Regression Estimation from an Individual Stable Sequence

Gusztáv Morvai, Sanjeev R. Kulkarni, Andrew B. Nobel *

January, 1999

Abstract

We consider univariate regression estimation from an individual (non-random) sequence $(x_1, y_1), (x_2, y_2), \ldots \in \mathbb{R} \times \mathbb{R}$, which is stable in the sense that for each interval $A \subseteq \mathbb{R}$, (i) the limiting relative frequency of A under $x_1, x_2, \ldots$ is governed by an unknown probability distribution μ, and (ii) the limiting average of those y_i with $x_i \in A$ is governed by an unknown regression function $m(\cdot)$.

A computationally simple scheme for estimating $m(\cdot)$ is exhibited, and is shown to be L_2 consistent for stable sequences $\{(x_i, y_i)\}$ such that $\{y_i\}$ is bounded and there is a known upper bound for the variation of $m(\cdot)$ on intervals of the form $(-i, i]$, $i \geq 1$. Complementing this positive result, it is shown that there is no consistent estimation scheme for the family of stable sequences whose regression functions have finite variation, even under the restriction that $x_i \in [0, 1]$ and y_i is binary-valued.

Appears in *Statistics*, vol. 33, pp.99-118, 1999.

Key words and phrases: nonparametric estimation, regression estimation, individual sequences, ergodic time series.

*Gusztáv Morvai is with Research Group for Informatics and Electronics of the Hungarian Academy of Sciences, Department of Computer Science and Information Theory, Technical University of Budapest 1521 Goldmann György tér 3, Budapest, Hungary. Email: morvai@inf.bme.hu Sanjeev Kulkarni is with the Department of Electrical Engineering, Princeton University, Princeton, NJ 08544. Email: kulkarni@ee.princeton.edu Andrew Nobel is with the Dept. of Statistics, University of North Carolina, Chapel Hill, NC 27599-3260. Email: nobel@stat.unc.edu

1 Introduction

Individual numerical sequences (binary and real-valued) have played an important role in the theory of data compression and computational complexity. The theory of lossless data compression developed by Ziv and Lempel [13], Ziv [24], and the complexity theory of Kolmogorov [9, 10] and Chaitin [3] are both formulated within a purely deterministic framework that is built around individual sequences. Subsequent work in these areas has considered useful notions of randomness, compressibility, and predictability. More recently, individual sequences have been studied in the context of statistical learning theory. In spite of the above research, there has been little consideration of individual sequences in the context of classical statistical estimation.

It is common in statistics to treat data, for the purposes of analysis, as a sequence of (typically independent) identically distributed random variables. This stochastic point of view collapses when one is faced with a *particular* collection of data, which is a fixed sequence of numbers or vectors from which we hope to learn something about the state of nature.

It is natural then to (re)formulate some classical statistical problems in terms of individual sequences. We concern ourselves here with the important problem of regression estimation. In the common statistical setting one is given n independent replicates $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a jointly distributed pair $(X, Y) \in \mathbb{R} \times \mathbb{R}$, and asked to find an estimate of the regression function $m(x) = E[Y|X = x]$. Justification for estimation of $m(x)$ comes from the fact that it minimizes $E(h(X) - Y)^2$ over all functions $h(\cdot)$ of X. Thus $m(x)$ is the least squares estimate of Y given X.

In this paper we present and analyze a simple regression estimation procedure that is applicable in a purely deterministic setting. By applying our estimates to individual sample paths, we easily establish their almost-sure consistency for ergodic processes having suitable one-dimensional distributions (the dependence structure of the process is unimportant). The approach and results of this paper are motivated by, and closely related to, recent results of [18] on density estimation from individual sequences.

For independent and weakly dependent stochastic data, a variety of estimation schemes have been proposed, including procedures based on histograms, kernels, neural networks, orthogonal series, wavelets, and nearest neighbors. For a description of some of these methods see, for example, Györfi, Härdle, Sarda, and Vieu [7], Roussas [21], Devroye-Krzyzak [5] and the references therein. Kulkarni and Posner [12] studied nearest neighbor regression estimates in the case where $x_1, x_2, \ldots$ are deterministic, but $Y_1, Y_2, \ldots$ are random

and conditionally independent given the x_i's. Yakowitz *et al.* [23] considered a family of truncated histogram regression estimates for processes with vector-valued covariates. For each constant $L > 0$ they exhibit a sequence of estimates that is almost surely pointwise consistent for every ergodic process $\{(X_i, Y_i)\}$ whose regression function satisfies a Lipschitz condition of the form $|g^*(x) - g^*(y)| \leq L||x - y||$. In practice, the constant L is known and fixed in advance of the data. Related work has been done in the area of nonparametric forecasting for a stationary process X_i. Cover [4] posed some natural questions which have been addressed by Bailey [2], Ryabko [22], and Ornstein [19], and more recently by Algoet [1], Morvai, Yakowitz, Györfi [16] and Morvai, Yakowitz, Algoet [15]. Nobel [17] has shown that no regression procedure is consistent for every bivariate ergodic process, even if one assumes that X_i is bounded and Y_i is binary valued. A similar negative result for individual sequences is established in Theorem 2 below.

In order to study regression estimation in a deterministic setting one must first specify how an individual sequence $(x_1, y_1), (x_2, y_2), \ldots$ can contain information about a regression function. In the present paper, following [18], it is required that suitable averages over the sequence are convergent or 'stable'. The deterministic setting of this paper is also in line with other recent work on individual sequences in information theory, statistics, and learning theory (cf. [24, 14, 8]). The principal contribution of the paper is to show how one may extract asymptotic information from the sequence in the absence of probabilistic inequalities, mixing conditions, rates of convergence, and so on. The deterministic setting is described in Section 2 and the principal results of the paper are stated in Section 3. Proofs of the principal results are given in Sections 4 and 5.

