On Competitive Prediction and Its Relation to Rate-Distortion Theory

Tsachy Weissman, *Member, IEEE,* and Neri Merhav, *Fellow, IEEE*

***Abstract*—Consider the normalized cumulative loss of a predictor F on the sequence $x^n = (x_1, \ldots, x_n)$, denoted $L_F(x^n)$. For a set of predictors $\mathcal{G}$, let $L(\mathcal{G}, x^n) = \min_{F \in \mathcal{G}} L_F(x^n)$ denote the loss of the best predictor in the class on x^n. Given the stochastic process $\boldsymbol{X} = X_1, X_2, \ldots$, we look at $EL(\mathcal{G}, X^n)$, termed the *competitive predictability of $\mathcal{G}$ on X^n*. Our interest is in the optimal predictor set of size M, i.e., the predictor set achieving $\min_{|\mathcal{G}| \leq M} EL(\mathcal{G}, X^n)$. When M is subexponential in n, simple arguments show that $\min_{|\mathcal{G}| \leq M} EL(\mathcal{G}, X^n)$ coincides, for large n, with the Bayesian envelope $\min_F EL_F(X^n)$. We investigate the behavior, for large n, of $\min_{|\mathcal{G}| \leq e^{nR}} EL(\mathcal{G}, X^n)$, which we term the *competitive predictability of $\boldsymbol{X}$ at rate R*. We show that whenever $\boldsymbol{X}$ has an autoregressive representation via a predictor with an associated independent and identically distributed (i.i.d.) innovation process, its competitive predictability is given by the distortion-rate function of that innovation process. Indeed, it will be argued that by viewing $\mathcal{G}$ as a rate-distortion codebook and the predictors in it as codewords allowed to base the reconstruction of each symbol on the past unquantized symbols, the result can be considered as the source-coding analog of Shannon's classical result that feedback does not increase the capacity of a memoryless channel. For a general process $\boldsymbol{X}$, we show that the competitive predictability is lower-bounded by the Shannon lower bound (SLB) on the distortion-rate function of $\boldsymbol{X}$ and upper-bounded by the distortion-rate function of any (not necessarily memoryless) innovation process through which the process $\boldsymbol{X}$ has an autoregressive representation. Thus, the competitive predictability is also precisely characterized whenever $\boldsymbol{X}$ can be autoregressively represented via an innovation process for which the SLB is tight. The error exponent, i.e., the exponential behavior of $\min_{|\mathcal{G}| \leq \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right)$, is also characterized for processes that can be autoregressively represented with an i.i.d. innovation process.**

***Index Terms*—Competitive prediction, error exponents, prediction with expert advice, rate distortion theory, redundancy, scandiction, strong converse.**

I. Introduction

THE problem of universal prediction, in both its deterministic setting (cf., e.g., [9], [12], [17], [25], [15]) and stochastic setting (cf. [1] and references therein), typically involves the construction of a predictor whose goal is to compete with a given comparison class $\mathcal{G}$ of predictors ("experts") in the sense of approaching the performance of the best predictor in the class $L(\mathcal{G}, x^n)$ whatever the data sequence x^n turns out to be. In the deterministic setting, the comparison class may represent a set of different approaches, or a set of prediction schemes which are limited in computational resources (cf. [26, Sec. I] for a broader discussion). In the stochastic setting, the comparison class typically consists of those predictors which are optimal for the sources with respect to which universality is sought.

In the choice of a reference class of predictors with which to compete, there is generally a tradeoff between the size of the class and the redundancy attainable relative to it. In the stochastic setting of universal prediction it is common practice, and often advantageous, that rather than taking the class of predictors corresponding to all sources in the uncertainty set, one takes a more limited class of "representatives" with the hope that the reduced redundancy will compensate for those sources that are not exactly covered. On the other extreme of this tradeoff, one takes a reference class richer than the class of sources, allotting more than one predictor to be specialized for each source, at the price of a higher redundancy. There are two issues which arise in this context. The first concerns the question of whether there is a size for a reference class which is in any sense optimal, when the dilemma is between a rich "cover" of the typical sequences one is going to encounter, and the redundancy which increases with the increase in the size of the set. The second concerns the following question. For a given size, what is the optimal predictor set? The first question has been extensively studied in the context of universal prediction (cf. [17] and reference therein), coding, and estimation (e.g., [20], [4], and references thereto and therein). The latter question is the motivation for our work.

It should be emphasized that in the prediction problem in the literature, especially that pertaining to computational learning theory (e.g., [23], [24], [9], [8], [7], [10], [15], and references therein), where the problem is formulated in terms of learning with expert advice, the class of experts is always assumed given and the questions typically asked concern optimal strategies per the given class. In such problems, one is not concerned with the question of *how* to choose the reference class of experts. Nevertheless, it is understood in such problems that there is no point in letting two experts be too similar and that, rather, an appropriate set should be chosen to efficiently cover the possible sequences one is likely to encounter. Our goal in this work is to gain some insight regarding the considerations for the choice of the expert class through an analysis of this problem in the probabilistic setting.

To answer this question, we shall turn to the most basic, nonuniversal setting and address the following problem: given

Manuscript received February 26, 2002; revised August 5, 2003. Most of this work was performed while T. Weissman was with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa.

T. Weissman is with the Department of Electrical Engineering, Stanford University, Stanford, CA 94305 USA (e-mail: tsachy@stanford.edu).

N. Merhav is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: merhav@ee.technion.ac.il).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Digital Object Identifier 10.1109/TIT.2003.820014

a probabilistic source sequence $X_1, \ldots, X_n$ and M, what is the predictor set of size M which is, in some sense, "best" for this source? In the problem we pose here, the object of interest which we seek to optimize is the predictor *set*. This problem will turn out to be intimately related with rate-distortion theory [13], [5] and, in a sense to be made precise, can be viewed as "rate distortion coding with feedback."

A brief account of the main gist of this work is as follows. Let

$$L_F(x^n) = \frac{1}{n}\sum_{t=1}^{n} \rho(x_t - F_t(x^{t-1}))$$

denote the normalized cumulative loss of the predictor F on the sequence $x^n = (x_1, \ldots, x_n)$, where ρ is a given loss function. For a predictor set $\mathcal{G}$, let further $L(\mathcal{G}, x^n) = \min_{F\in\mathcal{G}} L_F(x^n)$ denote the loss of the best predictor in the class on x^n. Given the stochastic process $\boldsymbol{X} = X_1, X_2, \ldots$, we look at $EL(\mathcal{G}, X^n)$, which will be termed the *competitive predictability of* $\mathcal{G}$ *on* X^n. Our interest is in the optimal predictor set of size M, i.e., the predictor set achieving $\min_{|\mathcal{G}|\le M} EL(\mathcal{G}, X^n)$. When M is subexponential in n, simple arguments will show that $\min_{|\mathcal{G}|\le M} EL(\mathcal{G}, X^n)$ coincides, for large n, with $\min_F EL_F(X^n)$, which, by classical results on prediction, is characterized by the Bayesian envelope of the process. When M grows exponentially in n, however, the problem significantly diverges from its classical origin. Thus, our interest is in the behavior, for large n, of $\min_{|\mathcal{G}|\le e^{nR}} EL(\mathcal{G}, X^n)$, which we term the *competitive predictability of* $\boldsymbol{X}$ *at rate* R. It will be shown that whenever $\boldsymbol{X}$ has an autoregressive representation with an associated independent and identically distributed (i.i.d.) innovation process, its competitive predictability is given by the distortion-rate function of the innovation process. For a general stochastic process $\boldsymbol{X}$, we find that its competitive predictability is lower-bounded by the Shannon lower bound (SLB) on its distortion-rate function. On the other hand, the competitive predictability of a process $\boldsymbol{X}$ will be argued to be upper-bounded by the distortion-rate function of the (not necessarily memoryless) innovation process of any autoregressive representation of $\boldsymbol{X}$. Various implications of these results will also be pointed out. In particular, we shall discuss an alternative viewpoint of the competitive predictability setting as a problem of rate-distortion theory with feedback. In this context, the fact that the competitive predictability is lower-bounded by the distortion-rate function of the source (in the generality mentioned above), can be thought of as the source-coding analog of Shannon's classical result that feedback does not increase the capacity of a memoryless channel [22]. The source-coding analog we present here, however, holds true much more generally than for the class of memoryless sources. We shall also characterize the error exponents for this problem, namely, the exponential behavior of $\min_{|\mathcal{G}|\le \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right)$. Specifically, it will be seen that when the source has an autoregressive representation with i.i.d. innovations, this exponent coincides with the Marton's exponent [16] of the innovations process.

