Universal Schemes for Sequential Decision from Individual Data Sequences

Neri Merhav, *Senior Member, IEEE*, and Meir Feder, *Senior Member, IEEE*

***Abstract*—Sequential decision algorithms are investigated, under a family of additive performance criteria, for individual data sequences, with various application areas in information theory and signal processing. Simple universal sequential schemes are known, under certain conditions, to approach optimality uniformly as fast as $n^{-1}\log n$, where n is the sample size. For the case of finite-alphabet observations, the class of schemes that can be implemented by finite-state machines (FSM's), is studied. It is shown that Markovian machines with sufficiently long memory exist that are asymptotically nearly as good as any given FSM (deterministic or randomized) for the purpose of sequential decision. For the continuous-valued observation case, a useful class of parametric schemes is discussed with special attention to the recursive least squares (RLS) algorithm.**

***Index Terms*—Sequential compound decision problem, empirical Bayes estimation, Bayes envelope, Bayes response, finite-state machines, Lempel–Ziv algorithm, recursive least squares.**

I. Introduction

MANY different problems that arise in information theory, signal processing, and control theory have the following generic form. An observer receives serially a sequence of measurements $x_1, x_2, \cdots$. At each time t, that is, after seeing x_{t-1}, he selects a *strategy* b_t from a given class $\boldsymbol{B}$ of permissible strategies, and the task is to minimize the long run time-average $n^{-1}\sum_{t=1}^{n} l(b_t, x_t)$ of a given loss function $l(\cdot,\cdot)$. If the measurements $\{x_t\}$ are governed by a known stationary ergodic probabilistic source and $\boldsymbol{B}$ allows any measurable function b_t of the past $(x_1,\cdots,x_{t-1})$, then the best strategy in the sense of minimizing the expected value of the time-average of the loss is clearly one that attains (or approaches) the least conditional expectation of $l(b_t, x_t)$ given the past. Moreover, this minimum loss is attainable almost surely [2] subject to certain regularity conditions on the loss function, even if the statistics of the source are not known *a priori*.

In this paper, we are concerned with the same sequential decision problem but in a deterministic rather than a probabilistic setting. The sequence $x_1, x_2, \cdots$ is considered as an individual, deterministic entity without any assumptions on the existence of an underlying statistical model. On the other hand, in order to incorporate in our model real-life limitations on computational power and memory resources, we confine the class $\boldsymbol{B}$ of the allowable nonanticipating strategies to certain structures with a limited number of degrees of freedom. Common examples of such classes include the class of strategies $b_t(x_1,\cdots,x_{t-1})$ that are implementable by deterministic finite-state machines (FSM's) (see, e.g., [11], [12], [24]–[26], [34], [49], [53]), parametric functions of the near past (in particular, linear functions [36], [38]), neural networks, etc.

For the case of FSM's, earlier work on data compression [53], gambling [11], and prediction [12] inspire the following extended setting of the deterministic sequential decision problem. For the first n outcomes $x_1,\cdots, x_n$, let $u_M(x_1,\cdots, x_n)$ be the minimum loss incurred by the best strategy that can be realized by a machine with M states ($M < n$). The limit supremum of this quantity as $n \to \infty$, that is $u_M(x_1, x_2,\cdots)$, describes the least asymptotic loss that an M-state strategy can guarantee. Finally,

$$u_\infty(x_1, x_2, \cdots) = \lim_{M\to\infty} u_M(x_1, x_2, \cdots)$$

is the least asymptotic loss achievable by an FSM with arbitrarily many states. While in this definition, the sequence of optimal M-state strategies may depend on the *entire* particular sequence of observables, we are primarily interested in a universal *sequential* decision scheme (strategy) that is independent of the particular sequence and yet attains $u_\infty(x_1, x_2, \cdots)$ in the long run. Later on we investigate the same problem when the class of strategies is extended to that of all *randomized* M-state machines. Analogous problems arise for parametric classes of strategies as described above. As an intermediate goal in the first problem, we seek a universal sequential strategy that asymptotically attains $u_1(x_1,\cdots,x_n)$, i.e., a scheme that is nearly as good as the best *fixed* (single-state) strategy for the given sequence.

It has been observed in some particular applications, that the dynamic selection of a strategy that best matches the data observed *so far* is asymptotically as good as the best fixed strategy that one could have used in retrospect. Moreover, in many cases, the performance of this dynamic strategy is within $O(n^{-1}\log n)$ close to optimality, uniformly for every possible data sequence of length n. Several useful examples of the deterministic sequential decision problem, where this phenomenon takes place, are the following.

The first example is related to universal data compression. Let $\boldsymbol{x}^n = (x_1, x_2, \cdots, x_n)$ be a given binary string to be

Manuscript received June 8, 1992; revised November 2, 1992. This work was supported in part by the Wolfson Research Awards administrated by the Israel Academy of Science and Humanities at Tel Aviv University. This work was presented in part at the 5th ACM Workshop on Computational Learning Theory, July 1992.

N. Merhav is with the Department of Electrical Engineering Technion—Israel Institute of Technology Haifa 32000, Israel.

M. Feder is with the Department of Electrical Engineering—Systems, Faculty of Electrical Engineering, Tel Aviv University, Tel Aviv 69 978, Israel.

IEEE 9209593.

compressed. Let $n_t(0)$ and $n_t(1)$ denote counts of "0" and "1," respectively, among the t first symbols of x^n $(t \leq n)$. Define $p_t(x) = (n_t(x) + 1/2)/(t+1)$, $x = 0, 1$, as the respective (biased) empirical probabilities of "0" and "1." A simple application of Stirling's formula (see, e.g., [31]) shows that

$$\frac{1}{n}\sum_{t=1}^{n} -\log p_{t-1}(x_t) \leq \frac{1}{n}\sum_{t=1}^{n} -\log p_n(x_t) + \frac{1}{2}\frac{\log n}{n} + O(\frac{1}{n}), \qquad (1)$$

where logarithms throughout the sequel are taken to the base 2 unless specified otherwise. The left-hand side of (1) corresponds to the normalized length of a codeword associated with a sequential optimal Shannon encoder which is based on current empirical letter probabilities from data observed so far. This length can be attained using, e.g., arithmetic coding techniques [33]. The first term on the right hand side of (1) is the empirical entropy associated with x^n, which corresponds to the minimum normalized codeword length associated with a fixed codebook that one could have achieved for a particular x^n if he knew in advance $\{p_n(x)\}_{x=0,1}$. The $O(n^{-1}\log n)$ term in (1) is the loss in performance due to sequentiality (see also [30], [31], [41], [42], [45]). Observe that (1) can be formalized as sequential minimization of $n^{-1}\sum_{t=1}^{n} l(b, x_t)$, where the per-letter loss function $l(b, x)$ is given in this case by

$$l(b, x) = \begin{cases} -\log b, & x = 0, \\ -\log(1-b), & x = 1, \end{cases} \qquad (2)$$

where the best choice of $b \in (0, 1]$ in the sense of minimizing $(t-1)^{-1}\sum_{\tau=1}^{t} -\log l(b, x_\tau)$ is $b = n_t(0)/t$ which is nearly $p_{t-1}(0)$.

Another interesting application of (1) and (2) is sequential gambling (see, e.g., [5], [11], [29]) where at each round t the player doubles the fraction of the current capital S_t wagered on the next outcome, i.e., $S_{t+1} = 2bS_t$ if $x_{t+1} = 0$ and $S_{t+1} = 2(1-b)S_t$ if $x_{t+1} = 1$. It is easy to see that the exponential growth rate $n^{-1}\log S_n$ of the capital is the time-average of $1 - l(b, x_t)$, where $l(\cdot, \cdot)$ is as in (2) and hence eq. (1) is meaningful for gambling as well.

Portfolio selection for optimal investment [1]–[3],[6] can be viewed as an extension of the previously described gambling problem, where the current capital S_t is distributed over m investment opportunities according to some portfolio $b \in \mathbb{R}^m$, a column vector of nonnegative weights summing to unity. The stock market on day t is characterized by a column vector $x_t \in \mathbb{R}^m$ with nonnegative components, x_t^i representing the return per monetary unit allocated to stock i on day t. The yield per unit invested is the weighted average of return ratios, i.e., the inner product $b^{\#}x_t$, where $\#$ denotes vector transposition. Thus, $S_n = S_0\prod_{t=1}^{n}(b^{\#}x_t)$ is the compounded capital after n investment days. Equivalently, the exponential growth rate $n^{-1}\log S_n$ of the capital is the time-average of $l(b, x_t) = \log(b^{\#}x_t)$. In [6] a sequential portfolio selection scheme has been proposed for arbitrary bounded sequences of market vectors, which is again as good as the optimal fixed investment policy up to a term of $O(n^{-1}\log n)$. The proof in [6], however, relies heavily on special properties of the per-letter loss function $l(b, x) = \log(b^{\#}x)$, considered in this specific case.

In [12] a result in the same spirit has been established for the problem of universal prediction of binary sequences, where predictors have been sought that uniformly minimize the fraction of prediction errors. The strategy b at time t is a choice of an estimate $\hat{x}_{t+1}$ of the next outcome x_{t+1} and $l(\hat{x}_{t+1}, x_{t+1})$ is the indicator function for $\hat{x}_{t+1} \neq x_{t+1}$ (i.e., the Hamming distance). Again, the techniques for deriving the results in [12] are specific to this particular loss function. It should be pointed out that the results in [12] are different from these of Ryabko [45], who focused on a probabilistic setting and considered the prediction problem as that of reliable estimation of the conditional probabilities of future outcomes, given the past, rather than that of estimating the outcomes themselves.

When $M = 1$, the previous examples can all be viewed as special cases of a more general setting, referred to as the *sequential compound decision problem*, which was first presented by Robbins [41] and has been thoroughly investigated later by many researchers (see, e.g., [4], [14]–[16], [21], [22], [39], [46], [47], [51], [52]). The setup of the sequential compound decision problem is somewhat more general in the sense that the observer may access only noisy versions of the measurements $\{x_t\}$ that appear in the loss function. Hannan [21] developed in the game theoretic level upper bounds on the decay rate of the difference between the average loss (or risk) associated with the best sequential strategy and that of the best fixed strategy. He has shown a convergence rate of $O(n^{-1/2})$ in the finite-alphabet, finite-strategy space case, and a rate of $O(n^{-1}\sum_{t=1}^{n} t^{-\alpha})$ in the continuous case, provided that the loss minimizing strategy b^*, as a functional of the underlying empirical measure (i.e., the Bayes response), satisfies a Lipschitz condition of order $\alpha > 0$. For $\alpha = 1$, this means a rate of $O(n^{-1}\log n)$. Van Ryzin [51] has shown that even in the former case the convergence rate can be more tightly upper bounded by $O(n^{-1}\log n)$ under some regularity conditions on the channel through which the observer receives the noisy measurements. Gilliland [15] further investigated convergence rates for the special case of squared-error-loss estimation, i.e., $l(b, x) = (b-x)^2$, under various assumption sets.

