ASYMPTOTIC PROPERTIES OF NONPARAMETRIC TIME SERIES PREDICTION

by

Young K. Truong
University of North Carolina, Chapel Hill

and

Charles S. Stone
University of California, Berkeley

Institute of Statistics Mimeo Series No. 1830

July 1987

ASYMPTOTIC PROPERTIES OF NONPARAMETRIC TIME SERIES PREDICTION

By YOUNG K. TRUONG
Department of Biostatistics
University of North Carolina, Chapel Hill

and

CHARLES S. STONE
Department of Statistics
University of California, Berkeley

Abstract. Let $(\mathbf{X}_t, Y_t)$ be an $(d+1)$ vector-valued stationary series, $t = 0, \pm 1, \pm 2, \ldots$ with $\mathbf{X}_t$ d vector-valued and Y_t real valued. Set $\theta(\mathbf{X}_0) = E(Y_0|\mathbf{X}_0)$ and let $\hat{\theta}_n(\cdot)$ be an estimator of $\theta(\cdot)$ based upon a realization $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of length n. Set $r = (2+d)^{-1}$. Under some regularity conditions, $\hat{\theta}_n(\cdot)$ can be chosen to achieve the optimal rate of convergence n^{-r} both pointwise and in L^2 norm restricted to compacts. Alternatively, set $\theta(\mathbf{X}_0) = \text{Median}(Y_0|\mathbf{X}_0)$ and let $\hat{\theta}_n(\cdot)$ be an estimator of $\theta(\cdot)$ based upon a realization $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ of length n. Under some regularity conditions, $\hat{\theta}_n(\cdot)$ can be chosen to achieve the same optimal rate of convergence n^{-r} both pointwise and in L^q norms $(1 \leq q < \infty)$ restricted to compacts.

Keywords. Kernel estimators, local averages, local medians, time series prediction, nonparametric regression.

1. INTRODUCTION

One of the most important problems in univariate time series analysis is that of *predicting* (or *forcasting*) a future value of a discrete time stationary random process from its past values. This prediction problem may be described as follows: Let n, m and d be positive integers and suppose that X_t, $t = 0, \pm 1, \pm 2, \ldots$ is a discrete time, strictly stationary random process. It is desired to predict the value of X_{n+m} from $X_{n-d+1}, \ldots, X_n$.

An m-th step prediction rule is a function $\theta(x_{n-d+1}, \ldots, x_n)$ of the past; that is, $\theta(\cdot)$ is a (Borel measureable) real-valued function defined on $\mathbf{R}^d$. Construction of such a rule can have two purposes: (i) to predict the future X_{n+m} as accurately as possible; (ii) to understand the structural relationship between the future X_{n+m} and the past $(X_{n-d+1}, \ldots, X_n)$.

For instance, a company may be interested in predicting the sale of a particular commoditity in the next year based upon the sales of the last few years. Here accuracy is the critical element. In the series of Wolfer's annual sunspot numbers (Morris, 1977), it was observed that the series is generated by a nonlinear mechanism. The point of constructing a prediction rule was to understand the relationship between say, this year's sunspot number and those of the last two years.

In bivariate time series analysis, one of the main concerns is to investigate the relationship between the input series $\{X_t; t = 0, \pm 1, \ldots\}$ and the output series $\{Z_t; t = 0, \pm 1, \ldots\}$. Here it is useful to consider Z_n as a function of $(X_{n-d+1}, \ldots, X_n)$.

Note that by letting $\mathbf{X}_n = (X_{n-d+1}, \ldots, X_n)$ and $Y_n = X_{n+m}$ in the univariate case, or $\mathbf{X}_n = (X_{n-d+1}, \ldots, X_n)$ and $Y_n = Z_n$ in the bivariate case; then the above set ups are special cases of the following situation: Let $(\mathbf{X}, Y)$ be a pair of random variables that are respectively d and 1 dimensional; the random variable Y is called the response and the random vector $\mathbf{X}$ is refered to as the predictor variable. One of the important problems in statistics is to construct a function $\theta(\cdot)$ in order to (i) study the relationship between the response and the explanatory variable or (ii) obtain the predictor $\theta(\mathbf{X})$ of Y based on $\mathbf{X}$.

The simplest and most widely used measure of accuracy of $\theta(\mathbf{X})$ as a predictor of Y

is the *Mean Square Error*, $E|Y - \theta(\mathbf{X})|^2$. The function $\theta(\cdot)$ which minimizes this measure of accuracy is the regression function of Y on $\mathbf{X}$, defined by $\theta(\mathbf{X}) = E(Y|\mathbf{X})$.

Recently, there has been an increasing interest in adopting the *Mean Absolute Deviation* $E|Y - \theta(\mathbf{X})|$ as a measure of accuracy, especially when outliers may be present (Bloomfield and Steiger, 1983). The optimal function $\theta(\cdot)$ is now defined so that $\theta(\mathbf{X})$ is the conditional median, Median$(Y|\mathbf{X})$, of Y given $\mathbf{X}$. Note that this function is not necessary uniquely defined.

In practice, it is necessary to construct estimators of these functions based on a set of observations. Time series prediction is the generic term revolving around the construction of estimators of these predictors based on a realization $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ from the stationary process $(\mathbf{X}_t, Y_t), t = 0, \pm 1, \ldots$.

Parametric Approach vs Nonparametric Approach

To estimate these predictors, the *parametric* approach starts with specific assumptions about the relationship between the response (or future) and the explanatory variables (or past) and about the variation in the response (future) that may or may not be accounted for by the explanatory variables. For instance, the standard regression method (or autoregressive method in time series) starts with an a priori model for the regression function $\theta(\cdot)$ which, by assumption or prior knowledge, is a linear function that contains finitely many unknown parameters. Under the assumption that the joint distribution is Gaussian, it is an optimal prediction rule; if the distribution is non-Gaussian, it is not generally posible to determine the funtion $\theta(\cdot)$; so one might settle for the *best* linear predctor. By constrast, in the *nonparametric* approach, the regression function will be estimated directly without assuming such an a priori model for $\theta(\cdot)$. As pointed out in Stone (1985), the nonparametric approach is more *flexible* than the standard regression method; *flexibility* means the ability of the model to provide accurate fits in a wide variety of realistic situations, inaccuracy here leading to *bias* in estimation. In recent years, nonparametric estimation has become an active area in time series analysis because of its flexibility in fitting data (Bierens, 1983; Collomb, 1982; Collomb and Hardle, 1984; Robinson, 1983).

The present approach deals with the asymptotic properties (in terms of rates of convergence) of a class of nonparametric estimators constructed by kernel methods based on local averages and local medians. It is hoped that the results obtained here serve as a starting point for further development and understanding of the sampling properties of more complicated nonparametric procedures involving robustification, local polynomial fits, additive regression, and spline approximation.

Some previous work on nonparametric estimation in time series will be surveyed in the next section.

2. DEVELOPMENTS IN TIME SERIES PREDICTION

The theory and practice of linear model fitting has now attained a refined state (Brillinger, 1980; Priestley, 1979); see, for example, the work of Akaike (1974a, 1974b) on the fundamental structural properties of these models and the definitive work of Hannan (1973) and Dunsmuir and Hannan (1976) on the inferential side. While the study of non-linear models in time series is still in its early stages, what has been learned so far is sufficient to indicate that this is a very rich and potentially rewarding field. Analysis of particular series have shown that non-linear models can provide better fits to the data (as one would expect) and, more importantly, that the structure underlying the data can not be captured by linear models.

So far, the study of non-linear models has been restricted to a few specific forms. For example, Priestley (1980), Tong and Lim (1980), Nicholls and Quinn (1980), and Haggan and Ozaki (1980, 1981) consider various non-linear filters of, possibly independent, identically distributed Gaussian random variables. In practice it may be difficult to decide a priori, which, if any, of these models is best suited to a given set of data.

Asymptotic results for the conditional expectation has been established by Doukhan and Ghindes (1980), Collomb (1982), Bierens (1983) and Robinson (1983) under various mixing conditions. In Robinson (1983), pointwise consistency and a central limit theorem was obtained for kernel estimators based on local averages under the α-mixing condition. Collomb (1982) and Bierens (1983) considered the uniform consistency and rate of convergence for kernel estimators based on local averages under the ϕ-mixing condition, which is considerably stronger than the α-mixing condition. Collomb and Härdle (1984) considered the uniform rate of convergence (also under ϕ-mixing) for a class of robust nonparametric estimators that did not include local medians.

