On Competitive Prediction and its Relation to Rate-Distortion Theory and to Channel Capacity Theory*

Tsachy Weissman Neri Merhav

February 17, 2002

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 $\mathbf{X} = X_1, X_2, \ldots$, 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 sub-exponential in n, simple arguments show that $\min_{|\mathcal{G}| \leq M} EL(\mathcal{G}, X^n)$ coincides, for large n, with the Bayesian envelope $\min_F EL_F(X^n)$. Our interest is in the behavior, for large n, of $\min_{|\mathcal{G}| \leq e^{nR}} EL(\mathcal{G}, X^n)$, which we term the *competitive predictability of* $\mathbf{X}$ *at rate* R. It is shown that under difference loss functions, the competitive predictability of $\mathbf{X}$ is lower bounded by the Shannon lower bound (SLB) on the distortion-rate function of $\mathbf{X}$ and upper bounded by the distortion-rate function of any (not necessarily memoryless) innovation process through which the process $\mathbf{X}$ has an autoregressive representation. This precisely characterizes the competitive predictability whenever $\mathbf{X}$ can be autoregressively represented via an innovation process for which the SLB is tight (e.g., when $\mathbf{X}$ is a Gaussian process under squared error loss). We next derive lower and upper bounds on the error exponents, i.e., on the exponential behavior of $\min_{|\mathcal{G}| \leq \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right)$, which are shown to be tight for many cases of interest. Finally, the universal setting is considered, where a predictor set is sought which minimizes its worst-case competitive predictability over all sources in a given family. The problem is shown to significantly diverge from its non-universal origin when the effective number of sources in the family grows exponentially with n. The optimal predictor set for this problem is shown to be related to the capacity-achieving code-book corresponding to the "channel" from the family of sources to their realizations.

Index Terms: Channel capacity, competitive prediction, error exponents, rate distortion theory, redundancy, scandiction, strong converse.

1 Introduction

The problem of universal prediction, in both its deterministic setting (cf., e.g., [12, 17, 21, 29, 19]) and stochastic setting (cf. [2] 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

*Authors are with the Department of Electrical Engineering, Technion- Israel Institute of Technology, Haifa 32000, Israel. tsachy@ee.technion.ac.il, merhav@ee.technion.ac.il.

(cf. [30, Section 1] 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. [21] and reference therein), coding, and estimation (e.g., [24, 6] and references thereto and therein). The latter question is the motivation for this work.

It should be emphasized that in the prediction problem in the literature, especially that pertaining to computational learning theory (e.g. [27, 28, 12, 11, 10, 13, 19] 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 first turn to the most basic, non-universal setting and address the following problem: given 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 [18, 7]. We shall later consider the universal case, where the sequence is known to be generated by a source belonging to some (exponentially large) uncertainty class. This problem will be seen to be connected, not only to rate-distortion theory, but also to channel capacity theory.

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 $x^n = (x_1, \ldots, x_n)$. 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 $\mathbf{X} = X_1, X_2, \ldots$, 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 $\min_{|\mathcal{G}|\leq M} EL(\mathcal{G}, X^n)$. When M is sub-exponential in n, simple arguments will show that $\min_{|\mathcal{G}|\leq M} EL(\mathcal{G}, X^n)$ 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 $\min_{|\mathcal{G}|\leq e^{nR}} EL(\mathcal{G}, X^n)$, which we term the *competitive predictability of* $\mathbf{X}$ *at rate* R. The competitive predictability of $\mathbf{X}$ will be shown to be lower bounded by the Shannon lower bound (SLB) on the distortion-rate function of $\mathbf{X}$. An upper bound will be seen to be given by the distortion-rate function of any (not necessarily memoryless) innovation process through which the process $\mathbf{X}$ has an autoregressive representation. This will lead to a precise characterization of the competitive predictability for all cases where $\mathbf{X}$ has an autoregressive representation via an innovation process for which the SLB is tight (e.g., when $\mathbf{X}$ is a Gaussian process and ρ is squared error). As will be discussed, this result has some rather surprising implications.

Next, we shall derive lower and upper bounds on the error exponents, i.e., on the exponential behavior of $\min_{|\mathcal{G}|\leq \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right)$. These bounds will be seen to be tight and to precisely characterize the error exponents for many cases of interest. As one example, we will obtain the precise competitive predictability exponent for any Gaussian process, a result which appears to be new even for $R = 0$.

Finally, we shall consider the universal setting, where one seeks a predictor set which minimizes its worst-case competitive predictability over all sources in a given family. The problem will be seen to significantly diverge from its non-universal origin when the effective number of sources in the family grows exponentially with n, a setting which naturally arises, e.g., in speech coding applications. The optimal predictor set for this problem will be obtained by a union over the optimal predictor sets corresponding to a certain representative subset of the original family of sources. This subset is induced by the capacity-achieving code-book corresponding to the "channel" from the family of sources to their realizations. As a concrete prototype, we shall give a detailed treatment of the case where the source sequence is known to be autoregressively generated via an arbitrarily varying innovation process. For this case the "achievable region" will be precisely characterized.

The remainder of this work is organized as follows. Section 2 will be dedicated to some notation, conventions, and preliminaries. In Section 3, we shall formulate the problem and present our main results for the non-universal case, outline the ideas behind their proofs, and discuss some of their implications. Section 4 will be dedicated to a presentation of our framework for the universal setting, a discussion of our general approach to the problem, a presentation of the main result, and an outline of its proof. The two subsequent sections will contain formal proofs and derivation of results: In Section 5, the results pertaining to the non-universal setting will be proven and some additional corollaries derived, and in Section 6, a formal proof of the main result from the universal setting will be given. Finally, in Section 7, we discuss some related directions for future research.

2 Notation, Conventions, and Preliminaries

Throughout, $\mathcal{X}$ will denote the source alphabet which will be either the real line or a finite set, in which case a group structure will be assumed so that addition and subtraction of elements are well-defined. 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 space of all (Borel, when $\mathcal{X} = \mathbb{R}$) 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 a sequence $\{P^{(n)}\}$, $P^{(n)} \in \mathcal{M}(\mathcal{X})$, let $P^{(n)} \to P$ denote weak convergence. For $P \in \mathcal{M}_n(\mathcal{X})$, let T_P denote the type of P, i.e., $T_P = \{x^n \in \mathcal{X}^n : p_{x^n} = P\}$. For $x^n, y^n \in \mathcal{X}^n$, with the usual abuse of notation, we write $\rho(x^n, y^n)$ for $\frac{1}{n}\sum_{i=1}^n \rho(x_i - y_i)$.

Assume throughout a fixed loss function ρ satisfying $\rho(0) = 0$. For $P \in \mathcal{M}(\mathcal{X})$, we let $H(P)$ denote the entropy (differential entropy, when $\mathcal{X} = \mathbb{R}$). 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., [21, 22]) to have a closed form representation. Specifically, for finite $\mathcal{X}$, let

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

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:

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

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., [22]):

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

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 analogue of (4) for this case is

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

For simplicity, our treatment of the universal competitive predictability problem in the case of an arbitrarily varying innovation process will be restricted to the case of a finite alphabet $\mathcal{X}$, as well as a finite state-space $\mathcal{S}$ associated with the arbitrarily varying innovation process. This will allow our analysis to heavily rely on the method of types a la Csiszár and Körner [16]. We shall thus adopt some of the notation and conventions of [16]. Specifically, for $P \in \mathcal{M}(\mathcal{S})$, a sequence $s^n \in \mathcal{S}^n$ is called P-typical with constant δ if $|\{1 \leq i \leq n : s_i = a\}/n - P(a)| \leq \delta$ for every $a \in \mathcal{X}$. The set of all sequences $s^n \in \mathcal{S}^n$ that are P-typical with constant δ are denoted by $T^n_{[P]_\delta}$. Further let, for the channel $V(w|s)$ and $s^n \in \mathcal{S}^n$, $T^n_{[V]_\delta}(s^n)$ denote the set of all $w^n \in \mathcal{X}^n$ that are V-typical under the condition $s^n \in \mathcal{S}^n$ with constant δ (cf. [16, Definition 2.9]). An immediate consequence of the definitions of δ-typical sequences is (cf. [16, Lemma 2.10]):

$$\text{If } s^n \in T^n_{[P]_\delta} \text{ and } w^n \in T^n_{[V]_{\delta'}}(s^n) \text{ then } (s^n, w^n) \in T^n_{[P\times V]_{\delta+\delta'}}. \tag{6}$$

We shall adopt throughout this work the "delta-convention" of [16]. Specifically, we assume a fixed sequence of positive reals $\{\delta_n\}_{n\geq 1}$ satisfying

$$\delta_n \to 0, \quad \sqrt{n}\delta_n \to \infty \quad \text{ as } \; n \to \infty \tag{7}$$

and, for any n, $P \in \mathcal{M}(\mathcal{S})$, channel V and $s^n \in \mathcal{S}^n$, we write $T^n_{[P]}$ for $T^n_{[P]_{\delta_n}}$ and $T^n_{[V]}(s^n)$ for $T^n_{[V]_{\delta_n}}(s^n)$. We omit the superscript n, writing $T_{[P]}$ and $T_{[V]}(s^n)$, when no confusion can arise.

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. $E_P\rho(W)$ denotes expectation when $W \sim P$. For $P \in \mathcal{M}(\mathcal{S})$, let $P \times V$ denote the distribution of the pair (S, W) when S is generated according to P and W is the output of the channel V whose input is S. We let $I(P; V)$ and $H(V|P)$ denote the mutual information between S and W, and the entropy of W given S, respectively, when jointly distributed according to $P \times V$. Finally, let $[c]_+ = \begin{cases} c & c \geq 0 \\ 0 & \text{otherwise.} \end{cases}$ and define the minimum over the empty set as ∞.

3 Competitive Predictability in the Non-Universal Setting

A Competitive Predictability Defined

A predictor F for a sequence $x_1, x_2, \ldots, x_t \in \mathcal{X}$, $|\mathcal{X}| < \infty$ ($|\cdot|$ denoting cardinality) is a sequence of functions $F = \{F_t\}_{t\geq 1}$, where $F_t : \mathcal{X}^{t-1} \to \mathcal{X}$. Let $\mathcal{F}$ denote the set of all such predictors and assume that the subtraction operation between elements of the alphabet $\mathcal{X}$ is well-defined (e.g., enough that $(\mathcal{X}, +)$ is a group). For a loss function $\rho : \mathcal{X} \to [0, \infty)$ and a sequence $x^n = (x_1, \ldots, x_n)$, denote the normalized cumulative loss of the predictor F by

$$L_F(x^n) \stackrel{\text{def}}{=} \frac{1}{n}\sum_{t=1}^{n} \rho(x_t - F_t(x^{t-1})). \tag{8}$$

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

$$L(\mathcal{G}, x^n) \stackrel{\text{def}}{=} \min_{F \in \mathcal{G}} L_F(x^n), \tag{9}$$

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

Suppose now that $\mathbf{X} = X_1, X_2, \ldots$ 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

$$\min_{|\mathcal{G}| \leq M} EL(\mathcal{G}, X^n) \tag{10}$$

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 predictor $F \in \mathcal{F}$ such that

$$\sup_{x^n} [L_F(x^n) - L(\mathcal{G}, x^n)] \leq K \sqrt{\frac{\log |\mathcal{G}|}{n}}, \tag{11}$$

where K is a constant depending only on the loss function ρ and the alphabet $\mathcal{X}$. A predictor F satisfying (11) can always be constructed via the exponential weighting approach (cf., e.g., [12, 13]). It follows from (11) that if M is sub-exponential 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. [2, 3] and references therein). Thus, the quantity in (10) can become interesting and significantly deviate from the classical optimal prediction problem when M grows exponentially in n. This is the motivation for focusing on the case where $M = e^{nR}$, $R > 0$.

