

# Universal Prediction of Individual Sequences

Meir Feder, *Member, IEEE*, Neri Merhav, *Member, IEEE*, and  
Michael Gutman, *Member, IEEE*

**Abstract**—The problem of predicting the next outcome of an individual binary sequence using finite memory, is considered. The finite-state predictability of an infinite sequence is defined as the minimum fraction of prediction errors that can be made by any finite-state (FS) predictor. It is proved that this FS predictability can be attained by universal sequential prediction schemes. Specifically, an efficient prediction procedure based on the incremental parsing procedure of the Lempel–Ziv data compression algorithm is shown to achieve asymptotically the FS predictability. Finally, some relations between compressibility and predictability are pointed out, and the predictability is proposed as an additional measure of the complexity of a sequence.

**Index Terms**—Predictability, compressibility, complexity, finite-state machines, Lempel–Ziv algorithm.

## I. INTRODUCTION

IMAGINE an observer receiving sequentially an arbitrary deterministic binary sequence  $x_1, x_2, \dots$ , and wishing to predict at time  $t$  the next bit  $x_{t+1}$  based on the past  $x_1, x_2, \dots, x_t$ . While only a limited amount of information from the past can be memorized by the observer, it is desired to keep the relative frequency of prediction errors as small as possible in the long run.

It might seem surprising, at first glance, that the past can be useful in predicting the future because when a sequence is arbitrary, the future is not necessarily related to the past. Nonetheless, it turns out that sequential (randomized) prediction schemes exist that utilize the past, whenever helpful in predicting the future, as well as any finite-state (FS) predictor. A similar observation has been made in data compression [1] and gambling [2]. However, while in these problems a conditional probability of the next outcome is estimated, here a decision is to be made for the *value* of this outcome, and thus it cannot be deduced as a special case of either of these problems.

Sequential prediction of binary sequences has been considered in [3]–[5], where it was shown that a universal predic-

tor, performing as well as the best fixed (or single-state) predictor, can be obtained using the theory of compound sequential Bayes decision rules developed in [6] and [7] and the approachability-excludability theory [8], [9]. In [5], this predictor is extended to achieve the performance of the best Markov predictor, i.e., an FS predictor whose state is determined by a finite number (order) of successive preceding outcomes. Our work extends these results by proving the existence and showing the structure of universal predictors that perform as well as *any* FS predictor and by providing a further understanding of the sequential prediction problem.

Analogously to the FS compressibility defined in [1], or the FS complexity defined in [2], we define the FS predictability of an infinite individual sequence as the minimum asymptotic fraction of errors that can be made by any FS predictor. This quantity takes on values between zero and a half, where zero corresponds to perfect predictability and a half corresponds to total unpredictability. While the definition of FS predictability enables a different optimal FS predictor for each sequence, we demonstrate *universal* predictors, independent of the particular sequence, that always attain the FS predictability.

This goal is accomplished in several steps. In one of these steps, an auxiliary result which might be interesting in its own right is derived. It states that the FS predictability can be always nearly attained by a Markov predictor. Furthermore, if the Markov order grows with time at an appropriate rate, then the exact value of the FS predictability is attained asymptotically. In particular, a prediction scheme, based on the Lempel–Ziv (LZ) parsing algorithm, can be viewed as such a Markov predictor with a time-varying order and hence attaining the FS predictability.

The techniques and results presented in this paper are not unique to the prediction problem, and they can be extended to more general sequential decision problems [10]. In particular, when these techniques are applied to the data compression problem, the LZ algorithm can be viewed as a universal Markov encoder of growing order which can be analyzed accordingly. This observation may add insight to why the LZ data compression method works well.

Finally, we introduce the notion of predictability as a reasonable measure of complexity of a sequence. It is demonstrated that the predictability of a sequence is not uniquely determined by its compressibility. Nevertheless, upper and lower bounds on the predictability in terms of the compressibility are derived which imply the intuitively appealing result that a sequence is perfectly predictable iff it is totally redundant and conversely, a sequence is totally unpredictable iff it is incompressible. Since the predictability is not uniquely

Manuscript received April 22, 1991; revised October 31, 1991. This work was supported in part by the Wolfson Research Awards administered by the Israel Academy of Science and Humanities, at Tel-Aviv University. This work was partially presented at the 17th Convention of Electrical and Electronics Engineers in Israel, May 1991.

M. Feder is with the Department of Electrical Engineering-Systems, Tel-Aviv University, Tel-Aviv, 69978, Israel.

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

M. Gutman was with the Department of Electrical Engineering, Technion—Israel Institute of Technology, Haifa, 32000, Israel. He is now with Intel Israel, P.O. Box 1659, Haifa, 31015, Israel.

IEEE Log Number 9106944.

determined by the compressibility, it is a distinct feature of the sequence that can be used to define a *predictive complexity* which is different from earlier definitions associated with the description length i.e., [11]–[13]. Our definition of predictive complexity is also different from that of [14] and [15], which again is a description length complexity but defined in a predictive fashion.

### II. FINITE-STATE PREDICTABILITY

Let  $\mathbf{x} = x_1, x_2, \dots$  be an infinite binary sequence. The *prediction rule*  $f(\cdot)$  of an FS predictor is defined by

$$\hat{x}_{t+1} = f(s_t), \quad (1)$$

where  $\hat{x}_{t+1} \in \{0, 1\}$  is the predicted value for  $x_{t+1}$ , and  $s_t$  is the current state which takes on values in a finite set  $\mathcal{S} = \{1, 2, \dots, S\}$ . We allow stochastic rules  $f$ , namely, selecting  $\hat{x}_{t+1}$  randomly with respect to a conditional probability distribution, given  $s_t$ . The state sequence of the finite-state machine (FSM) is generated recursively according to

$$s_{t+1} = g(x_t, s_t). \quad (2)$$

The function  $g(\cdot, \cdot)$  is called the *next-state function* of the FSM. Thus, an FS predictor is defined by a pair  $(f, g)$ .

Consider first a finite sequence  $x_1^n = x_1, \dots, x_n$  and suppose that the initial state  $s_1$  and the next-state function  $g$  (and hence the state sequence) are provided. In this case, as discussed in [2], the best prediction rule for the sequence  $x_1^n$  is deterministic and given by

$$\hat{x}_{t+1} = f(s_t) = \begin{cases} \text{“0”}, & \text{if } N_n(s_t, 0) > N_n(s_t, 1), \\ \text{“1”}, & \text{otherwise,} \end{cases} \quad (3)$$

where  $N_n(s, x)$ ,  $s \in \mathcal{S}$ ,  $x \in \{0, 1\}$  is the joint count of  $s_t = s$  and  $x_{t+1} = x$  along the sequence  $x_1^n$ . Note that this optimal rule depends on the entire sequence  $x_1^n$  and hence cannot be determined sequentially.

Applying (3) to  $x_1^n$ , the minimum fraction of prediction errors is

$$\pi(g; x_1^n) = \frac{1}{n} \sum_{s=1}^S \min \{N_n(s, 0), N_n(s, 1)\}, \quad (4)$$

where the notation emphasizes the dependence of (4) on the state sequence via the next-state function. Define the minimum fraction of prediction errors with respect to all FSM's with  $S$  states as the *S-state predictability* of  $x_1^n$ ,

$$\pi_S(x_1^n) = \min_{g \in G_S} \pi(g; x_1^n), \quad (5)$$

where  $G_S$  is the set of all  $S^{2S}$  next-state functions corresponding to  $S$ -state machines. The minimization of (5) is well defined since the set  $G_S$  is finite. The initial state  $s_1$  can be chosen arbitrarily since the search over  $G_S$  allows state permutations. The optimal  $g$  will depend, of course, on the sequence  $x_1^n$ .

Define the *asymptotic S-state predictability* of the infinite sequence  $\mathbf{x} = x_1, x_2, \dots$  as

$$\pi_S(\mathbf{x}) = \limsup_{n \rightarrow \infty} \pi_S(x_1^n), \quad (6)$$

and finally, define the *FS predictability* as

$$\pi(\mathbf{x}) = \lim_{S \rightarrow \infty} \pi_S(\mathbf{x}) = \lim_{S \rightarrow \infty} \limsup_{n \rightarrow \infty} \pi_S(x_1^n), \quad (7)$$

where the limit as  $S \rightarrow \infty$  always exists since the minimum fraction of errors, for each  $n$  and thus for its limit supremum, is monotonically nonincreasing with  $S$ . The definitions (5)–(7) are analogous to these associated with the FS compressibility, [1], [2], and [16].

Observe that, by definition,  $\pi(\mathbf{x})$  is attained by a sequence of FSM's that depends on the particular sequence  $\mathbf{x}$ . In what follows, however, we will present sequential prediction schemes that are universal in the sense of being independent of  $\mathbf{x}$  and yet asymptotically achieving  $\pi(\mathbf{x})$ .

### III. S-STATE UNIVERSAL SEQUENTIAL PREDICTORS

We begin with the case  $S = 1$ , i.e., single-state machines. From (3), the optimal single-state predictor employs counts  $N_n(0)$  and  $N_n(1)$  of zeros and ones, respectively, along the entire sequence  $x_1^n$ . It constantly predicts “0” if  $N_n(0) > N_n(1)$ , and “1”, otherwise. The fraction of errors made by this scheme is  $\pi_1(x_1^n) = n^{-1} \min \{N_n(0), N_n(1)\}$ . In this section, we first discuss how to achieve *sequentially*  $\pi_1(x_1^n)$  and later on extend the result to general  $S$ -state machines.