Under the α-mixing condition, the pointwise and the L^2 rates of convergence for nonparametric estimators of conditional expectations constructed by kernel methods based on local averages are described in Section 3. The pointwise and the L^q $(1 \leq q < \infty)$ rates of convergence for kernel estimators of the conditional medians based on local medians are also given in Section 3.

For this class of nonparametric estimators, the results presented there constitute an answer and an extension to one of the open questions of Stone (1982). In the random sample case, Härdle and Luckhaus (1984) considered the L^∞ rate of convergence for a class of robust nonparametric estimators including an estimator of the conditional median. But the problem of L^q $(1 \leq q < \infty)$ rates of convergence was still unsolved. Proofs of these results are given in Section 5.

3. NONPARAMETRIC TIME SERIES PREDICTION

Results on the local and global rates of convergence of nonparametric estimators of conditional expectations and conditional medians based on a realization of a discrete time stationary time series will be treated in this section. Recall that d is the dimensionality of the explainatory variable $\mathbf{X}$ and let U denote a nonempty bounded open neighborhood of the origin of $\mathbf{R}^d$. Let $\{(\mathbf{X}_i, Y_i), i = 0, \pm 1, \ldots\}$ be an $(d+1)$ vector-valued strictly stationary series and set $\theta(\mathbf{x}) = E(Y_0 | \mathbf{X}_0 = \mathbf{x})$ or, $\theta(\mathbf{x}) = \text{Median}(Y_0 | \mathbf{X}_0 = \mathbf{x})$. Let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ denote a realization of this process.

ASSUMPTION 1. *There is a positve constant M_0 such that*

$$|\theta(\mathbf{x}) - \theta(\mathbf{x}')| \leq M_0 \|\mathbf{x} - \mathbf{x}'\| \qquad \textit{for } \mathbf{x}, \mathbf{x}' \in U,$$

where $\|\mathbf{x}\| = (x_1^2 + \cdots + x_d^2)^{1/2}$ *for* $\mathbf{x} = (x_1, \ldots, x_d) \in \mathbf{R}^d$.

ASSUMPTION 2. *The distribution of* $\mathbf{X}_0$ *is absolutely continuous and its density* $f(\cdot)$ *is bounded away from zero and infinity. That is, there is a positive constant* M_1 *such that* $M_1^{-1} \leq f(\mathbf{x}) \leq M_1$ *for* $\mathbf{x} \in U$.

The following technical condition is required for bounding the variance of various terms in the proof. (See Lemmas 1–4.)

ASSUMPTION 3. *The conditional distribution of* $\mathbf{X}_1$ *given* $\mathbf{X}_0$ *is absolutely continuous and its density* $h(\cdot\,|\mathbf{x})$ *is bounded away from zero and infinity. That is* $M_1^{-1} \leq h(\mathbf{y}|\mathbf{x}) \leq M_1$ *for* $\mathbf{x}$ *and* $\mathbf{y} \in U$.

Collomb (1982) derived asymptotic properties for nonparametric estimators of conditional expectations based on bounded stationary time series. In order to extend the argument to include the unbounded time series, the following moment condition is required (Robinson, 1983).

ASSUMPTION 4. *There is a positive constant* $\nu > 2$ *such that*

$$\sup_{\mathbf{x} \in U} E(|Y_0|^{\nu} | \mathbf{X}_0 = \mathbf{x}) < \infty .$$

A weak dependence condition on the stationary sequence will now be described. Let $\mathcal{F}_t$ and $\mathcal{F}^t$ denote respectively the σ-fields generated by $\{(\mathbf{X}_i, Y_i) : -\infty < i \leq t\}$ and $\{(\mathbf{X}_i, Y_i) : t \leq i < \infty\}$. Given a positive integer k, set

$$\alpha(k) = \sup\{\,|\,P(A \cap B) - P(A)P(B)\,| : A \in \mathcal{F}_t \ \text{ and } \ B \in \mathcal{F}^{t+k}\,\} .$$

The stationary sequence is said to be α-mixing or strongly mixing if $\alpha(k) \to 0$ as $k \to \infty$ (Rosenblatt, 1956).

ASSUMPTION 5. *(i) The stationary sequence is* α*-mixing and*

$$\sum_{i \geq N} \alpha^{1-\frac{2}{\nu}}(i) = O(N^{-1}) \qquad \text{as } N \to \infty.$$

(ii) The stationary sequence is α-mixing and

$$\alpha(N) \sim \rho^N \qquad \text{as } N \to \infty \text{ for some } \rho \text{ with } 0 < \rho < 1.$$

A condition on the conditional distribution of Y given $\mathbf{X}$ is required to guarantee the uniqueness of the conditional median (uniqueness will ensure consistency) and also the achievability of the desired rate of conververgence. If the conditional density is not bounded away from zero around the median the desired rate of convergence will not be achievable. (The same condition is required in order to obtain the usual asymptotic result about the sample median in the univariate case.) In the following condition, $\theta(\cdot)$ denotes the conditional median.

ASSUMPTION 6. *The conditional distribution of Y_0 given $\mathbf{X}_0 = \mathbf{x}$ is absolutely continuous and its density $g(y|\mathbf{x}, \theta)$ is bounded away from zero and infinity over a neighborhood of the median θ. That is, there is a positive constant ϵ_0 such that $M_1^{-1} \leq g(y|\mathbf{x}, \theta) \leq M_1$ for $\mathbf{x} \in U$ and $y \in (\theta(\mathbf{x}) - \epsilon_0, \theta(\mathbf{x}) + \epsilon_0)$.*

Given positive numbers a_n and b_n, $n \geq 1$, let $a_n \sim b_n$ mean that a_n/b_n is bounded away from zero and infinity. Given random variables V_n, $n \geq 1$, let $V_n = O_{pr}(b_n)$ mean that the random variables $b_n^{-1}V_n$, $n \geq 1$ are bounded in probability or, equivalently, that

$$\lim_{c \to \infty} \limsup_{n} P(|V_n| > cb_n) = 0.$$

The kernel estimators of conditional expectations and conditional medians will now be described. Set $r = (2 + d)^{-1}$. For each $n \geq 1$, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a realization of the (strictly) stationary time series and let δ_n denote a sequence of positive numbers such that $\delta_n \sim n^{-r}$. (In the univariate case, set $\mathbf{X}_i = (X_{i-d+1}, \ldots, X_i)$ for $i \geq d$.) Set $I_n(\mathbf{x}) = \{i : 1 \leq i \leq n \text{ and } \|\mathbf{X}_i - \mathbf{x}\| \leq \delta_n\}$ and $N_n(\mathbf{x}) = \#(I_n(\mathbf{x}))$. Also set $\hat{\theta}_n(\mathbf{x}) = N_n(\mathbf{x})^{-1} \sum_{I_n(\mathbf{x})} Y_i$ if $\theta(\cdot)$ is the conditional expectation; and $\hat{\theta}_n(\mathbf{x}) = \text{Median}\{Y_i : i \in I_n(\mathbf{x})\}$ if $\theta(\cdot)$ is the conditional median.

In Theorems 1 and 2, $\theta(\cdot)$ denotes the conditional expectation and $\hat{\theta}_n$ its kernel estimator based on local averages.

THEOREM 1. *Suppose that Assumptions 1–4 and 5(i) hold. Then*

$$|\hat{\theta}_n(\mathbf{0}) - \theta(\mathbf{0})| = O_{pr}(n^{-r}).$$

The proof of this theorem, which will be given in Section 5, is basically a refinement of the corresponding one given in Stone (1980), with additional arguments involving asymptotic independence (see Lemmas 1–4).

Let C be a fixed compact subset of U having a nonempty interior and let $g(\cdot)$ be a real-valued function on $\mathbf{R}^d$. Set

$$\|g\|_q = \left\{ \int_C |g(\mathbf{x})|^q \, d\mathbf{x} \right\}^{\frac{1}{q}}, \qquad 1 \le q < \infty.$$

THEOREM 2. *Suppose that Assumptions 1–4 and 5(ii) hold. Then*

$$\|\hat{\theta}_n - \theta\|_2 = O_{pr}(n^{-r}).$$

The proof of this theorem will be given in Section 5. The argument is a refinement of the corresponding one for Theorem 1.

In Theorems 3 and 4, $\theta(\cdot)$ denotes the conditional median and $\hat{\theta}_n$ its kernel estimator based on local medians.

