Strongly-Consistent Nonparametric Forecasting and Regression for Stationary Ergodic Sequences

Sidney Yakowitz, László Györfi, John Kieffer and Gusztáv Morvai

J. Multivariate Anal. 71 (1999), no. 1, 24–41.

Abstract

Let $\{(X_i, Y_i)\}$ be a stationary ergodic time series with (X, Y) values in the product space $R^d \bigotimes R$. This study offers what is believed to be the first strongly consistent (with respect to pointwise, least-squares, and uniform distance) algorithm for inferring $m(x) = E[Y_0|X_0 = x]$ under the presumption that $m(x)$ is uniformly Lipschitz continuous. Autoregression, or forecasting, is an important special case, and as such our work extends the literature of nonparametric, nonlinear forecasting by circumventing customary mixing assumptions. The work is motivated by a time series model in stochastic finance and by perspectives of its contribution to the issues of universal time series estimation.

1 Introduction

Nonparametric regression has been applied to a variety of contexts, in particular to time series modeling and prediction. The present study contributes to the methodology by showing how a regression function can be consistently inferred from time series data under no process assumptions beyond stationarity and ergodicity. (A Lipschitz condition on the regression function itself will be imposed.)

Toward showing how our methodology can impinge on an established research area, we give one substantive application to a practical problem in stochastic finance: Many works, such as the Chapter entitled "Some Recent Developments in Investment Research" of the prominent text [5], argue for the need to move beyond the Black-Scholes stochastic differential equation. This and other studies suggest the so-called ARCH and GARCH extensions as a promising direction. The review of this approach by Bollerslev *et al.* [6] cites a litany of unresolved issues. Of particular relevance is the discussion of the need to account for persistency of the variance (Sections 2.6 and 3.6). (ARCH and GARCH models can be long-range dependent for certain ranges of parameters. In these cases, statistical analysis is delicate [8].)

The basic idea behind the ARCH/GARCH setup is that one must allow the asset volatility (variance) to change dynamically, and perhaps (GARCH) to depend on current and past volatility values. The review [6] documents (p. 30) that several authors have applied nonparametric and semiparametric regression, with some success, to infer the ARCH functions from data. These methods can fail if fairly stringent mixing conditions are not in force. Masry and Tjostheim [21], because of their rigorous consideration of consistency, sets the stage for appreciating the potential of the present investigation. They propose that both the asset dynamics and volatility of a nonlinear ARCH series

be inferred from nonparametric classes of regression functions. By imposing some fairly severe assumptions, which would be tricky to validate from data, these authors are able to assure that the ARCH process is strongly mixing (with exponentially decreasing parameter) and consequently standard kernel techniques are applicable.

On another avenue toward asset series modelling, decades ago, Mandelbrot suggested that fractal processes should be considered in this context. Fractals have been of interest to theorists and modellers alike in part because they can display persistency. In his 1999 study, "A Multifractal Walk down Wall Street," [20] Mandelbrot argues that conventional models for portfolio theory ignore soaring volatility, and that is akin to a mariner ignoring the possibility of a typhoon on the basis of the observation that weather is moderate 95% of the time.

Such persistence as exhibited in the models of finance calls into question whether various processes of interest are actually strongly mixing, a consistency requirement for conventional nonparametric regression techniques. We mention parenthetically that telecommunications modelers are increasingly turning toward long-range-dependent processes (e.g., [28] and [37])

As mentioned, the primary contribution of the present paper is an algorithm which is demonstratably consistent without imposition of mixing assumptions. The implication is that process assumptions such as in [21] are not required for our algorithm. The price paid for this flexibility is that convergence rates and asymptotic normality cannot be assured. This avenue is worthy of exploration, nevertheless, because the limits of process inference are clarified, and as a practical matter, future work might lead to methods which are reasonably efficient if the process does satisfy mixing assumptions, but simultaneously assures convergence when mixing fails.

The algorithm is of the series-expansion type. The foundational idea (after Kieffer

[17]) is that sometimes it is possible to bound the error of ignoring the series tail, and additionally assure that the leading coefficients are consistently estimated. Specific constructs are given for a partition-type estimator (Section 2) and for a kernel series (Section 3).

We close this introduction with a survey of the literature of nonparametric estimation for stationary series without mixing hypotheses.

Let Y be a real-valued random variable and let X be a d-dimensional random vector (*i.e.,* the observation or co-variate). We do not assume anything about the distribution of X. As is customary in regression and forecasting, the main aim of the analysis here is to minimize the mean-squared error :

$$\min_f E((f(X) - Y)^2)$$

over some space of real-valued functions $f(\cdot)$ defined on the range of X. This minimum is achieved by the regression function $m(x)$, which is defined to be the conditional distribution of Y given X:

$$m(x) = E(Y \mid X = x), \tag{1}$$

assuming the expectation is well-defined, i.e., if $E|Y| < \infty$. For each measurable function f one has

$$\begin{aligned} E((f(X) - Y)^2) &= E((m(X) - Y)^2) + E((m(X) - f(X))^2) \\ &= E((m(X) - Y)^2) + \int (m(x) - f(x))^2 \mu(dx), \end{aligned}$$

where μ stands for the distribution of the observation X. The second term on the right hand side is called *excess error* or *integrated squared error* for the function f, which is given the notation

$$J(f) = \int (m(x) - f(x))^2 \mu(dx). \tag{2}$$

Clearly, the mean squared error for f is close to that of the optimal regression function only if the excess error $J(f)$ is close to 0.

