Guessing Subject to Distortion *

Erdal Arikan† Neri Merhav‡

July 23, 1998

Abstract

We investigate the problem of guessing a random vector $\mathbf{X}$ within distortion level D. Our aim is to characterize the best attainable performance in the sense of minimizing, in some probabilistic sense, the number of required guesses $G(\mathbf{X})$ until the error falls below D. The underlying motivation is that $G(\mathbf{X})$ is the number of candidate code words to be examined by a rate-distortion block encoder until a satisfactory code word is found. In particular, for memoryless sources, we provide a single-letter characterization of the least achievable exponential growth rate of the ρth moment of $G(\mathbf{X})$ as the dimension of the random vector $\mathbf{X}$ grows without bound. In this context, we propose an asymptotically optimal guessing scheme that is universal both w.r.t. the information source and the value of ρ. We then study some properties of the exponent function $E(D,\rho)$ along with its relation to the source coding exponents. Finally, we provide extensions of our main results to the Gaussian case, guessing with side information, and sources with memory.

Index Terms: rate-distortion theory, fidelity criterion, source coding, guessing, source coding error exponent, side information.

*The work of N. Merhav was partially supported by the Wolfson Research Awards administered by the Israel Academy of Sciences and Humanities.

†Electrical-Electronics Engineering Department, Bilkent University, 06533 Ankara, Turkey.

‡Was on sabbatical leave at HP Labs, Palo Alto, CA 94304, USA. Now with the Department of Electrical Engineering and HP-ISC, Technion -I.I.T, Haifa 32000, Israel.

1 Introduction

Consider the following game: Bob draws a sample x from a random variable X. Then, Alice, who does not see x but wishes to learn it at least approximately, presents to Bob a (fixed) sequence of guesses $\hat{x}(1), \hat{x}(2), \ldots$. Bob checks the guesses successively until a guess $\hat{x}(i)$ is found such that $d(x, \hat{x}(i)) \leq D$ for some distortion measure d and distortion level D. Bob informs Alice of $\hat{x}(i)$ and in return Alice pays Bob an amount $G(x) = i$ equal to the number of guesses examined by Bob. What is the best Alice can do in designing a clever guessing list $\{\hat{x}(1), \hat{x}(2), ...\}$ so as to minimize the typical number of guesses $G(X)$ in some probabilistic sense? For the discrete distortionless case ($D = 0$), it is easy to see [2] that if the probability distribution P of X is known to Alice, the best she can do is simply to order her guesses according to decreasing probabilities. The extension to $D > 0$, however, seems to be more involved.

This game may serve as a model for certain betting games in which a player places a number of bets concerning the outcome of a chance event X, such as a horse race, and receives a payoff for each bet that is close enough to the actual outcome. The expected number of guesses $\mathbf{E}G(X)$ may serve as a measure of the number of bets to be placed for a fair chance of winning a payoff. This model may also be useful for studying pattern-matching and database search algorithms. Another motivation in studying this problem is its natural relevance to rate-distortion coding. Suppose that the random variable X to be guessed is a random N-vector $\mathbf{X}$, drawn by an information source, and to be encoded by a rate-distortion codebook. The number of guesses $G(\mathbf{X})$ is then interpreted as the number of candidate codebook vectors to be examined (and hence also the number of metric computations) before a satisfactory code word is found. It should be emphasized, however, that $G(\mathbf{X})$ indeed measures the search complexity only for a simple search algorithm that scans the codebook in a fixed order. In reality, the difference between the guessing problem and the search problem of lossy coding, is that in the latter, after each 'guess', we know the exact distortion, and not only whether or not it is below the desired threshold D. Therefore, in this context, the motivation of the guessing problem as a rate-distortion search problem should be considered relevant only w.r.t. this class of simple search schemes. Nevertheless, it serves as a first step towards possible further extensions that include classes of more sophisticated search algorithms (see also Section 7 below).

In an earlier related work, driven by a similar motivation among others, Merhav [14] has characterized the maximum achievable expectation of the number of code words that are within distance D from a randomly chosen source vector $\mathbf{X}$. The larger this number is, the easier it is, typically, to find quickly a suitable code word. In a more closely related work, Arikan [2] studied the guessing problem for discrete memoryless sources (DMS's) in the lossless case ($D = 0$). In particular, Arikan developed a single-letter characterization of the smallest attainable exponential growth rate of the ρth moment of the number of guesses $\mathbf{E}G(\mathbf{X})^\rho$ (ρ being an arbitrary nonnegative real) as the vector dimension N tends to infinity.

This work is primarily aimed at extending Arikan's study [2] to the lossy case $D > 0$, which is more difficult as mentioned above. In particular, our first result in Section 3 is that for a finite alphabet memoryless source P, the best attainable behavior of $\mathbf{E}G(\mathbf{X})^\rho$ is of the exponential order of $e^{NE(D,\rho)}$, where $E(D,\rho)$ is referred to as the *ρth order guessing exponent at distortion level D* (or simply, the *guessing exponent*), and given by

$$E(D,\rho) = \max_Q [\rho R(D,Q) - D(Q||P)], \tag{1}$$

where $R(D,Q)$ is the rate-distortion function of a memoryless source Q on $\mathcal{X}$ and $D(Q||P)$ is the relative entropy between Q and P. Thus for the special case $D = 0$, $R(D,Q)$ becomes the entropy $H(Q)$ and the maximization above gives ρ times Rényi's entropy [16] of order $1/(1+\rho)$ (see [2] for more detail). In view of this, $E(D,\rho)/\rho$, for $D > 0$, can be thought of as Rényi's analog to the rate-distortion function (see also [5]). We also demonstrate the existence of an asymptotically optimum guessing scheme that is universal both w.r.t. the underlying memoryless source P, and the moment order ρ. It is interesting to note that if $\rho = 1$, for example, then the guessing exponent $E(D,1)$ is in general larger than $R(D,P)$, in spite of the well known fact that a codebook whose size is exponentially $e^{NR(D,P)}$ is sufficient to keep the average distortion below D. In particular, $E(D,\rho)$ is in general positive at a certain range of distortion levels for which $R(D,P) = 0$. The roots of these phenomena lie in the tail behavior of the distribution of $G(\mathbf{X})$. We shall elaborate on this point later on.

In this context, we also study the closely related large deviations performance criterion, $\Pr\{G(\mathbf{X}) \geq e^{NR}\}$ for a given $R > R(D,P)$. Obviously, the exponential behavior of this probability is given by the source coding error exponent $F(R,D)$ [12], [4] for memoryless sources. It turns out, indeed, that there is an intimate relation between the guessing exponent considered here and the well-known source coding error exponent. In particular,

we show in Section 4 that for any fixed distortion level D, the ρth order guessing exponent $E(D,\rho)$ as a function of ρ is given by the one-sided Fenchel-Legendre transform (FLT) of the source coding error exponent $F(R,D)$ as a function of R. The inverse relation is that the FLT of $E(D,\rho)$ in ρ gives the lower convex hull of $F(R,D)$ in R. Moreover, since the above mentioned universal guessing scheme minimizes all moments of $G(\mathbf{X})$ simultaneously it also gives the best attainable large deviations performance, universally for every memoryless source P and every $R > R(D,P)$. We also establish relations to two other exponents in lossy source coding.

In Section 5, we study some basic properties of the function $E(D,\rho)$, such as monotonicity, convexity in both arguments, continuity, asymptotics, and others. Since no closed-form expression for $E(D,\rho)$ has been found in general, we also provide upper and lower bounds to $E(D,\rho)$, and a double maximum parametric representation, which might be suitable for iterative computation.

In Section 6, we provide several extensions and related results, including the memoryless Gaussian case, the case of a source with memory, and the case of incorporating side information.

Finally, in Section 7, we summarize our conclusions and share with the reader related open problems, some of which have resisted our best efforts so far.

2 Definitions and Notation Conventions

Consider an information source emitting symbols in an alphabet $\mathcal{X}$, and let $\hat{\mathcal{X}}$ denote a reproduction alphabet. When $\mathcal{X}$ is continuous, so will be $\hat{\mathcal{X}}$, and both will be assumed to be the entire real line. Let $d : \mathcal{X} \times \hat{\mathcal{X}} \to [0,\infty)$ denote a single-letter distortion measure. Let $\mathcal{X}^N$ and $\hat{\mathcal{X}}^N$ denote the Nth order Cartesian powers of $\mathcal{X}$ and $\hat{\mathcal{X}}$, respectively. The distortion between a source vector $\mathbf{x} = (x_1, ..., x_N) \in \mathcal{X}^N$ and a reproduction vector $\hat{\mathbf{x}} = (\hat{x}_1, ..., \hat{x}_N) \in \hat{\mathcal{X}}^N$ is defined as $d(\mathbf{x},\hat{\mathbf{x}}) = \sum_{i=1}^N d(x_i,\hat{x}_i)$.

Throughout the paper, scalar random variables will be denoted by capital letters while their sample values will be denoted by the respective lower case letters. A similar convention will apply to random N-dimensional vectors and their sample values, which will be denoted by boldface letters. Thus, for example, $\mathbf{X}$ will denote a random N-vector $(X_1, ..., X_N)$, and $\mathbf{x} = (x_1, ..., x_N)$ is a specific vector value in $\mathcal{X}^N$. Sources and channels will be denoted generically by capital letters, e.g., P, Q, and W. For memoryless sources and channels, the

respective lower case letters will denote the one-dimensional marginal probability density functions (PDF's) if the alphabet is continuous, or the one dimensional probability mass functions (PMF's) if it is discrete. Thus, a memoryless source P can be thought of as a vector (or a function) $\{p(x),\ x \in \mathcal{X}\}$. For N-vectors, the probability of the event $\mathbf{X} = \mathbf{x}$ will be denoted by $p^N(\mathbf{x})$, which in the memoryless case is given by $\prod_{i=1}^N p(x_i)$. Throughout this paper, P will denote the information source that generates the random variable X and the random vector $\mathbf{X}$ unless specified explicitly otherwise.

