Some Properties of Sequential Predictors for Binary Markov Sources

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

***Abstract*—Universal prediction of the next outcome of a binary sequence drawn from a Markov source with unknown parameters is considered. For a given source, the predictability is defined as the least attainable expected fraction of prediction errors. A lower bound is derived on the maximum rate at which the predictability is asymptotically approached uniformly over all sources in the Markov class. This bound is achieved by a simple majority predictor. For Bernoulli sources, bounds on the large deviations performance are investigated. A lower bound is derived for the probability that the fraction of errors will exceed the predictability by a prescribed amount $\Delta > 0$. This bound is achieved by the same predictor if Δ is sufficiently small.**

***Index Terms*— Predictability, universal prediction, Bernoulli processes, Markov sources, large deviations.**

I. Introduction

IN [1], universal finite-state (FS) predictors have been sought that minimize the asymptotic fraction of errors for an individual binary sequence. It has been shown in [1] that the best prediction performance is asymptotically attained by a (randomized) Markov predictor with a slowly growing order, i.e., a predictor based on current estimates of the conditional probabilities of the next outcome given the k preceding bits, where the order k increases gradually with time. A predictor based on the Lempel–Ziv (LZ) algorithm [2] has been demonstrated in [1] to be such a growing-order Markov predictor and hence to attain asymptotically the least possible fraction of errors made by any FS predictor, that is, the *FS predictability* [1]. Independently, in [3] a similar predictor (though nonrandomized) has been proposed with application to prefetching memory pages in computers, where the page sequence is modeled as being governed by a probabilistic unifilar FS source. It has been shown in [3] that the resulting expected fraction of errors (page faults) converges to the optimum. However, if the source is known to have no more than S states, then the LZ algorithm, which does not utilize this prior information, might yield a relatively slow convergence. A natural question that arises and that we shall be concerned with, is: how fast can the optimum performance of approached when the predictor knows the class of sources but not the parameter value?

Manuscript received February 7, 1992; revised September 28, 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, Tel Aviv University, Tel Aviv 69978, Israel.
M. Gutman is with Intel Israel, P.O. Box 1659, Haifa 31015, Israel.
IEEE Log Number 9207087.

In [4], a similar question has been addressed in the context of predicting Gaussian autoregressive moving average (ARMA) processes under the minimum-mean-square-error (mmse) criterion. It has been shown in [4] that no predictor exists that approaches the asymptotic mmse faster than $n^{-1} \log n$, n being the sample size, for all ARMA processes except for a collection of ARMA processes corresponding to a subset of parameter values whose volume is vanishingly small. This argument was based on an analogous result in universal data compression (proved in [4] as well), which rules out the existence of a lossless code whose compression ratio converges to the entropy faster than $n^{-1} \log n$ for a considerably large subset of parameter values. Note that an exception of a small subset of parameter values is necessary if *every* scheme is allowed, including the optimal scheme for a specific parameter value.

In this paper, an attempt is made to investigate, in the same spirit, fundamental limitations in universal prediction of finite-alphabet Markov sources, and in particular, binary Markov sources. We derive a lower bound on the rate at which the optimum prediction performance can be uniformly approached by any sequential predictor when the underlying Markovian source is known to be of order k, but otherwise unknown. However, in contrast to [4], here one cannot expect a nontrivial lower bound that holds simultaneously for *most* sources in the class. Consider, for example, a Bernoulli source parametrized by $\theta = \Pr\{x_t = 1\} = 1 - \Pr\{x_t = 0\}$. Here, predicting constantly "0" is a uniformly optimal strategy for every $0 \leq \theta \leq 1/2$, namely, for "half" of the sources in the class there cannot be a lower bound on the rate of approaching optimality. Thus, the bound here will hold for only half of the sources. In the Markovian case, the bound will still hold for a "considerably large" portion of the parameter space, i.e., for a fixed fraction of its volume. In either case, the corresponding bound is attained by a simple predictor based on a majority count.

Finally, we examine the achievable large deviations performance for Bernoulli sources under the criterion of minimizing the probability that the fraction of errors would exceed the optimum by a prescribed amount Δ. We derive an exponentially tight lower bound and show that it is uniformly attained by the majority predictor in some range $0 < \Delta \leq \Delta_\theta$, but not for $\Delta > \Delta_\theta$.

II. A Lower Bound on the Expected Fraction of Errors

We start with Bernoulli sources and later extend our discussion to Markov sources. Let $x_1, x_2, \cdots, x_n$, $x_t \in \{0, 1\}$, denote a binary n-tuple drawn from a Bernoulli source parametrized by $\theta = \Pr\{x_t = 1\}$. A predictor is a sequence of functions $f = (f_0, f_1, \cdots)$, f_t: $\{0, 1\}^t \to \{0, 1\}$, where at time t, the next outcome x_{t+1} is estimated by $\hat{x}_{t+1} = f_t(x_1, x_2, \cdots, x_t)$. Let $n_e(f) \triangleq \sum_{t=1}^n 1\{\hat{x}_t \neq x_t\}$ (where $1\{\cdot\}$ denotes the indicator function of an event). We are interested in minimizing $\pi_\theta(f) \triangleq \limsup_{n\to\infty} E_\theta n_e(f)/n$, where E_θ denotes expectation w.r.t θ. The *predictability*, defined as $\pi_\theta \triangleq \inf_f \pi_\theta(f)$, is obviously attained by the predictor $\hat{x}_{t+1} = 0$ if $\theta \leq 1/2$ and $\hat{x}_{t+1} = 1$ if $\theta > 1/2$. Hence, $\pi_\theta = \min\{\theta, \overline{\theta}\}$, where $\overline{\alpha}$ denotes $1 - \alpha$.

If θ is not known, then one does not know which one of these predictors to use, and therefore π_θ cannot be attained for every n and uniformly for every θ, but only asymptotically. We shall be interested in the rate at which π_θ can be uniformly attained when $n \to \infty$. Intuitively, the predictor

$$x^*_{t+1} = f^*_t(x_1, \cdots, x_t) = \begin{cases} 0, & \text{if } \hat{\theta}(t) < 1/2, \\ \text{flip a fair coin}, & \text{if } \hat{\theta}(t) = 1/2, \\ 1, & \text{if } \hat{\theta}(t) > 1/2, \end{cases} \tag{1}$$

where $\hat{\theta}(t) = t^{-1}\sum_{\tau=1}^t x_\tau$ is the current estimate of θ ($\hat{\theta}(0) \triangleq 1/2$), is in some sense the best one can use when θ is unknown. The following theorem consolidates this intuition.

