Some Stochastic Properties of Memoryless Individual Sequences

Andrew B. Nobel *

December 9, 2003

Abstract

An individual sequence of real numbers is memoryless if no continuous Markov prediction scheme of finite order can outperform the best constant predictor under the squared loss. It is established that memoryless sequences satisfy an elementary law of large numbers, and sliding-block versions of Hoeffding's inequality and the central limit theorem. It is further established that memoryless binary sequences have convergent sample averages of every order, and that their limiting distributions are Bernoulli. Several examples and sources of memoryless sequences are given, and it is shown how memoryless binary sequences may be constructed from aggregating methods for sequential prediction.

Appears in *IEEE Transactions on Information Theory*, vol. 50, pp.1497-1505, 2004.

Key words and phrases. Individual sequence, prediction, memoryless, central limit theorem.

*Andrew Nobel is with the Department of Statistics, University of North Carolina, Chapel Hill, NC 27599. Email: nobel@stat.unc.edu. His work was supported in part by NSF Grant DMS 9971964.

1 Introduction

Sequences of independent, identically distributed random variables occupy a central place in information theory, probability, and statistics. In applications where such sequences are required, for example in cryptography or numerical simulations, they are typically replaced by deterministic individual sequences (termed pseudo-random) that nevertheless exhibit some sort of random behavior. (*c.f.* [15]). In general, what constitutes appropriate random behavior will depend on the application and on the judgement of the researcher.

One way to assess the random behavior of an individual sequence is through prediction. Independent tosses of a fair coin are a worst case scenario for sequential prediction schemes, as the past outcomes of the tosses provide no information about future tosses. Thus one desirable property of an individual pseudo-random sequence is that its values cannot cannot be predicted by a reasonably rich family of prediction schemes. Of course, independent coin tosses also obey the law of large numbers and the central limit theorem, and we might wish an individual sequence to have some properties of this sort too. The goal of this paper is to show that, in fact, unpredictable sequences share some of the statistical properties of truly random ones.

Consider a standard version of the sequential prediction problem, in which a fixed, non-random, sequence of numbers is revealed, one at a time, to a forecaster. At each stage the forecaster predicts the next number in the sequence from those she has already observed. When the next number is revealed, the forecaster suffers a loss equal to the squared difference between her prediction and the revealed number. (The squared loss is also known as the Brier score in this context.) Now consider a forecaster whose predictions at each time are obtained by applying a fixed, bounded continuous function of finitely many arguments to the previous values of the sequence. If no such Markov forecaster can outperform the best constant prediction scheme, in terms of long run average loss, then we define the individual sequence of observations to be *memoryless*. Independent, identically distributed random variables provide a canonical model for memoryless sequences, and it is natural to ask if individual memoryless sequences share some of the asymptotic properties of purely random ones. We provide affirmative answers to several questions of this sort.

1.1 Memoryless Sequences

We now give a more formal account of memoryless sequences. Let $\mathbf{x} = x_1, x_2, \ldots$ be an individual sequence of real numbers. Following standard notation, let $x_i^j = x_i, x_{i+1}, \ldots, x_j$

when $i \leq j$. For each $k \geq 1$ let $\mathcal{C}_k = C_b(\mathbb{R}^k)$ be the family of bounded continuous functions $g : \mathbb{R}^k \to \mathbb{R}$, and let $\mathcal{C}_0$ be the family of constant functions. To each function $g \in \mathcal{C}_k$ there is an associated prediction scheme F_g defined by

$$F_g(x_1^{j-1}) \;=\; \begin{cases} 0 & \text{if } j \leq k \\ g(x_{j-k}^{j-1}) & \text{if } j > k \end{cases} . \tag{1}$$

For each $t \geq k+1$, F_g predicts x_t by $g(x_{t-k}^{t-1})$. As the value of F_g does not depend on side information or randomization, the scheme F_g represents a deterministic strategy for the sequential prediction of $\mathbf{x}$. Following standard terminology, F_g will be called a (continuous) Markov prediction scheme of order k. Markov prediction schemes of various sorts have received considerable attention in the literature. Markov predictors for binary sequences under Hamming loss were considered by Cover and Shenhar [9]. Connections between finite state and Markov schemes for the prediction of binary processes are studied by Feder, Merhav and Gutman [18]; see also the survey [28]. Markov schemes also play a role in the prediction of ergodic processes, see *e.g.* [32, 1, 30] and the references therein.

We restrict our attention to the ordinary squared loss, and our results rely in part on its numerical properties. (The question of whether results like those obtained here hold for general loss functions is open.) If a Markov prediction scheme F_g is applied successively to the first n terms of $\mathbf{x}$, then its average cumulative loss is given by

$$L_n(g) \;=\; L_n(g, \mathbf{x}) \;=\; \frac{1}{n}\sum_{t=1}^{n}(F_g(x^{t-1}) - x_t)^2 \;=\; \frac{1}{n}\sum_{t=k+1}^{n}(g(x_{t-k}^{t-1}) - x_t)^2 \;+\; \frac{1}{n}\sum_{t=1}^{k} x_t^2.$$

If g is identically equal to a constant a, we will write $L_n(a)$ and $L_n(a, \mathbf{x})$ for the average loss of F_g on x_1^n.

In order to compare the long-run average performance of two Markov prediction schemes F_g and F_h, of possibly different orders, it is natural to consider the asymptotic behavior of $L_n(h) - L_n(g)$. Let us say that h is as good as g if $\liminf_n(L_n(g) - L_n(h)) \geq 0$, in which case the average cumulative loss of g is, asymptotically, bounded below by the average cumulative loss of h. For more on this notion of comparative performance see [19, 13, 1] and the references therein.

Definition: A sequence $\mathbf{x}$ is memoryless if there exists a constant $c \in \mathbb{R}$ such that

$$\liminf_{n\to\infty} \left[L_n(g, \mathbf{x}) - L_n(c, \mathbf{x})\right] \;\geq\; 0 \tag{2}$$

for every $k \geq 0$ and every $g \in \mathcal{C}_k$. Neither $L_n(g, \mathbf{x})$ nor $L_n(c, \mathbf{x})$ is assumed to converge in the limit of increasing n.

By definition, a sequence is memoryless if the best constant prediction scheme is as good as any continuous Markov prediction scheme of finite order. Put another way, no continuous function of finitely many arguments can outperform the scheme that ignores the past and always predicts the next value of $\mathbf{x}$ by c. In this sense, a memoryless sequence is unpredictable by any finite state Markov predictor. One may also view (2) as a sequence (indexed by k) of asymptotic tests for a potential pseudo-random sequence $\mathbf{x}$.

The comparative measure used to define memoryless sequences is somewhat weak, as it involves only long run average loss. More stringent comparative performance measures (*c.f.* [13, 34, 35]) would give rise to a smaller class of sequences with different asymptotic properties; these variants will be explored in subsequent work.