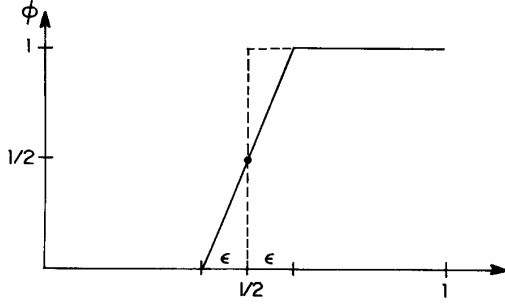
Consider the following simple prediction procedure. At each time instant  $t$ , update the counts  $N_t(0)$  and  $N_t(1)$  of zeros and ones observed so far in  $x_1^t$ . Choose a small  $\epsilon > 0$ , and let  $\hat{p}_t(x) = (N_t(x) + 1)/(t + 2)$ ,  $x = 0, 1$ , be the (biased) current empirical probability of  $x$ . Consider the symbol  $x$  with the larger count, i.e.,  $\hat{p}_t(x) \geq 1/2$ . If in addition  $\hat{p}_t(x) \geq 1/2 + \epsilon$ , guess that the next outcome will be  $x$ . If  $\hat{p}_t(x) \leq 1/2 + \epsilon$ , i.e., the counts are almost balanced, use a randomized rule for which the probability that the next outcome is  $x$  continuously decreases to  $1/2$ , as  $\hat{p}_t(x)$  approaches  $1/2$ . Specifically, the prediction rule is

$$\hat{x}_{t+1} = \begin{cases} \text{“0”}, & \text{with probability } \phi(\hat{p}_t(0)), \\ \text{“1”}, & \text{with probability } \phi(\hat{p}_t(1)) = 1 - \phi(\hat{p}_t(0)), \end{cases} \quad (8)$$

where  $\phi(\cdot)$  is given by

$$\phi(\alpha) = \begin{cases} 0, & 0 \leq \alpha < \frac{1}{2} - \epsilon, \\ \frac{1}{2\epsilon} \left[ \alpha - \frac{1}{2} \right] + \frac{1}{2}, & \frac{1}{2} - \epsilon \leq \alpha \leq \frac{1}{2} + \epsilon, \\ 1, & \frac{1}{2} + \epsilon < \alpha \leq 1, \end{cases} \quad (9)$$

and depicted in Fig. 1. In general, we allow an  $\epsilon = \epsilon_t$  that is varying with  $t$  in a fashion discussed below. We denote by  $\hat{\pi}_1(x_1^n)$  the expected fraction of errors made by this scheme over the sequence  $x_1^n$  where the expectation is with respect to the randomization in (8). The following theorem establishes the fact that  $\hat{\pi}_1(x_1^n)$  approaches  $\pi_1(x_1^n)$ , universally for all sequences.

Fig. 1. The function  $\phi(\alpha)$ .

**Theorem 1:** For any sequence  $x_1^n \in \{0, 1\}^n$ , and a fixed  $\epsilon > 0$  in (9)

$$\hat{\pi}_1(x_1^n) \leq \pi_1(x_1^n) + \frac{\epsilon}{1 - 2\epsilon} + \gamma_1(n, \epsilon), \quad (10)$$

where  $\gamma_1(n, \epsilon) = O(\log n/n)$ . Furthermore, for  $\epsilon = \epsilon_t = 1/(2\sqrt{t} + 2)$  in (9),

$$\hat{\pi}_1(x_1^n) \leq \pi_1(x_1^n) + \delta_1(n), \quad (11)$$

where  $\delta_1(n) = O(1/\sqrt{n})$ .

*Proof:* First observe that  $\pi_1(x_1^n)$  depends solely on the composition  $\{N_n(0), N_n(1)\}$ . We show in Appendix A that among all sequences of the same composition, and thus the same single-state predictability, the sequence for which the predictor (8) performs worst is

$$\tilde{x}_1^n = \frac{2N_n(1)}{0101 \cdots 01} \frac{N_n(0) - N_n(1)}{00 \cdots 00}, \quad (12)$$

where it is assumed, without loss of generality, that  $N_n(0) \geq N_n(1)$ . Clearly, the fraction of errors made by the predictor of (8) over  $\tilde{x}_1^n$  provides a uniform upper bound for  $\hat{\pi}_1(x_1^n)$ . In Appendix A, we also evaluate this average fraction of errors and find that for a fixed  $\epsilon$ ,

$$\hat{\pi}_1(\tilde{x}_1^n) \leq \frac{N_n(1)}{n} + \frac{\epsilon}{1 - 2\epsilon} + \frac{1}{8\epsilon} \frac{\ln(n+1)}{n} + \frac{1}{n} \cdot \frac{1 + 2\epsilon}{1 - 2\epsilon}, \quad (13)$$

while for  $\epsilon_t = 1/2\sqrt{t} + 2$

$$\hat{\pi}_1(\tilde{x}_1^n) \leq \frac{N_n(1)}{n} + \frac{\sqrt{n+1}}{n} + \frac{1}{2n}. \quad (14)$$

Denote

$$\gamma_1(n, \epsilon) \triangleq \frac{1}{8\epsilon} \cdot \frac{\ln(n+1)}{n} + \frac{1}{n} \cdot \frac{1 + 2\epsilon}{1 - 2\epsilon} = O\left(\frac{\log n}{n}\right), \quad (15)$$

and

$$\delta_1(n) \triangleq \frac{\sqrt{n+1}}{n} + \frac{1}{2n} = O\left(\frac{1}{\sqrt{n}}\right). \quad (16)$$

Since  $\hat{\pi}_1(x_1^n) \leq \hat{\pi}_1(\tilde{x}_1^n)$  and since we have assumed that  $N_n(1) \leq N_n(0)$ , then  $N_n(1)/n = \pi_1(\tilde{x}_1^n) = \pi_1(x_1^n)$  and the theorem follows.  $\square$

Several remarks are in order.

1) A natural choice of  $\phi(\cdot)$  could have been

$$\phi(\alpha) = \begin{cases} 0, & \alpha < \frac{1}{2}, \\ \frac{1}{2}, & \alpha = \frac{1}{2}, \\ 1, & \alpha > \frac{1}{2}. \end{cases} \quad (17)$$

However, this choice might be problematic for some sequences. For example, consider the sequence  $x_1^n = 0101 \cdots 01$ . While  $\pi_1(0101 \cdots) = 1/2$ , a predictor based on (17) makes errors 75% of the time on the average. The reason for this gap lies in the fact that  $\hat{p}_t(0)$ , in this example, converges to  $1/2$  which is a discontinuity point of (17). Thus, continuity of  $\phi(\cdot)$  is essential. Note that when  $\epsilon = \epsilon_t$  vanishes,  $\phi(\cdot) = \phi_t(\cdot)$  tends to a discontinuous function. Nevertheless, as discussed in Appendix A,  $\hat{\pi}_1(x_1^n)$  can be universally bounded in terms of  $\pi(x_1^n)$  provided that  $\epsilon_t$  does not go to zero faster than  $O(1/t)$ .

2) A sequential universal prediction scheme, referred to as Blackwell's procedure, has already been proposed [3]–[5], and shown to achieve the single-state predictability (or Bayes envelope in the terminology of [3]–[5]). Denote by  $\hat{\pi}^B(x_1^t)$  the fraction of errors made by this procedure over  $x_1^t$ . Blackwell's prediction rule at time  $t$  is determined by both the current fractions of zeros,  $\hat{p}_t(0)$ , and the current expected fraction of errors,  $\hat{\pi}_1^B(x_1^t)$ . It satisfies (see [3] and [5])

$$\hat{\pi}_1^B(x_1^n) - \pi_1(x_1^n) \leq \frac{3}{\sqrt{n}}, \quad \forall x_1^n \in \{0, 1\}^n. \quad (18)$$

The Blackwell predictor and its properties have been obtained using the theory developed in [6]–[9]. From Theorem 1 the performance of the predictor (8), in the case where  $\epsilon_t = O(1/\sqrt{t})$ , is equivalent to the performance of Blackwell's predictor—both converge to the predictability as  $O(1/\sqrt{n})$ , although the upper bound (14) on the performance of the predictor (8) exhibits a better coefficient of the  $1/\sqrt{n}$  term. Thus, Theorem 1 provides a less general derivation of the previous results, valid in the prediction problem, at the benefit of a conceptually simpler approach.

3) As observed in [4] and emphasized again here, a sequential predictor must be randomized for its performance to approach optimality, universally for all sequences. It was also proved there that the fastest rate to approach the predictability is  $O(1/\sqrt{t})$ . The bound in (14) which will be used throughout the rest of the paper, corresponds to this optimal rate and it is indeed better than (13). The result (13) is still interesting since it corresponds, for a fixed  $\epsilon$ , to a continuous function  $\phi(\cdot)$  and, in accordance with the more general results of [10], it shows a faster convergence rate,  $O(\log n/n)$ , but to a value slightly larger than the predictability.

- 4) It is also verified in Appendix A that the best sequence among all sequences of a given composition  $\{N_n(0), N_n(1)\}$ , in the sense of the smallest expected fraction of errors, has the form

$$\frac{N_n(0)}{000 \cdots 00} \frac{N_n(1)}{11 \cdots 1}. \quad (19)$$

The average number of errors made by the predictor of (8) over this sequence is at least  $N_n(1)$ . Combining this fact with (14), we conclude that for every  $x_1^n$

$$0 \leq \hat{\pi}_1(x_1^n) - \pi_1(x_1^n) \leq O\left(\frac{1}{\sqrt{n}}\right). \quad (20)$$

Thus, although both  $\pi_1(x_1^n)$  and  $\hat{\pi}_1(x_1^n)$  may not converge in general, their difference always converges to zero.

- 5) While Theorem 1 expresses the performance of the single-state sequential predictor in terms of its *expected* relative frequency of errors, it is easy to strengthen this theorem and to obtain an almost-sure result. Specifically, using the Borel–Cantelli lemma, one can prove that

$$\Pr \left\{ \lim_{n \rightarrow \infty} [\hat{\pi}_1(x_1^n) - \pi_1(x_1^n)] = 0 \right\} = 1. \quad (21)$$

The same comment holds throughout this paper.

We next describe a sequential predictor that achieves the performance  $\pi(g; x_1^n)$  for a given next state function  $g$ . Such a predictor has already been described in [5] for the case where  $g$  is Markovian, and it will be rederived here by a simple application of Theorem 1 and Jensen’s inequality. It follows from the observation that for each state  $s$  the optimal prediction rule  $\hat{x}_{t+1} = f(s)$  is fixed and so we can extend Theorem 1 straightforwardly by considering  $S$  sequential predictors of the form (8).