2 The Deterministic Setting

Let μ be a probability distribution on $(\mathbb{R}, \mathcal{B})$, and let $m : \mathbb{R} \to \mathbb{R}$ be a function satisfying $\int |m(x)|\mu(dx) < \infty$. Let $\mathbf{x} = (x_1, x_2, \ldots)$ and $\mathbf{y} = (y_1, y_2, \ldots)$ be infinite sequences of real numbers. For each interval $A \subseteq \mathbb{R}$ define the signed measure

$$\nu(A) = \int_A m(x)\mu(dx)\,.$$

For each $n \geq 1$ define the relative frequency

$$\hat{\mu}_n(A) = \frac{1}{n}\sum_{i=1}^{n} I\{x_i \in A\}\,,$$

and the joint sample average

$$\hat{\nu}_n(A) = \frac{1}{n}\sum_{i=1}^{n} y_i I\{x_i \in A\}\,.$$

The sequence $\mathbf{x}$ will be said to have *limiting distribution* $\mu(\cdot)$ if

$$\hat{\mu}_n(-\infty, t] \to \mu(-\infty, t] \;\text{ and }\; \hat{\mu}_n(\{t\}) \to \mu(\{t\}) \;\text{ for every } t \in \mathbb{R}, \tag{1}$$

and the pair $(\mathbf{x}, \mathbf{y})$ will be said to have *limiting regression* $m(\cdot)$ if

$$\hat{\nu}_n(-\infty, t] \to \nu(-\infty, t] \;\text{ and }\; \hat{\nu}_n(\{t\}) \to \nu(\{t\}) \;\text{ for every } t \in \mathbb{R}. \tag{2}$$

(Note that the second condition is superfluous in each case if μ is non-atomic.) By minor modification of a standard proof of the Glivenko Cantelli Theorem (such as that in Pollard [20]), one may show that if $\mathbf{x}$ has limiting distribution $\mu(\cdot)$ then in fact

$$\sup_{A \in \mathcal{A}} |\hat{\mu}_n(A) - \mu(A)| \to 0\,, \tag{3}$$

where $\mathcal{A}$ is the collection of all intervals of the form $(a, b]$ and $(-\infty, b]$ with $a, b \in \mathbb{R}$.

An individual sequence $(\mathbf{x}, \mathbf{y})$ satisfying (1) and (2) will be called *stable*. Let $\Omega(\mu, m)$ denote the set of stable sequences with limiting distribution μ and limiting regression m. Stationarity concerns only the asymptotic behavior of $\hat{\mu}_n$ and $\hat{\nu}_n$, which need not converge to their respective limits at any particular rate. No constraints are place on the mechanism by which the individual sequences $(\mathbf{x}, \mathbf{y})$ are produced. Note in particular that membership of $(\mathbf{x}, \mathbf{y})$ in $\Omega(\mu, m)$ is unaffected if one adds to $\mathbf{x}$ and $\mathbf{y}$ finite prefixes $x'_1, \ldots, x'_k$ and $y'_1, \ldots, y'_k$ having the same length. The next proposition, showing that the sample paths of ergodic processes are stable with probability one, follows easily from Birkhoff's ergodic theorem.

Proposition 1 *Let $(X_1, Y_1), (X_2, Y_2), \ldots$ be stationary such that $E|Y| < \infty$. Then $\{(X_i, Y_i)\}$ is stable with probability one. If in addition $\{(X_i, Y_i)\}$ is ergodic then $(\mathbf{X}, \mathbf{Y}) \in \Omega(\mu, m)$ with probability one, where $\mu(A) = P(X \in A)$, $m(x) = E(Y|X = x)$, $\mathbf{X} = (X_1, X_2, \ldots)$ and $\mathbf{Y} = (Y_1, Y_2, \ldots)$.*

Proof: Let $\mathcal{E}$ denote the invariant σ-algebra. By Gray [6] (Theorem 6.6.1 p. 204), for arbitrary Borel-measurable set $A \subset \mathbb{R}$ with probability one,

$$\hat{\mu}_n(A) \to P(X_1 \in A|\mathcal{E}) =: \mu_{\mathcal{E}}(A)$$

and

$$\hat{\nu}_n(A) \to E(Y_1 I_{\{X_1 \in A\}}|\mathcal{E}) =: \nu_{\mathcal{E}}(A).$$

If in addition $\{(X_i, Y_i)\}$ is ergodic then $\mathcal{E}$ is the trivial σ-algebra and so

$$\hat{\mu}_n(A) \to P(X_1 \in A)$$

and

$$\hat{\nu}_n(A) \to E(Y_1 I_{\{X_1 \in A\}}).$$

The rest follows from the standard proof of the Glivenko Cantelli Theorem cf. Pollard [20]. □

Remark 1. Note that for individual sequences,

$$\hat{\mu}_n(-\infty, t] \to \mu(-\infty, t] \text{ for all } t \in \mathbb{R}$$

does not necessarily imply

$$\hat{\mu}_n(\{t\}) \to \mu(\{t\}) \text{ for all } t \in \mathbb{R}.$$

Indeed, with $\mathbf{x} = (\frac{-1}{2}, \frac{-1}{3}, \ldots)$, $\hat{\mu}_n(-\infty, t] = 1$ for $t \geq 0$, while $\hat{\mu}_n(-\infty, t] \to 0$ for $t < 0$. Thus the limiting distribution μ should concentrate on the atom $\{0\}$, but $\hat{\mu}_n(\{0\}) = 0$ for all n.

3 Statement of Principal Results

Recall that the total variation of a real-valued function h defined on an interval $(a, b]$ is given by

$$V(h : a, b) = \sup \sum_{i=1}^{n} |h(t_i) - h(t_{i-1})|,$$

where the supremum is taken over all finite ordered sequences

$$a < t_0 < t_1 < \cdots < t_{n-1} < t_n = b.$$

Let $\mathbb{N}$ denote the positive intergers. For each non-decreasing function $\alpha : \mathbb{N} \to (0, \infty)$, let $\mathcal{F}(\alpha)$ denote the set of bounded measurable functions $m : \mathbb{R} \to \mathbb{R}$ such that $V(m : -i, i) < \alpha(i)$ for all $i \geq 1$. Let $\pi_0 = \{\mathbb{R}\}$, and for each $k \geq 1$ let π_k be the partition of $\mathbb{R}$ consisting of the dyadic intervals

$$A_{k,j} = \left(\frac{(j-1)}{2^k}, \frac{j}{2^k} \right] \quad -\infty < j < \infty.$$

Let $\pi_k[x]$ denote the unique cell of π_k containing $x \in \mathbb{R}$. Note that π_{k+1} refines π_k, and that for each x,

$$\lim_{k \to \infty} \text{len}(\pi_k[x]) = 0,$$

where $\mathrm{len}(A)$ denotes the length of an interval A.