B Main Results

Let ϕ^{-1} denote the generalized inverse function of ϕ defined by $\phi^{-1}(\alpha) \stackrel{\text{def}}{=} \inf\{d : \phi(d) > \alpha\}$. For a stochastic process $\mathbf{X}$, let

$$\underline{H}(\mathbf{X}) \stackrel{\text{def}}{=} \liminf_{n \to \infty} H(X^n), \tag{12}$$

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

$$D(\mathbf{X}, R) \stackrel{\text{def}}{=} \limsup_{n \to \infty} D(X^n, R), \tag{13}$$

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

$$W_t^F = X_t - F_t(X^{t-1}) \tag{14}$$

as the *innovation process*[1] *associated with the process* $\mathbf{X}$ *and the predictor* F. For this case, we shall say that $\mathbf{X}$ *has an autoregressive representation via the predictor* F *and innovation process* $\mathbf{W}^F$.

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

Theorem 1 *Let $\mathbf{X} = X_1, X_2, \ldots, X_i \in \mathcal{X}$, be any stochastic process and $R \geq 0$.*

Lower bound:

$$\liminf_{n\to\infty} \left[\min_{|\mathcal{G}|\leq \exp(nR)} EL(\mathcal{G}, X^n) \right] \geq \phi^{-1}(\underline{H}(\mathbf{X}) - R) \tag{15}$$

Upper bound:

$$\limsup_{n\to\infty} \left[\min_{|\mathcal{G}|\leq \exp(nR)} EL(\mathcal{G}, X^n) \right] \leq \inf_{F\in\mathcal{F}} D(\mathbf{W}^F, R), \tag{16}$$

where $\mathbf{W}^F$, on the right side of (16), is the innovation process associated with the process $\mathbf{X}$ and the predictor F.

Note that Theorem 1 is formulated in terms of the *expected* competitive predictability $EL(\mathcal{G}, X^n)$. As will be clear from its proof, however, a similar statement holds for the "with high probability" setting. Namely, the proof of the lower bound will actually be seen to imply that for any $\varepsilon > 0$

$$\lim_{n\to\infty} \min_{|\mathcal{G}|\leq \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) \geq \phi^{-1}(\underline{H}(\mathbf{X}) - R) - \varepsilon \right) = 1,$$

while the proof of the upper bound will be seen to easily be modified to establish that for any $\varepsilon > 0$

$$\lim_{n\to\infty} \min_{|\mathcal{G}|\leq \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) \leq \inf_{F\in\mathcal{F}} D(\mathbf{W}^F, R) + \varepsilon \right) = 1.$$

As was shown in [22], the entropy (rate) of a process equals the entropy (rate) of the associated innovation process relative to any predictor. Thus, given a process $\mathbf{X}$ and any predictor F, $\underline{H}(\mathbf{X}) = \underline{H}(\mathbf{W}^F)$. Recall that the SLB (cf. [7, 14, 25]) on the rate-distortion function of the source $\mathbf{X}$ at distortion level d is $\underline{H}(\mathbf{X}) - \phi(d)$. Furthermore, since the SLB is dependent on the source only through its entropy, it is clear that $R(\mathbf{W}^F, d) \geq \underline{H}(\mathbf{X}) - \phi(d)$ for all F or, in terms of the distortion-rate function,

$$D(\mathbf{W}^F, R) \geq \phi^{-1}(\underline{H}(\mathbf{X}) - R) \quad \forall F \tag{17}$$

Note that (17) follows from Theorem 1 as well since the left side is shown to be achievable while the right side is a lower bound on the achievable loss. In this context, we make the following observation.

Corollary 2 *Let $\mathbf{X}$ be a stochastic process and suppose there exists a predictor F such that $D(\mathbf{W}^F, R)$ meets the SLB with equality. Then*

$$\lim_{n\to\infty} \left[\min_{|\mathcal{G}|\leq \exp(nR)} EL(\mathcal{G}, X^n) \right] = D(\mathbf{W}^F, R). \tag{18}$$

In plain words, Corollary 2 tells us that whenever the process $\mathbf{X}$ has an autoregressive representation via any predictor F, i.e., $X_t = F_t(X^{t-1}) + W_t$, and $\{W_t\}$ has a rate-distortion function achieving the SLB with equality, the attainable limitation on competitive prediction of the source $\mathbf{X}$ is precisely the rate-distortion function of the innovation process. In subsection D we shall recall the general condition for the SLB to hold with equality, which is the case, e.g., for Bernoulli, Gaussian and Laplacian distributions under Hamming-, squared-, and absolute-error distortion measures, respectively.

For the case where the process $\mathbf{X}$ has an autoregressive representation with i.i.d. innovations, we obtain considerably more refined results on the error exponents for competitive prediction.

Theorem 3 *Let* $\mathbf{X}$ *be a stochastic process for which there exists a predictor* F *such that* $\mathbf{W}^F$ *is an i.i.d. process with a marginal distribution* Q.

Upper Bound:

$$\limsup_{n\to\infty}\left[-\frac{1}{n}\log\min_{|\mathcal{G}|\leq\exp(nR)}\Pr\left(L(\mathcal{G},X^n)>d\right)\right]\leq\min_{P:H(P)-\phi(d)\geq R}D(P\|Q) \tag{19}$$

Lower Bound:

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log\min_{|\mathcal{G}|\leq\exp(nR)}\Pr\left(L(\mathcal{G},X^n)>d\right)\right]\geq\begin{cases}F_d(R-0) & \text{if } R>0\\ \min_{P:E_P\rho(W)\geq d}D(P\|Q) & \text{if } R=0,\end{cases} \tag{20}$$

where $F_d(R)\stackrel{\text{def}}{=}\min_{P:R(P,d)\geq R}D(P\|Q)$ *is Marton's source coding error exponent function [20].*

Evidently, Theorem 3 gives the precise exponents when $\min_{P:H(P)\geq R+\phi(d)}D(P\|Q)$ coincides with the right side of (20). In particular, for rate $R=0$, this will be shown to happen whenever $Q=Q^{(\beta)}$ is a maximum-entropy distribution w.r.t. ρ, i.e.,

$$Q^{(\beta)}(w)=e^{-\beta\rho(w)+\lambda(\beta)}, \tag{21}$$

for some β positive, $\lambda(\beta)$ being the normalizing factor. We let Q_d denote the max-entropy distribution (21) when the parameter β is tuned so that $E_{Q^{(\beta)}}\rho(W)=d$.

Corollary 4 *Let* $\mathbf{X}$ *be a stochastic process for which there exists a predictor* F *such that* $\mathbf{W}^F$ *is an i.i.d. process with a maximum-entropy marginal* Q*. Then*

$$\begin{aligned}\lim_{n\to\infty}\left[-\frac{1}{n}\log\min_{G\in\mathcal{F}}\Pr\left(L_G(X^n)>d\right)\right] &= \lim_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left(L_F(X^n)>d\right)\right] &(22)\\ &= \begin{cases}D(Q_d\|Q) & \text{for } d>E_Q\rho(Z)\\ 0 & \text{otherwise.}\end{cases} &(23)\end{aligned}$$

Note that Corollary 4 explicitly addresses the case of one predictor, though it should be clear, following the discussion in subsection A, that the competitive predictability of any sub-exponential number of predictors would give rise to the same large deviations behavior. As far as the authors are aware, Theorem 3 at $R=0$ and Corollary 4 are the first to explicitly characterize the large deviations performance for the prediction problem.

More generally, observe that a sufficient condition for the right sides of (19) and (20) 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. This is easily seen to be the case, e.g., for the binary alphabet under Hamming loss, as will be made explicit in Section 5 (Corollary 8).

To simplify the exposition, we assume a finite alphabet. As will be elaborated on in Section 5, all the results carry over to the case where the alphabet is the real line and attention is restricted to $F_t:\mathbb{R}^{t-1}\to\mathbb{R}$ that are continuously differentiable for all t. Accordingly, throughout this work, when the alphabet is the real line, the assumption is that $\mathcal{F}$ consists of all the continuously differentiable predictors. As will be discussed in Section 5, much less than continuous differentiability is needed.

As an example of a concrete application of these results for a real-valued process, the following will be argued in subsection D to be a consequence of Theorem 1.

Corollary 5 *Let* $\mathbf{X}$ *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*

$$\lim_{n\to\infty}\left[\min_{|\mathcal{G}|\leq \exp(nR)} EL(\mathcal{G}, X^n)\right] = \sigma^2 2^{-2R}. \tag{24}$$

For the Gaussian case, the bounds on the error exponents given in (the continuous analogue of) Theorem 3, also turn out to be tight, leading to

Corollary 6 *Let* $\mathbf{X}$ *be a stationary Gaussian process as in Corollary 5. Then*

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log \min_{|\mathcal{G}|\leq \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right)\right] = \begin{cases} \frac{1}{2}\left[\log\left(\frac{\sigma^2}{2\cdot 2^{2R} d}\right) + \frac{2^{2R} d}{2\sigma^2}\right] & \text{if } d > \sigma^2 2^{-2R} \\ 0 & \text{otherwise} \end{cases} \tag{25}$$

In particular, taking $R = 0$ in Corollary 6, gives the complete characterization of the best attainable large deviations performance in prediction of Gaussian processes, a result which appears to be new (cf. discussion in [23]).

C Proof Ideas

At the heart of the proofs of the lower bounds presented above for the non-universal finite-alphabet case, lies the following "counting" argument. 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: $e_t = x_t - F_t(x^{t-1})$, $t = 1, \ldots, n$. To see this, one must simply notice that given e^n and F, x^n can be uniquely recovered[2]. Since, by definition, $L_F(x^n) = \frac{1}{n}\sum_{i=1}^{n} \rho(e_i)$, where e^n on the right side is the error sequence associated with x^n and F, it follows that for any F and $d \geq 0$

$$|\{x^n : L_F(x^n) \leq d\}| = \left|\left\{e^n : \frac{1}{n}\sum_{i=1}^{n}\rho(e_i) \leq d\right\}\right| \stackrel{\text{def}}{=} e^{n\phi_n(d)}. \tag{26}$$

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

$$|\{x^n : L(\mathcal{G}, x^n) \leq d\}| = \left|\bigcup_{F\in\mathcal{G}}\{x^n : L_F(x^n) \leq d\}\right| \leq \sum_{F\in\mathcal{G}} |\{x^n : L_F(x^n) \leq d\}| = |\mathcal{G}| e^{n\phi_n(d)}. \tag{27}$$

The fact that $\phi_n(d) \to \phi(d)$, combined with (27), leads to the following conclusions (stated qualitatively here and made precise in Section 5):

1. If $R + \phi(d) < \underline{H}(\mathbf{X})$, 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 $e^{n\underline{H}(\mathbf{X})}$ and hence, by the converse to the asymptotic equipartition property (AEP), $\Pr\left(L(\mathcal{G}, X^n) \leq d\right)$ 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] This is precisely the idea on which predictive coding techniques are based.