Theorem 1:

a) For every predictor f, and every $\theta \neq 1/2$ either

$$E_\theta n_e(f) \geq n\pi_\theta + c_0(\theta) - o(1)$$

or

$$E_{\overline{\theta}} n_e(f) \geq n\pi_{\overline{\theta}} + c_0(\overline{\theta}) - o(1),$$

where $c_0(\theta) = c_0(\overline{\theta}) = [2(1 - 2\pi_\theta)]^{-1}$.

b) The predictor f^* satisfies both

$$E_\theta n_e(f^*) \leq n\pi_\theta + c_0(\theta),$$

and

$$E_{\overline{\theta}} n_e(f^*) \leq n\pi_{\overline{\theta}} + c_0(\overline{\theta}).$$

Part a) tells us that every predictor must make on the average at least $c_0(\theta) = c_0(\overline{\theta})$ extra prediction errors beyond the minimum $n\pi_\theta = n\pi_{\overline{\theta}}$, for either θ or $\overline{\theta}$. Part b) implies that f^* is optimal in the sense of doing no worse for both sources. It follows from the simple inequality $0.5E_\theta n_e(f) + 0.5E_{\overline{\theta}} n_e(f) \geq 0.5E_\theta n_e(f^*) + 0.5E_{\overline{\theta}} n_e(f^*)$, which is justified below and has been observed independently by Rissanen [5]. Thus the convergence rate is $O(1/n)$.

In [1], [6]–[8] where prediction of *individual* sequences is considered, the convergence rate to the predictability (defined in [1] for the deterministic case) slows down to $O(1/\sqrt{n})$. The reason for the difference is that in the deterministic setup of [1], [6]–[8], a uniform upper bound is derived from a worst case analysis rather than the expectation over an ensemble of sequences.

Proof of Theorem 1: First we find a tight lower bound on $M(f) \triangleq 0.5E_\theta n_e(f) + 0.5E_{\overline{\theta}} n_e(f)$ for an arbitrary predictor f, and then we argue that this lower bound must hold for either $E_\theta n_e(f)$, or $E_{\overline{\theta}} n_e(f)$, or both, and hence for at least one half of the Bernoulli sources. We show that the tightest lower bound is $M(f^*)$ and hence it remains to evaluate the performance of f^*. First observe that by (1), $E_\theta n_e(f) = \sum_{t=1}^n P_\theta\{\hat{x}_t \neq x_t\}$, where $P_\theta\{\cdot\}$ denotes a probability w.r.t θ. Without loss of generality, let $\theta < 1/2$. Since x_t is independent of $x_1, x_2, \cdots, x_{t-1}$, and hence also of $\hat{x}_t$,

$$\begin{aligned} P_\theta\{\hat{x}_t \neq x_t\} &= P_\theta\{\hat{x}_t = 0\} \cdot P_\theta\{x_t = 1\} + P_\theta\{\hat{x}_t = 1\} \\ &\quad \cdot P_\theta\{x_t = 0\} \\ &= \theta \cdot [1 - P_\theta\{\hat{x}_t = 1\}] + (1 - \theta) \cdot P_\theta\{\hat{x}_t = 1\} \\ &= \theta + (1 - 2\theta) \cdot P_\theta\{\hat{x}_t = 1\} \\ &= \pi_\theta + (1 - 2\theta) \cdot P_\theta\{\hat{x}_t = 1\}. \end{aligned} \tag{2}$$

Similarly, $P_{\overline{\theta}}\{\hat{x}_t \neq x_t\} = \pi_{\overline{\theta}} + (1 - 2\theta) \cdot P_{\overline{\theta}}\{\hat{x}_t = 0\}$. The second term on the right-most side of (2) describes the excess in error probability beyond the predictability incurred by using a nonoptimal predictor for θ. Since $\pi_\theta = \pi_{\overline{\theta}}$, the minimization of $M(f)$ is equivalent to the minimization of $0.5P_\theta\{\hat{x}_t = 1\} + 0.5P_{\overline{\theta}}\{\hat{x}_t = 0\}$. This, in turn, can be thought of as a binary hypothesis testing problem where one seeks a rule f_t for deciding in favor of θ or $\overline{\theta}$ with priors $p(\theta) = p(\overline{\theta}) = 1/2$, and the goal is to minimize the error probability. This is accomplished by comparing the likelihood ratio $P_\theta(x_1, x_2, \cdots, x_{t-1})/P_{\overline{\theta}}(x_1, x_2, \cdots, x_{t-1})$ to unity, which is equivalent to f^*_{t-1} in (1). Thus the average of $E_\theta n_e(f)$ and $E_{\overline{\theta}} n_e(f)$ is minimized by f^* and either $E_\theta n_e(f)$ or $E_{\overline{\theta}} n_e(f)$ is not less than $E_\theta n_e(f^*) = E_{\overline{\theta}} n_e(f^*)$.