Swain [50], Johns [28], Gilliland and Hannan [17], Cover and Shenhar [7], Nogami [39], and Vardeman [52] have extended the scope of the sequential compound decision problem and developed sequential decision procedures whose performance is almost as good as that of the best kth-order *Markovian* (rather than fixed) strategy, i.e., the best strategy that depends at time t on the k preceding outcomes $x_{t-k}, x_{t-k+1}, \cdots, x_{t-1}$, and hence results in an average loss no greater than that of the best fixed strategy. While the Markovian strategy is intuitively appealing and plausible when the sequence is known to have a "Markov structure " [7], [32], it has not yet been justified rigorously for a general arbitrary sequence considered here.

As mentioned earlier, we study here the more general class of finite-state strategies. In fact, one important result

in this work links the performance attainable by the best kth-order Markovian strategy to that of the best M-state strategy Specifically, we assume that $\{x_t\}$ are directly accessible without noise and extend Theorem 2 of [12], showing (Section III) that for a given M-state machine, one can pick a sufficiently large k that is independent of the sequence, such that the best kth-order Markov machine performs to within ϵ as well as the M-state machine. This means that in the limit of indefinite increase in the number of states, a Markovian machine is as good as the best deterministic FSM. As a result, one can gradually increase the Markov order at a logarithmic rate independently of the particular sequence, and guarantee convergence to $u_\infty(x_1, x_2, \cdots)$. In Section IV, this result is further extended, and it is shown that Markovian machines with sufficiently long memory compete successfully with every *randomized* FSM in the sense of minimizing the expected value of $n^{-1}\sum_{t=1}^{n} l(b_t, x_t)$, where the expectation is with respect to the randomization.

This property of Markovian strategies is utilized in order to relate the least asymptotic loss achievable by FSM's over individual sequences to that of the probabilistic case where any limitations on the allowed nonanticipating strategies are relaxed. Specifically, following Algoet [2], where the Shannon–McMillan–Brieman theorem has been extended to a general sequential decision problem under a stationary ergodic regime, we show that these two quantities agree with probability one over an infinite sequence. This extends a result in the same spirit in the case of lossless coding [53].

Markovian schemes are useful also in continuous alphabet applications (Section V). One widespread example is prediction under the mean-squared error (mse) criterion, i.e., $l(b_t, x_t) = (x_t - b_t)^2$, where the strategy (namely, the predictor) b_t is given by a function $f(x_{t-k}, \cdots, x_{t-1})$ of the k most recent outcomes, e.g., a linear predictor [36], where $f(x_{t-k}, \cdots, x_{t-1}) = \sum_{i=1}^{k} c_i x_{t-i}$. The sequential version of this linear predictor leads to the recursive least squares (RLS) algorithm, which is here shown to be universal in the previous sense. Another example is vector quantization (VQ) (see e.g., [19], [27], [35], [37]), where $x \in \mathbb{R}^m$ and $l(b, x) = d(x, Q_b(x))$, $d(\cdot, \cdot)$ being a distortion measure and $Q_b(\cdot)$ a quantization function with quantization cells and centroids parametrized by b. Again, by allowing b to depend on the k preceding samples (or their quantized versions), we can implement a family of vector quantizers with memory [19], e.g., feedback quantizers, predictive quantizers, finite-state quantizers, etc.

II. Fixed Strategy

We start from some preliminaries and provide a sufficient condition for a sequential procedure to perform within $O(n^{-1}logn)$ as well as the best $fixed$ strategy. This condition is not entirely new and has appeared in several variations (see, e.g., [14], [16], [21], [46], [51]) for the finite alphabet case and the continuous alphabet case separately. Here, for the sake of convenience, we formulate it in a unified way that is suitable for both cases. This will serve as a background for the derivations that follow in Sections III, IV, and V.

Consider an arbitrary (deterministic) sequence of observations $\boldsymbol{x}^n = (x_1, x_2, \cdots, x_i, \cdots, x_n)$, x_i taking values in some alphabet $\boldsymbol{X}$. An observer wishes to select a member b from a set $\boldsymbol{B}$ of permissible strategies so as to minimize the time-average of a certain loss function $l(b, x_t)$, i.e., attain

$$u(\boldsymbol{x}^n) = \min_{b \in \boldsymbol{B}} \frac{1}{n} \sum_{t=1}^{n} l(b, x_t). \tag{3}$$

Unfortunately, since the best strategy b_n^* that attains $u(\boldsymbol{x}^n)$ depends, in general, on the entire sequence $\boldsymbol{x}^n$, it cannot be found in a sequential manner. A natural alternative is to adapt the strategy b, at each time instant t (before seeing x_t), to the data observed so far, i.e., to use at time t a strategy b_{t-1}^* that minimizes the quantity

$$\frac{1}{t-1} \sum_{\tau=1}^{t-1} l(b, x_\tau), \quad t > 1 \tag{4}$$

and an arbitrary strategy at time $t = 1$. The basic fact that is shown in this section is that, under certain regularity conditions, the sequence of strategies $\{b_{t-1}^*\}_{t=1}^n$ is asymptotically as good as b_n^*. More precisely, let

$$\hat{u}(\boldsymbol{x}^n) \triangleq \frac{1}{n} \sum_{t=1}^{n} l(b_{t-1}^*, x_t). \tag{5}$$

Then, the difference $\hat{u}(\boldsymbol{x}^n) - u(\boldsymbol{x}^n)$ vanishes as fast as $n^{-1}\log n$, uniformly for every sequence $\boldsymbol{x}^n$. This claim holds true whether or not $u(\boldsymbol{x}^n)$ and $\hat{u}(\boldsymbol{x}^n)$ converge as $n \to \infty$. Hence, no assumptions concerning asymptotic mean stationarity [20] and ergodicity of an underlying probability measure are required.

To formulate regularity conditions on $l(\cdot, \cdot)$ it will be convenient to consider the empirical probability measure, $P_n = n^{-1} \sum_{t=1}^{n} \delta_{x_t}$, (where δ_x, $x \in \boldsymbol{X}$, is the unit point mass at x) and to regard time-averages as expectations with respect to P_n, e.g., $n^{-1}\sum_{t=1}^{n} l(b, x_t) = E_{P_n} l(b, X)$, where X denotes a random variable (governed by P_n).

Let P be a probability measure defined on a measure space $(\boldsymbol{X}, \boldsymbol{F})$, $\boldsymbol{F}$ being a sigma-field generated from subsets of $\boldsymbol{X}$. Assume that P belongs to a set $\boldsymbol{\mathcal{P}}$ of probability measures defined as $\boldsymbol{\mathcal{P}} \triangleq \{P : \exists b \in \boldsymbol{B}, \ E_P l(b, X) < \infty\}$ and let

$$U(X) \triangleq \inf_{b \in \boldsymbol{B}} E_P l(b, X). \tag{6}$$

In the sequel, when we would like to stress the dependency of $U(X)$ upon P, we shall denote it by $U(P)$ with a slight abuse of notation. This quantity, called the *Bayes envelope* (see, e.g., [46]), can be thought of as a generalized notion of the entropy since in the special case (2) it agrees with the binary Shannon entropy. (This may serve as an intuitive explanation of the fact that sequential decision procedures have been proposed to assess the degree of "randomness" of a sequence [48]). It is very easy to see that $U(P)$ is always a concave functional [16]. Of course when P is the empirical measure P_n, then $U(P_n) = u(\boldsymbol{x}^n)$.

The following assumption on $l(\cdot, \cdot)$ will be made.

Assumption A: If $P \in \mathcal{P}$, then the infimum (6) is also a minimum, and there exists a minimizer $b^*(P)$, i.e., $E_P l(b^*(P), X) = U(P)$, such that for every $P \in \mathcal{P}$ and $x \in \mathcal{X}$,

$$\sup_{\alpha \in (0,1)} \frac{1}{\alpha} \left| l(b^*(P), x) - l(b^*((1-\alpha)P + \alpha\delta_x), x) \right| \le K, \quad (7)$$

for some $K < \infty$.

Assumption A is a version of the Lipschitz condition on $l(b^*(\cdot), x)$ as a functional of P, where distances between probability measures are restricted to convex combinations with unit point mass measures. The supremum over α can be replaced by a limit as $\alpha \to 0^+$ (Gateaux derivative), resulting in a slightly weaker version of the assumption A, at the expense of restricting $l(\cdot,\cdot)$ to be bounded. In the finite-alphabet case, Gilliland and Helmers [16, Theorem 2] provide necessary and sufficient conditions for $l(b^*(P), x)$ being continuous w.r.t. P in the "direction" $(1-\alpha)P + \alpha\delta_x$. It is easy to imply from the proof of [16, Theorem 2] (see also Samuel [46]) that in the finite-alphabet case, Assumption A is equivalent to the condition that $U(P)$ has derivatives w.r.t the letter probabilities, and they all satisfy a first-order Lipschitz condition.

The following theorem provides bounds on the average loss $\hat{u}(\boldsymbol{x}^n)$ associated with the sequential strategy selection procedure $b^*_{t-1} \equiv b^*(P_{t-1})$, in terms of the loss associated with the best fixed strategy $u(\boldsymbol{x}^n)$.

Theorem 1: Under Assumption A, for every $\boldsymbol{x}^n$,

$$u(\boldsymbol{x}^n) \le \hat{u}(\boldsymbol{x}^n) \le u(\boldsymbol{x}^n) + \frac{K}{n}[\ln(n) + 1], \quad (8)$$

where K is as in (7).

The theorem tells us that applying the best strategy $b^*(P_{t-1})$ for the data observed so far is not as good as the best fixed strategy, but it results in an average loss which is only $O(n^{-1}\log n)$ far away from optimality.