2. Suppose that the process $\mathbf{X}$ has been auto-regressively generated via the predictor F and the i.i.d. innovation process $\mathbf{W}^F$. 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 (27), is exponentially more than $|\{x^n : L(\mathcal{G}, x^n) \leq d\}|$ whenever $R + \phi(d) < H(P)$. Thus, $\Pr\left(L(\mathcal{G}, X^n) > d\right)$ 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 1, while the second observation leads to the large deviations converse (upper bound on the exponent) in Theorem 3.

The idea underlying the upper bounds of the non-universal setting is the following. Fix any random process $\mathbf{X}$, any predictor F, and $R \geq 0$. To any $\hat{w}^n \in \mathcal{X}^n$, we associate the predictor G, specified by

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

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 code-book $\mathcal{C}_n$ and the predictor F*. By the definition of $\mathbf{W}^F$ (recall equation (14)) it follows that for any $\mathcal{C}_n$, with probability 1,

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

where the $\mathcal{G}_n$ on the right side is that induced by $\mathcal{C}_n$ and F. Thus,

1. For large n, rate-distortion theory guarantees the existence of a code-book, $\mathcal{C}_n \subseteq \mathcal{X}^n$, for the innovation process $\mathbf{W}^F$ such that $|\mathcal{C}_n| \leq e^{nR}$ and $E\left[\min_{\hat{w}^n \in \mathcal{C}_n} \frac{1}{n} \sum_{t=1}^{n} \rho(W_t^F - \hat{w}_t)\right] \approx D(\mathbf{W}^F, R)$. Consequently, it follows from (29) that by letting $\mathcal{G}_n$ be the predictor set induced by $\mathcal{C}_n$ and F, we get $EL(\mathcal{G}_n, X^n) \approx D(\mathbf{W}^F, R)$.

2. Suppose that the process $\mathbf{X}$ has been autoregressively generated by the predictor F and the i.i.d. innovation process $\mathbf{W}^F$. For $R, d \geq 0$, take $\mathcal{C}_n$ to be a code-book which is optimal in Marton's error-exponent sense [20], i.e., for which $\Pr\left(\min_{\hat{w}^n \in \mathcal{C}_n} \frac{1}{n} \sum_{t=1}^{n} \rho(W_t^F - \hat{w}_t) > d\right) \approx e^{-nF_d(R)}$. For the $\mathcal{G}_n$ induced by $\mathcal{C}_n$ and F, (29) implies $\Pr\left(L(\mathcal{G}_n, X^n) > d\right) \approx e^{-nF_d(R)}$.

The above two observations lie at the heart of the proofs of the direct parts of Theorem 1 and Theorem 3, respectively.

As mentioned in subsection B, the results will be shown to carry over to the case of the continuous, real-valued, alphabet. For this case, in the line of argumentation described above the "counting" arguments are replaced by "volume-preservation" ones.

We mention in passing that "counting" and "volume-preservation" arguments of the type we employ here were used in a recent work [22] to characterize the fundamental limitations on *scandiction* performance, where a *scandictor* is any scheme for the sequential scanning and prediction of (usually multi-dimensional-) data. Indeed, it was shown in that work that equation (26) (and its volume-preservation analogue for the real-valued alphabet) continues to hold also for any *scandictor*. Consequently, all the above described results carry over to the more general setting of "competitive scandictability". Specifically, all the results remain valid when the minimum is taken, rather than only over predictor sets of size $|\mathcal{G}| \leq e^{nR}$, over all scandictor sets with the same size limitation. Similarly, those parts of the results pertaining to the auto-regressive representation of a process via a predictor remain true more generally for the autoregressive representation of the process via any scandictor (cf. [22] for the precise definition of this notion).

D Discussion of Results

As was discussed and shown in subsection B (Corollary 2, in particular), if the process $\mathbf{X}$ has an autoregressive representation via some predictor F and an innovation process achieving the SLB with equality at a certain rate R, then the competitive predictability of the process at rate R is completely characterized, namely, it is given by the distortion-rate function of the innovation process. It is thus of interest to recall the necessary and sufficient condition for the tightness of the SLB, in the case of an i.i.d. source (cf., e.g., [7, Theorem 4.3.1]). To this end, recall first that $\phi(d)$ is the entropy[3] of the maximum-entropy distribution (21), when the parameter β is tuned so that $E_{Q^{(\beta)}}\rho(Z) = d$ (cf., e.g., [22]), i.e., $\phi(d) = H(Q_d)$. The reason for the term "maximum-entropy distribution" is the property that

$$E_Q\rho(Z) \leq E_{Q^{(\beta)}}\rho(Z) \quad \Rightarrow \quad H(Q) \leq H(Q^{(\beta)}). \tag{30}$$

The condition for the tightness of the SLB can be summarized as

$$R(P, d) = H(P) - \phi(d) \quad \Leftrightarrow \quad \exists \hat{P} : P = \hat{P} * Q_d, \tag{31}$$

where $*$ denotes convolution. The conclusion, for our context, is that if the source $\mathbf{X}$ has an autoregressive representation with i.i.d. $\sim P$ innovations, and if P satisfies the right side of (31), then the competitive predictability of $\mathbf{X}$ at rate $R = R(P, d)$ is d.

Examples of distributions for which the SLB holds with equality include the Gaussian distribution under squared-error distortion, the Laplacian distribution under absolute-error distortion, the Bernoulli distribution under Hamming loss, and infinitely many more (cf. [7, 25]). Using such

[3] The differential entropy for the continuous alphabet.

distributions, we can construct an infinite spectrum of examples for which the bounds of Theorem 1 coincide or, in other words, for which Corollary 2 is applicable. Two such representative examples follow.

Example 1. Stationary Gaussian Source under Squared Error: If $\mathbf{X}$ is any zero-mean 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 auto-regressive 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 $\sigma^2 = \exp\left\{\frac{1}{2\pi}\int_0^{2\pi}\log f_{\mathbf{X}}(\lambda)d\lambda\right\}$, $f_{\mathbf{X}}$ being the density associated with the absolutely continuous component in the Lebesgue decomposition of the spectral measure of $\mathbf{X}$. The entropy rate and distortion-rate function of $\{W_t\}$ are easily seen to coincide with those of the i.i.d. $\sim N(0,\sigma^2)$ source. Thus, we obtain that the attainable lower bound to competitive prediction at rate R of any Gaussian source is given by $\sigma^2\exp(-2R)$, the distortion-rate function of the i.i.d. $N(0,\sigma^2)$ source, which is precisely Corollary 5.

Example 2. First-Order Symmetric Binary Markov Source under Hamming Loss: If $\mathbf{X}$ is a first-order Markov process taking values in $\{0,1\}$ with a symmetric transition matrix $\begin{pmatrix} 1-\varepsilon & \varepsilon \\ \varepsilon & 1-\varepsilon \end{pmatrix}$ ($\varepsilon \leq 1/2$), then $\mathbf{X}$ can clearly be represented autoregressively represented via $X_t = X_{t-1} + W_t$, $\{W_t\}$ being i.i.d. Bernoulli (ε) (and addition here is modulo 2). Since the Bernoulli source attains the SLB with equality, we find 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 $[0,1/2]$).

As described in the previous subsection, the predictor sets constructed for the upper bounds are those induced by rate-distortion code-books for the innovations. The predictors in these sets quantize the innovations in a data-independent way (as clearly $\hat{w}_t$ 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 where the innovations achieve the SLB with equality, such predictor sets are, in fact, optimal. To further ponder on the implications of this fact, we make the following two observations:

Distortion-Rate Source Coding with Perfect Past Side-Information: Suppose we wish to store the i.i.d. data $(W_1, \ldots, W_n)$ (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 $\hat{W}_1$ by January 1st (2002), $\hat{W}_2$ by January 2nd, and so forth. We know, however, that the original data $(W_1, \ldots, W_n)$ is 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? Corollary 2 implies that when the distribution of the W_i-s achieves the SLB with equality, there is nothing to gain from the *perfect* (as opposed to quantized) observations of the past source sequence. At first glance, this observation is not surprising because the W_i-s are independent so it seems natural that there is nothing to gain from observing the past sequence for reconstruction of the present symbol. This observation is, however, somewhat 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. The point to emphasize in the context of this example is that the sequence of predicted values of every predictor G_t (28), cannot be considered a code-word for X^n as it is autoregressively constructed from the clean (non-quantized) source, rendering the connection with rate-distortion theory rather intriguing.

Stationary Gaussian Source under Squared Error: In the context of Example 1 above, recall (e.g. [7]) that the distortion-rate function of the stationary Gaussian source with one-step prediction error σ^2 is given by

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

where $R_{crit} = \frac{1}{2} \log \left[\frac{\sigma^2}{\min_{\lambda \in [0,2\pi)} f_{\mathbf{X}}(\lambda)} \right]$. On the other hand, Corollary 5 tells us that the competitive predictability of the Gaussian source is $\sigma^2 2^{-2R}$ at *all* rates. Two conclusions this leads to are:

1. For $R \geq R_{crit}$, one can use codewords from the optimal R-D code-book as "predictors" and attain optimal competitive predictability performance.

2. For $R < R_{crit}$, codewords from the optimal R-D code-book are strictly sub-optimal if used as predictors. Reassuringly, this is in accordance to what we know to be the case for $R = 0$ (where the best predictor achieves distortion σ^2, yet the best code-word 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 sub-optimal. 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 [8, Subsection V.D].

4 Universal Case: Competitive Predictability w.r.t. an Exponentially Large Family of Sources

Suppose now that the process $\mathbf{X}$ is known to have been generated autoregressively via a certain predictor F and an innovation process $\mathbf{W}$ whose distribution, rather than being completely known, is only known to belong to some set of distributions. More concretely, suppose that the distribution of the innovation vector W^n generating X^n via F is known to lie in Θ_n. Our interest in this setting is to find, given a distortion level d, the smallest possible predictor set whose competitive predictability, for large enough n, does not exceed d for all sources $\theta \in \Theta_n$. Letting P_θ denote the probability measure corresponding to the innovation source θ, we formalize this as follows.

Definition 1 *The pair (R, d) will be said to be achievable w.r.t. $\{\Theta_n\}_{n\geq 1}$ if there exists a sequence of predictor sets $\{\mathcal{G}_n\}$ with $|\mathcal{G}_n| \leq e^{nR}$ such that*

$$\lim_{n\to\infty} \min_{\theta\in\Theta_n} P_\theta \left(L(\mathcal{G}_n, X^n) \leq d\right) = 1. \tag{32}$$

The achievable region is the closure of the set of all achievable pairs.

When the effective size[4] of Θ_n is sub-exponential in n, there is no essential price for universality, as the size of the predictor set resulting from the union of predictor sets of size $\approx e^{nR}$ corresponding to the (effective) different members of Θ_n is exponentially the same. Thus, when the Θ_n's are, for example, parametric classes corresponding to all possible i.i.d. innovation sources over the simplex, the problem of competitive predictability as introduced in Definition 1 does not digress from its non-universal origin. When the effective size of Θ_n *is* exponential, on the other hand, the problem becomes interesting both from a practical and a theoretical point of view.