The remainder of this work is organized as follows. Section II will be dedicated to some notation, conventions and preliminaries. In Section III, we shall formulate the problem and present our main results: Section III-A will formally define the problem and Section III-B will present the main results. Section IV will be dedicated to the proofs of the main results: Section IV-A will present proofs of the main results that are based on information-theoretic arguments and Section IV-B will present an alternative route to arrive at the results via "counting" and volume-preservation arguments. Section V will further discuss the main results and their implications. Finally, in Section VI, we summarize this work and discuss some related directions for future research.

II. Notation, Conventions, and Preliminaries

Throughout, $\mathcal{X}$ will denote the source alphabet where addition and subtraction of elements are assumed well defined. Unless otherwise indicated, $\mathcal{X}$ will also be assumed finite (in which case, addition and subtraction are normally defined modulo the alphabet size). Stochastic processes will be assumed to have $\mathcal{X}$-valued components. When considering processes with real-valued components, it will always be assumed that, for each n, the distribution of X^n is absolutely continuous with respect to the Lebesgue measure and has a finite differential entropy.

Assume throughout a fixed loss function $\rho : \mathcal{X} \to [0, \infty)$. For $P \in \mathcal{M}(\mathcal{X})$, we let $H(P)$ denote the associated entropy. Define

$$\phi(d) \stackrel{\text{def}}{=} \sup_{P: E_P\rho(Z)\le d} H(P) \tag{1}$$

where $E_P\rho(Z)$ denotes expectation when $Z \sim P$. The function $\phi(d)$ defined in (1) is well known (cf., e.g., [17], [18]) to be monotone, concave, and have a closed-form representation. Specifically, for finite $\mathcal{X}$, let

$$\lambda(\beta) = -\log\left[\sum_{x\in\mathcal{X}} e^{-\beta\rho(x)}\right], \qquad \beta > 0 \tag{2}$$

denote the log-moment generating function associated with the loss function ρ. Then, ϕ is given by the one-sided Fenchel–Legendre transform of λ

$$\phi(d) = \inf_{\beta>0}[\beta d - \lambda(\beta)], \qquad d \ge 0. \tag{3}$$

The significance of the function $\phi(d)$ is that it conveys the precise exponential behavior of the size of the n-dimensional ρ-ball (cf., e.g., [18])

$$\phi_n(d) \stackrel{\text{def}}{=} \frac{1}{n}\log\left|\left\{x^n \in \mathcal{X}^n : \frac{1}{n}\sum_{i=1}^{n}\rho(x_i) \le d\right\}\right| \stackrel{n\to\infty}{\longrightarrow} \phi(d). \tag{4}$$

We shall let ϕ^{-1} denote the inverse function of ϕ, defined (even when ϕ is not strictly increasing) by

$$\phi^{-1}(\alpha) = \sup\{d : \phi(d) < \alpha\}.$$

When $\mathcal{X} = \mathbb{R}$, the summation in (2) is replaced by integration and ρ will be assumed to be sufficiently steep such that the integral exists and is finite for all $\beta > 0$. For this case, ϕ is defined as in (3) and the analog of (4) for this case is

$$\phi_n(d) \stackrel{\text{def}}{=} \frac{1}{n}\log \mathrm{Vol}\left\{x^n \in \mathbb{R}^n : \frac{1}{n}\sum_{i=1}^{n}\rho(x_i) \le d\right\} \stackrel{n\to\infty}{\longrightarrow} \phi(d). \tag{5}$$

Throughout, capital letters will denote random variables while the respective lower case letters will denote individual sequences or specific sample values. For probability measures P and Q on $\mathcal{X}$, we let $R(P,\cdot)$ denote the rate-distortion function associated with P under the distortion measure ρ and $D(P\|Q)$ denote the Kullback–Leibler divergence between P and Q. Finally, we shall work throughout in nats so all the mentioned information-theoretic functionals are assumed to be defined with natural logarithms.

III. Competitive Predictability

A. Competitive Predictability Defined

A predictor G is a sequence of mappings $G = \{G_t\}_{t\geq 1}$, where $G_t : \mathcal{X}^{t-1} \to \mathcal{X}$. Let $\mathcal{F}$ denote the set of all predictors. For a loss function (difference distortion measure) $\rho : \mathcal{X} \to [0,\infty)$ and a sequence $x^n = (x_1,\ldots,x_n)$, denote the normalized cumulative loss of the predictor G by

$$L_G(x^n) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n}\rho(x_t - G_t(x^{t-1})). \tag{6}$$

For a predictor set $\mathcal{G} \subseteq \mathcal{F}$, define the *competitive predictability of $\mathcal{G}$ on x^n* by

$$L(\mathcal{G},x^n) \stackrel{\text{def}}{=} \min_{G\in\mathcal{G}} L_G(x^n) \tag{7}$$

i.e., the loss of the best predictor in $\mathcal{G}$ for the sequence x^n.

Suppose now that $\boldsymbol{X} = X_1, X_2, \ldots$ is a stochastic process and that, for a given M and sequence length n, we seek to minimize the competitive predictability over all predictor sets of size at most M. Since for any $\mathcal{G} \subseteq \mathcal{F}$, $L(\mathcal{G},X^n)$ is a random variable, a natural goal here would be to minimize the expected competitive predictability. In other words, we are interested in

$$\min_{|\mathcal{G}|\leq M} EL(\mathcal{G},X^n) \tag{8}$$

and in the predictor set $\mathcal{G}$ achieving it. Note, however, that for any given set of predictors $\mathcal{G}$, the theory of universal prediction guarantees the existence of a (possibly randomized) predictor F such that

$$\sup_{x^n}[EL_F(x^n) - L(\mathcal{G},x^n)] \leq K\sqrt{\frac{\log|\mathcal{G}|}{n}} \tag{9}$$

where K is a constant depending only on the loss function ρ and the alphabet $\mathcal{X}$ (and the expectation on the left-hand side of (9) is with respect to the randomization). A predictor F satisfying (9) can always be constructed via the exponential weighting approach (cf., e.g., [9], [10]). It follows from (9) that if M is subexponential in n, then $\min_{|\mathcal{G}|\leq M} EL(\mathcal{G},X^n)$ is asymptotically equivalent to $\min_{F\in\mathcal{F}} EL_F(X^n)$, where the latter is the "Bayesian envelope" of the classical problem of optimal prediction in the stochastic setting (cf. [1], [2], and references therein). Thus, the quantity in (8) can become interesting and significantly deviate from the classical optimal prediction problem only when M grows exponentially in n. This is the motivation for focusing on the case where $M = e^{nR}$, $R > 0$.