To complete the proof, it remains to prove part b). To do this, we evaluate the performance of f^* for $\theta < 1/2$. From (2), $E_\theta n_e(f^*) = n\pi_\theta + (1 - 2\theta) \cdot \sum_{t=1}^n P_\theta\{x^*_t = 1\}$, where the summation on the right-hand side converges to a constant $A \triangleq \sum_{t\geq 1}[P_\theta\{\hat{\theta}(t - 1) > 1/2\} + 0.5P_\theta\{\hat{\theta}(t - 1) = 1/2\}]$, which can be calculated by generating function techniques [9, ch. IV, section 17] in the following manner. Define $Y_t = 2x_t - 1$ and a random walk $S_t = \sum_{i=1}^t Y_i$. We wish to calculate $A = 0.5 + \sum_{t\geq 1}[P_\theta\{S_t > 0\} + 0.5P_\theta\{S_t = 0\}]$. Let $z = e^{j\omega}$ and $\phi(z) = E_\theta z^{Y_1} = \theta z + \overline{\theta} z^{-1}$. For $|r| \leq 1$ we first factor, in two different ways, the function $1 - r\phi(z)$ as a product $c(r)f_+(r, z)f_-(r, z)$, where f_+ and f_- contain positive and negative powers of z, respectively. A direct spectral factorization of the second-order polynomial in z, $1 - r\phi(z) = 1 - r\theta z - r\overline{\theta} z^{-1}$ yields $c(r) = 0.5(1 + \rho)$, $f_+(r, z) = 1 - (2\overline{\theta} r)^{-1} z(1 - \rho)$ and $f_-(r, z) = 1 - (2\theta r z)^{-1}(1 - \rho)$ where $\rho = [1 - 4\theta\overline{\theta} r^2]^{1/2}$. On the other hand, $1 - r\phi(z) = \exp[\log(1 - rE_\theta z^{Y_1})] = \exp[-\sum_{t\geq 1} t^{-1}(rE_\theta z^{Y_1})^t] = \exp[-\sum_{t\geq 1} r^t E_\theta z^{S_t}]$, where we have used the Taylor expansion of the logarithmic function and the fact that $[E_\theta z^{Y_1}]^t = E_\theta z^{S_t}$ for independent copies of Y_1. Now, $E_\theta z^{S_t}$ is composed from contributions of negative and positive powers of z in accordance to the sign of S_t. Thus, the exponent can be factored as cf_+f_- where $c(r) = \exp[-\sum_{t\geq 1} t^{-1} r^t P_\theta\{S_t = 0\}]$, $f_+(r, z) = \exp[-\sum_{t\geq 1} t^{-1} r^t E_\theta(z^{S_t} \cdot 1\{S_t > 0\})]$, and $f_-(r, z) =$

$\exp[-\sum_{t\geq 1} t^{-1}r^t E_\theta(z^{S_t}\cdot 1\{S_t<0\})]$. Therefore,

$$A = \frac{1}{2} - \frac{d}{dr}\log f_+(r,1)\bigg|_{r=1} - \frac{1}{2}\frac{d}{dr}\log c(r)\bigg|_{r=1} = \frac{1}{2(1-2\theta)^2}, \tag{3}$$

which yields the desired result and completes the proof of Theorem 1. □

The explicit calculation of $c_0(\theta)$ is more involved in the Markovian case. An alternative representation of $c_0(\theta)$ in the Bernoulli case, which will be extended to the Markov case, is given by $c_0(\theta) = 0.5\sum_{t=0}^{\infty} P_\theta\{\hat{\theta}(t) = 1/2\}$. This can be shown as an immediate corollary of the following lemma (whose proof appears in the appendix), and the fact that $n^{-1}E_\theta \min\{n(0), n(1)\}$ converges exponentially to π_θ.

Lemma 1: For a Bernoulli source θ, $E_\theta n_e(f^*) = E_\theta \min\{n(0), n(1)\} + 0.5E_\theta n^*$, where $n(0)$ and $n(1)$ are counts of zeroes and ones, respectively, along $x_1,\cdots,x_n$ and n^* is the number of times t that $\hat{\theta}(t) = 1/2$, i.e., $n^* = \sum_{t=0}^{n-1} 1\{\hat{\theta}(t) = 1/2\}$.

Next, we consider binary Markov sources. For simplicity, we shall confine our discussion to the first-order case, but the results will generalize straightforwardly to the kth-order case. A first-order Markov source is indexed by a vector $\theta = (\theta_0, \theta_1)$, where $\theta_0 = \Pr\{x_{t+1} = 1 \mid x_t = 0\}$ and $\theta_1 = \Pr\{x_{t+1} = 1 \mid x_t = 1\}$. To guarantee that the source is irreducible and aperiodic (see, e.g., [10]), we shall assume that $\theta \in \Theta \triangleq \{\theta\colon \theta_0 > \theta,\ \theta_1 < 1, \text{ and either } \theta_0 < 1 \text{ or } \theta_1 > 0\}$. This ensures the existence of stationary probabilities $\mu_\theta = \lim_{t\to\infty}\Pr\{x_t = 1\}$ and $\overline{\mu}_\theta = \lim_{t\to\infty}\Pr\{x_t = 0\}$ and hence enables the definition of the predictability π_θ as $\inf_f \pi_\theta(f)$, which is attained by $\hat{x}_{t+1} = f_t(x_1,\cdots,x_t) = g(x_t)$, where $g(x) \triangleq 1\{\theta_x \geq 1/2\}$, $x = 0, 1$. Consequently, $\pi_\theta = \overline{\mu}_\theta \min\{\theta_0, \overline{\theta}_0\} + \mu_\theta \min\{\theta_1, \overline{\theta}_1\}$. For an unknown θ, a natural extension of f^* to the Markov case is

$$x^*_{t+1} = f^*_t(x_1,\cdots,x_t) = \begin{cases} 0, & \text{if } \hat{\theta}_{x_t}(t) < 1/2,\\ \text{flip a fair coin}, & \text{if } \hat{\theta}_{x_t}(t) = 1/2,\\ 1, & \text{if } \hat{\theta}_{x_t}(t) > 1/2,\end{cases} \tag{4}$$

where $\hat{\theta}_x(t) = n_t(x,1)/n_t(x)$, $x = 0, 1$, $n_t(x,1)$ being the number of transitions from $x_\tau = x$ to $x_{\tau+1} = 1$, $\tau = 0, 1,\cdots,t-1$, and $n_t(x) = n_t(x,1) + n_t(x,0)$ is the number of occurrences of the symbol x in $x_0, x_1,\cdots,x_{t-1}$. If $n_t(x) = 0$, $\hat{\theta}_x(t) \triangleq 1/2$. Define the *prediction error redundancy* of f at θ as $R_n(f,\theta) = n^{-1}E_\theta n_e(f) - \pi_\theta$. For a given $\theta = (\theta_0, \theta_1)$, define the *reflection set* as $G(\theta) = \{(\theta_0,\theta_1), (\theta_0,\overline{\theta}_1), (\overline{\theta}_0,\theta_1), (\overline{\theta}_0,\overline{\theta}_1)\}$. The following is an extension of Theorem 1 to the Markovian case.