Integration w.r.t. a probability measure (e.g., $\int p(dx)f(x)$, $\int q^N(d\mathbf{x})f(\mathbf{x})$, etc.) will be interpreted as expectation w.r.t. this measure, which in the discrete case should be understood as an appropriate summation. Similar conventions will apply to conditional probability measures associated with channels. The probability of an event $A \subseteq \mathcal{X}^N$ will be denoted by $p^N\{A\}$, or by $\Pr\{A\}$ if there is no room for ambiguity regarding the underlying probability measure. The operator $\mathbf{E}\{\cdot\}$ will denote expectation w.r.t. the underlying source P unless otherwise specified.

For a memoryless source Q, let

$$H(Q) = -\int_{\mathcal{X}} q(dx) \ln q(x). \tag{2}$$

For two given memoryless sources P and Q on $\mathcal{X}$, let

$$D(Q||P) = \int_{\mathcal{X}} q(dx) \ln \frac{q(x)}{p(x)} \tag{3}$$

denote the relative entropy between Q and P. For a given memoryless source Q and a memoryless channel $W = \{w(\hat{x}|x),\ x \in \mathcal{X},\ \hat{x} \in \hat{\mathcal{X}}\}$, let $I(Q,W)$ denote the mutual information

$$I(Q,W) = \int_{\mathcal{X}} q(dx) \int_{\hat{\mathcal{X}}} w(d\hat{x}|x) \ln \frac{w(\hat{x}|x)}{\int_{\mathcal{X}} q(dx')w(\hat{x}|x')}. \tag{4}$$

The rate-distortion function $R(D,Q)$ for a memoryless source Q w.r.t. distortion measure d is defined as

$$R(D,Q) = \inf_W I(Q,W), \tag{5}$$

where the infimum is taken over all channels W such that

$$\Delta(Q,W) \stackrel{\triangle}{=} \int_{\mathcal{X}} q(dx) \int_{\hat{\mathcal{X}}} w(d\hat{x}|x) d(x,\hat{x}) \le D. \tag{6}$$

Comment: Throughout this paper we will assume that for every $x \in \mathcal{X}$, there exists $\hat{x} \in \hat{\mathcal{X}}$ with $d(x,\hat{x}) = 0$, that is, $d_{\min}(x) \stackrel{\triangle}{=} \min_{\hat{x} \in \hat{\mathcal{X}}} d(x,\hat{x}) = 0$ for all $x \in \mathcal{X}$. For distortion

measures that do not satisfy this condition, the parameter D should be henceforth thought of as the excess distortion beyond $d_{\min}(x)$.

Definition 1 *A D-admissible guessing strategy w.r.t. a source P is a (possibly infinite) ordered list $\mathcal{G}_N = \{\hat{\mathbf{x}}(1), \hat{\mathbf{x}}(2), ...\}$ of vectors in $\hat{\mathcal{X}}^N$, henceforth referred to as guessing code words, such that*

$$p^N\{d(\mathbf{X}, \hat{\mathbf{x}}(j)) \leq ND \text{ for some } j\} = 1. \tag{7}$$

Definition 2 *The guessing function $G_N(\cdot)$ induced by a D-admissible guessing strategy for N-vectors $\mathcal{G}_N$, is the function that maps each $\mathbf{x} \in \mathcal{X}^N$ into a positive integer, which is the index j of the first guessing code word $\hat{\mathbf{x}}(j) \in \mathcal{G}_N$ such that $d(\mathbf{x}, \hat{\mathbf{x}}(j)) \leq ND$. If no such guessing code word exists in $\mathcal{G}_N$ for a given $\mathbf{x}$, then $G_N(\mathbf{x}) \stackrel{\triangle}{=} \infty$.*

Thus, for a D-admissible guessing strategy, the induced guessing function takes on finite values with probability one.

Definition 3 *The optimum ρth order guessing exponent theoretically attainable at distortion level D is defined, whenever the limit exists, as*

$$\mathcal{E}_X(D, \rho) \stackrel{\triangle}{=} \lim_{N\to\infty} \frac{1}{N} \inf_{\mathcal{G}_N} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\}, \tag{8}$$

where the infimum is taken over all D-admissible guessing strategies.

The subscript X will be omitted whenever the source P, and hence also the random variable X associated with P, are clear from the context. Throughout the sequel, $o(N)$ will serve as a generic notation for a quantity that tends to zero as $N \to \infty$. For a finite set A, the cardinality will be denoted by $|A|$.

Another set of definitions and notation is associated with the method of types, which will be needed in some of the proofs for the finite alphabet case.

For a given source vector $\mathbf{x} \in \mathcal{X}^N$, the empirical probability mass function (EPMF) is the vector $Q_{\mathbf{x}} = \{q_{\mathbf{x}}(a), a \in \mathcal{X}\}$, where $q_{\mathbf{x}}(a) = N_{\mathbf{x}}(a)/N$, $N_{\mathbf{x}}(a)$ being the number of occurrences of the letter a in the vector $\mathbf{x}$. The set of all EPMF's of vectors in $\mathcal{X}^N$, that is, rational PMF's with denominator N, will be denoted by $\mathcal{Q}_N$. The type class $T_{\mathbf{x}}$ of a vector $\mathbf{x}$ is the set of all vectors $\mathbf{x}' \in \mathcal{X}^N$ such that $Q_{\mathbf{x}'} = Q_{\mathbf{x}}$. When we need to attribute a type class to a certain rational PMF $Q \in \mathcal{Q}^N$ rather than to a sequence in $\mathcal{X}^N$, we shall use the notation T_Q.

In the same manner, for sequence pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^N \times \mathcal{Y}^N$, the joint EPMF is the matrix $Q_{\mathbf{xy}} = \{q_{\mathbf{xy}}(a,b), a \in \mathcal{X}, b \in \mathcal{Y}\}$, where $q_{\mathbf{xy}}(a,b) = N_{\mathbf{xy}}(a,b)/N$, $N_{\mathbf{xy}}(a,b)$ being the number of joint occurrences of $x_i = a$ and $y_i = b$. The joint type class $T_{\mathbf{xy}}$ of $(\mathbf{x}, \mathbf{y})$ is the set of all pair sequences $(\mathbf{x}', \mathbf{y}') \in \mathcal{X}^N \times \mathcal{Y}^N$ for which $Q_{\mathbf{x}'\mathbf{y}'} = Q_{\mathbf{xy}}$.

Finally, a conditional type $T_{\mathbf{x}|\mathbf{y}}$ for a given $\mathbf{x}$ and $\mathbf{y}$, is the set of all sequences $\mathbf{x}'$ in $\mathcal{X}^N$ for which $(\mathbf{x}', \mathbf{y}) \in T_{\mathbf{xy}}$.

3 Guessing Exponents for Memoryless Sources

The main result in this section is a single-letter characterization of a lower bound to $\mathcal{E}(D, \rho)$ for memoryless sources, that is shown to be tight at least for the finite alphabet case. Specifically, for two given memoryless sources P and Q, and a given $\rho \geq 0$, let

$$E_X(D, \rho, Q) = \rho R(D, Q) - D(Q||P), \tag{9}$$

and let

$$E_X(D, \rho) = \sup_Q E_X(D, \rho, Q), \tag{10}$$

where the supremum is taken over all PDFs Q of memoryless sources for which $R(D, Q)$ and $D(Q||P)$ are well-defined and finite. Again, the subscript X of these two functions will be omitted whenever there is no room for ambiguity regarding the underlying source P that generates X.

We are now ready to state our main result in this section.

Theorem 1 *Let P be a memoryless source on $\mathcal{X}$.*

(a) (Converse part): Let $\{\mathcal{G}_N\}_{N \geq 1}$ be an arbitrary sequence of D-admissible guessing strategies, and let ρ be an arbitrary nonnegative real. Then,

$$\liminf_{N \to \infty} \frac{1}{N} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\} \geq E(D, \rho), \tag{11}$$

where G_N is the guessing function induced by $\mathcal{G}_N$.

(b) (Direct part): If $\mathcal{X}$ and $\hat{\mathcal{X}}$ are finite alphabets, then for any $D \geq 0$, there exists a sequence of D-admissible guessing strategies $\{\mathcal{G}_N^*\}_{N \geq 1}$ such that for every memoryless source P on $\mathcal{X}$ and every $\rho \geq 0$,

$$\limsup_{N \to \infty} \frac{1}{N} \ln \mathbf{E}\{G_N^*(\mathbf{X})^\rho\} \leq E(D, \rho), \tag{12}$$

where G_N^* is the guessing function induced by $\mathcal{G}_N^*$.

Corollary 1 *For a finite alphabet memoryless source, $\mathcal{E}(D,\rho)$ exists and is given by*

$$\mathcal{E}(D,\rho) = E(D,\rho). \tag{13}$$

Discussion: A few comments are in order in the context of this result.

First, observe that Theorem 1 is asymmetric in that part (a) is general while part (b) applies to the finite alphabet case only. This does not mean that part (b) is necessarily false when it comes to a general memoryless source. Nevertheless, so far we were unable to prove that it applies in general. The reason is primarily the fact that the method of types, which is used heavily in the proof below, does not lend itself easily to deal with the continuous case except for certain exponential families, like the Gaussian case, as will be discussed in Section 6.1.

Clearly, as one expects, in the finite alphabet lossless case ($D = 0$), the result of [2] is obtained as a special case since $\max_Q[\rho H(Q) - D(Q||P)]$ gives $\rho H_{1/(1+\rho)}(P)$, where $H_\theta(P)$ is Rényi's entropy [16] of order θ, defined as

$$H_\theta(P) = \frac{1}{1-\theta}\ln\sum_{x\in\mathcal{X}} p(x)^\theta. \tag{14}$$

As another point of view, Theorem 1 and its proof below remain valid if instead of the guessing problem, we consider the exponential behavior of $\mathbf{E}\{e^{\rho L(\mathbf{X})}\}$, that is, the characteristic function of the length $L(\mathbf{X})$ associated with variable length lossy coding subject to maximum distortion D. In this context, Theorem 1 serves as a tool to extend earlier results on the buffer overflow problem in lossless source coding (see, e.g., [10], [11], [15], [19]), where optimum performance is again characterized by Rényi's entropy.