THEOREM 3. *Suppose that Assumptions 1–3, 5(i) and 6 hold. Then*

$$|\hat{\theta}_n(\mathbf{0}) - \theta(\mathbf{0})| = O_{pr}(n^{-r}).$$

The proof of this theorem will be given in Section 5.

THEOREM 4. *Suppose that Assumptions 1–3, 5(ii) and 6 hold. Then*

$$\|\hat{\theta}_n - \theta\|_q = O_{pr}(n^{-r}), \qquad 1 \le q < \infty.$$

The proof of this theorem, which will be given in Section 5, uses a result on uniform consistency (Lemma 5) and the argument is a refinement of that in the i.i.d. case.

With a simple modification of Assumption 6, Theorems 3 and 4 are easily extended to yield rates of convergence for nonparametric estimators of other conditional quantiles.

4. DISCUSSION

For $n \geq 1$, let $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$ be a random sample of size n from the distribution of $(\mathbf{X}, Y)$ and let k denote a non-negative integer. Let $\theta(\cdot)$ be the regression function of Y on $\mathbf{X}$ and suppose that $\theta(\cdot)$ has bounded $(k+1)$th derivative. Set $r = p/(2p+d)$ where $p = k+1$. Stone (1980, 1982) showed that if $1 \leq q < \infty$, then n^{-r} is the optimal rate of convergence in both pointwise and L^q norms; while $(n^{-1}\log n)^{-r}$ is the optimal rate of convergence in L^∞ norm. To find an estimator of $\theta(\cdot)$ that achieves these optimal rates of concergence, given $\mathbf{x}$, let $\hat{P}_n(\cdot;\mathbf{x})$ be the polynomial on $\mathbf{R}^d$ of degree k that minimizes

$$\sum_{I_n(\mathbf{x})} [\, Y_i - \hat{P}_n(\mathbf{X}_i;\mathbf{x})\,]^2$$

and set $\hat{\theta}_n(\mathbf{x}) = \hat{P}_n(\mathbf{x};\mathbf{x})$ (if $q = \infty$, define $\hat{\theta}_n$ as above over a finite subset of C and then extend it to all of C by suitable interpolation). Note that this estimator can be easily obtained by solving the corresponding normal equation.

Based on results presented in the previous sections, the following generalization to the case of conditional medians seems plausible. Suppose that the conditional median $\theta(\cdot)$ has bounded pth derivative. To find an estimator that achieves the above L^q $(1 \leq q \leq \infty)$ rates of convergence, given $\mathbf{x}$, let $\hat{P}_n(\cdot;\mathbf{x})$ be a polynomial on $\mathbf{R}^d$ of degree k which minimizes

$$\sum_{I_n(\mathbf{x})} |Y_i - \hat{P}_n(\mathbf{X}_i;\mathbf{x})|$$

and set $\hat{\theta}_n(\mathbf{x}) = \hat{P}_n(\mathbf{x};\mathbf{x})$. Though there may not be a unique solution, this numerical optimization problem is readily solved by the simplex method (see, for example, Bloomfield and Steiger (1983)). The corresponding generalization to time series is straightforward.

One drawback that the nonparametric approach has is the high *dimensionality*, which can be thought of in terms of the *variance* in estimation. In other words: A *huge* data set may be required for nonparametric estimation of a function of many variables; otherwise the variance of the estimator may be unacceptably large. This drawback is serious especially in time series analysis where the future usually depends on much of the past.

A possible solution would be to use *additivity* as in Stone (1985) to alleviate *curse of dimensionality*. More formally, let $\theta(\cdot)$ be the regression function defined on $\mathbf{R}^d$ and suppose that θ is additive; that is, that there is smooth functions $\theta_1(\cdot), \ldots, \theta_d(\cdot)$ defined on $\mathbf{R}^1$ such that

$$\theta(x_1, \ldots, x_d) = \mu + \theta_1(x_1) + \cdots + \theta_d(x_d),$$

where $\mu = E(Y)$. Using B-splines, an estimator of $\theta(\cdot)$ can be constructed to achieve the optimal rates of convergence n^{-r}, where r now is equal to $p/(2p+1)$. The rates of convergence here do not depend on the dimensional parameter d. Another nice feature about this estimator is that it is smoother and is as flexible as ordinary nonparametric procedures constructed by the kernel method.

The corresponding methodology is generalized immediately to time series, and it is an interesting open problem to determine whether the asymptotic properties described above (with r independent of d) also hold in this context. Another interesting question is to extend Theorem 4 to include the L^∞ rate of convergence under the same asssumptions.

5. PROOF OF THEOREMS

For each $i = 1, \ldots, n$, set $K_i = 1_{\{\|\mathbf{X}_i\| \leq \delta_n\}}$. The following lemma is an immediate consequence of Assumptions 2 and 3.

Lemma 1. *There is a positive constant C_1 such that*

$$E(K_i K_{i+j}) \leq C_1 \delta_n^{2d}.$$

Lemma 2. $\mathrm{Var}(\sum_i K_i) = O(n\delta_n^d)$.

Proof. By Theorem A.5 of Hall and Heyde (1980, p.277), $|\mathrm{Cov}(K_i, K_{i+j})| \leq 4\alpha(j)$. Thus by Assumption 5(i) and Lemma 1,

$$\mathrm{Var}(\textstyle\sum_i K_i) = n\mathrm{Var}(K_1) + 2\sum_i \sum_j \mathrm{Cov}(K_i, K_{i+j})$$

$$= O\left(n\delta_n^d + n\textstyle\sum_1^n \min\left(\alpha(j), \delta_n^{2d}\right)\right) = O(n\delta_n^d).$$

The following result follows from Tchebychev's inequality, Lemma 2 and Assumption 2.

Lemma 3. *There is a positive constant k_1 such that*

$$\lim_n P(\textstyle\sum_i K_i \leq k_1 n\delta_n^d) = 0.$$

Lemma 4. $\mathrm{Var}(\sum_i K_i[\,Y_i - \theta(\mathbf{X}_i)\,]) = O(n\delta_n^d)$.

Proof. (Robinson, 1983) Let B be a positive constant and set

$$Y_i' = Y_i 1_{\{|Y_i| \leq B\}};\ Y_i'' = Y_i 1_{\{|Y_i| \geq B\}}.$$

$$\theta'(\mathbf{X}_i) = E[\,Y_i'|\mathbf{X}_i\,],\ \theta''(\mathbf{X}_i) = E[\,Y_i''|\mathbf{X}_i\,].$$

Then $Y_i = Y_i' + Y_i''$ and $\theta(\mathbf{X}_i) = \theta'(\mathbf{X}_i) + \theta''(\mathbf{X}_i)$.

Set $Z_i = Y_i' - \theta'(\mathbf{X}_i)$. Observe that $|Z_i| \leq 2B$ and $E(Z_i|\mathbf{X}_i) = 0$. By the argument used in the proof of Lemma 2,

$$\begin{aligned}\mathrm{Var}(\textstyle\sum_i K_i Z_i) &= n\mathrm{Var}(K_1 Z_1) + 2\textstyle\sum_i \sum_j \mathrm{Cov}(K_i Z_i, K_{i+j} Z_{i+j}) \\ &= O\left(n\delta_n^d + n\textstyle\sum_1^n \min(\alpha(j), \delta_n^{2d})\right) = O(n\delta_n^d). \qquad (5.1)\end{aligned}$$

Set $W_i = Y_i'' - \theta''(\mathbf{X}_i)$. Applying Holder's inequality twice,

$$\begin{aligned}&E(K_i|W_i|K_{i+j}|W_{i+j}|) \\ &= E\left[\,(K_i|W_i|^{\nu})^{\frac{1}{\nu}}(K_{i+j}|W_{i+j}|^{\nu})^{\frac{1}{\nu}}(K_iK_{i+j})^{1-\frac{2}{\nu}}K_i^{\frac{1}{\nu}}K_{i+j}^{\frac{1}{\nu}}\right] \\ &\leq \{E\,[\,K_i|W_i|^{\nu}]\}^{\frac{2}{\nu}}\,\{E[\,K_iK_{i+j}\,]\}^{1-\frac{2}{\nu}}. \qquad (5.2)\end{aligned}$$

By Corollary A.2 of Hall and Heyde (1980, p.278),

$$E(K_i|W_i|K_{i+j}|W_{i+j}|) \leq 4\,\{E(K_i|W_i|^{\nu})\}^{\frac{2}{\nu}}\,\{\alpha(j)\}^{1-\frac{2}{\nu}}. \qquad (5.3)$$