Specifically, let  $N_t(s, x)$ ,  $s \in \mathcal{S}$ ,  $x \in \{0, 1\}$ , denote the joint count of  $s$  and  $x$  along the sequence  $x_1^t$  and the corresponding state sequence  $s_1^t = s_1, \dots, s_t$  generated by  $g$ . Let  $\hat{p}_t(x|s) = (N_t(s, x) + 1)/(N_t(s) + 2)$ ,  $x = 0, 1$ , where  $N_t(s) = N_t(s, 0) + N_t(s, 1)$  is the number of occurrences of the state  $s$  along  $s_1^t$ . Consider the predictor,

$$\hat{x}_{t+1} = f(s_t) = \begin{cases} \text{“0”}, & \text{with probability } \phi(\hat{p}_t(0|s_t)), \\ \text{“1”}, & \text{with probability } \phi(\hat{p}_t(1|s_t)), \end{cases} \quad (22)$$

where the state sequence is generated by  $s_{t+1} = g(x_t, s_t)$  for the given  $g \in G_S$  and  $\phi(\cdot)$  is as in (9) with  $\epsilon_{N_t(s_t)}$ . Let  $\hat{\pi}(g; x_1^n)$  be the fraction of errors of the predictor (22). Now, decompose the sequence  $x_1^n$  into  $S$  subsequences  $x^n(s)$  of length  $N_n(s)$  according to the time instants where each state  $s = 1, \dots, S$  occurred, i.e.,  $x^n(s) = \{x_t, t: s_t =$

$s\}$ . Applying Theorem 1 to each  $x^n(s)$ , we find that

$$\begin{aligned} \hat{\pi}(g; x_1^n) &\leq \frac{1}{n} \sum_{s=1}^S [\min\{N_n(s, 0), N_n(s, 1)\} \\ &\quad + N_n(s) \cdot \delta_1(N_n(s))] \\ &= \pi(g; x_1^n) + \sum_{s=1}^S \frac{N_n(s)}{n} \cdot \delta_1(N_n(s)). \end{aligned} \quad (23)$$

By Jensen’s inequality and the concavity of the square root function,

$$\sum_{s=1}^S \frac{N_n(s)}{n} \cdot \delta_1(N_n(s)) \leq \frac{S}{n} \sqrt{\left(\frac{n}{S} + 1\right)} + \frac{1}{2n} \triangleq \delta_S(n). \quad (24)$$

Thus,  $\hat{\pi}(g; x_1^n)$  approaches  $\pi(g; x_1^n)$  at least as fast as  $O(S/n \cdot \sqrt{n/S}) = O(\sqrt{S/n})$ .

Next, we show how to achieve sequentially the  $S$ -state predictability for a predefined  $S$ . In general, the  $S$ -state predictability requires an optimization with respect to all  $g \in G_S$ . This optimization is bypassed, at a price of increased complexity as presented next.

Let us first define a *refinement* of an FS machine. Given an  $S$ -state machine characterized by a next-state function  $g$ , a refinement of  $g$  is a machine with  $\tilde{S} > S$  states characterized by  $\tilde{g}$ , such that at each time instant  $s_t = h(\tilde{s}_t)$  where  $s_t$  and  $\tilde{s}_t$  are the states at time  $t$  generated by  $g$  and  $\tilde{g}$ , respectively. Clearly, any two time instants corresponding to the same state  $\tilde{s}$  in the refined machine  $\tilde{g}$  also correspond to the same state  $s$  in the machine  $g$ . Thus,

$$\begin{aligned} \pi(g; x_1^n) &= \frac{1}{n} \sum_{s=1}^S \min\{N_n(s, 0), N_n(s, 1)\} \\ &= \frac{1}{n} \sum_{s=1}^S \min\left\{ \sum_{\tilde{s}: h(\tilde{s})=s} N_n(\tilde{s}, 0), \sum_{\tilde{s}: h(\tilde{s})=s} N_n(\tilde{s}, 1) \right\} \\ &\geq \frac{1}{n} \sum_{s=1}^S \sum_{\tilde{s}: h(\tilde{s})=s} \min\{N_n(\tilde{s}, 0), N_n(\tilde{s}, 1)\} \\ &= \frac{1}{n} \sum_{\tilde{s}=1}^{\tilde{S}} \min\{N_n(\tilde{s}, 0), N_n(\tilde{s}, 1)\} = \pi(\tilde{g}; x_1^n), \end{aligned} \quad (25)$$

i.e., refinement improves performance. Furthermore, combining (23) and (25) it follows that the sequential scheme attaining  $\pi(\tilde{g}; x_1^n)$  also attains  $\pi(g; x_1^n)$ , albeit at a slightly slower rate  $O(\sqrt{\tilde{S}/n})$  due to the effort to achieve the predictability of a machine with a larger number of states.

Consider now a refinement  $\tilde{g}$  of all  $M = S^{2S}$  possible  $S$ -state machines. The state  $\tilde{s}_t$  of  $\tilde{g}$ , at time  $t$ , is the vector  $(s_t^1, s_t^2, \dots, s_t^M)$ , where  $s_t^i$ ,  $i = 1, \dots, M$ , is the state at time  $t$  associated with the  $i$ th  $S$ -state machine  $g_i$ . Following the above discussion, it is clear that  $\pi(\tilde{g}; x_1^n) \leq \pi(g; x_1^n)$  for all  $g \in G_S$  and so  $\pi(\tilde{g}; x_1^n) \leq \pi_S(x_1^n)$ . Thus, the sequential scheme (22) based on  $\tilde{g}$  asymptotically attains  $\pi_S(x_1^n)$ . This prohibitively complex scheme achieves the predictability at a very slow rate, and it only achieves the  $S$ -state predictability for a prescribed  $S$ . This scheme only serves as a simple proof for the existence of universal schemes attaining

$\pi_S(x_1^n)$ . Later on we present much more efficient schemes that achieve the performance of any FS predictor, without even requiring an advance specification of  $S$ .

#### IV. MARKOV PREDICTORS

An important subclass of FS predictors is the class of Markov predictors. A Markov predictor of order  $k$  is an FS predictor with  $2^k$  states where  $s_t = (x_{t-1}, \dots, x_{t-k})$ . Similarly to (5), define the  $k$ th-order Markov predictability of the finite sequence  $x_1^n$  as

$$\mu_k(x_1^n) = \frac{1}{n} \sum_{x^k \in \{0,1\}^k} \min \{N_n(x^k, 0), N_n(x^k, 1)\}, \quad (26)$$

where  $N_n(x^k, x) = N_n(x^{k+1})$ ,  $x = 0, 1$ , is the number of times the symbol  $x$  follows the binary string  $x^k$  in  $x_1^n$ , and where for the initial Markov state we use the cyclic convention  $x_{-i} = x_{n-i}$ ,  $i = 1, \dots, k$ . (The choice of initial state does not affect the asymptotic value of  $\mu_k(x_1^n)$ . The cyclic convention is used for reasons that will be clarified later on.)

The asymptotic  $k$ th-order Markov predictability of the infinite sequence  $x$  is defined as

$$\mu_k(x) = \limsup_{n \rightarrow \infty} \mu_k(x_1^n), \quad (27)$$

and finally the Markov predictability of the sequence  $x$  is defined as

$$\mu(x) = \lim_{k \rightarrow \infty} \mu_k(x) = \lim_{k \rightarrow \infty} \limsup_{n \rightarrow \infty} \mu_k(x_1^n), \quad (28)$$

where the limit for  $k$  exists since a  $(k+1)$ st-order Markov predictor is a refinement of a  $k$ th-order predictor and so  $\mu_k(x)$  monotonically decreases with  $k$ .

We next prove that the Markov predictability and the FS predictability are equivalent. Thus, any scheme which attains  $\mu(x)$  also achieves  $\pi(x)$ . Observe first, that since the class of FSM's contains the subclass of Markov machines it is obvious that for any finite sequence  $x_1^n$  and  $S = 2^k$ ,

$$\mu_k(x_1^n) \geq \pi_S(x_1^n), \quad (29)$$

and therefore,  $\mu(x) \geq \pi(x)$ . The following theorem established a converse inequality.

**Theorem 2:** For all integers  $k \geq 0$ ,  $S \geq 1$  and for any finite sequence  $x_1^n \in \{0, 1\}^n$ ,

$$\mu_k(x_1^n) \leq \pi_S(x_1^n) + \sqrt{\frac{\ln S}{2(k+1)}}. \quad (30)$$

Note that Theorem 2 holds for any arbitrary integers  $k$  and  $S$ , and it becomes meaningful when  $2^k \gg S$  in contrast to (29) in which  $S = 2^k$ .

*Proof:* The idea in the proof is to consider a predictor which is a refinement of both the Markov machine and a given  $S$ -state machine. This refined predictor performs better than both machines. We will show, however, that when the Markov order  $k$  is large (relative to  $\ln S$ ) the performance of this refined machine with  $2^k \times S$  states is not *much better* than that of the Markov machine with  $2^k$  states. *A-fortiori*,

the  $S$ -state machine cannot perform much better than the Markov machine.

Let  $s_t$  be the state at time  $t$  of the  $S$ -state machine  $g$  and consider a machine  $\tilde{g}_j$  whose state at time  $t$  is  $\tilde{s}_t = (s_{t-j}, x_{t-j}, \dots, x_{t-1})$ . Clearly, for every positive integer  $j$ ,  $\tilde{g}_j$  is a refinement of  $g$ . As a result  $\pi(\tilde{g}_j; x_1^n) \leq \pi(g; x_1^n)$ , and so

$$\mu_j(x_1^n) - \pi(g; x_1^n) \leq \mu_j(x_1^n) - \pi(\tilde{g}_j; x_1^n). \quad (31)$$

In the following lemma, we upper bound  $\mu_j(x_1^n) - \pi(\tilde{g}_j; x_1^n)$  in terms of the difference between the respective empirical entropies<sup>1</sup>

$$\hat{H}(X | X^j) = - \sum_{x^j \in \{0,1\}^j} \frac{N_n(x^j)}{n} \sum_{x=0,1} \frac{N_n(x^j, x)}{N_n(x^j)} \cdot \log \frac{N_n(x^j, x)}{N_n(x^j)} \quad (32)$$

corresponding to the  $j$ th-order Markov machine, and

$$\hat{H}(X | X^j, \mathcal{S}) = - \sum_{\tilde{s} \in \tilde{S}^j} \frac{N_n(\tilde{s})}{n} \sum_{x=0,1} \frac{N_n(\tilde{s}, x)}{N_n(\tilde{s})} \cdot \log \frac{N_n(\tilde{s}, x)}{N_n(\tilde{s})}, \quad (33)$$

where  $\tilde{S}^j = \{1, \dots, S\} \times \{0, 1\}^j$ , corresponding to the refined machine  $\tilde{g}_j$ , and here  $\mathcal{S}$  denotes 2 random variable whose sample space is  $\{1, \dots, S\}$ .

