Limits to Classification and Regression Estimation from Ergodic Processes

Andrew B. Nobel [*]

June 19, 1998

Abstract

We answer two open questions concerning the existence of universal schemes for classification and regression estimation from stationary ergodic processes. It is shown that no measurable procedure can produce weakly consistent regression estimates from every bivariate stationary ergodic process, even if the covariate and response variables are restricted to take values in the unit interval. It is further shown that no measurable procedure can produce weakly consistent classification rules from every bivariate stationary ergodic process for which the response variable is binary valued. The results of the paper are derived via reduction arguments, and are based in part on recent work concerning density estimaton from ergodic processes.

Appears in Annals of Statistics, vol.27, pp.262-273

[*] Andrew Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. Email: nobel@stat.unc.edu. His research was supported in part by NSF Grant DMS-9501926. *AMS 1991 subject classifications.* Primary 62G07; secondary 60G10, 62M99. *Key words and phrases.* Classification, regression, ergodic processes, counterexamples, reduction arguments.

1 Introduction

Nonparametric classification and regression estimation are of fundamental importance in the theory and practice of statistics. Much of the existing theory for these problems is based on the assumption that the available data are independent and identically distributed (i.i.d.). The existence of weakly consistent procedures for classification and regression from any i.i.d. process was first established by Stone (1977) using nearest neighbor methods. The consistency of standard kernel regression estimators for i.i.d. processes was shown by Devroye and Wagner (1980), and Spiegelman and Sacks (1980).

Beginning with the papers of Roussas (1967, 1969) and Rosenblatt (1970), there has been a great deal of work on regression and density estimation from stationary, weakly dependent processes satisfying α, β, ρ, and related mixing conditions. The majority of this work is devoted to central limit theorems and rates of convergence for kernel and histogram type estimates. References and discussion can be found in the monographs of Györfi, Härdle, Sarda and Vieu (1989), and Rosenblatt (1991). Masry (1996) has recently analyzed local polynomial regression from α-mixing processes.

There is also a substantial body of work on regression and density estimation from stationary processes exhibiting long range (also called strong) dependence. For such processes one may obtain different asymptotic behavior than in the weakly dependent case. For an overview of these results and additional references, we refer the reader to Cheng and Robinson (1991), Ho (1995), and Hidalgo (1997).

There is, in addition to the above developments, a growing body of literature devoted to nonparametric estimation from stationary processes that exhibit very strong dependence, or that are assumed only to be ergodic. This work focuses primarily on consistency, as rates of convergence and central limit theorems are typically not available in these general settings.

Strengthening earlier work of Delecroix (1987), Györfi *et al* (1989) showed that for a bi-infinite stationary ergodic process $\{X_i\}_{i=-\infty}^{\infty}$ such that the conditional density of $X_1, \ldots, X_r$ given $X_0, X_{-1}, \ldots$ exists and is continuous for every $r \geq 1$, one can estimate $E(X_r | X_0, \ldots, X_{-k})$ for each $r, k \geq 1$ by means of kernel estimates with a suitable choice of bandwidth. Their estimates, which make use of observations occurring prior to time zero, converge pointwise with probability one. For related results, see also Delecroix and Rosa (1996). Yakowitz (1993) proposed an adaptive nearest neighbor estimate for the autoregression function $h(x) = E(X_1 | X_0 = x)$ of a real-valued process. For Markov chains satisfying a mild recurrence condition, he established in-probability and almost sure consistency of the estimate at continuity points of $h(\cdot)$.

When suitable constraints are placed on the dependence of the observations, few assumptions on the unknown regression or density function are required to ensure the consistency of the method under study. As an alternative, one may consider general ergodic processes, and place constraints instead on the family of candidate regression or density functions under study.

Yakowitz *et al.* (1997) 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 stationary ergodic process whose regression function g^* 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. Morvai, Kulkarni, and Nobel (1997) considered a family of adaptive histogram regression estimates for processes with real-valued covariates. Given positive constants $\alpha_1, \alpha_2, \ldots$, they exhibit a sequence of estimates that is strongly L_2 consistent for every stationary ergodic processes such that (i) the distribution of the covariate variable is non-atomic, and (ii) the variation of the regression function g^* on each interval $[-i, i]$ is less than α_i. In practice, the constants $\alpha_1, \alpha_2, \ldots$ are known and fixed in advance of the data. An analogous result for density estimation is given by Nobel *et al.* (1997).

The strongest positive results concerning nonparametric estimation from ergodic processes have been obtained for the problem of estimating the infinite order autoregression $Z_\infty = E(X_0|X_{-1}, X_{-2}, \ldots)$. Extending work of Ornstein (1978) for finite alphabet processes, Algoet (1992) defined estimates $\hat{Z}_n = \hat{Z}_n(X_{-1}, \ldots, X_{-n})$ such that $\hat{Z}_n \to Z_\infty$ with probability one for every stationary ergodic process $\{X_i\}_{i=-\infty}^{\infty}$ taking values in a bounded interval of the real line. Morvai, Yakowitz, and Györfi (1996) gave a simple proof of this result for a different sequence of estimates. Further work along these lines can be found in Morvai, Yakowitz, and Algoet (1997).