One important example for the practical significance of the setting where the data is known to be autoregressively generated via one certain predictor F, yet with an innovation process whose distribution lies in an exponentially large class, is that of multi-pulse and stochastically excited linear predictive coders (MELP) in the context of speech compression (cf., e.g., [9, 5, 1, 26] and the many references therein and thereto). In these problems, the data is assumed to be generated according to $X_t = F_t^{(\eta)}(X^{t-1}) + W_t$, where $\{F^{(\eta)}\}$ is typically the parametric family of linear predictors of a given order. The innovation process is allowed to be non-stationary, and in the language of signal processing, it is exactly the multipulse excitation. Since, as mentioned above, universality w.r.t. a smooth parametric family is not an issue in our case because its richness is sub-exponential, it is essentially like assuming that η is known. Since the family of possible distributions governing the non-stationary innovation process is exponentially large, this case naturally falls within the setting on which we shall focus henceforth, namely, that where the process $\mathbf{X}$ is generated via a known predictor F and an innovation process governed by a distribution from an exponentially large family. One important and naturally occurring (e.g., in the context of the speech compression setting) example for such an exponentially large family is that of the arbitrarily varying source (AVS), on which we shall focus in subsection B.

A Qualitative Approach to Derivation of Results

Our aim in this subsection is to schematically and informally present our basic approach to the problem. In particular, we shall see how the notion of channel capacity naturally arises.

A necessary condition for the competitive predictability value d to be achievable at rate R for

[4]By "effective size of Θ_n" we loosely mean here the size of a representative subset of sources in Θ_n needed to approximate Θ_n.

all sources in Θ_n is essentially the existence of a set $\mathcal{G}$ with $|\mathcal{G}| \leq e^{nR}$ such that

$$\bigcup_{\theta \in \Theta_n} \text{Typical set}_\theta \subseteq \{x^n : L(\mathcal{G}, x^n) \leq d\}, \tag{33}$$

where by "Typical set$_\theta$", we loosely mean the exponentially smallest set containing "most" of the probability mass under P_θ. Since $\mathrm{Vol}(\{x^n : L(\mathcal{G}, x^n) \leq d\}) \leq e^{n(R+\phi(d))}$, a necessary condition for (33) is

$$\mathrm{Vol}\left(\bigcup_{\theta \in \Theta_n} \text{Typical set}_\theta\right) \leq e^{n(R+\phi(d))}. \tag{34}$$

Thus, for any subset $\{\theta^{(i)}\}_{i=1}^M \subseteq \Theta_n$ such that the Typical set$_{\theta^{(i)}}$ are essentially disjoint:

$$\sum_{i=1}^{M} \mathrm{Vol}\left(\text{Typical set}_{\theta^{(i)}}\right) \stackrel{\cdot}{<} \mathrm{Vol}\left(\bigcup_{\theta \in \Theta_n} \text{Typical set}_\theta\right) \leq e^{n(R+\phi(d))}, \tag{35}$$

$\stackrel{\cdot}{<}$ denoting inequality up to sub-exponential terms. To construct such a set $\{\theta^{(i)}\}_{i=1}^M$, we think of the "channel" from $\theta \in \Theta_n$ into the realization of the source P_θ. In particular, if the capacity of this channel is $\approx C$, then we can find a channel code-book $\{\theta^{(i)}\}_{i=1}^{e^{nC}}$ for which $\{\text{Typical set}_{\theta^{(i)}}\}$ are essentially disjoint. If, in addition, the code-book has constant composition in the sense that $H_{\theta^{(i)}} \stackrel{\text{def}}{=} \frac{1}{n} \log \mathrm{Vol}\left(\text{Typical set}_{\theta^{(i)}}\right)$ are approximately all equal, say to H, then considering this code-book in (35) leads to the converse statement

$$C + H \stackrel{<}{\sim} R + \phi(d), \tag{36}$$

$\stackrel{<}{\sim}$ denoting inequality up to asymptotically negligible terms. For a direct result, suppose that the "channel" described above has a strong converse. This will essentially imply that if we take a code-book of size $\{\theta^{(i)}\}_{i=1}^{e^{n(C+\varepsilon)}}$ then $\bigcup_i \text{Typical set}_{\theta^{(i)}} = \bigcup_{\theta \in \Theta_n} \text{Typical set}_\theta$, i.e., covers the whole space of typical sequences corresponding to all sources. Thus, by taking, for each i, the optimal predictor set which guarantees competitive predictability level d for the source $P_{\theta^{(i)}}$, the predictor set obtained by the union of these sets is guaranteed of achieving competitive predictability level d for all sources $\theta \in \Theta_n$. Letting $R(P_\theta, d)$ denote the distortion-rate function of the source P_θ, it follows from the upper bound in Theorem 1 that there exists a predictor set of size $e^{nR(\theta^{(i)}, d)}$ whose competitive predictability is at most d under $P_{\theta^{(i)}}$. Thus, by unifying the predictor sets corresponding to the sources $\{\theta^{(i)}\}$, we obtain a predictor set whose size e^{nR} is $\stackrel{<}{\sim} e^{n(\max_i R(\theta^{(i)}, d) + C)}$ and which attains a competitive predictability value of at most d for all sources $\theta \in \Theta_n$. Note, in particular, that when the $R(\theta^{(i)}, d)$ are all equal and attain the SLB with equality, namely, $R(\theta^{(i)}, d) = H - \phi(d)$, we get $R \stackrel{<}{\sim} H - \phi(d) + C$, i.e., the reverse inequality to (36). Thus, when this is the case, the predictor set constructed this way is optimal and the achievable region (in the sense of Definition 1) is fully characterized: (R, d) is an achievable pair if and only if it satisfies (36).

The above is a description of our basic approach to characterizing the competitive predictability in the universal setting. To make it precise, the exact structure of the sources P_θ must be considered on a case-by-case basis. In the next subsection, we shall present our main result and informally outline how the approach presented above is specialized for the concrete problem where the innovations source is an AVS.

B The Arbitrarily Varying Innovation Process

The prototypical exemplar for an exponentially large class of sources is the case where the innovation process is an arbitrarily varying source. Specifically, suppose that the innovation process is known to be generated by an AVS characterized by the "channel" $V(w|s)$. I.e., for every $w^n \in \mathcal{X}^n$, $\Pr(W^n = w^n) = \prod_{i=1}^n V(w_i|s_i)$, where $s^n = \{s_i\}_{i=1}^n$ is an unknown state sequence, $s_i \in \mathcal{S}$, $\mathcal{S}$ being a finite state space, as well as $|\mathcal{X}| < \infty$. In accordance with the notation introduced above, we let P_{s^n} denote[5] the probability measure corresponding to the innovation source indexed by s^n. To present our main result for this setting, let $[P \times V]_{\mathcal{W}}$ denote the marginal distribution of W induced by $P \times V$. Our main result for this setting is the following.

Theorem 7 *Suppose that $V(\cdot|s)$ attains the SLB with equality for all $s \in \mathcal{S}$. The pair (R, d) is in the achievable region w.r.t. $\{\mathcal{S}^n\}$ (namely, w.r.t. all AVS's) if and only if*

$$\max_{P \in \mathcal{M}(\mathcal{S})} H\left([P \times V]_{\mathcal{W}}\right) \le R + \phi(d). \tag{37}$$

Note that, for a given rate R, the definition of an achievable loss value d requires the competitive predictability to be below d (with high probability) for all sources (worst case). As will be seen in the proof of the direct part of Theorem 7 in Section 6, however, when (R, d) satisfy (37) one can construct predictor sets having competitive predictability value significantly less than d for many of the sources. This will be done by constructing, for each type $P \in \mathcal{M}_n(\mathcal{S})$, a predictor set of exponential size e^{nR} with worst-case distortion d (relative to T_P) satisfying $H\left([P \times V]_{\mathcal{W}}\right) = R + \phi(d)$. Since $\mathcal{M}_n(\mathcal{S})$ is polynomial, the unification of these predictor sets gives one predictor set of the same rate R and worst-case distortion attaining (37) with equality, yet achieving lesser distortion on all types P' for which $H\left([P' \times V]_{\mathcal{W}}\right) < \max_{P \in \mathcal{M}(\mathcal{S})} H\left([P \times V]_{\mathcal{W}}\right)$. It will also be seen that the converse part of Theorem 7 holds for any AVS, regardless of whether V attains the SLB or not. In the remainder of this section, we informally outline the idea behind the proof of Theorem 7.

Note first that Vol(Typical set$_{s^n}$) $\approx e^{nH(W|S)}$, $H(W|S)$ denoting the conditional entropy when S, W are jointly distributed according to $p_{s^n} \times V$. For $P \in \mathcal{M}(\mathcal{S})$ let now $C(P)$ denote the capacity of the channel $V(w|s)$ when the codewords are constrained to be of type P and let $\{c^{(i)}\}_{i=1}^{e^{nC(P)}} \subseteq \mathcal{S}^n$ denote a code-book (approximately) achieving the $C(P)$-capacity of this channel. For each $1 \le i \le e^{nC(P)}$ we have Vol(Typical set$_{c^{(i)}}$) $\approx e^{nH(W|S)}$, where S in the conditional entropy is distributed according to P. Since $\{c^{(i)}\}_{i=1}^{e^{nC(P)}}$ is a channel code, the {Typical set$_{c^{(i)}}$} are essentially disjoint.

[5]Throughout, P_{s^n} is the probability measure corresponding to the innovation W^n when the state sequence is s^n. This should not be confused with $p_{s^n} \in \mathcal{M}_n(\mathcal{S})$, the empirical measure of s^n.

Thus, we get

$$\mathrm{Vol}\left(\bigcup_{s^n\in\mathcal{S}^n}\text{Typical set}_{s^n}\right)\geq \mathrm{Vol}\left(\bigcup_{1\leq i\leq e^{nC(P)}}\text{Typical set}_{c^{(i)}}\right)\approx e^{nC(P)}e^{nH(W|S)}=e^{nH([P\times V]_{\mathcal{W}})}. \tag{38}$$

Combined with (34), this gives

$$H([P\times V]_{\mathcal{W}})\leq R+\phi(d), \tag{39}$$

which, by the arbitrariness of P, implies the converse part of Theorem 7. Note that here, the assumption that $V(\cdot|s)$ attains the SLB with equality is not needed.

To establish the direct part of Theorem 7 we construct a predictor set as follows. For each P and each of the code-words $\{c^{(i)}\}_{i=1}^{e^{n(C(P)+\varepsilon)}}$ we construct the predictor set achieving competitive predictability d_P for the source $P_{c^{(i)}}$, made up of $e^{nR(c^{(i)},d_P)}$ predictors, where $R(c^{(i)},d_P)$ denotes the R-D function (at distortion level d_P) of the innovation source indexed by $c^{(i)}$. This is essentially guaranteed to be possible by the direct part of Theorem 1. We let $\mathcal{G}_P$ be the predictor set obtained by the union of all these sets over $1\leq i\leq e^{n(C(P)+\varepsilon)}$ and let $\mathcal{G}$ denote the union of $\mathcal{G}_P$ over all types P. By a strong converse to channel-coding with constant-composition codes (which we prove below), the predictor set $\mathcal{G}_P$ is guaranteed of achieving competitive predictability value of at least d_P for all sources indexed by state sequences which are P-typical (as discussed above, following (36)).