**Lemma 1:** For every integer  $j \geq 0$ , and every next-state function  $g \in G_S$ ,  $S \geq 1$ ,

$$\begin{aligned} & \mu_j(x_1^n) - \pi(\tilde{g}_j; x_1^n) \\ & \leq \sqrt{\frac{\ln 2}{2} [\hat{H}(X | X^j) - \hat{H}(X | X^j, \mathcal{S})]}. \end{aligned} \quad (34)$$

Lemma 1 is proved in Appendix B.

Now, since  $\mu_k(x_1^n) \leq \mu_j(x_1^n)$  for all  $j \leq k$ .

$$\begin{aligned} & \mu_k(x_1^n) - \pi(g; x_1^n) \\ & \leq \frac{1}{k+1} \sum_{j=0}^k [\mu_j(x_1^n) - \pi(g; x_1^n)] \\ & \leq \frac{1}{k+1} \sum_{j=0}^k [\mu_j(x_1^n) - \pi(\tilde{g}_j; x_1^n)] \\ & \leq \frac{1}{k+1} \sum_{j=0}^k \sqrt{\frac{\ln 2}{2} [\hat{H}(X | X^j) - \hat{H}(X | X^j, \mathcal{S})]} \\ & \leq \sqrt{\frac{\ln 2}{2} \left( \frac{1}{k+1} \sum_{j=0}^k [\hat{H}(X | X^j) - \hat{H}(X | X^j, \mathcal{S})] \right)}, \end{aligned} \quad (35)$$

where the second inequality follows from (31), the third inequality follows from Lemma 1 and the last inequality

<sup>1</sup> Throughout this paper,  $\log x = \log_2 x$  while  $\ln x = \log_e x$ .

follows from Jensen's inequality and the concavity of the square root function. By the chain rule of conditional entropies,

$$\sum_{j=0}^k \hat{H}(X | X^j) = \hat{H}(X^k, X) = \hat{H}(X^{k+1}), \quad (36)$$

$$\sum_{j=0}^k \hat{H}(X | X^j, \mathcal{S}) = \hat{H}(X, X^k | \mathcal{S}) = \hat{H}(X^{k+1} | \mathcal{S}). \quad (37)$$

The chain rule applies since the empirical counts are computed using the cyclic convention, resulting in a *shift invariant* empirical measure in the sense that the  $j$ th order marginal empirical measure derived from the  $k$ th order empirical measure ( $j \leq k$ ) is independent of the position of the  $j$ -tuple in the  $k$ -tuple. Now observe that

$$\begin{aligned} \hat{H}(X^{k+1}) - \hat{H}(X^{k+1} | \mathcal{S}) &= \hat{H}(X^{k+1}) \\ &- \hat{H}(X^{k+1}, \mathcal{S}) + \hat{H}(\mathcal{S}) \leq \hat{H}(\mathcal{S}) \leq \log S. \end{aligned} \quad (38)$$

Combining (35)–(38),

$$\mu_k(x_1^n) \leq \pi(g; x_1^n) + \sqrt{\frac{\ln S}{2(k+1)}}. \quad (39)$$

Since  $g \in G_S$  is arbitrary, the proof is complete.  $\square$

Having proved (30), one can take the limit supremum as  $n \rightarrow \infty$ , then the limit  $k \rightarrow \infty$  and finally the limit  $S \rightarrow \infty$  and obtain  $\mu(\mathbf{x}) \leq \pi(\mathbf{x})$ , which together with the obvious relation  $\mu(\mathbf{x}) \geq \pi(\mathbf{x})$  leads to

$$\mu(\mathbf{x}) = \pi(\mathbf{x}). \quad (40)$$

The fact that Markov machines perform asymptotically, as well as any FSM is not unique to the prediction problem. In particular, consider the data compression and the gambling problems where  $\hat{H}(X | X^k)$  and  $\hat{H}(X | \mathcal{S})$  quantify the performance of the  $k$ th order Markov machine and an FS machine with  $S$  states, respectively (see [2], [17], and [18]). Clearly,

$$\begin{aligned} \hat{H}(X | X^j) - \hat{H}(X | \mathcal{S}) \\ \leq \hat{H}(X | X^j) - \hat{H}(X | X^j, \mathcal{S}). \end{aligned} \quad (41)$$

Using (41) for all  $j \leq k$  and following the same steps as in (35)–(38),

$$\hat{H}(X | X^k) \leq \hat{H}(X | \mathcal{S}) + \frac{\log S}{k+1}. \quad (42)$$

This technique is further exercised in [10] to obtain similar relations between the performances of Markov machines and FS machines for a broad class of sequential decision problems.

Next we demonstrate a sequential universal scheme that attains  $\mu(\mathbf{x})$  and thus,  $\pi(\mathbf{x})$ . First observe that from the discussion in Section III, for a fixed  $k$ , the  $k$ th order Markov predictability can be achieved asymptotically by the predictor

(22) with  $s_t = (x_{t-k+1}, \dots, x_t)$ , i.e.,

$$\hat{x}_{t+1} = \begin{cases} \text{“0”}, & \text{with probability } \phi(\hat{p}_t(0 | x_t, \dots, x_{t-k+1})), \\ \text{“1”}, & \text{with probability } \phi(\hat{p}_t(1 | x_t, \dots, x_{t-k+1})), \end{cases} \quad (43)$$

where, e.g.,

$$\hat{p}_t(0 | x_t, \dots, x_{t-k+1}) = \frac{N_t(x_{t-k+1} \dots x_t, 0) + 1}{N_t(x_{t-k+1} \dots x_t) + 2},$$

and  $\phi(\cdot)$  is determined with  $\epsilon_{N_t(x_{t-k+1} \dots x_t)}$ .

To attain  $\mu(\mathbf{x})$ , the order  $k$  must grow as more data is available. Otherwise, if the order is at most  $k^*$ , the scheme may not outperform a Markov predictor of order  $k > k^*$ . Increasing the number of states corresponding to increasing the number of separate counters for  $N_t(x^k, x)$ . There are two conflicting goals: On one hand, one wants to increase the order rapidly so that a high-order Markov predictability is reached as soon as possible. On the other hand, one has to increase the order slowly enough to assure that there are enough counts in each state for a reliable estimate of  $\hat{p}_t(x | x_t, \dots, x_{t-k+1})$ . As will be seen, the order  $k$  must grow not faster than  $O(\log t)$  to satisfy both requirements.

More precisely, denote by  $\hat{\mu}_k(x_1^n)$  the expected fraction of errors of the predictor (43). Following (23) and (24),

$$\hat{\mu}_k(x_1^n) \leq \mu_k(x_1^n) + \delta_{2^k}(n), \quad (44)$$

where  $\delta_{2^k}(n) = O(\sqrt{2^k/n})$ . Suppose now that the observed data is divided into nonoverlapping segments,  $\mathbf{x} = \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$  and apply the  $k$ th order sequential predictor (43) to the  $k$ th segment,  $\mathbf{x}^{(k)}$ . Choose a sequence  $\alpha_k$  such that  $\alpha_k \rightarrow \infty$  monotonically as  $k \rightarrow \infty$ , and let the length of the  $k$ th segment, denoted  $n_k$ , be at least  $\alpha_k \cdot 2^k$ . By (44) and (24),

$$\begin{aligned} \hat{\mu}_k(\mathbf{x}^{(k)}) &\leq \mu_k(\mathbf{x}^{(k)}) + \delta_{2^k}(n_k) \leq \mu_k(\mathbf{x}^{(k)}) \\ &+ \frac{\sqrt{\alpha_k + 1}}{\alpha_k} + \frac{1}{2\alpha_k \cdot 2^k} = \mu_k(\mathbf{x}^{(k)}) + \xi(k), \end{aligned} \quad (45)$$

where  $\xi(k) = O(1/\sqrt{\alpha_k})$  and so  $\xi(k) \rightarrow 0$  as  $k \rightarrow \infty$ . Thus, in each segment the Markov predictability of the respective order is attained as  $k$  increases, in a rate that depends on the choice of  $\alpha_k$ .

Consider now a finite, arbitrarily long sequence  $x_1^n$ , where  $n = \sum_{k=1}^{k_n} n_k$  and  $k_n$  is the number of segments in  $x_1^n$ . The average fraction of errors made by the above predictor, denoted  $\hat{\mu}(x_1^n)$ , satisfies,

$$\begin{aligned} \hat{\mu}(x_1^n) &= \sum_{k=1}^{k_n} \frac{n_k}{n} \hat{\mu}_k(\mathbf{x}^{(k)}) \\ &\leq \sum_{k=1}^{k_n} \frac{n_k}{n} \mu_k(\mathbf{x}^{(k)}) + \sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k). \end{aligned} \quad (46)$$

Now, for any fixed  $k' < k_n$ ,

$$\begin{aligned} \hat{\mu}(x_1^n) &\leq \sum_{k=1}^{k'-1} \frac{n_k}{n} \mu_k(x^{(k)}) + \sum_{k=k'}^{k_n} \frac{n_k}{n} \mu_k(x^{(k)}) \\ &+ \sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k) \leq \frac{1}{2} \sum_{k=1}^{k'-1} \frac{n_k}{n} + \sum_{k=1}^{k_n} \frac{n_k}{n} \mu_{k'}(x^{(k)}) \\ &+ \sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k), \end{aligned} \quad (47)$$

where (47) holds since always  $\mu_i(x^{(k)}) \leq 1/2$ , since  $\mu_i(x^{(k)}) \leq \mu_j(x^{(k)})$  when  $i > j$ , and since adding positive terms only increases the right-hand side (RHS) of (47). Now, the term  $\sum_{k=1}^{k_n} (n_k/n) \mu_{k'}(x^{(k)})$  is the fraction of errors made in predicting  $x_1^n$  by a machine whose state is determined by the  $k'$ th Markov state and the current segment number. This is a refinement of the  $k'$ th-order Markov predictor. Thus,

$$\sum_{k=1}^{k_n} \frac{n_k}{n} \mu_{k'}(x^{(k)}) \leq \mu_{k'}(x_1^n). \quad (48)$$

Also since the  $\xi(k)$  is monotonically decreasing and since the length of each segment is monotonically increasing we can write

$$\sum_{k=1}^{k_n} \frac{n_k}{n} \xi(k) \leq \frac{1}{k_n} \sum_{k=1}^{k_n} \xi(k) \triangleq \bar{\xi}(k_n), \quad (49)$$

where by the Cesaro theorem  $\bar{\xi}(k_n) \rightarrow 0$  as  $k_n \rightarrow \infty$ . Thus, we can summarize the result of this section in the following theorem.