With respect to the statistical problem of regression estimation, let $(X_1, Y_1), \ldots, (X_n, Y_n) \ldots$ be a stationary ergodic time series with marginal component denoted as (X, Y). We study pointwise, $L_2(\mu)$, and L_∞ convergence of the regression estimate m_n to m. The estimator m_n is called *weakly universally consistent* if $J(m_n) \to 0$ in probability for all distributions of (X, Y) with $E|Y|^2 < \infty$. In the context of independent identically-distributed (i.i.d.) pairs (X, Y), Stone [35] first pointed out in 1977 that there exist weakly universally consistent estimators. Similarly, m_n is called *strongly universally consistent* if $J(m_n) \to 0$ a.s. for all distributions of (X, Y) with $E|Y|^2 < \infty$.

Following pioneering papers by Roussas [31] and Rosenblatt [30], a large body of literature has accumulated on consistency and asymptotic normality when the samples are correlated. In developments below, we will employ the notation,

$$x_m^n = (x_m, \ldots, x_n),$$

presuming that $m \leq n$.

The theory of nonparametric regression is of significance in time series analysis because, by considering samples $\{(X_{n-q}^n, X_{n+1})\}$ in place of the pairs $\{(X_{n-q}^n, Y_n)\}$, the regression problem is transformed into the *forecasting* (or *auto-regression*) problem. Thus, in forecasting, we are asking for the conditional expectation of the next observation, given the $q-$past, with q a positive integer, or perhaps infinity.

As mentioned, nearly all the works on consistent statistical methods for forecasting hypothesize *mixing conditions*, which are assumptions about how quickly dependency attenuates as a function of time separation of the observables. Under a variety of mixing assumptions, kernel and partitioning estimators are consistent, and have attractive rate

properties. The monograph by Györfi *et. al.* [14] gives a coverage of the literature of nonparametric inference for dependent series. In that work, the partition estimator is shown to be strongly consistent, provided $|Y|$ is a.s. bounded, under $\phi-$ mixing and, with some provisos, under $\alpha-$ mixing. A drawback to much of the literature on nonparametric forecasting is that mixing conditions are unverifiable by available statistical procedures. Consequently, some investigators have examined the problem:

> Let $\{X_i\}$ be a real vector-valued stationary ergodic sequence. Find a forecasting algorithm which is provably consistent in some sense.

Of course, some additional hypotheses regarding smoothness of the auto-regression function and moment properties of the variables will be allowed, but additional assumptions about attenuation of dependency are ruled out. A *forecasting algorithm* for

$$m(X_{-p}^{-1}) = E[X_0|X_{-p}^{-1}]$$

here means a rule giving a sequence $\{m_n\}$ of numbers such that for each n, m_n is a measureable function determined entirely by the data segment X_{-n}^{-1}.

For X binary, Ornstein [27] provided a (complicated) strongly-consistent estimator of

$E[X_0|X_{-\infty}^{-1}]$. Algoet [1] extended this approach to achieve convergence over real-valued time series and in this and [2], connected the universal forecasting problem with fundamental issues in portfolio and gambling analysis as well as data compression. Morvai *et al.* [22] offered another algorithm achieving strong consistency in the above sense. Their

algorithm is easy to describe and analyze, and such analysis shows, unfortunately, that its data requirements make it infeasible [23].

On the negative side, Bailey [4] and Ryabko [32] have proven that even over binary processes, there is no strongly consistent estimator for the dynamic problem of inferring $E[X_{n+1}|X_0^n]$, $n = 0, 1, 2, \ldots$.

We mention that for a real vector-valued Markov series with a stationary transition law, a strongly-consistent estimator is available for inferring $m(x) = E[X_0|X_{-1} = x]$ under the hypothesis that the sequence is Harris recurrent [38]. Admittedly this is a dependency condition, but the marginal (i.e., invariant) law need not exist: Positive recurrence is not hypothesized. It is difficult to imagine a Markov condition weaker than Harris recurrence under which statistical inference is assured.

It is to be noted that there are weakly-consistent estimators for the moving regression problem $E[X_{n+1}|X_0^n]$, $n = 0, 1, 2, \ldots$. It turns out that universal coding algorithms (e.g. [39]) of the information theory literature can be converted to weakly-universally consistent algorithms when the coordinate space is finite. Morvai *et al* [25] have given a weakly-consistent (and potentially computationally feasible) regression estimator for the moving regression problem when X takes values from the set of real numbers. That work offers a synopsis of the literature of weakly consistent estimation for stationary and ergodic time series. All the studies we have cited on consistency without mixing assumptions rely on algorithms which do not fall into any of the traditional classes (partitioning, kernel, nearest neighbor) mentioned in connection with i.i.d. regression.

From this point on, $\{(X_i, Y_i)\}$ will represent a time series with (X, Y) values in $R^d \bigotimes R$ which is stationary and ergodic, and such that $E|Y_i| < \infty$. In Section 2, we establish by means of a variation on the partitioning method, that we have a.s. convergence pointwise, and, in the case of bounded support, in uniform distance, provided that the regression

function $m(x) = E[Y_0|X_0 = x]$ satisfies a Lipschitz condition and a bound on the Lipschitz constant is known in advance. If furthermore $|Y|$ is known to be bounded (but perhaps the bound itself is not known), then our algorithm converges in $L_2(\mu)$. Section 3 provides analogous results for a truncated kernel-type estimate. In summary, we miss our goal of pointwise strong universal consistency only in that we must restrict attention to regression functions satisfying a uniform Lipschitz condition and the user must have a bound to the Lipschitz constant. ¿From counter-examples in Györfi *et al.* [16] one sees that some restrictions are needed.

Recently we have obtained an important preprint by Nobel *et al.* [26] which bears similarities with the present investigation. That study gives an algorithm for the long-standing problem of density estimation of the marginal of a stationary sequence. Somewhat analoguous to our conditions, Nobel *at al.* require that the density function be of bounded variation. The algorithm itself is based on different principles from the present paper. In the paper [24] by G. Morvai, S. Kulkarni, and A. Nobel, the ideas in [26] were extended for regression estimation.