Remark: It can readily be shown that a bounded sequence $\mathbf{x}$ is memoryless if and only if for some constant c the two-sided sequence $\tilde{\mathbf{x}} = \ldots, 0, 0, x_1, x_2, \ldots$ is such that

$$\liminf_{n\to\infty} \left[\frac{1}{n} \sum_{i=1}^{n} (g(\tilde{x}_{-\infty}^{i-1}) - \tilde{x}_i)^2 \;-\; \frac{1}{n} \sum_{i=1}^{n} (\tilde{x}_i - c)^2 \right] \;\geq\; 0$$

for every bounded function $g : \mathbb{R}^\infty \to \mathbb{R}$ that is continuous with respect to the usual product topology on $\mathbb{R}^\infty$. For bounded sequences, there is no loss of generality in considering continuous predictors with finitely many arguments.

1.2 Markov Sequences

It is possible to extend the notion of memoryless sequence to account for sequences with Markov structure. Define $\mathbf{x}$ to be k'th order Markov if k is the least integer $l \geq 0$ for which there exists $h \in \mathcal{C}_l$ such that

$$\liminf_{n\to\infty} \left[L_n(g, \mathbf{x}) - L_n(h, \mathbf{x}) \right] \;\geq\; 0$$

for every $g \in \cup_{r\geq 0} \mathcal{C}_r$. If such an optimal predictor exists, it can be shown that the residual sequence $\mathbf{y} = \{y_i = x_i - h(x_{i-k}^{i-1}) : i \geq k+1\}$ is memoryless with $c = 0$, and the results of the paper will apply to $\mathbf{y}$.

1.3 Overview of Some Related Work

Numerous examples of sequential prediction problems for stationary and more general processes can be found in the literature; see for example [32, 1, 34, 28] and the references therein. An account of stochastic and non-stochastic sequential decision problems, and their relation to calibration and foundational questions in Statistics, can be found in the

work of Dawid [11, 12, 13]. Merhav and Feder [28] give an overview of prediction from individual sequences.

While there are so-called "universal" prediction schemes for the family of stationary ergodic processes taking values in a given finite set [32], or satisfying suitable moment conditions [1, 2, 21, 30], no such schemes exist for individual sequences. In the latter context, attention shifts from absolute to comparative measures of performance. A central problem in the comparative framework is how to construct a single scheme that competes favorably with every scheme in a given family $\mathcal{F}$ on every individual sequence. In many cases, this may be accomplished by suitably combining, or aggregating, the decisions of the individual schemes in $\mathcal{F}$. Representative work on aggregating methods for a variety of settings and loss functions, and further references, can be found in [32, 39, 26, 7, 6, 22]. Foster and Vohra [20] give an account of the aggregating problem and its history. Cesa-Bianchi and Lugosi [4, 5] have applied probabilistic techniques to the analysis of regret bounds for individual sequences.

Dawid and Vovk [14] describe a game theoretic approach to probability they call the "prequential framework". They establish game theoretic generalizations of several classical results, including the martingale law of large numbers, the central limit theorem, and the law of the iterated logarithm. While their results apply to individual sequences of moves made in a three-player game, they do not identify or characterize particular sequences with stochastic properties. Vovk [40] establishes similar results in a related game theoretic setting. We describe these results in more detail. Consider a perfect information game, where the t'th round of play proceeds sequentially as follows: an expert issues a forecast $e_t \in [0,1]$, a learner issues a forecast $l_t \in [0,1]$, and a third player (termed "nature") reveals a value $\omega_t \in [0,1]$. The numbers chosen by the expert, forecaster, and nature can depend on all the prior values in the game. The predictive performance of the expert and learner is judged by their cumulative squared or logarithmic loss. It is shown in [40] that under either loss the learner has a strategy ensuring that, for every sequence of moves by the expert and nature, either the difference between his cumulative loss and that of the expert tends to minus infinity, or $n^{-1}\sum_{i=1}^{n}(e_t - w_t) \to 0$. In other words, if the expert's predictions do not match the revealed sequence, the learner's predictions will be superior to those of the expert. As a corollary of this result, one obtains the standard martingale law of large numbers. A similar result is established for the law of the iterated logarithm.

There is a large body of literature on the definition and characterization of random individual sequences, beginning with work of Von Mises [29], and continuing with work of

Kolmogorov [24], Martin-Löf [27], and others. An account of this work can be found in the survey [37], see also [42, 43]. In this context, Vovk [38, 40] establishes a law of large numbers and a law of the iterated logarithm for binary sequences with maximal algorithmic complexity (see Section 2.3 for further discussion). Ryabko [33] considers some connections between effective prediction, Kolmogorov complexity, and Hausdorff dimension. V'yugin [41] establishes an ergodic theorem for Martin-Löf random (typical) individual sequences when the relevant measure, transformation, and test function are all computable.

1.4 Outline

The next Section describes the principal results of the paper. An an elementary law of large numbers for memoryless sequences is presented in Section 2.1. It is shown in Section 2.2 that binary memoryless sequences have convergent relative frequencies of every order, and that for each fixed order the limiting measures are Bernoulli. Some connections between memoryless and incompressible sequences are also discussed. Additional stochastic properties of memoryless sequences are presented in Section 2.3, including sliding block versions of the central limit theorem and Hoeffding's inequality. Several examples of memoryless sequences, and a construction of memoryless binary sequences from aggregating methods for sequential prediction, are given in Section 3. Proofs of the principal results are given in Section 4

2 Overview of Principal Results

2.1 A Law of Large Numbers

Our first result shows that the constant c appearing in (2) is unique.

Lemma 1 *If* $\mathbf{x}$ *is memoryless, then the optimal constant predictor* c *is unique.*

Definition: If $\mathbf{x}$ is memoryless, let $c(\mathbf{x})$ be the (unique) centering constant satisfying (2). Note that $c(\alpha\,\mathbf{x}+\beta) = \alpha\,c(\mathbf{x}) + \beta$ for $\alpha, \beta \in \mathbb{R}$, where $\alpha\mathbf{x}+\beta = \alpha x_1 + \beta, \alpha x_2 + \beta, \ldots$.

Memoryless sequences satisfy an elementary law of large numbers, in which the centering constant plays the role of an expectation.

Proposition 1 *If* $\mathbf{x}$ *is memoryless, then* $n^{-1}\sum_{i=1}^{n} x_i \to c(\mathbf{x})$

On the other hand, the law of large numbers may not hold for continuous functions of the elements of $\mathbf{x}$. Here is a simple example. Let $X_1, X_2, \ldots$ be i.i.d. with $P\{X_i = 1\} = P\{X_i = -1\} = 1/2$. For $j \geq 0$ define $Y_i = X_i$ when $2^j \leq i < 2^{j+1}$ and j is even, and $Y_i = 0$ otherwise. Then, with probability one, $Y_1, Y_2, \ldots$ is memoryless but Y_i^2 fails to converge. (For a deterministic example, we may replace $\{Y_i\}$ by a translated version of one of the memoryless binary sequences described in Section 3.2.)

2.2 Memoryless Binary Sequences

A binary sequences $\mathbf{x} = x_1, x_2, \ldots \in \{0, 1\}$ is memoryless if there exists a constant $c \in [0, 1]$ such that (2) holds for every $k \geq 1$ and every function $g : \{0, 1\}^k \to [0, 1]$. In contrast with the real-valued case, memoryless binary sequences have weakly convergent empirical distributions of every finite order, and the weak limit of the m-dimensional distributions is a Bernoulli measure with $p = c(\mathbf{x})$.

