Weakly Convergent Nonparametric Forecasting of Stationary Time Series

Gusztáv Morvai, Sidney Yakowitz and Paul Algoet

IEEE Transactions on Information Theory Vol. 43, pp. 483-498, 1997.

Abstract

The conditional distribution of the next outcome given the infinite past of a stationary process can be inferred from finite but growing segments of the past. Several schemes are known for constructing pointwise consistent estimates, but they all demand prohibitive amounts of input data. In this paper we consider real-valued time series and construct conditional distribution estimates that make much more efficient use of the input data. The estimates are consistent in a weak sense, and the question whether they are pointwise consistent is still open. For finite-alphabet processes one may rely on a universal data compression scheme like the Lempel-Ziv algorithm to construct conditional probability mass function estimates that are consistent in expected information divergence. Consistency in this strong sense cannot be attained in a universal sense for all stationary processes with values in an infinite alphabet, but weak consistency can. Some applications of the estimates to on-line forecasting, regression and classification are discussed.

I. Introduction and Overview

We are motivated by some fundamental questions regarding inference of time series that were raised by T. Cover [9] and concerning which significant progress has been made during the intervening years. The time series is a stationary process $\{X_t\}$ with values in a set $\mathcal{X}$ which may be a finite set, the real line, or a finite dimensional euclidean space. For $t \geq 0$ let $X^t = (X_0, X_1, \ldots, X_{t-1})$ denote the t-past at time t. It is also convenient to consider the outcome $X = X_0$, the t-past $X^{-t} = (X_{-t}, \ldots, X_{-1})$ and the infinite past $X^- = (\ldots, X_{-2}, X_{-1})$ at time 0. The true process distribution P is unknown a priori but is known to fall in the class $\mathcal{P}_s$ of stationary distributions on the sequence space $\mathcal{X}^{\mathcal{Z}}$.

Cover's list of questions included the following: given that $\{X_t\}$ is a $\{0,1\}$-valued time series with an unknown stationary ergodic distribution P, is it possible to infer estimates $\hat{P}\{X_t = 1|X^t\}$ of the conditional probabilities $P\{X_t = 1|X^t\}$ from the past X^t such that

$$[\hat{P}\{X_t = 1|X^t\} - P\{X_t = 1|X^t\}] \to 0 \quad P\text{-almost surely as } t \to \infty? \tag{1}$$

D. Bailey [5] used the cutting and stacking technique of ergodic theory to prove that the answer is negative. A simple proof of this negative result is outlined in Proposition 3 of Ryabco [30]. Bailey [5] also discussed a result of Ornstein [22] that provides a positive answer to a less demanding question of Cover [9], namely whether there exist estimates $\hat{P}\{X = 1|X^{-t}\}$ based on the past X^{-t} such that for all $P \in \mathcal{P}_s$,

$$\hat{P}\{X = 1|X^{-t}\} \to P\{X = 1|X^-\} \quad P\text{-almost surely as } t \to \infty. \tag{2}$$

Ornstein constructed estimates $\hat{P}_k\{X = 1|X^{-\lambda(k)}\}$ which depend on finite past segments $X^{-\lambda(k)} = (X_{-\lambda(k)}, \ldots, X_{-1})$ and which converge almost surely to $P\{X = 1|X^-\}$ for every $P \in \mathcal{P}_s$. The length $\lambda(k)$ of the data record $X^{-\lambda(k)}$ depends on the data itself, i.e. $\lambda(k)$ is a stopping time adapted to the filtration $\{\sigma(X^{-t}) : t \geq 0\}$. To get estimates satisfying (2), simply define $\hat{P}\{X = 1|X^{-t}\}$ as the estimate $\hat{P}_k\{X = 1|X^{-\lambda(k)}\}$ where k is the largest integer such that $\hat{P}_k\{X = 1|X^{-\lambda(k)}\}$ can be evaluated from the data X^{-t} (that is, $X^{-\lambda(k)}$ is a suffix of the string X^{-t} but $X^{-\lambda(k+1)}$ is not.) The true conditional probability $P\{X = 1|X^{-t}\}$ converges to $P\{X = 1|X^-\}$ almost surely by the martingale convergence theorem and the estimate $\hat{P}\{X = 1|X^{-t}\}$ converges to the same limit, hence

$$[\hat{P}\{X = 1|X^{-t}\} - P\{X = 1|X^{-t}\}] \to 0 \quad P\text{-almost surely and in } L^1(P). \tag{3}$$

An on-line estimate $\hat{P}\{X_t = 1|X^t\}$ can be constructed at time t from the past X^t in the same way as $\hat{P}\{X = 1|X^{-t}\}$ was constructed from X^{-t}. By (3) and stationarity

$$[\hat{P}\{X_t = 1|X^t\} - P\{X_t = 1|X^t\}] \to 0 \quad \text{in } L^1(P) \text{ as } t \to \infty. \tag{4}$$

Thus the guessing scheme $\hat{P}\{X_t = 1|X^t\}$ is universally consistent in the weak sense of (4), although no guessing scheme can be universally consistent in the pointwise sense of (1).

Ornstein's result can be generalized when $\{X_t\}$ is a stationary process with values in a complete separable metric (Polish) space $\mathcal{X}$. Algoet [1] constructed estimates $\hat{P}_k(dx|X^{-\lambda(k)})$ that, with probability one under any $P \in \mathcal{P}_s$, converge in law to the true conditional distribution $P(dx|X^-)$ of $X = X_0$ given the infinite past. By setting $\hat{P}(dx|X^{-t}) = \hat{P}_k(dx|X^{-\lambda(k)})$ for $\lambda(k) \leq t < \lambda(k+1)$, one obtains estimates $\hat{P}(dx|X^{-t})$ that almost surely converge in law to the random measure $P(dx|X^-)$ in the space of probability distributions on $\mathcal{X}$. Thus for any bounded continuous function $h(x)$ and any stationary distribution $P \in \mathcal{P}_s$,

$$\int h(x)\, \hat{P}(dx|X^{-t}) \to \int h(x)\, P(dx|X^-) \quad P\text{-almost surely.} \tag{5}$$

A much simpler estimate $\hat{P}_k(dx|X^{-\lambda(k)})$ and convergence proof were obtained by Morvai, Yakowitz and Györfi [21]. Their estimate $\hat{P}_k\{X \in B|X^{-\lambda(k)}\}$ of the conditional probability of a subset $B \subseteq \mathcal{X}$ has the structure of a sample mean:

$$\hat{P}_k\{X \in B|X^{-\lambda(k)}\} = \frac{1}{k} \sum_{1 \leq i \leq k} 1\{X_{-\tau(i)} \in B\}, \tag{6}$$

where the $X_{-\tau(i)}$ are samples of the process at selected instants in the past and $\lambda(k)$ is the smallest integer t such that the indices $\{\tau(i) : 1 \leq i \leq k\}$ can be inferred from the segment X^{-t}. From careful reading of [21], one can surmise that $\lambda(k)$ will be huge for relatively small values of the sample size k. Morvai [20] applied the ergodic theorem for recurrence times of Ornstein and Weiss [24] and argued that if $\{X_t\}$ is a stationary ergodic finite-alphabet process with positive entropy rate H bits per symbol and C is a constant such that $1 < C < 2^H$, then, with probability one,

$$\lambda(k) \geq C^{C^{\cdot^{\cdot^{C}}}} \quad \text{eventually for large } k, \tag{7}$$

where the height of the exponential tower is $k - k_0$ for some number k_0 that depends on the process realization but not on k. To our knowledge, none of the strongly-consistent methods have been applied to any data sets, real or simulated.

Scarpellini [31] has applied the methods of Bailey [5] and Ornstein [22] to infer the conditional expectation $E\{X_\tau|\{X_s\}_{s\leq 0}\}$ of the outcome X_τ at some fixed time $\tau > 0$ given the infinite past of a stationary real-valued continuous-time process $\{X_t\}$ from past experience. The outcomes X_t are assumed to be bounded in absolute value by some fixed constant K. Scarpellini constructs estimates by averaging samples taken at a finite number of regularly spaced instants in the past and proves that the estimates converge almost surely to the desired limit $E\{X_\tau|\{X_s\}_{s\leq 0}\}$. His generalization of Ornstein's result is not quite straightforward, and the difficulty seems to be caused more by the continuity of the range space $[-K, K]$ than by the continuity of the time index t.

These works are of considerable theoretical interest because they point to the limits of what can be achieved by way of time series prediction. Pointwise consistency can be

attained for all stationary processes, but the estimates are based on enormous data records. It is hard to say how much raw data are really needed to get estimates with reasonable precision. The nonparametric class of all stationary ergodic processes is very rich and can model all sorts of complex nonlinear dynamics with long range dependencies and periodicities at many different time scales. It is hopeless to get efficient estimates with bounds on the convergence rate unless one has a priori information that winnows the range of possibilities to some manageable subclass. In the literature on nonparametric estimation (e.g. see Györfi, Härdle, Sarda and Vieu [15] and also Marton and Shields [19]), one imposes mixing conditions on the time series and then finds that the standard methods are consistent and achieve stated asymptotic rates of convergence. These approaches are preferable to the universal methods when one is assured of the mixing hypotheses. On the other hand, there is essentially no methodology for testing for mixing.

In the present study we relax the strong consistency requirement and push in the direction of greater efficiency. Rather than demanding strong consistency or pointwise convergence in (5), we shall be satisfied with weak consistency or mean convergence in $L^1(P)$. (Note that mean convergence is equivalent to convergence in probability because the random variables are uniformly bounded.) Being more tolerant in this way enables us to significantly reduce the data demands of the algorithm. The estimates will again be defined as empirical averages of sample values, but the length of the raw data segment that must be inspected to collect a given number of samples will grow only polynomially fast in the sample size (when $\mathcal{X}$ is a finite alphabet), rather than as a tower of exponentials in (7).

For processes with values in a finite set $\mathcal{X}$, weak consistency means that for any stationary distribution P on $\mathcal{X}^{\mathcal{Z}}$ and any $x \in \mathcal{X}$, the estimate $\hat{P}(x|X^{-t}) = \hat{P}\{X = x|X^{-t}\}$ will converge in mean to the true conditional probability $P(x|X^-) = P\{X = x|X^-\}$:

$$\hat{P}(x|X^{-t}) \to P(x|X^-) \quad \text{in } L^1(P), \text{ for any } x \in \mathcal{X}. \tag{8}$$

There exist estimates that are universally consistent in a stronger sense. Given a universal data compression algorithm or a universal parsimonious modeling scheme for stationary processes with values in the finite alphabet $\mathcal{X}$, we shall design estimates $\hat{P}(x|X^{-t})$ that are consistent in expected information divergence for all stationary P. The expectation of the Kullback-Leibler divergence between the conditional probability mass function $P(x|X^-)$ and the estimate $\hat{P}(x|X^{-t})$ will vanish in the limit as $t \to \infty$ for all $P \in \mathcal{P}_s$:

$$E_P\{I(P_{X|X^-}|\hat{P}_{X|X^{-t}})\} \to 0, \tag{9}$$

where

$$I(P_{X|X^-}|\hat{P}_{X|X^{-t}}) = \sum_{x \in \mathcal{X}} P(x|X^-) \log \left(\frac{P(x|X^-)}{\hat{P}(x|X^{-t})} \right). \tag{10}$$

Consistency in expected information divergence implies consistency in mean as in (8), and is equivalent to the requirement that for any stationary $P \in \mathcal{P}_s$ we have mean convergence

$$\log \hat{P}(X|X^{-t}) \to \log P(X|X^-) \quad \text{in } L^1(P). \tag{11}$$

The constructions of Ornstein [22] and Morvai, Yakowitz and Györfi [21] yield estimates $\hat{P}(x|X^{-t})$ such that (11) holds universally in the pointwise sense, but perhaps not in mean.

No estimates $\hat{P}(x|X^{-t})$ can be consistent in expected information divergence for all stationary processes with values in a countable infinite alphabet, but weak consistency as in (8) is universally achievable. Barron, Györfi and van der Meulen [7] consider an unknown distribution $P(dx)$ on an abstract measurable space $\mathcal{X}$ and construct estimates from independent samples so that the estimates are consistent in information divergence and in expected information divergence whenever $P(dx)$ has finite Kullback-Leibler divergence $I(P|M) < \infty$ relative to some known probability distribution $M(dx)$ on $\mathcal{X}$. In the present paper, the discussion of estimates that are consistent in expected information divergence is limited to the finite-alphabet case.