Theorem 2:

a) For every f, any $\theta \in \Theta$, and all n, there exists at least one point $\theta' \in G(\theta)$ such that $nR_n(f,\theta') \geq c_1(\theta') - o(1)$ where $c_1(\theta') = 0.5\sum_{x=0}^{1}\sum_{t=1}^{\infty} P_{\theta'}\{x_t = x, \hat{\theta}_x(t) = 1/2\}$.

b) The predictor f^* satisfies $nR_n(f^*,\theta) \leq c_1(\theta)$ for all $\theta \in \Theta$ and all n.

The theorem tells that the decay rate of $R_n(f,\theta)$ cannot be faster than that of f^* at least at one of the four points in $G(\theta)$, for every θ. In other words, $R_n(f^*,\theta) \leq R_n(f,\theta)$ for at least a "quarter" of the binary Markov sources. Note that this holds for all n.

Proof of Theorem 2: We prove that $R_n(f,\theta') \geq R_n(f^*,\theta')$ for at least one point $\theta' \in G(\theta)$. The proof of part b) is a straightforward extension of the proof of Theorem 1b), where the expression for $c_1(\theta)$ is obtained as a simple generalization of the expression for $c_0(\theta)$. For a given f,

$$\begin{aligned} &P_\theta\{\hat{x}_{t+1} \neq x_{t+1}\}\\ &= \sum_{a=0}^{1}\sum_{b=0}^{1} P_\theta(x_t = a)P_\theta(x_{t+1} = b|x_t = a)\\ &\quad\cdot P_\theta(\hat{x}_{t+1} = \overline{b}|x_{t+1} = b, x_t = a)\\ &\overset{\text{(a)}}{=} \sum_{a=0}^{1}\sum_{b=0}^{1} P_\theta(x_t = a)P_\theta(x_{t+1} = b|x_t = a)\\ &\quad\cdot P_\theta(\hat{x}_{t+1} = \overline{b}|x_t = a)\\ &\overset{\text{(b)}}{=} \pi_\theta + (1 - 2\min\{\theta_0, \overline{\theta}_0\})\\ &\quad\cdot P_\theta\{x_t = 0, \hat{x}_{t+1} \neq 1\{\theta_0 \geq 1/2\}\}\\ &\quad + (1 - 2\min\{\overline{\theta}_1, \tilde{\theta}_1\})\\ &\quad\cdot P_\theta\{x_t = 1, \hat{x}_{t+1} \neq 1\{\theta_1 \geq 1/2\}\}\\ &\triangleq \pi_\theta + r_t(f,\theta), \end{aligned} \tag{5}$$

where equality a) follows from Markovity and the fact that $\hat{x}_{t+1}$ depends only on $x_t, x_{t-1},\cdots$, equality b) is obtained similarly to (2), and $R_n(f,\theta) = n^{-1}\sum_{t=0}^{n-1} r_t(f,\theta)$. Clearly, $\min R_n(f,\theta)$ is obtained by minimizing each term of $r_t(f,\theta)$ individually. Every predictor f is a pair (g_0, g_1) of sequences of prediction functions associated with state $x_t = 0$ and $x_t = 1$, respectively. We shall denote $r_t(f,\theta)$ and P_θ by the more detailed notations $r_t(g_0, g_1, \theta_0, \theta_1)$ and $P_{\theta_0\theta_1}\{\cdot\}$, respectively. From (5),

$$\begin{aligned} &r_t(g_0, g_1, \theta_0, \theta_1)\\ &= (1 - 2\min\{\theta_0, \overline{\theta}_0\})\\ &\quad\cdot P_{\theta_0\theta_1}\{x_t = 0, \hat{x}_{t+1} \neq 1\{\theta_0 \geq 1/2\}\}\\ &\quad + (1 - 2\min\{\theta_1, \overline{\theta}_1\})\\ &\quad\cdot P_{\theta_0\theta_1}\{x_t = 1, \hat{x}_{t+1} \neq 1\{\theta_1 \geq 1/2\}\}, \end{aligned} \tag{6}$$

and similarly,

$$\begin{aligned} &r_t(g_0, g_1, \overline{\theta}_0, \theta_1)\\ &= (1 - 2\min\{\theta_0, \overline{\theta}_0\})\\ &\quad\cdot P_{\overline{\theta}_0\theta_1}\{x_t = 0, \hat{x}_{t+1} \neq 1\{\theta_0 < 1/2\}\}\\ &\quad + (1 - 2\min\{\theta_1, \overline{\theta}_1\})\\ &\quad\cdot P_{\overline{\theta}_0\theta_1}\{x_t = 1, \hat{x}_{t+1} \neq 1\{\theta_1 \geq 1/2\}\}. \end{aligned} \tag{7}$$

Similar to the hypothesis testing consideration of Theorem 1, the average of the first terms on the right-hand side of (6) and (7) is minimized if g_0 is replaced by g_0^*, which means using (4) when the current state x_t is zero. The second term in both (6) and (7), which corresponds to state one, remains unaffected. Hence,

$$\tfrac{1}{2}r_t(g_0, g_1, \theta_0, \theta_1) + \tfrac{1}{2}r_t(g_0, g_1, \overline{\theta}_0, \theta_1) \geq \tfrac{1}{2}r_t(g_0^*, g_1, \theta_0, \theta_1) + \tfrac{1}{2}r_t(g_0^*, g_1, \overline{\theta}_0, \theta_1). \quad (8)$$

Similarly,

$$\tfrac{1}{2}r_t(g_0, g_1, \theta_0, \overline{\theta}_1) + \tfrac{1}{2}r_t(g_0, g_1, \overline{\theta}_0, \overline{\theta}_1) \geq \tfrac{1}{2}r_t(g_0^*, g_1, \theta_0, \overline{\theta}_1) + \tfrac{1}{2}r_t(g_0^*, g_1, \overline{\theta}_0, \overline{\theta}_1). \quad (9)$$