By stationarity, the estimates of Algoet (1992) and Morvai *et al.* (1996) give in-probability consistent estimates of Z_∞ based on observations $X_1, \ldots, X_n$ extending forward in time. In fact, Bailey (1976) and Ryabko (1988) showed that, for observations of this sort, no procedure provides *almost surely* consistent estimates of Z_∞ from every ergodic process. Ryabko (1988) established a similar result concerning estimation of the one-step autoregression $E(X_1|X_0 = x)$ from $X_1, \ldots, X_n$. See also Györfi, Morvai and Yakowitz (1997) for a discussion and proofs of these results.

In spite of the positive results cited above, there are indications that for many nonparametric problems even weakly consistent schemes may not exist when the dependence of the observations is unrestricted. In a result attributed to Shields, it was shown by Györfi *et al.* (1989) that there exist histogram density estimates, consistent for every i.i.d. process, that fail to be consistent for a suitably constructed ergodic process. Györfi and Lugosi (1992) exhibited an ergodic process for which a standard kernel density estimates with bandwidths

$h_n \to 0$ and $nh_n \to \infty$ fail to be consistent. Györfi *et al.* (1997) exhibit histogram estimates of the one-step autoregression function that are consistent for a family of mixing processes, but fail to be consistent for a suitably constructed ergodic process.

While these negative results are suggestive, they leave open the possibility that more elaborate estimates, e.g. estimates that learn or adapt to the mixing structure of the observations, might be consistent for every ergodic process. The positive results for infinite-order auto-regression provide cause for some optimism in this regard. Thus we are led to the following question, versions of which appear in Györfi (1981) and many of the cited references on estimation from ergodic processes:

> Do there exist procedures for density estimation, regression, one-step autoregression, or classification, that are consistent for every stationary ergodic process?

The answer, in each case, is 'No'. The first such result was obtained by Adams and Nobel (1997), who showed that, for every $p \geq 1$, there is no weakly L_p consistent density estimation scheme for ergodic processes with p-integrable marginal densities. Strengthening this result, Adams (1997) showed that for any density estimation procedure there exists a family of isomorphic Bernoulli transformations, one of whose invariant densities the procedure fails to consistently identify. It is shown here that there are no universal schemes for regression estimation or classification from ergodic processes. These results are presented in Theorems 1 and 2, respectively. The common starting point for both conclusions is a simple decision problem for families of ergodic processes, and a modification of the result of Adams and Nobel (1997).

Adams (1997) independently established that no regression procedure is consistent for every bivariate ergodic process. Using renewal methods, Yakowitz and Heyde (1997) have independently shown that there is no weakly consistent procedure for estimating the one-step autoregression $E(X_1|X_0 = x)$ from every ergodic process for which $EX_i^2 < \infty$, and have also established a negative result for density estimation.

Preliminary definitions and the statements of Theorems 1 and 2 are given in the next section. Our method of proof is discussed in Section 2.3. Proofs of the main results are given in Section 3.

2 Statement of Results

In what follows we restrict our attention to bivariate processes taking values in the unit square. For the purposes of exhibiting negative results, this restriction entails no loss of generality.

2.1 Regression Estimation

In regression estimation one observes the initial sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a stationary bivariate process and constructs, based on that data, an estimate $\hat{g}_n$ of the regression function $g^*(x) = E(Y|X = x)$ associated with the marginal distribution of the process. Note that $g^*(\cdot)$ minimizes the expected squared difference $E(g(X) - Y)^2$ over all measurable functions $g(\cdot)$, and is therefore the optimal predictor of Y given X for the squared error loss function. The regression estimation problem and its potential solutions may be formalized as follows.

Let $\{(X_1, Y_1)\}_{i=1}^{\infty}$ be a bivariate stationary ergodic process taking values in $[0,1] \times [0,1]$, with regression function $g^*(x) = E(Y|X = x)$, and such that X_i has distribution μ. A sequence of measurable functions

$$\hat{g}_n : [0,1] \times ([0,1] \times [0,1])^n \to \mathbb{R}, \qquad n = 1, 2, \ldots$$

is a weakly consistent regression scheme for $\{(X_i, Y_i)\}_{i=1}^{\infty}$ if for each $\epsilon > 0$,

$$\mu\{\, x : |\, \hat{g}_n(x : X_1, Y_1, \ldots, X_n, Y_n) \; - \; g^*(x)\,| \; > \; \epsilon \,\} \; \to \; 0 \tag{1}$$

in probability as n tends to infinity. A scheme $\{\hat{g}_n\}$ is weakly consistent for a family $\mathcal{Q}$ of bivariate ergodic processes if it is weakly consistent for every member of $\mathcal{Q}$. When no confusion will arise, reference to the data will be omitted in expressions involving estimates such as $\hat{g}_n$.

Remark: A regression scheme is said to be weakly L_p consistent for a process $\{(X_i, Y_i)\}$ if as n tends to infinity,

$$\int |\, \hat{g}_n(x : X_1, Y_1, \ldots, X_n, Y_n) - g^*(x)\,|^p \, d\mu(x) \to 0 \tag{2}$$

in probability, and strongly L_p consistent if the convergence is with probability one. It follows from Chebyshev's inequality that weak L_p consistency implies (1) if $p \geq 1$.