Let $m \in \mathcal{F}(\alpha)$ be arbitrary. Let μ denote an arbitrary probability distribution on $\mathbb{R}$. Fix two numerical sequences $\mathbf{x}$ and $\mathbf{y}$ such that $(\mathbf{x}, \mathbf{y}) \in \Omega(\mu, m)$. For each $k \geq 1$ we define a histogram regression estimate based on π_k and adaptively chosen initial sequences of $\mathbf{x}$ and $\mathbf{y}$. For each $n \geq 1$, $k \geq 0$ define

$$\hat{m}_{k,n}(x) = \frac{\hat{\nu}_n(\pi_k[x])}{\hat{\mu}_n(\pi_k[x])}\,,$$

where by convention $0/0 = 0$. Note that $\hat{m}_{k,n}$ is piecewise constant on the cells of π_k. Let $\tau_0 = 1$ and for each $k \geq 1$ define

$$\tau_k = \min\left\{n > \tau_{k-1} \,:\, V(\hat{m}_{k,n} : -i, i) < 4\alpha(i) \text{ for all } 1 \leq i \leq k\right\}.$$

By Lemma 1, τ_k is well defined and finite. Note that $\tau_k \to \infty$. Define the estimate

$$\hat{m}_k = \hat{m}_{k,\tau_k}\,.$$

Note that $\hat{m}_k$ depends only on the pairs $(x_1, y_1), \ldots, (x_{\tau_k}, y_{\tau_k})$. To create a fixed sample size version of the estimate for $n \geq 1$ let

$$\kappa_n = \max\{k \geq 0 : \tau_k \leq n\}$$

and define

$$\tilde{m}_n = \hat{m}_{\kappa_n}.$$

The $L_2(\mu)$-consistency of the estimates is established in the following theorem.

Theorem 1 *Let $\alpha : \mathbb{N} \to (0, \infty)$ be a known, non-decreasing function. For every $m(\cdot) \in \mathcal{F}(\alpha)$, every probability distribution μ on $\mathbb{R}$, and every stable pair $(\mathbf{x}, \mathbf{y}) \in \Omega(\mu, m)$ such that the components of $\mathbf{y}$ are bounded,*

$$\int (\hat{m}_k(x) - m(x))^2 \mu(dx) \to 0 \quad \textit{and} \quad \int (\tilde{m}_n(x) - m(x))^2 \mu(dx) \to 0\,.$$

In other words, the estimates $\tilde{m}_n$ and $\hat{m}_k$ are $L_2(\mu)$-consistent.

Remark 2. Definition of $\hat{m}_k$ is based solely on $\alpha(\cdot)$ and the given numerical sequences. In advance of the data, one need only know a bound on the variation of its limiting regression on the intervals $(-i, i]$. The limiting distribution μ, the pre-asymptotic behavior of the individual sequences, and the bound on the y_i need not be known in advance.

Remark 3. Let $\mathcal{B}(M)$ denote the class of monotone bounded functions $m : \mathbb{R} \to \mathbb{R}$ such that $|m(x)| < M$ for all $x \in \mathbb{R}$. Since $B(M) \subset \mathcal{F}(\alpha)$ with $\alpha(n) = 2M$ Theorem 1 is applicable to $B(M)$.

Remark 4. Let $\Lambda(C)$ denote the class of Lipschitz continuous functions $m : \mathbb{R} \to \mathbb{R}$ such that $|m(x) - m(z)| < C|z - x|$ for all $x, z \in \mathbb{R}$. Since $L(M) \subset \mathcal{F}(\alpha)$ with $\alpha(n) = 2Cn + \epsilon$ where $0 < \epsilon < \infty$ is arbitrary, Theorem 1 is applicable to $\Lambda(C)$.

Theorem 1 and Poposition 1 imply the next corollary.

Corollary 1 *Let $\alpha : \mathbb{N} \to (0, \infty)$ be a known, non-decreasing function. For every stationary ergodic process $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathbb{R} \times \mathbb{R}$ such that X_i has distribution μ, Y is bounded with probability one, and $m(x) = E(Y_i | X_i = x) \in \mathcal{F}(\alpha)$,*

$$\int (\hat{m}_k(x) - m(x))^2 \mu(dx) \to 0 \quad \textit{and} \quad \int (\tilde{m}_n(x) - m(x))^2 \mu(dx) \to 0$$

with probability one.

Theorem 1 and Proposition 1 imply even more. We apply the same notations as in the proof of Proposition 1.

Corollary 2 *Let $\alpha : \mathbb{N} \to (0, \infty)$ be a known, non-decreasing function.*
Let $(X_1, Y_1), (X_2, Y_2), \ldots \in \mathbb{R} \times \mathbb{R}$ be a stationary process such that Y is bounded with probability one. Let $m_{\mathcal{E}} := \frac{d\nu_{\mathcal{E}}}{d\mu_{\mathcal{E}}}$, that is, $\nu_{\mathcal{E}}(A) = \int_A m_{\mathcal{E}}(x) \mu_{\mathcal{E}}(dx)$. Assume that $m_{\mathcal{E}}(\cdot) \in \mathcal{F}(\alpha)$ with probability one. Then

$$\int (\hat{m}_k(x) - m_{\mathcal{E}}(x))^2 \mu_{\mathcal{E}}(dx) \to 0 \quad \textit{and} \quad \int (\tilde{m}_n(x) - m_{\mathcal{E}}(x))^2 \mu_{\mathcal{E}}(dx) \to 0$$

with probability one.

The conditions in Theorem 1 cannot be significantly weakened.