Theorem 1 *If* $\mathbf{x}$ *is a memoryless binary sequence with* $c(\mathbf{x}) = p$*, then for every* $m \geq 1$ *and every sequence* $b_1, \ldots, b_m \in \{0, 1\}$

$$\lim_n \frac{1}{n} \sum_{i=0}^{n-1} I\left\{x_{i+1}^{i+m} = b_1^m\right\} \;=\; \prod_{j=1}^{m} p^{b_j} (1-p)^{1-b_j}.$$

The conclusion of Theorem 1 applies to any memoryless sequence taking values in a two-element set. The example following Proposition 1 shows that Theorem 1 may fail for sequences taking values in the set $\{-1, 0, +1\}$.

2.2.1 Connections with Finite-State Complexity and Compressibility

Memoryless binary sequences have direct connections with the the theories of finite state complexity, compressibility, and predictability. Ziv [44] defined a notion of finite state complexity for an individual sequence with values in a finite alphabet, and established the central role of this quantity in fixed-rate coding theorems for individual sequences. We briefly recall his definition for binary sequences. Fix $\mathbf{x} = x_1, x_2, \ldots$ with $x_i \in \{0, 1\}$. Let $h_k(\mathbf{x}) = k^{-1} \log_2 N_k(\mathbf{x})$, where $N_k(\mathbf{x}) \leq 2^k$ is the number of distinct binary k-blocks appearing in $\mathbf{x}$, and define $h(\mathbf{x}) = \lim_k h_k(\mathbf{x})$. For $\delta > 0$, let $H_\delta(\mathbf{x})$ be the infimum of $h(\mathbf{y})$ over all binary sequences $\mathbf{y}$ that are close to $\mathbf{x}$ in the sense that $\limsup_n n^{-1} \sum_{i=1}^n I\{x_i \neq y_i\} < \delta$. The finite state complexity of $\mathbf{x}$ is defined by $H(\mathbf{x}) = \lim_{\delta \to 0} H_\delta(\mathbf{x})$, and always lies between zero and one. By a straightforward modification of the proof of Theorem 5 in [44], one may establish the following result.

Proposition 2 *If* $\mathbf{x}$ *is a memoryless binary sequence with* $c(\mathbf{x}) = p$ *then* $H(\mathbf{x}) = -p \log p - (1-p)\log(1-p)$ *is the ordinary binary entropy of* p.

Ziv and Lempel [45] defined a related notion of finite state compressibility for individual sequences, and studied its connection with variable rate coding of individual sequences (see also [17]). Like the complexity $H(\mathbf{x})$, the compressibility $\rho(\mathbf{x})$ of a binary sequence $\mathbf{x}$ is between zero and one. It follows immediately from Theorem 3 of [45] that a memoryless binary sequence $\mathbf{x}$ with $c(\mathbf{x}) = p$ has compressibility $\rho(\mathbf{x})$ equal to the binary entropy of p.

Memoryless binary sequences with $c(\mathbf{x}) = 1/2$ play a distinguished role, analogous to that of a sequence of fair coin flips. If $\mathbf{x}$ is such a sequence, then $H(\mathbf{x}) = \rho(\mathbf{x}) = 1$, the maximum possible value. Moreover, as shown in [18], the finite-state predictability $\pi(\mathbf{x})$ of $\mathbf{x}$ is also maximized, and equal to $1/2$. It follows that no finite-state prediction scheme for $\mathbf{x}$ can do better under the Hamming loss than the simple scheme that always predicts 0, regardless of the past values of the sequence.

2.3 Sliding Block Central Limit Theorem

Here we present a sliding block central limit theorem for memoryless sequences. In order to establish this result for an unbounded sequence $\mathbf{x}$, we need to impose moment conditions on its entries.

Definition: A real valued sequence $\mathbf{x}$ has an empirical moment of order $t \geq 0$ if

$$\gamma(t) \;=\; \sup_{n \geq 1} \frac{1}{n} \sum_{i=1}^{n} |x_i|^t \;<\; \infty \tag{3}$$

and an empirical moment of order $t+$ if

$$\gamma(t, \phi) \;=\; \sup_{n \geq 1} \frac{1}{n} \sum_{i=1}^{n} |x_i|^t \phi(|x_i|) \;<\; \infty \tag{4}$$

for some strictly increasing continuous function ϕ such that $\phi(u) \to \infty$ as $u \to \infty$.

Condition (3) is equivalent to $\limsup_n n^{-1} \sum_{i=1}^{n} |x_i|^t < \infty$, and is implied by the convergence of $n^{-1} \sum_{i=1}^{n} |x_i|^t$ to a finite constant. Similar remarks apply to (4). Clearly, bounded sequences have empirical moments of every order.

Now let $\mathbf{x}$ be a fixed sequence of real numbers. For $i, m \geq 1$ define the normalized partial sums

$$s_{i,m} \;=\; \frac{1}{m^{1/2}} \sum_{j=1}^{m} x_{i+j} \, .$$

Let $\eta(n,m)$ be the empirical measure of $s_{1,m},\ldots,s_{n,m}$, *i.e.*, for each Borel subset A of $\mathbb{R}$,

$$\eta(n,m)(A) \;=\; \frac{1}{n}\sum_{i=1}^{n} I\{s_{i,m}\in A\}.$$

Note that $\eta(n,m)(A)$ depends only on $x_1,\ldots,x_{n+m}$. Let $\rho(\cdot;\cdot)$ be any metric for probability distributions on $(\mathbb{R},\mathcal{B})$ that is compatible with weak convergence, *i.e.*, $\nu_n \Rightarrow \nu$ if and only if $\rho(\nu_n;\nu)\to 0$. One example is the Prohorov metric,

$$\rho(\nu;\eta) \;=\; \inf\{\epsilon>0 \text{ s.t. } \nu(A)\le \eta(A^\epsilon) \text{ for every } A\in\mathcal{B}\},$$

where $A^\epsilon=\{u : |u-v|<\epsilon \text{ for some } v\in A\}$ is the ϵ-blow-up of A. (See Chapter 11 of [16] for more details.) Let $\mathcal{N}(\alpha,\sigma^2)$ denote a normal distribution with mean α and variance σ^2.

Theorem 2 *Let* $\mathbf{x}$ *be an individual numerical sequence having an empirical moment of order* $2+$*, and let* $\alpha=c(\mathbf{x})$*. If* $\mathbf{x}$ *and* $(\mathbf{x}-\alpha)^2=(x_1-\alpha)^2,(x_2-\alpha)^2,\ldots$ *are memoryless, then*

$$\lim_{m\to\infty}\limsup_{n\to\infty}\rho(\,\eta(n,m)\,;\,\mathcal{N}(\alpha,\sigma^2)\,) \;=\; 0.$$

where $\sigma^2=c((\mathbf{x}-\alpha)^2)$*. Equivalently, for every* $\epsilon>0$ *there exists a block length* m_0 *(depending on* ϵ*) and sample size* n_0 *(depending on* ϵ *and* m_0*) such that* $\rho(\,\eta(n,m_0),\mathcal{N}(\alpha,\sigma^2)\,)<\epsilon$ *for each* $n\ge n_0$*.*

Remark: Theorem 2 is expressed in terms of double limits: the first is taken as the number n of sliding blocks increases, and the second is taken with increasing block size m. The first limit accounts for the stochastic behavior of the sequence, essentially replacing the m-dimensional distribution of a random sequence with the limiting empirical distribution of the m blocks of $\mathbf{x}$. The second limit corresponds to increasing sample size.