Theorem 1 *There is no weakly consistent regression scheme for the family $\mathcal{Q}_0$ of bivariate stationary ergodic processes $\{(X_i, Y_i)\}_{i=1}^{\infty}$ such that $X_i, Y_i \in [0,1]$.*

Adams (1997) has independently established a similar result. Our proof shows that it is enough to consider processes for which $Y_i \in \{0,1\}$ is binary valued, and the distribution of X_i is equivalent to Lebesgue measure.

Corollary 1 *If $p \geq 1$ then there is no weakly L_p-consistent regression scheme for the family of stationary ergodic processes $\{(X_i, Y_i)\}_{i=1}^{\infty}$ such that $E|Y|^p < \infty$*

Proof: For each $p \geq 1$ the given family contains $\mathcal{Q}_0$. Thus any L_p consistent procedure for the family would be weakly consistent for $\mathcal{Q}_0$, contradicting Theorem 1.

Yakowitz and Heyde (1997) have independently considered estimation of the one-step autoregression function $h^*(x) = E(X_1|X_0 = x)$ from ergodic processes. Using a Markov chain construction, they have shown that there is no estimation procedure for one-step autoregression that is L_2-consistent for every ergodic processes $\{X_i\}$ such that $EX_i^2 < \infty$. As autoregression is a special case of regression, their result also implies Corollary 1 for values of $p \leq 2$.

2.2 Classification

In classification, the underlying object of interest is a jointly distributed pair (X, Y) in which the response variable Y takes values in the two-point set $\{0, 1\}$. A classification rule is a map $\phi : [0, 1] \to \{0, 1\}$ that assigns a fixed class label to each possible value of X. Each classification rule has probability of error $P\{\phi(X) \neq Y\}$. The minimum probability of error over all classification rules is achieved by the Bayes rule

$$\phi^*(x) = \begin{cases} 0 & \text{if } P(Y = 1|X = x) \leq 1/2 \\ 1 & \text{if } P(Y = 1|X = x) > 1/2 \end{cases} . \tag{3}$$

The value of $\phi^*(x)$ when $P(Y = 1|X = x) = 1/2$ does not affect its probability of error; here such 'ties' are broken arbitrarily in favor of class zero. For a comprehensive treatment of classification and pattern recognition in the context of i.i.d. processes, we refer the reader to the text of Devroye, Györfi, and Lugosi (1996).

Let $\{(X_i, Y_i)\}_{i=1}^{\infty}$ be a stationary ergodic process taking values in $[0, 1] \times \{0, 1\}$ such that (X_1, Y_1) has Bayes rule $\phi^*(x)$, and X_i is distributed according to μ. A sequence of measurable functions

$$\hat{\phi}_n : [0, 1] \times ([0, 1] \times \{0, 1\})^n \to \{0, 1\} \qquad n = 1, 2, \ldots$$

is a weakly consistent classification scheme for $\{(X_i, Y_i)\}$ if

$$\mu\{\, x : \hat{\phi}_n(x : X_1, Y_1, \ldots, X_n, Y_n) \neq g^*(x)\,\} \to 0 \tag{4}$$

in probability as n tends to infinity. A scheme $\{\hat{\phi}_n\}$ is weakly consistent for a family $\mathcal{Q}$ of ergodic processes with binary response variables if it is weakly consistent for every member of $\mathcal{Q}$.

Theorem 2 *There is no weakly consistent classification scheme for the family $\mathcal{Q}_1$ of bivariate stationary ergodic processes $\{(X_i, Y_i)\}_{i=1}^{\infty}$ taking values in $[0, 1] \times \{0, 1\}$.*

Remark: It follows from (3) that a consistent regression scheme $\{\hat{g}_n\}$ for $\mathcal{Q}_0$ can be converted into a consistent classification scheme for $\mathcal{Q}_1$ simply by setting $\hat{\phi}_n(x) = I\{\hat{g}_n(x) > 1/2\}$. Thus Theorem 2 implies Theorem 1. Separate arguments are given below in order to simplify the proofs.

The negative results presented here and in the work of Adams (1997) and Yakowitz and Heyde (1997), lead to several open questions. The following question is posed in several of the references cited above:

> What is the largest family of bivariate stationary ergodic processes for which there exists a consistent regression scheme?

It seems likely that no unique maximal family exists, but that there are many such families, no two of which are comparable. To be more concrete, one may ask the following:

> Is there a consistent regression procedure for the family of all stationary α-mixing processes, where we make no assumptions regarding the rate at which the mixing coefficients go to zero?

Naturally, one may ask analogous questions for other types of mixing. Answers to these questions might shed new light on the comparative strengths of different mixing conditions, and on the necessity of assumptions concerning mixing rates in order to obtain positive results.

2.3 Method of Proof

Let $\mathcal{P}_0$ and $\mathcal{P}_1$ be two disjoint families of stationary ergodic processes taking values in $[0,1]$, where different processes may be defined on different underlying probability spaces. The decision problem for $(\mathcal{P}_0, \mathcal{P}_1)$ is described as follows. At the outset, Player I is given complete information regarding the joint distributions of each process in $\mathcal{P}_0 \cup \mathcal{P}_1$. Then Player II selects a process $\{W_i\} \in \mathcal{P}_0 \cup \mathcal{P}_1$ that is unknown to Player I. For each $n \geq 1$ Player I is presented with the initial sequence $W_1, \ldots, W_n$ of the process chosen by Player II, and asked to decide whether $\{W_i\}$ belongs to $\mathcal{P}_0$ or to $\mathcal{P}_1$.