Theorem 2 *Let λ denote the uniform distribution on $[0, 1]$. There is no $L_2(\lambda)$ consistent regression procedure for the family of stable sequences $(\mathbf{x}, \mathbf{y})$ such that $x_i \in [0, 1]$ has limiting distribution λ, and $y_i \in \{0, 1\}$ has limiting regression m with $V(m : 0, 1) < \infty$.*

4 Proof of Theorem 1

Lemma 1 *Let $\alpha : \mathbb{N} \to (0, \infty)$ be a known, non-decreasing function. For every $m(\cdot) \in \mathcal{F}(\alpha)$, every probability distribution μ on $\mathbb{R}$, every stable pair $(\mathbf{x}, \mathbf{y}) \in \Omega(\mu, m)$, and for all $k \geq 0$, τ_k is well defined and finite.*

Proof: By definition $\tau_0 = 1$. Hence we may assume $k \geq 1$. Let f be any function with bounded variation $V(f : -i, i) < \infty$ on $(-i, i]$. Define

$$(f \circ \pi_k)(x) = \frac{1}{\mu(\pi_k[x])} \int_{\pi_k[x]} f(z)\mu(dz).$$

Note that $f \circ \pi$ is piecewise constant on the cells of π.

For f non-decreasing it is immediate that $V(f \circ \pi_k : -i, i) \leq V(f : -i, i)$. If f is not necessarily non-decreasing then $f(x) = u(x) - v(x)$ where $u(\cdot)$ and $v(\cdot)$ are non-decreasing, $V(u : -i, i) \leq V(f : -i, i)$ and $V(v : -i, i) \leq 2V(f : -i, i)$ (cf. Kolmogorov and Fomin [11]). It follows from the definition that $f \circ \pi_k = u \circ \pi_k - v \circ \pi_k$, and since u and v are non-decreasing, so are $u \circ \pi_k$ and $v \circ \pi_k$. Therefore

$$\begin{aligned} V(f \circ \pi_k : -i, i) &= V(u \circ \pi_k - v \circ \pi_k : -i, i) \\ &\leq V(u \circ \pi_k : -i, i) + V(v \circ \pi_k : -i, i) \\ &\leq V(u : -i, i) + V(v : -i, i) \\ &\leq 3V(f : -i, i) \end{aligned}$$

as the variation of the sum is less than the sum of the variations. Now note that since $V(m : -i, i) < \alpha(i)$ hence as $n \to \infty$

$$\begin{aligned} V(\hat{m}_{k,n} : -i, i) &= \sum_{j=-i2^k+1}^{i2^k-1} |\frac{\hat{\nu}_n(A_{k,j})}{\hat{\mu}_n(A_{k,j})} - \frac{\hat{\nu}_n(A_{k,j+1})}{\hat{\mu}_n(A_{k,j+1})}| \\ &\to \sum_{j=-i2^k+1}^{i2^k-1} |\frac{\nu(A_{k,j})}{\mu(A_{k,j})} - \frac{\nu(A_{k,j+1})}{\mu(A_{k,j+1})}| \\ &= V(m \circ \pi_k : -i, i) \\ &\leq 3V(m : -i, i) < 4\alpha(i). \end{aligned}$$

Thus τ_k is well defined and finite. □

Proof of Theorem 1: Fix a sequence $(\mathbf{x}, \mathbf{y})$ satisfying the conditions of the theorem. For each $k \geq 1$ define $g_k(x) = \hat{m}_k(x) - m(x)$. It follows from the definition of τ_k and the assumption that $m(\cdot) \in \mathcal{F}(\alpha)$ that

$$V(g_k : -i, i) \leq V(\hat{m}_k : -i, i) + V(m : -i, i) < 5\alpha(i) \quad \text{for } 1 \leq i \leq k.$$

Let $D/2 > 1$ be a common bound for $m(\cdot)$ and the elements of $\mathbf{y}$, so that $|g_k(x)| < D$ for each x.

Let $U = \{u_1, u_2, \ldots\}$ be those numbers u for which $\mu(\{u\}) > 0$. Then U is either finite or countably infinite. Note that μ may be decomposed as a sum $\mu_d + \mu_c$, where μ_d is a purely atomic measure supported on U, and μ_c is non-atomic. Fix $\epsilon \in (0, 1)$. Let $T \geq 1$ be an integer such that

$$\mu(\{x : |x| \geq T\}) < \frac{\epsilon}{D^2} \tag{4}$$

and let $J \geq 1$ be so large that

$$\sum_{i=J+1}^{|U|} \mu(\{u_i\}) < \frac{\epsilon}{D^2}, \tag{5}$$

where $|U|$ denotes the cardinality of U. For $k \geq 1$ define

$$\Delta(k) = \min\{\mu_c(A) : A \in \pi_k,\ A \subseteq (-T, T],\ \mu_c(A) > 0\}$$

and

$$\Theta(k) = \max\{\mu_c(A) : A \in \pi_k, A \subseteq (-T, T]\}.$$

Note that $\Theta(k) \geq \Delta(k) > 0$ for each k and that $\Theta(k)$ is a non-increasing function of k. Let

$$\Theta^* = \lim_{k \to \infty} \Theta(k)\,.$$

Suppose that $\Theta^* > 0$. Then there is a sequence of intervals $A_k \in \pi_k$ such that $\mu_c(A_k) \geq \Theta^*$ and $\mathrm{clos}(A_{k+1}) \subseteq \mathrm{clos}(A_k)$ for each $k \geq 1$, where $\mathrm{clos}(A)$ denotes the closure of A. As $\mathrm{len}(A_k) \to 0$, $\cap_k \mathrm{clos}(A_k)$ is a singleton $\{x_0\}$. Continuity of μ_c implies that $\mu_c(\{x_0\}) \geq \Theta^* > 0$, which contradicts the fact that μ_c is non-atomic. Therefore $\Theta^* = 0$. Let $K \geq 1$ be so large that

$$\Theta(K) < \frac{\epsilon^2}{10\alpha(T)D^2}. \tag{6}$$

Fix an atom $u \in U$. If $r \leq k$ then

$$\frac{\hat{\nu}_{\tau_k}(\{u\}) + D\hat{\mu}_{\tau_k}(\{u\})}{\hat{\mu}_{\tau_k}(\pi_r(u))} \leq \frac{\hat{\nu}_{\tau_k}(\pi_k(u)) + D\hat{\mu}_{\tau_k}(\pi_k(u))}{\hat{\mu}_{\tau_k}(\pi_k(u))} \leq \frac{\hat{\nu}_{\tau_k}(\pi_r(u)) + D\hat{\mu}_{\tau_k}(\pi_r(u))}{\hat{\mu}_{\tau_k}(\{u\})}.$$