Vovk [38] established a law of the iterative logarithm for binary sequences $\mathbf{x}$ such that $K(x_1^n)=n+O(1)$, where $K(x_1^n)$ denotes the Kolmogorov complexity (*c.f.* [10]) of the sequence $x_1,\ldots,x_n$. Sequences of this sort are called Kolmogorov random; by definition, the complexity of their prefixes grow at the largest possible rate. A related result, under a different notion of complexity, is given in [40]. In general, being Kolmogorov random is a stronger condition than being memoryless. To see this, consider the Champernowne sequence $\mathbf{x}^*$, obtained by concatenating the binary representations of the successive integers $0,1,2,\ldots$. We argue in Section 3.2 below that $\mathbf{x}^*$ is memoryless. On the other hand, as was noted in [44], $K(x_1^*,\ldots,x_n^*)=O(1)$, so that $\mathbf{x}^*$ is far from being Kolmogorov random. This distinction between Kolmogorov random and memoryless sequences arises from the

fact that memoryless sequences need only be "complex" with respect to continuous finite-order Markov schemes; they may be easy to predict if we allow any computable prediction scheme.

Arguments like those for Theorem 2 may also be used to derive a sliding block version of Hoeffding's inequality for sums of bounded independent random variables.

Proposition 3 *If* $\mathbf{x}$ *is a memoryless sequence with* $x_i \in [a,b]$*, then for every* $m \geq 1$ *and* $t > 0$*,*

$$\limsup_{n\to\infty} \frac{1}{n} \sum_{i=0}^{n-1} I\left\{ \left| \frac{1}{m} \sum_{j=1}^{m} x_{i+j} \right| > t \right\} \leq 2e^{-mt^2/2(b-a)^2}.$$

3 Examples of Memoryless Sequences

3.1 Sample Paths of Random Sequences

The most natural source of memoryless sequences are the sample paths of bounded martingale differences. The following well known stability result is an immediate consequence of Hoeffding's inequality [23].

Lemma A *If* $Y_1, Y_2, \ldots$ *is a bounded martingale difference sequence, then* $n^{-1}\sum_{i=1}^{n} Y_i \to 0$ *with probability one.*

Proposition 4 *If* $\mathbf{X} = X_1, X_2, \ldots$ *is a bounded martingale difference sequence, then* ω*-almost every trajectory* $\mathbf{x} = \mathbf{X}(\omega)$ *is a memoryless sequence with* $c(\mathbf{x}) = 0$*.*

We remark that the process $\mathbf{X}$ in Proposition 4 need not be stationary. Under suitable integrability conditions, one may establish analogous results for unbounded sequences using extensions of Lemma A such as those in [36, 1].

3.2 Non-Random Examples

Proposition 4 is non-constructive, and its conclusion is restricted to "typical" sample paths, which cannot be identified in a finitary fashion. One would evidently like to exhibit specific memoryless sequences in a way that does not rely on a prior stochastic process. A canonical example of this latter sort is the binary sequence $\mathbf{x}^*$ formed by concatenating, in order, the standard binary representations of the non-negative integers $0, 1, 2, 3, ...$, *i.e.* $\mathbf{x}^* = 011011100101....$ In [8], Champernowne showed that the sequence 012345... is base-10 normal; a straightforward modification of his argument shows that the limiting relative

frequency of any binary m-block in $\mathbf{x}^*$ is equal to 2^{-m}, and the memoryless property of $\mathbf{x}^*$ easily follows. As a corollary, we deduce that $\mathbf{x}^*$ also satisfies the sliding block CLT and Hoeffding inequality. A discussion of the compressibility and other coding-theoretic properties of $\mathbf{x}^*$ can be found in [44, 45]. Lehrer [25] gives a general method of constructing individual sequences that are normal (generic) for a given finite or countable alphabet stochastic process. For the simple case of a Bernoulli process with success probability p, his construction provides further examples of memoryless sequences.

For additional non-random examples of memoryless sequences, we turn to aggregating methods for sequential prediction schemes. Recall that aggregating methods combine, or aggregate, the forecasts of a family of prediction schemes to produce a single prediction scheme that competes favorably with every scheme in the family. Here it is shown how aggregating methods can be used to construct a memoryless binary sequence. The essential idea of the construction is this: if an individual sequence cannot be reliably predicted by an aggregate scheme, then it cannot be reliably predicted by any expert in the family.

For the purposes of constructing binary sequences we may, without loss of generality, restrict our attention to families D_k, $k \geq 1$, consisting of functions $h : \{0,1\}^k \to [0,1]$. Let $D_k^o \subseteq D_k$ be the countable sub-family consisting of functions taking values in the rational numbers. The next result follows immediately from existing results (c.f. [19]) on aggregating prediction schemes.

Theorem A *There exists a function (prediction scheme) $H : \cup_{k\geq 1}\{0,1\}^k \to [0,1]$ such that*

$$\liminf_{n\to\infty} \left[L_n(h,\mathbf{x}) \; - \; \frac{1}{n}\sum_{t=1}^{n} (H(x_1^t) - x_{t+1})^2 \right] \; \geq \; 0 \tag{5}$$

for every $h \in \cup_{k\geq 1} D_k^o$ and every sequence $\mathbf{x} \in \{0,1\}^\infty$.

Remark: It is clear from the definition of D_k^o that, in fact, (5) holds for every $h \in \cup_{k\geq 1} D_k$. Thus the asymptotic loss of the prediction scheme H is no worse than the asymptotic loss of any finite state Markov prediction scheme. It should be noted that H itself is *not* Markov: for each $t \geq 1$, the prediction $H(x_1^t)$ will depend on the entire past $x_1, \ldots, x_t$.

Using H, define an individual sequence $\mathbf{x} \in \{0,1\}^\infty$ recursively as follows. Let $x_1 = 0$, and for $t \geq 1$ set

$$x_t \; = \; \begin{cases} 0 & \text{if } H(x_1^{t-1}) > 1/2 \\ 1 & \text{if } H(x_1^{t-1}) \leq 1/2 \end{cases}$$

By definition of $\mathbf{x}$, for each $n \geq 1$,

$$L_n(H, \mathbf{x}) \;=\; \frac{1}{n}\sum_{t=1}^{n}(H(x^{t-1}) - x_t)^2 \;\geq\; \frac{1}{4} \;=\; L_n(1/2, \mathbf{x}).$$

It then follows from (5) that $\liminf_n [L_n(h, \mathbf{x}) - L_n(1/2, \mathbf{x})] \geq 0$ for every $h \in \cup_{k\geq 1} D_k$. Thus $\mathbf{x}$ is memoryless, with $c(\mathbf{x}) = 1/2$.

4 Proofs of Principal Results

4.1 Uniqueness of c

Lemma 1 If $\mathbf{x}$ is memoryless, then the optimal constant predictor c appearing in (2) is unique.