2 Truncated partitioning estimation

Let $(X_i, Y_i)_{i=1}^{\infty}$ be an ergodic stationary random sequence with $E|Y| < \infty$. Now we attack the problem of estimating the regression function $m(x)$ by combining partitioning estimation with a series expansion.

Let $\mathcal{P}_k = \{A_{k,i}$ i=1,$\dots\}$ be a nested cubic partition of R^d with volume $(2^{-k-2})^d$. Define $A_k(x)$ to be the partition cell of $\mathcal{P}_k$ into which x falls. Take

$$M_k(x) := E(Y|X \in A_k(x)). \tag{3}$$

One can show that

$$M_k(x) \to m(x) \tag{4}$$

for μ-almost all $x \in R^d$. (To see this, notice that $\{M_k(X), \sigma(A_k(X))\ k = 1, 2, \ldots\}$ is a martingale, $E|Y| < \infty$ implies $\sup_{k=1,2,\ldots} E|M_k(X)| < \infty$ and hence the martingale convergence theorem can be applied to achieve the desired result (4), cf. Ash [3] pp. 292.)

For $k \geq 2$ let

$$\Delta_k(x) = M_k(x) - M_{k-1}(x). \tag{5}$$

Our analysis is motivated by the representation,

$$m(x) = M_1(x) + \sum_{k=2}^{\infty} \Delta_k(x) = \lim_{k \to \infty} M_k(x) \tag{6}$$

for μ-almost all $x \in R^d$. Now let $L > 0$ be an arbitrary positive number. For integer $k \geq 2$ define

$$\Delta_{k,L}(x) = \text{sign}(M_k(x) - M_{k-1}(x)) \min(|M_k(x) - M_{k-1}(x)|, L2^{-k}). \tag{7}$$

Define

$$m_L(x) := M_1(x) + \sum_{i=2}^{\infty} \Delta_{i,L}(x). \tag{8}$$

Notice that $|\Delta_{i,L}(x)| \leq L2^{-i}$, and hence $m_L(x)$ is well defined for all $x \in S$, where S stands for the support of μ defined as

$$S := \{x \in R^d :\ \mu(A_k(x)) > 0 \text{ for all } k \geq 1.\} \tag{9}$$

By Cover and Hart [7], $\mu(S) = 1$.

The crux of the truncated partitioning estimate is inference of the terms $M_1(x)$ and $\Delta_{i,L}(x)$ for $i = 2, 3, \ldots$ in (8). Define

$$\hat{M}_{k,n}(x) := \frac{\sum_{j=1}^n Y_j 1_{\{X_j \in A_k(x)\}}}{\sum_{j=1}^n 1_{\{X_j \in A_k(x)\}}}. \tag{10}$$

If $\sum_{j=1}^n 1_{\{X_j \in A_k(x)\}} = 0$, then take $\hat{M}_{k,n}(x) = 0$. Now for $k \geq 2$, define

$$\hat{\Delta}_{k,n,L}(x) = \text{sign}(\hat{M}_{k,n}(x) - \hat{M}_{k-1,n}(x)) \min(|\hat{M}_{k,n}(x) - \hat{M}_{k-1,n}(x)|, L2^{-k}) \tag{11}$$

and for N_n a non-decreasing unbounded sequence of positive integers, define the estimator

$$\hat{m}_{n,L}(x) = \hat{M}_{1,n}(x) + \sum_{k=2}^{N_n} \hat{\Delta}_{k,n,L}(x). \tag{12}$$

Theorem 1 *Let $\{(X_i, Y_i)\}$ be a stationary ergodic time series with $E|Y_i| < \infty$. Assume $N_n \to \infty$. Then almost surely, for all $x \in S$*

$$\hat{m}_{n,L}(x) \to m_L(x). \tag{13}$$

If the support S of μ is a bounded subset of R^d then almost surely

$$\sup_{x \in S} |\hat{m}_{n,L}(x) - m_L(x)| \to 0. \tag{14}$$

If either (i) $|Y| \leq D < \infty$ almost surely (D need not be known) or (ii) μ is of bounded support then

$$\int (\hat{m}_{n,L}(x) - m_L(x))^2 \mu(dx) \to 0. \tag{15}$$

Proof First we prove that almost surely, for all $x \in S$, and for all $k \geq 1$,

$$\lim_{n \to \infty} |\hat{M}_{k,n}(x) - M_k(x)| = 0. \tag{16}$$

By the ergodic theorem, as $n \to \infty$, a.s.,

$$\frac{\sum_{j=1}^n 1_{\{X_j \in A_{k,i}\}}}{n} \to P(X \in A_{k,i}) = \mu(A_{k,i}).$$

Similarly,

$$\frac{\sum_{j=1}^{n} 1_{\{X_j \in A_{k,i}\}} Y_j}{n} \to E(Y 1_{\{X \in A_{k,i}\}}) = \int_{A_{k,i}} m(z)\mu(dz),$$

which is finite since $E|Y|$ is finite. Since there are countably many $A_{k,i}$, almost surely, for all $A_{k,i} \in \cup_v \mathcal{P}_v$ for which $\mu(A_{k,i}) > 0$:

$$\frac{\sum_{j=1}^{n} 1_{\{X_j \in A_{k,i}\}} Y_j}{\sum_{j=1}^{n} 1_{\{X_j \in A_{k,i}\}}} \to E(Y|X \in A_{k,i}).$$

Since for each $x \in S$, $\mu(A_k(x)) > 0$ and for some index i, $A_k(x) = A_{k,i}$, we have proved (16).