Proof of Theorem 1: The proof is similar to these in [21] and [46] and brought here for the sake of completeness. The theorem follows from two simple inequalities due to Hannan [21] (see also Gilliland [14]) that follow directly from the definition of $b^*(P)$ as a minimizer of the loss. First, note that by the definition of $b^*(P)$,

$$\begin{aligned} n \cdot u(\boldsymbol{x}^n) &= \sum_{t=1}^{n} l(b^*(P_n), x_t) \\ &\le \sum_{t=1}^{n} l(b^*(P_{n-1}), x_t) \\ &= \sum_{t=1}^{n-1} l(b^*(P_{n-1}), x_t) + l(b^*(P_n - 1), x_t) \\ &\le \sum_{t=1}^{n-1} l(b^*(P_{n-2}), x_t) + l(b^*(P_{n-1}), x_t) \\ &= \sum_{t=1}^{n-2} l(b^*(P_{n-2}),\ x_t) + \sum_{t=n-1}^{n} l(b^*(P_{t-1}), x_t), \end{aligned} \quad (9)$$

and so forth, ending up with

$$n \cdot u(\boldsymbol{x}^n) \le \sum_{t-1}^{n} l(b^*(P_{t-1}), x_t) = n \cdot \hat{u}(\boldsymbol{x}^n), \quad (10)$$

which completes the proof of the left inequality of Theorem 1. As for the right inequality, similarly to (9) and (10), we have

$$\begin{aligned} n \cdot u(\boldsymbol{x}^n) &= \sum_{t=1}^{n-1} l(b^*(P_n), x_t) + l(b^*(P_n), x_n) \\ &\ge \sum_{t=1}^{n-1} l(b^*(P_{n-1}), x_t) + l(b^*(P_n), x_n) \\ &= \sum_{t=1}^{n-2} l(b^*(P_{n-1}), x_t) + \sum_{t=n-1}^{n} l(b^*(P_t), x_t) \\ &\ge \sum_{t=1}^{n-2} l(b^*(P_{n-2}), x_t) + \sum_{t=n-1}^{n} l(b^*(P_t), x_t), \end{aligned} \quad (11)$$

and so forth, ending up with

$$u(\boldsymbol{x}^n) = \frac{1}{n}\sum_{t=1}^{n} l(b^*(P_n), x_t) \ge \frac{1}{n}\sum_{t=1}^{n} l(b^*(P_t), x_t). \quad (12)$$

By Assumption A and the fact that $P_t = (1 - t^{-1})P_{t-1} + t^{-1}\delta_{x_t}$, we have

$$|l(b^*(P_{t-1}), x_t) - l(b^*(P_t), x_t)| \le \frac{K}{t}. \quad (13)$$

Hence, the right-hand side of (12) is further underbounded as follows.

$$\begin{aligned} &\frac{1}{n}\sum_{t=1}^{n} l(b^*(P_t), x_t) \\ &\quad \ge \frac{1}{n}\sum_{t=1}^{n} l(b^*(P_{t-1}), x_t) \\ &\quad\quad - \frac{1}{n}\sum_{t=1}^{n} |l(b^*(P_{t-1}), x_t) - l(b^*(P_t), x_t)| \\ &\quad = \hat{u}(\boldsymbol{x}^n) - \frac{K}{n}\sum_{t=1}^{n}\frac{1}{t}. \end{aligned} \quad (14)$$

Finally, since

$$\begin{aligned} \sum_{t=1}^{n}\frac{1}{t} &= 1 + \sum_{t=2}^{n}\frac{1}{t} \le 1 + \sum_{t=2}^{n}\int_{t-1}^{t}\frac{du}{u} \\ &= 1 + \int_1^n \frac{du}{u} = 1 + \ln(n), \end{aligned}$$

the proof of Theorem 1 is complete. □

Assumption A (and similarly, assumptions made in [14], [16], [21]), though satisfied for a reasonably wide class of loss functions $l(\cdot,\cdot)$, is somewhat more demanding than necessary in the sense that it does not cover all these examples and yet the theorem holds for all of them. This assumption, however, makes the proof of the theorem intuitively appealing. It is based upon the following simple idea: If x_t was available

before the tth strategy had to be selected, then an average loss smaller than that of the best fixed strategy could have been achieved. Nonetheless, although x_t is yet unavailable when the tth decision is to be made, one can still approximate faithfully $b^*(P_t)$ by $b^*(P_{t-1})$ under appropriate continuity conditions. Furthermore, the approximation error, and hence also the loss in performance, behaves normally like $1/t$, whose time-average over $t = 1, 2, \cdots, n$, is $O(n^{-1}\log n)$. We now examine the examples described in Section I in the light of Theorem 1 and demonstrate that in some of these examples Assumption A is violated.

The logarithmic loss function (2), which arises in data compression and gambling applications, clearly does not satisfy condition A. Intuitively, however, the theorem still holds here because when $p_{t-1}(0)$ is small (and hence, $\mid \log p_{t-1}(0) \mid$ is large), then by definition, the relative frequency of zeros is small as well and hence their overall effect to the average $n^{-1}\sum_{t=1}^{n} \log p_{t-1}(x_t)$ is negligibly small.

The universal portfolio selection problem is associated with the function $l(b, x) = -\log(b^{\#}x)$ (see Section I), which again suffers from a singularity problem about the origin. This can be avoided if it is assumed that the components of x_t lie in some interval $[a, c]$, $0 < a \leq c < \infty$. Indeed, this assumption is made in [6] and hence the condition A holds. It should be pointed out that the sequential strategy selection procedure proposed in [6] is slightly different from $b^*(P_{t-1})$, but it is asymptotically equivalent.

The prediction problem [12] involves a zero-one loss function which is discontinuous, and hence obviously cannot satisfy the condition A. As an alternative to the definition of this loss function, one can define $l(b, x)$ of [12] by

$$l(b,x) = \begin{cases} \phi(b), & x = 0, \\ 1 - \phi,(b) & x = 1, \end{cases} \tag{15}$$

where $\phi(b)$ is a prediction error indicator, namely, a unit step function at $b = 1/2$ and $\boldsymbol{B} = [0, 1]$. Indeed, it is easy to see that if $b^*(P_{t-1}) = n_t(0)/t$ converges to the discontinuity point, then the theorem does not hold. Specifically, in [21], [46], [51] it has been shown that the this discontinuity causes the convergence to slow down to $O(n^{-1/2})$. For this reason, $\phi(b)$ is approximated in [12] by a continuous function $\phi_\epsilon(b)$ where the step is smoothed by a finite-slope line in the interval $b \in [1/2-\epsilon, 1/2+\epsilon]$, and the values of $\phi_\epsilon(b)$ between zero and one correspond to randomization. For $\phi_\epsilon(\cdot)$, however, there is no longer a continuous minimizer $b^*(P)$, and again the condition A is violated. Nevertheless, it is proved in [12] that $b^*(P_{t-1})$ is asymptotically ϵ-optimal using techniques different from those of the proof of Theorem 1.

For the prototype problems of sequential prediction, where $l(b, x) = |x - b|^\alpha$, $\alpha > 0$, and sequential VQ design (see Section I) with $l(b, x) = |x - Q_b(x)|^\alpha$, Condition A is met if the measurements $\{x_t\}$ are bounded uniformly, which is a fairly mild requirement in practice (see also Gilliland [14], [15]).

So far we have considered sequential schemes that compete successfully with any *fixed* strategy. In the remaining part of the paper, we shall extend Theorem 1 and further investigate properties of more interesting competing schemes that consist of a certain amount of memory of past data.

III. Deterministic Finite-State Machines

A commonly-used model for sequential machines with a limited amount of storage is a *finite-state machine* (FSM). A sequential decision strategy based on an M-state FSM is a triple $E = (\boldsymbol{S}, f, g)$, where $\boldsymbol{S}$ is a finite set of states with M elements, $f : \boldsymbol{S} \to \boldsymbol{B}$ is the *output function*, and $g : \boldsymbol{X} \times \boldsymbol{S} \to \boldsymbol{S}$ is the *next-state function*. When an input sequence $x_1, x_2, \cdots$ is fed into E, starting with an initial state $s_0 \in \boldsymbol{S}$, this FSM produces a sequence of output strategies $b_1, b_2, \cdots$ while going through a sequence of states, $s_1, s_2, \cdots$ according to the recursive rules

$$s_t = g(x_{t-1}, s_{t-1}) \tag{16a}$$

$$b_t = f(s_t), \tag{16b}$$

where s_t is the state of E at time t. This model has been adopted in all aforementioned applications where the input alphabet is finite, e.g., noiseless data compression [53], gambling [11], and prediction [12]. Although FSM's may be useful in some continuous input alphabet applications as well, e.g., finite-state vector quantization (FSVQ) [19], we shall assume in this section that the input alphabet $\boldsymbol{X}$ is finite.

The fixed strategy case, considered in Section II, is a special case of (16) where $M = 1$. For a given next-state function $g(\cdot, \cdot)$, Theorem 1 extends in a straightforward manner to the M-state model. Observe that for a given state $s \in \boldsymbol{S}$, the strategy $b = f(s)$ is again fixed. Thus, for each subsequence $\boldsymbol{x}^n(s) = \{x_t, t : s_t = s\}$, $s \in \boldsymbol{S}$, one can apply Theorem 1 and thereby asymptotically attain the least possible contribution of state s to the total average loss. Therefore, the minimum achievable average loss associated with a next-state function g with M states is given by

$$\begin{aligned} u(\boldsymbol{x}^n; g) &= \frac{1}{n}\sum_{s\in\boldsymbol{S}} \min_{b\in\boldsymbol{B}} \sum_{t:s_t=s} l(b, x_t) \\ &= \sum_{s\in\boldsymbol{S}} \frac{n_g(s)}{n} u(\boldsymbol{x}^n(s)), \end{aligned} \tag{17}$$

where $n_g(s)$ is the relative frequency of state s in the state sequence $\boldsymbol{s}^n = s_1, s_2, \cdots, s_n$ generated by g. Consider the following sequential scheme: At each time instant t, first update the state by $s_t = g(x_{t-1}, s_{t-1})$ and then apply the strategy that best fits the observations seen so far that are associated with s_t, namely, $\{x_\tau, \tau \leq t-1, \tau : s_\tau = s_t\}$. For an upper bound on the average loss $\hat{u}(\boldsymbol{x}^n; g)$ associated with this sequential procedure, one can apply Theorem 1 for each state separately (see also [12]). Since each state s contributes an unnormalized term of $K[\ln n_g(s) + 1]$ to the upper bound, it follows that

$$\begin{aligned} \hat{u}(\boldsymbol{x}^n; g) &\leq u(\boldsymbol{x}^n; g) + \frac{K}{n}\sum_{s\in\boldsymbol{S}} [\ln n_g(s) + 1] \\ &\leq u(\boldsymbol{x}^n; g) + \frac{KM}{n}\left[\ln(\frac{n}{M}) + 1\right], \end{aligned} \tag{18}$$

where for the last step we have used Jensen's inequality and the fact that $\sum_{s\in\boldsymbol{S}} n_g(s) = n$. Thus, the optimal performance

for a given $g(\cdot,\cdot)$ is attained to within an $O(n^{-1}M\log n/M)$ term. A more difficult problem, though, is to attain sequentially the performance of the *best* M-state machine, i.e., $u_M(\boldsymbol{x}^n) \stackrel{\Delta}{=} \min_{g\in G_M} u(\boldsymbol{x}^n;g)$, where G_M is the set of all M^{AM} possible next-state functions associated with M-state machines. The minimizing $g(\cdot,\cdot)$ obviously depends on the entire sequence $\boldsymbol{x}^n$ and hence cannot be found in a sequential manner but only by an exhaustive search over G_M once the sequence $\boldsymbol{x}^n$ is fully available. Nonetheless, in this section, it is shown, under some regularity conditions, that in the limit as $M \to \infty$ (with $n \gg M$), it is sufficient to focus on Markovian machines for the purpose of asymptotically minimizing $u(\boldsymbol{x}^n;g)$, among all M-state machines.

Markovian Machines

An important special case of an FSM (with $M = A^k$ states) is the kth-order Markovian machine, for which the state s_t at time t is defined by the k preceding input letters, i.e., $s_t = (x_{t-k},\cdots,x_{t-1})$ [7], [32], [39], [50]. When g is Markovian of order k, let us denote $u(\boldsymbol{x}^n;g)$ by $u(\boldsymbol{x}^n;\mathcal{MMM}_k)$.