**Theorem 3:** For any finite  $k$ , any finite  $S$ , and for any  $x_1^n \in \{0, 1\}^n$

$$\hat{\mu}(x_1^n) \leq \mu_k(x_1^n) + \bar{\xi}(k_n) \leq \pi_S(x_1^n) + \xi^*(n), \quad (50)$$

where both  $\bar{\xi}(k_n) \rightarrow 0$  and  $\xi^*(n) \rightarrow 0$  as  $n \rightarrow \infty$ .

*Proof:* The first inequality is achieved by combining (47)–(49), taking the limit supremum, and observing that  $k_n \rightarrow \infty$  as  $n \rightarrow \infty$ . For the second inequality we use (30) where we define  $\xi^*(n) = \bar{\xi}(k_n) + \sqrt{(\ln S)/2(k_n + 1)}$ .  $\square$

Note that this theorem implies that for any finite individual sequence

$$\hat{\mu}(x) \triangleq \limsup_{n \rightarrow \infty} \hat{\mu}(x_1^n) = \mu(x) = \pi(x).$$

In summary, then, we have shown that a sequential Markov predictor whose order is incremented from  $k$  to  $k + 1$  after observing at least  $n_k = \alpha_k \cdot 2^k$  data samples (i.e., a predictor whose order grows as  $O(\log t)$ ) achieves, within  $\xi^*(n)$ , the performance of any finite-state predictor.

## V. PREDICTION USING INCREMENTAL PARSING

In this section, we present a sequential predictor based on the incremental parsing algorithm, suggested by Lempel and Ziv [1], and show that it attains the FS predictability. The underlying idea is that the incremental parsing algorithm induces another technique for gradually changing the Markov order with time at an appropriate rate.

The LZ parsing algorithm parses an outcome sequence into distinct phrases such that each phrase is the shortest string which is not a previously parsed phrase. For example, the sequence 001010100... is parsed into  $\{0, 01, 010, 1, 0100, \dots\}$ . It is convenient to consider this procedure as a process of growing a tree, where each new phrase is represented by a leaf in the tree. The initial tree for binary sequences consists of a root and two leaves, corresponding to the phrases  $\{0, 1\}$ , respectively. At each step, the current tree is used to create an additional phrase by following the path (from the root to a leaf) that corresponds to the incoming symbols. Once a leaf has been reached, the tree is extended at that point, making the leaf an internal node, and adding its two offsprings to the tree. The process of growing a binary tree for the above example is shown in Fig. 2. The set of phrases which correspond to the leaves of the tree is called a dictionary. Note that in the process of parsing the sequence, each outcome  $x_t$  is associated with a node reached by the path corresponding to the string starting at the beginning of the phrase and ending at  $x_t$ .

Let  $K_j$  be the number of leaves in the  $j$ th step (note that  $K_j = j + 1$  for binary sequences) and assign a weight  $1/K_j$  to each leaf. This can be thought of as assigning a uniform probability mass function to the leaves. The weight of each internal node is the sum of weights of its two offsprings (see Fig. 2, for an example). Define the conditional probability  $\hat{p}_t^{LZ}(x_{t+1} | x_t^j)$  of a symbol  $x_{t+1}$  given its past as the ratio between the weight of the node corresponding to  $x_{t+1}$  (0 or 1) that follows the current node  $x_t$ , and the weight of the node associated with  $x_t$ . Note that if  $x_{t+1}$  is the first symbol of a new phrase, the node associated with  $x_t$  is the root. This definition of  $\hat{p}_t^{LZ}(x_{t+1} | x_t^j)$  as the conditional probability induced by the incremental parsing algorithm was originally made in [19] and [20].

In [2],  $\hat{p}_t^{LZ}(x_{t+1} | x_t^j)$  has been used for universal gambling where it was suggested to wager on "0" a fraction  $\hat{p}_t^{LZ}(0 | x_t^j)$  of the capital at time  $t$ . Here, we suggest this estimator for sequential prediction, according to

$$\hat{x}_{t+1} = \begin{cases} \text{"0"}, & \text{with probability } \phi_t(\hat{p}_t^{LZ}(0 | x_t^j)), \\ \text{"1"}, & \text{with probability } \phi_t(\hat{p}_t^{LZ}(1 | x_t^j)), \end{cases} \quad (51)$$

where  $\phi_t(\cdot)$  is as in (9) with a time-varying parameter  $\epsilon_t$  to be defined later. This predictor will henceforth be referred to as the incremental parsing (IP) predictor. We prove below that it attains  $\pi(x)$ . For this purpose it is useful to recall the counting interpretation of  $\hat{p}_t^{LZ}(\cdot | \cdot)$ .

In this interpretation, the outcomes are sorted into bins (or "contexts" in the terminology of [19], [20]). Each outcome  $x_t$  is classified into a bin determined by the string  $\nu$  starting at the beginning of the current phrase and ending at  $x_{t-1}$ . The string  $\nu$  will be referred to as the bin label. The first bin, labeled by the empty string, contains all the bits that appear at the beginning of a phrase. In the previous example,  $\{0, 01, 010, 1, 0100, \dots\}$ , the bits 00010... at locations 1, 2, 4, 7, 8, ... are the initial bits of parsed strings and belong to the first bin, the bits 111... at locations 3, 5, 9, ... belong to the bin labeled "0", and so on.

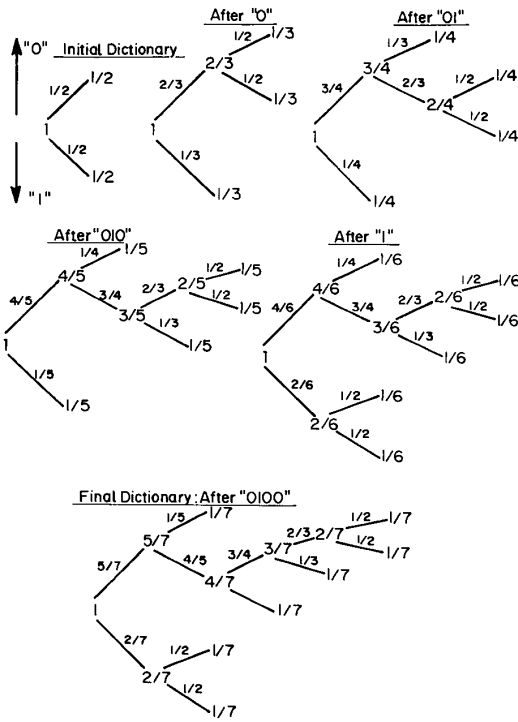


Fig. 2. Dictionary trees and probability estimate induced by the LZ scheme.

The procedure begins with the single bin labeled by the empty string. With each new phrase a new bin, labeled by that phrase, is added. Thus, we observe that the sequence  $x_1^n$  generates  $c + 1$  bins, where  $c = c(x_1^n)$  is the number of parsed strings in  $x_1^n$ . Actually we observe that the sequence is divided into at most  $c$  bins since at least the last bin, labeled by the string just parsed, will be empty.

A sequential probability estimate is defined for each bin as follows. Let  $N_t^j(x)$ ,  $j = 1, \dots, c$ , denote the number of symbols equal to  $x$  in the  $j$ th bin at time  $t$ . The probability estimate of the next bit  $x$  entering the  $j$ th bin is

$$\frac{N_t^j(x) + 1}{N_t^j(0) + N_t^j(1) + 2} = \frac{N_t^j(x) + 1}{N_t^j + 2}, \quad x = 0, 1,$$

where  $N_t^j = N_t^j(0) + N_t^j(1)$ . It turns out, as was previously observed in [19], that this probability estimate, at the current bin, equals to  $\hat{p}_t^{LZ}(x | x_1^t)$ . In the previous example, the sequence of bits in the first bin is 00010...; thus, the respective estimate of the probability that the next bit classified to this bin will be "0" are  $1/2, 2/3, 3/4, 4/5, 4/6, \dots$  which coincide with the corresponding estimates, indicated in Fig. 2, that the bits in locations 1, 2, 4, 7, 8, ... will be "0".

Following this interpretation, we set  $\epsilon_t = 1/(2\sqrt{N_t^j + 2})$  for determining the predictor (51) used in guessing the next bit to enter the  $j$ th bin. Having defined the IP predictor, let  $\pi^{IP}(x_1^n)$  denote its expected fraction of errors. We are now ready to present Theorem 4 which upper bounds  $\pi^{IP}(x_1^n)$ .

**Theorem 4:** For every sequence  $x_1^n$  and any integer  $k \geq 0$ ,

$$\pi^{IP}(x_1^n) \leq \mu_k(x_1^n) + \eta(n, k) \quad (52)$$

where for a fixed  $k$ ,  $\eta(n, k) = O(1/\sqrt{\log n})$ .