B. Main Results

For a stochastic process $\boldsymbol{X}$, let

$$\underline{H}(\boldsymbol{X}) \stackrel{\text{def}}{=} \liminf_{n\to\infty}\frac{1}{n}H(X^n) \tag{10}$$

where $H(X^n)$ on the right-hand side is the entropy of the random vector X^n and

$$D(\boldsymbol{X},R) \stackrel{\text{def}}{=} \limsup_{n\to\infty} D(X^n,R) \tag{11}$$

where the right-hand side is the distortion-rate function associated with X^n. For any predictor F, we let $\boldsymbol{W}^F = W_1^F, W_2^F, \ldots$ denote the process defined by

$$W_t^F = X_t - F_t(X^{t-1}). \tag{12}$$

We shall say, in this case, that *$\boldsymbol{X}$ has an autoregressive representation via the predictor F and innovation process*[1] *$\boldsymbol{W}^F$*.

Theorem 1: Let $\boldsymbol{X}$ be a stochastic process for which there exists a predictor F such that $\boldsymbol{W}^F$ is an i.i.d. process. Then, for $R \geq 0$

$$\lim_{n\to\infty}\left[\min_{|\mathcal{G}|\leq\exp(nR)} EL(\mathcal{G},X^n)\right] = D(\boldsymbol{W}^F,R).$$

In words, when $\boldsymbol{X}$ can be "whitened" by some predictor, its competitive predictability is given by the distortion-rate function of the corresponding innovation process. The proof of Theorem 2 below will be seen to imply that for the case where $\boldsymbol{X}$ itself is an i.i.d. process (and then $F_t(X^{t-1}) = 0$, so $\boldsymbol{W}^F = \boldsymbol{X}$), Theorem 1 holds for a general, not necessarily difference, distortion measure $\rho(X_t, G_t(X^{t-1}))$. For a general source, we have the following.

Theorem 2: Let $\boldsymbol{X} = X_1, X_2, \ldots, X_i \in \mathcal{X}$, be any stochastic process and $R \geq 0$. Then

$$\liminf_{n\to\infty}\left[\min_{|\mathcal{G}|\leq\exp(nR)} EL(\mathcal{G},X^n)\right] \geq \phi^{-1}(\underline{H}(\boldsymbol{X}) - R) \tag{13}$$

and

$$\limsup_{n\to\infty}\left[\min_{|\mathcal{G}|\leq\exp(nR)} EL(\mathcal{G},X^n)\right] \leq \inf_{F\in\mathcal{F}} D(\boldsymbol{W}^F,R). \tag{14}$$

It is easy to show (cf., e.g., [18]) that the entropy (rate) of a process equals the entropy (rate) of the associated innovation process relative to any predictor. Thus, given a process $\boldsymbol{X}$ and any predictor F, $\underline{H}(\boldsymbol{X}) = \underline{H}(\boldsymbol{W}^F)$. Recall that the SLB (cf. [5], [11], [21]) on the rate-distortion function of the source $\boldsymbol{X}$ at distortion level d is $\underline{H}(\boldsymbol{X}) - \phi(d)$. Furthermore, since the SLB is dependent on the source only through its entropy, it is clear that $R(\boldsymbol{W}^F,d) \geq \underline{H}(\boldsymbol{X}) - \phi(d)$ for all F or, in terms of the distortion-rate function

$$D(\boldsymbol{W}^F,R) \geq \phi^{-1}(\underline{H}(\boldsymbol{X}) - R) \qquad \forall F. \tag{15}$$

Note that (15) follows from Theorem 2 as well since the left-hand side is shown to be achievable while the right-hand side is a lower bound on the achievable loss. In this context, we make the following observation.

Corollary 3: Let $\boldsymbol{X}$ be a stochastic process and suppose there exists a predictor F such that $D(\boldsymbol{W}^F,R)$ meets the SLB with equality. Then

$$\lim_{n\to\infty}\left[\min_{|\mathcal{G}|\leq\exp(nR)} EL(\mathcal{G},X^n)\right] = D(\boldsymbol{W}^F,R). \tag{16}$$

[1]Note that in our definition, the innovation process can be a general process and, in particular, its components may not be independent.

Corollary 3 tells us that whenever the process $\boldsymbol{X}$ has an autoregressive representation via any predictor F, i.e., $X_t = F_t(X^{t-1}) + W_t$, the attainable limitation on its competitive prediction is given by the rate-distortion function of $\{W_t\}$ even when it is not an i.i.d. process, provided it satisfies the SLB with equality. In Section V, we will mention a few examples of families of sources for which the SLB is tight.

Returning to the case where the process $\boldsymbol{X}$ has an autoregressive representation with i.i.d. innovations, we have the following characterization of the error exponents for competitive prediction.

Theorem 4: Let $\boldsymbol{X}$ be a stochastic process for which there exists a predictor F such that $\boldsymbol{W}^F$ is an i.i.d. process with a marginal distribution Q. For $R > 0$

$$F_d(R-0) \le \liminf_{n\to\infty} \left[-\frac{1}{n} \log \min_{|\mathcal{G}|\le \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right) \right] \tag{17}$$

$$\le \limsup_{n\to\infty} \left[-\frac{1}{n} \log \min_{|\mathcal{G}|\le \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right) \right] \le F_d(R+0) \tag{18}$$

where $F_d(R) \stackrel{\text{def}}{=} \min_{P: R(P,d)\ge R} D(P\|Q)$ is Marton's error exponent function for the source Q [16].

Finally, we mention that the finite-alphabet assumption is made to minimize the technical details in the proofs. It will be clear, however, by examination of the proofs, that Theorem 1 and Theorem 2 (and hence also Corollary 3) are valid also in the case where the alphabet is the real line under standard mild regularity assumptions. For example, the following result is a consequence of the continuous-alphabet version of Theorem 1 (or of Corollary 3 when the distortion measure is squared error).

Corollary 5: Let $\boldsymbol{X}$ be a stationary Gaussian process and let $\sigma^2 = \exp\left\{\frac{1}{2\pi}\int_{-\pi}^{\pi} \ln f(\lambda) d\lambda\right\}$, where f is the density function associated with the absolutely continuous component in the Lebesgue decomposition of its spectral measure. Then

$$\lim_{n\to\infty} \left[\min_{|\mathcal{G}|\le \exp(nR)} EL(\mathcal{G}, X^n) \right] = D(\sigma^2, R) \tag{19}$$

with $D(\sigma^2, R)$ on the right-hand side of (19) denoting the distortion-rate function of a zero-mean Gaussian with variance σ^2 (under the given distortion measure).

We will also indicate in Section III-B how Theorem 4 can be extended to the continuous-alphabet setting. For the Gaussian case this will be shown to lead to the following.

Corollary 6: Let $\boldsymbol{X}$ be a stationary Gaussian process as in Corollary 5. Then, under squared error distortion

$$\lim_{n\to\infty} \left[-\frac{1}{n} \log \min_{|\mathcal{G}|\le \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right) \right] = \begin{cases} \frac{1}{2}\left[\log\left(\frac{\sigma^2}{2\cdot e^{2R} d}\right) + \frac{e^{2R} d}{2\sigma^2}\right], & \text{if } d > \sigma^2 e^{-2R} \\ 0, & \text{otherwise.} \end{cases} \tag{20}$$

Corollary 6 is a consequence of Paley–Wiener theory which guarantees that when $\sigma^2 > 0$, $\boldsymbol{X}$ can be represented as the response of a causal (minimum phase, invertible) filter to its innovation process which is white noise, and, by Gaussianity, also i.i.d. In particular, the case $R = 0$ in Corollary 6 characterizes the best attainable large deviations performance in prediction of Gaussian processes, a result which appears to be new (cf. discussion in [19]).