Again, we shall formulate regularity conditions in terms of empirical probability measures. For two probability mass functions P and Q on $\boldsymbol{X}$, let

$$\Delta(P\|Q) \stackrel{\Delta}{=} E_P l(b^*(Q),X) - E_P l(b^*(P),X). \tag{19}$$

This quantity expresses the loss of optimality in applying a strategy b that best matches Q, when the true underlying probability measure is P. Clearly, $\Delta(P\|Q) \geq 0$ with equality if $P = Q$, and (19) generalizes the notion of divergence in the sense that if $l(\cdot,\cdot)$ is as in (2), then $\Delta(P\|Q) \equiv D(P\|Q)$, the Kullback–Leibler informational divergence. We make the following assumption on $\Delta(P\|Q)$ which, in turn, induces an assumption on $l(\cdot,\cdot)$.

Assumption B: There exist positive constants C and δ such that for every two probability mass functions $P = \{p(x)\}_{x\in\boldsymbol{X}}$ and $Q = \{q(x)\}_{x\in\boldsymbol{X}}$ on the finite alphabet $\boldsymbol{X}$,

$$\Delta(P\|Q) \leq C\|P-Q\|^{\delta}, \tag{20}$$

where $\|P-Q\| \stackrel{\Delta}{=} \sum_{x\in\boldsymbol{X}} |p(x)-q(x)|$, namely, the variational distance between P and Q.

Observe that

$$\Delta(P\|Q) \leq |\, E_P l(b^*(Q),X) - E_Q l(b^*(Q),X)\,| + |\, E_Q l(b^*(Q),X) - E_P l(b^*(P),X)\,|.$$

If $l(\cdot,\cdot)$ is bounded by a constant L, then the first term is smaller than $L\cdot\|P-Q\|$ and the second term is the difference between the Bayes envelopes associated with P and with Q, respectively. Since the Bayes envelope is a concave functional of the underlying measure, it is also continuous, and hence the requirement that the second term would be bounded in terms of $\|P-Q\|^{\delta}$, where $\delta > 0$ is allowed to be arbitrarily small, is not highly restrictive.

Theorem 2: If assumption B is met, then for every $\boldsymbol{x}^n \in \boldsymbol{X}^n$ and every two positive integers k and M,

$$u(\boldsymbol{x}^n;\boldsymbol{\mathcal{M}}_k) \leq \min_{g\in G_M} u(\boldsymbol{x}^n;g) + \left(\frac{2C\ln M}{k+1}\right)^{\delta/2}. \tag{21}$$

The proof appears in the Appendix.

This is a generalized version of Theorem 2 in [12], which tells us that a kth-order Markovian machine of order k is within ϵ as good as the best M-state machine, uniformly for every $\boldsymbol{x}^n$, provided that $k > 2C\epsilon^{-2/\delta}\ln M$. It should be pointed out that this does *not* mean that in general M-state machines can be simulated by Markovian machines with sufficiently long memory (see [13, p. 156] for a counterexample).

Given an infinite sequence $\boldsymbol{x} = (x_1,x_2,\cdots)$, we define

$$u_\infty(\boldsymbol{x}) = \lim_{M\to\infty} \limsup_{n\to\infty} \min_{g\in G_M} u(\boldsymbol{x}^n;g). \tag{22}$$

This quantity, which depends solely on $\boldsymbol{x}$, is a generalization of both the *finite-state compressibility* of Ziv and Lempel [53], where $l(\cdot,\cdot)$ is as in (2), and the *finite-state predictability* of [12], where $l(\cdot,\cdot)$ is as in (15). The number $u_\infty(\boldsymbol{x})$ is by definition an asymptotic lower bound on the attainable performance of any FSM with arbitrarily many states, when fed with $\boldsymbol{x}$.

In light of the previous results, given an infinite input sequence $\boldsymbol{x}$, and provided that the assumptions A and B are met, it is possible to attain $u_\infty(\boldsymbol{x})$ using a machine with infinitely many states by chopping the data into exponentially growing segments, where at the kth segment, $k = 1,2,\cdots$, we use the next-state function of a kth-order Markovian machine and the appropriate sequential strategy as explained earlier. Following (18), after sufficiently long time, the average loss is essentially as low as that of the best Markovian machine with an arbitrarily long memory, which in turn (Theorem 2) is nearly as good as the best FSM with arbitrarily many states. An alternative policy of increasing k is the one induced by the incremental parsing procedure [53] applied to $\boldsymbol{x}$. The reader is referred to [12, Sections IV, V] for more details concerning these two methods of increasing the order k.

It is interesting to relate the quantity $u_\infty(\boldsymbol{x})$ to the best achievable performance in the probabilistic setup, as was done in the special case of FS compressibility [53]. Algoet [2] has studied the sequential decision problem for a stationary ergodic input $X_1, X_2,\cdots$. One of the results in [2] is an extension of the Shannon–McMillan–Breiman theorem to the case of a general loss function $l(\cdot,\cdot)$. Specifically, let us extend the definition (6) to the form

$$U(X|X^k) = \inf_{b\in\boldsymbol{B}} E\{l(b,X_0)|X_{-1},\cdots,X_{-k}\} \tag{23}$$

where we regard the input as a two sided process $\{\cdots,X_{-1},X_0,X_1,\cdots\}$ using the shift invariance property. Now, let

$$U(X|X^\infty) = \lim_{k\to\infty} U(X|X^k). \tag{24}$$

Under certain integrability conditions [2] $U(X|X^\infty)$ agrees with $\inf_b E\{l(b,X_0)|X_{-1},X_{-2},\cdots\}$. Now, in Theorem 6 of

[2] it has been shown that if $\Lambda(x) \triangleq \sup_{b\in B} |l(b,x)|$ satisfies

$$E\left[\Lambda(X)\cdot \log_+^{1+\epsilon} \Lambda(X)\right] < \infty \tag{25}$$

where $\log_+ a \triangleq \max\{0, \log a\}$ and ϵ is an arbitrarily small strictly positive number, then

$$\frac{1}{n}\sum_{t=1}^{n} l(b_t^*(X_1^{t-1}),\ X_t) = U(X|X^-), \tag{26}$$

almost surely and in L_1, where $b_t^*(\cdot)$ achieves (or approaches) $\inf_b E\{l(b, X_t)|X_1, \cdots, X_{t-1}\}$. Using this result and Theorem 2 we can now relate $u_\infty(\boldsymbol{x})$ to $U(X|X^\infty)$.

Theorem 3: Let $X_1, X_2, \cdots$ be a stationary ergodic process, and assume that the conditions (A), (B), and (25) are satisfied. Then, $u_\infty(X_1,\ X_2, \cdots) = U(X|X^\infty)$ almost surely.

The proof appears in the Appendix.

IV. Randomized Finite-State Machines

Let us return to the deterministic setting where $x_1,\ x_2, \cdots$ is a given individual sequence. So far we have considered nonrandom FSM's, where the next-state function g and the output function f of (16) are deterministic. A natural possible extension (especially for gambling, prediction and investment applications), is obtained by letting f and g be *stochastic* functions, namely, replacing f and g by conditional probability distributions $p(\cdot|s_t)$ and $q(\cdot \mid x_{t-1}, s_{t-1})$ for randomly selecting b_t and s_t. The performance will then be judged on a statistical basis, e.g., the expected value of $n^{-1}\sum_{t=1}^n l(b_t, x_t)$, or more generally, the expected value of $\psi(n^{-1}\sum_{t=1}^n l(b_t, x_t))$ for some monotonically increasing function ψ. (Note that the expectation is defined with respect to the ensemble of randomly chosen states and strategies while the sequence $\boldsymbol{x}$ is still considered fixed.)

The problem of designing a randomized FSM is that of selecting the best conditional probability distributions $\{p(b|s)\}_{b,s}$ and $\{q(s \mid x, s')\}_{x,s,s'}$ so as to minimize the expected value of ψ for a particular sequence $\boldsymbol{x}$, which is observed sequentially. Since the class of randomized FSM's contains deterministic FSM's as a subclass, it is not surprising that a good randomized FSM can do better, in general, than the best deterministic FSM with the same number of states. Indeed, randomized FSM's have been thoroughly investigated in certain applications of statistical inference (see, e.g., [24]–[26], [34], [49], and references therein), and were shown to outperform their deterministic counterparts. In [34], for instance, it has been shown that a randomized M-state machine for estimating the probability of a Bernoulli process is equivalent to a deterministic FSM with as many as $O(M \log M)$ states. However, in these examples the difference in performance between deterministic and randomized FSM's disappears once the limit $M \to \infty$ ($n \gg M$) is taken for both types of machines.

It is interesting to investigate a similar question for the sequential decision problem considered here: In the limit of arbitrarily many states, are deterministic FSM's as good as randomized FSM's or, rather, can performance still be gained by using randomized FSM's? The answer to this question turns out to depend on the particular risk function ψ under consideration.

Before we address this question, observe that with no loss in generality we can assume that the output function f is deterministic, i.e., $p(b|s)$ puts its entire mass on $b = f(s)$. To see this, recall that for a given state sequence $\boldsymbol{s}^n$, the best deterministic strategy is derived from the joint empirical probability measure associated with $(\boldsymbol{x}^n, \boldsymbol{s}^n)$ as explained in Section III, and any other strategy will yield a higher loss. It follows that any randomization, which puts a positive probability on values of b other than the optimal value, will result in an average loss larger than that of the best deterministic output function. On the other hand, the randomization in q might be helpful as it allows in general all M^n possible state sequences rather than only one state sequence with a constrained structure as in the deterministic case. In view of these facts, we henceforth assume that f is deterministic, and thus the only randomization is due to q.

Consider first the criterion of minimizing

$$E_q\{\frac{1}{n}\sum_{t=1}^{n} l(f(S_t), x_t)\}, \tag{27}$$

where $E_q\{\cdot\}$ denotes the expectation under q, and S_t denotes the random state at time t. We next show that deterministic FSM's with sufficiently many states are capable of doing nearly as well as any randomized FSM in the above sense. Specifically, we next extend Theorem 2 and prove that the best deterministic kth-order Markovian machine is to within ϵ as good as the best randomized M-state machine for sufficiently large k, but as in (21) it takes as many as $A^k = M^{2C\epsilon^{-2/\delta}\ln A}$ states (where A is the alphabet size) for a Markov machine to guarantee successful competition with the best M-state randomized FSM. The following theorem summarizes this fact.

Theorem 4: If assumption B is met, then for every $\boldsymbol{x}^n \in \boldsymbol{X}^n$ and every two integers k and M,

$$u(\boldsymbol{x}^n; \boldsymbol{\mathcal{M}}_k) \le \min_{f\in F_M}\ \inf_{q\in Q_M}\ E_q\left\{\frac{1}{n}\sum_{t=1}^{n} l(f(S_t), x_t)\right\} + \left(\frac{2C\ln M}{k+1}\right)^{\delta/2}, \tag{28}$$

where F_M is the class of all output functions $f : \boldsymbol{S} \to \boldsymbol{B}$ associated with $|\boldsymbol{S}| = M$ states and Q_M is the class of all conditional probability distributions $\{q(\cdot \mid x, s)\}_{x\in\boldsymbol{X}, s\in\boldsymbol{S}}$ associated with randomized M-state machines.