According to Assumption 2,

$$E(K_i|W_i|^s) = E(K_i E(|W_i|^s|K_i))$$
$$\leq M_1 \sup_{\|\mathbf{y}\|\leq \delta_n} Q(\mathbf{y}) \int K_i(\mathbf{x})\,d\mathbf{x} = O(\delta_n^d) \text{ for } 1 \leq s \leq \nu, \tag{5.4}$$

where $Q(\mathbf{y}) = E(|W_i|^s|\mathbf{X}_i = \mathbf{y})$ is bounded in $\mathbf{y} \in U$ by Assumption 4. By (5.2)–(5.4), Lemma 1 and Assumption 5(i) (note that $E(W_i|\mathbf{X}_i) = 0$),

$$\mathrm{Var}(\textstyle\sum_i K_i W_i) = n\mathrm{Var}(K_1 W_1) + 2\sum_i \sum_j \mathrm{Cov}(K_i W_i, K_{i+j} W_{i+j})$$
$$= O\left(n\delta_n^d + n(\delta_n^d)^{\frac{2}{\nu}} \textstyle\sum_1^n \min\left\{\alpha^{1-\frac{2}{\nu}}(j), (\delta_n^{2d})^{1-\frac{2}{\nu}}\right\}\right) = O(n\delta_n^d). \tag{5.5}$$

It follows from (5.1) and (5.5) that

$$\mathrm{Var}(\textstyle\sum_i K_i[\,Y_i - \theta(\mathbf{X}_i)\,]) \leq 2\{\mathrm{Var}(\sum_i K_i Z_i) + \mathrm{Var}(\sum_i K_i W_i)\}$$
$$= O(n\delta_n^d),$$

which completes the proof of Lemma 4.

Proof of Theorem 1. According to Assumption 1

$$|\theta(\mathbf{X}_i) - \theta(\mathbf{0})| \leq M_0 \delta_n \qquad \text{for } i \in I_n.$$

Thus

$$|N_n^{-1} \textstyle\sum_{I_n} [\,\theta(\mathbf{X}_i) - \theta(\mathbf{0})\,]| = O_{pr}(n^{-r}). \tag{5.6}$$

On the other hand,

$$P(N_n^{-1}|\textstyle\sum_{I_n}[\,Y_i - \theta(\mathbf{X}_i)]| \geq cn^{-r})$$
$$\leq P(N_n^{-1}|\textstyle\sum_{I_n}[\,Y_i - \theta(\mathbf{X}_i)]| \geq cn^{-r}; N_n > k_1 n\delta_n^d) + P(N_n \leq k_1 n\delta_n^d)$$
$$\leq P(|\textstyle\sum_{I_n}[\,Y_i - \theta(\mathbf{X}_i)]| \geq k_1 c n^{-r} n\delta_n^d) + P(N_n \leq k_1 n\delta_n^d).$$

Hence, by Lemma 3, Lemma 4 and Tchebychev's inequality,

$$|N_n^{-1} \textstyle\sum_{I_n} [\,Y_i - \theta(\mathbf{X}_i)\,]| = O_{pr}(n^{-r}). \tag{5.7}$$

The conclusion of Theorem 1 follows from (5.6) and (5.7).

Proof of Theorem 2. We may assume that C is contained in the interior of the cube $C_0 = [-\frac{1}{2}, \frac{1}{2}]^d \subset U$. According to Assumption 1, there is a positive constant k_1 such that $|\theta(\mathbf{X}_i) - \theta(\mathbf{x})| \leq k_1 \|\mathbf{X}_i - \mathbf{x}\| \leq k_1 \delta_n$, for $i \in I_n(\mathbf{x})$ and $\mathbf{x} \in C$. Thus there is a positive constant k_2 such that

$$\lim_n P\left(\left| N_n(\mathbf{x})^{-1} \sum_{I_n(\mathbf{x})} [\, \theta(\mathbf{X}_i) - \theta(\mathbf{x})\,] \right| \geq k_2 \delta_n \text{ for some } \mathbf{x} \in C \right) = 0. \tag{5.8}$$

Set $Z_n(\mathbf{x}) = \sum_{i \in I_n(\mathbf{x})} [\, Y_i - \theta(\mathbf{X}_i)\,]$ By Lemma 4 and Assumption 4 (since Assumption 5(ii) is stronger than Assumption 5(i))

$$E[\, Z_n^2(\mathbf{x})\,] = O(n\delta_n^d) \qquad \text{uniformly over } \mathbf{x} \in C.$$

Consequently,

$$E\left[\int_C |Z_n(\mathbf{x})|^2 \, d\mathbf{x} \right] = \int_C E[\, |Z_n(\mathbf{x})|^2\,]\, d\mathbf{x} = O(n\delta_n^d). \tag{5.9}$$

According to Assumption 5(ii), there is a positive constant k_3 such that

$$\lim_n P(\Omega_n) = 1, \tag{5.10}$$

where $\Omega_n = \{ N_n(\mathbf{x}) \geq k_3 n \delta_n^d \text{ for } \mathbf{x} \in C \}$. (See the Appendix for the proof.)

By (5.9) and (5.10),

$$\begin{aligned} &P\left(\left\{ \int_C \left| N_n(\mathbf{x})^{-1} \sum_{I_n(\mathbf{x})} [\, Y_i - \theta(\mathbf{X}_i)\,] \right|^2 d\mathbf{x} \right\}^{\frac{1}{2}} \geq c(n^{-1}\delta_n^{-d})^{\frac{1}{2}} \right) \\ &\leq P(\Omega_n^c) + P\left(\int_C |Z_n(\mathbf{x})|^2 \, d\mathbf{x} \geq c^2 k_3^2 n \delta_n^d \right) \\ &= P(\Omega_n^c) + \frac{O(1) n \delta_n^d}{c^2 n \delta_n^d} = o(1) \qquad \text{as } n, c \to \infty. \end{aligned} \tag{5.11}$$

It follows from (5.8) and (5.11) that

$$\lim_{c \to \infty} \lim_n P\left(\|\hat{\theta}_n - \theta\|_2 \geq c\left(\delta_n + (n^{-1}\delta_n^{-d})^{\frac{1}{2}} \right) \right) = 0.$$

The Conclusion of Theorem 2 now follows by choosing δ_n so that $\delta_n = (n^{-1}\delta_n^{-d})^{\frac{1}{2}}$, or equivalently, $\delta_n = n^{-r}$.

Proof of Theorem 3. Let B_{ni} be the event that $\|\mathbf{X}_i\| \le \delta_n$. According to Assumption 1, $\theta(\mathbf{X}_i) \le \theta(\mathbf{0}) + M_0\delta_n$ whenever $\|\mathbf{X}_i\| \le \delta_n$. Thus

$$\tfrac{1}{2} - P(Y_i \ge \theta(\mathbf{0}) + c\delta_n \,|\, B_{ni}) \ge P(M_0\delta_n \le Y_i - \theta(\mathbf{0}) \le c\delta_n \,|\, B_{ni}).$$

Hence by Assumption 6

$$\tfrac{1}{2} - P(Y_i \ge \theta(\mathbf{0}) + c\delta_n \,|\, B_{ni}) \ge (c - M_0)M_1^{-1}\delta_n \qquad \text{for } c_0 > M_0. \tag{5.12}$$

Set $K_i = 1_{\{\|\mathbf{X}_i\| \le \delta_n\}}$ and $Z_i = 1_{\{Y_i \ge \theta(\mathbf{0}) + c\delta_n\}} - P(Y_i \ge \theta(\mathbf{0}) + c\delta_n \,|\, \mathbf{X}_i)$. Then $E[\,Z_i\,] = 0$ and, by the first argument in the proof of Lemma 4,

$$\operatorname{Var}(\textstyle\sum_i K_i Z_i) = \operatorname{Var}(\textstyle\sum_{I_n} Z_i) = O(n\delta_n^d).$$

According to (5.12)

$$\begin{aligned} &\tfrac{1}{2} - N_n^{-1}\textstyle\sum_i K_i P(Y_i \ge \theta(\mathbf{0}) + c\delta_n \,|\, \mathbf{X}_i) \\ &= \tfrac{1}{2} - N_n^{-1}\textstyle\sum_i P(Y_i \ge \theta(\mathbf{0}) + c\delta_n \,|\, B_{ni}) \ge (c - M_0)M_1^{-1}\delta_n \qquad \text{for } c > M_0. \end{aligned}$$