*Proof:* Following the counting interpretation, the IP predictor is a set of sequential predictors each operating on a separate bin. Applying Theorem 1 to each one of the  $c$  bins and averaging over the bins similarly to (23), we get

$$\begin{aligned} \pi^{IP}(x_1^n) &\leq \frac{1}{n} \sum_{j=1}^c (\min \{N_n^j(0), N_n^j(1)\} + N_n^j \cdot \delta_1(N_n^j)) \\ &\leq \frac{1}{n} \sum_{j=1}^c \min \{N_n^j(0), N_n^j(1)\} + \delta_c(n), \end{aligned} \quad (53)$$

where the last inequality follows from the convexity of the logarithm and Jensen's inequality, as in (24). Now, for any  $k \geq 0$ , we can write

$$\begin{aligned} \sum_{j=1}^c \min \{N_n^j(0), N_n^j(1)\} &= \sum_{j \in J_1} \min \{N_n^j(0), N_n^j(1)\} \\ &\quad + \sum_{j \in J_2} \min \{N_n^j(0), N_n^j(1)\}, \end{aligned} \quad (54)$$

where  $J_1$  is the set of bins labeled by strings shorter than  $k$ , and  $J_2$  is the set of bins labeled by strings of length  $k$  or longer. Note that the first  $k$  bits in each phrase are allocated to bins in  $J_1$ . Thus, at most  $k \cdot c$  bits are allocated to bins in  $J_1$  and so the first term in the RHS of (54) is upper bounded by  $(k \cdot c)/2$ . As for the second term in the RHS of (54), observe that the  $k$ th-order Markov predictor divides the data into bins labeled by the  $k$  previous bits. Since a bin in  $J_2$  is labeled by at least  $k$  previous bits, the IP predictor serves as a refinement to the  $k$ th-order Markov predictor in  $J_2$ . Thus, the second term is smaller than the number of errors made by the  $k$ th-order Markov predictor over the bits in  $J_2$ , which in turn is smaller than  $n \cdot \mu_k(x_1^n)$ , the number of errors made by the Markov predictor over the entire sequence. Combining these observations we get,

$$\sum_{j=1}^c \min \{N_n^j(0), N_n^j(1)\} \leq \frac{kc}{2} + n \cdot \mu_k(x_1^n). \quad (55)$$

Substituting (55) into (53),

$$\pi^{IP}(x_1^n) \leq \mu_k(x_1^n) + k \cdot \frac{c}{2n} + \delta_c(n). \quad (56)$$

Since  $\delta_c(n) = O(\sqrt{c/n})$  and recalling that  $c/n \leq O(1/(\log n))$  (see [21]) the theorem follows.  $\square$

We have just shown that the IP predictor asymptotically outperforms a Markov predictor of any finite order and hence, by Theorem 2, it also attains the FS predictability. Note, however, that the rate at which the predictability is attained is  $O(1/\sqrt{\log n})$  which is slower than the rate  $O(\sqrt{(2^k)/n})$  of the predictor (43). The reason is that the IP predictor has effectively  $c \approx n/\log n$  states and so its equivalent Markov order is  $\log c \approx \log(n/\log n)$ .

A result concerning the compression performance of Lempel–Ziv algorithm, known as Ziv’s inequality and analogous to the theorem above, states that the compression ratio of the LZ algorithm is upper bounded by the  $k$ th-order empirical entropy plus an  $O((\log \log n)/(\log n))$  term. This result has been originally shown in [22] (see also [23, ch. 12]). It can be proved, more directly, by a technique similar to the above, utilizing (42) and the  $O((2^k/(n) \log(n/2^k))$  convergence rate observed in universal coding schemes for Markov sources, [20], [24]–[26], and the fact that the LZ algorithm has an equivalent order of  $\log c \approx \log(n/\log n)$ .

For each individual sequence, the compression ratio of the LZ algorithm is determined uniquely by the number of parsed strings  $c(x_1^n)$ , which is a relatively easily computable quantity. It is well known [1] that this compression ratio,  $n^{-1}c(x_1^n) \log c(x_1^n)$ , is a good estimator for the compressibility of the sequence (and the entropy of a stationary ergodic source). As will be evident from the discussion in the next section, the predictability of a sequence cannot be uniquely determined by its compressibility and hence neither by  $c(x_1^n)$ . It is thus an interesting open problem to find out an easily calculable estimator for the predictability.

Finally, the IP predictor proposed and analyzed here has been suggested independently in [27] as an algorithm for page prefetching into a cache memory. For this purpose, the algorithm in [27] was suggested and analyzed in a more general setting of nonbinary data, and in the case where one may predict that the next outcome lie in a set of possible values (corresponding to a cache size larger than 1). However, unlike our analysis which holds for any individual sequence, the analysis in [27] was performed under the assumption that the data is generated by a finite state *probabilistic* source.

## VI. PREDICTABILITY AND COMPRESSIBILITY

Intuitively, predictability is related to compressibility in the sense that sequences which are easy to compress seem to be also easy to predict and conversely, incompressible sequences are hard to predict. In this section we try to consolidate this intuition.

The definition of FS predictability is analogous to the definition of the FS compressibility  $\rho(x)$ , see [1], [2], and [16]. Specifically, the FS compressibility (or FS complexity, FS empirical entropy) was defined in [2] as

$$\rho(x) \triangleq \lim_{S \rightarrow \infty} \limsup_{n \rightarrow \infty} \min_{g \in G_S} \rho(g; x_1^n), \quad (57)$$

where

$$\rho(g; x_1^n) = \sum_{s=1}^S \frac{N_n(s)}{n} h\left(\frac{N_n(s, 0)}{N_n(s)}\right), \quad (58)$$

and where  $h(\alpha) = -\alpha \log \alpha - (1 - \alpha) \log(1 - \alpha)$  is the binary entropy function. This quantity represents the optimal data compression performance of any FS machine where the integer codeword length constraint is relaxed (this noninteger codeword length can be nearly obtained using, e.g., arithmetic coding [28]). Also, utilizing the relation between compression and gambling, [17], [18], the quantity  $1 - \rho(x)$  is the

optimal capital growth rate in sequential gambling over the outcome of the sequence using any FS machine.

As was observed in [2], by the concavity of  $h(\cdot)$  and Jensen’s inequality,

$$\begin{aligned} \rho(g; x_1^n) &= \sum_{s=1}^S \frac{N_n(s)}{n} h\left(\frac{N_n(s, \hat{x})}{N_n(s)}\right) \\ &\leq h\left(\sum_{s=1}^S \frac{N_n(s)}{n} \cdot \frac{N_n(s, \hat{x})}{N_n(s)}\right) \\ &= h\left(\sum_{s=1}^S \frac{N_n(s, \hat{x})}{n}\right) = h(\pi(g; x_1^n)), \quad (59) \end{aligned}$$

where  $\hat{x} = \arg \min N_n(s, x)$ . By minimizing over  $g \in G_S$ , taking the limit supremum as  $n \rightarrow \infty$ , and the limit  $S \rightarrow \infty$  for both sides of (59) and by the monotonicity of  $h(\cdot)$  in the domain  $[0, 1/2]$ ,

$$\pi(x) \geq h^{-1}(\rho(x)). \quad (60)$$

An upper bound on the predictability in terms of the compressibility can be derived as well. Since  $h(\alpha) \geq 2\alpha$  for  $0 \leq \alpha \leq 1/2$ ,

$$\begin{aligned} \rho(g; x_1^n) &= \sum_{s=1}^S \frac{N_n(s)}{n} h\left(\frac{N_n(s, \hat{x})}{N_n(s)}\right) \\ &\geq \sum_{s=1}^S \frac{N_n(s)}{n} \cdot 2 \frac{N_n(s, \hat{x})}{N_n(s)} = 2\pi(g; x_1^n) \quad (61) \end{aligned}$$

which leads to

$$\frac{1}{2}\rho(x) \geq \pi(x). \quad (62)$$

Both the upper bound and the lower bound as well as any point in the region in between, can be obtained by some sequence. Thus, the compressibility of the sequence does not determine uniquely its predictability. The achievable region in the  $\rho - \pi$  plane is illustrated in Fig. 3.

The lower bound (60) is achieved whenever  $\hat{p}_n(\hat{x} | s) = N(s, \hat{x})/N_n(s)$  are equal for all  $s$ . This is the case where the FS compressibility of the sequence is equal to the zero-order empirical entropy of the prediction error sequence (i.e., the prediction error sequence is “memoryless”). Only in this case, a predictive encoder based on the optimal FS predictor will perform as well as the optimal FS encoder.

The upper bound (62) is achieved when at some states  $\hat{p}_n(\hat{x} | s) = 0$  and in the remaining states  $\hat{p}_n(\hat{x} | s) = 1/2$ , i.e., in a case where the sequence can be decomposed by an FSM into perfectly predictable and totally unpredictable subsequences.

The upper and lower bounds coincide at  $(\rho = 0, \pi = 0)$  and  $(\rho = 1, \pi = 1/2)$  implying that a sequence is *perfectly predictable iff it is totally redundant*, and conversely, a sequence is *totally unpredictable iff it is incompressible*.

A new complexity measure may be defined based on the notion of predictability. Analogously to the complexity definitions of Solomonoff, [11], Kolmogorov, [12] and Chaitin, [13], we may define the predictive complexity as the minimum fraction of errors made by a universal Turing machine in sequentially predicting the future of the sequence. A point to observe is that while the complexities above are related to the description length (or program length) and hence, to each

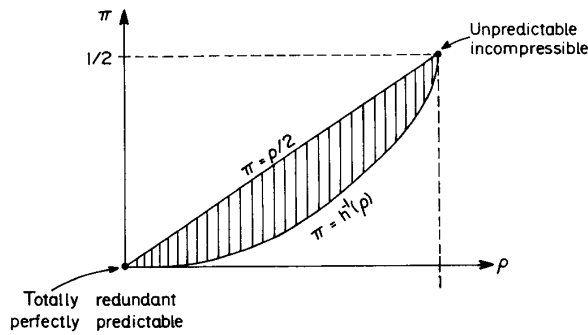


Fig. 3. Compressibility and predictability—achievable region.

[15] to the approach of this paper, regarding the prediction the discussion above, sequences that have the same Kolmogorov's complexity (description length) may have a different predictive complexity, and vice-versa.

A predictability definition can be made for probabilistic sources as well. The predictability of a binary random variable  $X$  will be

$$\pi(X) = E\{I(\Pr(X))\} = \min\{p, 1 - p\}, \quad (63)$$

where  $p$  is the probability that  $X = 0$  and  $I(\alpha) = 1$  for  $\alpha < 1/2$ ,  $I(\alpha) = 1/2$  for  $\alpha = 1/2$ , and  $I(\alpha) = 0$  for  $\alpha > 1/2$ . The conditional predictability of the random variable  $X_1$  given  $X_2$  is defined as

$$\begin{aligned} \pi(X_1 | X_2) &= E\{I(\Pr(X_1 | X_2))\} \\ &= E_{X_2}\{\pi(X_1 | X_2 = x_2)\}. \end{aligned} \quad (64)$$

One immediately observes that conditioning reduces the predictability i.e.,  $\pi(X_1) \geq \pi(X_1 | X_2) \geq \pi(X_1 | X_2, X_3)$  etc. These definitions can be generalized to random vectors, and stochastic processes. For example, the predictability of a stationary ergodic process  $\mathcal{X}$  is defined as

$$\begin{aligned} \pi(\mathcal{X}) &= \lim_{n \rightarrow \infty} \pi(X_{n+1} | X_n, \dots, X_1) \\ &= \pi(X_0 | X_{-1}, \dots). \end{aligned} \quad (65)$$

It will be interesting to further explore the predictability measure and its properties. For example, to establish the predictability as the minimum frequency of errors that can be made by any sequential predictor over the outcome of a general (ergodic) source, and convergence of the performance of prediction schemes to the source's predictability. Another problem of interest is the derivation of a tight lower bound on the rate at which the predictability can be approached asymptotically by a universal predictor for a parametric class of sources. This problem is motivated by an analogous existing result in data compression [14], which states that no lossless code has a compression ratio that approaches the entropy faster than  $(S/2n) \log n$ , where  $S$  is the number of parameters, except for a small subset of the sources corresponding to a small subset of the parameter space. A solution to this problem might follow from [15]. It is interesting, in general, to relate the results and approach of

other, the predictive complexity is a distinct measure. From problem. These issues as well as other topics mentioned above are currently under investigation.