The proof appears in the Appendix.

Theorem 4, therefore, enables one to extend the definition (22) to randomized FSM's and still attain the resulting lower bound using sequential Markov schemes that let k grow slowly with t, as described earlier.

In certain applications of the sequential decision problem, however, the criterion (27) is not really the relevant performance measure. In gambling and portfolio selection applications, for instance, a natural goal might be to maximize

the exponential growth rate of the expected fortune at time n, corresponding to

$$\max_q \frac{1}{n} \log E_q \exp_2\{-\sum_{t=1}^{n} l(f(S_t), x_t)\} \tag{29}$$

with the appropriate choice of the function $l(\cdot,\cdot)$ (see Section I). In data compression applications this criterion with $l(\cdot,\cdot)$ as in (2) corresponds to the negative normalized length function of a universal code for the class of finite-state sources, where the probability of an input string $\boldsymbol{x}^n$ is given by

$$P(\boldsymbol{x}^n) = \sum_{\boldsymbol{s}^n} \prod_{t=1}^{n} q(s_t|x_{t-1}, s_{t-1}) r(x_t|s_t), \tag{30}$$

where $r(0|s) = f(s)$, $r(1|s) = 1 - f(s)$, and $f(s) \in (0,1)$.

More generally, consider the performance criterion

$$\max_q \frac{1}{n\lambda} \log E_q \exp_2\{-\lambda \sum_{t=1}^{n} l(f(S_t), x_t)\}, \quad \lambda > 0 \tag{31}$$

that is, $\psi(z) = -2^{-n\lambda z}$, which is an extension of both (27) (for $\lambda \to 0$) and (29) (for $\lambda = 1$). Equation (31) is also the log-moment generating function of the random variable $n^{-1}\sum_{t=1}^{n} l(f(S_t), x_t)$, and therefore if it can be minimized *uniformly* for all $\lambda > 0$, this will yield a good large deviations behavior of $n^{-1}\sum_{t=1}^{n} l(f(S_t), x_t)$, because the expression in (31) plays a role in the Chernoff bound on $\Pr\{n^{-1}\sum_{t=1}^{n} l(f(S_t), x_t) < \mu\}$ for every real μ. This probability, in turn, is a reasonable objective function to maximize.

It turns out that if one adopts (31) as a performance criterion, then randomized FSM's may perform better than deterministic FSM's even in the limit $M \to \infty$. Specifically, we next demonstrate by a counterexample that no matter how many states are allowed, there is no deterministic FSM that attains (31) uniformly for every $\lambda > 0$. Assume that $l(b,x) \geq 0$ and $\boldsymbol{x}$ is such that $u_\infty(\boldsymbol{x}) > 0$. Consider a randomized two-state FSM with $q(s \mid x, s') = 1/2$, $x \in \boldsymbol{X}$, $s, s' \in \boldsymbol{S}$. Then,

$$\begin{aligned} &\frac{1}{n\lambda} \log E_q \exp_2\left[-\lambda \sum_{t=1}^{n} l(f(S_t), x_t)\right] \\ &\quad \geq \frac{1}{n\lambda} \log \max_{\boldsymbol{s}^n} \left[2^{-n} \exp_2\left(-\lambda \sum_{t=1}^{n} l(f(s_t), x_t)\right)\right] \\ &\quad = -\min_{\boldsymbol{s}^n} \frac{1}{n} \sum_{t=1}^{n} l(f(s_t),\ x_t) - \frac{1}{\lambda}. \end{aligned} \tag{32}$$

Let $\boldsymbol{X} = \boldsymbol{S} = \boldsymbol{B} = \{0,1\}$, $f(0) = 0$, $f(1) = 1$, and let $l(\cdot,\cdot)$ be the Hamming distance. Thus, the first term on the right-most side of (32) is zero. Now, if λ is chosen larger than $1/u_\infty(\boldsymbol{x})$, then we have demonstrated a simple two-state randomized machine for which

$$\frac{1}{n\lambda} \ln E_q \exp_2\left[-\lambda \sum_{t=1}^{n} l(f(S_t), x_t)\right] > -u_\infty(\boldsymbol{x}), \tag{33}$$

while for the best deterministic FSM, the value $-u_\infty(\boldsymbol{x})$ cannot be exceeded (by definition), even in the limit. The reason for this phenomenon is as follows. If λ is very large, the exponential risk function becomes sensitive to $n^{-1}\sum_t l(b_t, x_t)$. Thus, if there exists a state sequence that yields an average loss smaller than that of any deterministic FSM, then the gain in the exponential risk is so large that even if this state sequence possesses a low probability, its contribution to the expected risk is significant.

In spite of this fact, it is interesting to note that, at least in the case where $l(\cdot,\cdot)$ is as in (2) and $\lambda = 1$, the best performance attainable by a randomized FSM in the sense of (31) can be still be approached by a sufficiently complex deterministic FSM. This follows from the fact [40], [54] that for any finite-state source of the form (30), $-\log P(\boldsymbol{x}^n) \geq U_{\mathrm{LZ}}(\boldsymbol{x}^n) - n\epsilon_n$, where $\epsilon_n \to 0$ and $U_{\mathrm{LZ}}(\boldsymbol{x}^n)$ is the codeword length function associated with Lempel–Ziv algorithm, which in turn is asymptotically lower bounded by the (deterministic) FS compressibility of the infinite sequence.

Finally, we comment that there exists a universal randomized sequential scheme that asymptotically attains (31) uniformly for every $\lambda > 0$. This scheme works as follows: At time instant t randomly select the next state s_{t+1} from the probability distribution

$$p_t(s_t = s \mid x_{t-1} = x,\ s_{t-1} = s') = \frac{n_{t-1}(x, s, s') + 1/2}{n_{t-1}(x, s') + M/2}, \tag{34}$$

where $n_{t-1}(x, s, s')$ is the joint count of $(x_\tau = x, s_{\tau+1} = s, s_\tau = s')$ in $(\boldsymbol{x}^{t-1}, \boldsymbol{s}^{t-1})$ and $n_{t-1}(x, s') = \sum_{s \in \boldsymbol{S}} n_{t-1}(x, s, s')$. The strategy at time t is chosen with respect to the subsequence $\{x_\tau,\ \tau : s_\tau = s_t,\ \tau \leq t-1\}$, as explained in Section III. Note that (34) is in the spirit of the universal predictive measure developed in [45]. However, unlike in [45], (34) serves here as a random mechanism for selecting *states* w.r.t. a given deterministic rather than a random sequence.

While the expected value of $\exp_2\{-\lambda\sum_{t=1}^{n} l(f(S_t), x_t)\}$ in this scheme is exponentially equivalent to $\max_q E_q \exp_2\{-\lambda\sum_{t=1}^{n} l(f(S_t), x_t)\}$, as shown in the Appendix, the main drawback of this scheme is that it does not have an "ergodic property" in the sense that $(n\lambda)^{-1}\log \max_q E_q \exp_2\{-\lambda\sum_{t=1}^{n} l(f(S_t), x_t)\}$ is rarely attained in a single experiment. The reason is that (34) induces a nonergodic probability distribution on $\boldsymbol{s}^n$. Intuitively, (34) describes a self-generating mechanism for selecting states in the sense that at each time instant t it depends on the past realizations $S_1, S_2, \cdots, S_{t-1}$. Thus, if a certain state, for instance, is assigned a low conditional probability at an early time instant t, it will not be likely to appear later on, and hence its conditional probability will reduce even further resulting in a "positive feedback" effect, which makes the convergence to the optimal loss very unstable.

A possible alternative to the above scheme which is applicable in gambling and investment applications is to divide the initial capital into a large number of portions corresponding to a sensibly dense finite grid of points in the space Q_M of all possible M-state conditional distributions q, and to apply in parallel all randomized strategies associated with the grid points. The exponential rate of the total fortune will be dominated by that of the best grid point Q_M, which, in turn,

is close to optimum by continuity considerations. This idea is in the same spirit as in [6].

V. Parametric Schemes

In continuous-alphabet applications of the sequential decision problem considered here, e.g., linear prediction, filtering, system identification, vector quantization, portfolio selection etc., a natural analogue of the FSM that was studied in Section III is the dynamical system model, described as in (16), where x_t, s_t and b_t are now interpreted as vectors in Euclidean spaces, $\boldsymbol{X} = \mathbb{R}^l$, $\boldsymbol{S} = \mathbb{R}^k$, and $\boldsymbol{B} = \mathbb{R}^m$. In particular, the case where f and g are linear mappings, i.e.,

$$s_t = As_{t-1} + Bx_{t-1} \tag{35a}$$

$$b_t = Cs_t, \tag{35b}$$

where A, B and C are, respectively, $k \times k$, $k \times l$ and $m \times k$ matrices, is of great interest, because of its mathematical tractability. Here, the class of allowed schemes is a parametric class, where the parameters are the entries of A, B and C, and the problem is again that of the best choice of these parameters so as to minimize the time-average of some loss function $l(b_t, x_t)$.

An important special case of (35) is the linear prediction problem where x_t and b_t are scalars ($l = m = 1$), A is a fixed $k \times k$ matrix, whose all entries are zero except for the lower off-diagonal entries which are set to unity, $B = (1, 0, 0, \cdots, 0)^{\#}$ is a k dimensional vector, and C is a k dimensional vector whose elements $\{c_i\}_{i=1}^k$ are to be chosen. In this case, $b_t = \sum_{t=1}^k c_i x_{t-i}$, namely, a linear Markovian strategy depending only on the k preceding letters. In the linear prediction problem b_t plays the role of a linear predictor $\hat{x}_t$ of x_t and its performance is measured by the squared error criterion $l(b, x) = (x - b)^2$ (see , e.g., [15]). Observe that once the form of the predictor has been chosen, then the problem of minimizing $n^{-1} \sum_{t=1}^n (x_t - \sum_{i=1}^k c_i x_{t-1})^2$ can be viewed again in the fixed strategy setup of Section II, where we redefine the strategy $\tilde{b}$ as C and the measurements $\tilde{x}_t$ as vectors $(x_{t-k}, \cdots, x_t)$. Hence, one can use the techniques developed in Section II to compare the performance of the recursive least squares (RLS) algorithm [18], [23] which best matches $C = C^t$ at each time instant t to the data see n so far, with the batch procedure, which minimizes the time-average of the squared error along the entire sequence $\boldsymbol{x}^n$. By doing this, we find that the RLS algorithm is universal in the sense of asymptotically attaining the minimum prediction error uniformly for bounded sequences $\boldsymbol{x}^n$. Since the algorithm is recursive there is an explicit expression for the difference $C^{t+1} - C^t$, and hence the result of Theorem 1 can be further developed here. We next demonstrate this point by combining techniques similar to those developed in [15] with the special structure of the RLS algorithm.