It was mentioned briefly in the Introduction and should be emphasized again that $E(D,\rho)$ is in general larger than $\rho R(D,P)$. The latter is the exponential behavior that could have been expected at a first glance on the problem, because exponentially $e^{NR(D,P)}$ code words are known to suffice in order to keep the average distortion less than D. The intuition behind the larger exponential order that we obtain is that, while in the classical rate-distortion problem performance is judged on the basis of the *coding rate*, which is roughly speaking, equivalent to $\mathbf{E}\log G_N(\mathbf{X})$, here the criterion is $\mathbf{E}G_N(\mathbf{X})^\rho$ or equivalently, $\log\mathbf{E}G_N(\mathbf{X})^\rho$, which assigns much more weight to large values of the random variable $G_N(\mathbf{X})$. To put this even more in focus, observe that while in the ordinary source coding setting, the contribution of non-typical sequences can be ignored by using the asymptotic

equipartition property (AEP), here the major contribution is provided by *non-typical* sequences, in particular, sequences whose empirical PMF is close to Q^*, the maximizer of $E(D,\rho,Q)$, which in general may differ from P. Furthermore, while the above explanation is valid even in the lossless case $D=0$, the fact that we are dealing here with the lossy case $D>0$ gives another aspect to the difference between the classical source coding problem and the guessing problem: In source coding, essentially $e^{NR(D,P)}$ codewords suffice in order to guarantee *average* distortion within D, namely, if the rate is fixed, the distortion is a random variable whose expectation can be made arbitrarily close to D. This is achieved essentially by covering only the set of typical sequences by spheres of radius D. However, if we insist on *fixed* (or *maximum*) distortion less than D for *every* realization of the source, like in the guessing problem discussed here, then we must cover the *entire* space by a number of spheres that exponentially exceeds $e^{NR(D,P)}$ in general. (For example, when the source has unbounded support, it takes infinitely many spheres to cover the space.) Even then, if the rate-distortion codewords are encoded by a suitable variable-length code (entropy coding), then an average rate (approximately given by $N^{-1}\mathbf{E}\log G_N(\mathbf{X})$) that asymptotically attains the rate-distortion bound, can be achieved. In summary, the important point here is the following: While the source coding problem is 'insensitive' to whether we are dealing with fixed distortion or average distortion (because this difference can be traded for average rate as opposed to fixed rate), the guessing problem is sensitive to the difference between the two cases. This is because the performance criterion (moments of $G_N(\mathbf{X})$) is different than the one in source coding.

Note that part (b) of the Theorem actually states that there exists a *universal* guessing scheme, because it tells us that there exists a single scheme that is asymptotically optimum for every P and every ρ. Specifically, the proposed guessing scheme is composed from ordering codebooks that correspond to type classes Q in an increasing order of $R(D,Q)$ (see proof of part (b) below). This can be viewed as an extension of [18] from the lossless to the lossy case, as universal ordering of sequences in decreasing probabilities was carried out therein according to increasing empirical entropy $H(Q)$.

As an alternative proof to part (b), one can show the existence of an optimal *source-specific* guessing scheme using the classical random coding technique. Of course, once we have a universal scheme, there is no reason to bother about a source-specific scheme for the purpose of proving Theorem 1. The interesting point here, however, is that the

optimal random coding distribution for guessing is, in general, different than that of the ordinary rate-distortion coding problem. While in the latter, we use the output distribution corresponding to the test channel of $R(D,P)$, here it is best to use the one that corresponds to $R(D,Q^*)$, where Q^* maximizes $E(D,\rho,Q)$. Since optimum guessing codebooks have different statistics than optimum ordinary rate-distortion codebooks in general, it seems, at first glance, that guessing and source coding are conflicting goals. Nevertheless, it is possible to enjoy the benefits of both by interlacing the code words of a good rate-distortion code and a good guessing list. Since the index of each code word is at most doubled by this interlacing, it essentially neither affects the behavior of $\mathbf{E}\ln G_N(\mathbf{X})$, nor that of $\ln \mathbf{E}G_N(\mathbf{X})^\rho$. Thus the main message to be conveyed at this point is that if one wishes not only to attain the rate-distortion function, but also to minimize the expected number of candidate code words to be examined by the encoder, then good guessing code words must be included in the codebook in addition to the usual rate-distortion code words. In this context, it should be mentioned that the asymptotically optimum universal guessing scheme proposed in the proof of part (b) below attains also the rate-distortion function when used as a codebook followed by appropriate entropy coding.

The remaining part of this section is devoted to the proof of Theorem 1.

Proof of Theorem 1. We begin with part (a). Let $\mathcal{G}_N$ be an arbitrary D-admissible guessing strategy with guessing function G_N. Then, for any memoryless source Q,

$$\begin{aligned}
\mathbf{E}[G_N(\mathbf{X})^\rho] &= \int_{\mathcal{X}^N} p^N(d\mathbf{x}) G_N(\mathbf{x})^\rho \\
&= \int_{\mathcal{X}^N} q^N(d\mathbf{x}) \exp[-\ln \frac{q^N(\mathbf{x})}{p^N(\mathbf{x}) G_N(\mathbf{x})^\rho}] \\
&\geq \exp[-ND(Q||P) + \rho \int_{\mathcal{X}^N} q^N(d\mathbf{x}) \ln G_N(\mathbf{x})], \qquad (15)
\end{aligned}$$

where we have used Jensen's inequality in the last step.

The underlying idea behind the remaining part of the proof is that $\ln G_N(\mathbf{x})$ is essentially a length function associated with a certain entropy encoder that operates on the guessing list, and therefore the combination of the guessing list and the entropy coder can be thought of as a rate-distortion code. Thus, by the converse to the rate-distortion coding theorem, the expectation of $\ln G_N(\mathbf{X})$ w.r.t. a source Q essentially cannot be smaller than $NR(D,Q)$. Specifically, if we define

$$\alpha_i = \int_{\mathbf{x}:G_N(\mathbf{x})=i} q^N(d\mathbf{x}), \qquad (16)$$

then we have

$$\int_{\mathbf{x}} q^N(d\mathbf{x}) \ln G_N(\mathbf{x}) = \sum_i \alpha_i \ln i. \tag{17}$$

For a given $\delta > 0$, consider the following probability assignment on the positive integers:

$$\beta_i = \frac{C(\delta)}{i^{1+\delta}}, \qquad i = 1, 2, ..., \tag{18}$$

where $C(\delta)$ is a normalizing constant such that $\sum_i \beta_i = 1$. Consider a lossless code for the positive integers $\{i\}$ with length function $\lceil -\log_2 \beta_i \rceil$ bits, which when applied to the index $i = G_N(\mathbf{x})$ of the guessing code word for $\mathbf{x}$, gives a variable length rate-distortion code with maximum per-letter distortion D. Thus, by the converse to the rate-distortion coding theorem,

$$\begin{aligned} NR(D,Q) \log_2 e &\leq \sum_i \alpha_i \lceil -\log_2 \beta_i \rceil \\ &\leq 1 + (1+\delta) \sum_i \alpha_i \log_2 i - \log_2 C(\delta), \end{aligned} \tag{19}$$

which then gives

$$\sum_i \alpha_i \ln i \geq \frac{NR(D,Q) + \ln C(\delta) - \ln 2}{1+\delta}. \tag{20}$$

Combining this inequality with eqs. (15) and (17) yields

$$\ln \mathbf{E}[G_N(\mathbf{X})^\rho] \geq -ND(Q||P) + \frac{\rho[NR(D,Q) + \ln C(\delta) - \ln 2]}{1+\delta}. \tag{21}$$

Dividing by N and taking the limit infimum of both sides as $N \to \infty$, we get

$$\liminf_{N\to\infty} \frac{1}{N} \ln \mathbf{E}[G_N(\mathbf{X})^\rho] \geq \frac{\rho R(D,Q)}{1+\delta} - D(Q||P). \tag{22}$$

Since the left-hand side does not depend on δ, we may now take the limit of the right-hand side as $\delta \to 0$, and obtain

$$\liminf_{N\to\infty} \frac{1}{N} \ln \mathbf{E}[G_N(\mathbf{X})^\rho] \geq E(D,\rho,Q). \tag{23}$$

Finally, since the left-hand side does not depend on Q, we can take the supremum over all allowable PDF's Q, and thereby obtain $E(D,\rho)$ as a lower bound. This completes the proof of part (a).

To prove part (b), we shall invoke the *type covering lemma* due to Csiszár and Körner [6, p. 181] (see also [20] for a refined version), stating that every type class T_Q can be entirely covered by exponentially $e^{NR(D,Q)}$ spheres of radius ND in the sense of the distortion measure d. More precisely, the type covering lemma is the following.