ACKNOWLEDGMENT

The authors acknowledge T. M. Cover for bringing [3]–[9] to their attention and for helpful discussions. Useful suggestions made by J. Ziv and O. Zeitouni are gratefully acknowledged. Interesting discussions with A. Lempel at the early phase of the research are also appreciated.

APPENDIX A

*The Worst Sequence of a Given Composition:* Assume, without loss of generality, that  $N_n(0) \geq N_n(1)$  and construct a state diagram where the state at time  $t$  corresponds to the absolute difference  $C_t = |N_t(0) - N_t(1)|$ . Clearly,  $C_0 = 0$  and  $C_n = N_n(0) - N_n(1)$ . The final state  $C_n$  is the same for all sequences of the same composition and hence, the same single-state predictability  $\pi_1(x_t^n)$ . However, the exact trellis of  $\{C_t\}$  depends on the particular sequence.

Define an *upward* loop in the trellis as a pattern  $(C_t = k, C_{t+1} = k + 1, C_{t+2} = k)$  for some integer  $k > 0$ , and similarly a *downward* loop as  $(C_t = k, C_{t+1} = k - 1, C_{t+2} = k)$ . Replacing an upward loop by a downward loop corresponds to changing "01" to "10" or vice-versa, which does not affect the composition of the sequence, but as we show next, it can only increase the *loss* or the expected number of errors.

Assume first that  $N_t(0) > N_t(1)$ . The loss incurred along the upward loop at time  $t$  is

$$\alpha = I\left(\frac{N_t(0) + 1}{t + 2}\right) + 1 - I\left(\frac{N_t(0) + 2}{t + 3}\right), \quad (A.1)$$

where  $I(\cdot) = 1 - \phi(\cdot)$ . The loss incurred along the downward loop is

$$\beta = 1 - I\left(\frac{N_t(0) + 1}{t + 2}\right) + I\left(\frac{N_t(0) + 1}{t + 3}\right). \quad (A.2)$$

We want to show that  $\alpha \leq \beta$ . We may write

$$\begin{aligned} \alpha - \beta &= \phi_{t+1}\left(\frac{N_t(0) + 2}{t + 3}\right) \\ &\quad + \phi_{t+1}\left(\frac{N_t(0) + 1}{t + 3}\right) - 2\phi_t\left(\frac{N_t(0) + 1}{t + 2}\right), \end{aligned}$$

where  $\phi_t(\cdot)$  denotes the function  $\phi(\cdot)$  of (9) with a possibly time varying  $\epsilon = \epsilon_t$ . Observe that  $(N_t(0) + 2)/(t + 3) > (N_t(0) + 1)/(t + 2) > (N_t(0) + 1)/(t + 3) \geq 1/2$  and consider the following two possible cases. In the first case  $(N_t(0) + 1)/(t + 2) > 1/2 + \epsilon_t$  and when  $\epsilon_t$  is nonincreasing we have  $(N_t(0) + 2)/(t + 3) > 1/2 + \epsilon_{t+1}$  and so  $\phi_t((N_t(0) + 1)/(t + 2)) = \phi_{t+1}((N_t(0) + 2)/(t + 3)) = 1$ , in which case  $\alpha - \beta = \phi_{t+1}((N_t(0) + 1)/(t + 3)) - 1 \leq 0$ . In the second case,  $(N_t(0) + 1)/(t + 2) \leq 1/2 + \epsilon_t$  and we can replace  $\phi_t((N_t(0) + 1)/(t + 2))$  by  $1/2\epsilon_t[(N_t(0) + 1)/(t + 2) - 1/2] + 1/2$ . Also, note that a continuation of the sloping part of  $\phi(\cdot)$  to the right serves as an upper bound

to this function and so, e.g.,  $\phi_{t+1}((N_t(0) + 2)/(t + 3)) \leq 1/(2\epsilon_{t+1})[(N_t(0) + 2)/(t + 3) - 1/2] + 1/2$ , and we can write in this case:

$$\begin{aligned} \alpha - \beta &\leq \frac{1}{2\epsilon_{t+1}} \left[ \frac{N_t(0) + 2}{t + 3} - \frac{1}{2} + \frac{N_t(0) + 1}{t + 3} - \frac{1}{2} \right] \\ &\quad - \frac{2}{2\epsilon_t} \left[ \frac{N_t(0) + 1}{t + 2} - \frac{1}{2} \right] \\ &= \frac{N_t(0) - t}{2\epsilon_{t+1}(t + 3)} - \frac{N_t(0) - t}{2\epsilon_t(t + 2)}. \end{aligned}$$

Thus, as long as  $\epsilon_t(t + 2) \leq \epsilon_{t+1}(t + 3)$ , we have again  $\alpha \leq \beta$ . In other words, whenever both  $\epsilon_t$  is nonincreasing and the function  $f(t) = \epsilon_t(t + 2)$  is nondecreasing, (e.g., a constant  $\epsilon$  or an  $\epsilon_t$  that monotonically goes to zero, say, as  $C/\sqrt{t + 2}$ , but not faster than  $C/(t + 2)$ ), the loss incurred in the upward loop is smaller than that of the downward loop.

When  $N_t(1) > N_t(0)$  then  $\alpha$  is the loss incurred in the downward loop and  $\beta$  is the loss incurred in the upward loop. Using similar observations we can show that  $\beta \leq \alpha$  under the similar condition on  $\epsilon_t$ . Again, the loss incurred over the downward loop can only be greater than the loss incurred in the upward loop. We note that the proof above can be generalized to any nondecreasing  $\phi(\alpha)$ , concave for  $\alpha \geq 1/2$ , such that  $\phi(\alpha) = 1 - \phi(1 - \alpha)$ .

Now given any sequence  $x_1^n$  and its trellis, one can replace every upward loop by a downward loop and thereby increase the average loss at each step. After a finite number of such steps one reaches a sequence such that the first  $N_n(1)$  pairs of bits  $(x_{2t-1}, x_{2t})$  are all either "01" or "10", and the remaining  $N_n(0) - N_n(1)$  bits are all "0". For such a sequence, all upward loops correspond to  $k = 0$  and hence cannot be replaced by downward loops. Thus, every sequence of this structure, in particular the sequence (12), incurs the same maximal loss. Note that by replacing any downward loop with an upward loop one ends up with a sequence of the form (19) which has, as a result, the smallest loss among all the sequences of the given composition.

*Proof of (13):* Since the sequence of (12), denoted  $\tilde{x}_1^n$ , is the worst sequence for a given value of  $\pi_1(x_1^n)$  then  $\hat{\pi}_1(x_1^n) \leq \hat{\pi}_1(\tilde{x}_1^n)$ . The average loss over  $\tilde{x}_1^n$  is

$$\begin{aligned} \hat{\pi}_1(\tilde{x}_1^n) &= \sum_{t=0}^{n-1} l\left(\frac{N_t(\tilde{x}_{t+1}) + 1}{t + 2}\right) \\ &= \sum_{k=1}^{N_n(1)} \left[ l\left(\frac{k}{2k}\right) + l\left(\frac{k}{2k+1}\right) \right] \\ &\quad + \sum_{k=1}^{N_n(0) - N_n(1)} l\left(\frac{N_n(1) + k}{2N_n(1) + k + 1}\right) \\ &= \frac{N_n(1)}{2} + \sum_{k=1}^{N_n(1)} l\left(\frac{k}{2k+1}\right) \end{aligned}$$

$$\begin{aligned} &+ \sum_{k=1}^{N_n(0) - N_n(1)} l\left(\frac{N_n(1) + k}{2N_n(1) + k + 1}\right) \\ &\triangleq \frac{N_n(1)}{2} + A + B. \end{aligned} \quad (\text{A.3})$$

Consider first the case where  $\epsilon$  is fixed. To overbound  $A = \sum_{k=1}^{N_n(1)} l(k/(2k+1))$ , we observe that  $l(k/(2k+1)) \leq 1/2 + (1/2\epsilon) \cdot (1/2 - k/(2k+1)) = 1/2 + (1/4\epsilon) \cdot 1/(2k+1)$ . Thus,

$$\begin{aligned} A &\leq \frac{N_n(1)}{2} + \frac{1}{4\epsilon} \sum_{k=1}^{N_n(1)} \frac{1}{2k+1} \\ &\leq \frac{N_n(1)}{2} + \frac{1}{4\epsilon} \int_0^{N_n(1)} \frac{du}{2u+1} \\ &\leq \frac{N_n(1)}{2} + \frac{1}{8\epsilon} \ln(2N_n(1) + 1) \\ &\leq \frac{N_n(1)}{2} + \frac{1}{8\epsilon} \ln(n + 1), \end{aligned} \quad (\text{A.4})$$

where in the last inequality we used the fact that  $2N_n(1) \leq n$ . As for  $B = \sum_{k=1}^{N_n(0) - N_n(1)} l((N_n(1) + k)/(2N_n(1) + k + 1))$ , some of the terms are zero and the arguments of  $l(\cdot)$  for the nonzero terms must satisfy  $(N_n(1) + k)/(2N_n(1) + k + 1) \leq 1/2 + \epsilon$  and hence, for these terms

$$k \leq \frac{4\epsilon N_n(1)}{1 - 2\epsilon} + \frac{1 + 2\epsilon}{1 - 2\epsilon} \triangleq K.$$

Also, the nonzero terms are smaller than  $1/2$  since the argument of  $l(\cdot)$  is greater than  $1/2$ . Thus,

$$\begin{aligned} B &\leq \sum_{k=1}^K \frac{1}{2} = \frac{2\epsilon N_n(1)}{1 - 2\epsilon} + \frac{1 + 2\epsilon}{2(1 - 2\epsilon)} \\ &\leq \frac{\epsilon}{1 - 2\epsilon} \cdot n + \frac{1 + 2\epsilon}{2(1 - 2\epsilon)}. \end{aligned} \quad (\text{A.5})$$