Combining (8) and (9), we get

$$\begin{aligned}
&\tfrac{1}{4}\textstyle\sum_{\theta' \in G(\theta)} r_t(f, \theta') \\
&\geq \tfrac{1}{2}[\tfrac{1}{2}(r_t(g_0^*, g_1, \theta_0, \theta_1) + r_t(g_0^*, g_1, \theta_0, \overline{\theta}_1)) \\
&\quad + \tfrac{1}{2}(r_t(g_0^*, g_1, \overline{\theta}_0, \theta_1) + r_t(g_0^*, g_1, \overline{\theta}_0, \overline{\theta}_1))] \\
&\geq \tfrac{1}{2}[\tfrac{1}{2}(r_t(g_0^*, g_1^*, \theta_0, \theta_1) + r_t(g_0^*, g_1^*, \theta_0, \overline{\theta}_1)) \\
&\quad + \tfrac{1}{2}(r_t(g_0^*, g_1^*, \overline{\theta}_0, \theta_1) + r_t(g_0^*, g_1^*, \overline{\theta}_0, \overline{\theta}_1))] \\
&= \tfrac{1}{4}\textstyle\sum_{\theta' \in G(\theta)} r_t(f^*, \theta'). \qquad (10)
\end{aligned}$$

where the second inequality follows from the hypothesis testing consideration applied to g_1^*, i.e., the predictor (4) at state $x_t = 1$. Taking a time average of $r_t(f, \theta')$ results in $R_n(f, \theta') \geq R_n(f^*, \theta')$ for at least one θ' in $G(\theta)$, completing the proof of Theorem 2. □

One might wonder whether optimality of f^* on a quarter of Θ, as was mentioned earlier, is the strongest possible statement that can be made when allowing simultaneously both every θ and every f. There seems to be two answers to this question.

1) Strictly speaking, one answer is yes since there exists a predictor $\tilde{f}$ for which $r_t(\tilde{f}, \theta') < r_t(f^*, \theta')$ at *three* points of $G(\theta)$ for *every* θ. Thus the lower bound can be violated simultaneously for three quarters of Θ. A counterexample is a predictor that knows θ is outside one quarter of the parameter space, say, $\theta \notin Q \triangleq \{\theta\colon \theta_0 > 1/2, \theta_1 > 1/2\}$. Such prior information improves the prediction performance for every $\theta \in Q^c$ as shown in the appendix.
2) Theorem 2a) can be modified to hold for at least *three* quarters of Θ at the expense of decreasing $c_1(\theta')$. An alternative form of Theorem 2a) is the following: For every predictor f, every parameter θ, and for at least three points θ' in $G(\theta)$,

$$nR_n(f, \theta') \geq \tilde{c}_1(\theta') \triangleq \tfrac{1}{2}\min_{x \in \{0,1\}} \sum_{t=1}^{\infty} P_{\theta'}\{x_t = x, \hat{\theta}_x(t) = \tfrac{1}{2}\}. \quad (11)$$

(The proof appears in the Appendix.) This means that the lower bound given in the right-hand side of (11) holds for three quarters of the binary first-order Markov sources. Again, note that this result cannot be strengthened as there exists a predictor that indeed violates (11) at one point $\theta' \in \Theta$: the optimal predictor for $\theta \in G(\theta)$, which satisfies $nR_n(f, \theta) = 0$ simultaneously for all sources in the same quarter as θ.

In the extension of Theorem 2 to kth-order Markov sources parametrized by $\theta = (\theta_0, \theta_1, \cdots, \theta_{2^k-1})$, where

$$\theta_i \triangleq \Pr\{x_t = 1 | (x_{t-k}, \cdots, x_{t-1}) = \text{binary expansion of } i\}, \quad i = 0, 1, \cdots, 2^k - 1,$$

$G(\theta)$ includes all 2^k points of the form $\tilde{\theta} = (\tilde{\theta}_0, \tilde{\theta}_1, \cdots, \tilde{\theta}_{2^k-1})$, where $\tilde{\theta}_i$ is either θ_i or $\overline{\theta}_i$. Here, the extension of Theorem 2a), which describes the behavior of f^* for kth-order Markov sources, holds for a fraction 2^{-2^k} of Θ, but the extension of (11) holds for a fraction $(1 - 2^{-2^k})$ of Θ.

III. Large Deviations Performance

We now return to Bernoulli sources and evaluate the large deviations performance of f^*, i.e., the exponential decay rate of $P_\theta\{n_e(f) > n(\pi_\theta + \Delta)\}$ for a prescribed $\Delta \in (0, 1/2 - \pi_\theta)$. We show that f^* attains the optimal error exponent for a certain range $0 < \Delta \leq \Delta_\theta$. However, if $\Delta > \Delta_\theta$ this is no longer true. We first derive an exponentially tight upper bound for the error exponent.

Theorem 3: For every Bernoulli source θ and any predictor f,

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\ln P_\theta\{n_e(f) > n\zeta\}\right] \leq D(\zeta\|\pi_\theta),$$

where $\zeta \triangleq \pi_\theta + \Delta < 1/2$ and $D(\zeta\|\pi_\theta) \triangleq \zeta\ln(\zeta/\pi_\theta) + \overline{\zeta}\ln(\overline{\zeta}/\overline{\pi}_\theta)$.

The bound is exponentially tight because, for $\theta \leq 1/2$ and $f = 0$, $n_e(f) = n(1)$ and the large deviations behavior of $n(1)$ is obviously characterized by $D(\zeta\|\theta)$.

Proof of Theorem 3: Define $E = \{\boldsymbol{x}\colon \min\{n(0), n(1)\} \geq n(\zeta + \epsilon)\}$, $F = \{\boldsymbol{x}\colon n_e(f) \geq \min\{n(0), n(1)\} - \epsilon\}$, and $G = \{\boldsymbol{x}\colon n_e(f) \geq n\zeta\}$. Since $G \supseteq E \cap F$ then $P_\theta(G) \geq P_\theta(E \cap F) = [1 - P_\theta(F^c|E)]P_\theta(E)$. Now $P_\theta(E) = P_\theta\{\zeta + \epsilon \leq n(1)/n \leq 1 - \zeta - \epsilon\} \doteq \exp[-nD(\zeta + \epsilon\|\pi_\theta)]$, where the notation $a_n \doteq b_n$ means that $n^{-1}\log(a_n/b_n) \to 0$. The exponent on the right-hand side can be made arbitrarily close to $D(\zeta\|\pi_\theta)$ by choosing ϵ sufficiently small. Thus, to complete the proof it suffices to show that $P_\theta(F^c|E) \to 0$ as $n \to \infty$. To see this, divide the space of binary n-tuples into *types*, where the type $T_{\boldsymbol{x}}$ associated with a binary n-tuple $\boldsymbol{x} = (x_1, \cdots, x_n)$ is the set of all n-sequences with the same composition $\{n(0), n(1)\}$ as that of $\boldsymbol{x}$. Now

$$\begin{aligned}
P_\theta(F^c|E) &= \sum_{T_{\boldsymbol{x}} \subset E} P_\theta(F^c|T_{\boldsymbol{x}}) \cdot \frac{P_\theta(T_{\boldsymbol{x}})}{P_\theta(E)} \\
&\leq (n+1) \cdot \max_{T_{\boldsymbol{x}}} P_\theta(F^c|T_{\boldsymbol{x}}). \qquad (12)
\end{aligned}$$