Particularly, almost surely, for all $x \in S$, and for all $k \geq 2$,

$$\hat{M}_{1,n}(x) \to M_1(x) \tag{17}$$

and

$$\hat{\Delta}_{k,n,L}(x) \to \Delta_{k,L}(x). \tag{18}$$

Let integer $R > 1$ be arbitrary. Let n be so large that $N_n > R$. For all $x \in S$,

$$\begin{aligned}
&|\hat{m}_{n,L}(x) - m_L(x)| \\
&\leq |\hat{M}_{1,n}(x) - M_1(x)| + \sum_{k=2}^{N_n} |\hat{\Delta}_{k,n,L}(x) - \Delta_{k,L}(x)| + \sum_{k=N_n+1}^{\infty} |\Delta_{k,L}(x)| \\
&\leq |\hat{M}_{1,n}(x) - M_1(x)| + \sum_{k=2}^{R} |\hat{\Delta}_{k,n,L}(x) - \Delta_{k,L}(x)| + \sum_{k=R+1}^{\infty} (|\hat{\Delta}_{k,n,L}(x)| + |\Delta_{k,L}(x)|) \\
&\leq |\hat{M}_{1,n}(x) - M_1(x)| + \sum_{k=2}^{R} |\hat{\Delta}_{k,n,L}(x) - \Delta_{k,L}(x)| + 2L \sum_{k=R+1}^{\infty} 2^{-k} \\
&\leq |\hat{M}_{1,n}(x) - M_1(x)| + \sum_{k=2}^{R} |\hat{\Delta}_{k,n,L}(x) - \Delta_{k,L}(x)| + L2^{-(R-1)}.
\end{aligned} \tag{19}$$

By (17) and (18), almost surely, for all $x \in S$,

$$|\hat{M}_{1,n}(x) - M_1(x)| + \sum_{k=2}^{R} |\hat{\Delta}_{k,n,L}(x) - \Delta_{k,L}(x)| \to 0. \tag{20}$$

By (19), almost surely, for all $x \in S$,

$$\limsup_{n\to\infty} |\hat{m}_{n,L}(x) - m_L(x)| \le L2^{-(R-1)}. \tag{21}$$

Since R was arbitrary, (13) is proved.

Now we prove (14). Assume the support S of μ is bounded. Let $\mathcal{A}_k$ denote the set of hyper-cubes from partition $\mathcal{P}_k$ with nonempty intersection with S. That is, define

$$\mathcal{A}_k = \{A \in \mathcal{P}_k : \ A \cap S \neq \emptyset\}. \tag{22}$$

Since S is bounded, $\mathcal{A}_k$ is a finite set. For $A \in \mathcal{P}_k$ let $a(A)$ be the center of A. Then almost surely,

$$\begin{aligned}
&\sup_{x\in S}\left(|\hat{M}_{1,n}(x) - M_1(x)| + \sum_{k=2}^{R} |\hat{\Delta}_{k,n,L}(x) - \Delta_{k,L}(x)|\right) \\
&\quad\le \max_{A\in\mathcal{A}_1} |\hat{M}_{1,n}(a(A)) - M_1(a(A))| + \sum_{k=2}^{R} \max_{A\in\mathcal{A}_k} |\hat{\Delta}_{k,n,L}(a(A)) - \Delta_{k,L}(a(A))| \qquad (23)\\
&\quad\to 0 \qquad (24)
\end{aligned}$$

keeping in mind that only finitely many terms are involved in the maximization operation. The rest of the proof goes virtually as before.

Now we prove (15).

$$|\hat{m}_{n,L}(x) - m_L(x)|^2 \le 2\left(|\hat{M}_{1,n}(x) - M_1(x)|^2 + |M_1(x) + \sum_{k=2}^{N_n} \hat{\Delta}_{k,n,L}(x) - m_L(x)|^2\right).$$

If condition (i) holds, then for the first term we have dominated convergence

$$|\hat{M}_{1,n}(x) - M_1(x)|^2 \le (2D)^2,$$

and for the second one, too:

$$\begin{aligned}
&|M_1(x) + \sum_{k=2}^{N_n} \hat{\Delta}_{k,n,L}(x) - m_L(x)| \\
&\quad\le \sum_{k=2}^{\infty}(|\hat{\Delta}_{k,n,L}(x)| + |\Delta_{k,L}(x)|) \\
&\quad\le L,
\end{aligned}$$

and thus (15) follows by Lebesgue's dominated convergence theorem,

$$0 = \int \lim_{n\to\infty} |\hat{m}_{n,L}(x) - m_L(x)|^2 \mu(dx) = \lim_{n\to\infty} \int |\hat{m}_{n,L}(x) - m_L(x)|^2 \mu(dx)$$

almost surely. If condition (ii) holds then (15) follows from (14).

□

Corollary 1 *Assume $m(x)$ is Lipschitz continuous with Lipschitz constant C. With the choice of $L \geq C\sqrt{d}$, for all $x \in S$, $m_L(x) = m(x)$ and Theorem 1 holds with $m_L(x)$ replaced by $m(x)$.*

Proof Since $m(x)$ is Lipschitz with constant $L/\sqrt{d}$, for $x \in S$,

$$\begin{aligned} |M_k(x) - m(x)| &\leq |\frac{\int_{A_k(x)} m(y)\mu(dy)}{\mu(A_k(x))} - m(x)| \\ &\leq \frac{1}{\mu(A_k(x))} \int_{A_k(x)} |m(y) - m(x)|\mu(dy) \\ &\leq \frac{1}{\mu(A_k(x))} \int_{A_k(x)} (L/\sqrt{d})(2^{-k-2}\sqrt{d})\mu(dy) \\ &= L2^{-k-2} \end{aligned}$$

and $M_k(x) \to m(x)$. For $x \in S$ we get

$$\begin{aligned} |M_k(x) - M_{k-1}(x)| &\leq |M_k(x) - m(x)| + |m(x) - M_{k-1}(x)| \\ &\leq L2^{-k-2} + L2^{-k-1} \\ &< L2^{-k}. \end{aligned}$$

Thus $m(x) = M_1(x) + \sum_{k=2}^{\infty} \Delta_k(x)$ and $\Delta_{k,L}(x) = \Delta_k(x)$ for all $x \in S$. Hence for all $x \in S$,

$$m_L(x) = M_1(x) + \sum_{k=2}^{\infty} \Delta_{k,L}(x) = M_1(x) + \sum_{k=2}^{\infty} \Delta_k(x) = m(x)$$

and Corollary 1 is proved.