The two subsections in the following section are dedicated to proving the above results. In Section IV-A we give information-theoretic proofs, hinging upon a view of the problem as "rate-distortion with feedback," a viewpoint that will be expounded in Section V. Section IV-B describes an alternative approach to the derivation of the results via counting or "volume-preservation" arguments. While the information-theoretic argumentation leads, in general, to stronger results, the merit of the latter approach is that it gives geometric intuition for the converse parts of the results, an intuition generally lacking in information-theoretic proofs.

Finally, we remark in the context of the "rate-distortion with feedback viewpoint" that Theorem 4 can be interpreted as saying that feedback does not improve rate-distortion performance even in the error-exponent sense, in contrast to the channel-coding problem where feedback is known to improve the error-exponent (cf., e.g., [13, Example 5.14]).[2]

IV. Proofs

A. Information Theoretic Proofs

In this subsection, we give information-theoretic proofs to the main results.

Proof of Theorem 1: The direct part follows from the upper bound in Theorem 2 (which we prove subsequently). For the converse, fix n and let $\boldsymbol{W}^F = (W_1, W_2, \ldots)$ be the i.i.d. $\sim Q$ innovation vector so that $X_t = F_t(X^{t-1}) + W_t$, $1 \le t \le n$. In this proof, let $R(\cdot)$, $D(\cdot)$ denote the rate-distortion function associated with Q and its inverse (for the difference distortion measure ρ). X^t is completely determined by W^t (and *vice versa*), hence we shall write $X^t(W^t)$ when we want to make this dependence explicit. Fix further a predictor set $\mathcal{G} = \{G^{(1)}, G^{(2)}, \ldots, G^{(\lfloor e^{nR} \rfloor)}\}$ and define the integer random variable

$$J = J(X^n) = \arg \min_{1\le i\le e^{nR}} L_{G^{(i)}}(X^n)$$

(resolving ties arbitrarily) so that, in particular, $L_{G^{(J)}}(X^n) = L(\mathcal{G}, X^n)$. Finally, for $1 \le i \le e^{nR}$ and $1 \le t \le n$, define $g_t(\cdot,\cdot)$ by

$$g_t(W^{t-1}, i) = G_t^{(i)}(X^{t-1}(W^{t-1})) - F_t(X^{t-1}(W^{t-1})).$$

Then

$$\begin{aligned} nR &\ge H(J) \\ &\ge I(X^n; J) \\ &= H(X^n) - H(X^n|J) \\ &= \sum_{t=1}^{n} [H(X_t|X^{t-1}) - H(X_t|X^{t-1}, J)] \end{aligned}$$

[2]Of course, the duality between channel coding with feedback and our setting is merely a qualitative one. The error criterion in the former setting is the block probability of error while in our setting it is the cumulative per-letter distortion, so technically the two problems are quite different.

$$
\begin{aligned}
&= \sum_{t=1}^{n} [H(F_t(X^{t-1}) + W_t | X^{t-1}) \\
&\quad - H(F_t(X^{t-1}) + W_t | X^{t-1}, J)] \\
&= \sum_{t=1}^{n} [H(W_t | X^{t-1}) - H(W_t | X^{t-1}, J)] \\
&= \sum_{t=1}^{n} [H(W_t) - H(W_t | W^{t-1}, J)] \qquad (21) \\
&= \sum_{t=1}^{n} I(W_t; W^{t-1}, J) \\
&\geq \sum_{t=1}^{n} I(W_t; g_t(W^{t-1}, J)) \\
&\geq \sum_{t=1}^{n} R\left(E\rho\left(W_t - g_t(W^{t-1}, J)\right)\right) \\
&= \sum_{t=1}^{n} R\left(E\rho\left(X_t - G_t^{(J)}(X^{t-1})\right)\right) \\
&\geq nR\left(\frac{1}{n}\sum_{t=1}^{n} E\rho\left(X_t - G_t^{(J)}(X^{t-1})\right)\right) \\
&= nR\left(EL_{G^{(J)}}(X^n)\right) \\
&= nR\left(EL(\mathcal{G}, X^n)\right) \qquad (22)
\end{aligned}
$$

where (21) follows by the independence of W_t and X^{t-1}, and by the one-to-one correspondence between X^n and W^n. The remaining equalities and inequalities follow from standard information-theoretic identities, from the fact that $X_t = F_t(X^{t-1}) + W_t$, and from convexity. Consequently, $EL(\mathcal{G}, X^n) \geq D(R)$, completing the proof by the arbitrariness of $\mathcal{G}$. □

Proof of Lower Bound in Theorem 2: Let $J = J(X^n)$ be defined as in the preceding proof. Then

$$
\begin{aligned}
nR &\geq H(X^n) - H(X^n|J) \qquad (23) \\
&= H(X^n) - \sum_{t=1}^{n} H(X_t | X^{t-1}, J) \\
&\geq H(X^n) - \sum_{t=1}^{n} H(X_t | G_t^{(J)}(X^{t-1})) \\
&= H(X^n) - \sum_{t=1}^{n} H(X_t - G_t^{(J)}(X^{t-1}) | G_t^{(J)}(X^{t-1})) \\
&\geq H(X^n) - \sum_{t=1}^{n} H(X_t - G_t^{(J)}(X^{t-1})) \\
&\geq H(X^n) - \sum_{t=1}^{n} \phi(E\rho(X_t - G_t^{(J)}(X^{t-1}))) \qquad (24) \\
&\geq H(X^n) - n\phi\left(\frac{1}{n}\sum_{t=1}^{n} E\rho(X_t - G_t^{(J)}(X^{t-1}))\right) \qquad (25) \\
&= H(X^n) - n\phi\left(EL(\mathcal{G}, X^n)\right) \qquad (26)
\end{aligned}
$$

where (23) follows as in the first few lines of (22), (24) follows from the definition of ϕ, and (25) follows from its concavity. Since $\mathcal{G}$ is an arbitrary predictor set of size $\leq e^{nR}$ we obtain

$$\min_{|\mathcal{G}| \leq \exp(nR)} EL(\mathcal{G}, X^n) \geq \phi^{-1}(H(X^n) - R) \qquad (27)$$

which implies (13). □

Proof of the Upper Bound in Theorem 2: Fix a predictor F and $R \geq 0$. To any $\hat{w}^n \in \mathcal{X}^n$, we associate the predictor G specified by

$$G_t(x^{t-1}) = F_t(x^{t-1}) + \hat{w}_t. \qquad (28)$$

In this way, for any $\mathcal{C}_n \subseteq \mathcal{X}^n$, we can look at the predictor set $\mathcal{G}_n$ consisting of the predictors associated with the members of $\mathcal{C}_n$. We shall refer to $\mathcal{G}_n$ as the predictor set induced by the codebook $\mathcal{C}_n$ and the predictor F. By the definition of $\boldsymbol{W}^F$ (recall (12)) it follows that for any $\mathcal{C}_n$, with probability 1

$$\min_{\hat{w}^n \in \mathcal{C}_n} \frac{1}{n}\sum_{t=1}^{n} \rho(W_t^F - \hat{w}_t) = L(\mathcal{G}_n, X^n) \qquad (29)$$

where the $\mathcal{G}_n$ on the right-hand side is that induced by $\mathcal{C}_n$ and F. rate-distortion theory guarantees (cf. [5], [14]) the existence of a sequence of codebooks $\{\mathcal{C}_n\}$, $\mathcal{C}_n \subseteq \mathcal{X}^n$ with $|\mathcal{C}_n| \leq e^{nR}$, for which

$$\limsup_{n\to\infty} E\left[\min_{\hat{w}^n \in \mathcal{C}_n} \frac{1}{n}\sum_{t=1}^{n} \rho(W_t^F - \hat{w}_t)\right] \leq D(\boldsymbol{W}^F, R). \qquad (30)$$