A decision procedure Ψ is a sequence of measurable functions $\hat{\psi}_n : [0,1)^n \to \{0,1\}$ for $n \geq 1$. Ψ is consistent for $(\mathcal{P}_0, \mathcal{P}_1)$ if for each $j \in \{0,1\}$ and each $\{W_i\} \in \mathcal{P}_j$,

$$\mathbb{P}\{\,\hat{\psi}_n(W_1, \ldots, W_n) \,=\, j\} \,\to\, 1 \quad \text{as} \quad n \to \infty.$$

Here $\mathbb{P}$ is the probability measure governing the process $\{W_i\}$. One may show (c.f. Barron (1985) and Adams and Nobel (1997)) that if $\mathcal{P}_0$ and $\mathcal{P}_1$ are countable, and if any two processes in their union differ in some k-dimensional distribution, then there is a consistent decision procedure Ψ for $(\mathcal{P}_0, \mathcal{P}_1)$.

For $k \geq 1$ let π_k be the partition of $[0, 1)$ whose cells $C_{j,k} = [j2^{-k}, (j+1)2^{-k})$, $j = 0, \ldots, 2^k - 1$, are dyadic intervals of order k. Define sets

$$A_k = \bigcup_{j=0}^{2^{(k-1)}-1} C_{2j,k} \qquad B_k = \bigcup_{j=0}^{2^{(k-1)}-1} C_{2j+1,k}$$

containing alternating even and odd cells of π_k. Note that $A_k \cap B_k = \emptyset$ and that $\lambda(A_k) = \lambda(B_k) = 1/2$, where λ denotes Lebesgue measure on $[0, 1)$ equipped with its Borel subsets $\mathcal{B}$. Let $P_0 \equiv 1$ be the uniform distribution on $[0, 1)$, and for each $s \geq 12$ let P_s be the distribution on $[0, 1)$ having density $2I_{A_k}$ if $s = 2k$, and density $2I_{B_k}$ if $s = 2k + 1$. (The indexing ensures that $k \geq 6$.) Let $\mathcal{P}_0^*$ be the family of all stationary ergodic processes $\{W_i\}$ taking values in $[0, 1)$ such that $W_1 \sim P_0$. Let $\mathcal{P}_1^*$ be the family of all stationary ergodic processes $\{W_i\}$ taking values in $[0, 1)$ such that $W_1 \sim P_s$ for some $s \geq 12$.

Using a cutting and stacking argument, Adams and Nobel (1997) showed that there is no weakly consistent density estimation scheme for the family $\mathcal{P}_0^* \cup \mathcal{P}_1^*$. By a routine modification of their argument, one may establish the following basic result. As the essential features of the proof are unchanged, we omit the details, and refer the interested reader to Theorem 1 of Adams and Nobel (1997).

Theorem A *There is no consistent decision procedure for* $(\mathcal{P}_0^*, \mathcal{P}_1^*)$.

Many consistency results rely, in a direct or indirect way, on information about the rate at which the average of a function, applied to the given observations, converges to its expected value. The ergodic theorem ensures the asymptotic convergence of averages to expected values under very weak conditions. However, without assumptions about the observations, one cannot say, even for bounded functions, how fast this convergence takes place. This possibility of arbitrarily slow convergence lies behind the failure of any procedure to distinguish between $\mathcal{P}_0^*$ and $\mathcal{P}_1^*$. In particular, if $\Psi = \{\psi_n : n \geq 1\}$ succeeds in identifying each process in $\mathcal{P}_1^*$, then one may construct a 'misleading' process $\{W_i\} \in \mathcal{P}_0^*$ that cannot be consistently identified by Ψ. When applied to $\{W_i\}$, the procedure decides, infinitely often, that it is viewing different processes in $\mathcal{P}_1^*$. In fact, these processes look more and more like a process in $\mathcal{P}_0^*$, but $\{W_i\}$ reveals its marginal distribution more slowly than Ψ makes decisions.

The proof of Theorem 1 shows that any weakly consistent regression scheme for the family $\mathcal{Q}_0$ can be converted into a consistent decision procedure for for $(\mathcal{P}_0^*, \mathcal{P}_1^*)$. This is accomplished in part by converting the observed sequence $W_1, \ldots, W_n$ into the initial sequence $(X_1, Y_1), \ldots, (X_n, Y_n)$ of a process in $\mathcal{Q}_0$. It then follows from Theorem A that there can be no weakly consistent regression scheme for $\mathcal{Q}_0$. Theorem 2 is proved in a similar fashion.