The organization of the paper is as follows. In Section II we describe an algorithm for constructing estimates $\hat{P}_k(dx|X^{-\lambda(k)})$ and prove weak consistency for all stationary real-valued time series. The method and its proof applies to time series with values in any σ-compact Polish space. In Section III we transform the estimates $\hat{P}_k(dx|X^{-\lambda(k)})$ into estimates $\hat{P}(dx|X^{-n})$ by letting k depend on n. We choose an increasing sequence $k(n)$ and define the estimate $\hat{P}(dx|X^{-n})$ as $\hat{P}_{k(n)}(dx|X^{-\lambda(k(n))})$ if $\lambda(k(n)) \leq n$ and as some default measure $Q(dx)$ otherwise. If $k(n)$ grows sufficiently slowly with n then the data requirement $\lambda(k(n))$ will seldom exceed the available length n and the estimates $\hat{P}(dx|X^{-n})$ will be weakly consistent just like the estimates $\hat{P}_{k(n)}(dx|X^{-\lambda(k(n))})$. Section IV is about modeling and data compression and about estimates that are consistent in expected information divergence for stationary processes with values in a finite alphabet. In Section V, we shift $\hat{P}(dx|X^{-t})$ from time 0 to time t and show that the shifted estimates $\hat{P}(dx_t|X^t)$ can be used for sequential forecasting or on-line prediction. We show that one can make sequential decisions based on the shifted estimates $\hat{P}(dx_t|X^t)$ so that the average loss per decision converges in mean to the minimum long run average loss that could be attained if one could make decisions with knowledge of the true conditional distribution of the next outcome given the infinite past at each step. In particular, the average rate of incorrect guesses in classification and the average of the mean squared error in regression converge to the minimum that could be attained if the infinite past were known to begin with.

We would like to alert the reader about some of our notational conventions. Only one level of subscripts or superscripts is allowed in equations that are embedded in the text and so we are often forced to adopt the flat functional notation $\lambda(k)$, $\lambda(k(n))$, $\ell(k)$, $J(k)$,

$\tau(k,j)$, etc. However, the equations sometimes look better with nested subscripts and superscripts and therefore we prefer to write λ_k, $\lambda_{k(n)}$, ℓ_k, J_k, τ_j^k, etc. in the displayed equations. We hope that mixing of these notational conventions will not be a source of confusion but rather will improve the readability of the paper. Logarithms and entropy rates are taken in base 2 unless specified otherwise, and exponential growth rates are really doubling rates.

II. Learning the Conditional Distribution $P(dx|X^-)$

Let $\{X_t\}$ be a real-valued stationary time series. The process distribution is unknown but shift-invariant. We wish to infer the conditional distribution of $X = X_0$ given the infinite past X^- from past experience. We show that it is very easy to construct weakly consistent estimates $\hat{P}_k(dx|X^{-\lambda(k)})$ depending on finite past data segments $X^{-\lambda(k)}$ such that for every bounded continuous function $h(x)$ on $\mathcal{X}$ and any stationary distribution $P \in \mathcal{P}_s$,

$$\lim_k \int h(x)\hat{P}_k(dx|X^{-\lambda(k)}) = \int h(x)\, P(dx|X^-) \quad \text{in } L^1(P). \tag{12}$$

The estimates $\hat{P}_k(dx|X^{-\lambda(k)})$ will be defined in terms of quantized versions of the process $\{X_t\}$. Let $\mathcal{X}$ denote the real line and let $\{\mathcal{B}_k\}_{k\geq 1}$ be an increasing sequence of finite subfields that asymptotically generate the Borel σ-field on $\mathcal{X}$. Let $x \mapsto [x]^k$ denote the quantizer that maps any point $x \in \mathcal{X}$ to the atom of $\mathcal{B}_k$ that happens to contain x. For any integer $\ell \geq 1$ let $[X^{-\ell}]^k$ denote the quantized sequence $([X_{-\ell}]^k, \ldots, [X_{-1}]^k)$. Given any integer $J \geq 1$, one may search backwards in time and collect J samples of the process at times when the quantized ℓ-past looks exactly like the quantized ℓ-past at time 0. Let $\lambda = \lambda(k,\ell,J)$ denote the length of the data segment $X^{-\lambda} = (X_{-\lambda}, \ldots, X_{-1})$ that must be inspected to find these J samples and let $\hat{P}_{k,\ell,J}(dx|X^{-\lambda})$ denote the empirical distribution of those samples. Then $\hat{P}_{k,\ell,J}(dx|X^{-\lambda})$ will be a good estimate of $P(dx|X^-)$ if the sample size J, the context length ℓ and the quantizer index k are sufficiently large. In fact, if k and ℓ are fixed and the sample size J tends to infinity then by the ergodic theorem, $\hat{P}_{k,\ell,J}(dx|X^{-\lambda(k,\ell,J)})$ will converge in law to $P(dx|[X^{-\ell}]^k)$. If we now refine the context by increasing k and ℓ, then $P(dx|[X^{-\ell}]^k)$ will converge in law to $P(dx|X^-)$ by the martingale convergence theorem. The question is how to turn this limit of limits into a single limit by letting k, ℓ and J increase simultaneously to infinity. We must make k and ℓ large to reduce the bias and we must make J large to reduce the variance of the estimates. We will let ℓ and J grow with k and show that if $\ell(k)$ and $J(k)$ are monotonically increasing to infinity then the empirical conditional distribution estimate $\hat{P}_k(dx|X^{-\lambda(k)}) = \hat{P}_{k,\ell(k),J(k)}(dx|X^{-\lambda(k,\ell(k),J(k))})$ converges weakly to $P(dx|X^-)$. After this brief outline we now proceed with a detailed development.

Let $\{\ell_k\}_{k\geq 1}$ and $\{J_k\}_{k\geq 1}$ be two nondecreasing unbounded sequences of positive integers. We often write $\ell(k)$ and $J(k)$ instead of ℓ_k and J_k. For fixed $k \geq 1$ let $\{-\tau_j^k\}_{j\geq 0}$ and $\{\tilde{\tau}_j^k\}_{j\geq 0}$

denote the sequences of past and future recurrence times of the pattern $[X^{-\ell(k)}]^k$. Thus we set $\tau_0^k = \tilde{\tau}_0^k = 0$ and for $j = 1, 2, \ldots$ we inductively define

$$\tau_j^k = \min\{t > \tau_{j-1}^k : ([X_{-\ell_k - t}]^k, \ldots, [X_{-1-t}]^k) = ([X_{-\ell_k}]^k, \ldots, [X_{-1}]^k)\}, \tag{13}$$

$$\tilde{\tau}_j^k = \min\{t > \tilde{\tau}_{j-1}^k : ([X_{-\ell_k}]^k, \ldots, [X_{-1}]^k) = ([X_{-\ell_k + t}]^k, \ldots, [X_{-1+t}]^k)\}. \tag{14}$$

The random variables $\tau(k,j) = \tau_j^k$ and $\tilde{\tau}(k,j) = \tilde{\tau}_j^k$ are finite almost surely by Poincaré's recurrence theorem for the quantized process $\{[X_t]^k\}$, cf. Theorem 6.4.1 of Gray [14]. The lengths $\lambda_k = \lambda(k)$ and estimates $\hat{P}_k(dx|X^{-\lambda(k)})$ are now defined by the formulas

$$\lambda_k = \lambda(k) = \ell(k) + \tau(k, J_k), \tag{15}$$

$$\hat{P}_k(dx|X^{-\lambda_k}) = \frac{1}{J_k} \sum_{1 \le j \le J_k} \delta_{X_{-\tau(k,j)}}(dx), \tag{16}$$

where $\delta_\xi(dx)$ is the Dirac measure that places unit mass at the point $\xi \in \mathcal{X}$. Thus for any Borel set B, the conditional probability estimate

$$\hat{P}_k\{X \in B|X^{-\lambda_k}\} = \frac{1}{J_k} \sum_{1 \le j \le J_k} 1\{X_{-\tau(k,j)} \in B\} \tag{17}$$

is obtained by searching for the J_k most recent occurrences of the pattern $[X^{-\ell(k)}]^k$ and calculating the relative frequency with which the next realized symbols $X_{-\tau(k,j)}$ hit the set B. We shall prove that $\hat{P}_k(dx|X^{-\lambda(k)})$ is a weakly consistent estimate of $P(dx|X^-)$. The precise statement and the proof are broken down in two parts.

Theorem 1A. *For any set B in the generating field $\bigcup_k \mathcal{B}_k$ and any stationary process distribution $P \in \mathcal{P}_s$ we have mean convergence*

$$\lim_k \hat{P}_k\{X \in B|X^{-\lambda_k}\} = P\{X \in B|X^-\} \quad \textit{in } L^1(P). \tag{18}$$

The proof is somewhat technical and is placed in the Appendix. In the second part we argue that the estimators $\hat{P}_k(dx|X^{-\lambda(k)})$ can be employed to infer the regression function $E\{h(X)|X^-\} = \int h(x)\, P(dx|X^-)$ of any bounded continuous function $h(x)$ given the past.

Theorem 1B. *Let $\{X_t\}$ be a real-valued stationary time series. If the fields $\mathcal{B}_k$ are generated by intervals and the estimator $\hat{P}_k(dx|X^{-\lambda(k)})$ is defined as in (16) then for any bounded continuous function $h(x)$ on $\mathcal{X}$,*

$$\lim_k \int h(x) \hat{P}_k(dx|X^{-\lambda_k}) = \int h(x)\, P(dx|X^-) \quad \textit{in } L^1(P). \tag{19}$$

Proof: Pick some bound M such that $|h(x)| \le M$ on $\mathcal{X}$. Given $\epsilon > 0$ there exists an integer κ and a finite interval K in the field $\mathcal{B}_\kappa$ such that

$$P\{X \in K\} > 1 - \frac{\epsilon}{M}. \tag{20}$$

If necessary we increase κ until κ is sufficiently large so that there exists a $\mathcal{B}_\kappa$-measurable function $g(x)$ such that $|h(x) - g(x)| \le \epsilon$ on K. Assuming $g(x) = 0$ outside K, we have

$$|h(x) - g(x)| \le f(x) = \epsilon\, 1\{x \in K\} + M\, 1\{x \notin K\}. \tag{21}$$

Let $\hat{P}_k$ and P^- be shorthand for $\hat{P}_k(dx|X^{-\lambda(k)})$ and $P(dx|X^-)$. Then

$$\left| \int h\, d\hat{P}_k - \int h\, dP^- \right| \le \int |h - g|\, d\hat{P}_k + \left| \int g\, d\hat{P}_k - \int g\, dP^- \right| + \int |g - h|\, dP^-. \tag{22}$$

The function $g(x)$ is a finite linear combination of indicator functions of $\mathcal{B}_\kappa$-measurable subsets, and Theorem 1A implies that $\int g\, d\hat{P}_k$ converges to $\int g\, dP^-$ in L^1:

$$E\left| \int g\, d\hat{P}_k - \int g\, dP^- \right| \to 0. \tag{23}$$

The function $f(x)$ is $\mathcal{B}_\kappa$-measurable and bounded, hence $\int f\, d\hat{P}_k$ converges to $\int f\, dP^-$ in L^1 and the expectations converge:

$$E \int f\, d\hat{P}_k \to E \int f\, dP^- = Ef. \tag{24}$$

Since $|h - g| \le f$ and $Ef \le \epsilon\, P\{X \in K\} + M\, P\{X \notin K\} < 2\epsilon$ by (20) and (21), it follows from (22), (23) and (24) that

$$E\left| \int h\, d\hat{P}_k - \int h\, dP^- \right| \le 2\epsilon + \epsilon + 2\epsilon \quad \text{eventually for large } k. \tag{25}$$

Thus $E|\int h\, d\hat{P}_k - \int h\, dP^-| \to 0$, and this is the desired conclusion (19). ∎

Theorem 1B holds in general if $\mathcal{X}$ is a σ-compact Polish space and the fields $\mathcal{B}_k$ are suitably chosen. Indeed, let $\{K_k\}_{k \ge 1}$ be an increasing sequence of compact subsets with union $\bigcup_k K_k = \mathcal{X}$. For any fixed k one may cover K_k with a finite collection of open balls having diameter less than ϵ_k, where $\epsilon_k \searrow 0$ as $k \to \infty$. Let $\mathcal{B}_k$ denote the smallest field containing $\mathcal{B}_{k-1}$ and the sets $B \cap K_k$ where B ranges over all balls in the finite cover of K_k. (We start with the trivial field $\mathcal{B}_0 = \{\emptyset, \mathcal{X}\}$.) Any bounded continuous function $h(x)$ on $\mathcal{X}$ is uniformly continuous on each compact subset of $\mathcal{X}$. If $|h(x)| \le M$ and $\epsilon > 0$, then for sufficiently large κ there exists some compact subset K in $\mathcal{B}_\kappa$ such that $P\{X \notin K\} \le \epsilon/M$ and $h(x)$ oscillates less than ϵ on each atom of $\mathcal{B}_\kappa$ that is contained in K. Thus there exists a $\mathcal{B}_\kappa$-measurable function $g(x)$ such that $|h(x) - g(x)| < \epsilon$ on K and $g(x) = 0$ outside K. We can then proceed as in the proof of Theorem 1B to prove that for any bounded continuous function $h(x)$,

$$\int h(x)\, \hat{P}_k(dx|X^{-\lambda(k)}) \to E\{h(X)|X^-\} \quad \text{in } L^1. \tag{26}$$

III. Truncation of the Search Depth

The estimates $\hat{P}_k(dx|X^{-\lambda(k)})$ are based on finite but random length segments of the past. We shall transform these into estimates $\hat{P}(dx|X^{-n})$ that depend on finite past segments with deterministic length but that still are weakly consistent. The details are somewhat more involved than for the strongly consistent estimates in Section I. In terms of the empirical conditional distribution $\hat{P}_{k,\ell,J}(dx|X^{-\lambda(k,\ell,J)})$ that was defined in the outline of Section II, the question is how fast k, ℓ and J may increase with n so that $\lambda(k(n),\ell(n),J(n)) \leq n$ with high probability. The weak consistency of the estimates $\hat{P}_{k(n),\ell(n),J(n)}(dx|X^{-\lambda(k(n),\ell(n),J(n))})$ will not suffer if we redefine the estimates by assigning some default measure $Q(dx)$ in those rare cases when the search depth $\lambda(k(n),\ell(n),J(n))$ exceeds the available record length n. It is difficult to say what the optimal growth path is for $k(n)$, $\ell(n)$ and $J(n)$ without prior information about the spatial and temporal dependency structure of the process.