□

Remark 1. If there is no truncation, that is if $L = \infty$, then $\hat{m}_n = \hat{M}_{N_n,n}$. In this case, $\hat{m}_n$ is the standard partitioning estimate (defined, for example in [14]). It is known that there is an ergodic process (X_i, Y_i) with Lipschitz continuous $m(x)$ with constant $C = 1$ such that a classical partitioning estimate is not even weakly consistent. (cf. Györfi, Morvai, Yakowitz [16]).

Remark 2. Our consistency is not universal, however, since m is hypothesized to be Lipschitz continuous.

Remark 3. N_n can be data dependent, provided $N_n \to \infty$ a.s.

Remark 4. The methodology here is applicable to linear auto-regressive processes. Let $\{Z_i\}$ be i.i.d. random variables with $EZ = 0$ and $Var(Z) < \infty$. Define

$$W_{n+1} = a_1 W_n + a_2 W_{n-1} + \ldots + a_K W_{n-K+1} + Z_{n+1} \tag{25}$$

where $\sum_{i=1}^{K} |a_i| < 1$. Equation (25) yields a stationary ergodic solution. Assume $K \leq d$. Let $Y_{n+1} = W_{n+1}$, and $X_{n+1} = (W_n, \ldots, W_{n-d+1})$. Now

$$\begin{aligned} m(X_{n+1}) = E(Y_{n+1}|X_{n+1}) &= E(W_{n+1}|W_n, \ldots, W_{n-d+1}) \\ &= a_1 W_n + a_2 W_{n-1} + \ldots a_K W_{n-K+1}. \end{aligned}$$

The regression function $m(x)$ is Lipschitz continuous with constant $C = 1$, since for $x = (x_1, \ldots, x_d)$ and $z = (z_1, \ldots, z_d)$,

$$|m(x) - m(z)| \leq \sum_{i=1}^{K} |a_i||x_i - z_i| \leq \max_{1 \leq i \leq d} |x_i - z_i| \leq \|x - z\|.$$

3 Truncated kernel estimation

Let $K(x)$ be a non-negative continuous kernel function with

$$b 1_{\{x \in S_{0,r}\}} \leq K(x) \leq 1_{\{x \in S_{0,1}\}},$$

where $0 < b \leq 1$ and $0 < r < 1$. ($S_{z,r}$ denotes the closed ball around z with radius r.) Choose

$$h_k = 2^{-k-2}$$

and

$$M_k^*(x) = \frac{E(YK(\frac{X-x}{h_k}))}{E(K(\frac{X-x}{h_k}))} = \frac{\int m(z)K(\frac{z-x}{h_k})\mu(dz)}{\int K(\frac{z-x}{h_k})\mu(dz)}. \tag{26}$$

Let

$$\Delta_k^*(x) = M_k^*(x) - M_{k-1}^*(x). \tag{27}$$

As a motivation, we note that Devroye [9] yields (4), and therefore (6), too. Now for $k \geq 2$, define

$$\Delta_{k,L}^*(x) = \text{sign}(M_k^*(x) - M_{k-1}^*(x)) \min(|M_k^*(x) - M_{k-1}^*(x)|, L2^{-k}). \tag{28}$$

Define

$$m_L^*(x) := M_1^*(x) + \sum_{i=2}^{\infty} \Delta_{i,L}^*(x). \tag{29}$$

Put

$$\hat{M}_{k,n}^*(x) := \frac{\sum_{j=1}^n Y_j K(\frac{X_j-x}{h_k})}{\sum_{j=1}^n K(\frac{X_j-x}{h_k})}$$

where we use the convention that $0/0 = 0$. Now for $k \geq 2$, introduce

$$\hat{\Delta}_{k,n,L}^*(x) = \text{sign}(\hat{M}_{k,n}^*(x) - \hat{M}_{k-1,n}^*(x)) \min(|\hat{M}_{k,n}^*(x) - \hat{M}_{k-1,n}^*(x)|, L2^{-k}) \tag{30}$$

and

$$\hat{m}_{n,L}^*(x) = \hat{M}_{1,n}^*(x) + \sum_{k=2}^{N_n} \hat{\Delta}_{k,n,L}^*(x). \tag{31}$$

Redefine the support S of μ as

$$S := \{x \in R^d : \ \mu(S_{x,1/k}) > 0 \text{ for all } k \geq 1\}. \tag{32}$$

By Cover and Hart [7], $\mu(S) = 1$.