As k tends to infinity, stationarity implies that

$$\hat{\mu}_{\tau_k}(\pi_r(u)) \to \mu(\pi_r(u)),\quad \hat{\nu}_{\tau_k}(\pi_r(u)) \to \nu(\pi_r(u)),\quad \hat{\mu}_{\tau_k}(\{u\}) \to \mu(\{u\}).$$

As r tends to infinity, continuity of the measures μ and ν implies that

$$\mu(\pi_r(u)) \to \mu(\{u\}),\quad \nu(\pi_r(u)) \to \nu(\{u\}).$$

From these relations we conclude that

$$\lim_{k\to\infty} \frac{\hat{\nu}_{\tau_k}(\pi_k(u))}{\hat{\mu}_{\tau_k}(\pi_k(u))} = \frac{\nu(\{u\})}{\mu(\{u\})}. \tag{7}$$

By (3), (7) and (2) there exists $K' \geq \max(K, T)$ such that for all indices $k \geq K'$,

$$\sup_{A\in\mathcal{A}} |\hat{\mu}_{\tau_k}(A) - \mu(A)| < \frac{\epsilon}{4D}\Delta(K), \tag{8}$$

$$|g_k(u_i)|^2 < \frac{\epsilon}{J} \quad \text{for } i = 1, \ldots, J, \tag{9}$$

and

$$|\int_A \hat{m}_k d\hat{\mu}_{\tau_k} - \int_A m d\mu| \; < \; \frac{\epsilon}{4}\Delta(K) \tag{10}$$

for every cell $A \in \pi_K$ with $A \subseteq (-T, T]$.

Fix $k \geq K'$, and let $A \in \pi_K$ be such that $\mu(A) > 0$ and $A \subseteq (-T, T]$. Inequalities (8) and (10) imply that

$$\begin{aligned} |\int_A g_k(x)\mu(dx)| &\leq |\int_A \hat{m}_k d\mu - \int_A \hat{m}_k d\hat{\mu}_{\tau_k}| + |\int_A \hat{m}_k d\hat{\mu}_{\tau_k} - \int_A m d\mu| \\ &\leq D \sup_{A'\in\mathcal{A}} |\hat{\mu}_{\tau_k}(A') - \mu(A')| + \frac{\epsilon}{4}\Delta(K) \\ &\leq \frac{\epsilon}{2}\Delta(K)\,, \end{aligned}$$

and therefore

$$\left|\frac{\int_A g_k(x)\mu(dx)}{\mu(A)}\right| \leq \frac{\epsilon}{2}. \tag{11}$$

Consider those points

$$H_k = \{x \in \mathbb{R} : |g_k(x)| > \epsilon\}$$

for which g_k exceeds ϵ, and define

$$\mathcal{H}_k = \{A \in \pi_K \,:\, A \cap H_k \neq \emptyset,\, A \subseteq (-T, T],\, \mu(A) > 0\}.$$

If $A \in \mathcal{H}_k$ then there exists $x \in A$ such that $|g_k(x)| > \epsilon$. Assume without loss of generality that $g_k(x) > \epsilon$. By virtue of (11) there exists $z \in A$ such that $g_k(z) \leq \epsilon/2$, and therefore, $|g_k(x) - g_k(z)| > \epsilon/2$ for some $x, z \in A$. Consequently

$$\frac{\epsilon}{2}|\mathcal{H}_k| \leq V(g_k : -T, T) < 5\alpha(T)$$

from which follows that

$$|\mathcal{H}_k| < \frac{10\alpha(T)}{\epsilon}. \tag{12}$$

Consider now the $L_2(\mu)$ error of $\hat{m}_k$. From the definition of $\mathcal{H}_k$ and inequalities (12), (6), (5), (9), and (4) it follows that

$$\begin{aligned}\int |g_k(x)|^2\mu(dx) &\leq \sum_{A\in\mathcal{H}_k}\int_A D^2 d\mu_c + \sum_{A\in\mathcal{H}_k}\int_A |g_k(x)|^2 d\mu_d(x) \\ &+ \sum_{A\notin\mathcal{H}_k, A\subseteq(-T,T]}\int_A \epsilon^2 d\mu + \int_{|x|\geq T} D^2 d\mu \\ &\leq \epsilon + \sum_{i=1}^{J}|g_k(u_i)|^2 + \sum_{i=J+1}^{|U|} D^2\mu_d(\{u_i\}) + \epsilon^2 + \epsilon \\ &\leq 4\epsilon + \epsilon^2.\end{aligned}$$

Letting $k\to\infty$ and $\epsilon\to 0$ shows that $\int |g_k(x)|^2\mu(dx)\to 0$. Since $\kappa_n \nearrow \infty$, the $L_2(\mu)$ convergence of $\tilde{m}_n$ to m is immediate. □

5 Proof of Theorem 2

Proof of Theorem 2: For $k\geq 1$ define the k'th Rademacher function as

$$h_k(x) = \begin{cases} 1 & \text{if } 2j2^{-k}\leq x < (2j+1)2^{-k} \text{ for some } 0\leq j < 2^{k-1} \\ 0 & \text{otherwise ,}\end{cases}$$

and let $h_0(x) = 0.5I\{x\in[0,1]\}$. Define $\mathcal{F}_0 = \{h_0, h_1, h_2, \ldots\}$ and let $\mathcal{F}_1 = \{h_1, h_2, \ldots\}$. Let λ denote the uniform distribution on $[0,1]$. We will prove even more than stated in Theorem 2, namely:

There is no $L_2(\lambda)$ consistent regression estimation procedure for the family

$$\Omega^* = \bigcup_{m\in\mathcal{F}_0}\Omega(\lambda, m)\cap\{(\mathbf{x},\mathbf{y}) : x_n\in[0,1], y_n\in\{0,1\} \text{ for all } n\geq 1\}.$$