Since for each P the corresponding $\{c^{(i)}\}$ have empirical distributions P, the $R(c^{(i)},d)$-s all equal $R(c^{(1)},d)$ and hence $|\mathcal{G}_P|\approx e^{n(C(P)+R(c^{(1)},d_P))}$. One can show that if for each s $R(V(\cdot|s),d)=H(V(\cdot|s))-\phi(d)$ (namely, achieves the SLB with equality), then $R(c^{(1)},d)=\left(\sum_{s\in\mathcal{S}}H(V(\cdot|s))p_{c^{(1)}}(s)\right)-\phi(d)=H(W|S)-\phi(d)$, where (S,W) are jointly distributed as the input-output pair of the channel $V(w|s)$ with the input distribution P. Thus $\frac{1}{n}\log|\mathcal{G}_P|\approx C(P)+R(c^{(1)},d_P)\approx H([P\times V]_{\mathcal{W}})-\phi(d_P)$, so that the rate of the predictor set $|\mathcal{G}|$ is essentially $R=\max_P[H([P\times V]_{\mathcal{W}})-\phi(d_P)]$. In particular, constructing a predictor set for $d_P=d$ for all P (or better yet, as we shall do in Section 6 and as discussed above, we take $d_p\leq d\stackrel{\text{def}}{=}\max_{P'}d_{P'}$ with $H([P\times V]_{\mathcal{W}})-\phi(d_P)$ constant in P), gives a worst-case distortion, over all sources $s^n\in\mathcal{S}^n$, of d and rate $\max_P H([P\times V]_{\mathcal{W}})-\phi(d)$. Consequently, the rate of this predictor set and the worst-case distortion it achieves, d, satisfy (37) with equality. This is the idea behind the construction, made precise in the formal proof of Theorem 7, to which Section 6 is dedicated.

5 Proofs for Results in Non-Universal Setting

A Proof of Theorem 1

Proof of lower bound: Fix d for which $R+\phi(d)<\underline{H}(\mathbf{X})$. It will be enough to show that the left side of (15) 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

$$\liminf_{n\to\infty} EL(\mathcal{G}_n, X^n) \geq d. \tag{40}$$

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

$$A_n = \{x^n : L(\mathcal{G}_n, x^n) \leq d\}. \tag{41}$$

By (27),

$$|A_n| \leq e^{n(R+\phi_n(d))}. \tag{42}$$

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

$$|A_n| \leq e^{n(\underline{H}(\mathbf{X})-\varepsilon)}. \tag{43}$$

The proof of the strong converse to the AEP for i.i.d. sources (cf., e.g., [14, 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 $|\tilde{A}_n| \leq e^{n(\underline{H}(\mathbf{X})-\varepsilon)}$, $\Pr(X^n \in \tilde{A}_n) \to 0$ (exponentially rapidly). Thus, by (43) and the strong converse,

$$\Pr(L(\mathcal{G}_n, X^n) \leq d) = \Pr(X^n \in A_n) \to 0. \tag{44}$$

The proof is completed by noting that (44) implies (40). □

Proof of the upper bound: Fix an arbitrary predictor $F \in \mathcal{F}$. It will suffice to establish the existence of a sequence $\{\mathcal{G}_n\}$ of predictor sets with $|\mathcal{G}_n| \leq e^{nR}$ for which

$$\limsup_{n\to\infty} EL(\mathcal{G}_n, X^n) \leq D(\mathbf{W}^F, R). \tag{45}$$

Recall first that rate-distortion theory guarantees the existence of a sequence of code-books $\{\mathcal{C}_n\}$, $\mathcal{C}_n \subseteq \mathcal{X}^n$ with $|\mathcal{C}_n| \leq e^{nR}$ (which we fix henceforth), for which

$$\limsup_{n\to\infty} E\left[\min_{\hat{w}^n \in \mathcal{C}_n} \frac{1}{n}\sum_{t=1}^{n} \rho(W_t^F - \hat{w}_t)\right] \leq D(\mathbf{W}^F, R). \tag{46}$$

Let now $\mathcal{G}_n$ be the predictor set induced by the code-book $\mathcal{C}_n$ and the predictor F (recall (28)). The predictor sets $\{\mathcal{G}_n\}$ satisfy $|\mathcal{G}_n| \leq e^{nR}$ and (29). This, by (46), implies (45) and completes the proof. □

B Proof of Theorem 3

Proof of the lower bound: The continuity of $H(\cdot)$ and $D(\cdot\|Q)$ imply that the right side of (19) 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 side of (19) 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

$$\limsup_{n\to\infty} \left[-\frac{1}{n}\log \Pr\left(L(\mathcal{G}_n, X^n) > d\right)\right] \leq D(P\|Q). \tag{47}$$

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

$$H(P^{(n)}) \geq R + \phi_n(d) + \varepsilon \tag{48}$$

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

$$\begin{aligned}
\Pr\left(L(\mathcal{G}_n, X^n) > d\right) &= \Pr\left(A_n^c\right) && (49)\\
&\geq \Pr\left(A_n^c \cap \{(W_1^F, \ldots, W_n^F) \in T_{P^{(n)}}\}\right) && (50)\\
&= \Pr\left(\{(W_1^F, \ldots, W_n^F) \in T_{P^{(n)}}\}\right) - \Pr\left(A_n \cap \{(W_1^F, \ldots, W_n^F) \in T_{P^{(n)}}\}\right) &&\\
&\geq (|T_{P^{(n)}}| - |A_n|)e^{-n[D(P^{(n)}\|Q)+H(P^{(n)})]} && (51)\\
&\geq ((n+1)^{-|\mathcal{X}|}e^{nH(P^{(n)})} - e^{n(R+\phi_n(d))})e^{-n[D(P^{(n)}\|Q)+H(P^{(n)})]} && (52)\\
&\geq ((n+1)^{-|\mathcal{X}|}e^{nH(P^{(n)})} - e^{n(H(P^{(n)})-\varepsilon)})e^{-n[D(P^{(n)}\|Q)+H(P^{(n)})]} && (53)\\
&\geq \frac{1}{2}(n+1)^{-|\mathcal{X}|}e^{-nD(P^{(n)}\|Q)}, && (54)
\end{aligned}$$

where A_n was defined in (41) and the c superscript denotes complementation. Inequality (51) follows from the 1-1 correspondence between source sequences and innovation sequences. Inequality (52) follows from the bound in (42), and inequality (53) follows from (48). Considering the two ends of the above chain implies that the left side of (47) is upper bounded by $\limsup_{n\to\infty} D(P^{(n)}\|Q)$ which, in turn, implies (47) by the continuity of $D(\cdot\|Q)$.

Proof of the upper bound: Assume first $R > 0$. It will suffice to establish the existence of a sequence $\{\mathcal{G}_n\}$ of predictor sets with $|\mathcal{G}_n| \leq e^{nR}$ for which

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left(L(\mathcal{G}_n, X^n) > d\right)\right] \geq F_d(R-0). \tag{55}$$

The proof proceeds similarly as that of the direct part of Theorem 1, where we construct the $\{\mathcal{G}_n\}$ induced by a sequence of rate-distortion code-books $\{\mathcal{C}_n\}$ and the predictor F. The only difference is that here we take a sequence which is optimal in Marton's error exponent sense. Specifically, we recall from [20, Theorem 1] the existence of a sequence of code-books $\{\mathcal{C}_n\}$, $\mathcal{C}_n \subseteq \mathcal{X}^n$ with $|\mathcal{C}_n| \leq e^{nR}$, for which

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left(\min_{\hat{w}^n\in\mathcal{C}_n}\frac{1}{n}\sum_{t=1}^{n}\rho(W_t^F - \hat{w}_t) > d\right)\right] \geq F_d(R-0). \tag{56}$$

Thus, the predictor set sequence $\{\mathcal{G}_n\}$ constructed via the sequence satisfying (56), by (29), satisfies (20) and completes the proof for $R > 0$. For the case $R = 0$, it will clearly suffice to establish the existence of a predictor, say G, for which

$$\liminf_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left(L_G(X^n) > d\right)\right] \geq \min_{P:E_P\rho(Z)\geq d} D(P\|Q). \tag{57}$$

Letting $G = F$ (F being the predictor corresponding to the auto-regressive presentation of $\mathbf{X}$ via the i.i.d. process $\mathbf{W}^F$), clearly $\{L_G(X^n) > d\} = \{1/n\sum_{i=1}^n \rho(W_i^F) > d\}$ so (57) holds for this choice of G by classical large deviations theory for i.i.d. random variables. □

C The Case $\mathcal{X} = \mathbb{R}$

The above proofs were suited for the case of $\mathcal{X}$ finite. Suppose now that $\mathcal{X} = \mathbb{R}$, fix an arbitrary predictor $F \in \mathcal{F}$, and consider the transformation taking the source sequence x^n into the sequence of prediction errors $e^n = (x_1 - F_1, x_2 - F_2(x_1), \ldots, x_n - F_n(x^{n-1}))$. The argumentation in subsection 3.C carries over verbatim to conclude that this transformation is one-to-one and onto. Furthermore, the Jacobian of this transformation is readily seen to be lower-triangular with diagonal entries all equal to 1, implying that this mapping is volume-preserving. Thus, we get the analogue of (26) (cf. [22, Theorem 5]),

$$\mathrm{Vol}\left(\{x^n : L_F(x^n) \le d\}\right) = \mathrm{Vol}\left(\left\{e^n : \frac{1}{n}\sum_{i=1}^n \rho(e_i) \le d\right\}\right) \stackrel{\mathrm{def}}{=} e^{n\phi_n(d)}, \tag{58}$$

leading to the analogue of (27),

$$\mathrm{Vol}\left(\{x^n : L(\mathcal{G}, x^n) \le d\}\right) \le |\mathcal{G}| e^{n\phi_n(d)}. \tag{59}$$

The above proofs carry over by replacing throughout (4) with (5), (27) with (59), $|\cdot|$ with $\mathrm{Vol}(\cdot)$, and entropies with differential entropies.

Note that the continuous differentiability of the predictors was needed for the existence of the Jacobian, though ultimately the point was the volume-preservation property of the transformation taking the source sequence into the prediction error sequence. For this property to reign, clearly less is needed, e.g., piecewise continuous differentiability. It is, in fact, the unproven conjecture of the authors that the volume-preservation property holds for *all* predictors (consisting of measurable functions).

D Tightness of the Large Deviations Bounds

Corollary 4 can be shown to follow from the "Pythagorean Theorem of the Divergence" [15]. In the proof that follows we derive it from first principles.