Consequently, by Lemma 3 and Tchebychev's inequality

$$\begin{aligned} P(\hat{\theta}_n(\mathbf{0}) \ge \theta(\mathbf{0}) + c\delta_n) &\le P\left(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i \ge \theta(\mathbf{o}) + c\delta_n\}} \ge \tfrac{1}{2}\right) \\ &\le P(N_n^{-1}\textstyle\sum_{I_n} Z_i \ge \tfrac{1}{2} - N_n^{-1}\textstyle\sum_i P(Y_i \ge \theta(\mathbf{0}) + c\delta_n \,|\, B_{ni})) \\ &\le P(N_n^{-1}\textstyle\sum_{I_n} Z_i \ge (c - M_0)M_1^{-1}\delta_n) \\ &\le P(N_n^{-1}\textstyle\sum_{I_n} Z_i \ge (c - M_0)M_1^{-1}\delta_n); N_n \ge \tfrac{1}{2}n\delta_n^d) + P(N_n < \tfrac{1}{2}n\delta_n^d) \\ &= \frac{O(1)}{(c - M_0)^2}\frac{n\delta_n^d}{(n\delta_n^d\delta_n)^2} + o(1) = o(1) \qquad \text{as } n, c \to \infty, \end{aligned}$$

since δ_n is chosen so that $n\delta_n^d\delta_n^2 = 1$, or equivalently, $\delta_n = n^{-r}$. This completes the proof of Theorem 3.

Proof of Theorem 4. The proof of the theorem depends on the following result.

Lemma 5. *Suppose that Assumptions 1–3, 5(ii) and 6 hold. Then*

$$\lim_{n\to\infty} P(\sup_{\mathbf{x}\in C} |\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| \geq \xi) = 0 \qquad \textit{for } \xi > 0.$$

Proof. Without loss of generality it can be assumed that $C = [-\frac{1}{2}, \frac{1}{2}]^d$. Set $L_n = [n^{2r}]$. Let W_n be the collection of $(2L_n + 1)^d$ points in C each of whose coordinates is of the form $j/(2L_n)$ for some integer j such that $|j| \leq L_n$. Then C can be written as the union of $(2L_n)^d$ subcubes, each having length $2\lambda_n = (2L_n)^{-1}$ and all of its vertices in W_n. For each $\mathbf{x} \in C$ there is a subcube Q_w with center w such that $\mathbf{x} \in Q_w$. Let C_n denote the collection of the centers of these subcubes. Let ξ be a positive constant. Then

$$P\left(\sup_{\mathbf{x}\in C} |\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| \geq \xi\right) = P\left(\max_{C_n} \sup_{\mathbf{x}\in Q_w} |\hat{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})| \geq \xi\right).$$

It follows from $\lambda_n \sim n^{-2r}$ and Assumption 1 that $|\theta(\mathbf{x}) - \theta(w)| \leq M_0\|\mathbf{x} - w\| \leq M_0\delta_n < \xi$ for $\mathbf{x} \in Q_w$, $w \in C_n$ (for n sufficiently large). Therefore, to prove the lemma, it is sufficient to show that

$$\lim_n P\left(\max_{w\in C_n} \sup_{\mathbf{x}\in Q_w} |\hat{\theta}_n(\mathbf{x}) - \theta(w)| \geq \xi\right) = 0 \qquad \text{for } \xi > 0. \tag{5.13}$$

To prove (5.13), let $\eta \equiv \sqrt{d}$, $\mathbf{x} \in Q_w$ and $N_n' \equiv N_n'(w) = \#\{i : \|\mathbf{X}_i - w\| \leq \delta_n - \eta\lambda_n\}$. Now $N_n \equiv N_n(\mathbf{x}) = \#\{i : \|\mathbf{X}_i - \mathbf{x}\| \leq \delta_n\} \geq N_n'$ for $\mathbf{x} \in Q_w$, $\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \geq \xi\} \subseteq \{N_n^{-1} \sum_{I_n} 1_{\{Y_i \geq \theta(w)+\xi\}} \geq \frac{1}{2}\} \subseteq \{\sum_{I_n^*} 1_{\{Y_i \geq \theta(w)+\xi\}} \geq \frac{1}{2}N_n'\}$, where $I_n^* \equiv I_n^*(w) = \{i : 1 \leq i \leq n \text{ and } \|\mathbf{X}_i - w\| \leq \delta_n + \eta\lambda_n\}$. Thus

$$\cup_{Q_w} \left\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \geq \xi\right\} \subseteq \left\{\sum_{I_n^*} 1_{\{Y_i \geq \theta(w)+\xi\}} \geq \tfrac{1}{2}N_n'\right\}. \tag{5.14}$$

Set $N_n^* \equiv N_n^*(w) = \#I_n^*(w)$. By Assumptions 2, 3, 5(ii) and Markov's inequality there are positive constants k_1 and k_2 such that

$$\lim_n P(\Psi_n) = 1, \tag{5.15}$$

where $\Psi_n = \left\{N_n^*(w) - N_n'(w) \leq k_1 n^r \text{ and } N_n^*(w) \geq k_2 n\delta_n^d \text{ for all } w \in C_n\right\}$.

Indeed, note that $N_n^* - N_n' = \#\{i : \delta_n - \eta\lambda_n \le \|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n\}$ with $q_n = P(\delta_n - \eta\lambda_n \le \|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n) \sim ((\delta_n + \eta\lambda_n)^d - (\delta_n - \eta\lambda_n)^d) \sim \delta_n^d(\lambda_n/\delta_n)$ as $n \to \infty$. It follows from $n\delta_n^{d+2} \sim 1$ and $\lambda_n \sim n^{-2r}$ that $nq_n = O(n^r) \to \infty$ as $n \to \infty$. Thus by Lemma A of the Appendix and Markov's inequality

$$\begin{aligned} P\left(N_n^*(w) - N_n'(w) \ge 2nq_n \quad \text{for some } w \in C_n\right) &= [n^{2r}]^d \max_{C_n}(nq_n)^{-2k} E(N_n^* - N_n')^{2k} \\ &= [n^{2r}]^d O(nq_n)^{-k} = [n^{2r}]^d O(n^r)^{-k} \to 0, \end{aligned}$$

for k large enough. Similarly, $\lim_n P(N_n^*(w) \le \frac{1}{2}np_n(w) \quad \text{for some } w) = 0$, where $p_n(w) = P(\|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n) \sim \delta_n^d$. Thus (5.15) is proven.

Note that $n^r N_n^{*-1} \le n^r(k_2 n\delta_n^d)^{-1} \sim \delta_n$ on Ψ_n. According to (5.14) and (5.15), there is a positive constant k_3 such that

$$\begin{aligned} &P\left(\max_{C_n} \sup_{x \in Q_w} [\,\hat{\theta}_n(\mathbf{x}) - \theta(w)\,] \ge \xi\right) \\ &\le P\left(\cup_{C_n} \cup_{Q_w} \left\{\hat{\theta}_n(\mathbf{x}) - \theta(w) \ge \xi\right\}\right) \\ &\le P\left(\cup_{C_n} \left\{\textstyle\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+\xi\}} \ge \frac{1}{2}N_n'\right\}\right) \\ &\le P\left(\cup_{C_n} \left\{\textstyle\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+\xi\}} \ge \frac{1}{2}N_n^* - \frac{1}{2}k_1 n^r\right\} \cap \Psi_n\right) + P(\Psi_n^c) \\ &\le P\left(\cup_{C_n} \left\{N_n^{*-1} \textstyle\sum_{I_n^*} 1_{\{Y_i \ge \theta(w)+\xi\}} \ge \frac{1}{2} - k_3\delta_n\right\}\right) + P(\Psi_n^c). \end{aligned} \tag{5.16}$$

Let $B_{ni} = B_{ni}(w)$ denote the event $\|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n$ for $i = 1, \dots, n$. According to Assumption 1, $\theta(\mathbf{X}_i) \le \theta(w) + M_0(\delta_n + \eta\lambda_n)$ whenever $\|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n$. Thus

$$\tfrac{1}{2} - P(Y_i \ge \theta(w) + \xi \mid B_{ni}) \ge P(M_0(\delta_n + \eta\lambda_n) \le Y_i - \theta(w) \le \xi \mid B_{ni}).$$

Hence by Assumption 6, there is a positive constant M_2 such that

$$\tfrac{1}{2} - P(Y_i \ge \theta(w) + \xi \mid B_{ni}) \ge M_2(\xi \wedge \epsilon_0) \qquad \text{for } n \text{ sufficient large.} \tag{5.17}$$