Theorem 2 *Let $\{(X_i, Y_i)\}$ be a stationary ergodic time series with $E|Y_i| < \infty$. Assume $N_n \to \infty$. Then almost surely, for all $x \in S$,*

$$\hat{m}^*_{n,L}(x) \to m^*_L(x). \tag{33}$$

If the support S of μ is a bounded subset of R^d then almost surely

$$\sup_{x \in S} |\hat{m}^*_{n,L}(x) - m^*_L(x)| \to 0. \tag{34}$$

If either (i) $|Y| \leq D < \infty$ almost surely (D need not be known) or (ii) μ is of bounded support then

$$\int (\hat{m}^*_{n,L}(x) - m^*_L(x))^2 \mu(dx) \to 0. \tag{35}$$

Proof We first prove that (16) holds with $\hat{M}^*_{k,n}$ and M^*_k. Let

$$g_{k,n}(x) = \frac{1}{n} \sum_{j=1}^{n} Y_j K\left(\frac{X_j - x}{h_k}\right)$$

and

$$g_k(x) = E\left(Y K\left(\frac{X - x}{h_k}\right)\right).$$

Similarly put

$$f_{k,n}(x) = \frac{1}{n} \sum_{j=1}^{n} K\left(\frac{X_j - x}{h_k}\right)$$

and

$$f_k(x) = E K\left(\frac{X - x}{h_k}\right).$$

We have to show that almost surely, for all $k \geq 1$, and for all $x \in S$, both $g_{k,n}(x) \to g_k(x)$ and $f_{k,n}(x) \to f_k(x)$. Consider $g_{k,n}(x)$ with k fixed. Let $Q \subseteq R^d$ denote the set of vectors with rational coordinates. (Note that the set Q has countably many elements.) By the ergodic theorem, almost surely, for all $r \in Q$,

$$g_{k,n}(r) \to g_k(r).$$

Let $\delta > 0$ be arbitrary. Let integers $Z-1 > M > 0$ be so large that $E\left(|Y|1_{\{X \notin S_{0,M}\}}\right) < \delta$. By ergodicity, almost surely,

$$\sup_{x \notin S_{0,Z}} |g_{k,n}(x)| \leq \frac{1}{n}\sum_{i=1}^{n} |Y_i|1_{\{X_i \notin S_{0,M}\}} \to E\left(|Y|1_{\{X \notin S_{0,M}\}}\right) < \delta.$$

Since $K_{h_k}(x) = K(\frac{x}{h_k})$ is continuous and $K_{h_k}(x) = 0$ if $\|x\| > h_k$ and hence $K_{h_k}(x)$ is uniformly continuous on R^d. Define

$$U_k(u) = \sup_{x,z \in R^d : \|x-z\| \leq u} |K_{h_k}(x) - K_{h_k}(z)|.$$

Let $B_\delta \subseteq S_{0,Z} \cap Q$ be a finite subset of vectors with rational coordinates such that

$$\sup_{x \in S_{0,Z}} \min_{r \in B_\delta} U_k(\|x - r\|) < \delta.$$

For $x \in S_{0,Z}$, let $r(x)$ denote one of the closest rational vector $r \in B_\delta$ to x. Now

$$\begin{aligned}
\sup_{x \in S_{0,Z}} |g_{k,n}(x) - g_{k,m}(x)| &\leq \sup_{x \in S_{0,Z}} |g_{k,n}(x) - g_{k,n}(r(x))| \\
&+ \sup_{x \in S_{0,Z}} |g_{k,n}(r(x)) - g_{k,m}(r(x))| \\
&+ \sup_{x \in S_{0,Z}} |g_{k,m}(r(x)) - g_{k,m}(x)| \\
&\leq \delta\frac{1}{n}\sum_{i=1}^{n} |Y_i| + \max_{r \in B_\delta} |g_{k,m}(r) - g_{k,n}(r)| + \delta\frac{1}{m}\sum_{i=1}^{n} |Y_i|.
\end{aligned}$$

Combining the results, by the ergodic theorem, for almost all $\omega \in \Omega$, there exists $N(\omega)$ such that for all $m > N$, and $n > N$,

$$\sup_{x \in R^d} |g_{k,n}(x) - g_{k,m}(x)| \leq \sup_{x \in S_{0,Z}} |g_{k,n}(x) - g_{k,m}(x)|$$

$$
\begin{aligned}
&+ \quad \sup_{x \notin S_{0,Z}} |g_{k,n}(x) - g_{k,m}(x)| \\
&\leq \quad 2\delta E|Y| + 3\delta.
\end{aligned}
$$

Since δ was arbitrary, for almost all $\omega \in \Omega$, for every $\epsilon > 0$, there exists an integer $N_\epsilon(\omega)$ such that for all $m > N_\epsilon(\omega)$, $n > N_\epsilon(\omega)$:

$$
\sup_{x \in R^d} |g_{k,n}(x) - g_{k,m}(x)| < \epsilon. \tag{36}
$$

As a consequence, almost surely, the sequence of functions $\{g_{k,n}\}_{n=1}^{\infty}$ converges uniformly. Since all $g_{k,n}$ are continuous, the limit function must be also continuous. Since almost surely, for all $r \in Q$, $g_{k,n}(r) \to g_k(r)$, and by the Lebesgue dominated convergence g_k is continuous, the limit function must be g_k. Since there are countably many k, almost surely, for all $k \geq 1$,

$$
\sup_{x \in R^d} |g_{k,n}(x) - g_k(x)| \to 0.
$$

The same argument implies that almost surely, for all $k \geq 1$,

$$
\sup_{x \in R^d} |f_{k,n}(x) - f_k(x)| \to 0.
$$

We have proved (16). The rest of the proof of (33) goes as in the proof of Theorem 1. Now we prove (34). Since now, by assumption, the support is bounded, and since it is closed, and hence it is compact. Now note that there must exist an $\epsilon > 0$ such that $\inf_{x \in S} f_k(x) > \epsilon$. (Otherwise, there would be a sequence $x_i \in S$ such that $\liminf_{i \to \infty} f_k(x_i) = 0$. Continuity on a compact set would imply that there would be an $x \in S$ such that $f_k(x) = 0$ in contradiction to the hypothesis that $x \in S$.) By uniform convergence, for large n, $\inf_{x \in S} f_{k,n}(x) > \epsilon/2$. Thus