Proof: If c_1 and c_2 satisfy (2) for some memoryless sequence $\mathbf{x}$, then

$$\begin{aligned} 0 \;&\leq\; \liminf_{n\to\infty} [L_n(c_1) - L_n(c_2)] \;\leq\; \limsup_{n\to\infty} [L_n(c_1) - L_n(c_2)] \\ &=\; -\liminf_{n\to\infty} [L_n(c_2) - L_n(c_1)] \;\leq\; 0. \end{aligned}$$

Thus $L_n(c_1) - L_n(c_2) \to 0$ as $n \to \infty$. Let $c = (c_1 + c_2)/2$. An elementary calculation shows that $L_n(c) = L_n(c_1)/2 + L_n(c_2)/2 - (c_1 - c_2)^2/4$, and therefore

$$L_n(c) - L_n(c_1) \;=\; \frac{1}{2}\left(L_n(c_2) - L_n(c_1)\right) - \frac{1}{4}(c_1 - c_2)^2.$$

It then follows from the optimality of c_1 that

$$0 \leq \liminf_{n\to\infty} [L_n(c) - L_n(c_1)] \;=\; -\frac{1}{4}(c_1 - c_2)^2$$

from which we conclude that $c_1 = c_2$.

The following elementary lemma provides a connection between memoryless sequences and martingale differences (see Proposition 3 below). We give a proof for completeness.

Lemma 2 *Let $\mathbf{x}$ be any real valued sequence and let $\mathcal{Y} \subseteq \mathbb{R}^\infty$ be any family of bounded sequences that is closed under scaler multiplication, i.e., if $\mathbf{y} \in \mathcal{Y}$ then $\alpha\mathbf{y} \in \mathcal{Y}$ for every $\alpha \in \mathbb{R}$. Then*

$$\liminf_n \left[\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2 - \frac{1}{n}\sum_{i=1}^{n} x_i^2\right] \;\geq\; 0 \quad \textit{for every} \quad \mathbf{y} \in \mathcal{Y} \tag{6}$$

if and only if

$$\frac{1}{n}\sum_{i=1}^{n} x_i\, y_i \;\to\; 0 \;\; \textit{for every} \;\; \mathbf{y} \in \mathcal{Y}. \tag{7}$$

Proof: Note that $(x_i - y_i)^2 - x_i^2 = y_i^2 - 2\,x_i\,y_i$, so clearly (7) implies (6). Suppose that $n^{-1}\sum_{i=1}^n x_i\,y_i \not\to 0$ for some $\mathbf{y} \in \mathcal{Y}$. Replacing $\mathbf{y}$ by $-\mathbf{y}$ if necessary, we may assume that $\limsup_n n^{-1}\sum_{i=1}^n x_i\,y_i \geq \delta$ for some $\delta > 0$. For each $\alpha > 0$,

$$\begin{aligned}
\liminf_n \left[\frac{1}{n}\sum_{i=1}^n (x_i - \alpha y_i)^2 - \frac{1}{n}\sum_{i=1}^n x_i^2\right] &= \liminf_n \left[\frac{\alpha^2}{n}\sum_{i=1}^n y_i^2 - \frac{2\alpha}{n}\sum_{i=1}^n x_i\,y_i\right] \\
&\leq \liminf_n \left[\alpha^2\,||\mathbf{y}||_\infty^2 - \frac{2\alpha}{n}\sum_{i=1}^n x_i\,y_i\right] \\
&= \alpha^2\,||\mathbf{y}||_\infty^2 \;-\; 2\alpha \limsup_n \left[\frac{1}{n}\sum_{i=1}^n x_i\,y_i\right] \\
&\leq \alpha^2\,||\mathbf{y}||_\infty^2 \;-\; 2\alpha\delta.
\end{aligned}$$

The last term above is negative if $0 < \alpha < 2\delta/||\mathbf{y}||_\infty^2$, and therefore (6) fails to hold.

Proposition 1 If $\mathbf{x}$ is memoryless, then $n^{-1}\sum_{i=1}^n x_i \to c(\mathbf{x})$.

Proof: Immediate from Lemma 2.

4.2 Empirical Measures and Weak Convergence

Every individual sequence $\mathbf{x}$ is naturally associated with a family of empirical measures, and the stochastic properties of the sequence are determined by the limiting behavior of these measures with increasing sample size. For $n, k \geq 1$ define the n-sample, k-dimensional empirical measure of $\mathbf{x}$ by

$$\mu_{n,k}(A) \;=\; \frac{1}{n}\sum_{i=0}^{n-1} I\{(x_{i+1}, \ldots, x_{i+k}) \in A\} \quad A \subseteq \mathbb{R}^k$$

We will make use of some basic definitions and results from the standard theory of weak convergence for random vectors. We recall a few key definitions below, and refer the interested reader to [16, 3] for more details. A sequence $\{\nu_n : n \geq 1\}$ of probability measures on $\mathbb{R}^k$ is said to converge weakly to a limiting probability measure ν, written $\nu_n \Rightarrow \nu$, if $\int g\,d\nu_n \to \int g\,d\nu$ for every $g \in \mathcal{C}_k$. (In what follows, $\{\nu_n\}$ will most often be the k-dimensional empirical measures of a given sequence $\mathbf{x}$, or a subsequence of these measures.) Likewise, a sequence of random vectors $X_1, X_2, \ldots \in \mathbb{R}^k$ converges weakly to X if $Eg(X_n) \to Eg(X)$ for every $g \in \mathcal{C}_k$. To every measure ν on $\mathbb{R}^k$ there correspond random variables $X_1, \ldots, X_k$ defined on some underlying probability space and having ν as their joint distribution; in this case we will write $\nu_n \Rightarrow \nu$ equivalently as $\nu_n \Rightarrow (X_1, \ldots, X_k)$.

A sequence $\{\nu_n : n \geq 1\}$ of probability measures on $\mathbb{R}^k$ is said to be (uniformly) tight if for every $\epsilon > 0$ there is a compact set $K \subseteq \mathbb{R}^k$ such that $\sup_{n\geq 1} \nu(K^c) \leq \epsilon$.

Prohorov's theorem states that $\{\nu_n\}$ is tight if and only if every subsequence $\{\nu_{n_k}\}$ has a further subsequence $\{\nu_{m_k}\}$ that is weakly convergent. Note that tightness of the empirical measures $\{\mu_{n,1}\}$ implies tightness of $\{\mu_{n,k}\}$ for each $k \geq 1$. A sequence $\mathbf{x}$ is bounded if $||\mathbf{x}||_\infty = \sup_{i\geq 1} |x_i| < \infty$. The empirical measures of a bounded sequence are automatically tight.

4.3 Memoryless Sequences and Martingale Differences

Lemma 3 *Let $\mathbf{x}$ be an individual sequence having an empirical moment of order $1+$. Then the following are equivalent.*

(a) $\mathbf{x}$ is memoryless and $c(\mathbf{x}) = 0$.

(b) For every $j \geq 1$ and every $g \in \mathcal{C}_j$, $n^{-1}\sum_{i=0}^{n-1} g(x_{i+1}^{i+j})\, x_{i+j+1} \to 0$.