3 Proofs of Theorems

Let X be a random variable defined on $([0,1), \mathcal{B})$ that maps each of the intervals $[0, \frac{1}{2})$, $[\frac{1}{2}, \frac{3}{4})$, $[\frac{3}{4}, \frac{7}{8})$, $\ldots$ onto $[0,1)$ in an affine fashion:

$$X(w) = \sum_{k=0}^{\infty} 2^{k+1} \left(w - \frac{2^k - 1}{2^k} \right) I \left\{ \frac{2^k - 1}{2^k} \leq w < \frac{2^{k+1} - 1}{2^{k+1}} \right\}.$$

For constants $\alpha, \beta \in \mathbb{R}$ and $A \subseteq [0,1)$, define $\alpha + \beta A = \{\alpha + \beta x : x \in A\}$. It follows from the definition of X that for every $A \subseteq [0,1)$,

$$X^{-1}A \;=\; \bigcup_{j=1}^{\infty} ((1 - 2^{-j}) + 2^{-(j+1)} A) \;=\; \frac{1}{2} A \cup \left(\frac{1}{2} + \frac{1}{4} A \right) \cup \left(\frac{3}{4} + \frac{1}{8} A \right) \cup \cdots \qquad (5)$$

Let μ_s be the distribution of X under P_s so that $\mu_s(A) = P_s(X^{-1}A)$. Let λ denote Lebesgue measure on $[0,1)$. One may readily verify X is uniformly distributed under P_0, so that $\mu_0 = \lambda$.

Lemma 1 *For $s = 0, 12, 13, \ldots$ each of the following hold.*

(a) $\mu_s(A) \leq 2\lambda(A)$ for each $A \in \mathcal{B}$

(b) μ_s dominates λ, i.e. for each $A \in \mathcal{B}$ $\mu_s(A) = 0$ implies $\lambda(A) = 0$.

Proof: Both statements are clear if $s = 0$, so let $s \geq 12$. Note that $P_s(A) \leq 2\lambda(A)$ for every $A \in \mathcal{B}$ since the density of P_s is at most 2. Therefore (5) implies that

$$\begin{aligned}
\mu_s(A) \;&=\; P_s\left(\frac{1}{2}A\right) + P_s\left(\frac{1}{2} + \frac{1}{4}A\right) + P_s\left(\frac{3}{4} + \frac{1}{8}A\right) + \cdots \\
&\leq\; 2\lambda\left(\frac{1}{2}A\right) + 2\lambda\left(\frac{1}{2} + \frac{1}{4}A\right) + 2\lambda\left(\frac{3}{4} + \frac{1}{8}A\right) + \cdots \\
&=\; \lambda(A) + \frac{1}{2}\lambda(A) + \frac{1}{4}\lambda(A) + \cdots \;=\; 2\lambda(A)\,,
\end{aligned}$$

which establishes (a). Suppose that $s \geq 12$ is even, so that P_s has density $2I_{A_k}$. Then since X maps the last interval of A_k onto $[0,1)$, one has $\mu_s(A) \geq 2^{-(k-1)}\lambda(A)$ for each Borel set A, and (b) follows. A similar inequality holds if $s \geq 12$ is odd. □

Proof of Theorem 1: Define the random variable $Y(\omega) = I\{0 \leq \omega < 1/2\}$ on $([0,1), \mathcal{B})$. For each $s = 0, 12, 13, \ldots$ let $g_s(x) = P_s(Y = 1 \,|\, X = x)$ be the conditional expectation of

Y given X under P_s. By definition, $g_s(x)$ satisfies

$$\int_{X^{-1}A} Y(\omega) dP_s(\omega) = \int_A g_s(x) d\mu_s(x) \tag{6}$$

for every set $A \in \mathcal{B}$ (c.f. Ash (1972)). Using (6) and (5), one may verify that $\int_A g_0(x)dx = \frac{1}{2}\lambda(A)$ for every $A \in \mathcal{B}$, which implies that $g_0(x) = 1/2$ for λ-almost every $x \in [0,1)$.

Fix $s \geq 12$, and assume for the moment that s is even, so that μ_s has density $2I_{A_k}$ for some $k \geq 6$. As Y is non-negative, equation (6) with $A = \{g_s < 0\}$ shows that $g(x) \geq 0$ for μ_s-almost every $x \in [0,1)$. Since $X^{-1}B_{k-1} \subseteq (B_k \cap [0, \frac{1}{2})) \cup [\frac{1}{2}, 1)$,

$$\int_{B_{k-1}} g_s d\mu_s \leq \int_{B_k \cap [0,\frac{1}{2})} Y dP_s + \int_{[\frac{1}{2},1)} Y dP_s \ = \ P_s(B_k \cap [0, \tfrac{1}{2})) \ = \ 0\,,$$

and therefore $g_s(x) = 0$ for μ_s-almost every $x \in B_{k-1}$. Lemma 1(b) implies that $g_s(x) = 0$ for λ-almost every $x \in B_{k-1}$, which in turn implies that

$$\lambda\{g_s = 0\} \ \geq \ \lambda(B_{k-1}) \ = \ \frac{1}{2}\,. \tag{7}$$

The same inequality can be established when s is odd by reversing the roles of A_k and B_k.