Let now $\{\mathcal{G}_n\}$ be the predictor sets induced, respectively, by the codebooks $\{\mathcal{C}_n\}$ that satisfy (30). The predictor sets $\{\mathcal{G}_n\}$ satisfy $|\mathcal{G}_n| \leq e^{nR}$ and, by (29) and (30), also

$$\limsup_{n\to\infty} EL(\mathcal{G}_n, X^n) \leq D(\boldsymbol{W}^F, R) \qquad (31)$$

implying (14) by the arbitrariness of F. □

Proof of Theorem 4: The direct part (inequality (17)) follows similarly to the direct part of Theorem 2. This time one takes predictor sets induced by $\{\mathcal{C}_n\}$ and F, where $\{\mathcal{C}_n\}$ are optimal codebooks in Marton's error exponent sense [16] for the i.i.d. source $\boldsymbol{W}^F$. The proof of the converse part, equipped with Theorem 1, is analogous to the proof of the converse in [16]. Its outline is as follows. Fix a $P \in \mathcal{M}(\mathcal{X})$ such that $R < R(P, d)$, a predictor set $\mathcal{G}$ with $|\mathcal{G}| \leq e^{nR}$, and $\delta > 0$. Consider the set

$$A = \{w^n : L(\mathcal{G}, X^n(w^n)) > d\} \qquad (32)$$

where $X^n(\cdot)$ in (32) denotes the (one-to-one) transformation taking the innovation sequence into the source sequence under the predictor F. Define further[3]

$$B = \left\{w^n : \left|\frac{1}{n}\log\frac{P(w^n)}{Q(w^n)} - D(P\|Q)\right| \leq \delta\right\}. \qquad (33)$$

Since $R < R(P, d)$, the converse part of Theorem 1 implies that

$$P(A) \geq \alpha(P, d, R) > 0. \qquad (34)$$

On the other hand, by the law of large numbers, for all large n

$$P(B) \geq 1 - \alpha(P, d, R)/2 \qquad (35)$$

which, with (34), implies

$$P(A \cap B) \geq \alpha(P, d, R)/2. \qquad (36)$$

[3]With a slight abuse of notation we write $P(w^n), Q(w^n)$ for $\prod_{i=1}^{n} P(w_i), \prod_{i=1}^{n} Q(w_i)$, respectively.

Hence, for large enough n

$$\begin{aligned} Q(A) &\geq Q(A\cap B) && (37)\\ &= \sum_{w^n\in A\cap B} P(w^n)\frac{Q(w^n)}{P(w^n)} && (38)\\ &\geq \sum_{w^n\in A\cap B} P(w^n)e^{-n(D(P\|Q)+\delta)} && (39)\\ &= P(A\cap B)e^{-n(D(P\|Q)+\delta)} && (40)\\ &\geq \frac{1}{2}\alpha(P,d,R)e^{-n(D(P\|Q)+\delta)} && (41)\end{aligned}$$

implying the result by the arbitrariness of $|\mathcal{G}| \leq e^{nR}$, of P (satisfying $R < R(P,d)$), and of $\delta > 0$. □

B. Counting and Volume-Preservation Arguments

An alternative route to the derivation of some of the results is through the use of "counting" arguments, which were recently applied to obtain results for the scandiction problem in [18]. Applied to our present setting, the idea is as follows. Every predictor F defines a one-to-one correspondence from $\mathcal{X}^n$ into itself by mapping $x^n \in \mathcal{X}^n$ into $e^n \in \mathcal{X}^n$ according to

$$e_t = x_t - F_t(x^{t-1}), \qquad t = 1,\ldots,n.$$

To see this, one must simply notice that given e^n and F, x^n can be uniquely recovered.[4] Since $L_F(x^n) = \frac{1}{n}\sum_{i=1}^n \rho(e_i)$, where e^n on the right-hand side is the error sequence associated with x^n and F, it follows that for any F and $d \geq 0$

$$|\{x^n : L_F(x^n)\leq d\}| = \left|\left\{e^n : \frac{1}{n}\sum_{i=1}^n \rho(e_i)\leq d\right\}\right| \stackrel{\text{def}}{=} e^{n\phi_n(d)}. \tag{42}$$

Consequently, for any set of predictors $\mathcal{G}\subseteq\mathcal{F}$

$$\begin{aligned}|\{x^n : L(\mathcal{G},x^n)\leq d\}| &= \left|\bigcup_{F\in\mathcal{G}}\{x^n : L_F(x^n)\leq d\}\right|\\ &\leq \sum_{F\in\mathcal{G}}|\{x^n : L_F(x^n)\leq d\}| = |\mathcal{G}|e^{n\phi_n(d)}. \end{aligned}\tag{43}$$

The fact that $\phi_n(d)\to\phi(d)$, combined with (43), leads to the following conclusions.

1) If $R+\phi(d) < \underline{H}(\boldsymbol{X})$, then for large n and any predictor set $\mathcal{G}$ with $|\mathcal{G}| \leq e^{nR}$, the set $|\{x^n : L(\mathcal{G},x^n) \leq d\}|$ is exponentially smaller than $e^{n\underline{H}(\boldsymbol{X})}$ and, hence, by the converse to the asymptotic equipartition property (AEP), $\Pr(L(\mathcal{G},X^n)\leq d)$ is very (exponentially) small. In simple words, the set $\mathcal{G}$ is too small to cover the set of typical sequences of X^n.

2) Suppose that the process $\boldsymbol{X}$ has been autoregressively generated via the predictor F and the i.i.d. innovation process $\boldsymbol{W}^F$. For large n, any predictor set $\mathcal{G}$ with $|\mathcal{G}| \leq e^{nR}$ and any probability measure P on $\mathcal{X}$, if $R+\phi(d) < H(P)$, then "most" of the innovations sequences with empirical measure (close to) P are such that the source sequence that they generate lies outside the set $\{x^n : L(\mathcal{G},x^n)\leq d\}$. This is because the size of the set of sequences with empirical measure close to P (and, by the above observations, the size of the set of source sequences generated by these innovations sequences) is $\approx e^{nH(P)}$ which, by (43), is exponentially more than $|\{x^n : L(\mathcal{G},x^n)\leq d\}|$ whenever $R+\phi(d) < H(P)$. Thus, $\Pr(L(\mathcal{G},X^n)>d)$ is essentially lower-bounded by the probability that the innovations vector W^n will be P-typical, namely, $\approx e^{-nD(P\|Q)}$ (Q being the marginal distribution of the innovations process).

The observation made in the first of the above items leads to the converse part of Theorem 2 , while the second observation leads to the following upper bound on the error exponent of Theorem 4 (under the assumption of Theorem 4):

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log\min_{|\mathcal{G}|\leq\exp(nR)}\Pr(L(\mathcal{G},X^n)>d)\right] \leq \min_{P:H(P)-\phi(d)\geq R} D(P\|Q). \tag{44}$$

Appendix A contains the rigorous proofs based on this argumentation, as well as a discussion of a few examples where the upper bound in (44) is tight.