This statement says that even for the countable class $\mathcal{F}_0$ of regression functions there is no $L_2(\lambda)$ consistent estimation procedure. We briefly describe the main idea of the proof. Let $\Phi = \{\phi_1, \phi_2, \ldots\}$ be any regression estimation procedure. If Φ fails to be consistent for some sequence $(\mathbf{x},\mathbf{y})\in\bigcup_{m\in\mathcal{F}_1}$ with $x_i\in[0,1]$ and $y_i\in\{0,1\}$, there is nothing to prove. Assuming then that Φ is consistent for every such sequence, we construct a stable sequence $(\mathbf{x}^*,\mathbf{y}^*)$ such that $\phi_n(\cdot : (x_1^*, y_1^*), \ldots, (x_n^*, y_n^*))$ fails to converge. The sequence $(\mathbf{x}^*,\mathbf{y}^*)$ has limiting distribution λ and limiting regression h_0. It is constructed by 'splicing' together longer and longer blocks of stable sequences $(\mathbf{x}^{(k)},\mathbf{y}^{(k)})\in\Omega(h_k,\lambda)$. When applied to the resulting sequence, the procedure Φ first produces estimates close to h_1; as the sample size is

increased Φ produces estimates close to h_2, then h_3, and so on. As the h_i's fail to converge, so to do the estimates $\phi_n(\cdot : (x_1^*, y_1^*), \ldots, (x_n^*, y_n^*))$, $n \geq 1$.

Note that each h_j is supported on $[0,1]$ and that $\int |h_j(x) - h_k(x)|^2 \lambda(dx) = 0.5$ whenever $j \neq k$, and $j \geq 1$, $k \geq 1$. Let

$$\nu_k(A) = \int_A h_k(x)\lambda(dx)$$

and for each finite sequence $(x_1, y_1), \ldots, (x_m, y_m) \in [0,1] \times \{0,1\}$ let

$$\Delta(x_1, \ldots, x_m) = \sup_{A \in \mathcal{A}} \left| \frac{1}{m} \sum_{j=1}^{m} I\{x_j \in A\} - \lambda(A) \right| = \sup_{A \in \mathcal{A}} |\hat{\mu}_m(A) - \lambda(A)|$$

and

$$\begin{aligned} \tilde{\Delta}_k((x_1, y_1), \ldots, (x_m, y_m)) &= \sup_{A \in \mathcal{A}} \left| \frac{1}{m} \sum_{j=1}^{m} y_j I\{x_j \in A\} - \nu_k(A) \right| \\ &= \sup_{A \in \mathcal{A}} \left| \frac{1}{m} \sum_{j=1}^{m} I\{y_j = 1, x_j \in A\} - \nu_k(A) \right| \\ &= \sup_{A \in \mathcal{A}} |\hat{\nu}_m(A) - \nu_k(A)| \end{aligned}$$

where $\mathcal{A}$ is the collection of all intervals of the form $(a, b]$ and $(-\infty, b]$ with $a, b \in \mathbb{R}$.

A minor modification of a standard proof of the Glivenko Cantelli Theorem (e.g. using the bracketing approach found in Pollard [20]) shows that

$$\Delta(x_1, \ldots, x_m) \to 0 \text{ and } \tilde{\Delta}_k((x_1, y_1), \ldots, (x_m, y_m)) \to 0 \tag{13}$$

for all $(\mathbf{x}, \mathbf{y}) \in \Omega(\lambda, h_k) \cap \{(\mathbf{x}, \mathbf{y}) : x_n \in (0,1), y_n \in \{0,1\} \text{ for all } n \geq 1\}$.

Suppose now that $\Phi = \{\phi_1, \phi_2, \ldots\}$ is consistent for $\mathcal{F}_1$. For each $k \geq 1$ select a sequence

$$(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) = ((x_1^{(k)}, y_1^{(k)}), (x_2^{(k)}, y_2^{(k)}), \ldots)$$

such that

$$(\mathbf{x}^{(k)}, \mathbf{y}^{(k)}) \in \Omega(h_k, \lambda) \cap \{(\mathbf{x}, \mathbf{y}) : x_n \in (0,1), y_n \in \{0,1\} \text{ for all } n \geq 1\}$$

and

$$x_i^{(k)} = x_j^{(l)} \text{ if and only if } i = j,\ k = l \tag{14}$$

(e.g. typical sample sequences from independent i.i.d. time series

$$(X_1^{(k)}, h_k(X_1^{(k)})), (X_2^{(k)}, h_k(X_2^{(k)})), \ldots$$

where $X_i^{(k)}$ has distribution λ cf. Proposition 1). Define

$$l_k = \min\left\{L : \sup_{m\geq L} \Delta(x_1^{(k)},\ldots,x_m^{(k)}) \leq \frac{1}{k+1}\right\} \tag{15}$$

$$\tilde{l}_k = \min\left\{L : \sup_{m\geq L} \tilde{\Delta}_k((x_1^{(k)},y_1^{(k)}),\ldots,(x_m^{(k)},y_m^{(k)})) \leq \frac{1}{k+1}\right\}. \tag{16}$$

By (13), both l_k and $\tilde{l}_k$ are finite. Consider the infinite sequence $(\mathbf{x}^{(1)},\mathbf{y}^{(1)})$. As $h_1 \in \mathcal{F}_1$, and Φ is consistent for $\mathcal{F}_1$ by assumption,

$$\lim_{n\to\infty} \int |\phi_n(x : (x_1^{(1)},y_1^{(1)})\ldots,(x_n^{(1)},y_n^{(1)})) - h_1(x)|^2\lambda(dx) = 0.$$