Let $\mathcal{P} = \mathcal{P}_0 \cup \mathcal{P}_1$ and let $\mathcal{Q}$ be the family of processes $\{(X_i, Y_i)\}_{i=1}^{\infty}$ such that $X_i = X(W_i)$ and $Y_i = Y(W_i)$ for some $\{W_i\} \in \mathcal{P}$. Note that each element of $\mathcal{Q}$ is ergodic and that $\mathcal{Q} \subseteq \mathcal{Q}_0$. Assume by way of contradiction that $\{\hat{g}_n\}$ is a weakly consistent regression scheme for $\mathcal{Q}$. Using $\{\hat{g}_n\}$ one may construct a decision procedure for $(\mathcal{P}_0, \mathcal{P}_1)$ as follows. Given the initial sequence $W_1, \ldots, W_n$ of a process in $\mathcal{P}$, form $X_i = X(W_i)$ and $Y_i = Y(W_i)$ for $i = 1, \ldots, n$, and let

$$\hat{\psi}_n(W_1, \ldots, W_n) = \begin{cases} 0 & \text{if } \ \lambda\{x : \hat{g}_n(x : X_1, Y_1, \ldots, X_n, Y_n) \leq 1/4\} \ \leq \ 1/4 \\ 1 & \text{otherwise} \end{cases}\,.$$

Note that the necessity of considering Lebesgue measure arises from the fact that the distribution μ_s of X_i is not known when implementing the rule $\hat{\psi}_n$.

Let $\{W_i\}$ be an element of $\mathcal{P}$ with $W_1 \sim P_s$, and let $\{(X_i, Y_i)\}$ be the process it generates in $\mathcal{Q}$. The weak consistency of $\{\hat{g}_n\}$ implies that $\mu_s\{|\hat{g}_n - g_s| \geq 1/4\}$ tends to zero in probability as $n \to \infty$, and by Lemma 1(b) the same is true of $\lambda\{|\hat{g}_n - g_s| \geq 1/4\}$. Thus when $s = 0$,

$$\lambda\{\hat{g}_n \leq 1/4\} \ \leq \ \lambda\{|\hat{g}_n - 1/2| \geq 1/4\} \ = \ \lambda\{|\hat{g}_n - g_0| \geq 1/4\}\,,$$

which tends in probability to zero. On the other hand, when $s \geq 12$, the inequality (7) implies that

$$\begin{aligned} \lambda\{\hat{g}_n \leq 1/4\} \ &\geq \ \lambda\{\hat{g}_n \leq 1/4,\, g_s = 0\} \ \geq \ \lambda\{|\hat{g}_n - g_s| \leq 1/4\} - \lambda\{g_s = 0\} \\ &\geq \ \lambda\{|\hat{g}_n - g_s| \leq 1/4\} - \frac{1}{2} \end{aligned}$$

which tends in probability to one half. Thus $\{\hat{\psi}_n\}$ is a consistent decision procedure for $(\mathcal{P}_0, \mathcal{P}_1)$. This contradicts Theorem A and completes the proof. $\Box$

Proof of Theorem 2: Let X be defined on $([0,1), \mathcal{B})$ as above and define $Y(\omega) = I\left\{\frac{1}{2} \leq \omega < \frac{31}{32}\right\}$. Note that for $k \geq 6$, each of the intervals $[\frac{1}{2}, \frac{3}{4})$, $[\frac{3}{4}, \frac{7}{8})$, $[\frac{7}{8}, \frac{15}{16})$, and $[\frac{15}{16}, \frac{31}{32})$ comprising the support of Y is a disjoint union of two or more cells of π_k. For each $s = 0, 12, 13, \ldots$ let $\eta_s(x) = P_s(Y = 1 | X = x)$ be the regression function of (X, Y) under P_s with associated Bayes rule

$$\phi_s^*(x) = \begin{cases} 0 & \text{if } \eta_s(x) \leq 1/2 \\ 1 & \text{if } \eta_s(x) > 1/2 \end{cases} .$$

As $P_0 = \mu_0 = \lambda$, equation (6) implies that for every Borel set A,

$$\begin{aligned} \int_A \eta_0(x) dx &= \int_{X^{-1}(A)} Y(w) dw \\ &= \lambda\left(\frac{1}{2} + \frac{1}{4}A\right) + \cdots + \lambda\left(\frac{15}{16} + \frac{1}{32}A\right) \\ &= \lambda(A)\left(\frac{1}{4} + \frac{1}{8} + \frac{1}{16} + \frac{1}{32}\right) \;=\; \frac{15}{32}\lambda(A) . \end{aligned}$$

Therefore $\eta_0(x) = 15/32$, and $\phi_0^*(x) = 0$ for λ-almost every $x \in [0,1)$.