The special case of finite alphabet processes is most interesting and it is simpler because only 2 of the 3 parameters k, ℓ, J play a role. We do not need an index for subfields of $\mathcal{X}$ because the obvious choice for $\mathcal{B}_k$ is the field of all subsets of $\mathcal{X}$. Also, it is convenient to choose the block length ℓ_k equal to k so that τ_j^k is the time for j recurrences of X^{-k}.

In Section A we recall the ergodic theorem for recurrence times that was derived by Wyner and Ziv [34] and by Ornstein and Weiss [24] for finite alphabet processes. In Section B we define conditional probability mass function estimates $\hat{P}(x|X^{-n})$ and we prove consistency in mean if the block length $k(n)$ and the sample size $J_{k(n)}$ grow deterministically and sufficiently slowly with n. In Section C we discuss generalizations for real-valued processes.

A. Recurrence Times

Let $\{X_t\}$ be a stationary ergodic process with values in a finite set $\mathcal{X}$. Starting at time $\tau_0^k = 0$, the successive recurrence times τ_j^k of the k-block X^{-k} are defined as follows:

$$\tau_j^k = \inf\{t > \tau_{j-1}^k : (X_{-k-t}, \ldots, X_{-1-t}) = (X_{-k}, \ldots, X_{-1})\}. \tag{27}$$

If $P\{X^{-k} = x^{-k}\} > 0$ then by the results of Kac [17] (see also Willems [33], Wyner and Ziv [34]),

$$E\{\tau_1^k | X^{-k} = x^{-k}\} = \frac{1}{P\{X^{-k} = x^{-k}\}}. \tag{28}$$

Let H denote the entropy rate of the stationary ergodic process $\{X_t\}$ in bits per symbol:

$$H = \lim_k -\frac{1}{k} E\{\log P(X^k)\} = \lim_k -\frac{1}{k} E\{\log P(X^{-k})\}. \tag{29}$$

Wyner and Ziv [34], Theorem 3, invoked Kac's result and the Shannon-McMillan-Breiman theorem to prove that τ_1^k cannot grow faster than exponentially with limiting rate H

($\limsup_k k^{-1} \log \tau_1^k \le H$ almost surely). Ornstein and Weiss [24] then argued that τ_1^k will grow exponentially fast almost surely with limiting rate exactly equal to H:

$$k^{-1} \log \tau_1^k \to H \quad \text{almost surely.} \tag{30}$$

Now suppose a sample of size J_k is desired. The total time needed to find $J_k = J(k) \ge 1$ instances of the pattern X^{-k} is equal to the recurrence time $\tau_{J(k)}^k$. The ratio $\tau_{J(k)}^k / J_k$ can be interpreted as the average inter-recurrence time:

$$\frac{\tau_{J(k)}^k}{J_k} = \frac{1}{J_k} \sum_{1 \le j \le J_k} (\tau_j^k - \tau_{j-1}^k). \tag{31}$$

We claim that like τ_j^k, the average inter-recurrence time $\tau_{J(k)}^k / J_k$ cannot grow faster than exponentially with limiting rate H. The proof is based on Kac's result and the lemma that was developed by Algoet and Cover [3] to give a simple proof of the Shannon-McMillan-Breiman theorem and a more general ergodic theorem for the maximum exponential growth rate of compounded capital invested in a stationary market.

Theorem 2. *Let $\{X_t\}$ be a stationary ergodic process with values in a finite set $\mathcal{X}$ and with entropy rate H bits per symbol. If $\Delta_k = \Delta(k)$ is a sequence of numbers such that $\sum_k 2^{-\Delta(k)} < \infty$, then for arbitrary $J(k) = J_k > 0$ we have*

$$\log \left(\frac{\tau_{J(k)}^k}{J_k} \right) \le -\log P(X^{-k}) + \Delta_k \quad \textit{eventually for large } k, \tag{32}$$

and consequently

$$\limsup_k \frac{1}{k} \log \left(\frac{\tau_{J(k)}^k}{J_k} \right) \le H \quad \textit{almost surely.} \tag{33}$$

Proof: The inter-recurrence times $\tau_j^k - \tau_{j-1}^k$ are identically distributed with the same conditional distribution given X^{-k} as the first recurrence time τ_1^k. By Kac's result,

$$E\{\tau_{J(k)}^k | X^{-k}\} P(X^{-k}) = J_k \, E\{\tau_1^k | X^{-k}\} P(X^{-k}) = J_k. \tag{34}$$

(A referee pointed out that a result like this was also proved by Gavish and Lempel [13].) Thus the random variable $Z_k = P(X^{-k}) \, \tau_{J(k)}^k / J_k$ has expectation

$$E\{Z_k\} = E\left\{ P(X^{-k}) \, E\left\{ \frac{\tau_{J(k)}^k}{J_k} \middle| X^{-k} \right\} \right\} = 1. \tag{35}$$

By the Markov inequality,

$$P\{\log Z_k > \Delta_k\} = P\{Z_k > 2^{\Delta_k}\} \le 2^{-\Delta_k} E\{Z_k\} = 2^{-\Delta_k}, \tag{36}$$

and by the Borel-Cantelli lemma $\log Z_k \le \Delta_k$ eventually for larger k. This proves (32). Assertion (33) follows from (32) upon dividing both sides by k and taking the lim sup as

$k \to \infty$. Indeed, $-k^{-1} \log P(X^{-k}) \to H$ almost surely by the Shannon-McMillan-Breiman theorem and one may choose $\Delta_k = 2 \log k$ so that $\Delta_k / k \to 0$. ∎

It is worthwhile to observe that Theorem 2 can be generalized if the process $\{X_t\}$ is stationary but not necessarily ergodic. Let P be a stationary distribution and let P_ω denote the ergodic mode of the actual process realization ω. Then by the ergodic decomposition theorem (see Theorem 7.4.1 of Gray [14]) and the monotone convergence theorem,

$$\begin{aligned} P\{X^{-k} = x^{-k}\} E\{\tau^k_{J(k)} | X^{-k} = x^{-k}\} &= \sum_{1 \le t < \infty} t P\{X^{-k} = x^{-k}, \tau^k_{J(k)} = t\} \\ &= \sum_{1 \le t < \infty} \int t P_\omega\{X^{-k} = x^{-k}, \tau^k_{J(k)} = t\} P(d\omega) \\ &= \int \sum_{1 \le t < \infty} t P_\omega\{X^{-k} = x^{-k}, \tau^k_{J(k)} = t\} P(d\omega) \\ &= \int P_\omega\{X^{-k} = x^{-k}\} E_\omega\{\tau^k_{J(k)} | X^{-k} = x^{-k}\} P(d\omega) \\ &= \int J_k P(d\omega) \\ &= J_k. \end{aligned} \tag{37}$$

It follows that $E\{P(X^{-k})\, \tau^k_{J(k)}\} = J_k$ and

$$\log(\tau^k_{J(k)} / J_k) \le -\log P(X^{-k}) + \Delta_k \quad \text{eventually for large } k. \tag{38}$$

The Shannon-McMillan-Breiman theorem for stationary nonergodic processes asserts that $P(X^{-k})$ decreases exponentially fast with limiting rate $H(P_\omega)$, so one may conclude that

$$\limsup_k \frac{1}{k} \log \left(\frac{\tau^k_{J(k)}}{J_k} \right) \le H(P_\omega) \quad \text{almost surely.} \tag{39}$$

Thus the average inter-recurrence time $\tau^k_{J(k)} / J_k$ cannot grow faster than exponentially with limiting rate $H(P_\omega)$, the entropy rate of the ergodic mode P_ω.

B. Conditional Probability Mass Function Estimates

In the finite alphabet case, the general estimator $\hat{P}_k(dx | X^{-\lambda(k)})$ that was defined in (16) reduces to the conditional probability mass function estimate

$$\hat{P}_k(x | X^{-\lambda(k)}) = \frac{1}{J_k} \sum_{1 \le j \le J_k} 1\{X_{-\tau(k,j)} = x\}. \tag{40}$$

Here $k = \ell_k$ is the block length and the sample size J_k is monotonically increasing. The recurrence times τ^k_j of the k-block X^{-k} were defined inductively for $j = 1, 2, 3, \ldots$ in (27).

We choose a slowly increasing sequence of block lengths $k(n)$ and set $\hat{P}(x | X^{-n})$ equal to $\hat{P}_{k(n)}(x | X^{-\lambda(k(n))})$ if this estimate can be computed from the available data segment X^{-n}.

Otherwise, if $\lambda_{k(n)} > n$, we truncate the search and define $\hat{P}(x|X^{-n})$ as the default measure $Q(x) = 1/|\mathcal{X}|$. Thus for $n \geq 0$, we define

$$\hat{P}(x|X^{-n}) = \begin{cases} \hat{P}_{k(n)}(x|X^{-\lambda(k(n))}) & \text{if } \lambda(k(n)) \leq n, \\ Q(x) & \text{otherwise.} \end{cases} \tag{41}$$

If $k(n)$ grows sufficiently slowly then truncation is a rare event and $\hat{P}(x|X^{-n})$ coincides most of the time with the weakly consistent estimator $\hat{P}_{k(n)}(x|X^{-\lambda(k(n))})$. The question is how fast the block length $k(n)$ and the sample size $J_{k(n)}$ may grow to get consistent estimates. To answer this question, we use our results about recurrence times.

The inter-recurrence times $\tau_j^k - \tau_{j-1}^k$ have the same conditional distribution and hence the same conditional expectation given X^{-k} as the first recurrence time τ_1^k. The expected inter-recurrence time is bounded as follows:

$$E\left\{\frac{\tau_{J(k)}^k}{J_k}\right\} = E\{\tau_1^k\} = \sum_{x^{-k}:\, P\{X^{-k}=x^{-k}\}>0} P\{X^{-k} = x^{-k}\}\, E\{\tau_1^k | X^{-k} = x^{-k}\} \leq |\mathcal{X}|^k. \tag{42}$$

If $\epsilon_k > 0$ then by the Markov inequality

$$P\left\{\frac{\tau_{J(k)}^k}{J_k} > \frac{|\mathcal{X}|^k}{\epsilon_k}\right\} \leq \epsilon_k. \tag{43}$$

If $\epsilon_k \to 0$ then $P\{\tau_{J(k)}^k > J_k|\mathcal{X}|^k/\epsilon_k\} \to 0$ and if $\sum_k \epsilon_k < \infty$ then $\tau_{J(k)}^k \leq J_k|\mathcal{X}|^k/\epsilon_k$ eventually for large k by the Borel-Cantelli lemma. This is similar to (32) with $\epsilon_k = 2^{-\Delta(k)}$. Since $\lambda(k) = k + \tau_{J(k)}^k$, we see that

$$P\{\lambda(k(n)) \leq n\} \to 1 \quad \text{as } n \to \infty \tag{44}$$

if J_k and $k(n)$ are chosen so that for some $\epsilon_k > 0$ with $\epsilon_k \to 0$,

$$k(n) + J_{k(n)}|\mathcal{X}|^{k(n)}/\epsilon_{k(n)} \leq n \quad \text{eventually for large } n. \tag{45}$$

It suffices that $k(n) = (1-\epsilon)\log_{|\mathcal{X}|} n$ for some $0 < \epsilon < 1$ and $J_k = o(|\mathcal{X}|^{k\epsilon/(1-\epsilon)})$ so that $J_{k(n)} = o(n^\epsilon)$. (Noninteger values are rounded down to the nearest integer, as usual.) We can be slightly more aggressive.

Theorem 3. *Let $\{X_t\}$ be a stationary process with values in a finite set $\mathcal{X}$ and choose $Q(x) = |\mathcal{X}|^{-1}$ as default measure in (41). If the block length $k(n)$ and the sample size $J_{k(n)}$ are monotonically increasing to infinity and satisfy*

$$J_{k(n)}\, |\mathcal{X}|^{k(n)} = \mathcal{O}(n), \tag{46}$$

then the estimates $\hat{P}(x|X^{-n})$ in (41) are consistent in mean:

$$\hat{P}(x|X^{-n}) \to P(x|X^-) \quad \text{in } L^1(P). \tag{47}$$

In particular, the estimates $\hat{P}(x|X^{-n})$ are consistent in mean if the block length is $k(n) = (1-\epsilon)\log_{|\mathcal{X}|} n$ and the sample size is $J_{k(n)} = n^{\epsilon}$ for some $0 < \epsilon < 1$.