(c) For $m \geq 1$ every weak limit $X_1, \ldots, X_m$ of $\{\mu_{n,m} : n \geq 1\}$ is a stationary martingale difference sequence.

Proof: Let $\mathcal{Y}$ be the set of sequences $\mathbf{y} = y_1, y_2, \ldots$ such that $y_1 = \cdots = y_m = 0$ and $y_j = g(x_{j-m}^{j-1})$ for $j > m$ and some fixed $g \in \mathcal{C}_m$. Then (b) follows from (a) using Lemma 2 and the memoryless property of $\mathbf{x}$.

Suppose that (b) holds and that $\mu_{n_k,m} \Rightarrow (X_1, \ldots, X_m)$. Then for each $s, j \geq 1$ with $s + j \leq m$, and every $g \in \mathcal{C}_j$,

$$\begin{aligned} Eg(X_s, \ldots, X_{s+j}) &= \lim_k \frac{1}{n_k} \sum_{i=0}^{n_k-1} g(x_{i+s}, \ldots, x_{i+s+j}) \\ &= \lim_k \frac{1}{n_k} \sum_{i=0}^{n_k-1} g(x_{i+1}, \ldots, x_{i+j+1}) \;=\; Eg(X_1, \ldots, X_{j+1}). \end{aligned}$$

It follows that $(X_s, \ldots, X_{s+j})$ has the same joint distribution as $(X_1, \ldots, X_{j+1})$, and as this is true for each choice of s, j above, $X_1, \ldots, X_m$ is stationary. To establish the martingale difference property, let $(X_{1,k}, \ldots, X_{m,k}) \sim \mu_{n_k,m}$. Fix $1 \leq j < m$ and $g \in \mathcal{C}_j$. Then $g(X_{1,k}, \ldots, X_{j,k})\, X_{j+1,k} \Rightarrow g(X_1, \ldots, X_j)\, X_{j+1}$ by the continuous mapping theorem, and the $1+$ empirical moment condition ensures that the convergent sequence is uniformly integrable. Therefore, (b) implies that

$$\begin{aligned} Eg(X_1, \ldots, X_j)\, X_{j+1} &= \lim_k Eg(X_{1,k}, \ldots, X_{j,k})\, X_{j,k} \qquad (8) \\ &= \lim_k \frac{1}{n_k} \sum_{i=0}^{n_k-1} g(x_{i+1}^{i+j})\, x_{i+j+1} \;=\; 0. \end{aligned}$$

Suppose now that $f : \mathbb{R}^j \to [0,1]$ is any measurable function and let $g_1, g_2, \ldots \in \mathcal{C}_j$ be continuous functions with values in $[0,1]$ such that $g_k(X_1^j) \to f(X_1^j)$ with probability one.

Using (8) we find that

$$|Ef(X_1^j)\,X_{j+1}| \;=\; |Ef(X_1^j)\,X_{j+1} \;-\; Eg_k(X_1^j)\,X_{j+1}| \;\leq\; E[\,|X_{j+1}|\,|f(X_1^j) - g_k(X_1^j)|\,]$$

The last term above tends to zero as $k \to \infty$ by the dominated convergence theorem. It follows that $EX_{j+1}I_A = 0$ for every event A in the sigma-field generated by $X_1, \ldots, X_j$, and consequently $E[X_{j+1}\,|\,X_1^j] = 0$. That (c) implies (a) may be established by similar methods, but as this will not be needed in what follows, we omit the proof.

4.4 Binary Sequences

Theorem 1 If $\mathbf{x}$ is a memoryless binary sequence with $c(\mathbf{x}) = p$, then for every $m \geq 1$ and every sequence $b_1, \ldots, b_m \in \{0, 1\}$,

$$\lim_n \frac{1}{n} \sum_{i=0}^{n-1} I\left\{x_{i+1}^{i+m} = b_1^m\right\} \;=\; \prod_{j=1}^{m} p^{b_j}\,(1-p)^{1-b_j}.$$

Proof: By a routine tightness argument, it suffices to show that for each $m \geq 1$ every weak limit $X_1, \ldots, X_m$ of $\{\mu_{n,m} : n \geq 1\}$ is a Bernoulli sequence with parameter p. To this end, note that $X_j \in \{0, 1\}$ with probability one. By Lemma 3,

$$P(X_j = 1 \mid X_1^{j-1}) \;=\; E(X_j \mid X_1^{j-1}) \;=\; c(\mathbf{x}) \;=\; p,$$

and therefore $P(X_j = 0 \mid X_1^{j-1}) = (1-p)$. The result follows by an application of the chain rule for probabilities.

4.5 Sliding Block Central Limit Theorem

Theorem 2 Let $\mathbf{x}$ be an individual numerical sequence having an empirical moment of order 2+, and let $\alpha = c(\mathbf{x})$. If $\mathbf{x}$ and $(\mathbf{x} - \alpha)^2 = (x_1 - \alpha)^2, (x_2 - \alpha)^2, \ldots$ are memoryless, then

$$\lim_{m\to\infty} \limsup_{n\to\infty} \rho(\,\eta(n,m)\,;\,\mathcal{N}(\alpha, \sigma^2)\,) \;=\; 0. \tag{9}$$

where $\sigma^2 = c((\mathbf{x} - \alpha)^2)$.

Proof: By considering the shifted sequence $(\mathbf{x} - \alpha)$, we may assume that $c(\mathbf{x}) = 0$. For each $m \geq 1$ select a subsequence $\{n_j\}$, depending on m, such that

$$\lim_{j\to\infty} \rho(\,\eta(n_j,m);\,\mathcal{N}(0,\sigma^2)\,) \;=\; \limsup_{n\to\infty} \rho(\,\eta(n,m);\,\mathcal{N}(0,\sigma^2)\,),$$

and such that the m-dimensional empirical measures $\{\mu_{n_j,m} : j \geq 1\}$ converge weakly to a jointly distributed sequence $(X_{1,m}, \ldots, X_{m,m})$. Let $S_m = m^{-1/2}(X_{1,m} + \cdots + X_{m,m})$. By definition,

$$\rho(\, S_m; \mathcal{N}(0,\sigma^2)\,) \;=\; \limsup_{n\to\infty} \rho(\, \eta(n,m); \mathcal{N}(0,\sigma^2)\,),$$

so it suffices to show that $S_m \Rightarrow \mathcal{N}(0,\sigma^2)$ as $m \to \infty$. By Lemma 3, $X_{1,m}, \ldots, X_{m,m}$ is a stationary martingale difference sequence. The desired weak convergence of S_m will follow from the standard martingale central limit theorem (see Chapter 8 in Pollard [31]) if one can show that as $m \to \infty$,

$$\frac{1}{m}\sum_{j=1}^{m} E(X_{j,m}^2 \,|\, X_{1,m}, \ldots, X_{j-1,m}) \;\to\; \sigma^2 \quad \text{in probability} \tag{10}$$

and for every $\epsilon > 0$,

$$\frac{1}{m}\sum_{j=1}^{m} E(X_{j,m}^2 I\{|X_{j,m}| > \sqrt{m}\epsilon\} \,|\, X_{1,m}, \ldots, X_{j-1,m}) \;\to\; 0 \quad \text{in probability.} \tag{11}$$

These relations are established below.