Lemma 1 *([6], [20]): For any $Q \in \mathcal{Q}^N$ and distortion level $D \geq 0$, there exists a codebook $\mathcal{C}_Q \subset \hat{\mathcal{X}}^N$ such that for every $\mathbf{x} \in T_Q$,*

$$\min_{\hat{\mathbf{x}} \in \mathcal{C}_Q} d(\mathbf{x}, \hat{\mathbf{x}}) \leq ND, \tag{24}$$

and at the same time,

$$\frac{1}{N} \ln |\mathcal{C}_Q| \leq R(D, Q) + o(N). \tag{25}$$

For every $Q \in \mathcal{Q}^N$, let $\mathcal{C}_Q$ denote a certain codebook in $\hat{\mathcal{X}}^N$ that satisfies the type covering lemma. Let us now order the rational PMF's in $\mathcal{Q}^N$ as $\{Q_1, Q_2, ...\}$ according to increasing value of $R(D, Q)$, that is, $R(D, Q_i) \leq R(D, Q_{i+1})$ for all $i < |\mathcal{Q}^N|$. Our guessing list $\mathcal{G}_N^*$ is composed of the ordered concatenation of the corresponding codebooks $\mathcal{C}_{Q_1}, \mathcal{C}_{Q_2}, ...$, where the order of guessing code words within each $\mathcal{C}_{Q_i}$ is immaterial. We now have

$$\begin{aligned}
\mathbf{E}[G_N^*(\mathbf{X})^\rho] &= \sum_{\mathbf{x} \in \mathcal{X}^N} p^N(\mathbf{x}) G_N^*(\mathbf{x})^\rho \\
&= \sum_i \sum_{\mathbf{x} \in T_{Q_i}} p^N(\mathbf{x}) G_N^*(\mathbf{x})^\rho \\
&\leq \sum_i \sum_{\mathbf{x} \in T_{Q_i}} p^N(\mathbf{x}) \left(\sum_{j \leq i} |\mathcal{C}_{Q_j}| \right)^\rho \\
&\leq \sum_i \exp[-ND(Q_i||P)] \left(\sum_{j \leq i} |\mathcal{C}_{Q_j}| \right)^\rho \\
&\leq \sum_{Q \in \mathcal{Q}^N} \exp\{-ND(Q||P) + \rho N[R(D, Q) + o(N)]\} \\
&\leq \exp\{N[E(D, \rho) + o(N)]\},
\end{aligned} \tag{26}$$

where we have used the facts [6] that $p^N(T_Q) \leq \exp[-ND(Q||P)]$ and that $|\mathcal{Q}^N|$ grows polynomially in N. Taking the logarithms of both sides, dividing by N, and passing to the limit as $N \to \infty$, give the assertion of part (b), and thus completes the proof of Theorem 1. □

4 Relations to Other Exponents in Lossy Source Coding

In this section, we demonstrate that the guessing exponent function $E(D, \rho)$ is intimately related to optimum exponents associated with certain other problems in lossy source coding. These relations will help us to investigate the properties of $E(D, \rho)$ in Section 5. Here and

throughout the sequel, we confine our attention to finite alphabet memoryless sources unless specified otherwise.

Intuitively, the moments of $G_N(\mathbf{X})$ are closely related to the cumulative distribution function of this random variable, and hence to the tail behavior, or equivalently, the large deviations performance $\Pr\{G_N(\mathbf{X}) \geq e^{NR}\}$, for $R > R(D,P)$. Obviously, the best attainable exponential rate of this probability is given by the *source coding error exponent* [12], [4, Theorem 6.6.4], which is the best attainable exponential rate of the probability that a codebook of size e^{NR} would fail to encode a randomly drawn source vector with distortion less than or equal to ND. The source coding error exponent at rate R and distortion level D, $F(R,D)$ is given by

$$F(R,D) = \min_{Q:R(D,Q)\geq R} D(Q||P). \tag{27}$$

Using the same technique as in the proof of Theorem 1(b), it is easy to see that the universal guessing scheme proposed therein $\mathcal{G}_N^*$ attains the best attainable large deviations performance in Marton's sense [12], that is,

$$\begin{aligned} F(R-0,D) &\leq \liminf_{N\to\infty}\left[-\frac{1}{N}\ln\Pr\{G_N^*(\mathbf{X}) \geq e^{NR}\}\right] \\ &\leq \limsup_{N\to\infty}\left[-\frac{1}{N}\ln\Pr\{G_N^*(\mathbf{X}) \geq e^{NR}\}\right] \\ &\leq F(R+0,D), \end{aligned} \tag{28}$$

where $F(R+0,D)$ and $F(R-0,D)$ are limits of $F(R+\epsilon,D)$ as $\epsilon \to 0$, along positive values of ϵ and negative values of ϵ, respectively.[1] This follows from the simple fact that by construction of $\mathcal{G}_N^*$, the event $\{\mathbf{x}:\ G_N^*(\mathbf{x}) \geq e^{NR}\}$ is essentially equivalent to the event $\{\mathbf{x}:\ R(D,Q_{\mathbf{x}}) \geq R\}$, where $Q_{\mathbf{x}}$ is the empirical PMF associated with $\mathbf{x}$. This result is not very surprising if we recall that $\mathcal{G}_N^*$ asymptotically minimizes all nonnegative moments of $G_N(\mathbf{X})$ simultaneously. The natural question that arises at this point is: what is the relation between the guessing exponent $E(D,\rho)$ and the source coding error exponent $F(R,D)$?

The following theorem tells that for a fixed distortion level D, the guessing exponent $E(D,\rho)$, as a function of ρ, is the one-sided Fenchel-Legendre transform (FLT) of $F(R,D)$ as a function of R. (See also [5, Theorem 1] for the lossless case). As for the inverse relation, the FLT of $E(D,\rho)$ as a function of ρ is the lower convex hull of $F(R,D)$ as a function of

[1]The function $F(R,D)$ may not be continuous in general (see Ahlswede [1]). However, monotonicity guarantees continuity everywhere except for countably many points. Sufficient conditions for everywhere continuity are discussed in [1] and [12].

D. Thus, if $F(R,D)$ is itself convex in R, the inverse FLT relation holds as well. It is easy to show that $F(R,D)$ is convex in R whenever $R(D,Q)$ meets the Shannon lower bound for every Q (e.g., binary source and Hamming distortion measure). This follows from the fact that $F(R,0)$ is always convex, and that in this case, $F(R,D) = F(R+\phi(D),0)$ for some function ϕ.

Theorem 2 *For a given finite alphabet memoryless source P and distortion level D,*

$$E(D,\rho) = \sup_{R\geq 0}[\rho R - F(R,D)], \quad \textit{for all } \rho \geq 0, \tag{29}$$

and

$$\tilde{F}(R,D) = \sup_{\rho\geq 0}[\rho R - E(D,\rho)], \quad \textit{for all } R \geq 0, \tag{30}$$

where $\tilde{F}(R,D)$ is the lower convex hull of $F(R,D)$ in R.

Proof. Eq. (29) is obtained as follows.

$$\begin{aligned}
\sup_{R\geq 0}[\rho R - F(R,D)] &= \sup_{R\geq 0} \max_{Q:R(D,Q)\geq R}[\rho R - D(Q||P)] \\
&= \max_{Q} \max_{0\leq R\leq R(D,Q)}[\rho R - D(Q||P)] \\
&= \max_{Q}[\rho R(D,Q) - D(Q||P)] \\
&= E(D,\rho).
\end{aligned} \tag{31}$$

Eq. (30) is a version of the duality lemma of the FLT [7, p. 135, Theorem 4.5.10], [17, p. 104, Theorem 12.2 and the preceding discussion]. Although the duality lemma therein refers to the *two-sided* FLT (i.e., with suprema taken over the entire real line) as opposed to the one-sided FLT considered here, eq. (30) can be obtained as a special case since F is monotone in R. Nevertheless, for the sake of convenience and completeness, we prove in the Appendix the following duality lemma specifically for the one-sided FLT.

Lemma 2 *Let $f(x)$ be an arbitrary nondecreasing function defined for $x \geq 0$, and let*

$$f^*(y) = \sup_{x\geq 0}[xy - f(x)] \tag{32}$$

be the one-sided FLT of f. Let f^{**} be the one-sided FLT of f^*, i.e., $f^{**}(x) = \sup_{y\geq 0}[xy - f^*(y)]$. Then, f^{**} equals the lower convex-hull of f.

This competes the proof of Theorem 2. □

Another related problem in lossy source coding is the following: For a given N-vector $\mathbf{x}$ and a codebook $\mathcal{C}_N$ of e^{NR} code words in $\hat{\mathcal{X}}^N$, let $d(\mathbf{x}, \mathcal{C}_N)$ denote the minimum of $d(\mathbf{x}, \hat{\mathbf{x}})$, over $\hat{\mathbf{x}} \in \mathcal{C}_N$. Suppose we would like to characterize the smallest attainable asymptotic exponential rate of the characteristic function of $d(\mathbf{X}, \mathcal{C}_N)$, i.e.,

$$\mathcal{J}(R, s) = \lim_{N\to\infty} \frac{1}{N} \min_{\mathcal{C}_N} \ln \mathbf{E}\{e^{s\, d(\mathbf{X}, \mathcal{C}_N)}\}, \qquad s > 0, \tag{33}$$

provided that the limit exists. By using the same techniques as above, it is easy to show that for memoryless sources with finite $\mathcal{X}$ and $\hat{\mathcal{X}}$, $\mathcal{J}(R, s)$ exists and is given by

$$\mathcal{J}(R, s) = J(R, s) = \max_Q [sD(R, Q) - D(Q||P)], \tag{34}$$

where Q is again a memoryless source on $\mathcal{X}$, and $D(R, Q)$ is its distortion-rate function. Thus, this problem can be thought of as being dual to the guessing problem in the sense that $J(R, s)$ has the same form as $E(D, \rho)$ except that the rate-distortion function is replaced by the distortion-rate function. Moreover, while $E(D, \rho)$ and $F(R, D)$ are a one-sided FLT pair provided that F is convex, it is easy to see that $J(R, s)$ and $F(R, D)$ are also a one-sided FLT pair under a similar condition on $F(R, D)$ as a function of D. Thus, in this case, $J(R, s)$ and $E(D, \rho)$ can be thought of as a two-dimensional FLT pair.