Proof: If the entropy rate H is strictly less than $\log|\mathcal{X}|$ and R is any constant such that $H < R < \log|\mathcal{X}|$ then by (33), $\tau^k_{J(k)}$ is asymptotically bounded by $J_k 2^{Rk}$. It follows that

$$\tau^{k(n)}_{J(k(n))} \leq J_{k(n)} 2^{Rk(n)} \leq J_{k(n)} |\mathcal{X}|^{k(n)} 2^{(R-\log|\mathcal{X}|)k(n)} = o(n). \tag{48}$$

It is necessary for (46) that $k(n) < \log_{|\mathcal{X}|} n$ eventually for large n since $J_{k(n)} \to \infty$ by assumption. Thus $\lambda(k(n)) = k(n) + \tau^{k(n)}_{J(k(n))} = o(n)$ and $\lambda(k(n))$ is upper bounded by n eventually for large n. If $H = \log|\mathcal{X}|$ then there is no guarantee that we can collect $J_{k(n)}$ samples from X^{-n}, but the estimate $\hat{P}(x|X^{-n})$ will nevertheless be consistent in mean if the default measure is $Q(x) = |\mathcal{X}|^{-1}$ because the outcomes X_t happen to be independent identically distributed according to this distribution $Q(x)$ when $H = \log|\mathcal{X}|$. ∎

The estimates $\hat{P}_k(x|X^{-\lambda(k)})$ in (40) are consistent in the pointwise sense under certain conditions. For example, if $\{X_t\}$ is a stationary finite-state Markov chain with order K then the empirical estimates $\hat{P}_k(x|X^{-\lambda(k)})$ are averages of bounded random variables $1\{X_{-\tau(k,j)} = x\}$ $(j = 1, 2, \ldots, J_k)$ that are conditionally independent and identically distributed given X^{-K} when $k \geq K$. It follows that the estimates $\hat{P}_k(x|X^{-\lambda(k)})$ converge exponentially fast in the number of samples J_k to the conditional probability $P\{x|X^{-K}\} = P\{x|X^{-}\}$ and therefore the estimates are pointwise consistent. It is not known whether the estimates $\hat{P}_k(x|X^{-\lambda(k)})$ converge in the pointwise sense for all finite-alphabet stationary time series.

If we know the entropy rate H in advance we can make use of it. In this case, weak consistency is guaranteed if $k(n) = (1-\epsilon)(\log n)/R$ for some $R > H$ and $J_{k(n)} = n^{\epsilon}$. Indeed, if $H < r < R$ then $\lambda(k(n)) < n$ eventually for large n since

$$\begin{aligned} \lambda(k(n)) &= k(n) + \tau^{k(n)}_{J(k(n))} \\ &\leq k(n) + J(k(n)) 2^{rk(n)} \\ &= \mathcal{O}(\log n) + n^{\epsilon} n^{(1-\epsilon)r/R} \\ &= o(n). \end{aligned} \tag{49}$$

If the entropy rate is not known in advance then we must be prepared to deal with the worst case of nearly maximum entropy rate. The estimates will be wasteful if the entropy rate is low because they exploit only a small portion of the available data segment X^{-n} when $H < \log|\mathcal{X}|$. If $k(n) = (1-\epsilon)\log_{|\mathcal{X}|} n$ and $J_{k(n)} = n^{\epsilon}$ then the length of the useful portion is about

$$\tau^{k(n)}_{J(k(n))} \approx J_{k(n)} 2^{Hk(n)} = n^{\epsilon + (1-\epsilon)H/\log|\mathcal{X}|} = n^{\alpha}, \tag{50}$$

where $\alpha = \epsilon + (1-\epsilon)H/\log|\mathcal{X}|$ varies linearly between $\epsilon < \alpha \leq 1$ as $0 < H \leq \log|\mathcal{X}|$.

The length $\lambda(k) = k + \tau^k_{J(k)}$ of the data record $X^{-\lambda(k)}$ that must be examined to collect J_k samples of the pattern X^{-k} grows approximately like $J_k 2^{Hk}$, which is polynomial in J_k if J_k grows exponentially fast with k. Also, the length n of the segment X^{-n} is just polynomial in the sample size $J_{k(n)}$ if $J_{k(n)} = n^\epsilon$. The strongly consistent estimates of Morvai, Yakowitz and Györfi [21] are much less efficient: they collect J samples from a data record whose length grows like a tower of exponentials in (7). Their samples are very sparse because extremely stringent demands are placed on the context where those samples are taken. For the weakly consistent estimates of the present study, the demands on context are much less severe and so the samples are much more abundant although perhaps less trustworthy. Thus universal prediction is not hopelessly out of computational reach as it might seem for an algorithm whose input demands grow as a tower of exponentials in (7).

C. Weak Consistency for Real-valued Processes

When $\mathcal{X}$ is the real line or a σ-compact Polish space, the estimate $\hat{P}_k(dx|X^{-\lambda(k)})$ is defined by the formula in (16). We now choose a nondecreasing unbounded sequence $k(n)$ and we define $\hat{P}(dx|X^{-n})$ as the empirical conditional distribution $\hat{P}_{k(n)}(dx|X^{-\lambda(k(n))})$ if this estimate can be computed from the available data segment X^{-n}. Otherwise, if $\lambda_{k(n)} > n$, we truncate the search and define $\hat{P}(dx|X^{-n})$ as some default measure $Q(dx)$. Thus

$$\hat{P}(dx|X^{-n}) = \begin{cases} \hat{P}_{k(n)}(dx|X^{-\lambda_{k(n)}}) & \text{if } \lambda_{k(n)} \le n, \\ Q(dx) & \text{otherwise.} \end{cases} \tag{51}$$

If $k(n)$ grows slowly then truncation is rare and $\hat{P}(dx|X^{-n})$ coincides most of the time with the estimator $\hat{P}_{k(n)}(dx|X^{-\lambda(k(n))})$ which is weakly consistent. The question is how slowly the partition index $k(n)$, the block length $\ell(k(n))$ and the sample size $J(k(n))$ must grow with n to get consistent estimates of $P(dx|X^-)$. It suffices that $P\{\lambda(k(n)) < n\} \to 1$.

Theorem 4. *Let $\{X_t\}$ be a real-valued stationary ergodic time series and choose $\mathcal{B}_k$, ℓ_k and J_k as before. Let Ξ_k denote the set of atoms of the finite field $\mathcal{B}_k$ and choose a nondecreasing unbounded sequence of integers $k(n)$ and numbers $\epsilon_k \to 0$ such that*

$$n \ge \ell_{k(n)} + J_{k(n)} |\Xi_{k(n)}|^{\ell_{k(n)}} / \epsilon_{k(n)} \quad \textit{eventually for large } n. \tag{52}$$

Then $P\{n \ge \lambda_{k(n)}\} \to 1$ as $n \to \infty$, and the estimates $\hat{P}(dx|X^{-n})$ are weakly consistent: for every set B in the generating field $\bigcup_k \mathcal{B}_k$ we have

$$\hat{P}\{X \in B|X^{-n}\} \to P\{X \in B|X^-\} \quad \textit{in } L^1(P), \tag{53}$$

and for every bounded continuous function $h(x)$ we have

$$\int h(x)\, \hat{P}(dx|X^{-n}) \to \int h(x)\, P(dx|X^-) \quad \textit{in } L^1(P). \tag{54}$$

Proof: The inter-recurrence times $\tau_j^k - \tau_{j-1}^k$ $(j = 1, 2, 3, \ldots)$ are identically distributed conditionally given the pattern $[X^{-\ell(k)}]^k$. By Kac's result,

$$E\{\tau_{J(k)}^k | [X^{-\ell_k}]^k\} = J_k \, E\{\tau_1^k | [X^{-\ell_k}]^k\} = \frac{J_k}{P([X^{-\ell_k}]^k)}. \tag{55}$$

It follows that

$$E\{\tau_{J(k)}^k\} = J_k E\{\tau_1^k\} = \sum_{[x^{-\ell(k)}]^k} P([x^{-\ell(k)}]^k) E\{\tau_{J(k)}^k | [x^{-\ell(k)}]^k\} \le J_k |\Xi_k|^{\ell(k)}. \tag{56}$$

(The sum is taken over $[x^{-\ell(k)}]^k$ such that $P([x^{-\ell(k)}]^k) = P\{[X^{-\ell(k)}]^k = [x^{-\ell(k)}]^k\}$ is strictly positive.) By the Markov inequality,

$$P\{\lambda_k > \ell_k + J_k |\Xi_k|^{\ell(k)}/\epsilon_k\} = P\{\tau_{J(k)}^k > J_k |\Xi_k|^{\ell(k)}/\epsilon_k\} \le \frac{E\{\tau_{J(k)}^k\}}{J_k |\Xi_k|^{\ell(k)}/\epsilon_k} \le \epsilon_k. \tag{57}$$

Assertions (53) and (54) follow from Theorem 1A and 1B because $P\{\ell_k + J_k |\Xi_k|^{\ell_k}/\epsilon_k \ge \lambda_k\} \to 1$ and hence, in view of assumption (52),

$$P\{n \ge \ell_{k(n)} + J_{k(n)} |\Xi_{k(n)}|^{\ell_{k(n)}}/\epsilon_{k(n)} \ge \lambda_{k(n)}\} \to 1 \quad \text{as } n \to \infty. \tag{58}$$

This completes the proof of the theorem. ∎

The theorem remains valid in the stationary non-ergodic case. Indeed, let P be a stationary distribution and let P_ω denote the ergodic mode of ω. Then one may argue as above that $P_\omega\{\lambda_k \le \ell_k + J_k |\Xi_k|^{\ell(k)}/\epsilon_k\} \to 1$. By the ergodic decomposition theorem and Lebesgue's dominated convergence theorem,

$$\begin{aligned} \lim_k P\{\lambda_k \le \ell_k + J_k |\Xi_k|^{\ell(k)}/\epsilon_k\} &= \lim_k \int P_\omega\{\lambda_k \le \ell_k + J_k |\Xi_k|^{\ell(k)}/\epsilon_k\} P(d\omega) \\ &= \int \lim_k P_\omega\{\lambda_k \le \ell_k + J_k |\Xi_k|^{\ell(k)}/\epsilon_k\} P(d\omega) \\ &= \int 1 P(d\omega) = 1. \end{aligned} \tag{59}$$

Thus the conclusions of the theorem also hold for stationary nonergodic processes.

IV. The Information Theoretic Point of View

In this section we discuss conditional distribution estimates $\hat{P}(dx|X^{-n})$ that are consistent in expected information divergence. Such estimates are also weakly consistent, but the converse is not necessarily true. It is possible to construct estimator sequences that are consistent in expected information divergence for all stationary processes with values in a finite alphabet, but not for all stationary processes with values in a countable infinite alphabet. There are connections with universal gambling or modeling schemes and with

universal noiseless data compression algorithms for finite alphabet processes. For more information on these subjects see Rissanen and Langdon [28] and Algoet [1].

A. Consistency in Expected Information Divergence

The Kullback-Leibler information divergence between two probability distributions P and Q on a measurable space $\mathcal{X}$ is defined as follows: if P is dominated by Q then

$$I(P|Q) = E_P\left\{\log\left(\frac{dP}{dQ}\right)\right\}, \tag{60}$$

otherwise $I(P|Q) = \infty$. The variational distance is defined as

$$\|P - Q\| = \sup_{-1\le h(x)\le 1}\left|\int h\,dP - \int h\,dQ\right|, \tag{61}$$

where the supremum is taken over all measurable functions $h(x)$ such that $|h(x)| \le 1$. If $p = dP/d\mu$ and $q = dQ/d\mu$ are the densities of P and Q relative to a dominating σ-finite measure μ then $\|P - Q\| = \int |p - q|\,d\mu$. Exercise 17 on p. 58 of Csiszár and Körner [11] asserts that

$$\frac{\log e}{2}\,\|P - Q\|^2 \le I(P|Q). \tag{62}$$

It follows that $I(P|Q) \ge 0$ with equality iff $P = Q$. Pinsker [26], pp. 13–15 proved the existence of a universal constant $\Gamma > 0$ such that

$$I(P|Q) \le E_P\left\{\left|\log\left(\frac{dP}{dQ}\right)\right|\right\} \le I(P|Q) + \Gamma\sqrt{I(P|Q)}. \tag{63}$$

Barron [6] simplified Pinsker's argument and proved that the constant $\Gamma = \sqrt{2}$ is best possible when natural logarithms are used in the definition of $I(P|Q)$.