For $s \geq 12$ each of the intervals $[\frac{1}{2}, \frac{3}{4})$, $\ldots$, $[\frac{15}{16}, \frac{31}{32})$ has the same probability under $P_s(\cdot)$ and $\lambda(\cdot)$, and consequently

$$P_s\{Y = 1\} = \int \eta_s(x) d\mu_s(x) = \frac{15}{32} . \tag{8}$$

Suppose that s is even, so that $P_s = 2I_{A_k}$ for some $k \geq 6$. We seek a subset of $[0,1)$ whose measure is relatively large, but over which the integral of η_s is relatively small compared to 1/2. It is readily verified that $2^{-r}B_l = B_{l+r} \cap [0, 2^{-r})$ for each $r \geq 1$, and that $\lambda(C \cap B_r) = \frac{1}{2}\lambda(C)$ if $C \in \pi_l$ with $l < r$. Consider B_{k-2}. By (5),

$$\mu_s(B_{k-2}) \;=\; P_s(X \in B_{k-2}) \;=\; \sum_{j=0}^{\infty} P_s(\,(1 - 2^{-j}) \;+\; 2^{-(j+1)}B_{k-2}\,).$$

The $j = 0$ term in the above sum is

$$P_s(B_{k-1} \cap [0, 1/2) \;=\; 2\int_0^{1/2} I_{A_k} I_{B_{k-1}} dx \;=\; \int_0^1 I_{A_k} I_{B_{k-1}} dx \;=\; \frac{1}{2}\lambda(B_{k-1}) \;=\; 1/4 .$$

The $j = 1$ term is $P_s(B_k \cap [1/2, 1))$, which is zero as $A_k \cap B_k = \emptyset$. The $j = 2$ term can be evaluated as follows:

$$P_s(3/4 + B_{k+1} \cap [0, 1/8)) \;\;=\;\; P_s(B_{k+1} \cap [3/4, 7/8))$$

$$
\begin{aligned}
&= 2\int_{3/4}^{7/8} I_{A_k} I_{B_{k+1}} dx \\
&= 2 \sum_{C \in \pi_k} I\{C \subseteq A_k \cap [3/4, 7/8)\} \cdot \int I_C I_{B_{k+1}} dx \\
&= \sum_{C \in \pi_k} I\{C \subseteq A_k \cap [3/4, 7/8)\} \cdot \lambda(C) \\
&= \frac{1}{2}\lambda([3/4, 7/8)) \;=\; \frac{1}{16}.
\end{aligned}
$$

Similar arguments show that the remaining terms in the sum are $1/32$, $1/64$, ... and therefore $\mu_s(B_{k-2}) = 3/8$. In conjunction with (6) these calculations also show that

$$
\begin{aligned}
\int_{B_{k-2}} \eta_s(x) d\mu_s(x) \;&=\; \int_{X^{-1}B_{k-2}} Y(w) dP_s(w) \;=\; \sum_{j=1}^{4} P_s(\,(1-2^{-j}) \,+\, 2^{-(j+1)} B_{k-2}\,) \\
&= \frac{1}{16} + \frac{1}{32} = \frac{3}{32}\,.
\end{aligned}
$$

Combining these calculations with equation (8) gives

$$
\begin{aligned}
\frac{15}{32} &= \int \eta_s(x) d\mu_s(x) \\
&= \int_{B_{k-2}} \eta_s(x) d\mu_s(x) + \int_{B_{k-2}^c} \eta_s(x) d\mu_s(x) \\
&\leq \int_{B_{k-2}} \eta_s(x) d\mu_s(x) + \frac{1}{2}\mu_s(B_{k-2}^c) + \mu_s\{\eta_s > 1/2\} \\
&= \frac{13}{32} + \mu_s\{\eta_s > 1/2\}\,.
\end{aligned}
$$

By the above and Lemma 1(a),

$$
\frac{1}{32} \;\leq\; \frac{1}{2}\mu_s\{\eta_s > 1/2\} \;\leq\; \lambda\{\eta_s > 1/2\} \;=\; \lambda\{\phi_s^* = 1\}\,. \tag{9}
$$

A similar argument shows that the same inequality holds when $s \geq 12$ is odd.

Now we proceed as in the proof of Theorem 1. Let $\mathcal{P}$ and $\mathcal{Q}$ be defined as before using $Y(\omega) = I\{\omega \in [1/2, 31/32)\}$, and assume that $\{\hat{\phi}_n\}$ is a weakly consistent classification scheme for $\mathcal{Q}$. Given the initial sequence $W_1, \ldots, W_n$ of a process in $\mathcal{P}$, form $X_i = X(W_i)$ and $Y_i = Y(W_i)$, and define

$$
\hat{\psi}_n(W_1, \ldots, W_n) = \begin{cases} 0 & \text{if } \lambda\{\, x : \hat{\phi}_n(x : X_1, Y_1, \ldots, X_n, Y_n) \,=\, 1\,\} \,\leq\, 1/64 \\ 1 & \text{otherwise} \end{cases}\,.
$$

If $\{W_i\} \in \mathcal{P}_0$ then the consistency of $\{\hat{\phi}_n\}$ implies that $\lambda\{\hat{\phi}_n = 1\} \;=\; \mu_0\{\hat{\phi}_n \neq \phi_0^*\}$ tends to zero in probability, so that the same is true of $\hat{\psi}_n$. If $\{W_i\} \in \mathcal{P}_1$ with $W_1 \sim P_s$, then the consistency of $\{\hat{\phi}_n\}$ implies that $\mu_s\{\hat{\phi}_n \neq \phi_s^*\} \to 0$ in probability. Consequently $\lambda\{\hat{\phi}_n \neq \phi_s^*\} \to 0$ by Lemma 1(a), and $\hat{\psi}_n \to 1$ in probability by virtue of (9). This completes the proof. □

Acknowledgments

The author wishes to thank Terrence Adams for many useful discussions, and his helpful comments and suggestions.