Finally, to complete the picture, let us consider now another related problem which corresponds to minimizing a linear combination of the rate and the distortion. Let $\mathcal{C}_N$ denote a code book as before, and for a given source vector $\mathbf{x}$, let $V(\mathbf{x}, \mathcal{C}_N) = \min_{\hat{\mathbf{x}}\in\mathcal{C}_N} [\rho L(\hat{\mathbf{x}}) + sd(\mathbf{x}, \hat{\mathbf{x}})]$, where $L(\hat{\mathbf{x}})$ is the coding length after entropy coding, and ρ and s are given nonnegative reals. It can be easily shown by using the same techniques that the best attainable exponential behavior of $\mathbf{E}\{e^{V(\mathbf{X}, \mathcal{C}_N)}\}$ among all codebooks $\mathcal{C}_N$, is given by

$$K(s, \rho) = \max_Q \min_W [\rho I(Q, W) + s\Delta(Q, W) - D(Q||P)]. \tag{35}$$

Now, $E(D, \rho)$ is given in terms of $K(s, \rho)$ as follows.

$$\begin{aligned}
E(D, \rho) &= \max_Q [\rho R(D, Q) - D(Q||P)] \\
&= \max_Q \min_{\{W:\Delta(Q,W)\le D\}} [\rho I(Q, W) - D(Q||P)] \\
&= \max_Q \min_W \begin{cases} \rho I(Q, W) - D(Q||P) & \text{if } \Delta(Q, W) \le D \\ \infty & \text{elsewhere} \end{cases} \\
&= \max_Q \min_W \sup_{s\ge 0} [\rho I(Q, W) - D(Q||P) + s(\Delta(Q, W) - D)]
\end{aligned}$$

$$= \sup_{s\geq 0} \max_{Q} \min_{W} [\rho I(Q,W) - D(Q||P) + s(\Delta(Q,W) - D)]$$
$$= \sup_{s\geq 0} [K(s,\rho) - sD], \tag{36}$$

which means that $E(D,\rho)$ can be thought of as the vertical axis intercept of the supporting line of slope D to the curve $K(s,\rho)$ vs. s for fixed ρ. The significance and the implications of this representation of $E(D,\rho)$ will be further discussed in the next section. Also in this context, an important property of $K(s,\rho)$ is that it is monotonically increasing and concave in each argument, as will be restated and proved in the next section. Similarly as in the proof of eq. (30) in Theorem 2, monotonicity and concavity of $K(s,\rho)$ in s for fixed ρ leads to the inverse relation

$$K(s,\rho) = \inf_{D\geq 0} [E(D,\rho) + sD], \tag{37}$$

which means that $K(s,\rho)$ can be also interpreted as the vertical axis intercept of the supporting line of slope $-s$ to the curve $E(D,\rho)$ vs. D for fixed ρ. Similar relations hold between $K(s,\rho)$ and $J(R,s)$ for fixed s, by replacing s and D with ρ and R, respectively. All the relations among the four bivariate functions $E(D,\rho)$, $F(R,D)$, $J(R,s)$, and $K(s,\rho)$ are summarized in Fig. 1. Again, it should be kept in mind that the transform relations in the directions from $E(D,\rho)$ to $F(R,D)$ and from $J(R,s)$ to $F(R,D)$ hold subject to convexity conditions.

5 Properties of the Guessing Exponent Function

In this section, we study some more basic properties of the guessing exponent function $E(D,\rho)$ for finite alphabet memoryless sources and finite reproduction alphabets. We begin by listing a few simple facts about $E(D,\rho)$, some of which follow directly from known properties of the rate-distortion function.

Proposition 1 *The guessing exponent $E(D,\rho)$ has the following properties:*

(a) $E(D,\rho)$ is nonnegative; $E(0,\rho) = \rho H_{1/(1+\rho)}(P)$; $E(D,0) = 0$; the smallest distortion level $D_0(\rho)$ beyond which $E(D,\rho) = 0$ is given by

$$D_0(\rho) = \sup\{D:\ a(D) < \rho\}, \tag{38}$$

where $a(D) \stackrel{\triangle}{=} \inf_{R\geq 0} F(R,D)/R$.

(b) $E(D,\rho)$ is a strictly decreasing, convex function of D in $[0, D_0(\rho))$, for any fixed $\rho > 0$.

(c) For fixed D, $E(D,\rho)$ is a strictly increasing, convex function of ρ in the range of ρ where $E(D,\rho) > 0$.

(d) $E(D,\rho)$ is continuous in $D \geq 0$ and in $\rho \geq 0$.

(e) $E(D,\rho) \geq \rho R(D,P)$; $\lim_{\rho\to 0} E(D,\rho)/\rho = R(D,P)$.

(f) $E(D,\rho) \leq \rho R_{\max}(D)$, where $R_{\max}(D) = \max_Q R(D,Q)$; $\lim_{\rho\to\infty} E(D,\rho)/\rho = R_{\max}(D)$.

The proof appears in the Appendix.

We are not aware of the existence of a closed-form expression for $E(D,\rho)$ in general. Parts (e) and (f) of Proposition 1 suggest a lower and an upper bound, respectively. Another simple and useful lower bound, which is sometimes tight and then gives a closed-form expression to $E(D,\rho)$, is induced from the Shannon lower bound to $R(D,Q)$ [3, Sect. 4.3.1]. The Shannon lower bound applies to difference distortion measures, i.e., distortion measures $d(x,\hat{x})$ that depend only on the difference $x - \hat{x}$ (for a suitable definition of subtraction of elements in $\hat{\mathcal{X}}$ from elements in $\mathcal{X}$).

Theorem 3 *For a difference distortion measure,*

$$E(D,\rho) \geq \max\{0, \rho H_{1/(1+\rho)}(P) - \rho\phi(D)\}, \tag{39}$$

where $\phi(D)$ is the maximum entropy of the random variable $(X-\hat{X})$ subject to the constraint $Ed(X-\hat{X}) \leq D$. Equality is attained if the distortion measure is such that the Shannon lower bound $R(D,Q) \geq \max\{0, H(Q) - \phi(D)\}$ is met with equality for every Q.

Proof.

$$\begin{aligned}
E(D,\rho) &= \max_Q [\rho R(D,Q) - D(Q||P)] \\
&\geq \max_Q [\rho \max\{0, H(Q) - \phi(D)\} - D(Q||P)] \\
&= \max_Q \max\{-D(Q||P), \rho[H(Q) - \phi(D)] - D(Q||P)\} \\
&= \max\{\max_Q [-D(Q||P)], \max_Q [\rho[H(Q) - \phi(D)] - D(Q||P)]\} \\
&= \max\{0, \rho H_{1/(1+\rho)}(P) - \rho\phi(D)\}.
\end{aligned} \tag{40}$$

□

Note, that if the distortion measure d is such that the Shannon lower bound is tight for all Q, e.g., binary sources and the Hamming distortion measure (see also the Gaussian case, Sect. 6.1), we have a closed-form expression for $E(D,\rho)$, and hence also for $D_0(\rho)$ as

$$D_0(\rho) = \phi^{-1}(H_{1/(1+\rho)}(P)). \tag{41}$$

Moreover, the PMF Q^* that attains $E(D,\rho)$ does not depend on D. Fig. 2 illustrates curves of $E(D,\rho)$ vs. D for a binary source with letter probabilities 0.4 and 0.6 and the Hamming distortion measure. As can be seen, $E(D,\rho)$ becomes zero at different distortion levels $D_0(\rho)$ depending on ρ. Since $E(D,\rho) \geq \rho R(D,P)$, then $D_0(\rho)$ is never smaller than $D_{\max}$, the smallest distortion at which $R(D,P) = 0$.

As mentioned earlier, $E(D,\rho)$ does not always have a known closed-form expression. To obtain an alternative characterization of $E(D,\rho)$, which may be more suitable than the saddle-point form (35) for determining $E(D,\rho)$, we cite without proof the following result from Gallager [9, Theorem 9.4.1, p. 459].

Lemma 3 *For any Q and $r \geq 0$,*

$$\min_W [I(Q,W) + r\Delta(Q,W)] = \max_{f \in \mathcal{F}_r} \left[H(Q) + \sum_x q(x) \ln f(x) \right] \tag{42}$$

where $\mathcal{F}_r$ is the set of all vectors $f = \{f(x), x \in \mathcal{X}\}$ with nonnegative components such that $\sum_{x \in \mathcal{X}} f(x) e^{-r d(x,\hat{x})} \leq 1$ for all $\hat{x} \in \hat{\mathcal{X}}$. Any feasible W and f achieve, respectively, the minimum and the maximum in (42) iff they satisfy for all $x, \hat{x}$,

$$w(\hat{x}|x) q(x) = m(\hat{x}) f(x) e^{-r d(x,\hat{x})}, \tag{43}$$

where $m(\hat{x}) \stackrel{\triangle}{=} \sum_x q(x) w(\hat{x}|x)$.

Substituting (42) in (35) with $r = s/\rho$, we obtain a characterization of $K(S,\rho)$ as a double-maximum,

$$K(s,\rho) = \max_Q \max_{f \in \mathcal{F}_{s/\rho}} \left[\rho H(Q) + \rho \sum_x q(x) \ln f(x) - D(Q||P) \right], \text{æ} \tag{44}$$

which appears amenable to iterative numerical computation. (It is noteworthy for computational purposes that the maximum here is achieved by a unique pair (Q,f), as will be discussed later in this section.) Once $K(s,\rho)$ is determined, $E(D,\rho)$ can be found by line search over $s \geq 0$ using the right-most side of eq. (36).

A straightforward calculation shows that, for fixed f, the maximum over Q in (44) is achieved by

$$q(x) = cp(x)^{1/(1+\rho)} f(x)^{\rho/(1+\rho)}, \tag{45}$$

where c is a normalizing constant so that $\sum_{x\in\mathcal{X}} q(x) = 1$. Substituting this into (44) and using (36), we obtain the following expression for $E(D,\rho)$.

Theorem 4 *For all $D \geq 0$ and $\rho > 0$, the guessing exponent is given by*

$$E(D,\rho) = \sup_{s\geq 0} \max_{f\in\mathcal{F}_{s/\rho}} \left[(1+\rho) \ln \sum_{x\in\mathcal{X}} p(x)^{1/(1+\rho)} f(x)^{\rho/(1+\rho)} - sD \right]. \tag{46}$$

Necessary and sufficient conditions for $f \in \mathcal{F}_{s/\rho}$ to achieve the maximum are that there exist a W satisfying the condition (43) with $r = s/\rho$ and Q given by (45).