We mention that the counting arguments previously described carry over to the continuous case where $\boldsymbol{X}=\mathbb{R}$ by replacing the counting arguments with volume-preservation ones. Specifically, suppose that $\mathcal{X}=\mathbb{R}$, fix an arbitrary predictor[5] $F\in\mathcal{F}$, and consider the transformation taking the source sequence x^n into the sequence of prediction errors

$$e^n = (x_1 - F_1, x_2 - F_2(x_1),\ldots,x_n - F_n(x^{n-1})).$$

As in the finite-alphabet setting, this transformation can be seen to be one-to-one and onto. Furthermore, assuming first that the mappings F_t defining the predictor are continuously differentiable, the Jacobian matrix of this transformation is readily seen to be lower-triangular with diagonal entries all equal to 1, implying that this mapping is volume preserving. In fact, for *any* predictor, not necessarily consisting of continuously differentiable functions, Tamás Linder has shown the authors a proof that this volume-preservation property holds. Thus, we get the analog of (42) (cf. [18, Theorem 5])

$$\begin{aligned}&\mathrm{Vol}(\{x^n : L_F(x^n)\leq d\})\\ &\quad= \mathrm{Vol}\left(\left\{e^n : \frac{1}{n}\sum_{i=1}^n\rho(e_i)\leq d\right\}\right) \stackrel{\text{def}}{=} e^{n\phi_n(d)}\end{aligned}\tag{45}$$

leading to the analog of (43)

$$\mathrm{Vol}(\{x^n : L(\mathcal{G},x^n)\leq d\}) \leq |\mathcal{G}|e^{n\phi_n(d)}. \tag{46}$$

The proof ideas described above carry over by replacing throughout (4) with (5), (43) with (46), $|\cdot|$ with $\mathrm{Vol}(\cdot)$, and entropies with differential entropies.

[4] This is the idea on which predictive coding techniques are based.

[5] In this case, a predictor F is a sequence $\{F_t\}$ of measurable functions where $F_t : \mathbb{R}^{t-1}\to\mathbb{R}$.

V. Discussion of Results

A. Distortion-Rate Source Coding With Perfect Past Feedback

The information-theoretic proofs of the main results, as presented in Section III-A, suggest an alternative viewpoint of competitive predictability as characterizing optimum performance in the following lossy source coding scenario: Suppose we wish to store the data $(X_1, \ldots, X_n)$ (with distortion) in our computer and the memory at our disposal is nR bits. Suppose further that we are required to give the reconstructed symbol $\hat{X}_1$ by January 1st (2004), $\hat{X}_2$ by January 2nd, and so forth. We know, however, that the original data $(X_1, \ldots, X_n)$ are going to be posted on the Internet (to which our computer has access), a little while later, say, starting January 2nd, one new symbol every day. The question is: how should we use our available computer memory so that the overall distortion of the reconstructed symbols is minimized? Our results (Theorem 1 and Corollary 3 , respectively) imply that, when $\boldsymbol{X}$ is an i.i.d. source, or is a source attaining the SLB with equality, there is nothing to gain from the *perfect* (as opposed to quantized) observations of the past source sequence. It may seem natural, at first glance, at least for the i.i.d. case, that there is nothing to gain from observing the past sequence for reconstruction of the present symbol. This observation is, however, rather surprising in the context of Shannon theory which tells us that other sequence components are very relevant for the coding of each symbol, even when the source is i.i.d. Note that the problem described in this example can be thought of as "rate-distortion coding with feedback," and the conclusion as the source-coding analog of Shannon's result that feedback does not increase the capacity of a memoryless channel [22]. Furthermore, our results establish the existence of a large class of processes *with* memory, namely, those for which the SLB holds with equality,[6] for which feedback does not improve rate-distortion performance. In general, however, our results imply that the competitive predictability of a process is strictly below its distortion rate function or, in other words, in general, feedback does enhance rate distortion performance (as it does the capacity of a general, nonmemoryless channel). Indeed, according to Corollary 3, if the process $\boldsymbol{X}$ has an autoregressive representation via some predictor F and an innovation process which is either i.i.d. or achieves the SLB with equality at a certain rate R, then the competitive predictability of the process at rate R is given by the distortion-rate function of the innovation process which, in general, may be strictly lower than that of the original process.

Evidently, the family of processes whose competitive predictability is completely characterized (in Theorem 1 and Corollary 3) is a rather large one. Following are a few examples for the kind of processes that it includes.

Example 1. Stationary Gaussian Source: If $\boldsymbol{X}$ is any stationary Gaussian source then, by letting

$$F_t(X^{t-1}) = E(X_t|X^{t-1})$$

it is well known that we have the autoregressive representation

$$X_t = F_t(X^{t-1}) + W_t$$

where the W_t's are independent zero-mean Gaussian, with decreasing variances converging to[7]

$$\sigma^2 = \exp\left\{\frac{1}{2\pi}\int_0^{2\pi} \log f_{\boldsymbol{X}}(\lambda)d\lambda\right\}$$

$f_{\boldsymbol{X}}$ being the density associated with the absolutely continuous component in the Lebesgue decomposition of the spectral measure of $\boldsymbol{X}$. Consequently, the distortion-rate function of $\{W_t\}$ coincides with that of the i.i.d. $\sim N(0, \sigma^2)$ source. Thus, we obtain from Theorem 1 that the attainable lower bound to competitive prediction at rate R of any Gaussian source is given by $D(\sigma^2, R)$, the rate-distortion function of the Gaussian random variable with variance σ^2 (and the given distortion measure). This is precisely Corollary 5. For squared error distortion, we get that the competitive predictability is explicitly given by $\sigma^2 \exp(-2R)$. Note that the result for this particular case is also implied by Corollary 3 , since, for squared error distortion, the i.i.d. Gaussian source attains the SLB with equality. We recall, in the context of this example, that the distortion-rate function, under squared error loss, of the stationary Gaussian source with one-step prediction error σ^2 is given by (cf., e.g., [5])

$$D(R) = \begin{cases} \sigma^2 e^{-2R}, & R \geq R_{\text{crit}} \\ \text{more than } \sigma^2 e^{-2R}, & R < R_{\text{crit}} \end{cases}$$

where

$$R_{\text{crit}} = \frac{1}{2}\ln\left[\frac{\sigma^2}{\min_{\lambda\in[0,2\pi)} f_{\boldsymbol{X}}(\lambda)}\right].$$

On the other hand, Corollary 5 tells us that the competitive predictability of the Gaussian source is $\sigma^2 e^{-2R}$ at *all* rates. Two conclusions this leads to are as follows.

1) For $R \geq R_{\text{crit}}$, one can use codewords from the optimal rate-distortion codebook as "predictors" and attain optimal competitive predictability performance.
2) For $R < R_{\text{crit}}$, codewords from the optimal rate-distortion codebook are strictly suboptimal if used as predictors. Reassuringly, this agrees with what we know to be the case for $R = 0$ (where the best predictor achieves distortion σ^2, yet the best codeword achieves distortion $\text{Var}(X_1)$).

The first conclusion is surprising because for a general stationary Gaussian process, which may be far from memoryless, one would expect a predictor set consisting of memoryless predictors to be strictly suboptimal. This counter-intuitive fact may be connected to the confounding relation between the rate-distortion function of the Gaussian source and that of its innovation process, as discussed in [6, Sec. V-D].

Example 2. First-Order Symmetric Binary Markov Source (and Hamming Loss): If $\boldsymbol{X}$ is a first-order Markov process taking values in $\{0, 1\}$ with a symmetric transition matrix

$$\begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$$

[6]The reader is referred to [5] for a reminder of the wide range of circumstances under which a process satisfies the SLB with equality.

[7]The reason the variances of the W_t's are only converging to σ^2 rather than being precisely equal to it is that our formulation assumes that the predictor F_t depends only on X^{t-1}, rather than on $X_{-\infty}^{t-1}$.

($\varepsilon \leq 1/2$), then $\boldsymbol{X}$ can clearly be represented autoregressively via $X_t = X_{t-1} + W_t$, $\{W_t\}$ being i.i.d. Bernoulli (ε) (and addition here is modulo 2). Thus, Theorem 1 characterizes the competitive predictability of this process. Furthermore, for Hamming loss, the innovation process satisfies the SLB with equality so the same conclusion can be reached via Corollary 3. Specifically, we obtain that the attainable lower bound to competitive prediction at rate R of this source is given by the distortion-rate function of the Bernoulli (ε) source at R (namely, $h^{-1}(h(\varepsilon) - R)$, $h^{-1}(\cdot)$ being the inverse function of $h(\cdot)$ restricted to $[0, 1/2]$).