References

[1] Adams, T.M. (1997) Families of ergodic processes without consistent density or regression estimates. Preprint.

[2] Adams, T.M. and Nobel, A.B. (1997). On density estimation from ergodic processes. To appear in *Ann. Probab.*.

[3] Algoet, P. (1992). Universal schemes for prediction, gambling and portfolio selection. *Ann. Probab.*, 20 901-941. Correction: *ibid*, 23 474-478, 1995.

[4] Ash, R. (1972). *Real Analysis and Probability.* Academic Press, New York.

[5] Bailey, D.H. (1976). *Sequential Schemes for Classifying and Predicting Ergodic Processes.* Ph.D. Thesis, Stanford University, Dept. of Mathematics.

[6] A.R. Barron (1985). Logically smooth density estimation. Technical Report TR 56, Department of Statistics, Stanford University.

[7] Cheng, B.C. and Robinson, P.M. (1991) Density estimation in strongly dependent non-linear time series. *Stat. Sinica*, 1 335-359.

[8] Delecroix, M. (1987). *Sur l'estimation et la prévision non-paramétrique des processus ergodiques.* Ph.D. Thesis, University of Lille Flandres Artois, Lille, France.

[9] Delecroix, M. and Rosa A.C. (1996). Nonparametric estimation of a regression function and its derivatives under an ergodic hypothesis. *Nonparametric Stat.*, 6 367-382.

[10] Devroye, L. and Wagner, T. (1980). Distribution-free consistency results in nonparametric discrimination and regression function estimation. *Ann. Stat.*, 8 231-239.

[11] Devroye, L., Györfi, L., and Lugosi, G. (1996). *A Probabalistic Theory of Pattern Recognition.* Springer, New York.

[12] Györfi, L. (1981). Strongly consistent density estimate from ergodic sample. *J. Multivariate Analysis*, 11 81-84.

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

[14] Györfi, L. and Lugosi, G. (1992). Kernel density estimation from ergodic sample is not universally consistent. *Comput. Stat. Data Anal.*, 14 437-442.

[15] Györfi, L., Morvai, G., and Yakowitz, S. (1998). Limits to consistent on-line forecasting for ergodic time series. *IEEE Trans. Info. Theory*, 44 886-892.

[16] Hidalgo, J. (1997) Non-parametric estimation with strongly dependent time multivariate time series. *J. Time Sers. Anal.*, 18(2) 95-122.

[17] Ho, H.C. (1995). On the strong uniform consistency of density estimation for strongly dependent sequences. *Stat. and Prob. Letters*, 22 149-156.

[18] Masry, E. (1996). Multivariate local polynomial regression for time series: uniform strong consistency and rates. *J. Time Sers. Anal.*, 17 571-599.

[19] Morvai, G., Yakowitz, S., and Györfi, L. (1996). Nonparametric inference for ergodic, stationary time series. *Ann. Stat.*, 24(1) 370-379.

[20] Morvai, G., Yakowitz, S., and Algoet, P. (1997). Weakly convergent nonparametric forecasting of stationary time series. *IEEE Trans. Info. Theory*, 43(2) 483-498.

[21] Morvai, G., Kulkarni, S., and Nobel, A.B. (1997). Regression estimation from an individual stationary sequence. Submitted for publication.

[22] Nobel, A.B., Morvai, G., and Kulkarni, S. (1997). Density estimation from an individual numerical sequence. *IEEE Trans. Info. Theory*, 44 537-541.

[23] Ornstein, D.S. (1978). Guessing the next output of a stationary process *Israel J. Math*, 30 292-296.

[24] Rosenblatt, M. (1970). Density estimates and Markov sequences. In *Nonparametric Techniques in Statistical Inference*, M. Puri editor, Cambridge Univ. Press, London, 199-210.

[25] Rosenblatt, M. (1991). *Stochastic Curve Estimation.* NSF-CBMS Regional Conference Series in Probability and Statistics, Institute of Mathematical Statistics, Hayward, CA..

[26] Roussas, G. (1967). Nonparametric estimation in Markov processes. *Ann. Inst. Statist. Math.*, 21 73-87.

[27] Roussas, G. (1969). Nonparametric estimation of the transition distribution function of a Markov process. *Ann. Math. Stat.*, 40 1386-1400.

[28] Ryabko, B. Ya. (1988). Prediction of random sequences and universal coding. *Problems of Info. Trans.*, 24 87-96.

[29] Spiegelman, C. and Sacks, J. (1980). Consistent window estimation in nonparametric regression. *Ann. Stat.*, 8 240-246.

[30] Stone, C. (1977) Consistent nonparametric regression. *Ann. Stat.*, 5 595-620.

[31] Yakowitz, S. (1993). Nearest neighbor regression estimation for null-recurrent Markov time series. *Stoc. Proc. Appl.*, 48 311-318.

[32] Yakowitz, S., Györfi, L., Kieffer, J., and Morvai, G. (1997). Strongly-consistent nonparametric estimation of smooth regression functions for stationary ergodic sequences. Under revision, *J. Multivar. Anal.*.

[33] Yakowitz, S. and Heyde, C. (1997). Long range dependency effects with implications for forecasting and queueing inference. Submitted for publication.