Let $\{o_m(1) : m \geq 1\}$ denote any sequence of constants tending to zero as m tends to infinity. By stationarity and Lemma 5, the expectation of the average in condition (11) is bounded by

$$E(X_{1,m}^2 I\{|X_{1,m}| > \sqrt{m}\epsilon\}) \;\leq\; \frac{E[X_{1,m}^2\, \phi^{1/2}(|X_{1,m}|)]}{\phi^{1/2}(\sqrt{m}\epsilon)} \;\leq\; \frac{\gamma(2,\phi^{1/2})}{\phi^{1/2}(\sqrt{m}\epsilon)} \;=\; o_m(1), \tag{12}$$

and (11) follows. Now let $\lambda_1 < \lambda_2 < \cdots$ be any sequence of constants such that $\lambda_m = o(m^{1/4})$ and $\lambda_m \to \infty$. Fix $m \geq 1$ and define $\tilde{X}_{j,m} = X_{j,m}\, I\{|X_{j,m}| \leq \lambda_m\}$. Let $\mathcal{S}_{j,m}$ denote the sigma field generated by $X_{1,m}, \ldots, X_{j,m}$. An elementary argument shows that

$$E\left|\frac{1}{m}\sum_{j=1}^{m} E(X_{j,m}^2 \,|\, \mathcal{S}_{j-1,m}) - \sigma^2\right| \;\leq\; A_m + B_m + C_m + E\left|\frac{1}{m}\sum_{j=1}^{m} X_{j,m}^2 - \sigma^2\right|$$

where

$$A_m \;=\; E\left|\frac{1}{m}\sum_{j=1}^{m} E(X_{j,m}^2\, I\{|X_{j,m}| > \lambda_m\} \mid \mathcal{S}_{j-1,m})\right|$$

$$B_m \;=\; E\left|\frac{1}{m}\sum_{j=1}^{m}\Big[E(\tilde{X}_{j,m}^2 \mid \mathcal{S}_{j-1,m}) - \tilde{X}_{j,m}^2\Big]\right| \qquad C_m \;=\; E\left|\frac{1}{m}\sum_{j=1}^{m} \tilde{X}_{j,m}^2\, I\{|X_{j,m}| > \lambda_m\}\right|.$$

By arguments like that for (12), $A_m, C_m \leq E(X_{1,m}^2 I\{|X_{1,m}| > \lambda_m\}) = o_m(1)$. As for B_m, the terms appearing in the corresponding sum are orthogonal and bounded by λ_m^2, so that

$$B_m^2 \;\leq\; \frac{1}{m^2}\sum_{j=1}^{m} E\Big[E(\tilde{X}_{j,m}^2 \,|\, \mathcal{S}_{j-1,m}) - \tilde{X}_{j,m}^2\Big]^2 \;\leq\; \frac{\lambda_m^4}{m} \;=\; o_m(1).$$

Putting these bounds together,

$$E\left|\frac{1}{m}\sum_{j=1}^{m} E(X_{j,m}^2 \,|\, \mathcal{S}_{j-1,m}) - \sigma^2\right| \;\le\; E\left|\frac{1}{m}\sum_{j=1}^{m} X_{j,m}^2 - \sigma^2\right| \;+\; o_m(1).$$

Let $\mathcal{S}'_{j,m}$ denote the sigma field generated by $X_{1,m}^2, \ldots, X_{j,m}^2$. An argument analogous to that above shows that

$$E\left|\frac{1}{m}\sum_{j=1}^{m} X_{j,m}^2 - \sigma^2\right| \;\le\; E\left|\frac{1}{m}\sum_{j=1}^{m} E(X_{j,m}^2 \,|\, \mathcal{S}'_{j-1,m}) - \sigma^2\right| \;+\; o_m(1).$$

Since $X_{1,m}^2 - \sigma^2, \ldots, X_{m,m}^2 - \sigma^2$ is a martingale difference sequence, the second term above is identically zero. Together, the last two displays imply (10).

The proof of Proposition 3 is similar to that of Theorem 2 and is ommitted.

4.6 Sample Paths of Martingale Differences

Proposition 4 If $\mathbf{X} = X_1, X_2, \ldots$ is a bounded martingale difference sequence, then ω-almost every trajectory $\mathbf{x} = \mathbf{X}(\omega)$ is a memoryless sequence with $c(\mathbf{x}) = 0$.

Proof: Fix $b < \infty$ such that $P(X_i \in [-b, b]) = 1$. Let $\mathcal{C}_k(b)$ be the family of continuous functions $g : [-b, b]^k \to \mathbb{R}$. As $[-b, b]$ is compact, $\mathcal{C}_k(b)$ contains a countable subfamily $\mathcal{C}_k^o(b)$ that is dense in the supremum norm. Lemma A implies that $n^{-1}\sum_{i=0}^{n-1} g(X_{i+1}^{i+k})\, X_{i+k+1} \to 0$ with probability one for every $g \in \cup_k \mathcal{C}_k^o(b)$, and therefore every $g \in \cup_k \mathcal{C}_k(b)$. The result follows from Lemma 3.

4.7 Tightness and Empirical Moments

Here we establish two technical results that provide a connection between tightness and empirical moments.

Lemma 4 *If* $\mathbf{x}$ *has an empirical moment of order* $0+$*, then the empirical measures* $\{\mu_{n,k}\}$ *are tight for each* $k \ge 1$*.*

Proof: It suffices to show that $\{\mu_{n,1}\}$ are tight. If (4) holds with $s = 0$, then for fixed $a > 0$ and each $n \ge 1$,

$$\mu_{n,1}([-a, a]^c) \;=\; \frac{1}{n}\sum_{i=1}^{n} I\{|x_i| > a\} \;\le\; \frac{1}{n}\sum_{i=1}^{n} \frac{\phi(|x_i|)}{\phi(a)}.$$

Thus $\sup_n \mu_{n,1}([-a, a]^c) \le \gamma(0, \phi)/\phi(a)$, which tends to zero as a tends to infinity.

Lemma 5 *Let* $\mathbf{x}$ *be an individual sequence having an empirical moment of order* $1+$ *with (continuous) growth function* ϕ*. If* $X_n \sim \mu_{n,1}$ *for* $n \geq 1$*, then* $\{X_n\}$ *and* $\{|X_n|\phi^{1/2}(|X_n|)\}$ *are uniformly integrable, and every weak limit* X *of* $\{X_n\}$ *satisfies* $E|X|\phi^{1/2}(|X|) \leq \gamma(1, \phi^{1/2})$.

Proof: For each $n \geq 1$ and $a > 0$,

$$E|X_n|\,\phi^{1/2}(|X_n|)I\{|X_n| > a\} \;=\; \frac{1}{n}\sum_{i=1}^{n} |x_i|\,\phi^{1/2}(|x_i|)I\{|x_i| > a\} \;\leq\; \frac{\gamma(1,\phi)}{\phi^{1/2}(a)}$$

which tends to zero uniformly in n as a tends to infinity. This establishes the second claim above; the first follows similarly. If $X_{n_k} \Rightarrow X$ then $|X_{n_k}|\,\phi^{1/2}(|X_{n_k}|) \Rightarrow |X|\,\phi^{1/2}(|X|)$ by the continuous mapping theorem, and the uniform integrability of $\{|X_{n_k}|\phi^{1/2}(|X_{n_k}|)\}$ ensures that

$$E|X|\,\phi^{1/2}(|X|) \;=\; \lim_k E|X_{n_k}|\,\phi^{1/2}(|X_{n_k}|) \;\leq\; \gamma(1,\phi^{1/2}),$$

which is finite by assumption.