Since all sequences of a given type are equiprobable, we see that $P_\theta(F^c|T_{\boldsymbol{x}})$ is just the fraction of $T_{\boldsymbol{x}}$-typical sequences with $n_e(f) < \min\{n(0), n(1)\} - n\epsilon$. We claim that this fraction is exponentially small. Indeed, a type $T_{\boldsymbol{x}}$ that corresponds to a composition $\{n(0) = n\alpha, n(1) = n\overline{\alpha}\}$ contains $n!/[(n\alpha)!(n\overline{\alpha})!] \doteq e^{nh(\alpha)}$ sequences where $h(\alpha) = -\alpha\ln\alpha - \overline{\alpha}\ln\overline{\alpha}$. Without loss of generality, assume that $\alpha \leq 1/2$ and observe that for a given f, the mapping from $\boldsymbol{x}$ to the error sequence $e_1, e_2, \cdots, e_n$, (where $e_t = |x_t - \hat{x}_t|$) is one-to-one. The number of error sequences with less than $n(\alpha-\epsilon)$ ones (prediction errors) is less than $e^{nh(\alpha-\epsilon)}$ [11, lemma 2.3.5]. Thus, the fraction of sequences in F^c is exponentially never larger than $\exp\{-n[h(\alpha)-h(\alpha-\epsilon)]\}$. This completes the proof of Theorem 3. □

We next present the upper bound associated with f^*.

Theorem 4: For a Bernoulli source θ, and $\pi_\theta < \zeta \le \zeta_\theta \triangleq 0.5\sqrt{\pi_\theta/\overline{\pi}_\theta}$,

$$\lim_{n\to\infty}\left[-\frac{1}{n}\ln P_\theta\{n_e(f^*) > n\zeta\}\right] \ge D(\zeta\|\pi_\theta). \tag{13}$$

Proof: Again, assume that $0 \le \theta \le 1/2$ and hence $\pi_\theta = \theta$. It is shown in the appendix (proof of Lemma 1) that $n_e(f^*) \le \min\{n(0), n(1)\} + n^*$, where $n^* = \sum_{t=0}^{n-1} 1\{\hat{\theta}(t) = 1/2\}$. Let Q denote the random variable $n(1)/n$ and let $\tilde{Q} \triangleq \min\{Q, \overline{Q}\}$. Similarly, let q denote a particular value of Q and let $\tilde{q} = \min\{q, \overline{q}\}$. By Lemma 1, the large deviation event $G^* = \{\boldsymbol{x}: n_e(f^*) > n\zeta\}$ is a subset of $\{\boldsymbol{x}: n^* \ge n\alpha\}$ where $\alpha \triangleq \zeta - \tilde{Q}$. For a given q (i.e., for a given type $T_{\boldsymbol{x}}$), we first upper bound $P_\theta\{n^* > n\alpha | Q = q\}$. Since all q-typical sequences are equiprobable given $Q = q$, this is just the fraction of q-typical sequences for which $n^* > n\alpha$. Let $n_t(x)$, $x = 0, 1$ denote the count of the symbol x in $x_1, x_2, \cdots, x_t$. Note that if there are at least $n\alpha$ occurrences of $\hat{\theta}(t) = 1/2$, i.e., $n_t(0) = n_t(1)$, then there must be at least one occurrence for some $t \ge 2n\alpha$, as this event can occur only at even time instants. Thus,

$$G^* \subseteq \{\boldsymbol{x}: n^* \ge n\alpha\} \subseteq \bigcup_{t=2n\alpha}^{n} \{\boldsymbol{x}: n_t(0) = n_t(1)\}. \tag{14}$$

Since the number of sequences with a given $Q = q$ is larger than $(n+1)^{-1}\exp[nh(q)]$, [12] where $h(q) \triangleq -q\ln q - \overline{q}\ln\overline{q}$, it follows from (14) and the union bound that

$$\begin{aligned}
&P_\theta\{G^* | Q = q\} \\
&\le (n+1)\exp[-nh(q)]\sum_{t=2n\alpha}^{n} 2^t \binom{n-t}{\max\{n(0), n(1)\} - t/2} \\
&\doteq (n+1)^2 \exp[-nh(q)] \\
&\quad \cdot \sum_{t=2n\alpha}^{n} \exp\left[t\ln 2 + (n-t)h\left(\frac{\max\{n(0), n(1)\} - t/2}{n-t}\right)\right] \\
&\le (n+1)^2 \exp\left\{n \cdot \max_{2\alpha\le\xi\le 1}\right. \\
&\qquad \left.\cdot \left[\xi\ln 2 + (1-\xi)\cdot h\left(\frac{\tilde{q} - \xi/2}{1-\xi}\right) - h(\tilde{q})\right]\right\} \\
&\doteq \exp[nC(\tilde{q}, \alpha)],
\end{aligned} \tag{15}$$

where

$$C(\tilde{q}, \alpha) \triangleq \begin{cases} 0, & \text{if } \alpha < 0, \\ 2\alpha\ln 2 + (1-2\alpha) + \\ h[(\tilde{q}-\alpha)/(1-2\alpha)] - h(\tilde{q}), & \text{if } 0 \le \alpha \le \tilde{q}. \end{cases} \tag{16}$$