$$
\sup_{x \in S} \left| \frac{g_{k,n}(x)}{f_{k,n}(x)} - \frac{g_k(x)}{f_k(x)} \right| \quad \leq \quad \sup_{x \in S} \left| \frac{g_{k,n}(x)(f_k(x)/f_{k,n}(x)) - g_k(x)}{f_k(x)} \right|
$$

$$
\begin{aligned}
&\leq \frac{1}{\epsilon}\sup_{x\in S}\left|\frac{f_k(x)}{f_{k,n}(x)}\right||g_{k,n}(x)-g_k(x)|+|g_k(x)|\left|\frac{f_k(x)}{f_{k,n}(x)}-1\right| \\
&\leq \frac{2}{\epsilon^2}\sup_{x\in S}|g_{k,n}(x)-g_k(x)|+\sup_{x\in S}|g_k(x)|\frac{2}{\epsilon}\sup_{x\in S}|f_{k,n}(x)-f_k(x)| \\
&\to 0.
\end{aligned}
$$

Thus almost surely, for all $k \geq 1$,

$$\sup_{x\in S}|\hat{M}^*_{k,n}(x)-\hat{M}^*_k(x)| \to 0.$$

Almost surely, for arbitrary integer $R > 2$,

$$\sup_{x\in S}\left(|\hat{M}^*_{1,n}(x)-M^*_1(x)|+\sum_{k=2}^{R}|\hat{\Delta}^*_{k,n,L}(x)-\Delta^*_{k,L}(x)|\right) \to 0.$$

The rest of the proof goes exactly as in Theorem 1.

□

Corollary 2 *Assume $m(x)$ is Lipschitz continuous with Lipschitz constant C. With the choice of $L \geq C$ for all $x \in S$, $m^*_L(x) = m(x)$ and Theorem 2 holds with $m^*_L(x)$ substituted by $m(x)$.*

Proof Since $m(x)$ is Lipschitz with constant C, for $x \in S$,

$$
\begin{aligned}
|M^*_k(x)-m(x)| &\leq \left|\frac{\int m(z)K(\frac{z-x}{h_k})\mu(dz)}{\int K(\frac{z-x}{h_k})\mu(dz)}-m(x)\right| \\
&\leq \frac{\int |m(z)-m(x)|K(\frac{z-x}{h_k})\mu(dz)}{\int K(\frac{z-x}{h_k})\mu(dz)} \\
&\leq Ch_k \\
&\leq L2^{-k-2},
\end{aligned}
$$

therefore

$$|M^*_k(x)-M^*_{k-1}(x)| < L2^{-k}.$$

The rest of the proof goes as in Corollary 1.

□

4 Conclusions

This contribution is part of a long-standing endeavor of the authors to extend nonparametric forecasting methodology to the most lenient assumptions possible. The present work does push into new territory: strong consistency for finite regression under a Lipschitz assumption. The computational aspects have not been explored, but the algorithms are so close to their traditional partitioning and kernel counterparts that it is evident that they could be implemented and in fact, might be competitive.

The fundamental formula (8) leading to the truncated histogram approach was motivated by a representation used in a related but non-constructive setting by Kieffer [17]. The essence is to see that an infinite-dimensional nonparametric space may sometimes be decomposed into sums of terms in finite dimensional spaces, with tails of the summations being *a priori* asymptotically bounded over the regression class of interest. Through different devices, two ideas for obtaining such tail bounds for the partition and kernel methods have been presented.

Our contribution has been to apply the idea with Lipschitz continuity assuring the negligibility. Thus, results here are fundamentally intertwined with the Lipschitz bounds. Perhaps other useful expansions are possible. The interplay of finite subspaces and *a priori* bounded tails has proven a bit delicate. Sections 2 and 3 present different attacks to the error-bounding problem. The obvious nearest-neighbor estimator did not yield to this technique because the radii are random and do not necessarily decrease rapidly enough to assure bounded tails. The device which was successful here may find other applications. Evidently, a similar investigation could be carried out for regression classes having Fourier expansions with coefficients vanishing sufficiently quickly.

It is well-known (e.g., [33]) that universal convergence rates under the generality

of mere ergodicity do not exist. An avenue which would be worth exploring is that of adapting universal algorithms, such as explored and referenced here, so that they asymptotically attain the fastest possible convergence if, unknown to the statistician, the time series happens to fall into a mixing class. The design should be such that consistency is still assured if mixing rates do not hold.

References

[1] P. H. Algoet, Universal schemes for prediction, gambling and portfolio selection, *Annals Probab.,* vol 20, pp. 901–941, 1992. Correction: *ibid.*, vol. 23, pp. 474–478, 1995.

[2] P. H. Algoet, The strong law of large numbers for sequential decisions under uncertainty, *IEEE Trans. Inform. Theory,* vol. 40, pp. 609–634, May 1994.

[3] R. B. Ash, *Real Analysis and Probability.* Academic Press, 1972.

[4] D. H. Bailey, *Sequential Schemes for Classifying and Predicting Ergodic Processes.* Ph. D. thesis, Stanford University, 1976.

[5] Z. Bodie, A. Kane, A, and A. Marcus, *Investments,* 3rd ed. Irwin, Chicago 1996.

[6] T. Bollerslev, R. Y. Chou, and K. F. Kroner,*ARCH modeling in finance*, Journal of Econometrics, 52, pp. 5-59, 1992.

[7] T. M. Cover and P. Hart *Nearest neighbor pattern classification*, IEEE Transactions on Information Theory, IT-13, pp. 21–27, 1967.

[8] R. Davis and T. Mikosch, *The sample autocorrelations of heavy-tailed processes with applications to ARCH,* Annals of Statistics, 26, pp. 2049-2080, 1999.