Let $\{X_t\}$ be a stationary process with values in a complete separable metric space $\mathcal{X}$. The divergence between the true conditional distribution $P(dx|X^-)$ and an estimate $\hat{P}(dx|X^{-t})$ is a nonnegative function of the past X^- which vanishes iff $P(dx|X^-) = \hat{P}(dx|X^{-t})$ P-almost surely. We say that the estimates $\hat{P}(dx|X^{-t})$ are consistent in information divergence for a class Π of stationary distributions on $\mathcal{X}^{\mathcal{Z}}$ if for any $P \in \Pi$,

$$I(P_{X|X^-}|\hat{P}_{X|X^{-t}}) \to 0 \quad P\text{-almost surely.} \tag{64}$$

We say that $\hat{P}(dx|X^{-t})$ is consistent in expected information divergence for the class Π if for any $P \in \Pi$,

$$E_P\{I(P_{X|X^-}|\hat{P}_{X|X^{-t}})\} \to 0. \tag{65}$$

Such estimates are weakly consistent for all distributions in the class Π. Indeed, if $h(x)$ is any bounded measurable function on $\mathcal{X}$ with norm $\|h\|_\infty = \sup_x |h(x)|$ then

$$|\textstyle\int h(x)\,P(dx|X^-) - \int h(x)\,\hat{P}(dx|X^{-t})| \le \|h\|_\infty \|P_{X|X^-} - \hat{P}_{X|X^{-t}}\|. \tag{66}$$

Applying the Csiszár-Kemperman-Kullback inequality (62), we see that

$$\left| \int h(x)\, P(dx|X^-) - \int h(x)\, \hat{P}(dx|X^{-t}) \right|^2 \le \frac{2\|h\|_\infty^2}{\log e}\, I(P_{X|X^-}|\hat{P}_{X|X^{-t}}). \tag{67}$$

If $\hat{P}(dx|X^{-t})$ is consistent in expected information divergence for Π then $\int h(x)\, \hat{P}(dx|X^{-t})$ converges in $L^2(P)$ and also in $L^1(P)$ to $\int h(x)\, P(dx|X^-)$ whenever $P \in \Pi$.

Suppose the outcomes X_t are independent with identical distribution P_X on $\mathcal{X}$. Barron, Györfi and van der Meulen [7] have constructed estimates $\hat{P}(dx|X^{-t})$ that are consistent in information divergence and in expected information divergence when the true distribution P_X has finite information divergence $I(P_X|M_X) < \infty$ relative to some known normalized reference measure M_X. Györfi, Páli and van der Meulen [16] assume that $\mathcal{X}$ is the countable set of integers and argue that for arbitrary conditional probability mass function estimates $\hat{P}(x|X^{-n})$, there exists some distribution P_X with finite entropy such that

$$I(P_X|\hat{P}_{X|X^{-n}}) = \infty \quad \text{almost surely for all } n. \tag{68}$$

Therefore, it is impossible to construct estimates $\hat{P}(dx|X^{-t})$ that are consistent in information divergence or in expected information divergence for all independent identically distributed processes with values in an infinite space. For stationary processes with values in a finite alphabet, the constructions of Ornstein [22] and Morvai, Yakowitz and Györfi [21] yield estimates $\hat{P}(x|X^{-t})$ such that $\log \hat{P}(x|X^{-t})$ converges almost surely to $\log P(x|X^-)$. It is still an open question as to whether these estimates are consistent in information divergence or whether modifications are needed to get such consistency. (The difficulty is that small changes in $\hat{P}(x|X^{-n})$ cause huge changes in $\log \hat{P}(x|X^{-n})$ when $\hat{P}(x|X^{-n})$ is small.) However, it is easy to construct estimates $\hat{P}(x|X^{-t})$ that are consistent in expected information divergence.

B. Consistent Estimates for Finite-alphabet Processes

Let $\{X_t\}$ be a stationary process with values in a finite set $\mathcal{X}$. We shall construct conditional probability mass function estimates $\hat{P}(x|X^{-n})$ that are consistent in expected information divergence for any stationary $P \in \mathcal{P}_s$. Such estimates also converge to $P(x|X^-)$ in mean: for any stationary $P \in \mathcal{P}_s$ and $x \in \mathcal{X}$ we have

$$\hat{P}(x|X^{-n}) \to P(x|X^-) \quad \text{in } L^1(P). \tag{69}$$

An observation of Perez [25] implies that consistency in expected information divergence is equivalent to mean consistency of $\log \hat{P}(X|X^{-n})$.

Theorem 5. *Let $\{X_t\}$ be a stationary process with values in a finite alphabet $\mathcal{X}$. A sequence of conditional probability mass function estimates $\hat{P}(x|X^{-n})$ is consistent in expected information divergence iff we have mean convergence*

$$\log \hat{P}(X|X^{-n}) \to \log P(X|X^-) \quad \textit{in } L^1. \tag{70}$$

Proof: Pinsker's inequality (63) for $P(x|X^-)$ and $\hat{P}(x|X^{-n})$ asserts that

$$\begin{aligned} I(P_{X|X^-}|\hat{P}_{X|X^{-n}}) &\leq E\left\{\left|\log\left(\frac{P(X|X^-)}{\hat{P}(X|X^{-n})}\right)\right|\Big|X^-\right\} \\ &\leq I(P_{X|X^-}|\hat{P}_{X|X^{-n}}) + \Gamma\sqrt{I(P_{X|X^-}|\hat{P}_{X|X^{-n}})}. \end{aligned} \tag{71}$$

Taking expectations and using concavity of the square root function, we obtain

$$\begin{aligned} E\{I(P_{X|X^-}|\hat{P}_{X|X^{-n}})\} &\leq E\left|\log\left(\frac{P(X|X^-)}{\hat{P}(X|X^{-n})}\right)\right| \\ &\leq E\{I(P_{X|X^-}|\hat{P}_{X|X^{-n}})\} + \Gamma\sqrt{E\{I(P_{X|X^-}|\hat{P}_{X|X^{-n}})\}} \end{aligned} \tag{72}$$

by Jensen's inequality. This suffices to prove the theorem. ∎

To construct the estimates $\hat{P}(x|X^{-n})$, we start with probability mass functions $Q(x^n)$ on the product spaces $\mathcal{X}^n$ such that for every stationary distribution P on $\mathcal{X}^{\mathcal{Z}}$,

$$n^{-1}I(P_{X^n}|Q_{X^n}) \to 0 \quad \text{as } n \to \infty. \tag{73}$$

Several methods are known for constructing such models $Q(x^n)$ – see Section C below. By Pinsker's inequality, convergence of the means in (73) is equivalent to mean convergence

$$\frac{1}{n}\log\left(\frac{P(X^n)}{Q(X^n)}\right) \to 0 \quad \text{in } L^1(P). \tag{74}$$

Let now $Q(x|x^{-t})$ denote a shifted copy of the conditional probability mass function $Q(x_t|x^t)$ that appears in the chain rule expansion $Q(x^n) = \prod_{0\leq t<n} Q(x_t|x^t)$. The estimate $\hat{P}(x|X^{-n})$ is defined in terms of $Q(x^n)$ as

$$\hat{P}(x|X^{-n}) = \frac{1}{n}\sum_{0\leq t<n} Q(x|X^{-t}). \tag{75}$$

Theorem 6. *Let $\mathcal{X}$ be a finite alphabet and let $\{Q(x^n)\}_{n\geq 1}$ be a model sequence such that (73) or (74) holds for all $P \in \mathcal{P}_s$. Then the conditional probability mass function estimates $\hat{P}(x|X^{-n})$ are consistent in expected information divergence for the class $\mathcal{P}_s$ of all stationary process distributions on $\mathcal{X}^{\mathcal{Z}}$.*

Proof: The Kullback-Leibler divergence functional is convex in both arguments. By the definition (75) of $\hat{P}(x|X^{-n})$ and by Jensen's inequality,

$$I(P_{X|X^-}|\hat{P}_{X|X^{-n}}) \leq \frac{1}{n}\sum_{0\leq t<n} I(P_{X|X^-}|Q_{X|X^{-t}}). \tag{76}$$

Now we take expectations with respect to some distribution $P \in \mathcal{P}_s$. By stationarity and the chain rule expansion of information divergence, we obtain

$$\begin{aligned} E_P\{I(P_{X|X^-}|\hat{P}_{X|X^{-n}})\} &\leq \frac{1}{n}\sum_{0\leq t<n} E_P\{I(P_{X|X^-}|Q_{X|X^{-t}})\} \\ &= \frac{1}{n}\sum_{0\leq t<n} E_P\{I(P_{X_t|X^-X^t}|Q_{X_t|X^t})\} \\ &= \frac{1}{n}E_P\{I(P_{X^n|X^-}|Q_{X^n})\} \\ &= \frac{1}{n}E_P\{I(P_{X^n|X^-}|P_{X^n})\} + \frac{1}{n}I(P_{X^n}|Q_{X^n}). \end{aligned} \tag{77}$$

Observe that

$$E_P\{I(P_{X^n|X^-}|P_{X^n})\} = H(X^n) - H(X^n|X^-) \tag{78}$$

where $H(X^n) = E_P\{-\log P(X^n)\}$ and $H(X^n|X^-) = E_P\{-\log P(X^n|X^-)\}$. The entropy rate of the process is defined as $H = H(X|X^-) = n^{-1}H(X^n|X^-) = \downarrow \lim_n n^{-1}H(X^n)$, so one may conclude that

$$\frac{1}{n}E_P\{I(P_{X^n|X^-}|P_{X^n})\} = \frac{1}{n}H(X^n) - H \to 0 \quad \text{as } n \to \infty. \tag{79}$$

It follows from (77) and (79) that the estimates $\hat{P}(x|X^{-n})$ are consistent in expected information divergence, as claimed. ∎

The procedure which constructs $\hat{P}(x|x^{-n})$ from the models $Q(x^n)$ can be reversed. Indeed, let $\{\hat{P}(x|x^{-t}\}_{t\geq 0}$ be a sequence such that for every stationary distribution $P \in \mathcal{P}_s$, the expected information divergence of $P(x|X^-)$ relative to $\hat{P}(x|X^{-t})$ is finite for all t and vanishes in the limit as $t \to \infty$. Let $\hat{P}(x_t|x^t)$ be constructed from the t-past at time t in the same way as $\hat{P}(x|x^{-t})$ was constructed from the t-past at time 0. The Kullback-Leibler information divergence of the true marginal distribution $P(x^n)$ with respect to the compounded model $\hat{P}(x^n) = \prod_{0\leq t<n}\hat{P}(x_t|x^t)$ admits the chain rule expansion

$$I(P_{X^n}|\hat{P}_{X^n}) = \sum_{0\leq t<n} E_P\{I(P_{X_t|X^t}|\hat{P}_{X_t|X^t})\}. \tag{80}$$

By stationarity

$$E_P\{I(P_{X_t|X^t}|\hat{P}_{X_t|X^t})\} = E_P\{I(P_{X|X^{-t}}|\hat{P}_{X|X^{-t}})\}. \tag{81}$$

The divergence of $P(x|X^{-t})$ relative to $\hat{P}(x|X^{-t})$ is bounded by the divergence of $P(x|X^-)$ relative to $\hat{P}(x|X^{-t})$ since we have the decomposition

$$I(P_{X|X^-}|\hat{P}_{X|X^{-t}}) = I(P_{X|X^-}|P_{X|X^{-t}}) + I(P_{X|X^{-t}}|\hat{P}_{X|X^{-t}}). \tag{82}$$

From (80), (81) and (82) one may conclude that

$$I(P_{X^n}|\hat{P}_{X^n}) \leq \sum_{0\leq t<n} E_P\{I(P_{X|X^-}|\hat{P}_{X|X^{-t}})\}. \tag{83}$$

If the expected divergence between $P(x|X^-)$ and $\hat{P}(x|X^{-t})$ is finite and vanishes in the limit as $t \to \infty$ then the models $\hat{P}(x^n) = \prod_{0 \le t < n} \hat{P}(x_t|x^t)$ have vanishing expected per-symbol divergence: for all $P \in \mathcal{P}_s$ we have

$$n^{-1} I(P_{X^n}|\hat{P}_{X^n}) \to 0. \tag{84}$$

The results of Shields [32] imply that there can be no universal bound on the speed of convergence in expected information divergence. Indeed, if the expected divergence of $P(x|X^-)$ relative to $\hat{P}(x|X^{-t})$ were always $\mathcal{O}(\beta_t)$ where $\beta_t \to 0$, then we could construct a modeling scheme $\{\hat{P}(x_t|x^t)\}_{t \ge 0}$ such that the divergence of $P(x^n)$ relative to $\hat{P}(x^n) = \prod_{0 \le t < n} \hat{P}(x_t|x^t)$ would be $\mathcal{O}(\beta_0 + \ldots + \beta_{n-1})$. The per-symbol divergence of $P(x^n)$ relative to $\hat{P}(x^n)$ would vanish with universal rate $\mathcal{O}[n^{-1}(\beta_0 + \ldots + \beta_{n-1})]$, which is impossible.