The RLS algorithm provides recursively the instantaneous estimate $C^t = \text{argmin}_C \sum_{\tau \le t} (x_\tau - \sum_{i=1}^k c_i x_{\tau-i})^2$ using the matrix inversion lemma. Specifically,

$$C^t = C^{t-1} + K_t(x_t - \sum_{i=1}^k c_i^{t-1} x_{t-i}) \triangleq C^{t-1} + K_t \delta_t, \tag{36}$$

where the k-dimensional vector K_t is jointly updated with the inverse of the unnormalized autocorrelation matrix $P_t = R_t^{-1}$, $(R_t)_{ij}$ being $\sum_{\tau \le t} x_{\tau-i} x_{\tau-j}$, in the following manner.

$$K_t = \frac{P_{t-1}\underline{x}_t}{1 + \underline{x}_t^{\#} P_{t-1} \underline{x}_t}, \tag{37}$$

and

$$P_t = P_{t-1} - \frac{P_{t-1}\underline{x}_t \underline{x}_t^{\#} P_{t-1}}{1 + \underline{x}_t^{\#} P_{t-1} \underline{x}_t}, \tag{38}$$

where $\underline{x}_t \triangleq (x_{t-1}, \cdots, x_{t-k})$. Let λ_t denote the smallest eigenvalue of $\bar{R}_t \triangleq t^{-1} R_t$ and assume that $| x_t | \le G < \infty$ for all t. Similarly to the proof of Theorem 1, we have that

$$\frac{1}{n} \sum_{t=1}^n (x_t - \underline{x}_t^{\#} C^{t-1})^2 \ge \frac{1}{n} \sum_{t=1}^n (x_t - \underline{x}_t^{\#} C^n)^2 \ge \frac{1}{n} \sum_{t=1}^n (x_t - \underline{x}_t^{\#} C^t)^2, \tag{39}$$

where the left-most side corresponds to the average squared error associated with the RLS algorithm, the central expression is the average squared error of the best linear predictor calculated in batch, and the right-most side is associated with an auxiliary anticipating version of the RLS algorithm which has access to x_t at time t. Thus, the difference between the left-most side and the right-most side of (39) serves as an upper bound on the difference between the average squared error of the RLS algorithm and that of the batch predictor. This in turn, is upper bounded as follows. Let $\epsilon_t \triangleq x_t - \underline{x}_t^{\#} C^t$. Then,

$$\begin{aligned}
&\frac{1}{n} \sum_{t=1}^n (x_t - \underline{x}_t^{\#} C^{t-1})^2 - \frac{1}{n} \sum_{t=1}^n (x_t - \underline{x}_t^{\#} C^t)^2 \\
&= \frac{1}{n} \sum_{t=1}^n \underline{x}_t^{\#} (C^t - C^{t-1})(\epsilon_t + \delta_t) \\
&= \frac{1}{n} \sum_{t=1}^n \underline{x}_t^{\#} K_t \delta_t (\epsilon_t + \delta_t) \\
&\le \frac{1}{n} \sum_{t=1}^n \frac{\underline{x}_t^{\#} P_{t-1} \underline{x}_t}{1 + \underline{x}_t^{\#} P_{t-1} \underline{x}_t} |\delta_t \cdot (\epsilon_t + \delta_t)| \\
&= \frac{1}{n} \sum_{t=1}^n \frac{\underline{x}_t^{\#} \bar{R}_{t-1}^{-1} \underline{x}_t}{t - 1 + \underline{x}_t^{\#} \bar{R}_{t-1}^{-1} \underline{x}_t} |\delta_t \cdot (\epsilon_t + \delta_t)| \\
&\le \frac{1}{n} \sum_{t=1}^n \frac{kG^2/\lambda_{t-1}}{t - 1 + kG^2/\lambda_{t-1}} |\delta_t \cdot (\epsilon_t + \delta_t)| \\
&= \frac{kG^2}{n} \sum_{t=1}^n \frac{1}{(t-1)\lambda_{t-1} + kG^2} |\delta_t \cdot (\epsilon_t + \delta_t)|.
\end{aligned} \tag{40}$$

Suppose now that there exists a positive number λ_∞ such that $\lambda_t \ge \lambda_\infty$ for all sufficiently large t. Since $| x_t | \le G$ by assumption, the instantaneous errors δ_t and ϵ_t are bounded as well, and (40) is bounded by an harmonic series similarly to (14) resulting in a bound proportional to $kG^2 \lambda_\infty^{-1} n^{-1} \log n$. Note, that even if λ_t tends to zero but slower than $1/t$, then still (40) may decay with n but not as fast as $n^{-1} \log n$.

The sensitivity factor λ_∞^{-1} (see also [6]) can be controlled by orthonormalizing the linear space spanned by all permissible predictors. Specifically, as an alternative to the linear predictor above, consider a predictor of the form

$$b_t = \sum_{i=1}^{r} c_i \phi_i(x_{t-1}, \cdots, x_{t-k}), \tag{41}$$

where $\{\phi_i(\cdot)\}_{i=1}^r$ is a family of functions with disjoint supports. For example, consider a uniform partition of $[-G,\ G]$ to intervals of width Δ, which induces a uniform partition of $[-G,\ G]^k$ into k dimensional cubes. If ϕ_i is defined as the indicator function of the ith cube, $i = 1, 2, \cdots, r = (2G/\Delta)^k$, then (41) can approximate a wide class of *nonlinear* smooth functions on the k dimensional cube $[-G,\ G]^k$, and at the same time the problem of selecting $\{c_i\}_{i=1}^r$ is associated with solving *linear* equations, provided that the mean square error criterion is adopted. Furthermore, since $\{\phi_i\}$ are orthonormal, there is no need for matrix inversion and no sensitivity problems occur. The weakness of (41) is of course the typically huge number of parameters $r = (2G/\Delta)^k$ needed to cover faithfully the domain $[-G,\ G]^k$ resulting in a relatively large $O(r \log n/n)$ term as in (18),

APPENDIX

For the proofs of the previous theorems, it will be more convenient to regard the average loss as a functional of the empirical measure extracted from $(\boldsymbol{x}^n, \boldsymbol{s}^n)$ (with respect to a given g) rather than a direct function of $\boldsymbol{x}^n$. To guarantee desirable shift invariant properties of these empirical measures (see also [12]), these will be defined with the cyclic convention that (x_n, s_n) precedes (x_1, s_1). In order that the state sequence will be cyclic as well, i.e., $s_1 = g(x_n, s_n)$, assume without loss of generality that g is irreducible (i.e., all states communicate) and add $\boldsymbol{x}^n$ with an appropriate suffix of length l (which is independent of n) such that $s_1 = g(x_{n+l}, s_{n+l})$. Of course, for $n \gg l$ this suffix does not affect the empirical measure.

The following notation will be used. Let $n_g(x, s)$ denote the joint count of $x_t = x$ and $s_t = s$ in the pair sequence $(\boldsymbol{x}^n, \boldsymbol{s}^n)$ with this cyclic convention. Let $p_n^g(x, s) \triangleq n_g(x, s)/n$, $p_n^g(s) \triangleq \sum_{x \in \boldsymbol{X}} p_n^g(x, s)$, and $p_n^g(x|s) \triangleq p_n^g(x, s)/p_n^g(s)$. For a given $s \in \boldsymbol{S}$ the conditional probability distribution $\{p_n^g(x|s)\}_{x \in \boldsymbol{X}}$ will be denoted by $P_{n,s}^g$. When the state sequence is not generated by a deterministic next-state function, i.e., in the randomized case, the superscript g will be omitted. Let X and S denote random variables governed by the joint probability distribution $p_n^g(x, s)$. Note that $u(\boldsymbol{x}^n; g)$ can be rewritten as a functional of the empirical conditional distributions $P_{n,s}^g$ and hence, will be denoted also by $U_n^g(X|S)$, i.e., a conditional Bayes envelope. Specifically,

$$u(\boldsymbol{x}^n; g) \equiv U_n(X|S) = \sum_{s \in \boldsymbol{S}} p_n^g(s) \inf_{b \in \boldsymbol{B}} E_{P_{n,s}^g} l(b,\ X).$$

When g is Markovian of order k, then $u(\boldsymbol{x}^n; \boldsymbol{\mathcal{M}}_k)$ will be denoted also by $U_n(X|X^k)$, where X^k denotes a random k-tuple governed by the empirical probability of k-tuples extracted from $\boldsymbol{x}^n$.

Proof of Theorem 2: The proof technique is similar to that of [12, Theorem 2]. Let g be an arbitrary next-state function of an M-state machine. Let $U_n^g(X \mid X^k, \tilde{S})$ be the conditional Bayes envelope associated with a combined $(M \cdot A^k)$-state machine, where the current state is $s_t^k = (s_{t-k}, x_{t-k}, x_{t-k+1}, \cdots, x_{t-1})$, s_{t-k} being the state (at time $(t-k)$) associated with $g(\cdot, \cdot)$. Intuitively, this machine performs better than the M-state machine associated with $g(\cdot, \cdot)$ because

$$\begin{aligned} s_t = g^k(s_t^k) &= g^k(s_{t-k}, x_{t-k}, \cdots, x_{t-1}) \\ &\triangleq g(g(\cdots g(s_{t-k}, x_{t-k}), x_{t-k+1}), \cdots, x_{t-1}) \end{aligned}$$

is a many-to-one mapping from s_t^k to s_t and hence s_t^k contains more information of the past. Mathematically, we have

$$\begin{aligned} U_n^g(X \mid S) &= \sum_s p_n^g(s) \cdot E_{P_{n,s}^g} l(b^*(P_{n,s}^g), X) \\ &= \sum_s \sum_{s^k:\ g^k(s^k)=s} p_n^g(s^k) E_{P_{n,s^k}^g} l(b^*(P_{n,s}^g), X) \\ &\geq \sum_s \sum_{s^k:\ g^k(s^k)=s} p_n^g(s^k) E_{P_{n,s^k}^g} l(b^*(P_{n,s^k}^g), X) \\ &= U_n^g(X \mid S^k) = U_n(X|X^k, \tilde{S}), \end{aligned} \tag{A.1}$$

where S^k is a random vector, governed by P_n, consisting of X^k and the state $\tilde{S}$ preceding X^k. In a similar manner it is obvious that $U_n^g(X \mid S^k) \leq U_n(X \mid X^k)$ but we show that the difference is small. To this end, we upper bound the difference $U_n(X \mid X^k) - U_n^g(X \mid S^k)$.