Set $K_i = 1_{\{\|\mathbf{X}_i - w\| \le \delta_n + \eta\lambda_n\}}$ and $Z_i = 1_{\{Y_i \ge \theta(w)+\xi\}} - P(Y_i \ge \theta(w) + \xi \mid \mathbf{X}_i)$. Then $E(Z_i) = 0$. By (5.17)

$$\tfrac{1}{2} - N_n^{*-1} \textstyle\sum_{I_n^*} P(Y_i \ge \theta(w) + \xi \mid \mathbf{X}_i) \ge M_2(\xi \wedge \epsilon_0). \tag{5.18}$$

It now follows from (5.18) that there is a positive constant M_3 such that

$$\begin{aligned}
&P\left(\cup_{C_n}\left\{N_n^{*-1}\textstyle\sum_{I_n^*}1_{\{Y_i\geq\theta(w)+\xi\}}\geq\frac{1}{2}-k_3\delta_n\right\}\right)\\
&\leq P\left(\cup_{C_n}\left\{N_n^{*-1}\textstyle\sum_{I_n^*}Z_i\geq\frac{1}{2}-N_n^{*-1}\sum_{I_n^*}P(Y_i\geq\theta(w)+\xi\,|\,\mathbf{X}_i)-k_3\delta_n\right\}\right)\\
&\leq[n^{2r}]^d\max_{C_n}P\left(N_n^{*-1}\textstyle\sum_{I_n^*}Z_i\geq M_2(\xi\wedge\epsilon_0)-k_3\delta_n\right)\\
&\leq[n^{2r}]^d\max_{C_n}P\left(N_n^{*-1}\textstyle\sum_{I_n^*}Z_i\geq M_3(\xi\wedge\epsilon_0)\right)\quad\text{for } n \text{ sufficient large.}
\end{aligned}\tag{5.19}$$

Set $p_n = p_n(w) = P(\|\mathbf{X}_i - w\| \leq \delta_n + \eta\lambda_n)$ (by stationary, p_n does not depend on i). Then $p_n \sim \delta_n^d$. Let k be a positive integer. Note that $\sum_{I_n^*} Z_i = \sum_i K_i Z_i$, $E(K_i Z_i) = 0$ and $E|K_i Z_i| = O(\delta_n^d)$. By Lemma A (see the Appendix)

$$\begin{aligned}
E|\textstyle\sum_{I_n^*}Z_i|^{2k} &= E|\textstyle\sum_i K_i Z_i|^{2k} = O(n\delta_n^d)^k,\\
E|N_n^* - np_n|^{2k} &= O(np_n)^k \qquad \text{on } C_n.
\end{aligned}$$

Consequently, by Markov's inequality

$$\begin{aligned}
&P\left(N_n^{*-1}\textstyle\sum_{I_n^*}Z_i\geq M_3(\xi\wedge\epsilon_0)\right)\\
&\leq P\left(N_n^{*-1}\textstyle\sum_{I_n^*}Z_i\geq M_3(\xi\wedge\epsilon_0); N_n^*\geq\frac{1}{2}np_n\right)+P\left(N_n^*<\frac{1}{2}np_n\right)\\
&\leq\frac{E|\sum_i K_iZ_i|^{2k}}{(\frac{1}{2}M_3(\xi\wedge\epsilon_0)np_n)^{2k}}+\frac{E|N_n^*-np_n|^{2k}}{(\frac{1}{2}np_n)^{2k}}=O(n\delta_n^d)^{-k}\text{ for } w\in C_n.
\end{aligned}\tag{5.20}$$

Note that δ_n is chosen so that $n\delta_n^d \sim \delta_n^{-2} \sim n^{2r}$. It follows from (5.20) that there is a positive integer k such that

$$[n^{2r}]^d\max_{C_n}P\left(N_n^{*-1}\textstyle\sum_{I_n^*}Z_i\geq M_3(\xi\wedge\epsilon_0)\right)\leq[n^{2r}]^dO(n^{2r})^{-k}\to 0\text{ as } n\to\infty. \tag{5.21}$$

Hence by (5.15), (5.16), (5.19) and (5.21)

$$\lim_n P\left(\max_{C_n}\sup_{\mathbf{x}\in Q_w}[\hat{\theta}_n(\mathbf{x})-\theta(w)]\geq\xi\right)=0\qquad\text{for } \xi>0. \tag{5.22}$$

Similarly,

$$\lim_n P\left(\max_{C_n}\sup_{\mathbf{x}\in Q_w}[\hat{\theta}_n(\mathbf{x})-\theta(w)]\geq-\xi\right)=0\qquad\text{for } \xi>0. \tag{5.23}$$

It follows from (5.22) and (5.23) that (5.13) is valid. This completes the proof of the lemma.

The proof of Theorem 4 will now be given. By Assumption 1, $\theta(\cdot)$ is bounded on C (compact). Thus it follows from Lemma 5 that there is a positive constant $T \geq 1$ such that

$$\lim_n P(\Phi_n) = 1 \tag{5.24}$$

where $\Phi_n \equiv \{\|\hat{\theta}_n\|_\infty \leq T\}$. For $i = 1, \ldots, n$, set

$$Y_i' = \begin{cases} -T & \text{if } Y_i \leq -T; \\ Y_i & \text{if } |Y_i| \leq T; \\ T & \text{if } Y_i \geq T. \end{cases}$$

Put $\bar{\theta}_n(\mathbf{x}) \equiv \text{Median}\{Y_i' : i \in I_n(\mathbf{x})\}$. Note that $\bar{\theta}_n(\mathbf{x}) = \hat{\theta}_n(\mathbf{x})$ except on Φ_n^c for $\mathbf{x} \in C$. Thus by (5.24), in order to prove the theorem, it is sufficient to show

$$\lim_n P\left(\|\bar{\theta}_n - \theta\|_q \geq cn^{-r}\right) = 0. \tag{5.25}$$

To prove (5.25), we may assume that C is contained in the interior of the cube $C_0 = [-\frac{1}{2}, \frac{1}{2}]^d \subset U$. By (5.15) or Lemma A, there is a positive constant k_4 such that

$$\lim_n P(\Omega_n) = 1, \tag{5.26}$$

where $\Omega_n = \{N_n(\mathbf{x}) \geq k_4 n \delta_n^d \text{ for } \mathbf{x} \in C\}$.

Write $P_{\Omega_n}(\cdot) = P(\cdot\,; \Omega_n) = P(\cdot \cap \Omega_n)$ and $E_{\Omega_n}(W) = \int w\, dP_{\Omega_n}$, where W is a real-valued random variable. By (5.26), there is a sequence of positive numbers $\epsilon_n \to 0$ such that

$$P\left(\int_C |\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q \, d\mathbf{x} \geq (cn^{-r})^q\right) \leq P\left(\int_C |\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q \, d\mathbf{x} \geq (cn^{-r})^q; \Omega_n\right) + \epsilon_n$$
$$\leq \frac{E_{\Omega_n}\left[\int_C |\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|^q \, d\mathbf{x}\right]}{(cn^{-r})^q} + \epsilon_n. \tag{5.27}$$

Set $Z_n(\mathbf{x}) = |\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x})|$. By Assumption 1, $Z_n(\mathbf{x})$ is bounded by T for $\mathbf{x} \in C$. Thus there is a positive constant k_5 such that

$$
\begin{aligned}
E_{\Omega_n}[Z_n^q(\mathbf{x})] &= \int_0^T qt^{q-1}P_{\Omega_n}\left(Z_n(\mathbf{x}) > t\right)\,dt \\
&= \int_0^{2M_0\delta_n} qt^{q-1}P_{\Omega_n}\left(Z_n(\mathbf{x}) > t\right)\,dt + \int_{2M_0\delta_n}^T qt^{q-1}P_{\Omega_n}\left(Z_n(\mathbf{x}) > t\right)\,dt \\
&\le k_5\delta_n^q + \int_{2M_0\delta_n}^T qt^{q-1}P_{\Omega_n}\left(Z_n(\mathbf{x}) > t\right)\,dt. \qquad (5.28)
\end{aligned}
$$

By Assumptions1–3, 5(ii) and 6, there is a positive number k_6 such that

$$\int_{2M_0\delta_n}^T qt^{q-1}P_{\Omega_n}\left(Z_n(\mathbf{x}) > t\right)\,dt \le k_6\delta_n^q \qquad \text{for } \mathbf{x} \in C. \qquad (5.29)$$

(The proof of (5.29) will be given shortly.) It follows from (5.28) and (5.29) that there is a positive constant k_7 such that

$$E_{\Omega_n}[Z_n^q(\mathbf{x})] \le k_7\delta_n^q.$$

Thus there is a positive constant k_8 such that

$$E_{\Omega_n}\left[\int_C Z_n^q(\mathbf{x})\,d\mathbf{x}\right] = \int_C E_{\Omega_n}[Z_n^q(\mathbf{x})]\,d\mathbf{x} \le k_7\delta_n^q. \qquad (5.30)$$

The conclusion of Theorem 4 follows from (5.27) and (5.30).