To obtain bounds on the per-symbol divergence one must restrict the process distribution to some manageable class. In particular, suppose Π is a class of Markov processes that is smoothly parametrized by k free parameters and consider models $\hat{P}(x^n)$ for which the per-symbol divergence attains Rissanen's [27] lower bound:

$$\frac{1}{n} I(P_{X^n}|\hat{P}_{X^n}) = \frac{k \log n}{2n}(1 + o(1)). \tag{85}$$

If we set $Q(x^n) = \hat{P}(x^n)$ and define $\hat{P}(x|X^{-n})$ as in (75), then (77) reduces to the bound

$$E_P\{I(P_{X|X^-}|\hat{P}_{X|X^{-n}})\} \le \frac{k \log n}{2n}(1 + o(1)), \quad P \in \Pi. \tag{86}$$

It is often possible to construct a prequential modeling scheme $\{\hat{P}(x_t|x^t)\}_{t \ge 0}$ such that the expected divergence of $P(x_t|X^t)$ relative to $\hat{P}(x_t|X^t)$ vanishes like $(k \log e)/(2t)$ for all process distributions in the class Π. An incremental bound of order $(k \log e)/(2t)$ yields a normalized cumulative bound of order $n^{-1} \sum_{t<n} (k \log e)/(2t) \approx (k \log n)/(2n)$. By shifting $\hat{P}(x_n|X^n)$ we obtain estimates $\hat{P}(x|X^{-n})$ such that the expected divergence of $P(x|X^-)$ relative to $\hat{P}(x|X^{-n})$ vanishes like $(k \log e)/(2n)$. This bound of order $(k \log e)/(2n)$ for $\hat{P}(x|X^{-n})$ is clearly better than the bound $(k \log n)/(2n)$.

C. Modeling and Data Compression

Any universal data compression scheme for stationary processes with finite alphabet $\mathcal{X}$ can be used as a basis for the construction of models $Q(x^n)$ satisfying (73) or (74). Indeed, let $l(x^n)$ denote the length of a uniquely decipherable block-to-variable-length binary code for sequences $x^n \in \mathcal{X}^n$. The redundancy of the code for X^n is defined as the difference between the actual codeword length $l(X^n)$ and the ideal description length $-\log P(X^n)$:

$$r(X^n) = l(X^n) + \log P(X^n). \tag{87}$$

The expected redundancy $E_P\{r(X^n)\}$ is equal to the information divergence between the true probability mass function $P(x^n)$ and the model

$$Q'(x^n) = 2^{-l(x^n)}, \quad x^n \in \mathcal{X}^n. \tag{88}$$

For a universal noiseless coding scheme, the expected per-symbol redundancy will vanish:

$$\frac{1}{n}E_P\{r(X^n)\} = \frac{1}{n}I(P_{X^n}|Q'_{X^n}) \to 0 \quad \text{for all } P \in \mathcal{P}_s. \tag{89}$$

The model $Q'(x^n)$ is not necessarily normalized but is always a subprobability measure, by the Kraft-McMillan inequality. However, if (89) holds for a sequence of subnormalized models $Q'(x^n)$ then (89) will certainly hold for the normalized models

$$Q(x^n) = \frac{Q'(x^n)}{\sum_{\xi^n \in \mathcal{X}^n} Q'(\xi^n)}. \tag{90}$$

Theorem 4 of Algoet [1] implies that for any stationary ergodic distribution P, the per-symbol description length of uniquely decipherable codes is asymptotically bounded below almost surely by the entropy rate $H(P) = \lim_n n^{-1}E_P\{-\log P(X^n)\}$:

$$\liminf_n n^{-1}l(X^n) \geq H(P) \quad P\text{-almost surely.} \tag{91}$$

It is well known that there exist universal noiseless codes for which the per-symbol description length almost surely approaches the entropy rate of the ergodic mode P_ω with probability one under any stationary distribution P:

$$n^{-1}l(X^n(\omega)) \to H(P_\omega) \quad P\text{-almost surely, for all } P \in \mathcal{P}_s. \tag{92}$$

This is true in particular for the data compression algorithm of Ziv and Lempel [36], by Theorem 12.10.2 of Cover and Thomas [10] or by the results of Ornstein and Weiss [24]. Other examples of noiseless codes satisfying (92) for every stationary ergodic P have been proposed by Ryabco [29], Ornstein and Shields [23], and Algoet [1]. Choosing the best among the given code with length $l(x^n)$ and a fixed-length code with length $\lceil n \log |\mathcal{X}| \rceil$ and adding one bit of preamble to indicate which code is better, one obtains a uniquely decipherable code with length

$$l'(x^n) = 1 + \min\{l(x^n), \lceil n \log |\mathcal{X}| \rceil\}. \tag{93}$$

The codeword may expand by one bit, but the per-symbol description length is now bounded by $\log |\mathcal{X}| + 2n^{-1}$ and (92) holds universally not only in the pointwise sense but also in mean. The corresponding models $Q(x^n)$ are universal in the sense that for any $P \in \mathcal{P}_s$,

$$\frac{1}{n}\log\left(\frac{P(X^n)}{Q(X^n)}\right) \to 0 \quad P\text{-almost surely and in } L^1(P). \tag{94}$$

Ryabco [29] and Algoet [1] have constructed probability measures Q with marginals $Q(x^n)$ such that the pointwise convergence in (94) holds for every stationary $P \in \mathcal{P}_s$. Each marginal $Q(x^n)$ is equal to the compounded product $Q(x^n) = \prod_{0 \le t < n} Q(x_t|x^t)$, and Ryabco's scheme has the extra property that when P is finite order Markov,

$$\log \left(\frac{P(X_t|X^t)}{Q(X_t|X^t)} \right) \to 0 \quad P\text{-almost surely.} \tag{95}$$

Rissanen and Langdon [28] and Langdon [18] previously observed that the Lempel-Ziv algorithm defines a sequential predictive modeling scheme $Q = \{Q(x_t|x^t)\}$. The per-symbol divergence vanishes pointwise in the Cesàro mean sense, for every $P \in \mathcal{P}_s$:

$$\frac{1}{n} \log \left(\frac{P(X^n)}{Q(X^n)} \right) = \frac{1}{n} \sum_{0 \le t < n} \log \left(\frac{P(X_t|X^t)}{Q(X_t|X^t)} \right) \to 0 \quad P\text{-almost surely.} \tag{96}$$

However, the pointwise convergence in (95) must fail for some $P \in \mathcal{P}_s$ because the quality of the predictive model $Q(x_t|X^t)$ degrades whenever the Lempel-Ziv incremental parsing procedure comes to the end of a phrase. The leaves of the dictionary tree and the nodes with few descendants are exactly those where empirical evidence is still lacking to make a reliable forecast. The number of times a node has been visited is equal to the number of leaves in the subtree rooted at that node, and if this number is small then the predictive model for the next symbol is a poor estimate based on few samples.

If the estimates $\hat{P}(x|X^{-t})$ are universally consistent in expected information divergence then $\log[P(X|X^-)/\hat{P}(X|X^{-t})] \to 0$ in $L^1(P)$ for all stationary $P \in \mathcal{P}_s$ by Theorem 5. Thus the shifted estimates $\hat{P}(x_t|X^t)$ are universally consistent in the sense that for all $P \in \mathcal{P}_s$,

$$\log \left(\frac{P(X_t|X^t)}{\hat{P}(X_t|X^t)} \right) \to 0 \quad \text{in } L^1(P). \tag{97}$$

Bailey [5] and Ryabco [30] proved that no modeling scheme Q exists such that the pointwise convergence in (95) holds for every stationary ergodic distribution P. The argument of [30] shows that for any modeling scheme Q there exists a stationary ergodic distribution P on $\mathcal{X}^{\mathcal{Z}}$ where $\mathcal{X} = \{a, b, c\}$ such that P fails to satisfy both (95) and the statement

$$P(X_t|X^t) - Q(X_t|X^t) \to 0 \quad P\text{-almost surely.} \tag{98}$$

The offending P is determined by a Markov chain with a countable set of states $\{0, 1, 2, \ldots\}$. Given that the Markov chain is in state i, it moves to state 0 with probability $1/2$ and generates the letter a, or it moves to state $i + 1$ with probability $1/2$ and generates the letter b or c with conditional probability Δ_i and $(1 - \Delta_i)$, where Δ_i is a parameter equal to either $1/3$ or $2/3$. The distribution of the Markov chain is determined by the infinite sequence $\Delta = (\Delta_0, \Delta_1, \ldots)$. If the Markov chain is started in its stationary distribution then the resulting distribution P_Δ on the sequence space $\mathcal{X}^\infty$ is stationary ergodic. Exact

prediction is impossible when the Markov chain visits a state i which it has not visited before, because the predictor doesn't know whether the probability $\Delta_i/2$ of next seeing symbol b is equal to 1/3 or 1/6. The Markov chain will visit states with arbitrarily large labels i, and the predictor must make inaccurate predictions infinitely often with positive probability under distribution P_Δ for some $\Delta = (\Delta_0, \Delta_1, \ldots)$.

V. Application to Online Prediction

In this section we discuss some applications of the estimates $\hat{P}(dx|X^{-t})$ to on-line prediction, regression and classification. We deal with special cases of a sequential decision problem that can be formulated abstractly as follows.

Let $\{X_t\}$ be a stationary process with values in the space $\mathcal{X}$ and let $l(x, a)$ be a loss function on $\mathcal{X} \times \mathcal{A}$ where $\mathcal{A}$ is a space of possible actions. We assume that $\mathcal{X}$ is a complete and $\mathcal{A}$ is a compact separable metric space and the loss function $l(x, a)$ is bounded and continuous on $\mathcal{X} \times \mathcal{A}$. We wish to select nonanticipating actions $A_t = A_t(X^t)$ with knowledge of the past $X^t = (X_0, \ldots, X_{t-1})$ so as to minimize the long run average loss per decision:

$$\limsup_n \frac{1}{n} \sum_{0 \le t < n} l(X_t, A_t) = \text{Min!} \tag{99}$$

If the process distribution is known a priori then the optimum strategy is to select actions $A_t^* = \arg\min_{a \in \mathcal{A}} E\{l(X_t, a)|X^t\}$ that attain the minimum conditional expected loss given the available information X^t at each time t. Suppose P is stationary and let $L(X_t|X^t)$ denote the expectation of the minimum conditional expected loss given the t-past at time t:

$$L(X_t|X^t) = E\{l(X_t, A_t^*)\} = \inf_{A_t = A_t(X^t)} E\{l(X_t, A_t)\}. \tag{100}$$

Similarly let $L(X|X^{-t})$ and $L(X|X^-)$ denote the minimum expected loss given the t-past and the minimum expected loss given the infinite past at time 0. By stationarity $L(X_t|X^t) = L(X|X^{-t})$, and $L(X|X^{-t})$ is clearly monotonically decreasing to a limit which by continuity must be $L(X|X^-)$. Thus for any stationary distribution P one may define

$$L^*(P) = \downarrow \lim_t L(X_t|X^t) = \downarrow \lim_t L(X|X^{-t}) = L(X|X^-). \tag{101}$$

If P is stationary ergodic then the minimum long run average loss is well defined and almost surely equal to $L^*(P) = L(X|X^-)$ by Theorem 6 of Algoet [2]:

$$\frac{1}{n} \sum_{0 \le t < n} l(X_t, A_t^*) \to L^*(P) \quad P\text{-almost surely and in } L^1(P). \tag{102}$$

Now suppose the process distribution is unknown a priori. It is shown in Section V.B of Algoet [2] that there exist nonanticipating actions $\hat{A}_t^* = \hat{A}_t^*(X^t)$ which attain the minimum long run average loss $L^*(P)$ with probability one under any stationary ergodic process

distribution P on $\mathcal{X}^{\mathcal{Z}}$. The actions $\hat{A}_t^*$ are constructed by a plug-in approach as follows. Choose estimates $\hat{P}(dx|X^{-t})$ that converge in law to the conditional distribution $P(dx|X^-)$ with probability one under any stationary P and construct $\hat{P}(dx_t|X^t)$ from X^t in the same way as $\hat{P}(dx|X^{-t})$ was computed from X^{-t}. Then $\hat{A}_t^*$ is defined as an action that attains the minimum conditional expected loss given X^t under $\hat{P}(dx_t|X^t)$:

$$\hat{A}_t^* = \arg\min_{a\in\mathcal{A}} \int l(x_t, a)\, \hat{P}(dx_t|X^t). \tag{103}$$

The average loss incurred by the actions $\hat{A}_t^*$ converges pointwise to the minimum long run average loss $L^*(P)$.

In this paper we rely on conditional distribution estimates $\hat{P}(dx|X^{-t})$ that are weakly consistent but hopefully more efficient than the pointwise consistent estimates of [22], [1], [21]. We limit our attention to certain on-line prediction problems, when $\mathcal{X} = \mathcal{A}$ is a compact separable metric space and the loss $l(x, \hat{x})$ is a continuous increasing function of the distance between the outcome x and the prediction $\hat{x}$. In classification problems $\mathcal{X} = \mathcal{A}$ is a finite set, $l(x, \hat{x}) = 1\{x \neq \hat{x}\}$ is the Hamming distance, and we wish to predict each outcome X_t with knowledge of the past X^t so as to minimize the long run average rate of incorrect guesses. In regression problems $\mathcal{X}$ is a finite closed interval, $l(x, \hat{x})$ is the squared euclidean distance, and the goal is to predict X_t from the past X^t so that the long run average of the squared prediction error is smallest possible. We show that if the estimates $\hat{P}(dx|X^{-t})$ are weakly consistent, then the minimum long run average loss in regression and classification is universally attained in the sense of mean convergence in $L^1(P)$. The proof is based on the following generalization of von Neumann's mean ergodic theorem, which parallels Breiman's [8] generalization of Birkhoff's pointwise ergodic theorem. See also Perez [25].