Proof of Corollary 4: Standard large deviations theory for i.i.d. random variables gives

$$\lim_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left(L_F(X^n) > d\right)\right] = \lim_{n\to\infty}\left[-\frac{1}{n}\log\Pr\left(\sum_{i=1}^n \rho(W_i^F) > d\right)\right] = \min_{P: E_P\rho(Z)\ge d} D(P\|Q). \tag{60}$$

Hence, by the converse part of Theorem 3 for $R = 0$, it will suffice to show that

$$\min_{P:H(P)\ge\phi(d)} D(P\|Q) = \min_{P:E_P\rho(Z)\ge d} D(P\|Q) = \begin{cases} D(Q_d\|Q) & \text{for } d > E_{Z\sim Q}\rho(Z) \\ 0 & \text{otherwise.}\end{cases} \tag{61}$$

Since (61) clearly holds for $d \le E_{Z\sim Q}\rho(Z)$, assume $d > E_{Z\sim Q}\rho(Z)$. As $\min_{P:H(P)\ge\phi(d)} D(P\|Q) \ge \min_{P:E_P\rho(Z)\ge d} D(P\|Q)$ (since by the max-entropy property (30) $H(P) \ge \phi(d)$ implies $E_P\rho(Z) \ge d$) and $H(Q_d) = \phi(d)$, it will suffice to show that $\min_{P:E_P\rho(Z)\ge d} D(P\|Q) = D(Q_d\|Q)$ or, in other words, that

$$D(P\|Q) \ge D(Q_d\|Q) \tag{62}$$

for any distribution P with $E_P\rho(W) \geq d$. Letting P denote any such distribution and β, β' the parameters for which $Q_d = Q^{(\beta')}$ and $Q = Q^{(\beta)}$ we have

$$\begin{aligned}
D(P\|Q^{(\beta)}) &= E_P \log \frac{P(W)}{Q^{(\beta)}(W)} && (63)\\
&= E_P \log \left[\frac{P(W)}{Q^{(\beta')}(W)} \frac{Q^{(\beta')}(W)}{Q^{(\beta)}(W)} \right] && (64)\\
&= D(P\|Q^{(\beta')}) + E_P[-\beta'\rho(W) + \lambda_\rho(\beta') + \beta\rho(W) - \lambda_\rho(\beta)] && (65)\\
&= D(P\|Q^{(\beta')}) + (\beta - \beta')E_P\rho(W) + \lambda_\rho(\beta') - \lambda_\rho(\beta) && (66)\\
&\geq D(P\|Q^{(\beta')}) + (\beta - \beta')d + \lambda_\rho(\beta') - \lambda_\rho(\beta) && (67)\\
&= D(P\|Q^{(\beta')}) + D(Q^{(\beta')}\|Q^{(\beta)}), && (68)
\end{aligned}$$

where the inequality follows since $E_P\rho(W) \geq d$ and $\beta - \beta' > 0$. □

Corollary 4 gives us the precise error exponent at rate $R = 0$ when the source has an auto-regressive representation via an i.i.d. innovation process with a maximum-entropy distribution. It shows that, in this case, the achiever of $\min_{P: E_P\rho(Z)\geq d} D(P\|Q)$ is a maximum-entropy distribution, implying it is also the achiever of $\min_{P: H(P)\geq \phi(d)} D(P\|Q)$ and, hence, the upper and lower bounds on the exponent, (19) and (20), coincide at $R = 0$. As discussed in subsection 3.B, for general $R \geq 0$ and for a similar reasoning, a sufficient condition for (19) and (20) 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. This leads to the precise characterization of the exponent.

Corollary 8 *Let $\mathcal{X} = \{0, 1\}$ and ρ be Hamming. Let $\mathbf{X}$ be a stochastic process for which there exists a predictor F such that $\mathbf{W}^F$ is an i.i.d. Bernoulli(q) ($q \leq 1/2$) process. Then for all $q \leq d \leq 1/2$*

$$\lim_{n\to\infty} \left[-\frac{1}{n} \log \min_{|\mathcal{G}|\leq \exp(nR)} \Pr\left(L(\mathcal{G}, X^n) > d\right) \right] = D(h^{-1}(h(d) + R)\|q), \tag{69}$$

where $D(p\|q)$ is an abbreviation of D(Bernoulli(p)$\|$Bernoulli(q)).

Another case for which the bounds in Theorem 3 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., [4]). 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 3. Consequently, Theorem 3 gives the precise large deviations asymptotics for the competitive predictability of any process having an auto-regressive 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 side of (25) is nothing but $D(N(0, \tilde{\sigma}^2)\|N(0, \sigma^2))$, $\tilde{\sigma}^2$ tuned such that $R(N(0, \tilde{\sigma}^2), d) = R$).

6 Proof of Main Result in the Universal Setting

We now prove Theorem 7, where $\mathcal{X}$ and $\mathcal{S}$ are assumed finite. Theorem 7 is a direct consequence of the following two lemmas (recall Definition 1 for the meaning of "achievable pair").

Lemma 9 *Let $P^{(n)} \in \mathcal{M}_n(\mathcal{S})$, be such that $P^{(n)} \to P$, $P \in \mathcal{M}(\mathcal{S})$. If (R,d) is achievable w.r.t. $\{T_{P^{(n)}}\}$ then*

$$H\left([P \times V]_{\mathcal{W}}\right) \leq R + \phi(d). \tag{70}$$

Clearly, if (R,d) is achievable w.r.t. $\{\mathcal{S}^n\}$ then it is achievable w.r.t. any $\{T_{P^{(n)}}\}_{n\geq 1}$ as in Lemma 9 and hence the converse part of Theorem 7 follows.

For any $R \geq 0$ and $s^n \in T_P$ we define $d^R_{s^n} \stackrel{\text{def}}{=} \phi^{-1}([H\left([P \times V]_{\mathcal{W}}\right) - R]_+)$. The direct part of Theorem 7 is a consequence of the next lemma.

Lemma 10 *Suppose that $V(\cdot|s)$ attains the SLB with equality for all $s \in \mathcal{S}$. Fix an arbitrary $\varepsilon > 0$ and suppose that $\max_{P\in\mathcal{M}(\mathcal{S})} H\left([P \times V]_{\mathcal{W}}\right) - \phi(d) + \varepsilon \leq R$. There exists a sequence of predictor sets $\{\mathcal{G}_n\}$ with $|\mathcal{G}_n| \leq e^{n(R+\varepsilon)}$ such that*

$$\lim_{n\to\infty} \min_{s^n\in\mathcal{S}^n} P_{s^n}\left(L(\mathcal{G}_n, X^n) \leq d^R_{s^n}\right) = 1. \tag{71}$$

Note that for fixed R, if d satisfies (37), then $d^R_{s^n} \leq d$ for all $s^n \in \mathcal{S}^n$. Hence, Lemma 10 implies that if d satisfies (37) then (R,d) is in the achievable region, namely, the direct part of Theorem 7. The Lemma, however, is somewhat stronger and more informative as it implies the existence of rate-R predictor sets such that the achievable distortion is below the worst-case distortion for many sources. Specifically, for such predictor sets, if the state sequence belongs to type P then the distortion achieved is $\phi^{-1}([H\left([P \times V]_{\mathcal{W}}\right) - R]_+)$ which, for most types, will be less than the d dictated by (37).

Proof of Lemma 9: Fix a pair (R,d) which is achievable w.r.t. $\{T_{P^{(n)}}\}$ and an arbitrary $\varepsilon > 0$. By Definition 1 there exists a sequence of predictor sets $\{\mathcal{G}_n\}$ with $|\mathcal{G}_n| \leq e^{nR}$ and an n_0 such that for all $n \geq n_0$ and $s^n \in T_{P^{(n)}}$

$$P_{s^n}\left(L(\mathcal{G}_n, X^n) \leq d\right) \geq 1 - \varepsilon. \tag{72}$$

On the other hand, for each $s^n \in \mathcal{S}^n$

$$\begin{aligned} P_{s^n}\left(L(\mathcal{G}_n, X^n) \leq d\right) &\leq P_{s^n}\left(W^n \in T^c_{[V]}(s^n)\right) + P_{s^n}\left(\{W^n \in T_{[V]}(s^n)\} \cap \{L(\mathcal{G}_n, X^n) \leq d\}\right) \\ &\leq \frac{|\mathcal{X}||\mathcal{S}|}{4n\delta_n^2} + P_{s^n}\left(\{W^n \in T_{[V]}(s^n)\} \cap \{L(\mathcal{G}_n, X^n) \leq d\}\right). \end{aligned} \tag{73}$$

Hence, by (72) and (73), for $s^n \in T_{P^{(n)}}$

$$P_{s^n}\left(\{W^n \in T_{[V]}(s^n)\} \cap \{L(\mathcal{G}_n, X^n) \leq d\}\right) \geq 1 - \varepsilon - \frac{|\mathcal{X}||\mathcal{S}|}{4n\delta_n^2}. \tag{74}$$

From (74) and [16, Lemma 2.14] it follows that for $s^n \in T_{P^{(n)}}$

$$\frac{1}{n}\log|T_{[V]}(s^n)\cap\{W^n : L(\mathcal{G}_n, X^n)\leq d\}| \geq H(V|P^{(n)}) - \varepsilon_n, \tag{75}$$

where $\varepsilon_n \to 0$ is independent of s^n. For an arbitrary $\delta > 0$, the fact that $P^{(n)} \to P$, combined with Lemma 14 of the Appendix, imply the existence of $\{s^n_{(i)}\}_{i=1}^{e^{n(I(P;V)-\delta)}} \subseteq T_{P^{(n)}}$ such that

$$\left|T_{[V]}(s^n_{(i)})\cap\bigcup_{j\neq i}T_{[V]}(s^n_{(j)})\right| \leq \exp\left(n[H(V|P) - I(P;V) + I(P;V) - \frac{\delta}{2} + \eta_n]\right) \leq \exp\left(n[H(V|P) - \frac{\delta}{3}]\right), \tag{76}$$

provided that $n \geq n_1(|\mathcal{S}|, |\mathcal{X}|, \delta)$. Note that it follows from (76) and (75) that for all $1 \leq i \leq e^{n(I(P;V)-\delta)}$

$$\left|\{W^n : L(\mathcal{G}_n, X^n)\leq d\}\cap\left\{T_{[V]}(s^n_{(i)})\setminus\bigcup_{j\neq i}T_{[V]}(s^n_{(j)})\right\}\right| \geq e^{n(H(V|P)-\delta)} \tag{77}$$

provided $n \geq n_2(|\mathcal{S}|, |\mathcal{X}|, \delta)$. Consequently,

$$\begin{aligned}
|\{W^n : L(\mathcal{G}_n, X^n)\leq d\}| &\geq \left|\{W^n : L(\mathcal{G}_n, X^n)\leq d\}\cap\bigcup_i T_{[V]}(s^n_{(i)})\right| &(78)\\
&\geq \left|\{W^n : L(\mathcal{G}_n, X^n)\leq d\}\cap\bigcup_i \left\{T_{[V]}(s^n_{(i)})\setminus\bigcup_{j\neq i}T_{[V]}(s^n_{(j)})\right\}\right| &(79)\\
&= \sum_i\left|\{W^n : L(\mathcal{G}_n, X^n)\leq d\}\cap\left\{T_{[V]}(s^n_{(i)})\setminus\bigcup_{j\neq i}T_{[V]}(s^n_{(j)})\right\}\right| &(80)\\
&\geq e^{n(I(P;V)-\delta)}e^{n(H(V|P)-\delta)} = e^{n(H([P\times V]_{\mathcal{W}})-2\delta)}, &(81)
\end{aligned}$$

where (81) follows from (77). On the other hand, in the proof of Theorem 1 it was seen that $|\{W^n : L(\mathcal{G}_n, X^n)\leq d\}| \leq e^{n(R+\phi_n(d))}$. Combined with (81), this implies that for all sufficiently large n

$$R + \phi_n(d) \geq H\left([P\times V]_{\mathcal{W}}\right) - 2\delta. \tag{82}$$

Letting $n \to \infty$, the arbitrariness of $\delta > 0$ completes the proof. □

A Proof of Lemma 10

We begin with a strong converse for constant-composition codes.