Finally, (5.29) will be proven. Set $D_i = 1_{\{Y_i \ge \theta(\mathbf{x})+t\}} - P(Y_i \ge \theta(\mathbf{x}) + t \mid \mathbf{X}_i)$ and let B_{ni} be the event $\|\mathbf{X}_i - \mathbf{x}\| \le \delta_n$ for $i = 1, \ldots, n$. Put $R_i = \frac{1}{2} - P(Y_i \ge \theta(\mathbf{x}) + t \mid B_{ni})$. By Assumption 6, there is a positive constant k_9 such that

$$N_n^{-1}\textstyle\sum_{I_n} R_i \ge k_9 T^{-1}(t - M_0\delta_n) \qquad \text{for } M_0 \le t \le T,\ T > 1.$$

Thus (since $\{Y_i' > \theta(\mathbf{x}) + t\} \subset \{Y_i > \theta(\mathbf{x}) + t\}$)

$$
\begin{aligned}
P_{\Omega_n}\left(\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x}) > t\right) &\le P_{\Omega_n}\left(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i' > \theta(\mathbf{x})+t\}} \ge \frac{1}{2}\right) \\
&\le P_{\Omega_n}\left(N_n^{-1}\textstyle\sum_{I_n} 1_{\{Y_i > \theta(\mathbf{x})+t\}} \ge \frac{1}{2}\right) \\
&\le P_{\Omega_n}\left(N_n^{-1}\textstyle\sum_{I_n} D_i \ge N_n^{-1}\sum_{I_n} R_i\right) \\
&\le P\left(\textstyle\sum_{I_n} D_i \ge k_9 T^{-1}(t - M_0\delta_n) n\delta_n^d\right). \qquad (5.31)
\end{aligned}
$$

Note that $\sum_{I_n} D_i = \sum_i K_i D_i$, $E(K_i D_i) = 0$ and $E|K_i D_i| = O(\delta_n^d)$. By Lemma A,

$$E|\sum_{I_n} D_i|^{2k} \le E|\sum_i K_i D_i|^{2k} = O(n\delta_n^d)^k.$$

Consequently, by Markov's inequality

$$\begin{aligned} P\left(\sum_{I_n} D_i \ge k_9 T^{-1}(t - M_0\delta_n)n\delta_n^d\right) &\le \frac{E\left|\sum_{I_n} D_i\right|^{2k}}{(k_9 T^{-1}(t - M_0\delta_n)n\delta_n^d)^{2k}} \\ &\le \frac{O(n\delta_n^d)^k}{(k_9 T^{-1}(t - M_0\delta_n)n\delta_n^d)^{2k}}. \end{aligned} \tag{5.32}$$

By (5.31) and (5.32), there is a positive constant k_{10} such that (note that $n\delta_n^d \sim \delta_n^{-2}$)

$$P_{\Omega_n}\left(\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x}) > t\right) \le k_{10}[T^{-1}(t - M_0\delta_n)]^{-2k}\delta_n^{2k}, \tag{5.33}$$

and, similarly,

$$P_{\Omega_n}\left(\bar{\theta}_n(\mathbf{x}) - \theta(\mathbf{x}) < -t\right) \le k_{10}[T^{-1}(t - M_0\delta_n)]^{-2k}\delta_n^{2k}, \tag{5.34}$$

It now follows from (5.33) and (5.34) that there is a positive constant k_{11} such that (make a change of variable $t = M_0\delta_n(s+1)$)

$$\begin{aligned} \int_{2M_0\delta_n}^{T} t^{q-1} P_{\Omega_n}\left(Z_n(\mathbf{x}) > t\right) dt &\le k_{10} \int_{2M_0\delta_n}^{T} t^{q-1}[T^{-1}(t - M_0\delta_n)]^{-2k}\delta_n^{2k}\, dt \\ &\le k_{11}\delta_n^q \int_1^\infty (s+1)^{q-1}s^{-2k}\, ds \\ &= O(\delta_n^q) \qquad \text{for } k > q, \end{aligned}$$

as desired. This completes the proof of (5.29).

APPENDIX

Proof of (5.10). Given positive numbers a_n and b_n, $n \ge 1$, let $a_n \simeq b_n$ mean that $a_n/b_n \to 1$ as $n \to \infty$. Write $[-\frac{1}{2}, \frac{1}{2}]^d$ as the disjoint union of M_n^d cubes C_{nv} of length $\simeq \delta_n/d$ where $M_n \simeq d\delta_n^{-1}$ and $v = 1, 2, \ldots, M_n^d$. Set $K_{iv} = 1_{\{\mathbf{X}_i \in C_{nv}\}}$, $\mu = \mu_v = E[K_{iv}] \sim \delta_n^d$ and $N_{nv} = \#\{i : 1 \le i \le n \text{ and } \mathbf{X}_i \in C_{nv}\} = \sum_i K_{iv}$. To prove (5.10), it suffices to show that

$$\lim_n P\left(N_{nv} \ge \tfrac{1}{2}n\mu \quad \text{for } v = 1, 2, \ldots, M_n^d\right) = 1. \tag{A.1}$$

Under the assumption of independence, there are several known results that can be used to prove this: Vapnik and Cervonenkis' inequality (see Theorem 12.2 of Breiman et al., 1984); Bernstein's inequality (see Theorem 3 of Hoeffding, 1963); Markov's inequality applied to sufficient high order moments; and Lemma 1 of Stone (1982). Collomb (1982) obtained a Bernstein type inequality for dependent random variables satisfying a ϕ-mixing condition, which is stronger than α-mixing and is too restrictive for many applications. In particular, this ϕ-mixing condition is equivalent to m-dependence for stationary Gaussian time series.

In what follows, we will prove (A.1) by calculating sufficiently high order (centered) moments of N_{nv} under Assumption 2 and Assumption 5 (ii). Let $\{\nu_n\}$ be a sequence of positive numbers such that $\nu_n \sim n^{-\gamma}$ for some $\gamma \in (0,1)$.

Lemma A. *Let $V_{n1}, \ldots, V_{nn}$ be uniformly bounded random variables such that V_{ni} has mean zero and is a function of $\mathbf{X}_i$. Suppose that $E|V_{ni}| \leq \nu_n$ and $E|V_{ni}V_{nj}| \leq \nu_n^2$ for $i, j = 1, \ldots, n$. Let k be a positive integer. Then*

$$E(\textstyle\sum_i V_{ni})^k = O(n\nu_n)^{\frac{k}{2}} \qquad \text{as } n \to \infty .$$

Proof. In the following discussion, write V_i for V_{ni}. Observe that

$$E(\textstyle\sum_i V_i)^k \leq k! \sum\sum \left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right|, \tag{A.2}$$

where the indices in the first sum on the right side of (A.2) are on values of t, $\tau_1, \ldots, \tau_t$ constrained by $\tau_1, \ldots, \tau_t > 0$ and $\tau_1 + \cdots + \tau_t = k$ for $t = 1, \ldots, k$ and, the indices in the second sum are on values of $i_1, \ldots, i_t$ constrained by $i_1, \ldots, i_t > 0$ and $i_1 + \cdots + i_t < n$. Let N be a positive integer less than n. Partition the second sum in (A.2) into a finite number of sums such that the indices in each of these sums are constrained by: certain of the indices are larger than N and all others are less than or equal to N. More precisely, let $\psi_t = (\phi_1, \ldots, \phi_t)$ be a t-tuple of 0's and 1's and let $\sum_{\psi_t} |E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_t}^{\tau_t})|$ mean that (a) if $\phi_l = 1$, then the index i_l in the sum ranges over $N+1, \ldots, n$; (b) if $\phi_l = 0$, then the index i_l in the sum ranges over $1, \ldots, N$. Thus

$$\left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| = \textstyle\sum_{all\,\psi_t} \sum_{\psi_t} \left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| . \tag{A.3}$$

Let ψ_t be fixed. By induction on m, where $m = \tau_1 + \cdots + \tau_t$,

$$\textstyle\sum_{\psi_t} \left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| = O(n\nu_n)^{\frac{m}{2}}. \tag{A.4}$$