Therefore there is an integer $n_1 \geq \max(l_2,\tilde{l}_2)$ and a corresponding initial segment $(\mathbf{v}^{(1)},\mathbf{w}^{(1)}) = ((x_1^{(1)},y_1^{(1)})\ldots,(x_{n_1}^{(1)},y_{n_1}^{(1)}))$ of $(\mathbf{x}^{(1)},\mathbf{y}^{(1)})$ such that

$$\int |\phi_{n_1}(x : (\mathbf{v}^{(1)},\mathbf{w}^{(1)})) - h_1(x)|^2\lambda(dx) \leq \frac{1}{40}$$

and

$$\Delta(\mathbf{v}^{(1)}) \leq \frac{1}{2}$$

and

$$\tilde{\Delta}_1((\mathbf{v}^{(1)},\mathbf{w}^{(1)})) \leq \frac{1}{2}.$$

Let $n_0 = 0$ and let n_1 be as defined above. Now suppose that for all $1 \leq j \leq k$ one has constructed sequences $(\mathbf{v}^{(j)},\mathbf{w}^{(j)})$ of finite length n_j in such a way that

$$(\mathbf{v}^j,\mathbf{w}^j) = (v_1^{(j-1)},w_1^{(j-1)}),\ldots,(v_{n_{j-1}}^{(j-1)},w_{n_{j-1}}^{(j-1)}),(x_1^{(j)},y_1^{(j)}),\ldots,(x_{n_j-n_{j-1}}^{(j)},y_{n_j-n_{j-1}}^{(j)}), \tag{17}$$

$$\int |\phi_{n_j}(x : (\mathbf{v}^{(j)},\mathbf{w}^{(j)})) - h_j(x)|^2\lambda(dx) \leq \frac{1}{40}, \tag{18}$$

$$\Delta(\mathbf{v}^{(j)}) \leq (j+1)^{-1} \tag{19}$$

$$\tilde{\Delta}_j((\mathbf{v}^{(j)},\mathbf{w}^{(j)})) \leq (j+1)^{-1} \tag{20}$$

$$n_j \geq j\cdot\max(l_{j+1},\tilde{l}_{j+1}). \tag{21}$$

As $(\mathbf{v}^{(k)},\mathbf{w}^{(k)})$ is finite, the concatenation

$$(v_1^{(k)},w_1^{(k)}),\ldots,(v_{n_k}^{(k)},w_{n_k}^{(k)}),(x_1^{(k+1)},y_1^{(k+1)}),(x_2^{(k+1)},y_2^{(k+1)}),\ldots$$

is contained in $\Omega(h_{k+1},\lambda)$. It follows from the consistency of Φ that for all large enough n

$$(v_1^{(k)},w_1^{(k)}),\ldots,(v_{n_k}^{(k)},w_{n_k}^{(k)}),(x_1^{(k+1)},y_1^{(k+1)}),(x_2^{(k+1)},y_2^{(k+1)}),\ldots,(x_{n-n_k}^{(k+1)},y_{n-n_k}^{(k+1)})$$

satisfies (17), (18), (19) and (20) with j replaced by $k+1$. Select $n_{k+1} > n_k$ so large that the same is true of (21).

As $(\mathbf{v}^{(k+1)}, \mathbf{w}^{(k+1)})$ is an extension of $(\mathbf{v}^{(k)}, \mathbf{w}^{(k)})$, repeating the above process indefinitely yields an infinite sequence $(\mathbf{x}^*, \mathbf{y}^*)$. By construction, the functions $\phi_n(\cdot) = \phi(\cdot : (x_1^*, y_1^*), \ldots, (x_n^*, y_n^*))$ do not converge in $L_2(\lambda)$. Indeed, it follows from (18) and from the inequality $a^2 \geq d^2/5 - b^2 - c^2$ whenever $(a+b+c)^2 = d^2$ that

$$\begin{aligned}
\int |\phi_{n_k}(x) - \phi_{n_l}(x)|^2 \lambda(dx) &\geq \frac{1}{5} \int |\phi_k(x) - \phi_l(x)|^2 \lambda(dx) \\
&- \int |\phi_k(x) - \phi_{n_k}(x)|^2 \lambda(dx) \\
&- \int |\phi_{n_l}(x) - \phi_l(x)|^2 \lambda(dx) \\
&\geq \frac{1}{10} - \frac{1}{40} - \frac{1}{40} \\
&\geq \frac{1}{20}
\end{aligned}$$

whenever $k \neq l$, $k \geq 1$, $l \geq 1$.

It remains to show that the limiting distribution of $\mathbf{x}^*$ is λ and the limiting regression of $(\mathbf{x}^*, \mathbf{y}^*)$ is h_0. To this end, fix $k > 1$ and let $A \subseteq [0,1]$ be an arbitrary interval. It is easily verified that

$$|\nu_k(A) - \nu_0(A)| \leq 2^{-k+1} \leq \frac{2}{k}. \tag{22}$$

Let $\hat{\mu}_n(A)$ and $\hat{\nu}_n(A)$ be evaluated on $((x_1^*, y_1^*), \ldots, (x_n^*, y_n^*))$, and for each $1 \leq r \leq n_{k+1} - n_k$ define

$$\hat{\mu}'_{r,k}(A) = \frac{1}{r} \sum_{j=n_k+1}^{n_k+r} I\{x_i^* \in A\})$$

and

$$\hat{\nu}'_{r,k}(A) = \frac{1}{r} \sum_{j=n_k+1}^{n_k+r} y_i^* I\{x_i^* \in A\}.$$

The equations

$$\begin{aligned}
\hat{\mu}_{n_k+r}(A) &= \frac{n_k}{n_k+r} \cdot \hat{\mu}_{n_k}(A) + \frac{r}{n_k+r} \cdot \hat{\mu}'_{r,k}(A) \\
\hat{\nu}_{n_k+r}(A) &= \frac{n_k}{n_k+r} \cdot \hat{\nu}_{n_k}(A) + \frac{r}{n_k+r} \cdot \hat{\nu}'_{r,k}(A)
\end{aligned}$$

imply the bounds

$$\begin{aligned}
|\hat{\mu}_{n_k+r}(A) - \lambda(A)| &\leq \frac{n_k}{n_k+r} \cdot |\hat{\mu}_{n_k}(A) - \lambda(A)| + \frac{r}{n_k+r} \cdot |\hat{\mu}'_{r,k}(A) - \lambda(A)| \\
&\triangleq I + II
\end{aligned}$$

$$\begin{aligned} |\hat{\nu}_{n_k+r}(A) - \nu_0(A)| &\leq \frac{n_k}{n_k+r} \cdot |\hat{\nu}_{n_k}(A) - \nu_0(A)| + \frac{r}{n_k+r} \cdot |\hat{\nu}'_{r,k}(A) - \nu_0(A)| \\ &\triangleq III + IV. \end{aligned}$$