Lemma 11 *For any $P \in \mathcal{M}_n(\mathcal{S})$, let $\{s^n_{(i)}\}_{i=1}^M \subseteq T_P$ be an arbitrary code-book of composition P. Let P_C be the associated probability of correct decision under the optimal (maximum-likelihood) decision rule for the channel $V(w|s)$. If $M \geq e^{n(I(P;V)+2\varepsilon)}$ then*

$$P_C \leq e^{-n\varepsilon} + \frac{K}{n\varepsilon^2}, \tag{83}$$

where $K = K(V)$ is a constant dependent only on the channel V.

Proof: Let $\{s^n_{(i)}\}_{i=1}^M \subseteq T_P$ be a code-book of composition P. Let $\{\Lambda_i\}_{i=1}^M$ be a partition of $\mathcal{W}^n$ induced by an optimal (maximum-likelihood) decision rule associated with this code-book and the channel V. Finally, let $P_{e|i}$ denote the probability of error when the channel input is $s^n_{(i)}$. For any $q \in \mathcal{M}(\mathcal{W})$ and $I \geq 0$

$$\begin{aligned}
\frac{e^{nI}}{M} &= \frac{e^{nI}}{M} \sum_{i=1}^M \sum_{w^n \in \Lambda_i} q^n(w^n) && (84)\\
&\geq \frac{e^{nI}}{M} \sum_{i=1}^M \left(\sum_{w^n \in \Lambda_i \cap \{V(w^n|s^n_{(i)}) \leq q^n(w^n)e^{nI}\}} q^n(w^n) \right) && (85)\\
&\geq \frac{1}{M} \sum_{i=1}^M \left(\sum_{w^n \in \Lambda_i \cap \{V(w^n|s^n_{(i)}) \leq q^n(w^n)e^{nI}\}} V(w^n|s^n_{(i)}) \right) && (86)\\
&= \frac{1}{M} \sum_{i=1}^M \left[1 - \Pr\left(\text{error} \cup \{V(W^n|s^n_{(i)}) > q^n(W^n)e^{nI}\} | s^n_{(i)} \right) \right] && (87)\\
&\geq \frac{1}{M} \sum_{i=1}^M \left[1 - P_{e|i} - \Pr\left(V(W^n|s^n_{(i)}) > q^n(W^n)e^{nI} | s^n_{(i)} \right) \right] && (88)\\
&= P_C - \Pr\left(\frac{V(W^n|s^n_{(1)})}{q^n(W^n)} > e^{nI} \middle| s^n_{(1)} \right), && (89)
\end{aligned}$$

where the last equality follows since $\Pr\left(V(W^n|s^n) > q^n(W^n)e^{nI}|s^n\right)$ depends on s^n only through its type. Thus, for $M = e^{nR}$, we get

$$P_C \leq e^{-n(R-I)} + \Pr\left(\frac{1}{n} \sum_{i=1}^n \log \frac{V(W_i|s_i)}{q(W_i)} > I \middle| s^n \right), \tag{90}$$

where $s^n \in T_P$. In particular, taking $q = [P \times V]_{\mathcal{W}}$, the expectation of $\frac{1}{n}\sum_{i=1}^n \log \frac{V(W_i|s_i)}{q(W_i)}$ for a state sequence $s^n \in T_P$ is precisely $I(P;V)$ and its variance is K/n, K being a constant dependent on the channel V only. Thus, letting $I = I(P;V) + \varepsilon$, we obtain by Chebyshev's inequality

$$P_C \leq e^{-n(R-I(P;V)-\varepsilon)} + \frac{K}{n\varepsilon^2}.\square \tag{91}$$

The significance of Lemma 11, for our problem, is that it implies that the typical set of any source P_{s^n}, $s^n \in T_P$, is covered by the union of the typical sets of the sources indexed by the words $\{s^n_{(i)}\}_{i=1}^M \subseteq T_P$. More precisely:

Lemma 12 *For any $P \in \mathcal{M}_n(\mathcal{S})$ and $\varepsilon > 0$ there exists a set $\{s^n_{(i)}\}_{i=1}^M \subseteq T_P$, $M = e^{n(I(P;V)+\varepsilon)}$, such that for* every $s^n \in T_P$

$$P_{s^n}\left(W^n \in \bigcup_i T_{[V]}(s^n_{(i)}) \right) \geq 1 - \alpha_n, \tag{92}$$

where $\alpha_n \to 0$ is a sequence which depends only on ε and V.

Proof: For any code-book $\{\tilde{s}^n_{(i)}\}_{i=1}^M \subseteq \mathcal{S}^n$, consider the following sub-optimal decoding rule:

$$\hat{i}(W^n) = \arg\min\{1 \le i \le M : W^n \in T_{[V]}(\tilde{s}^n_{(i)})\}, \tag{93}$$

where an error is declared if the set on the right side is empty. For $P \in \mathcal{M}_n(\mathcal{S})$, let $\tilde{P}^n_E(P, M)$ denote the minimal probability of error among all code-books of size M which have constant composition P, i.e., $\{\tilde{s}^n_{(i)}\}_{i=1}^M \subseteq T_P$, when the decoding rule in (93) is used. It is easy to verify that $\tilde{P}^n_E(P, M)$ is monotonically increasing in M as, given a code-book of any size, a code-book with lower probability of error (under the decoding rule of (93)) results by expurgating that word in the code-book with highest $P_{e|i}$. Now, Lemma 11 implies that

$$\tilde{P}^n_E(P, e^{n(I(P;V)+\varepsilon)}) \ge 1 - e^{-n\varepsilon/2} - \frac{4K}{n\varepsilon^2}, \tag{94}$$

where K is a constant depending on V only. Let $\{s^n_{(i)}\}_{i=1}^M \subseteq T_P$, $M = e^{n(I(P;V)+\varepsilon)}$, be the code-book achieving $\tilde{P}^n_E(P, e^{n(I(P;V)+\varepsilon)})$. It will suffice to show that for every $s^n \in T_P$

$$P_{s^n}\left(W^n \in \bigcup_{i=1}^M T_{[V]}(s^n_{(i)})\right) \ge 1 - e^{-n\varepsilon/2} - \frac{4K}{n\varepsilon^2} - 2\varepsilon_n, \tag{95}$$

where $\varepsilon_n \to 0$ is the sequence from [16, Lemma 2.12]. Assume, by contradiction, the existence of $\tilde{s}^n \in T_P$ for which

$$P_{\tilde{s}^n}\left(W^n \in \bigcup_{i=1}^M T_{[V]}(s^n_{(i)})\right) < 1 - e^{-n\varepsilon/2} - \frac{4K}{n\varepsilon^2} - 2\varepsilon_n \tag{96}$$

and let $\{s^n_{(i)}\}_{i=1}^{M+1}$ be the code-book obtained by appending to the code-book achieving $\tilde{P}^n_E(P, e^{n(I(P;V)+\varepsilon)})$ the code-word $s^n_{(M+1)} = \tilde{s}^n$. This gives a code-book of size $M+1$ whose probability of error under decoding rule (93) is given by

$$\frac{M}{M+1}\tilde{P}^n_E(P, M) + \frac{1}{M+1}P_{e|M+1}, \tag{97}$$

where $P_{e|M+1}$ is the probability of error given that $s^n_{(M+1)} = \tilde{s}^n$ was transmitted, so that

$$\begin{aligned}
P_{e|M+1} &= \Pr\left(\left\{W^n \in \bigcup_{i=1}^M T_{[V]}(s^n_{(i)})\right\} \cup \left\{W^n \in T^c_{[V]}(s^n_{(M+1)})\right\} | s^n_{(M+1)}\right) && (98)\\
&\le \Pr\left(W^n \in \bigcup_{i=1}^M T_{[V]}(s^n_{(i)}) | s^n_{(M+1)}\right) + \Pr\left(W^n \in T^c_{[V]}(s^n_{(M+1)}) | s^n_{(M+1)}\right) && (99)\\
&= P_{\tilde{s}^n}\left(W^n \in \bigcup_{i=1}^M T_{[V]}(s^n_{(i)})\right) + \Pr\left(W^n \in T^c_{[V]}(s^n_{(M+1)}) | s^n_{(M+1)}\right) && (100)\\
&\le 1 - e^{-n\varepsilon/2} - \frac{4K}{n\varepsilon^2} - 2\varepsilon_n + \varepsilon_n && (101)\\
&< \tilde{P}^n_E(P, M), && (102)
\end{aligned}$$

where the last inequality follows from (94). This implies that the expression in (97) is strictly less than $\tilde{P}^n_E(P, M)$. On the other hand, the expression in (97) is the probability of error associated with a code-book of size $M+1$ so it is *lower* bounded by $\tilde{P}_E^n(P, M+1)$, thus contradicting the fact that $\tilde{P}_E^n(P, M)$ is increasing in M. $\square$

We are now almost in a position to give the proof of Lemma 10 via the construction of the predictor set outlined in subsection A. We first need to state the following version of the type covering lemma of [16] to the case of the AVS.

Lemma 13 *Assume that $V(\cdot|s)$ attains the SLB with equality for all $s \in \mathcal{S}$. For $d \geq 0$ and $\varepsilon > 0$ we have the following. For every $P \in \mathcal{M}_n(\mathcal{S})$ and $s^n \in T_P$ there exists a set $B \subseteq \mathcal{X}^n$ such that*

$$\rho(w^n, B) \stackrel{\text{def}}{=} \min_{\tilde{w}^n \in B} \rho(w^n, \tilde{w}^n) \leq d \quad \textit{for every} \quad w^n \in T_{[V]}(s^n) \tag{103}$$

and

$$\frac{1}{n} \log |B| \leq H(V|P) - \phi(d) + \varepsilon \tag{104}$$

provided that $n \geq n_0(\rho, \varepsilon)$.

Noting that $H(V|P) - \phi(d)$ is nothing but the rate-distortion function of W^n when distributed under P_{s^n}, the proof is a straightforward extension of that of [16, Lemma 4.1] and is therefore omitted.