[9] L. Devroye, *On the almost everywhere convergence of nonparametric regression function estimates*, Annals of Statistics, 9, pp. 1310–1319, 1981.

[10] L. Devroye and L. Györfi, *Distribution-free exponential bound on the L_1 error of partitioning estimates of a regression function*, In: Proceedings of the Fourth Pannonian Symposium on Mathematical Statistics, eds. F. Konecny, J. Mogyoródi, W. Wertz, pp. 67–76, Akadémiai Kiadó, Budapest, Hungary, 1983.

[11] L. Devroye and A. Krzyżak, *An equivalence theorem for L_1 convergence of the kernel regression estimate*, Journal of Statistical Planning and Inference, 23, pp. 71–82, 1989.

[12] L. Devroye and T. J. Wagner, *Distribution-free consistency results in nonparametric discrimination and regression function estimation*, Annals of Statistics, 8, pp. 231–239, 1980.

[13] W. Greblicki, A. Krzyżak and M. Pawlak, *Distribution-free pointwise consistency of kernel regression estimate*, Annals of Statistics, 12, pp. 1570–1575, 1984.

[14] L. Györfi, W. Härdle, P. Sarda and Ph. Vieu, *Nonparametric Curve Estimation from Time Series*. Berlin: Springer-Verlag, 1989.

[15] L. Györfi, *Universal consistencies of a regression estimate for unbounded regression functions*, In: Nonparametric Functional Estimation, ed. G. Roussas, pp. 329–338, NATO ASI Series, Kluwer, Berlin, 1991.

[16] L. Györfi, G. Morvai and S. Yakowitz, *Limits to consistent on-line forecasting for ergodic time series*, IEEE Transactions on Information Theory, IT-44, pp. 886-892, 1998.

[17] J. Kieffer, *Estimation of a convex real parameter of an unknown information source*, The Annals of Probability, 7, 882-886, 1979.

[18] S.R. Kulkarni and S.E. Posner, *Rates of convergence of nearest neighbour estimation under arbitrary sampling*, IEEE Transaction on Information Theory, IT-41, pp, 1028-1039, 1995.

[19] A. Krzyżak and M. Pawlak, *Distribution-free consistency of a nonparametric kernel regression estimate and classification*, IEEE Transactions on Information Theory, IT-30, pp. 78–81, 1984.

[20] B. B. Mandelbrot, *A Multifractal Walk down Wall Street*, Scientific American, XXX, pp. 70–73, 1999.

[21] E. Masry and D. Tjostheim,*Nonparametric estimation and identification of nonlinear ARCH time series*, Econometric Theory, 11, pp. 258-289, 1995.

[22] G. Morvai, S. Yakowitz, and L. Györfi, *Nonparametric inferences for ergodic, stationary time series,* Annals of Statistics, 24, pp. 370-379, 1996.

[23] G. Morvai, *Estimation of Conditional Distributions for Stationary Time Series.* Ph. D. Thesis, Technical University of Budapest, 1996.

[24] G. Morvai, S. Kulkarni and A. Nobel, *Regression estimation from an individual stable sequence,* Statistics 33, no. 2, pp. 99–118, 1999.

[25] G. Morvai, S. Yakowitz, and P. Algoet, *Weakly convergent nonparametric forecasting of stationary time series,* IEEE Trans. Inform. Theory, 43, pp. 483–498, March, 1997.

[26] A. Nobel, G. Morvai and S. Kulkarni, *Density estimation from an individual numerical sequence,* IEEE Trans. Inform. Theory, 44, pp. 537-541, March, 1998.

[27] D. Ornstein, *Guessing the next output of a stationary process,* Israel J. of Math., **30**, pp. 292-296, 1978.

[28] S. Resnick, *Heavy tail modeling and teletraffic data,* Ann. Math. Statist., **25**, pp. 1805-1849, 1997.

[29] P. Robinson, *Time series with strong dependence,* in *Advances in Econometrics: Sixth World Congress*, C.A. Sims, ed., Vol. I, Cambridge University Press, pp. 47-95, 1994.

[30] M. Rosenblatt, *Density estimates and Markov sequences.* In *Nonparametric Techniques in Statistical Inference*, M. Puri, Ed. London: Cambridge University, pp. 199-210, 1970.

[31] G. Roussas, *Non-parametric estimation of the transition distribution of a Markov processes*, Annals of Inst. Statist. Math. 21, pp.73-87, 1969.

[32] B. Ya. Ryabko, *Prediction of random sequences and universal coding,* Problems of Inform. Trans., 24, pp. 87-96, Apr.-June 1988.

[33] P. Shields, *Universal redundancy rates don't exist,* IEEE Trans. Inform. Theory, 39, pp. 520-524, 1993.

[34] C. Spiegelman and J. Sacks, *Consistent window estimation in nonparametric regression*, Annals of Statistics, 8, pp. 240–246, 1980.

[35] C. J. Stone, *Consistent nonparametric regression*, Annals of Statistics, 8, pp. 1348–1360, 1977.

[36] G. S. Watson, *Smooth regression analysis*, Sankhya Series A, **26,** pp. 359–372, 1964.

[37] W. Willinger, M. S. Taqqu, W. E. Leland, and D. V. Wilson, *Self-similarity in high-speed packet traffic-analysis and modeling on ethernet traffic measurements*, Statistical Science, **10**, pp. 67-85, 1995.

[38] S. Yakowitz, *Nearest neighbor regression estimation for null-recurrent Markov time series*, Stochastic Processes and their Applications, **37**, pp. 311-318, 1993.

[39] J. Ziv and A. Lempel, Compression of individual sequences by variable rate coding, *IEEE Trans. Inform. Theory*, **IT-24**, pp. 530-536, Sept. 1978.