Theorem 4 can be used also to obtain lower bounds to $E(D,\rho)$ by selecting an arbitrary feasible f. In certain simple cases, as explored in the following examples, the optimal f can be guessed.

Example 1: The lossless case. Let $\mathcal{X} = \hat{\mathcal{X}}$, $d(x,\hat{x}) = 0$ for $x = \hat{x}$, and $d(x,\hat{x}) = \infty$ for $x \neq \hat{x}$. Here, the only interesting distortion level for guessing is $D = 0$. It is easy to verify that $K(s,\rho)$ is achieved by $f(x) = 1$ for all $s \geq 0$. For $D = 0$, we obtain from (46) that

$$E(0,\rho) = (1+\rho) \ln \left[\sum_x P(x)^{1/(1+\rho)} \right], \tag{47}$$

which agrees with the result in [2].

Comment: In the above example, if the distortion measure is modified so that it is finite but non-trivial in the sense that $0 < d(x,\hat{x}) < \infty$ for $x \neq \hat{x}$, then $E(0,\rho)$ is still given by the above form.

Example 2: The Hamming distortion measure. Let $\mathcal{X} = \hat{\mathcal{X}}$ be finite alphabets with size $K \geq 2$, $d(x,\hat{x}) = 0$ if $x = \hat{x}$, and $d(x,\hat{x}) = 1$ if $x \neq \hat{x}$. For $\rho > 0$ fixed and $s \geq 0$ arbitrary, the f with uniform components given by

$$f(x) = f_s \stackrel{\triangle}{=} \frac{1}{1 + (K-1)e^{-s/\rho}}, \quad \text{all } x \in \mathcal{X} \tag{48}$$

is feasible, and for this choice eq. (46) is maximized over $s \geq 0$ by

$$s^* = \rho \ln \frac{(K-1)(1-D)}{D}, \tag{49}$$

for D in the range $0 \leq D \leq (K-1)/K$. (At $D = 0$, we interpret s^* to be ∞.) Using s^* and $f(x) = f_{s^*}$ in (46), we have for any $\rho \geq 0$, and $0 \leq D \leq (K-1)/K$,

$$E(D, \rho) \geq \rho[H_{1/(1+\rho)}(P) - h(D) - D\ln(K-1)], \tag{50}$$

where $h(D) = -D\ln(D) - (1-D)\ln(1-D)$. It is easy to see that the condition for equality in (50) will be satisfied if and only if

$$q^*(x) \geq \frac{1}{e^{s^*/\rho} + (K-1)} = \frac{D}{K-1}, \quad \text{all } x \in \mathcal{X}, \tag{51}$$

where $q^*(x)$ is as defined in (45). Thus, equality holds in (50) for all $D \geq 0$ sufficiently small. In particular, for P the uniform distribution, equality holds for all $0 \leq D \leq (K-1)/K$. Note also that eq. (50) coincides with the Shannon lower bound, as for the Hamming distortion measure, $\phi(D) = h(D) + D\ln(K-1)$. $\square$

As already pointed out in the previous section, $K(s, \rho)$ can be given a geometric interpretation, in view of (37), as the vertical axis intercept of supporting line of slope $-s$ to the curve $E(D, \rho)$ vs. D for fixed $\rho \geq 0$. The proof of the inverse relation (37) as well as the one between $J(R, s)$ and $K(s, \rho)$ rely on the following properties of $K(s, \rho)$.

Lemma 4 *The function $K(s, \rho)$ is monotonically increasing and concave in each argument.*

The proof appears in the Appendix.

The next result establishes the uniqueness of the PMF Q that achieves $E(D, \rho)$ in its various possible representations. This signifies, e.g., that the maximum in $\max_Q E(D, \rho, Q)$ is achieved by a unique type class, with clear coding implications.

Proposition 2 *For any fixed distortion level in the range $0 \leq D < D_0(\rho)$, there exists a unique Q^* that achieves the maximum in $E(D, \rho) = \max_Q E(D, \rho, Q)$. The PMF Q^* also achieves uniquely the maximum in (35) and in (44) for each $s \in \mathcal{S}(D) \stackrel{\triangle}{=} \{s \geq 0 : \ E(D, \rho) = K(s, \rho) - sD\}$. Furthermore, the maximum in (44) is achieved by a unique pair (Q^*, f_s) for each $s \in \mathcal{S}(D)$.*

The proof is given in the Appendix.

By using the uniqueness of Q^*, it can be shown also that for bounded distortion measures, $E(D, \rho)$ is differentiable w.r.t. both arguments. The derivative w.r.t. D is given by $\rho\partial R(D, Q^*)/\partial D$, and the derivative w.r.t. ρ is given by $R(D, Q^*)$. In view of parts (c), (e),

and (f) of Proposition 1, this means that the slope of the curve $E(D,\rho)$ vs. ρ for fixed D grows monotonically and continuously from $R(D,P)$ to $R_{\max}(D)$ as ρ grows from zero to infinity.

The following example shows that, similarly as the rate-distortion function, $E(D,\rho)$ may not be differentiable w.r.t. D if the distortion measure is unbounded. Strictly speaking, in Example 1 above the distortion measure is unbounded as well. The difference, however, is that in Example 1 we have examined only the point $D=0$ as there was no other point of finite distortion level.

Example 3: Unbounded distortion measure. (cf. [9, Prob. 9.4, p. 567]). Let $\mathcal{X}=\{1,2,3,4\}$, $\hat{\mathcal{X}}=\{1,2,3,4,5,6,7\}$, and let the distortion matrix $\{d(x,\hat{x})\}$ be given by

$$\begin{bmatrix} 0 & \infty & \infty & \infty & 1 & \infty & 3 \\ \infty & 0 & \infty & \infty & 1 & \infty & 3 \\ \infty & \infty & 0 & \infty & \infty & 1 & 3 \\ \infty & \infty & \infty & 0 & \infty & 1 & 3 \end{bmatrix}. \tag{52}$$

It is easy to verify that $K(s,\rho)$ is achieved by an f with equal components, $f(x)=f_s$, where

$$f_s = \begin{cases} 0.25e^{(3s/\rho)} & \text{if } 0 \le s \le \frac{\rho}{2}\ln(2), \\ 0.5e^{(s/\rho)} & \text{if } \frac{\rho}{2}\ln(2) < s \le \rho\ln(2), \\ 1 & \text{if } s > \rho\ln(2). \end{cases} \tag{53}$$

Substituting the resulting $K(s,\rho)$ in (36), we obtain

$$E(D,\rho) = \begin{cases} \rho(2-D)\ln(2) & \text{if } 0 \le D \le 1, \\ \frac{1}{2}\rho(3-D)\ln(2) & \text{if } 1 < D \le 3, \\ 0 & \text{if } D > 3. \end{cases} \tag{54}$$

6 Related Results and Extensions

In this section we provide several extensions and variations on our previous results for other situations of theoretical and practical interest.

6.1 Memoryless Gaussian Sources

We mentioned in the Discussion after Theorem 1 that we do not have an extension of the direct part to general continuous alphabet memoryless sources. However, for the special case of a Gaussian memoryless source and the mean squared error distortion measure, this can still be done relatively easily by applying a continuous alphabet analog to the method of types.

Theorem 5 *If $\mathcal{X} = \hat{\mathcal{X}} = \mathbb{R}$, P is a memoryless, zero-mean Gaussian source, and $d(x, \hat{x}) = (x - \hat{x})^2$, then $\mathcal{E}(D, \rho)$ exists and is given by*

$$\mathcal{E}(D, \rho) = E(D, \rho), \tag{55}$$

where the supremum in the definition of $E(D, \rho)$ is now taken over all memoryless, zero-mean Gaussian sources Q.

Comment: For two zero-mean, Gaussian memoryless sources P and Q with variances σ_p^2 and σ_q^2, respectively, $D(Q||P)$ is given by

$$D(Q||P) = \frac{1}{2}\left(\frac{\sigma_q^2}{\sigma_p^2} - \ln\frac{\sigma_q^2}{\sigma_p^2} - 1\right). \tag{56}$$

Since

$$R(D, Q) = \max\left\{0, \frac{1}{2}\ln\frac{\sigma_q^2}{D}\right\} \tag{57}$$

agrees with the Shannon lower bound, then by Theorem 3, we obtain the closed-form expression

$$E(D, \rho) = \max\left\{0, \frac{1}{2}\left[\rho\ln\frac{\sigma_p^2}{D} + (1+\rho)\ln(1+\rho) - \rho\right]\right\}. \tag{58}$$

Note that the slope of $E(D, \rho)$ as a function of ρ for fixed D, grows without bound as $\rho \to \infty$. This happens because $R_{\max}(D) = \infty$ in this case (see Proposition 1(f)).

The remaining part of this subsection is devoted to the proof of Theorem 5.

Proof of Theorem 5. Since the converse part of Theorem 1 applies to memoryless sources in general, it suffices to prove the direct part. This in turn will be obtained as a simple extension of the proof of Theorem 1(b), provided that we have a suitable version of the type covering lemma for Gaussian sources. Another slight complication is that, unlike in the finite alphabet case, here we have infinitely many (rather than polynomially many) such type classes to take into account.