Indeed, (A.4) is valid for $m = 1, 2$. ($\sum_{i,j} |E(V_i V_j)| = O(n \sum_i \min(\alpha(i), \nu_n^2)) = O(n\nu_n)$.) Suppose $m > 2$ and assume that (A.4) holds for $\tau_1, \ldots, \tau_t$ with $\tau_1 + \cdots + \tau_t \leq m - 1$. Set $N = [m\gamma^{-1}(\gamma + 1) \log \nu_n / (2 \log \rho)]$. If $\phi_j = 0$ for $1 \leq j \leq t$ or $\phi_1 = 1$, then by the assumption that variables V_i's are bounded by 1 (say) and $m > 2$,

$$\begin{aligned}\textstyle\sum_{\psi_t} \left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| &\leq N^{t-1} n\nu_n \\ &\leq (\log n)^t n\nu_n = o(n\nu_n)^{\frac{m}{2}-1} n\nu_n = o(n\nu_n)^{\frac{m}{2}}.\end{aligned}$$

So suppose $\phi_j = 1$ for some $2 \leq j \leq t$ and set $b = \min\{j : 2 \leq j \leq t, \phi_j = 1\}$. Since the V_i's are bounded by 1, it follows from Theorem A.5 of Hall and Heyde (1980, p.277) that

$$\begin{aligned}&\left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_{b-1}}^{\tau_{b-1}} V_{i_1+\cdots+i_b}^{\tau_b} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| \\ &\leq \left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_{b-1}}^{\tau_{b-1}})\right| \left|E(V_{i_1+\cdots+i_b}^{\tau_b} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| + 4\alpha(i_b).\end{aligned}$$

Consequently, by the inductive hypothesis,

$$\begin{aligned}&\textstyle\sum_{\psi_t} \left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_{b-1}}^{\tau_{b-1}} V_{i_1+\cdots+i_b}^{\tau_b} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| \\ &\leq \textstyle\sum_{\psi_t} \left|E(V_{i_1}^{\tau_1} \cdots V_{i_1+\cdots+i_{b-1}}^{\tau_{b-1}})\right| \left|E(V_{i_1+\cdots+i_b}^{\tau_b} \cdots V_{i_1+\cdots+i_t}^{\tau_t})\right| + 4\textstyle\sum_{\psi_t} \alpha(i_b) \\ &\leq O(n\nu_n)^{(\tau_1+\cdots+\tau_{b-1})/2} O(n\nu_n)^{(\tau_b+\cdots+\tau_t)/2} + 4n^{t-1} \textstyle\sum_{i>N} \alpha(i) \\ &= O(n\nu_n)^{\frac{m}{2}},\end{aligned}$$

for it follows from $N = [m\gamma^{-1}(\gamma + 1) \log \nu / (2 \log \rho)]$ and Condition 3.5 (ii) that (with $t \leq m$) $n^t \sum_{i>N} \alpha(i) \leq n^m \sum_{i>N} \alpha(i) \sim n^m \nu_n^{m(\gamma+1)/2\gamma} \sim (n\nu_n)^{m/2}$. This completes the proof of (A.4).

The conclusion of the lemma follows from (A.2)–(A.4).

(A.1) will now be proven. Set $V_i \equiv V_{iv} = K_{iv} - \mu$. Then $E|V_i| \leq \mu \sim \delta_n^d \sim n^{-d/(2+d)}$. By Markov's inequality and Lemma A,

$$\begin{aligned}P(N_{nv} \leq \tfrac{1}{2} n\mu) &= P(\textstyle\sum_i V_i \leq -\tfrac{1}{2} n\mu) \\ &\leq (\tfrac{1}{2} n\mu)^{-2k} E(\textstyle\sum_i V_i)^{2k} = O(n\mu)^{-k}.\end{aligned}$$

Thus there is a positive integer k (large enough) such that

$$\begin{aligned} P\left(N_{nv} \geq \tfrac{1}{2}n\mu \ \text{ for } v = 1, \ldots, M_n^d\right) &\geq 1 - M_n^d P\left(N_{nv} \leq \tfrac{1}{2}n\mu\right) \\ &\geq 1 - M_n^d O(n\mu)^{-k} \\ &= 1 - O(\delta_n^{-d}\delta_n^{2k}) \to 1 \qquad \text{as } n \to \infty, \end{aligned}$$

since $M_n^d \sim \delta_n^{-d}$ and $n\mu \sim \delta_n^{-2}$. This completes the proof of (A.1).

REFERENCES

AKAIKE, H. (1974a) Markov representation of stochastic processes and its application to the analysis of the autoregressive moving-average process. *Ann. Inst. Statist. Math.* **26**, 263–387.

AKAIKE, H. (1974b) Stochastic theory of minimal realisation. *I.E.E.E. Trans. Automatic Control AC*-**19**, 263–387.

BIERENS, H. J. (1983) Uniform consistency of kernel estimators of a regression function under generalized conditions. *J. Amer. Statist. Assoc.* **78**, 699–707.

BLOOMFIELD, P. and STEIGER, (1983) *Least Absolute Deviations.* Birkhauser, Boston.

BREIMAN, L., FRIEDMAN, J. H., OLSHEN, R. A. and STONE, C. J. (1984) *Classification and Regression Trees.* Wadsworth, Belmont.

BRILLINGER, D. R. (1980) *Time series: Data analysis and theory.* Holden-Day, San Francisco.

COLLOMB, B. G. (1982) Prediction non-parametrique: Etude de L'erreur quadratique du predictogramme. *C. R. de l'Academie des Sciences, Paris.* **294**, 59–62.

COLLOMB, B. G. and HÄRDLE, W. (1984) Strong uniform convergence rates in robust nonparametric time series analysis and prediction: kernel regression estimation from dependent observations. *Preprint.*

DOUKHAN, P. and GHINDES, M. (1980) Estimations dans le processus $X_{n+1} = f(X_n) + \epsilon_n$. *C. R. de l'Academie des Sciences, Paris.* **291**, 61–64.

DUNSMUIR, W. and HANNAN, E. J. (1976) Vector linear time series models. *Adv. Appl. Prob.* **8**, 339–364.

HAGGAN, V. and OZAKI, T. (1980) Amplitude-dependent exponential AR model fitting for non-linear random vibrations. In *Time Series*, edited by O. D. Anderson. North-Holland: Amsterdam.

HAGGAN, V. and OZAKI, T. (1981) Modelling non-linear random vibrations using an amplitude-dependent autoregressive model. *Biometrika* **68**, 189–196.

HALL, P. and HEYDE, C. C. (1980) *Martingale limit theory and its applications.* Academic Press, New York.

HANNAN, E. J. (1973) The asymptotic theory of linear time series models. *J. Appl. Prob.* **10**, 130–145.

HÄRDLE, W. and LUCKHAUS, S. (1984) Uniform consistency of a class of regression function estimators. *Ann. Statist.* **12**, 612–623.

HOEFFDING, W. (1963) Probability inequalities for sums of bounded random variables. *J. Amer. Statist. Assoc.* **58**, 13–30.

MORRIS, J. (1977) Forcasting the sunspot cycles (with discussion). *J. R. Statist. Soc. A* **140**, 437–468.

NICHOLLS, D. F. and QUINN, B. G. (1980) The estimation of random coefficient autoregressive models. I. *J. Time Series Anal.* **1**, 37–46.

PRIESTLEY, M. B. (1979) *Time series and spectral analysis.* Academic Press, New York.

PRIESTLEY, M. B. (1980) State-dependent models: A general approach to non-linear time series analysis. *J. Time Series Anal.* **1**, 47–71.

ROBINSON, P. M. (1983) Nonparametric estimators for time series. *J. Time Series Anal.* **4**, 185–207.

ROSENBLATT, M. (1956) A central limit theorem and a strong mixing condition. *Proc. Nat. Acad. Sci., USA* **42**, 43–47.

STONE, C. J. (1980) Optimal rates of convergence for nonparametric estimators. *Ann. Statist.* **8**, 1348–1360.

STONE, C. J. (1982) Optimal global rates of convergence for nonparametric regression. *Ann. Statist.* **10**, 1040–1053.

STONE, C. J. (1985) Additive regression and other nonparametric models. *Ann. Statist.* **13**, 689–705.

TONG, H. and LIM, K. S. (1980) Threshold Autoregression, limit cycles and cyclical data. *J. Roy. Statist. Soc. B* **42**, 245–292.