Example 3. Hidden Markov Processes: It is easy to see that if a process satisfies the SLB with equality, then its "symbol-by-symbol" addition to any other process (independent of it) attains the SLB with equality as well.[8] Thus, Corollary 3 implies that, for example, the competitive predictability of any hidden Markov process with an additive noise channel from state to observation having noise components satisfying the SLB with equality is given by the distortion rate function of the noise. Furthermore, this is the value of the competitive predictability of any other process that can be represented autoregressively with such an innovation process.

The predictor sets constructed for the upper bounds are those induced by rate-distortion codebooks for the innovations. The predictors in these sets quantize the innovations in a data-independent way (since $\hat{w}_t$ in (28) does not depend on x^{t-1}), not making full use of their "predictive power." It would therefore seem natural to expect such predictor sets to be suboptimal. It is thus remarkable that in the above examples, as well as in all other cases covered by Theorem 1 and Corollary 3, such predictor sets are, in fact, optimal.

VI. Summary and Future Directions

The notion of "competitive predictability," as introduced in this work, seems like a novel one on two levels. The first is that it considers the problem of prediction relative to a reference class whose effective size is exponential with the sequence length, rather than the subexponential size usually assumed in the setting of prediction relative to a set of "experts." The other innovative perspective taken in this work is that the object of optimization is the predictor *set*, rather than one predictor.

We have been able to characterize the (asymptotic) value of the optimization problem associated with the competitive predictability of a wide family of processes, which includes any process possessing an autoregressive representation with an associated innovation process which is either i.i.d. or satisfies the Shannon lower bound with equality. In general, it was shown that the competitive predictability of any process is lower-bounded by the Shannon lower bound of the process (at that rate), while it is upper-bounded by the distortion rate function of the innovation process associated with any autoregressive representation for the process.

Complete characterization of the competitive predictability value for processes for which the indicated lower and upper bounds do not coincide is one open direction for future work.

Another interesting direction would be to consider the universal setting, where one seeks a predictor set minimizing its worst case competitive predictability over all sources in a given family. The problem can be shown to significantly diverge from its nonuniversal origin when the effective number of sources in the family grows exponentially with n, a setting arising naturally, e.g., in speech coding applications. It would be interesting to understand whether the optimal predictor set for this problem, analogously as in the case of universal lossless coding, is obtained by a union over the optimal predictor sets corresponding to a representative subset of the original family of sources that can be interpreted as the capacity-achieving codebook corresponding to the "channel" from the family of sources to their realizations. To the authors' knowledge, even the more basic problem of characterizing the redundancy in universal lossy source coding, say for the arbitrarily varying source, is still an open problem.

Appendix
Proofs Via Counting Arguments

We shall use the following notation. For any n and $x^n \in \mathcal{X}^n$, let

$$p_{x^n} \stackrel{\text{def}}{=} \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$$

denote the empirical measure induced by x^n. Let $\mathcal{M}(\mathcal{X})$ denote the set of all probability measures on $\mathcal{X}$ and

$$\mathcal{M}_n(\mathcal{X}) \stackrel{\text{def}}{=} \{p_{x^n} : x^n \in \mathcal{X}^n\}$$

denote the subset of $\mathcal{M}(\mathcal{X})$ consisting of nth-order empirical measures. For $P \in \mathcal{M}_n(\mathcal{X})$, let T_P denote the type class of P, i.e., $T_P = \{x^n \in \mathcal{X}^n : p_{x^n} = P\}$.

Proof of Lower Bound in Theorem 2

We give the proof assuming a finite alphabet. Fix d for which $R + \phi(d) < \underline{H}(\boldsymbol{X})$. It will be enough to show that the left-hand side of (13) is lower-bounded by d. Thus, given any sequence $\{\mathcal{G}_n\}$ of predictor sets with $|\mathcal{G}_n| \leq e^{nR}$, it remains to show that

$$\liminf_{n\to\infty} EL(\mathcal{G}_n, X^n) \geq d. \tag{A1}$$

To this end define, for each n, $A_n \subseteq \mathcal{X}^n$ via

$$A_n = \{x^n : L(\mathcal{G}_n, x^n) \leq d\}. \tag{A2}$$

By (43)

$$|A_n| \leq e^{n(R+\phi_n(d))}. \tag{A3}$$

Since $R + \phi(d) < \underline{H}(\boldsymbol{X})$ and $\phi_n(d) \to \phi(d)$ (4), it follows that for sufficiently small $\varepsilon > 0$ and all sufficiently large n

$$|A_n| \leq e^{n(\underline{H}(\boldsymbol{X})-\varepsilon)}. \tag{A4}$$

[8] To see this, note that the fact that the first process satisfies the SLB with equality implies that it can be represented as the output of an additive memoryless channel whose components have a maximum-entropy distribution with respect to (w.r.t.) the relevant loss function. Consequently, the process resulting from the said "symbol-by-symbol" addition is also an output of the same additive memoryless channel (whose input now is the "symbol-by-symbol" sum of the input process associated with the first process and the second process), and hence attains the SLB with equality as well.

The proof of the strong converse to the AEP for i.i.d. sources (cf., e.g., [11, Ch. 3, Problem 7]) easily extends to a general source, asserting that for any $\varepsilon > 0$ and sequence $\{\tilde{A}_n\}$, $\tilde{A}_n \subseteq \mathcal{X}^n$ with $|\tilde{A}_n| \leq e^{n(\underline{H}(\boldsymbol{X})-\varepsilon)}$, $\Pr(X^n \in \tilde{A}_n) \to 0$ (exponentially rapidly). Thus, by (A4) and the strong converse

$$\Pr(L(\mathcal{G}_n, X^n) \leq d) = \Pr(X^n \in A_n) \to 0. \tag{A5}$$

The proof is completed by noting that (A5) implies (A1). □

Proof of (44) (An Upper Bound to the Exponent of Theorem 4)

We now apply the counting argumentation to prove that if $\boldsymbol{X}$ is a stochastic process for which there exists a predictor F such that $\boldsymbol{W}^F$ is an i.i.d. process with a marginal distribution Q then

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log \min_{|\mathcal{G}|\leq\exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right)\right] \leq \min_{P:H(P)-\phi(d)\geq R} D(P\|Q). \tag{A6}$$

The continuity of $H(\cdot)$ and $D(\cdot\|Q)$ imply that the right-hand side of (18) equals $\inf_{P:H(P)>R+\phi(d)} D(P\|Q)$. Hence, for a fixed $P \in \mathcal{M}(\mathcal{X})$ with $H(P) > R + \phi(d)$, it will be enough to show that the left-hand side of (18) is upper-bounded by $D(P\|Q)$. Given any sequence $\{\mathcal{G}_n\}$ of predictor sets with $|\mathcal{G}_n| \leq e^{nR}$, it thus remains to show that

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log \Pr\left(L(\mathcal{G}_n, X^n) > d\right)\right] \leq D(P\|Q). \tag{A7}$$