Lemma. *Suppose $(\Omega, \mathcal{F}, P, T)$ is a stationary ergodic system. If g and $\{g_t\}_{t\geq 0}$ are integrable random variables such that $g_t \to g$ in $L^1(P)$, then*

$$\frac{1}{n} \sum_{0\leq t<n} g_t \circ T^t \to E\{g\} \quad \textit{in } L^1(P). \tag{104}$$

Proof: The mean ergodic theorem asserts that

$$\frac{1}{n} \sum_{0\leq t<n} g \circ T^t \to E\{g\} \quad \text{in } L^1(P), \tag{105}$$

and it is clear that

$$\frac{1}{n} \sum_{0\leq t<n} [g_t \circ T^t - g \circ T^t] \to 0 \quad \text{in } L^1(P) \tag{106}$$

since the triangle inequality, stationarity and the assumption $E|g_t - g| \to 0$ imply that

$$E\Big|\frac{1}{n} \sum_{0\leq t<n} [g_t \circ T^t - g \circ T^t]\Big| \leq \frac{1}{n} \sum_{0\leq t<n} E|g_t \circ T^t - g \circ T^t| = \frac{1}{n} \sum_{0\leq t<n} E|g_t - g| \to 0. \tag{107}$$

Addition of (105) and (106) yields (104). ∎

A. Regression

Let $\{X_t\}$ be a stationary ergodic real-valued time series with finite variance. We wish to predict each outcome X_t with knowledge of the past X^t so that the squared prediction error $|X_t - \hat{X}_t|^2$ is smallest possible in the long run average sense. The minimum long run average is equal to the minimum mean squared error given the infinite past, that is the variance of the innovation $X - E\{X|X^-\}$. If the outcomes X_t are independent and identically distributed then the sample mean $\hat{X}_t = (X_0 + \ldots + X_{t-1})/t$ is an optimal estimator in the long run. It is challenging to construct on-line predictors $\hat{X}_t$ that asymptotically attain the minimum squared prediction error in a universal sense for all stationary ergodic real-valued processes with finite variance. Here, we consider the simple case of stationary processes with values in a finite interval $\mathcal{X} = [-K, K]$. We do not assume that K is known a priori.

Let $\{\hat{P}(dx|X^{-t})\}_{t\geq 0}$ denote a weakly consistent sequence of conditional distribution estimates as in Section III. Since $h(x) = x$ is a bounded continuous function on $\mathcal{X}$, it follows from Theorem 4 that $\hat{X}_{-t} \to \hat{X}$ in probability where

$$\hat{X}_{-t} = \int x\, \hat{P}(dx|X^{-t}), \quad \hat{X} = E\{X|X^-\} = \int x\, P(dx|X^-). \tag{108}$$

Note that $\hat{X}_{-t}$ is not an estimate of X_{-t} but an estimate of $X = X_0$ based on the t-past X^{-t}. At time t we consider the conditional distribution estimate $\hat{P}(dx_t|X^t)$ and the predictor

$$\hat{X}_t = \int x_t\, \hat{P}(dx_t|X^t). \tag{109}$$

By construction $\hat{X}_t$ is the sample mean of some subset of the past outcomes $X_0, \ldots, X_{t-1}$, except in rare cases when $\hat{X}_t$ is equal to the default value $\int x\, Q(dx)$. The obvious choice for $Q(dx)$ is the Dirac measure that places unit mass at $x = 0$, so that $\int x\, Q(dx) = 0$. For any stationary ergodic process distribution P on $\mathcal{X}^{\mathcal{Z}}$ we have

$$|X - \hat{X}_{-t}|^2 \to |X - \hat{X}|^2 \quad \text{in } L^1(P), \tag{110}$$

and consequently, by the Lemma,

$$\frac{1}{n} \sum_{0\leq t<n} |X_t - \hat{X}_t|^2 \to E|X - \hat{X}|^2 \quad \text{in } L^1(P). \tag{111}$$

B. On-line Prediction and Classification

Let $\{X_t\}$ be a random process with values in a finite set $\mathcal{X}$. We wish to predict the outcomes X_t with knowledge of the past X^t so as to minimize the long run average rate of incorrect guesses. The best predictor for $X = X_0$ given the infinite past X^- is given by

$$\hat{X} = \arg\max_{x\in\mathcal{X}} P\{X = x|X^-\}. \tag{112}$$

If the process distribution P is stationary ergodic then the minimum long run average rate of prediction errors is equal to the error probability

$$P\{X \neq \hat{X}\} = 1 - E\{P\{X = \hat{X}|X^-\}\} = 1 - E\left\{\max_{x \in \mathcal{X}} P\{X = x|X^-\}\right\}. \tag{113}$$

If the process distribution is unknown, we choose some conditional probability mass estimates $\hat{P}\{X = x|X^{-t}\}$ that converge in mean to $P\{X = x|X^-\}$ for every stationary process distribution $P \in \mathcal{P}_s$ and $x \in \mathcal{X}$:

$$\hat{P}\{X = x|X^{-t}\} \to P\{X = x|X^-\} \quad \text{in } L^1(P). \tag{114}$$

We construct $\hat{P}\{X_t = x|X^t\}$ from the past X^t in the same way as $\hat{P}\{X = x|X^{-t}\}$ was computed from X^{-t} and we define the predictor

$$\hat{X}_t = \arg\max_{x \in \mathcal{X}} \hat{P}\{X_t = x|X^t\}, \tag{115}$$

Theorem 7. *Let $\{X_t\}$ be a stationary ergodic process with values in a finite set $\mathcal{X}$. If the conditional probability estimates $\hat{P}\{X = x|X^{-t}\}$ are weakly consistent, then the predictor $\hat{X}_t$ achieves the minimum long run average rate of incorrect guesses in probability. Thus for any stationary ergodic distribution P on $\mathcal{X}^{\mathcal{Z}}$ we have mean convergence*

$$\frac{1}{n} \sum_{0 \leq t < n} 1\{X_t \neq \hat{X}_t\} \to P\{X \neq \hat{X}\} \quad \textit{in } L^1(P). \tag{116}$$

Proof: Observe that $\hat{X}_t(\omega) = \hat{X}_{-t}(T^t\omega)$ where T is the left shift on $\mathcal{X}^{\mathcal{Z}}$ and where

$$\hat{X}_{-t} = \arg\max_{x \in \mathcal{X}} \hat{P}\{X = x|X^{-t}\}. \tag{117}$$

For any stationary ergodic P we have, by weak consistency of $\hat{P}\{X = x|X^{-t}\}$ and continuity of the maximum function,

$$\max_{x \in \mathcal{X}} \hat{P}\{X = x|X^{-t}\} \to \max_{x \in \mathcal{X}} P\{X = x|X^-\} \quad \text{in probability} \tag{118}$$

or equivalently

$$\hat{P}\{X = \hat{X}_{-t}|X^{-t}\} \to P\{X = \hat{X}|X^-\} \quad \text{in } L^1(P). \tag{119}$$

Since $[P\{X = x|X^-\} - \hat{P}\{X = x|X^{-t}\}] \to 0$ in $L^1(P)$ by weak consistency and

$$|P\{X = \hat{X}_{-t}|X^-\} - \hat{P}\{X = \hat{X}_{-t}|X^{-t}\}| \leq \sum_{x \in \mathcal{X}} |P\{X = x|X^-\} - \hat{P}\{X = x|X^{-t}\}|, \tag{120}$$

we see that

$$[P\{X = \hat{X}_{-t}|X^-\} - \hat{P}\{X = \hat{X}_{-t}|X^{-t}\}] \to 0 \quad \text{in } L^1(P). \tag{121}$$

It follows from (119) and (121) that

$$P\{X = \hat{X}_{-t}|X^-\} \to P\{X = \hat{X}|X^-\} \quad \text{in } L^1(P) \tag{122}$$

and consequently, by the Lemma,

$$\frac{1}{n}\sum_{0\le t<n} P\{X_t \neq \hat{X}_t|X^-X^t\} \to E\{P\{X \neq \hat{X}|X^-\}\} = P\{X \neq \hat{X}\} \quad \text{in } L^1(P). \tag{123}$$

Now observe that

$$\Delta_t = 1\{X_t \neq \hat{X}_t\} - P\{X_t \neq \hat{X}_t|X^-X^t\} \tag{124}$$

is a bounded martingale difference sequence with respect to the σ-fields $\sigma(X^-X^t)$ and hence

$$\frac{1}{n}\sum_{0\le t<n}\Delta_t = \frac{1}{n}\sum_{0\le t<n}[1\{X_t \neq \hat{X}_t\} - P\{X_t \neq \hat{X}_t|X^-X^t\}] \to 0 \quad \text{in } L^1(P) \tag{125}$$

(and also P-almost surely). In fact, the Cesàro means of Δ_t vanish exponentially fast by Azuma's [4] exponential inequalities for bounded martingale differences. Addition of (123) and (125) yields the conclusion (116). ∎

Feder, Merhav and Gutman [12] used the Lempel-Ziv algorithm as a method for sequential prediction of individual sequences.

C. Problems with Side Information

A well studied problem in statistical decision theory, pattern recognition and machine learning is to infer the class label X_t of an item at time t from a covariate or feature vector Y_t and a training set $X^tY^t = (X_0, Y_0, \ldots, X_{t-1}, Y_{t-1})$. It is often reasonable to assume that the successive pairs (X_t, Y_t) are independent and identically distributed, but sometimes defective items tend to come in batches or in periodic runs and in those cases it may be profitable to exploit dependencies between new items and recent or not so recent items. Here we assume that the pair process $\{(X_t, Y_t)\}$ is stationary ergodic and we try to exploit statistical dependencies of arbitrarily long range, although we have no idea what kind of dependencies to expect a priori. The minimum long run average misclassification rate is again equal to $P\{X \neq \hat{X}\}$, but now $\hat{X}$ is the best predictor of $X = X_0$ given the infinite past $X^-Y^- = (\ldots, X_{-2}, Y_{-2}, X_{-1}, Y_{-1})$ and the side information $Y = Y_0$:

$$\hat{X} = \arg\max_{x\in\mathcal{X}} P\{X = x|X^-Y^-Y\}. \tag{126}$$

The minimum misclassification rate will be asymptotically attained in probability by the predictors

$$\hat{X}_t = \arg\max_{x\in\mathcal{X}} \hat{P}\{X_t = x|X^tY^tY_t\}, \tag{127}$$

where $\hat{P}\{X_t = x|X^tY^tY_t\}$ is a shifted version of a conditional probability estimate $\hat{P}\{X = x|X^{-t}Y^{-t}Y\}$ such that for any stationary process distribution P on $(\mathcal{X} \times \mathcal{Y})^{\mathcal{Z}}$,

$$\hat{P}\{X = x|X^{-t}Y^{-t}Y\} \to P\{X = x|X^-Y^-Y\} \quad \text{in } L^1(P). \tag{128}$$

Such estimates $\hat{P}\{X = x|X^{-t}Y^{-t}Y\}$ can be constructed by generalizing the methods of Sections II and III.

In fact, let $\mathcal{X}$ and $\mathcal{Y}$ be complete separable metric spaces and let $\{\mathcal{B}_k\}_{k\geq 1}$ and $\{\mathcal{C}_k\}_{k\geq 1}$ be increasing sequences of finite subfields that asymptotically generate the Borel σ-fields on $\mathcal{X}$ and $\mathcal{Y}$. We assume that $\mathcal{X}$ is σ-compact and the fields $\mathcal{B}_k$ are constructed as in the paragraph after Theorem 1B. Let $[x]^k$ and $[y]^k$ denote the atoms of $\mathcal{B}_k$ and $\mathcal{C}_k$ that contain the points $x \in \mathcal{X}$ and $y \in \mathcal{Y}$, and consider the sequence of past recurrence times $\tau_j^k = \tau(k, j)$ of the pattern $[X^{-\ell(k)}Y^{-\ell(k)}Y]^k$. Then for every stationary process distribution P on $(\mathcal{X} \times \mathcal{Y})^{\mathcal{Z}}$,

$$\hat{P}(dx|X^{-\lambda_k}Y^{-\lambda_k}Y) = \frac{1}{J_k} \sum_{1\leq j\leq J_k} \delta_{X_{-\tau(k,j)}}(dx) \tag{129}$$

is a weakly consistent estimate of the true conditional distribution $P(dx|X^-Y^-Y)$. Thus all results in this paper remain valid if the decisions can be made with knowledge of not only the past but also side information.