Note that for $\alpha > \tilde{q}$ the set $\{\boldsymbol{x}: n^* > n\alpha\}$ is empty because $n^* \le \min\{n(0), n(1)\}$. The last step in (15) is obtained by maximizing the previous exponent function in the usual way. Since $P_\theta\{Q = q\} \le e^{-nD(q\|\theta)}$,

$$\begin{aligned}
P_\theta(G^*) &= \sum_{q=0/n}^{n/n} P_\theta\{Q = q\} \cdot P_\theta\{G^* | Q = q\} \\
&\le (n+1)^2 \sum_q \exp[-nD(q\|\theta)] \cdot \exp[-nC(\tilde{q}, \zeta - \tilde{q})] \\
&\le (n+1)^3 \exp\left\{-n \inf_{q:\, \zeta/2 \le \tilde{q} \le \zeta}\right. \\
&\qquad \left.\cdot [D(q\|\theta) + C(\tilde{q}, \zeta - \tilde{q})]\right\}.
\end{aligned} \tag{17}$$

The lower limit $\zeta/2$ is obtained from the fact that $n^* \le \min\{n(0), n(1)\}$ and hence $\alpha = \zeta - \tilde{q} \le \tilde{q}$. For the upper limit, observe that for sequences with $\tilde{q} > \zeta$ (which means $\alpha < 0$), the event $n^* > n\alpha$ obviously holds. These sequences contribute a probability which is exponentially equivalent to $e^{-nD(\zeta\|\theta)}$. Since the exponent on the right-most side of (17) never exceeds $D(\zeta\|\theta)$ (set $q = \tilde{q} = \zeta$ in (17)), the maximum of the previous function can be found for $\tilde{q} \le \zeta$. From standard extremum analysis of this function, we find that for $\theta < \zeta \le \zeta_\theta$, the minimum is obtained at $q = \zeta$ and its value is $D(\zeta\|\theta)$. This completes the proof of Theorem 4. □

This interesting phenomenon, that the error exponent is optimal for small threshold values $\Delta = \zeta - \theta$ but suboptimal for large values of Δ, is not a consequence of a possible looseness of the upper bound. A lower bound on $\Pr\{n_e(f^*) > n\zeta\}$ reveals the same effect. The intuition is that $n_e(f^*)$ is composed from $\min\{n(0), n(1)\}$ and n^*. When Δ is small, then $D(\zeta\|\theta)$, which characterizes the large deviations behavior of $\min\{n(0), n(1)\}$, is small as well. Thus, the large deviations behavior of $n_e(f^*)$ is dominated by that of $\min\{n(0), n(1)\}$. On the other hand, if Δ grows beyond a certain point, then the large deviations properties of n^* affect the performance.

Another aspect of the large deviations performance of f^* is its *competitive optimality*. Specifically, since $n_e(f^*) \le \min\{n(0), n(1)\} + n^*$ and for every competing predictor $n_e(f) \ge \min\{n(0), n(1)\} - n\epsilon$ except for an exponentially small minority of sequences from each type, we see that $P_\theta\{n_e(f^*) \ge n_e(f) + n\epsilon\}$ decays exponentially with n for every $\epsilon > 0$. An immediate conclusion, by the Borel–Cantelli lemma, is that $\limsup_{n\to\infty} n^{-1}[n_e(f^*) - n_e(f)] \le 0$ with probability one.

Appendix

Proof of Lemma 1: The predictor f^* can be described by a trellis diagram (see also [1, Appendix A]) in the following manner. Let $n_t(0)$ and $n_t(1)$ denote current counts at time t of zeros and ones, respectively. Define $C_t = |n_t(0) - n_t(1)|$ as the state and observe that every increment in C_t, i.e., a transition ($C_t = k$, $C_{t+1} = k+1$), $k > 0$, corresponds to a correct prediction of f^* and every decrement is associated with an error. The exception is $C_t = 0$ that must be followed by an increment ($C_{t+1} = 1$) whether or not the prediction at time t is correct. Assume, without loss of generality, that

$n(0) \le n(1)$. Clearly, $C_0 = 0$ and $C_n = n(1) - n(0)$. Let I and D denote the number of increments and decrements of C_t, respectively. Then, obviously $I + D = n$ and $I - D = C_n = n(1) - n(0)$, which together imply that $I = n(1)$ and $D = n(0)$. Thus, $n_e(f^*)$ involves errors associated with $D = n(0) = \min\{n(0), n(1)\}$ decrements of C_t plus errors that may occur when $C_t = 0$, which happens n^* times along the sequence. Since a fair coin is flipped whenever $C_t = 0$, then on the average $n^*/2$ additional errors appear. □

A Counterexample $\tilde{f}$: Define $R = \{\theta: \theta_0 > 1/2\}$, $B = \{\theta: \theta_1 < 1/2\}$, $Q = \{\theta: \theta_0 > 1/2 \text{ and } \theta_1 > 1/2\}$, $U = Q \cap \{\theta: \theta_0 < \theta_1\}$ and $V = Q - U$. Let $\hat{\theta}(t) = (\hat{\theta}_0(t), \hat{\theta}_1(t))$ denote the estimator of $\theta = (\theta_0, \theta_1)$ as in (18). The predictor $\tilde{f}$ is defined by $\tilde{x}_{t+1} = 1\{\hat{\theta}(t) \in Q^c\} \cdot x^*_{t+1} + 1\{\hat{\theta}(t) \in U\} \cdot x^{0,1}_{t+1} + 1\{\hat{\theta}(t) \in V\} \cdot x^{1,0}_{t+1}$ where x^*_{t+1} is as in (4), and $x^{a,b}_{t+1}$, $a, b = 0, 1$, is defined as $x^{a,b}_{t+1} = a$ when $x_t = 0$ and $x^{a,b}_{t+1} = b$ when $x_t = 1$. In other words, $\tilde{x}_{t+1}$ is the same as x^*_{t+1} as long as $\hat{\theta}(t)$ falls in the permissible domain Q^c. If $\hat{\theta}(t)$ happens to fall in Q, then the smaller between $\hat{\theta}_0(t)$ and $\hat{\theta}_1(t)$ is assumed to be in the wrong half interval and the predictor is "corrected" accordingly. Let $\theta = (\theta_0, \theta_1)$ satisfy $0 < \theta_0 < 1/2$ and $0 < \theta_1 < 1/2$. We next compare the performance of $\tilde{f}$ to that of f^* at (θ_0, θ_1) and its two reflections $(\overline{\theta}_0, \theta_1)$ and $(\theta_0, \overline{\theta}_1)$ and show that $\tilde{f}$ outperforms f^* at these three points. For θ in the lower left quarter of Θ, this result is obvious by making the two following observations. First, $r_t(f, \theta)$ depends of f only through the probability that f does not agree with the optimal predictor for that quarter (see (6), (7)), namely, the probability that $\hat{x}_{t+1} \ne 0$. Secondly, $\{\tilde{x}_{t+1} \ne 0\} \subset \{x^*_{t+1} \ne 0\}$. For $(\theta_0, \overline{\theta}_1)$ in the upper left quarter of Θ, we have from (6),