By virtue of (19), (20) and (22)

$$I \leq |\hat{\mu}_{n_k}(A) - \lambda(A)| \leq \frac{1}{k+1}$$

and

$$III \leq |\hat{\nu}_{n_k}(A) - \nu_k(A)| + |\nu_0(A) - \nu_k(A)| \leq \frac{1}{k+1} + \frac{2}{k}.$$

If $n_{k+1} - n_k \geq r \geq \max(l_{k+1}, \tilde{l}_{k+1})$ then by (15) and (16)

$$\Delta(x^*_{n_k+1}, \ldots, x^*_{n_k+r}) = \Delta(x_1^{(k+1)}, \ldots, x_r^{(k+1)}) \leq \frac{1}{k+2}$$

$$\begin{aligned} \tilde{\Delta}_{k+1}((x^*_{n_k+1}, y^*_{n_k+1}), \ldots, (x^*_{n_k+r}, y^*_{n_k+r})) &= \tilde{\Delta}_{k+1}((x_1^{(k+1)}, y_1^{(k+1)}), \ldots, (x_r^{(k+1)}, y_r^{(k+1)})) \\ &\leq \frac{1}{k+2} \end{aligned}$$

and therefore

$$II \leq |\hat{\mu}'_{r,k}(A) - \lambda(A)| \leq \frac{1}{k+2}$$

$$IV \leq |\hat{\nu}'_{r,k}(A) - \nu_{(k+1)}(A)| + |\nu_{(k+1)}(A) - \nu_0(A)| \leq \frac{1}{k+2} + \frac{2}{k+1}.$$

On the other hand, if $0 < r < \max(l_{k+1}, \tilde{l}_{k+1})$ then (21) implies that

$$\max(II, IV) \leq \frac{2r}{n_k+r} \leq \frac{2r}{kr+r} = \frac{2}{k+1}.$$

These bounds ensure that, since A was an arbitrary interval,

$$\max\left\{\sup_{A\in\mathcal{A}} |\hat{\mu}_n(A) - \lambda(A)| : n_k < n \leq n_{k+1}\right\} \leq \frac{6}{k}$$

$$\max\left\{\sup_{A\in\mathcal{A}} |\hat{\nu}_n(A) - \nu_0(A)| : n_k < n \leq n_{k+1}\right\} \leq \frac{6}{k}$$

and consequently,

$$\lim_{n\to\infty} \sup_{A\in\mathcal{A}} |\hat{\mu}_n(A) - \lambda(A)| = 0$$

and

$$\lim_{n\to\infty} \sup_{A\in\mathcal{A}} |\hat{\nu}_n(A) - \nu_0(A)| = 0.$$

Finally, by (14), for all $t \in \mathbb{R}$

$$\hat{\mu}_n(\{t\}) \to \lambda(\{t\}) = 0 \text{ and } \hat{\nu}_n(\{t\}) \to \nu_0(\{t\}) = 0. \quad \square$$

Acknowledgements

The first author wishes to thank András Antos for helpful discussions.

References

[1] P. H. Algoet. Universal schemes for prediction, gambling and portfolio selection *Ann. Probab.*, 20:901-941, 1992.

[2] D. Bailey. Sequential schemes for classifying and predicting ergodic processes. Ph.D. dissertation, Dept. Math, Stanford University.

[3] G.J. Chaitin. On the length of programs for computing binary sequences. *J. Assoc. Comp. Mach.*, 13:547-569, 1966.

[4] T. M. Cover. Open problems in information theory. in *1975 IEEE Joint Workshop on Information Theory* 35-36 IEEE Press, 1975.

[5] L. Devroye and A. Krzyzak. An equivalence theorem for L_1 convergence of the kernel regression estimate. *J. of Statistical Planning and Inference*, 23:71-82, 1989.

[6] R.M. Gray. *Probability, Random Processes, and Ergodic Properties.* Springer-Verlag, New York, 1988.

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

[8] D. Haussler, J. Kivinen, and M. Warmuth. Tight worst-case loss bounds for predicting with expert advice. *Proc. European Conference on Computational Learning Theory*, 1994.

[9] A.N. Kolmogorov. Three approaches to the quantitative definition of information. *Problems of Information Transmission*, 1: 4-7, 1965.

[10] A.N. Kolmogorov. Logical basis for information theory and probability theory. *IEEE Trans. Info. Theory*, IT-14: 662-664,1968.

[11] A.N. Kolmogorov and S.V. Fomin. *Introductory Real Analysis.* Dover, Mineola, 1970.

[12] S. R. Kulkarni and S. E. Posner. Rates of convergence for nearest neighbor estimation under arbitrary sampling. *IEEE Trans. on Information Theory*, IT-41:1028-1039, 1995.

[13] A.Lempel and J.Ziv. On the complexity of finite sequences. *IEEE Trans. on Information Theory*, IT-22: 75-81, 1976.

[14] N. Merhav, M. Feder, and M. Gutman. Universal prediction of individual sequences. *IEEE Trans. on Information Theory*, IT-38:1258-1270, 1992.

[15] G. Morvai, S. Yakowitz and P. Algoet. Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Information Theory* IT-43:483-498, 1997.

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

[17] A.B. Nobel Limits to classification and regression estimation from ergodic processes. To appear in Ann. Stat. v.27.

[18] A.B. Nobel, G. Morvai, S. Kulkarni. Density estimation from an individual numerical sequence. *IEEE Trans. Information Theory*, 44:537-541, 1998.

[19] D. Ornstein. Guessing the next output of a stationary process. *Israel J. Math.*, 30:292-396, 1974.

[20] D. Pollard. *Convergence of Stochastic Processes*. Springer-Verlag, New York, 1984.

[21] G. Roussas, ed. *Nonparametric functional estimation and related topics.* Kluwer, Netherlands, 1991.

[22] B. Ya. Ryabko. Prediction of random sequences and universal coding. *Problems of Inform. Trans.*, 24: 87-96, 1988.

[23] S. Yakowitz, L. Györfi, J. Kieffer, and G. Morvai (1997). Strongly-consistent nonparametric estimation of smooth regression functions for stationary ergodic sequences. Submitted for publication.

[24] J. Ziv. Coding theorems for individual sequences. *IEEE Trans. on Information Thoery*, IT-24:405-412, 1978.