References

[1] P.H. Algoet, "The strong law of large numbers for sequential decisions under uncertainty," *IEEE Trans. Info. Theory*, vol.40, pp.609-633, 1994.

[2] P.H. Algoet, "Universal schemes for learning the best nonlinear predictor given the infinite past and side information," *IEEE Trans. Info. Theory*, vol. 45, pp. 1165-1185, 1999.

[3] P. Billingsley, *Probability and Measure*, Third edition. Wiley, New York, 1995.

[4] N. Cesa-Bianchi and G. Lugosi, "On the prediction of individual sequences," *Ann. Stat.*, vol.27, pp.1865-1895, 2000.

[5] N. Cesa-Bianchi and G. Lugosi, "Worst-case bounds for the logarithmic loss of predictors," *Machine Learning*, vol.43, pp.247-264, 2001.

[6] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, and M.K. Warmuth, "On-line prediction and conversion strategies," *Machine Learning*, vol.25, pp.71-110, 1996.

[7] N. Cesa-Bianchi, Y. Freund, D. Haussler, D.P. Helmbold, R.E. Schapire, and M.K. Warmuth, "How to use expert advice," *J. Assoc. Comp. Mach.*, vol.44, pp.427-485, 1997.

[8] D.G. Champernowne, "The construction of decimals normal in the scale of ten," *J. London Math. Soc.*, vol.8, pp.254-260, 1933.

[9] T.M. Cover and A. Shenhar, "Compound Bayes predictors for sequences with apparent Markov structure," *IEEE Trans. Syst., Man, and Cyber.*, vol.7, pp.421-424, 1977.

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

[11] A.P. Dawid, "The well-calibrated Bayesian (with discussion)," *J. Amer. Stat. Assoc.*, vol.77, pp.605-613, 1982.

[12] A.P. Dawid, "Statistical theory. The prequential approach (with discussion)," *J. Roy. Statist. Soc. A*, vol.147, pp.278-292, 1984.

[13] A.P. Dawid, "Calibration-based empirical probability (with discussion)," *Ann. Statist.*, vol.13, pp.1251-1285, 1985.

[14] A.P. Dawid and V.G. Vovk, "Prequential probability: principles and properties," *Bernoulli*, vol.5, pp.125-162, 1999.

[15] L. Devroye, *Non-Uniform Random Variate Generation*, Springer-Verlag, New York, 1986.

[16] R. Dudley, *Real Analysis and Probability*, Chapman and Hall, New York, 1989.

[17] M. Feder, "Gambling using a finite state machine," *IEEE Trans. Info. Theory*, vol.37, pp.1459-1465, 1991.

[18] M. Feder, N. Merhav and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Info. Theory*, vol.38, pp.1258-1270, 1992.

[19] D.P. Foster, "Prediction in the worst case," *Ann. Statist.*, vol.19, pp.1084-1090, 1991.

[20] D.P. Foster and R. Vohra, "Regret in the on-line decision problem," *Games and Economic Behavior*, vol.29, pp.1084-1090, 1999.

[21] L. Györfi and G. Lugosi, "Strategies for sequential prediction of stationary time series," in *Modeling Uncertainty: an Examination of its Theory, Methods, and Applications*, M. Dror, P. L'Ecuyer and F. Szidarovszky Eds, Kluwer, 2001.

[22] D.H. Haussler, J. Kivenen, and M.K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. Info. Theory*, vol.44, pp.1906-1925, 1998.

[23] W. Hoeffding, "Probability inequalities for sums of bounded random variables," *Journal of the American Statistical Association*, vol.58, pp.13-30, 1963.

[24] A.N. Kolmogorov, "Three approaches to the definition of the concept 'quantity of information'," *Problems Inform. Trans.*, vol.1, pp.1-7, 1965.

[25] E. Lehrer, "The game of normal numbers," Preprint, 2001.

[26] N. Littlestone and M.K. Warmuth, "The weighted majority algorithm," *Info. and Comput.*, vol.108, pp.212-261, 1994.

[27] P. Martin-Löf, "On the concept of a random sequence," *Theory Probab. Appl.*, vol.11, pp.177-179, 1966.

[28] N. Merhav and M. Feder, "Universal Prediction," *IEEE Trans. Info. Theory*, vol.44, pp.2124-2147, 1998.

[29] R. von Mises, "Grundlagen der Wahrscheinlichkeitsrechnung," *Math Z.*, vol.5, pp.52-99, 1919.

[30] A.B. Nobel, "On optimal sequential prediction for general processes," *IEEE Trans. Info. Theory*, vol.49, pp.83-98, 2003.

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

[32] B.Y. Ryabko, "Prediction of random sequences and universal coding," *Problems Inform. Trans.*, vol.24, 87-96, 1988.

[33] B.Y. Ryabko, "The complexity and effectiveness of prediction algorithms," *J. Complexity*, vol.10, 281-295, 1994.

[34] K. Skouras and A.P. Dawid, "On efficient point prediction systems," *J. Roy. Statist. Soc. B*, vol.60, pp.765-780, 1998.

[35] K. Skouras and A.P. Dawid, "On efficient probability forecasting systems," *Biometrika*, vol.86, pp.765-784, 1999.

[36] W.F. Stout, *Almost Sure Convergence*, Academic Press, New York, 1974.

[37] V.A. Uspenskii, A.L. Semenov and A.Kh. Shen, "Can an individual sequence of zeroes and ones be random?," *Russian Math. Surveys*, vol.45, pp.121-189, 1990.

[38] V. Vovk, "The law of the iterated logarithm for random Kolmogorov, or chaotic, sequences," *Theory Probab. Appl.*, vol.32, pp.413-425, 1987.

[39] V. Vovk, "Aggregating strategies," In *Proc. 3'rd Annual Workshop on Computational Learning Theory*, pp. 371-383, Morgan Kaufman, San Mateo, 1990

[40] V. Vovk, "Probability theory for the Brier game," *Theor. Comp. Sci.*, vol.261, pp.57-79, 2001.

[41] V.V. V'yugin, "Effective convergence in probability and an ergodic theorem for individual random sequences," *Theory Probab. Appl.*, vol.42, pp.39-50, 1997.

[42] E.-H Yang and S. Shen, "Chaitin complexity, Shannon information content of a single event, and infinite random sequences–Part one," *Science in China Series A*, vol.34, pp.1183-1193, 1991.

[43] E.-H Yang and S. Shen, "Chaitin complexity, Shannon information content of a single event, and infinite random sequences–Part two," *Science in China Series A*, vol.34, pp.1307-1309, 1991.

[44] J. Ziv, "Coding theorems for individual sequences," *IEEE Trans. Info. Theory*, vol.24, pp.405-412, 1978.

[45] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Info. Theory*, vol.24, pp.530-536, 1978.