Combining (A.3), (A.4), and (A.5), we get

$$\begin{aligned} \hat{\pi}_1(\hat{x}_1^n) &\leq N_n(1) + \frac{\epsilon}{1 - 2\epsilon} \cdot n \\ &\quad + \frac{1}{8\epsilon} \ln(n + 1) + \frac{1 + 2\epsilon}{2(1 - 2\epsilon)}, \end{aligned} \quad (\text{A.6})$$

which completes the proof of (13).  $\square$

*Proof of (14):* When choosing  $\epsilon_t = 1/2\sqrt{t+2}$  we have  $l(k/(2k+1)) \leq 1/2 + 1/(2\epsilon_{2k-1}) \cdot (1/2 - k/(2k+1)) = 1/2 + 1/2 \cdot 1/\sqrt{2k+1}$ , and so

$$\begin{aligned} A &\leq \frac{N_n(1)}{2} + \frac{1}{2} \sum_{k=1}^{N_n(1)} \frac{1}{\sqrt{2k+1}} \\ &\leq \frac{N_n(1)}{2} + \frac{1}{2} \int_0^{N_n(1)} \frac{du}{\sqrt{2u+1}} \\ &\leq \frac{N_n(1)}{2} + \frac{\sqrt{n+1}}{2} - \frac{1}{2}. \end{aligned} \quad (\text{A.7})$$

Also in this case, the arguments of  $l(\cdot)$  for the nonzero terms in  $B$  must satisfy

$$\frac{N_n(1) + k}{2N_n(1) + k + 1} \leq \frac{1}{2} + \frac{1}{2\sqrt{2N_n(1) + k + 1}}, \quad k \geq 0,$$

which implies that  $k^2 - 3k \leq 2N_n(1)$ ,  $k \geq 0$ . Straightforward calculations show that the number of these nonzero components, denoted  $K$ , which is the maximal  $k$  satisfying the previous conditions, is upper bounded by

$$K \leq \sqrt{(2N_n(1) + \frac{9}{4}) + \frac{3}{2}} \leq \sqrt{(2N_n(1) + 1) + 2},$$

where the second inequality holds since  $N_n(1) \geq 0$ . Now each of these  $K$  nonzero terms is smaller than  $1/2$  and so

$$B \leq \sum_{k=1}^K \frac{1}{2} = \frac{1}{2} \sqrt{(2N_n(1) + 1) + 2} + 1 \leq \frac{1}{2} \sqrt{n+1} + 1. \tag{A.8}$$

Combining (A.3), (A.7), and (A.8), we get

$$\hat{\pi}_1(\tilde{x}_1^n) \leq \frac{N_n(1)}{2} A + B \leq N_n(1) + \sqrt{n+1} + \frac{1}{2}, \tag{A.9}$$

which completes the proof of (14).

Note that slightly different expressions for the loss may be obtained by choosing  $\epsilon_t = \epsilon_0 \sqrt{2}/\sqrt{t+2}$  with arbitrary  $\epsilon_0$ ; however, in all these cases the excess loss beyond  $\pi_1(x_1^n)$  decays like  $O(1/\sqrt{n})$ , and our choice of  $\epsilon_0 = 1/2\sqrt{2}$  leads to the tightest bound on the coefficient of  $1/\sqrt{n}$  that is attained in the previous technique.  $\square$

APPENDIX B

*Proof of Lemma 1:* The proof is based on Pinsker's inequality (see, e.g., [29, ch. 3, problem 17]) asserting that for every  $0 \leq p \leq 1$  and  $0 \leq q \leq 1$ ,

$$D(p||q) \triangleq p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \frac{2}{\ln 2} (p-q)^2.$$

Since  $\min\{p, 1-p\} - \min\{q, 1-q\} \leq |p-q|$ ,

$$p \log \frac{p}{q} + (1-p) \log \frac{1-p}{1-q} \geq \frac{2}{\ln 2} (\min\{p, 1-p\} - \min\{q, 1-q\})^2. \tag{B.1}$$

Let  $\hat{p}_n(\cdot)$  denote an empirical measure based on  $x_1^n$  where e.g.,  $\hat{p}_n(\tilde{s}, x) = \hat{p}_n(s, x^j, x) = (N_n(\tilde{s}, x)/n$ . Now,

$$\begin{aligned} & \hat{H}(X|X^j) - \hat{H}(X|X^j, \mathcal{S}) \\ &= \sum_{\tilde{s}} \hat{p}_n(\tilde{s}) \sum_{x=0,1} \hat{p}_n(x|\tilde{s}) \log \frac{\hat{p}_n(x|\tilde{s})}{\hat{p}_n(x|x^j)} \\ &\geq \frac{2}{\ln 2} \sum_{\tilde{s}} \hat{p}_n(\tilde{s}) \\ &\quad \cdot \left[ \min_{x \in \{0,1\}} \hat{p}_n(x|\tilde{s}) - \min_{x \in \{0,1\}} \hat{p}_n(x|x^j) \right]^2 \end{aligned}$$

$$\begin{aligned} & \geq \frac{2}{\ln 2} \left[ \sum_{\tilde{s}} \hat{p}_n(\tilde{s}) \left[ \min_{x \in \{0,1\}} \hat{p}_n(x|\tilde{s}) \right] \right. \\ & \quad \left. - \sum_{\tilde{s}} \hat{p}_n(\tilde{s}) \left[ \min_{x \in \{0,1\}} \hat{p}_n(x|x^j) \right] \right]^2, \tag{B.2} \end{aligned}$$

where the first inequality follows from (B.1) and the second follows from the convexity of the square function and Jensen's inequality. Noticing that

$$\begin{aligned} & \sum_{\tilde{s}} \hat{p}_n(\tilde{s}) \left[ \min_{x \in \{0,1\}} \hat{p}_n(x|x^j) \right] \\ &= \sum_{x^j} \hat{p}_n(x^j) \left[ \min_{x \in \{0,1\}} \hat{p}_n(x|x^j) \right] = \mu_j(x_1^n), \end{aligned}$$

and

$$\sum_{\tilde{s}} \hat{p}_n(\tilde{s}) \left[ \min_{x \in \{0,1\}} \hat{p}_n(x|\tilde{s}) \right] = \pi(\tilde{g}_j; x_1^n),$$

completes the proof of the lemma.  $\square$

REFERENCES

- [1] J. Ziv and A. Lempel, "Compression of individual sequences via variable rate coding," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 530-536, Sept. 1978.
- [2] M. Feder, "Gambling using a finite-state machine," *IEEE Trans. Inform. Theory*, vol. 37, pp. 1459-1465, Sept. 1991.
- [3] 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, 1957, no. 39, pp. 97-139.
- [4] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *Proc. 4th Prague Conf. Inform. Theory, Statistical Decision Functions, Random Processes*, 1965, Prague: Publishing House of the Czechoslovak Academy of Sciences, Prague, 1967, pp. 263-272.
- [5] 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.
- [6] H. Robbins, "Asymptotically subminimax solutions of compound statistical decision problems," in *Proc. Second Berkley Symp. Math. Statist. Prob.*, 1951, pp. 131-148.
- [7] 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, 1955.
- [8] D. Blackwell, "An analog to the minimax theorem for vector payoffs," *Pac. J. Math.*, vol. 6, pp. 1-8, 1956.
- [9] —, "Controlled random walk," in *Proc. Int. Congr. Mathematicians, 1954*, Vol. III. Amsterdam, North Holland, 1956, pp. 336-338.
- [10] N. Merhav and M. Feder, "Universal schemes for sequential decision from individual data sequences," submitted to *Inform. Comput.*, 1991.
- [11] R. J. Solomonoff, "A formal theory of inductive inference, pts. 1 and 2," *Inform. Contr.*, vol. 7, pp. 1-22 and pp. 224-254, 1964.
- [12] A. N. Kolmogorov, "Three approaches to the quantitative definition of information," *Probl. Inform. Transm.*, vol. 1, pp. 4-7, 1965.
- [13] G. J. Chaitin, "A theory of program size formally identical to information theory," *J. ACM*, vol. 22, pp. 329-340, 1975.
- [14] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629-636, 1984.
- [15] —, "Stochastic complexity and modelling," *Ann. Statist.*, vol. 14, pp. 1080-1100, 1986.
- [16] J. Ziv, "Coding theorems for individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 405-412, July 1978.
- [17] T. M. Cover, "Universal gambling schemes and the complexity measures of Kolmogorov and Chaitin," Tech. Rep. 12, Dept. of Statistics, Stanford Univ., Oct. 1974.
- [18] T. M. Cover and R. King, "A convergent gambling estimate of the

- entropy of English," *IEEE Trans. Inform. Theory*, vol. IT-24, pp. 413-421, July 1978.
- [19] G. G. Langdon, "A note on the Lempel-Ziv model for compressing individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 284-287, Mar. 1983.
- [20] J. Rissanen, "A universal data compression system," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 656-664, July 1983.
- [21] A. Lempel and J. Ziv, "On the complexity of finite sequences," *IEEE Trans. Inform. Theory*, vol. IT-22, pp. 75-81, Jan. 1976.
- [22] E. Plotnik, M. J. Weinberger, and J. Ziv, "Upper bounds on the probability of sequences emitted by finite-state sources and the redundancy of the Lempel-Ziv algorithm," *IEEE Trans. Inform. Theory*, vol. 38, pp. 66-72, Jan. 1992.
- [23] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. New York: Wiley, 1991.
- [24] L. D. Davisson, "Universal lossless coding," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 783-795, Nov. 1973.
- [25] L. D. Davisson, "Minimax noiseless universal coding for Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-29, pp. 211-215, Mar. 1983.
- [26] J. Rissanen, "Complexity of strings in the class of Markov sources," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 526-532, July 1986.
- [27] J. S. Vitter and P. Krishnan, "Optimal prefetching via data compression," tech. rep. CS-91-46, Brown Univ.; also summarized in *Proc. Conf. Foundation of Comput. Sci. (FOCS)*, 1991, pp. 121-130.
- [28] J. Rissanen, "Generalized Kraft's inequality and arithmetic coding," *IBM J. Res. Dev.*, vol. 20, no. 3, pp. 198-203, 1976.
- [29] I. Csiszár and J. Körner, *Information Theory—Coding Theorems for Discrete Memoryless Systems*. New York: Academic Press, 1981.