Appendix

Let $\{\mathcal{B}_k\}_{k\geq 1}$ be an increasing sequence of finite subfields that asymptotically generate the Borel σ-field on $\mathcal{X}$, and suppose the empirical conditional distributions $\hat{P}_k(dx|X^{-\lambda(k)})$ are defined as in (16). Let T denote the left shift on the two-sided sequence space $\mathcal{X}^{\mathcal{Z}}$.

Theorem 1A. *If $\{X_t\}$ is a stationary process with values in a complete separable metric (Polish) space $\mathcal{X}$ then for every set B in the generating field $\bigcup_k \mathcal{B}_k$, we have*

$$\lim_k \hat{P}_k\{X \in B|X^{-\lambda_k}\} = P\{X \in B|X^-\} \quad \textit{in } L^1. \tag{130}$$

Proof: It follows from the martingale convergence theorem that

$$\lim_k P\{X \in B|[X^{-\ell_k}]^k\} = P\{X \in B|X^-\} \tag{131}$$

almost surely and in L^1. Thus it suffices to show that $E|\Theta_k| \to 0$ where

$$\Theta_k = \frac{1}{J_k} \sum_{0\leq j\leq J_k} 1\{X_{-\tau(k,j)} \in B\} - P\{X_0 \in B|[X^{-\ell_k}]^k\}. \tag{132}$$

We claim that $E|\Theta_k| = E|\tilde{\Theta}_k|$ where

$$\tilde{\Theta}_k = \frac{1}{J_k} \sum_{0\leq j\leq J_k} 1\{X_{\tilde{\tau}(k,j)} \in B\} - P\{X_0 \in B|[X^{-\ell_k}]^k\}. \tag{133}$$

Indeed, for any measurable function $g(\Theta) \geq 0$ (including $g(\Theta) = |\Theta|$) and for any integer sequence $0 = t_0 < t_1 < \ldots < t_{J(k)} = t$, we have

$$[1\{\tau_j^k = t_j,\ 0 \leq j \leq J_k\}\, g(\Theta_k)] = [1\{\tilde{\tau}_{J_k-j}^k = t - t_j,\ 0 \leq j \leq J_k\}\, g(\tilde{\Theta}_k)] \circ T^{-t} \tag{134}$$

and consequently, by stationarity,

$$\begin{aligned} Eg(\Theta_k) &= \sum_t \sum_{0=t_0<t_1<\ldots<t_{J_k}=t} E\{1\{\tau_j^k = t_j,\ 0 \leq j \leq J_k\}\, g(\Theta_k)\} \\ &= \sum_t \sum_{0=t_0<t_1<\ldots<t_{J_k}=t} E\{1\{\tilde{\tau}_{J_k-j}^k = t - t_j,\ 0 \leq j \leq J_k\}\, g(\tilde{\Theta}_k)\} \\ &= \sum_t \sum_{0=\tilde{t}_0<\tilde{t}_1<\ldots<\tilde{t}_{J_k}=t} E\{1\{\tilde{\tau}_i^k = \tilde{t}_i,\ 0 \leq i \leq J_k\}\, g(\tilde{\Theta}_k)\} = Eg(\tilde{\Theta}_k). \end{aligned} \tag{135}$$

Observe that $\{\tilde{\tau}_j^k - 1\}_{j\geq 0}$ is an increasing sequence of stopping times adapted to the filtration $\{\mathcal{F}_t^k\}_{t\geq 0}$ where

$$\mathcal{F}_t^k = \sigma(\ldots, [X_{-1}]^k, [X_0]^k, [X_1]^k, \ldots, [X_{t-1}]^k). \tag{136}$$

Let $\tilde{\mathcal{F}}_j^k$ denote the σ-field of events that are expressible in terms of the quantized random variables $[X_t]^k$ at times $t < \tilde{\tau}_j^k$. Thus $\tilde{\mathcal{F}}_j^k$ is the σ-field of events F such that $F \bigcap \{\tilde{\tau}_j^k = t\}$

belongs to $\mathcal{F}_t^k$ for all $t \geq 0$, and $\tilde{\mathcal{F}}_j^k$ is generated by the family of events $\{F_t \bigcap \{\tau_j^k = t\} : F_t \in \mathcal{F}_t^k,\ t \geq 0\}$. One may decompose $\tilde{\Theta}_k$ into the sum

$$\tilde{\Theta}_k = \frac{1}{J_k} \sum_{0 \leq j \leq J_k} (\Delta_j^k + \Phi_j^k), \tag{137}$$

where

$$\Delta_j^k = 1\{X_{\tilde{\tau}(k,j)} \in B\} - P\{X_{\tilde{\tau}(k,j)} \in B | \tilde{\mathcal{F}}_j^k\}, \tag{138}$$

$$\Phi_j^k = P\{X_{\tilde{\tau}(k,j)} \in B | \tilde{\mathcal{F}}_j^k\} - P\{X_0 \in B | [X^{-\ell_k}]^k\}. \tag{139}$$

Notice that $\{\Delta_j^k\}_{j \geq 0}$ is a martingale difference sequence with respect to the filtration $\{\tilde{\mathcal{F}}_j^k\}_{j \geq 0}$ (in the sense that Δ_j^k is $\tilde{\mathcal{F}}_{j+1}^k$-measurable and $E\{\Delta_j^k | \tilde{\mathcal{F}}_j^k\} = 0$ for all $j \geq 0$). Since $|\Delta_j^k| \leq 1$ and the random variables Δ_j^k are orthogonal, we see that

$$E\Big|\frac{1}{J_k} \sum_{0 \leq j \leq J_k} \Delta_j^k\Big|^2 = \frac{1}{J_k^2} \sum_{0 \leq j \leq J_k} E|\Delta_j^k|^2 \leq \left(\frac{1 + J_k}{J_k^2}\right) \tag{140}$$

and consequently (since $(E|Z|)^2 \leq E\{|Z|^2\}$ and $J_k \to \infty$),

$$E\Big|\frac{1}{J_k} \sum_{0 \leq j \leq J_k} \Delta_j^k\Big| \leq \frac{\sqrt{1 + J_k}}{J_k} \to 0. \tag{141}$$

Also observe that for any measurable function $g(\Phi) \geq 0$ and any integer $t \geq 0$,

$$[1\{\tau_j^k = t\}\, g(\Phi_0^k)] = [1\{\tilde{\tau}_j^k = t\}\, g(\Phi_j^k)] \circ T^{-t} \tag{142}$$

and consequently, by stationarity,

$$\begin{aligned} Eg(\Phi_j^k) &= \sum_{t \geq 0} E\{1\{\tilde{\tau}_j^k = t\}\, g(\Phi_j^k)\} \\ &= \sum_{t \geq 0} E\{1\{\tau_j^k = t\}\, g(\Phi_0^k)\} = Eg(\Phi_0^k). \end{aligned} \tag{143}$$

In particular, setting $g(\Phi) = |\Phi|$ proves that $E|\Phi_j^k| = E|\Phi_0^k|$. By the martingale convergence theorem

$$\Phi_0^k = P\{X_0 \in B | [X^{-\infty}]^k\} - P\{X_0 \in B | [X^{-\ell_k}]^k\} \to 0 \quad \text{almost surely and in } L^1 \tag{144}$$

and consequently

$$E\Big|\frac{1}{J_k} \sum_{0 \leq j \leq J_k} \Phi_j^k\Big| \leq \frac{1}{J_k} \sum_{0 \leq j \leq J_k} E|\Phi_j^k| = \left(\frac{1 + J_k}{J_k}\right) E|\Phi_0^k| \to 0. \tag{145}$$

The desired conclusion $E|\Theta_k| \to 0$ follows since

$$E|\Theta_k| = E|\tilde{\Theta}_k| \leq E\Big|\frac{1}{J_k} \sum_{0 \leq j \leq J_k} \Delta_j^k\Big| + E\Big|\frac{1}{J_k} \sum_{0 \leq j \leq J_k} \Phi_j^k\Big| \to 0. \tag{146}$$

■

Acknowledgment

The authors wish to express their gratitude to László Györfi for his encouragement and suggestions regarding this investigation. The first author thanks Tamás Szabados for the many discussions on Polish spaces. The second author's efforts have been partially supported by NSF Grant INT 92 01430 as well as NIH Grant R01 AI37535.

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] P. H. Algoet and T. M. Cover, "Asympotic optimality and asymptotic equipartition properties of log-optimum investment," *Annals Probab.,* vol. 16, pp. 876–898, 1988.

[4] K. Azuma, "Weighted sums of certain dependent random variables," *Tohoku Mathematical Journal,* vol. 37, pp. 357–367, 1967.

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

[6] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.,* vol. 14, pp. 336–342, 1986.

[7] A. R. Barron, L. Györfi and E. C. van der Meulen, "Distribution estimation consistent in total variation and in two types of information divergence," *IEEE Trans. Inform. Theory,* vol. 38, pp. 1437–1454, Sept. 1992.

[8] L. Breiman, "The individual ergodic theorem of information theory," *Ann. Math. Statist.,* vol. 28, pp. 809–811, 1957. Correction: *ibid.,* vol. 31, pp. 809–810, 1960.

[9] T. M. Cover, "Open problems in information theory," in *1975 IEEE Joint Workshop on Information Theory,* pp. 35–36. New York: IEEE Press, 1975.

[10] T. M. Cover and J. Thomas, *Elements of Information Theory.* New York: Wiley, 1991.

[11] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* Budapest: Akadémiai Kiadó, 1981.

[12] M. Feder, N. Merhav and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory,* vol. 38, pp. 1258–1270, July 1992.

[13] A. Gavish and A. Lempel, "Match-length functionals for data compression," presented at *IEEE Int. Symp. Inform. Theory*, Trondheim, Norway, Jun. 27–Jul. 1, 1994.

[14] R. M. Gray, *Probability, Random Processes, and Ergodic Theory.* New York: Springer-Verlag, 1988.

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

[16] L. Györfi, I. Páli and E. C. van der Meulen, "There is no universal source code for infinite source alphabet," *IEEE Trans. Inform. Theory*, vol. 40, pp. 267–271, Jan. 1994.

[17] M. Kac, "On the notion of recurrence in discrete stochastic processes," *Bull. Amer. Math. Soc.*, vol. 53, pp. 1002–1010, Oct. 1947.

[18] G. G. Langdon, Jr., "A note on the Lempel-Ziv model for compressing individual sequences," *IEEE Trans. Inform. Theory,* vol. IT-29, pp. 284–287, Mar. 1994.

[19] K. Marton and P. C. Shields, "Entropy and the consistent estimation of joint distributions," *Annals Probab.,* vol. 22, pp. 960–977, Apr. 1994.

[20] G. Morvai, *Estimation of Conditional Distributions for Stationary Time Series.* Ph. D. Thesis, Technical University of Budapest, 1994.

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

[22] D. S. Ornstein, "Guessing the next output of a stationary process," *Israel J. Math.,* vol. 30, pp. 292–296, 1978.

[23] D. S. Ornstein and P. C. Shields, "Universal almost sure data compression," *Annals Probab.*, vol. 18, pp. 441–452, 1990.

[24] D. S. Ornstein and B. Weiss, "Entropy and data compression schemes," *IEEE Trans. Inform. Theory,* vol. 39, pp. 78–83, Jan. 1993.

[25] A. Perez, "On Shannon-McMillan's limit theorem for pairs of stationary processes," *Kybernetika,* vol. 16, nr. 4, pp. 301–314, 1980.

[26] M. S. Pinsker, *Information and Information Stability of Random Variables and Processes.* Translated and edited by A. Feinstein. San Francisco: Holden-Day, 1964.

[27] J. Rissanen, "Stochastic complexity and modeling," *Ann. Statist.,* vol. 14, pp. 1080–1100, 1986.

[28] J. Rissanen and G. G. Langdon, Jr., "Universal modeling and coding," *IEEE Trans. Inform. Theory*, vol. IT-27, pp. 12–23, Jan. 1981.

[29] B. Ya. Ryabco, "Twice-universal coding," *Problems of Inform. Trans.,* vol. 20, pp. 173–177, July-Sept. 1984.

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

[31] B. Scarpellini, "Conditional expectations of stationary processes," *Z. Wahrscheinlichkeitstheorie verw. Gebiete,* vol. 56, pp. 427–441, 1981.

[32] P. C. Shields, "Universal redundancy rates don't exist," *IEEE Trans. Inform. Theory*, vol. 39, pp. 520–524, Mar. 1993.

[33] F. M. J. Willems, "Universal data compression and repetition times," *IEEE Trans. Inform. Theory,* vol. 35, pp. 54–58, Jan. 1989.

[34] A. D. Wyner and J. Ziv, "Some asymptotic properties of entropy of a stationary ergodic data source with applications to data compression," *IEEE Trans. Inform. Theory,* vol. 35, pp. 1250–1258, Nov. 1989.

[35] S. Yakowitz, L. Györfi, and G. Morvai, "An algorithm for nonparametric forecasting for ergodic, stationary time series," presented at *IEEE Int. Symp. Inform. Theory,* Trondheim, Norway, Jun. 27–Jul. 1, 1994.

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