Proof of Lemma 10: Fix an arbitrary $\varepsilon > 0$ and $R \geq \max_{P \in \mathcal{M}(\mathcal{S})} H([P \times V]_{\mathcal{W}}) - \phi(d) + \varepsilon$. Note that $R \geq I(P;V) + \varepsilon$ for all $P \in \mathcal{M}(\mathcal{S})$ as $\max_{P \in \mathcal{M}(\mathcal{S})} H([P \times V]_{\mathcal{W}}) - \phi(d) \geq H([P \times V]_{\mathcal{W}}) - \phi(d) = I(P;V) + H(V|P) - \phi(d) \geq I(P;V)$. Note also that $H(V|s) - \phi(d) \geq 0$ for all s (as the left side is the rate-distortion function of $V(\cdot|s)$) and, consequently, $H(V|P) - \phi(d) \geq 0$. For each type $P \in \mathcal{M}_n(\mathcal{S})$ we construct a predictor set $\mathcal{G}_n(P)$ as follows. Take the set $\{s_{(i)}^n\}_{i=1}^M \subseteq T_P$, $M = e^{n(I(P;V)+\varepsilon)}$ from Lemma 12. For each code-word in that set $s_{(i)}^n \in T_P$, let the predictor set, $\mathcal{G}(s_{(i)}^n)$, be that induced by the set B from Lemma 13 corresponding to distortion level $d = d^R_{s^n_{(i)}}$ and to $s^n = s^n_{(i)}$, i.e., the set B satisfying

$$\rho(w^n, B) \leq d^R_{s^n_{(i)}} \quad \text{for every} \quad w^n \in T_{[V]}(s^n_{(i)}) \tag{105}$$

and

$$\frac{1}{n} \log |B| \leq H(V|P) - \phi(d^R_{s^n_{(i)}}) + \varepsilon = H(V|P) - [H([P \times V]_{\mathcal{W}}) - R]_+ + \varepsilon \leq R - I(P;V) + \varepsilon \tag{106}$$

Note that Lemma 13 asserts the existence of the set B satisfying (105) and (106) for n sufficiently large, dependent only on ε. So we assume henceforth that n is sufficiently large. Let now $\mathcal{G}_n(P) = \bigcup_{i=1}^{e^{n(I(P;V)+\varepsilon)}} \mathcal{G}(s^n_{(i)})$. Clearly, by (106), $|\mathcal{G}_n(P)| \leq e^{n(I(P;V)+\varepsilon)} e^{n(R - I(P;V)+\varepsilon)} = e^{n(R+2\varepsilon)}$. Finally, let $\mathcal{G}_n = \bigcup_{P \in \mathcal{M}_n(\mathcal{S})} \mathcal{G}_n(P)$, so that we have $|\mathcal{G}_n| \leq \exp\{n(R + 2\varepsilon + \frac{1}{n}|\mathcal{S}|\log(n+1))\}$. Now, for every $P \in \mathcal{M}_n(\mathcal{S})$ and $s^n \in T_P$,

$$\begin{aligned} P_{s^n}\left(L(\mathcal{G}_n, X^n) \leq d^R_{s^n}\right) &\geq P_{s^n}\left(L(\mathcal{G}_n(P), X^n) \leq d^R_{s^n}\right) && (107) \\ &\geq P_{s^n}\left(W^n \in \bigcup_i T_{[V]}(s^n_{(i)})\right), && (108) \end{aligned}$$

where $s^n_{(i)}$ are those through which $\mathcal{G}_n(P)$ was constructed and the inequality in (108) follows since, by the construction of $\mathcal{G}(s^n_{(i)})$, if the innovation vector w^n lies in $T_{[V]}(s^n_{(i)})$ for some i then there exists a predictor $G \in \mathcal{G}(s^n_{(i)})$ for which $L_G(x^n) \leq d^R_{s^n_{(i)}}$ (x^n being the source sequence associated with the innovation vector w^n). Combining (108) with Lemma 12 gives $\max_{s^n \in \mathcal{S}^n} P_{s^n}\left(L(\mathcal{G}_n, X^n) \leq d^R_{s^n}\right) \geq 1 - \alpha_n$ and completes the proof. $\Box$

7 Future Directions

The most interesting question remaining open in the context of this work is whether the lower bound on competitive predictability in Lemma 1 is tight in general, even when the rate-distortion function of the innovation process does not attain the SLB with equality. If tight, this would imply that, in general, predictor sets induced by rate-distortion code-books are strictly sub-optimal, as one would intuitively suspect. If the lower bound is not tight in general then the question of the optimality of predictor sets induced by rate-distortion code-books would still remain open and interesting.

Another interesting direction would be to extend the scope of the problem to general, non-difference, distortion measures, where it is unclear if and how the volume-preservation arguments can be applied.

Appendix

A Sphere-Packing Lemma

Lemma 14 *For every $R > 0$, $\delta > 0$, and every $P \in \mathcal{M}_n(\mathcal{S})$ satisfying $H(P) > R$, there exist $e^{n(R-\delta)}$ distinct sequences $s^n_{(i)} \in T_P$ such that for every pair of stochastic matrices $V : \mathcal{S} \to \mathcal{X}$, $\hat{V} : \mathcal{S} \to \mathcal{X}$ and every i*

$$\left| T_{[V]}(s^n_{(i)}) \cap \bigcup_{j \neq i} T_{[\hat{V}]}(s^n_{(j)}) \right| \leq \exp\Big(n[H(V|P) - I(P;\hat{V}) + R + \eta_n]\Big), \tag{A.1}$$

provided that $n \geq n_0(|\mathcal{S}|, |\mathcal{X}|, \delta)$, where $\eta_n \to 0$ depends only on $|\mathcal{S}|$ and $|\mathcal{X}|$.

Note that Lemma 14 is essentially [16, Lemma 5.1]. The difference is in that here we take $T_{[V]}$ instead of T_V.

Proof of Lemma 14: [16, Lemma 5.1] guarantees that for every $R > 0$, $\delta > 0$, and every $P \in \mathcal{M}_n(\mathcal{S})$ satisfying $H(P) > R$, there exist $e^{n(R-\delta)}$ distinct sequences $s^n_{(i)} \in T_P$ such that for every pair of stochastic matrices $U : \mathcal{S} \to \mathcal{X}$, $\hat{U} : \mathcal{S} \to \mathcal{X}$ and every i

$$\left| T_U(s^n_{(i)}) \cap \bigcup_{j \neq i} T_{\hat{U}}(s^n_{(j)}) \right| \leq \exp\Big(n[H(U|P) - I(P;\hat{U}) + R + \varepsilon_n]\Big), \tag{A.2}$$

provided that $n \geq n_0(|\mathcal{S}|, |\mathcal{X}|, \delta)$, where $\varepsilon_n \to 0$ is our sequence from the ε-convention. Now, for the given V (resp. $\hat{V}$), $T_{[V]}(\cdot)$ (resp. $T_{[\hat{V}]}(\cdot)$) is the union of at most $(n+1)^{|\mathcal{S}||\mathcal{X}|}$ disjoint U-shells (resp. $\hat{U}$-shells) $T_U(\cdot)$ (resp. $T_{\hat{U}}(\cdot)$). Hence, taking the set $\{s^n_{(i)}\}_{i=1}^{e^{n(R-\delta)}}$ that satisfies (A.2) we have for all i

$$\left| T_{[V]}(s^n_{(i)}) \cap \bigcup_{j \neq i} T_{[\hat{V}]}(s^n_{(j)}) \right| \tag{A.3}$$

$$= \left| \left(\bigcup_U T_U(s^n_{(i)}) \right) \cap \left(\bigcup_{\hat{U}} \bigcup_{j \neq i} T_{\hat{U}}(s^n_{(j)}) \right) \right| \tag{A.4}$$

$$= \left| \bigcup_U \bigcup_{\hat{U}} \left(T_U(s^n_{(i)}) \cap \bigcup_{j \neq i} T_{\hat{U}}(s^n_{(j)}) \right) \right| \tag{A.5}$$

$$\leq (n+1)^{2|\mathcal{S}||\mathcal{X}|} \exp\left(n[\max_{U,\hat{U}}(H(U|P) - I(P;\hat{U})) + R + \varepsilon_n] \right), \tag{A.6}$$

where the unions and maximization are over those U and $\hat{U}$ associated with $T_{[V]}$ and $T_{[\hat{V}]}$, respectively. Standard continuity arguments give (cf., e.g., proof of [16, Lemma 2.13])

$$|(H(V|P) - I(P;\hat{V})) - \max_{U,\hat{U}}(H(U|P) - I(P;\hat{U}))| \leq -2|\mathcal{S}||\mathcal{X}|\delta_k \log \delta_k, \tag{A.7}$$

where $\{\delta_k\}$ is the sequence from the δ-convention. Combining (A.7) with (A.6) completes the proof. $\square$

References

[1] S. T. Alexander. A simple noniterative speech excitation algorithm using the lpc residual. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, ASSP–33(2):432–435, April 1985.

[2] P. H. Algoet. The strong law of large numbers for sequential decisions under uncertainty. *IEEE Trans. Inform. Theory*, 40(3):609–633, May 1994.

[3] P. H. Algoet. Universal schemes for learning the best nonlinear predictor given the infinite past and side information. *IEEE Trans. Inform. Theory*, 45(4):1165–1185, May 1999.

[4] E. Arikan and N. Merhav. Guessing subject to distortion. *IEEE Trans. Inform. Theory*, IT-44(3):1041–1056, May 1998.

[5] B. S. Atal. High quality speech at low bit rates: Multi-pulse and stochastically excited linear predictive coders. *Proc. 1986 IEEE-IECE-J-ASJ*, 1986.

[6] A. Barron, J. Rissanen, and B. Yu. The minimum description length principle in coding and modeling. *IEEE Trans. Inform. Theory*, 44(6):2743–2760, October 1998.

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

[8] T. Berger. Lossy source coding. *IEEE Trans. Inform. Theory*, 44(6):2693–2723, October 1998.

[9] M. Berouti, H. Garten, P. Kabal, and P. Mermelstein. Efficient computation and encoding of the multipulse excitation for lpc. *Proc. ICASSP*, 1984.

[10] N. Cesa-Bianchi, Y. Freund, D.P. Helmbold, D. Haussler, R. Schapire, and M.K. Warmuth. How to use expert advice. *Journal of the ACM*, 44(3):427–485, 1997.

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

[12] N. Cesa-Bianchi and G. Lugosi. On sequential prediction of individual sequences relative to a set of experts. *Ann. Stat.*, 27(6):1865–1895, 1999.

[13] N. Cesa-Bianchi and G. Lugosi. Potential-based algorithms in on-line prediction and game theory. *Proceedings of the 14th Annual conference on Computational Learning Theory*, July 2001.

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

[15] I. Csiszar. I-divergence geometry of probability distributions and minimization problems. *Annals of Probability*, 3(1):146–158, February 1975.

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

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

[18] R. G. Gallager. *Information Theory and Reliable Communication.* Wiley, New York, 1968.

[19] D. Haussler, J. Kivinen, and M.K. Warmuth. Sequential prediction of individual sequences under general loss functions. *IEEE Trans. Inform. Theory*, 44:1906–1925, September 1998.

[20] K. Marton. Error exponent for source coding with a fidelity criterion. *IEEE Trans. Inform. Theory*, IT-20:197–199, March 1974.

[21] N. Merhav and M. Feder. Universal prediction. *IEEE Trans. Inform. Theory*, IT-44(6):2124–2147, October 1998.

[22] N. Merhav and T. Weissman. Scanning and prediction in multi-dimensional data arrays. *IEEE Trans. Inform. Theory*, August 2001. Submitted.

[23] A. Puhalskii and V. Spokoiny. On large deviation efficiency in statistical inference. *Bernoulli*, 4(2):203–272, 1998.

[24] J. Rissanen. Universal coding, information, prediction, and estimation. *IEEE Trans. Inform. Theory*, IT-30(4):629–636, July 1984.

[25] Kenneth Rose. A mapping approach to rate-distortion computation and analysis. *IEEE Trans. Inform. Theory*, 40(6):1939–1952, November 1994.

[26] I. M. Transoco and B. S. Atal. Efficient procedures for finding the optimal innovation in stochastic coders. *Proc. ICASSP–86*, pages 2375–2378, 1986.

[27] V. Vovk. Aggregating strategies. *Proc. 3rd Annu. Workshop on Computational Learning Theory*, pages 371–383, 1990. San Mateo, CA: Kaufmann.

[28] V. Vovk. A game of prediction with expert advice. *Proc. 8th Annu. Workshop on Computational Learning Theory.*, pages 51–60, 1995. New York NY.

[29] M.J. Weinberger, N. Merhav, and M. Feder. Optimal sequential probability assignment for individual sequences. *IEEE Trans. Inform. Theory*, 40:384–396, March 1994.

[30] T. Weissman and N. Merhav. Universal prediction of individual binary sequences in the presence of noise. *IEEE Trans. Inform. Theory*, 47(6):2151–2173, September 2001.