Let us first define the notion of a Gaussian type class. For a given value of $\sigma^2 > 0$ and $0 < \epsilon < 1$, a *Gaussian type class* $T^\epsilon(\sigma^2)$ is defined as the set of all N-vectors $\mathbf{x}$ with the property $|\mathbf{x}^t\mathbf{x} - N\sigma^2| \le N\epsilon\sigma^2$, where $\mathbf{x}$ is understood as a column vector and the superscript t denotes vector transposition. It is easy to show (see Appendix) that the volume of $T^\epsilon(\sigma^2)$ is upper bounded by

$$\mathrm{Vol}\{T^\epsilon(\sigma^2)\} \le [2\pi e\sigma^2(1+\epsilon)]^{N/2}. \tag{59}$$

Consider next, the forward test channel W of $R(D,Q)$, defined by

$$\hat{X} = \begin{cases} (1-\frac{D}{\sigma_q^2})X + V & \text{if } D < \sigma_q^2 \\ 0 & \text{if } D \geq \sigma_q^2 \end{cases}, \tag{60}$$

where $X \sim \mathcal{N}(0,\sigma_q^2)$, $V \sim \mathcal{N}(0, D - D^2/\sigma_q^2)$ and $V - X$. For $\sigma_q^2 > D$ and $0 < \epsilon < 1$, we next define the conditional type of an N-vector $\hat{\mathbf{x}}$ given an N-vector $\mathbf{x}$ w.r.t. W as

$$\begin{aligned} T_{\mathbf{x}}^{\epsilon}(W) \quad = \quad & \left\{ \hat{\mathbf{x}} : \ \hat{\mathbf{x}} = (1-\frac{D}{\sigma_q^2})\mathbf{x} + \mathbf{v}; \ |\mathbf{v}^t\mathbf{v} - N(D - \frac{D^2}{\sigma_q^2})| \leq N\epsilon(D - \frac{D^2}{\sigma_q^2}), \right. \\ & \left. |\mathbf{v}^t\mathbf{x}| \leq \epsilon\sqrt{N(D-\frac{D^2}{\sigma_q^2})\mathbf{x}^t\mathbf{x}} \right\}. \end{aligned} \tag{61}$$

It is shown in the Appendix that

$$\text{Vol}\{T_{\mathbf{x}}^{\epsilon}(W)\} \geq \left(1 - \frac{3}{N\epsilon^2}\right)\left[2\pi e^{1-\epsilon}\left(D - \frac{D^2}{\sigma_q^2}\right)\right]^{N/2}. \tag{62}$$

We now want to prove that $T^{\epsilon}(\sigma_q^2)$ can be covered by exponentially $\exp\{NR(D,Q)\}$ code vectors $\{\hat{\mathbf{x}}(i)\}$ within Euclidean distance essentially as small as $\sqrt{ND}$. For $\sigma_q^2 \leq D$, this is trivial as the vector $\hat{\mathbf{x}} = 0$ represents any $\mathbf{x} \in T^{\epsilon}(\sigma_q^2)$ within distortion $D + \epsilon$. Assume next, that $\sigma_q^2 > D$ and let $0 < \epsilon < 1$. Let us construct a grid S of all vectors in the Euclidean space $\mathcal{R}^N$ whose components are integer multiples of 2δ for some small $0 < \delta << \sqrt{D}$. Consider the N-dimenional cubes of size δ, centered at the grid points. For a given code $\mathcal{C} = \{\hat{\mathbf{x}}(1), \ldots, \hat{\mathbf{x}}(M)\}$, let $U(D)$ denote the subset of cubes in $T^{\epsilon}(\sigma_q^2)$ for which the cube center $\mathbf{x}_0$ satisfies $||\mathbf{x}_0 - \hat{\mathbf{x}}(i)||^2 > N(D+\mu)$ for all $i = 1, ..., M$, where μ is a small positive real which will be specified later. This means that $U(D)$ is the set of cubes in $T^{\epsilon}(\sigma_q^2)$ whose centers are not covered by $\mathcal{C}$ within distortion $D + \mu$.

Consider the following random coding argument. Let $\hat{\mathbf{X}}(1), \ldots, \hat{\mathbf{X}}(M)$ denote i.i.d. vectors drawn uniformly in $T^{\xi}(\sigma_q^2 - D)$, where $\xi = \epsilon(1 + 4\sqrt{D/(\sigma_q^2 - D)})$. If we show that $\mathbf{E}|U(D)| < 1$, then there must exist a code for which $U(D)$ is empty, which means that all cube centers are covered within distortion $D + \mu$, and therefore, by the triangle inequality, $T^{\xi}(\sigma_q^2)$ is entirely covered by M spheres within distortion $(\sqrt{D+\mu} + \delta)^2$. Now,

$$\begin{aligned} \mathbf{E}|U(D)| \quad &= \quad \mathbf{E}\left\{ \sum_{\mathbf{x}_0 \in S \cap T^{\epsilon}(\sigma_q^2)} \prod_{i=1}^{M} 1\{||\mathbf{x}_0 - \hat{\mathbf{X}}(i)||^2 > N(D+\mu)\} \right\} \\ &= \quad \sum_{\mathbf{x}_0 \in S \cap T^{\epsilon}(\sigma_q^2)} \left[1 - \Pr\left\{||\mathbf{x}_0 - \hat{\mathbf{X}}(1)||^2 \leq N(D+\mu)\right\}\right]^M. \end{aligned} \tag{63}$$

It is easy to verify that $T_{\mathbf{x}_0}^{\epsilon}(W)$ is a subset of $T^{\xi}(\sigma_q^2 - D)$ for the above defined value of ξ and for $\mathbf{x}_0 \in T^{\epsilon}(\sigma_q^2)$. In a similar manner, it is easy to check that for a given $\mathbf{x}_0$, the set

$T^{\epsilon}_{\mathbf{x}_0}(W)$ has only $\hat{\mathbf{x}}$-vectors with $||\mathbf{x}_0 - \hat{\mathbf{x}}||^2 \leq N(D+\mu)$, where $\mu = \epsilon D(1+4\sqrt{D/\sigma_q^2})$. Since the codewords are selected randomly w.r.t. a uniform distribution within $T^{\xi}(\sigma_q^2 - D)$, then

$$\begin{aligned}\Pr\{||\mathbf{x}_0 - \hat{\mathbf{X}}(1)||^2 \leq N(D+\mu)\} &\geq \frac{\mathrm{Vol}\{T^{\epsilon}_{\mathbf{x}_0}(W)\}}{\mathrm{Vol}\{T^{\xi}(\sigma_q^2 - D)\}} \\ &\geq \left(1 - \frac{3}{N\epsilon^2}\right) \exp\left\{-N\left(\frac{1}{2}\ln\frac{\sigma_q^2}{D} + \eta\right)\right\} \qquad (64)\end{aligned}$$

where $\eta = \epsilon + \ln(1+\xi)$, and where we have used the above bounds on the volumes. Thus,

$$\begin{aligned}\mathbf{E}|U(D)| &\leq |S \cap T^{\epsilon}(\sigma_q^2)| \cdot \left[1 - \left(1 - \frac{3}{N\epsilon^2}\right) \exp\left\{-N\left(\frac{1}{2}\ln\frac{\sigma_q^2}{D} + \eta\right)\right\}\right]^M \\ &\leq (2\delta)^{-N}[2\pi e\sigma_q^2(1+\epsilon)]^{N/2} \cdot \\ &\quad \exp\left[-M\left(1 - \frac{3}{N\epsilon^2}\right) \exp\left\{-N\left(\frac{1}{2}\ln\frac{\sigma_q^2}{D} + \eta\right)\right\}\right], \qquad (65)\end{aligned}$$

where we have used the facts that $1 - u \leq e^{-u}$ and that the number of cubes in $T^{\epsilon}(\sigma_q^2)$ cannot exceed the ratio between the volume of $T^{\epsilon}(\sigma_q^2)$ and the volume of a cube $(2\delta)^N$. It is readily seen that for $\mathbf{E}|U(D)| \to 0$ as $N \to \infty$, it is sufficient that M would be of the exponential order of $\exp\{N[0.5\ln(\sigma_q^2/D) + 2\eta)]\}$.

Thus, we have proved that, given the fact that $\mathbf{x} \in T^{\epsilon}(\sigma_q^2)$, $\epsilon < \mu + \delta^2$, there exists a $(\sqrt{D+\mu} + \delta)^2$-admissible guessing strategy such that $G_N(\mathbf{x}) = 1$ if $\sigma_q^2 \leq D$ and $G_N(\mathbf{x}) \leq \exp\{N[0.5\ln(\sigma_q^2/D) + 2\eta]\}$ for $\sigma_q^2 > D$. Equivalently, for $D > (\sqrt{\mu} + \delta)^2$ there is a D-admissible guessing stategy with $G_N(\mathbf{x}) \leq \exp\{N[\max\{0, 0.5\ln(\sigma_q^2/((\sqrt{D} - \delta)^2 - \mu))\} + 2\eta]\}$. Thus, by letting δ and ϵ (and hence also μ and η) be arbitrary small, we can make the exponential order of $G_N(\mathbf{x})$ arbitrarily close to $\exp[NR(D,Q)]$, where Q is a zero-mean memoryless Gaussian source with variance σ_q^2.