Let $0 \leq j \leq k$, and consider the difference between the associated empirical conditional Shannon entropies $H_n(X \mid X^j) - H_n^g(X \mid S^j)$, (which are obtained as a special case where $l(\cdot, \cdot)$ is logarithmic). Since S^j contains X^j as a component,

$$\begin{aligned} &H_n(X \mid X^j) - H_n^g(X \mid S^j) \\ &\quad = \sum_{s^j \in \boldsymbol{S} \times \boldsymbol{X}^j} p_n^g(s^j) \sum_{x \in \boldsymbol{X}} p_n^g(x \mid s^j) \log \frac{p_n^g(x \mid s^j)}{p_n(x \mid x^j)} \\ &\quad \geq \frac{1}{2 \ln 2} \sum_{s^j} p_n^g(s^j) \left[\sum_x |p_n^g(x \mid s^j) - p_n(x \mid x^j)| \right]^2 \\ &\quad \geq \frac{1}{2C \ln 2} \cdot \sum_{s^j} p_n^g(s^j) \left[\Delta(P_{n,s^j}^g \| P_{n,x^j}) \right]^{2/\delta}, \end{aligned} \tag{A.2}$$

where we have used Pinsker's inequality [9, ch. 3, problem 17] for the first inequality and the assumption of Theorem 2 for the second. Next, assume without loss of generality that $\delta \leq 2$. (If conversely, $\delta > 2$, then use $\|P-Q\|^\delta \leq \|P-Q\|^{\delta-2} \|P-Q\|^2 \leq 2^{\delta-2} \|P-Q\|^2$, absorb $2^{\delta-2}$ in C, and set $\delta = 2$ in (20)). Thus, (A.2) can be further lower bounded using Jensen's inequality:

$$\begin{aligned} &H_n(X \mid X^j) - H_n^g(X \mid S^j) \\ &\quad \geq \frac{1}{2C \ln 2} \left[\sum_{s^j} p_n^g(s^j) \cdot \Delta(P_{n,s^j}^g \| P_{n,x^j}) \right]^{2/\delta} \\ &\quad = \frac{1}{2C \ln 2} \left[U_n(X \mid X^j) - U_n^g(X \mid S^j) \right]^{2/\delta}. \end{aligned} \tag{A.3}$$

Finally,

$$\begin{aligned} U_n(X \mid X^k) - U_n^g(X \mid S) &\le \frac{1}{k+1}\sum_{j=0}^{k}\left[U_n(X \mid X^j) - U_n^g(X \mid S^j)\right] \\ &\le \left(\frac{1}{k+1}\sum_{j=0}^{k}\left[U_n(X \mid X^j) - U_n^g(X \mid S^j)\right]^{\frac{2}{\delta}}\right)^{\frac{\delta}{2}} \\ &\le \left(\frac{2C\ln 2}{k+1}\sum_{j=0}^{k}\left[H_n(X \mid X^j) - H_n^g(X \mid S^j)\right]\right)^{\frac{\delta}{2}} \\ &= \left(\frac{2C\ln 2}{k+1}\left[H_n(X^{k+1}) - H_n^g(X^{k+1} \mid \tilde{S})\right]\right)^{\frac{\delta}{2}} \\ &= \left[\frac{2C\ln 2}{k+1}\,H_n^g(S)\right]^{\delta/2} \le \left(\frac{2C\ln 2\log M}{k+1}\right)^{\delta/2} \\ &= \left(\frac{2C\ln M}{k+1}\right)^{\delta/2}, \end{aligned} \tag{A.4}$$

where the first inequality follows from the fact that conditioning reduces the Bayes envelope (similarly to (A.1)), the second inequality results from Jensen's inequality, and third inequality is implied by (A.3). Since g is an arbitrary M-state machine this completes the proof of Theorem 2. □

Proof of Theorem 3: >From Theorem 2 and the fact that a Markov strategy is a special case of a FS strategy it is apparent that an equivalent definition of $u_\infty(\boldsymbol{x})$ is

$$u_\infty(\boldsymbol{x}) = \lim_{k\to\infty}\limsup_{n\to\infty} u(\boldsymbol{x}^n; \boldsymbol{\mathcal{M}}_k). \tag{A.5}$$

Since for any nonanticipating kth order Markov strategy [2, (3.6)]

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n} l(b_t, X_t) \ge U(X|X^\infty), \tag{A.6}$$

almost surely, then clearly,

$$\limsup_{n\to\infty} u(X_1, \cdots, X_n; \boldsymbol{\mathcal{M}}_k) \ge U(X|X^\infty), \tag{A.7}$$

almost surely. By taking the limit as $k \to \infty$, we get $u_\infty(X_1, X_2, \cdots) \ge U(X|X^\infty)$. Thus, to complete the proof, it remains to show that converse inequality $u_\infty(X_1, X_2, \cdots) \le U(X|X^\infty)$ also holds, almost surely. Let $b^*(x_{t-k}, \cdots, x_{t-1})$ be the best kth-order Markov strategy in the sense of achieving the infimum of $E\{l(b, X_t)|x_{t-k}, \cdots, x_{t-1}\}$. Since $u(\boldsymbol{x}^n; \boldsymbol{\mathcal{M}}_k)$ is attained by the best kth-order Markov strategy for the *given individual sequence,* it is clear that for every realization $X_1, X_2, \cdots$,

$$u(X_1, \cdots, X_n; \boldsymbol{\mathcal{M}}_k) \le \frac{1}{n}\sum_{t=1}^{n} l(b^*(X_{t-k}, \cdots, X_{t-1}), X_t). \tag{A.8}$$

Now, by Birkhoff's ergodic theorem, the right hand side of (A.8) tends to $U(X|X^k)$ almost surely as $n \to \infty$. Thus,

$$\limsup_{n\to\infty} u(X_1, \cdots, X_n; \boldsymbol{\mathcal{M}}_k) \le U(X|X^k), \tag{A.9}$$

almost surely. Finally, by taking the limit as $k \to \infty$ on both sides of (A.9), we get $u_\infty(X_1, X_2, \cdots) \le U(X|X^\infty)$, which completes the proof of Theorem 3. □

Proof of Theorem 4: Note that

$$\begin{aligned} E_q\left\{\frac{1}{n}\sum_{t=1}^{n} l(f(S_t), x_t)\right\} &= E_q\sum_{x,s} p_n(x,s)l(f(s),x) \\ &\triangleq \sum_{x,s}\mu_n^q(x,s)l(f(s),x) \triangleq E_{\mu_n^q}\, l(f(S),X), \end{aligned} \tag{A.10}$$

where $\mu_n^q(x,s) \triangleq E_q p_n(x,s)$ is a joint probability of x and s induced by the expected empirical measure P_n with respect to q. Let (X, S, S^j) be a triple of random variables induced by μ_n^q, i.e., the expectations (w.r.t. q) of the relative frequencies of the joint events $\{x_t = x, S_t = s, S_t^j = s^j\}$, $x \in \boldsymbol{X}$, $s \in \boldsymbol{S}$, $s^j \in \boldsymbol{S} \times \boldsymbol{X}^j$, where $S_t^j = (S_{t-j}, x_{t-j}, x_{t-j+1}, \cdots, x_{t-1})$. Observe (A.11) (see the equation at the bottom of the page) where $\Pr\{\cdot\}$ is with respect to q. Equation (A.11) implies that $X \ominus S^j \ominus S$ is a Markov chain under μ_n^q.

$$\begin{aligned} \mu_n^q(X = x,\ S = s,\ S^j = (\sigma, x^j)) &= \frac{1}{n}\sum_{t=1}^{n} E_q\delta(s_t^j = (\sigma, x^j), s_t = s, x_t = x) \\ &= \frac{1}{n}\sum_{t:\ x_t = x,\ x_{t-j}^{t-1} = x^j} \Pr\{s_{t-j} = \sigma,\ s_t = s\} \\ &= \frac{1}{n}\sum_{t:\ x_t = x,\ x_{t-j}^{t-1} = x^j} \Pr\{s_{t-j} = \sigma\}\cdot\Pr\{s_t = s \mid s_{t-j} = \sigma,\ x_{t-j}^{t-1} = x^j,\ x_t = x\} \\ &= \frac{1}{n}\sum_{t:\ x_t = x,\ x_t = j^j - 1 = x^j} \Pr\{s_{t-j} = \sigma\}\cdot\Pr\{s_t = s \mid s_{t-j} = \sigma,\ x_{t-j}^{t-1} = x^j\}, \end{aligned} \tag{A.11}$$

Let $\mu^q_{n,s} \triangleq \{\mu^q_n(x \mid s)\}_{x\in\boldsymbol{X}}$, $\mu^q_{n,s^j} \triangleq \{\mu^q_n(x \mid s^j)\}_{x\in\boldsymbol{X}}$ and $\mu^q_{n,s,s^j} \triangleq \{\mu^q_n(x \mid s, s^j)\}_{x\in\boldsymbol{X}}$. Then, for every $q \in Q_M$,

$$\begin{aligned} V^q_n(X \mid S) &\triangleq \sum_{x,s} \mu^q_n(x,s) l(b^*(\mu_n, s^q), x) \\ &= \sum_{s,s^j} \mu^q_n(s,\ s^j) \sum_x \mu^q_n(x \mid s^j) l(b^*(\mu^q_{n,s}), x) \\ &\geq \sum_{s,s^j} \mu^q_n(s, s^j) \sum_x \mu^q_n(x \mid s^j) l(b^*(\mu^q_{n,s^j}), x) \\ &= E_{\mu^q_n} l(b^*(\mu^q_{n,S^j}),\ X), \\ &\triangleq V^q_n(X \mid S^j), \end{aligned} \tag{A.12}$$

which is analogous of (A.1) with the deterministic next-state function g replaced by the randomized rule q. The rest of the proof is identical to the proof of Theorem 2, where p's should be replaced by μ's, g's are substituted by q's, U's are changed to V's, and Shannon entropies are now defined with respect to μ^q_n. □

The Performance of the Proposed Universal Randomized Scheme

We first derive an upper bound on (31) and then demonstrate that our scheme attains this upper bound asymptotically. For every randomized M-state machine, we have

$$\begin{aligned} &\frac{1}{n\lambda} \log E_q \exp_2\left[-\lambda \sum_{t=1}^n l(f(S_t), x_t)\right] \\ &\leq \frac{1}{n\lambda} \log \max_{f\in F_M} \sup_{q\in Q_M} \sum_{\boldsymbol{s}^n} \left[\prod_{t=1}^n q(s_{t+1} \mid x_t, s_t)\right] \\ &\quad \cdot \exp_2\left[-\lambda \sum_{t=1}^n l(f(s_t), x_t)\right] \\ &\leq \frac{1}{n\lambda} \log \sum_{\boldsymbol{s}^n} \left[\sup_{q\in Q_M} \prod_{t=1}^n q(s_{t+1} \mid x_t, s_t)\right] \\ &\quad \cdot \exp_2\left[-\lambda \min_{f\in F_M} \sum_{t=1}^n l(f(s_t), x_t)\right] \\ &= \frac{1}{n\lambda} \log \sum_{\boldsymbol{s}^n} \exp_2\{-n[H_n(S \mid X, S') + \lambda U_n(X \mid S)]\}, \end{aligned} \tag{A.13}$$

where $U_n(X \mid S)$ is defined for a given state sequence $\boldsymbol{s}^n$, similarly to $U^q_n(X \mid S)$, but with respect to the empirical probability distribution $\{p_n(x,s)\}_{x\in\boldsymbol{X},\ s\in\boldsymbol{S}}$ induced by the pair sequence $(\boldsymbol{x}^n, \boldsymbol{s}^n)$. $H_n(S \mid X, S')$ is the empirical conditional Shannon entropy associated with the empirical probability distribution

$$\begin{aligned} &p_n(X = x, S = s, S' = s') \\ &\qquad = n^{-1} \sum_t \delta(x_t = x,\ s_{t+1} = s,\ s_t = s'). \end{aligned}$$