$$r_t(\tilde{f}, \theta_0, \overline{\theta}_1) = (1 - 2\theta_0) \cdot P_{\theta_0\overline{\theta}_1}\{x_t = 0, \hat{\theta}(t) \in R - U\} + (1 - 2\theta_1) \cdot P_{\theta_0\overline{\theta}_1}\{x_t = 1, \hat{\theta}(t) \in B \cup V\}, \quad \text{(A.1)}$$

while

$$r_t(f^*, \theta_0, \overline{\theta}_1) = (1 - 2\theta_0) \cdot P_{\theta_0\overline{\theta}_1}\{x_t = 0, \hat{\theta}(t) \in R\} + (1 - 2\theta_1) \cdot P_{\theta_0\overline{\theta}_1}\{x_t = 1, \hat{\theta}(t) \in B\}. \quad \text{(A.2)}$$

The difference is

$$\begin{aligned} &r_t(f^*, \theta_0, \overline{\theta}_1) - r_t(\tilde{f}, \theta_0, \overline{\theta}_1) \\ &= (1 - 2\theta_0) \cdot P_{\theta_0\overline{\theta}_0}\{x_t = 0, \hat{\theta}(t) \in U\} \\ &\quad - (1 - 2\theta_1) \cdot P_{\theta_0\overline{\theta}_1}\{x_t = 1, \hat{\theta}(t) \in V\} \\ &> (1 - 2\theta_0) \cdot \sum_{x:\, \hat{\theta}(t) \in U} P_{\theta_0\overline{\theta}_1}(x_1, \cdots, x_{t-1}) \\ &\quad \cdot P_{\theta_0\overline{\theta}_1}(x_t = 0 | x_{t-1}) \\ &\quad - (1 - 2\theta_1) \cdot P_{\theta_0\overline{\theta}_1}\{\hat{\theta}(t) \in V\} \\ &\ge (1 - 2\theta_0)\theta_1 \cdot P_{\theta_0\overline{\theta}_1}\{\hat{\theta}(t) \in U\} \\ &\quad - (1 - 2\theta_1) \cdot P_{\theta_0\overline{\theta}_1}\{\hat{\theta}(t) \in V\}. \end{aligned} \quad \text{(A.3)}$$

The large deviations theory for discrete Markov sources implies that $P_{\theta_0\overline{\theta}_1}\{\hat{\theta}(t) \in V\}$ is exponentially negligible relative to $P_{\theta_0\overline{\theta}_1}\{\hat{\theta}(t) \in U\}$, and hence that (A.3) is positive when t is sufficiently large. A similar consideration holds for $(\overline{\theta}_0, \theta_1)$ due to symmetry.

Proof of (11): The idea is to observe that at each state $x_t = x$ both the next outcome and the best prediction strategy behave like these of a Bernoulli process with a parameter θ_x. Any predictor (g_0, g_1) can be improved if g_0 is replaced by the optimal strategy for state "0", i.e., $\hat{x}_{t+1} = 1\{\theta_0 \ge 1/2\}$ while at state $x_t = 1$ the strategy g_1 remains unchanged. A straightforward application of Theorem 1a) to state "1" implies that for every value of θ_0 and for half of the values of θ_1, i.e., for half the sources, $nR_n(f, \theta) \ge c^1_1(\theta) \triangleq 0.5 \sum_{t=1}^{\infty} P_\theta\{x_t = 1, \hat{\theta}_1(t) = 1/2\}$. Interchanging the roles of state "0" and state "1" in the previous argument, we conclude similarly that for every θ_1 and for half the values of θ_0, $nR_n(f, \theta) \ge c^0_1(\theta) \triangleq 0.5 \sum_{t=1}^{\infty} P_\theta\{x_t = 0, \hat{\theta}_0(t) = 1/2\}$. Thus, when the right-hand side is replaced by $\tilde{c}_1(\theta) = \min\{c^0_1(\theta), c^1_1(\theta)\}$, the inequality holds simultaneously for at least 3/4 of the sources in Θ. □

ACKNOWLEDGMENT

The authors wish to thank P. Algoet for helpful suggestions he has provided.

REFERENCES

[1] M. Feder, N. Merhav, and M. Gutman, "Universal prediction of individual sequences," *IEEE Trans. Inform. Theory*, vol. 38, pp. 1258–1270, July 1992.
[2] 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.
[3] J. S. Vitter and P. Krishnan, "Optimal prefetching via data compression," Tech. Rep. No. CS-91-46, Dept. of Comput. Sci., Brown Univ., July 1991. (Also summarized in *Proc. FOCS-91*, pp. 121–130, 1991.)
[4] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July, 1984.
[5] ——, private communication.
[6] J. F. Hannan, "Approximation to Bayes risk in repeated plays. In contribution to theory of games," *Ann. Math. Studies*, vol. 3, no. 39, pp. 97–139, 1957.
[7] T. M. Cover, "Behavior of sequential predictors of binary sequences," in *Proc. 4th Prague Conf. Inform. Theory, Statistical Decision Functions, Random Processes*, 1965, pp. 263–272.
[8] 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, June 1977.
[9] F. Spitzer, *Principles of Random Walk.* New York: Springer-Verlag, 1976.
[10] F. T. Leighton and R. L. Rivest, "Estimating a probability using finite memory," *IEEE Trans. Inform. Theory*, vol. IT-32, pp. 733–742, Nov. 1986.
[11] R. M. Gray, *Entropy and Information Theory.* New York: Springer-Verlag, 1990.
[12] I. Csiszár and J. Korner, *Information Theory: Coding Theorems for Discrete Memoryless Systems.* New York: Academic, 1981.