Let $\{P^{(n)}\}$ be any sequence such that $P^{(n)} \in \mathcal{M}_n(\mathcal{X})$ and $P^{(n)} \to P$. The fact that $H(P) > R + \phi(d)$ and the continuity of $H(\cdot)$ guarantee the existence of $\varepsilon > 0$ such that

$$H(P^{(n)}) \geq R + \phi_n(d) + \varepsilon \tag{A8}$$

for all sufficiently large n. Now, for all sufficiently large n

$$\begin{aligned}
&\Pr\left(L(\mathcal{G}_n, X^n) > d\right) \\
&\quad = \Pr\left(A_n^c\right) && \text{(A9)}\\
&\quad \geq \Pr\left(A_n^c \cap \{(W_1^F, \ldots, W_n^F) \in T_{P^{(n)}}\}\right) && \text{(A10)}\\
&\quad = \Pr\left(\{(W_1^F, \ldots, W_n^F) \in T_{P^{(n)}}\}\right) \\
&\qquad - \Pr\left(A_n \cap \{(W_1^F, \ldots, W_n^F) \in T_{P^{(n)}}\}\right) \\
&\quad \geq \left(|T_{P^{(n)}}| - |A_n|\right) e^{-n[D(P^{(n)}\|Q)+H(P^{(n)})]} && \text{(A11)}\\
&\quad \geq \left((n+1)^{-|\mathcal{X}|} e^{nH(P^{(n)})} - e^{n(R+\phi_n(d))}\right) \\
&\qquad \times e^{-n[D(P^{(n)}\|Q)+H(P^{(n)})]} && \text{(A12)}\\
&\quad \geq \left((n+1)^{-|\mathcal{X}|} e^{nH(P^{(n)})} - e^{n(H(P^{(n)})-\varepsilon)}\right) \\
&\qquad \times e^{-n[D(P^{(n)}\|Q)+H(P^{(n)})]} && \text{(A13)}\\
&\quad \geq \frac{1}{2}(n+1)^{-|\mathcal{X}|} e^{-nD(P^{(n)}\|Q)} && \text{(A14)}
\end{aligned}$$

where A_n was defined in (A2) and the c superscript denotes complementation. Inequality (A11) follows from the one-to-one correspondence between source sequences and innovation sequences. Inequality (A12) follows from the bound in (A3), and inequality (A13) follows from (A8). Considering the two ends of the above chain implies that the left-hand side of (A7) is upper-bounded by $\limsup_{n\to\infty} D(P^{(n)}\|Q)$ which, in turn, implies (A7) by the continuity of $D(\cdot\|Q)$. □

By observation of the right-hand side of (A6) it is clear that a sufficient condition for (A6) and (17) to coincide is that the distribution achieving $\min_{P:R(P,d)\geq R} D(P\|Q)$ will be one for which the SLB holds with equality. One situation where this would always be the case is the binary alphabet under Hamming loss, as the SLB is achieved with equality for *all* Bernoulli sources. Another case where these bounds would coincide is that of Gaussian processes under squared error loss. Indeed, when Q is Gaussian, $\min_{P:R(P,d)\geq R} D(P\|Q)$ is known to be achieved by the Gaussian distribution P whose variance is tuned such that $R(P, d) = R$ (cf., e.g., [3]). Since the Gaussian distribution achieves the SLB with equality, we have for this case

$$\min_{P:R(P,d)\geq R} D(P\|Q) = \min_{P:H(P)-\phi(d)\geq R} D(P\|Q)$$

namely, equality between the upper and lower bounds of Theorem 4. Consequently, Theorem 4 gives the precise large deviations asymptotics for the competitive predictability of any process having an autoregressive representation with i.i.d. Gaussian innovations. Since this, in particular, is the case for any stationary Gaussian source, we obtain Corollary 6 (where the right-hand side of (20) is nothing but $D(N(0,\tilde{\sigma}^2)\|N(0,\sigma^2))$, $\tilde{\sigma}^2$ tuned such that $R(N(0,\tilde{\sigma}^2), d) = R$).

ACKNOWLEDGMENT

The authors are grateful to Tamás Linder for showing them a proof of the volume-preservation property for general predictors. The final version of the manuscript benefitted greatly from the insightful comments of the anonymous referees. We wish to thank Meir Feder for a very thorough reading of the manuscript and for the very thoughtful comments that he has provided, which resulted in significant improvement of the results.

REFERENCES

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

[2] ——, "Universal schemes for learning the best nonlinear predictor given the infinite past and side information," *IEEE Trans. Inform. Theory*, vol. 45, pp. 1165–1185, May 1999.

[3] E. Arikan and N. Merhav, "Guessing subject to distortion," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 1041–1056, May 1998.

[4] A. Barron, J. Rissanen, and B. Yu, "The minimum description length principle in coding and modeling," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2743–2760, Oct. 1998.

[5] T. Berger, *Rate-Distortion Theory: A Mathematical Basis for Data Compression*. Englewood Cliffs, N.J: Prentice-Hall, 1971.

[6] T. Berger and J. D. Gibson, "Lossy source coding," *IEEE Trans. Inform. Theory*, vol. 44, pp. 2693–2723, Oct. 1998.

[7] N. Cesa-Bianchi, Y. Freund, D. P. Helmbold, D. Haussler, R. Schapire, and M. K. Warmuth, "How to use expert advice," *J. ACM*, vol. 44, no. 3, pp. 427–485, 1997.

[8] N. Cesa-Bianchi and G. Lugosi, "Minimax regret under log loss for general classes of experts," in *Proc. 12th Annu. Workshop on Computational Learning Theory*, New York, 1999, pp. 12–18.

[9] ——, "On sequential prediction of individual sequences relative to a set of experts," *Ann. Statist.*, vol. 27, no. 6, pp. 1865–1895, 1999.

[10] ——, "Potential-based algorithms in on-line prediction and game theory," in *Proc. 5th Europ. Conf. Computational Learning Theory (EuroCOLT 2001)*, Amsterdam, The Netherlands, July 2001.

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

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

[13] R. G. Gallager, *Information Theory and Reliable Communication*. New York: Wiley, 1968.
[14] T. S. Han, "An information-spectrum approach to source coding theorems with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. 43, pp. 1145–1164, July 1997.
[15] D. Haussler, J. Kivinen, and M. K. Warmuth, "Sequential prediction of individual sequences under general loss functions," *IEEE Trans. Inform. Theory*, vol. 44, pp. 1906–1925, Sept. 1998.
[16] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 197–199, Mar. 1974.
[17] N. Merhav and M. Feder, "Universal prediction," *IEEE Trans. Inform. Theory*, vol. IT-44, pp. 2124–2147, Oct. 1998.
[18] N. Merhav and T. Weissman, "Scanning and prediction in multi-dimensional data arrays," *IEEE Trans. Inform. Theory*, vol. 49, pp. 65–82, Jan. 2003.
[19] A. Puhalskii and V. Spokoiny, "On large deviation efficiency in statistical inference," *Bernoulli*, vol. 4, no. 2, pp. 203–272, 1998.
[20] J. Rissanen, "Universal coding, information, prediction, and estimation," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 629–636, July 1984.
[21] K. Rose, "A mapping approach to rate-distortion computation and analysis," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1939–1952, Nov. 1994.
[22] C. E. Shannon, "The zero-error capacity of a noisy channel," *IRE Trans. Inform. Theory*, vol. IT-2, pp. 8–19, Sept. 1956.
[23] V. Vovk, "Aggregating strategies," in *Proc. 3rd Annu. Workshop on Computational Learning Theory*, San Mateo, CA, 1990, pp. 371–383.
[24] ——, "A game of prediction with expert advice," in *Proc 8th Annu. Workshop on Computational Learning Theory.*, New York, 1995, pp. 51–60.
[25] M. J. Weinberger, N. Merhav, and M. Feder, "Optimal sequential probability assignment for individual sequences," *IEEE Trans. Inform. Theory*, vol. 40, pp. 384–396, Mar. 1994.
[26] T. Weissman and N. Merhav, "Universal prediction of individual binary sequences in the presence of noise," *IEEE Trans. Inform. Theory*, vol. 47, pp. 2151–2173, Sept. 2001.