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 (5.30). Further, let $V_j = \sum_{k=1}^{d_j} v_{jk}$, $V_j^* = \sum_{k=1}^{d_j} v_{jk}^*$ $(j=1,\dots,\Delta)$. We define

$$t_{jk} = \sum_{l\le k} v_{jl} \quad , \quad t_{jk}^* = \sum_{l\le k} v_{jl}^*,$$

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

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

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

and

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

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

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

and a "fine structure" term

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

We can easily show that

$$\max_{j,k}\left\{|t_{jk} - t_{jk}^*|\right\} = \sum_{l\le k}(\varepsilon_{jl}^2 - v_{jl}) + \sum_{l\le k}(\widehat{\varepsilon}_{jl}^2 - \varepsilon_{jl}^2) = \tilde{O}\left((Th)^{1/2}T^\delta, T^{-\lambda}\right), \tag{5.31}$$

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

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

Therefore we have

$$\sup_x\{|S_{21}(x)|\} \;=\; \tilde{O}\left((Th)^{-1}T^{\delta}, T^{-\lambda}\right). \qquad (5.32)$$

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 (5.31) 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

$$S_{22}(x) \;=\; \tilde{O}\left((Th)^{-1}T^{\delta}, T^{-\lambda}\right). \qquad (5.33)$$

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

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

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

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

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

$$\begin{aligned} \sup_{x \in [a,b]} \{|S_1(x)|\} \;&\le\; \sup_x \left\{\sum_{t=1}^{T-1} |\overline{w}_h(x, x_{t-1}) - \overline{w}_h(x, x_t)||P_t - \tilde{P}_t| \;+\; |\overline{w}_h(x, x_{T-1})||P_T - \tilde{P}_T|\right\} \\ &=\; \tilde{O}\left((Th)^{-1}T^{\delta}, T^{-\lambda}\right). \end{aligned} \qquad (5.34)$$

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

$$\sup_{x \in [a,b]} \{|S_3(x)|\} \;=\; \tilde{O}\left((Th)^{-1}T^{\delta}, T^{-\lambda}\right). \qquad (5.35)$$

The assertion follows now from (5.29) and (5.32) to (5.35). $\square$

Proof of Theorem 4.2. This proof is analogous to that of Theorem 2.3 in Neumann and Polzehl (1995) and is essentially based on the fact that

$$\sup_{x \in [a,b]} \{|\hat{v}(x) \;-\; v(x)|\} \;=\; \tilde{O}\left(T^{\delta}(Th)^{-3/2}, T^{-\lambda}\right).$$

$\square$

References

1. Amosova, N. N. (1972). On limit theorems for probabilities of moderate deviations. *Vestnik Leningrad. Univ.* **13**, 5–14. (in Russian)
2. Csörgő, M. and Révész, P. (1981). *Strong Approximations in Probability and Statistics.* Academic Press, New York.
3. Doukhan, P. (1995). *Mixing: Properties and Examples. Lecture Notes in Statistics* **85**, Springer, New York.
4. Fan, J. (1992). Design-adaptive nonparametric regression. *J. Amer. Statist. Assoc.* **87**, 998–1004.
5. Fan, J. (1993). Local linear regression smoothers and their minimax efficiecies. *Ann. Statist.* **21**, 196–216.
6. Fan, J. and Gijbels, I. (1992). Variable bandwidth and local linear regression smoothers. *Ann. Statist.* **20**, 2008–2036.
7. Fan, J. and Gijbels, I. (1995). *Local Polynomial Modeling and Its Application – Theory and Methodologies.* Chapman and Hall, New York.
8. Franke, J., Kreiss, J.-P. and Mammen, E. (1996). Bootstrap of kernel smoothing in nonlinear time series, Preprint.
9. Hall, P. (1991). On the distribution of suprema. *Probab. Theory Rel. Fields* **89**, 447–455.
10. Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Application.* Academic Press, New York.
11. Härdle, W. and Mammen, E. (1993) Comparing nonparametric versus parametric regression fits. *Ann. Statist.* **21**, 1926–1947.
12. Härdle, W. and Tsybakov, A. B. (1995). Local polynomial estimators of the volatility function in nonparametric autoregression, Discussion Paper 42, SFB 373, Berlin.
13. Mammen, E. (1992). *When Does the Bootstrap Work? Lecture Notes in Statistics* **77**, Springer, New York.
14. Masry, E. and Tjøstheim, D. (1995). Nonparametric estimation and identification of nonlinear ARCH time series. *Econometric Theory* **11**, 258–289.
15. Neumann, M. H. and Polzehl, J. (1995). Simultaneous bootstrap confidence bands in nonparametric regression, Manuscript.
16. Robinson, P. M. (1983). Nonparametric estimators for time series. *J. Time Ser. Anal.* **4**, 185–207.
17. Sakhanenko, A. I. (1991). On the accuracy of normal approximation in the invariance principle. *Siberian Advances in Mathematics* **1**, 58–91.
18. Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics.* Wiley, New York.
19. Skorokhod, A. V. (1965). *Studies in the Theory of Random Processes.* Addison-Wesley, Reading, Massachussetts.
20. Stone, C. J. (1977). Consistent nonparametric regression. *Ann. Statist.* **5**, 595–620.
21. Tjøstheim, D. (1994). Non-linear time series: a selective review. *Scand. J. Statist.* , 97–130.
22. Tjøstheim, D. and Auestad, B. (1994). Nonparametric identification of nonlinear time series: projections. *J. Am. Stat. Assoc.* **89**, 1398–1409.
23. Wu, C. F. J. (1986). Jackknife, bootstrap and other resampling methods in regression analysis. *Ann. Statist.* **14**, 1261–1343.