Consider now the sequential randomized scheme proposed in Section IV. First recall that, similarly to (1), for every $(\boldsymbol{x}^n, \boldsymbol{s}^n)$,

$$\begin{aligned} &\prod_{t=1}^n p_t(s_t \mid x_{t-1}, s_{t-1}) \\ &\geq \exp_2\left\{-n[H_n(S \mid X, S') + O(\frac{M}{n} \log \frac{n}{M})]\right\} \end{aligned} \tag{A.14}$$

and that similarly to (18) for a given state sequence $\boldsymbol{s}^n$

$$\frac{1}{n} \sum_{t=1}^n l(b^*(P_{t-1,s_t}), x_t) \leq U_n(X \mid S) + O(\frac{M}{n} \log \frac{n}{M}). \tag{A.15}$$

Thus, for the randomized scheme considered here we have

$$\begin{aligned} &\frac{1}{n\lambda} \log E \exp_2\left[-\lambda \sum_{t=1}^n l(b_t, x_t)\right] \\ &= \frac{1}{n\lambda} \log \sum_{\boldsymbol{s}^n} \left[\prod_{t=1}^n p_t(s_t \mid x_{t-1}, s_t)\right] \\ &\quad \cdot \exp_2\left[-\lambda \sum_{t=1}^n l(b^*(P_{t-1,s_t}), x_t)\right] \\ &\geq \frac{1}{n\lambda} \log \sum_{\boldsymbol{s}^n} \exp_2\{-n[H_n(S \mid X, S') \\ &\qquad + \lambda U_n(X \mid S) + O(\frac{M}{n} \log \frac{n}{M})]\} \\ &\geq \frac{1}{n\lambda} \log \sum_{\boldsymbol{s}^n} \exp_2\{-n[H_n(S \mid X, S') \\ &\qquad + \lambda U_n(X \mid S)]\} - O(\frac{M}{n} \log \frac{n}{M}), \end{aligned} \tag{A.16}$$

which agrees with the upper bound (A.13) up to a term of $O(M/n \log n/M)$. Since the scheme does not depend on λ it attains (31) uniformly.

ACKNOWLEDGMENT

The authors are grateful to Dr. M. Gutman for useful discussions in the course of this work. Useful comments made by the anonymous references are greatly appreciated.

REFERENCES

[1] P. H. Algoet, "Universal schemes for prediction, gambling, and portfolio selection," *Ann. Probab.*, Apr. 1992.

[2] ——, "The strong law of large numbers for sequential decisions under uncertainty," preprint.

[3] P. H. Algoet and T. M. Cover, "Asymptotic optimality and asymptotic equipartition properties of log-optimum investment," *Ann. Probab.*, vol. 16, no. 2, pp. 876–898, 1988.

[4] D. Blackwell, "An analog to the minimax theorem for vector payoffs," *Pac. J. Math.*, vol. 6, pp. 1–8, 1956.

[5] T. M. Cover, "Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin," Tech. Rep. 12, Dept. of Statist., Stanford Univ., 1974.

[6] ——, "Universal portfolios," in *Math. Finance*, vol. 1, no. 1, pp. 1–29, Jan. 1991.

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

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

[9] I. Csiszár and J. Körner, *Information Theory: Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.

[10] L. D. Davisson, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory,* vol. IT-29, no. 2, pp. 211–215, 1983.
[11] M. Feder, "Gambling using a finite–state machine," *IEEE Trans. Inform. Theory,* vol. IT-37, no. 5, pp. 1459–1465, Sept. 1991.
[12] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory,* vol. 38, no. 4, pp. 1258–1270, July 1992.
[13] A. Gill, *Introduction to the Theory of Finite-State Machines.* New York: McGraw Hill, 1962.
[14] D. C. Gilliland, "Asymptotic risk stability resulting from play against the past in a sequence of decision problems," *IEEE Trans. Inform. Theory,* vol. IT-18, no. 5, pp. 614–617, Sept. 1972.
[15] ——, "Sequential compound estimation," *Ann. Math. Statist.,* vol. 39, no. 6, pp. 1890–1904, 1968.
[16] D. C. Gilliland and M. K. Helmers, "On the continuity of the Bayes response," *IEEE Trans. Inform. Theory,* vol. IT-24, no. 4, pp. 506–508, July 1978.
[17] D. C. Gilliland and J. F. Hannan, "On an extended compound decision problem," *Ann. Math. Statist.,* vol. 40, no. 5, pp. 1536–1541, 1969.
[18] G. C. Goodwin and R. L. Payne, *Dynamic System Identification: Experiment Design and Data Analysis.,* (Mathematics in Science and Engin ϵ ring), vol. 136. New York: Academic Press, 1977.
[19] R. M. Gray, "Vector quantization," *IEEE ASSP Mag.,* vol. 1, no. 2, pp. 4–29, 1984.
[20] ——, *Probability, Random Processes, and Ergodic Properties.* New York: Springer-Verlag, 1988.
[21] J. F. Hannan, "Approximation to Bayes risk in repeated plays," in *Contributions to the Theory of Games, vol. III, Annals of Mathematics Studies.* Princeton, NJ: Princeton Univ. Press,1957, no. 39, pp. 97–139.
[22] J. F. Hannan and H. Robbins, "Asymptotic solutions of the compound decision problem for two completely specified distributions," *Ann. Math. Statist.,* vol. 26, pp. 37–51, 1957.
[23] S. Haykin, *Adaptive Filter Theory.* Englewood Cliffs, NJ: Prentice-Hall, 1986.
[24] M. E. Hellman, "The effects of randomization on finite-memory decision schemes," *IEEE Trans. Inform. Theory,* IT-18, no. 4, pp. 499–502, 1972.
[25] M. E. Hellman and T. M. Cover, "Learning with finite memory," *Ann. Math. Statist.,* vol. 41, no. 3, pp. 765–782, 1970.
[26] ——, "On memory saved by randomization," *Ann. Math. Statist.,* vol. 42, no. 3, pp. 1075–1078, 1971.
[27] N. S. Jayant and P. Noll, *Digital Coding of Waveforms.* Englewood Cliffs, NJ: Prentice-Hall, 1984.
[28] M. V. Johns, Jr., "Two-action compound decision problems," *Proc. Fifth Berkeley Symposium Mathematical and Statistical Probability.* Berkeley, CA: Univ. of California Press, 1967vol. 1, pp. 463–478.
[29] J. L. Kelly, Jr., "A new interpretation of information rate," *Bell Syst. Tech. J.,* vol. 35, pp. 917–926, 1956.
[30] R. E. Krichevsky, "The relation between redundancy coding and the reliability of information from a source," *Probl. Inform. Transm.,* vol. 4, no. 3, pp. 37–45, 1968.
[31] R. E. Krichevsky and V. K. Trofimov, "The performance of universal encoding," *IEEE Trans. Inform. Theory,* IT-27, no. 2, pp. 199–207, Mar. 1981.
[32] R. Krzystofowicz, "Markovian forecast processes," *J. Amer. Statist. Assoc.,* vol. 82, no. 397, pp. 31–37, Mar. 1987.
[33] J. Rissanen and G. G. Langdon, Jr., "Universal modeling and coding," *IEEE Trans. Inform. Theory,* vol. IT-27, no. 1, pp. 12–23, Jan. 1981.
[34] F. T. Leighton and R. L. Rivest, "Estimating a probability using finite memory," *IEEE Trans. Inform. Theory,* vol. IT-32, no. 6, pp. 733–742, Nov. 1986.
[35] Y. Linde, A. Buzo, and R. M. Gray, "An algorithm for vector quantizer design," *IEEE Trans. Commun.,* vol. COM-28, no. 1, pp. 84–95, 1980.
[36] J. Makhoul, "Linear prediction: A tutorial review," *Proc. IEEE,* vol. PROC-63, no. 4, 1975.
[37] J. Makhoul, S. Roucos, and H. Gish, "Vector quantization in speech coding," *Proc. IEEE,* vol. PROC-73, no. 11, pp. 1551–1588, 1985.
[38] J. D. Markel and A. H. Gray, Jr., *Linear Prediction of Speech.* New York: Springer-Verlag, 1976.
[39] Y. Nogami, "The k-extended set-compound estimation problem in a nonregular family of distributions over $[\theta, \theta + 1)$," *Ann. Inst. Statist. Math.,* vol. 31A, pp. 169–176, 1979.
[40] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and on the redundancy of the Lempel–Ziv algorithm," *IEEE Trans. Inform. Theory,* vol. 38, no. 1, pp. 66–72, Jan. 1992.
[41] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory,* vol. IT-30, no. 4, pp. 629–636, July 1984.
[42] ——, "Stochastic complexity and modeling," *Ann. Statist.,* vol. 14, no. 3, pp. 1080–1100, 1986.
[43] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," in *Proc. 2nd Berkeley Symp. Math. Statist. Probab.,* 1951, pp. 131–148.
[44] B. Ya. Ryabko, "Twice-universal coding," *Probl. Inform. Transm.,* vol. 20, pp. 173–177, July–Sept. 1984.
[45] ——, "Prediction of random sequences and universal coding," *Probl. Inform. Transm.,* vol. 24, pp. 87–96, Apr.–June 1988.
[46] E. Samuel, "Asymptotic solutions of the sequential compound decision problem," *Ann. Math. Statist.,* vol. 34, pp. 1079–1095, 1963.
[47] ——, "Convergence of the losses of certain decision rules for the sequential compound decision problem," *Ann. Math. Statist.,* vol. 35, pp. 1606–1621, 1964.
[48] C. P. Schnorr, "A unified approach to the definition of random sequences," *Math. Syst. Theory,* vol. 5, no. 3, pp. 246–258, 1971.
[49] B. O. Shubert "Finite-memory classification of Bernoulli sequences using reference samples," *IEEE Trans. Inform. Theory,* vol. IT-20, no. 3, pp. 384–387, 1974.
[50] D. D. Swain, "Bounds and rates of convergence for the extended compound estimation problem in the sequence case," Tech. Rep. no. 81, Dept. of Statistics, Stanford Univ., Stanford, CA, 1965.
[51] J. Van Ryzin, "The sequential compound decision problem with $m \times n$ finite loss matrix," *Ann. Math. Statist.,* vol. 37, pp. 954–975, 1966.
[52] S. B. Vardeman, "Admissible solutions of k-extended finite state set and sequence compound decision problems," *J. Multivariate Anal.,* vol. 10, pp. 426–441, 1980.
[53] J. Ziv and A. Lempel, "Compression of individual sequences via variable-rate coding," *IEEE Trans. Inform. Theory,* vol. IT-24, no. 5, pp. 530–536, Sept. 1978.
[54] J. Ziv, "Compression, tests for randomness, and estimating the statistical model of an individual sequence," in *Sequences,* R. M. Capocelli, Ed. New York: Springer-Verlag, 1990, pp. 366–373.