For a given $0 < \Delta < D$, consider now the grid $\sigma_q^2(i) = D + \Delta i$, $i = 1, 2, ...$ Clearly, the sphere $\{\mathbf{x} : \mathbf{x}^t\mathbf{x} \leq ND\}$ together with the sets $T_i \stackrel{\triangle}{=} T^{\Delta/D}(\sigma_q^2(i))$, $i = 1, 2, ...$, entirely cover the space $\mathcal{R}^N$. With this choice, we have $\epsilon = \Delta/D$ and $\sigma_q^2 - D \geq \Delta$, and so, ξ and μ are uniformly upper bounded by $\Delta/D + 4\sqrt{\Delta/D}$ and 5Δ, respectively, independently of i. Therefore, similarly as in the proof of eq. (59), it is easy to see that the probability of T_i decays exponentially at the rate of $D(Q_i||P)$ (within a term that tends to zero as $\Delta \to 0$ independently of i), where Q_i is a zero-mean Gaussian source with variance $\sigma_q^2(i)$ (see eq. (56)). Consider now a guessing list whose first guess is $\hat{\mathbf{x}} = 0$, followed by code vectors of a code $\mathcal{C}_1$ that covers T_1 within distortion D, then a code $\mathcal{C}_2$ that covers T_2, and so on. Since

the codes are in the order of increasing exponential size, we have $G_N(\mathbf{x}) = 1$ for $\mathbf{x}^t\mathbf{x} \leq ND$, and $G_N(\mathbf{x}) \leq 1 + \sum_{j=1}^{i} |\mathcal{C}_j| \leq 1 + i|\mathcal{C}_i|$ for $\mathbf{x} \in T_i$. Therefore,

$$\mathbf{E}\{G_N(\mathbf{X})^\rho\} \leq 1 + \sum_{i=1}^{\infty} \Pr\{T_i\}(i|\mathcal{C}_i|)^\rho. \tag{66}$$

From the above considerations, it follows that the product $\Pr\{T_i\}|\mathcal{C}_i|^\rho$ is upper bounded by $\exp\{N[\rho R(D, Q_i) - D(Q_i||P) + \zeta(\Delta)]\}$ where $\zeta(\Delta) \to 0$ as $\Delta \to 0$, and so,

$$\mathbf{E}\{G_N(\mathbf{X})^\rho\} \leq 1 + \sum_{i=1}^{\infty} \exp\{N[\rho R(D, Q_i) - D(Q_i||P) + \zeta(\Delta) + \rho \ln i/N]\}. \tag{67}$$

Note that the exponential rate of each term of the last expression, as a function of i, is of the form $U_i = \ln(Ai + B) - Ci - D$, where A, B, and C are positive reals and D is immaterial since it represents multiplication by a constant factor. It is shown in the Appendix that

$$\lim_{N\to\infty} \frac{1}{N} \ln \left\{ \sum_{i=1}^{\infty} \exp[N(\ln(Ai + B) - Ci)] \right\} = \max_{i\geq 1} [\ln(Ai + B) - Ci] \tag{68}$$

Finally, from the continuity of the function $\rho R(D, Q) - D(Q||P)$ as a function of σ_q^2 in the Gaussian case, it follows that in the limit $N \to \infty$, followed by the limit of dense grids ($\Delta \to 0$), the maximum of $\rho R(D, Q_i) - D(Q_i||P) + \rho \ln i/N$ over i (which is $\max_i[\ln(Ai+B)-Ci]$) tends to the maximum of $\rho R(D, Q) - D(Q||P)$ over the continuum. □

6.2 Sources with Memory

A natural extension of Theorem 1 is to certain classes of stationary sources with memory. It is easy to extend Theorem 1 to stationary finite alphabet sources with the following property: There exists a finite positive number B such that for all m, n, $\mathbf{u} \in \mathcal{X}^m$, and $\mathbf{v} \in \mathcal{X}^n$,

$$|\ln P(\mathbf{X}_1^n = \mathbf{v}|\mathbf{X}_{-m+1}^0 = \mathbf{u}) - \ln P(\mathbf{X}_1^n = \mathbf{v})| \leq B, \tag{69}$$

where $\mathbf{X}_i^j$, for $i \leq j$, denotes $(X_i, ..., X_j)$. This assumption is clearly met, e.g., for Markov processes.

Theorem 6 *Let P be a finite alphabet stationary source with the above property for a given B. Then, $\mathcal{E}(D, \rho)$ exists and is given by*

$$\mathcal{E}(D, \rho) = \lim_{k\to\infty} E^k(D, \rho), \tag{70}$$

where

$$E^k(D, \rho) = \frac{1}{k} \max_Q [\rho R^k(D, Q) - D^k(Q||P)], \tag{71}$$

Q is a probability measure on $\mathcal{X}^k$, $D^k(Q||P)$ is the unnormalized divergence between Q and the kth order marginal of P, the maximum is over all kth order marginal PMF's, and $R^k(D,Q)$ is the rate-distortion function associated with a k-block memoryless source Q w.r.t. the alphabet $\mathcal{X}^k$ and the distortion measure induced by d additively over a k-block.

Proof. Assume, without essential loss of generality, that k divides N, and parse $\mathbf{x}$ into N/k non-overlapping blocks of length k, denoted $\mathbf{x}_{ik+1}^{ik+k}$, $i = 0, 1, ..., N/k-1$. Then, by the above property of P, we have

$$p^N(\mathbf{x}) \geq e^{-NB/k} \prod_{i=0}^{N/k-1} p^k(\mathbf{x}_{ik+1}^{ik+k}), \tag{72}$$

and so, by invoking the converse part of Theorem 1 to block memoryless sources, we get

$$\liminf_{N\to\infty} \frac{1}{N} \inf_{\mathcal{G}_N} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\} \geq E^k(D,\rho) - \frac{B}{k}. \tag{73}$$

Since this is true for every positive integer k, then

$$\liminf_{N\to\infty} \frac{1}{N} \inf_{\mathcal{G}_N} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\} \geq \limsup_{k\to\infty} E^k(D,\rho). \tag{74}$$

On the other hand, since

$$p^N(\mathbf{x}) \leq e^{NB/k} \prod_{i=0}^{N/k-1} p^k(\mathbf{x}_{ik+1}^{ik+k}), \tag{75}$$

then if we apply the universal guessing strategy $\mathcal{G}_N^*$ w.r.t. a superalphabet of k-blocks, then by invoking the direct part of Theorem 1 w.r.t. $\mathcal{X}^k$, we get

$$\limsup_{N\to\infty} \frac{1}{N} \inf_{\mathcal{G}_N} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\} \leq E^k(D,\rho) + \frac{B}{k}, \tag{76}$$

which then leads to

$$\limsup_{N\to\infty} \frac{1}{N} \inf_{\mathcal{G}_N} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\} \leq \liminf_{k\to\infty} E^k(D,\rho). \tag{77}$$

Combining eqs. (74) and (77), we conclude that both $N^{-1} \inf_{\mathcal{G}_N} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\}$ and $E^k(D,\rho)$ converge, and to the same limit. This completes the proof of Theorem 6. $\square$

Finally, it should be pointed out that a similar result can be further extended to a broader class of mixing sources by creating "gaps" between successive k-blocks. The length of each such gap should grow with k in order to make the successive blocks asymptotically independent, but at the same time should be kept small relative to k so that the distortion incurred therein would be negligibly small.

6.3 Guessing with Side Information

Another direction of extending our basic results for DMS's is in exploring the most efficient way of using side information. Consider a source that emits a sequence of i.i.d. pairs of symbols (X_i, Y_i) in $\mathcal{X} \times \mathcal{Y}$ w.r.t. to some joint probability measure $p(x, y)$. The guesser now has to guess $\mathbf{X} \in \mathcal{X}^N$ within distortion level D upon observing the statistically related side information $\mathbf{Y} \in \mathcal{Y}^N$.

Definition 4 *A D-admissible guessing strategy with side information $\mathcal{G}_N$ is a set $\{\mathcal{G}_N(\mathbf{y}), \mathbf{y} \in \mathcal{Y}^N\}$, such that for every $\mathbf{y} \in \mathcal{Y}^N$ with positive probability, $\mathcal{G}_N(\mathbf{y}) = \{\hat{\mathbf{x}}_{\mathbf{y}}(1), \hat{\mathbf{x}}_{\mathbf{y}}(2), ...\}$, is a D-admissible guessing strategy w.r.t. $p^N(\cdot|\mathbf{Y} = \mathbf{y})$.*

Definition 5 *The guessing function $G_N(\mathbf{x}|\mathbf{y})$ induced by a D-admissible guessing strategy with side information $\mathcal{G}_N$ maps $(\mathbf{x}, \mathbf{y}) \in \mathcal{X}^N \times \mathcal{Y}^N$ into a positive integer j, which is the index of the first guessing code word $\hat{\mathbf{x}}_{\mathbf{y}}(j) \in \mathcal{G}_N(\mathbf{y})$ such that $d(\mathbf{x}, \hat{\mathbf{x}}_{\mathbf{y}}(j)) \leq ND$. If no such code word exists in $\mathcal{G}_N(\mathbf{y})$, then $G_N(\mathbf{x}|\mathbf{y}) \stackrel{\triangle}{=} \infty$.*

Similarly as in Section 3, let us define

$$\mathcal{E}_{X|Y}(D, \rho) = \lim_{N \to \infty} \frac{1}{N} \inf_{\mathcal{G}_N} \ln \mathbf{E}\{G_N(\mathbf{X}|\mathbf{Y})^\rho\}, \tag{78}$$

provided that the limit exists, and where the infimum is over all D-admissible guessing strategies with side information. By using the same techniques as before, it can be easily shown that for a memoryless source P, if $\mathcal{X}$, $\hat{\mathcal{X}}$, and $\mathcal{Y}$ are all finite alphabets, then $\mathcal{E}_{X|Y}(D, \rho)$ exists and is given by

$$\mathcal{E}_{X|Y}(D, \rho) = E_{X|Y}(D, \rho) \stackrel{\triangle}{=} \sup_Q [\rho R_{X|Y}(D, Q) - D(Q||P)], \tag{79}$$

where $Q = \{q(x, y),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ is a joint PMF on $\mathcal{X} \times \mathcal{Y}$, $D(Q||P)$ is defined as the relative entropy between the joint PMF's, and $R_{X|Y}(D, Q)$ is the rate-distortion function of X given Y defined as

$$R_{X|Y}(D, Q) = \inf_W \sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{\hat{x} \in \hat{\mathcal{X}}} q(x, y) w(\hat{x}|x, y) \ln \frac{w(\hat{x}|x, y)}{\sum_{x' \in \mathcal{X}} q(x', y) w(\hat{x}|x', y)}, \tag{80}$$

where the infimum is over all channels W such that

$$\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} \sum_{\hat{x} \in \hat{\mathcal{X}}} q(x, y) w(\hat{x}|x, y) d(x, \hat{x}) \leq D. \tag{81}$$

It is straightforward to see that $E_{X|Y}(D,\rho) \le E_X(D,\rho)$ with equality when X and Y are independent under P.

For the proof of the direct part, we need the following version of the type covering lemma.

Lemma 5 *Let $T_{\mathbf{x}|\mathbf{y}}$ be a conditional type where $\mathbf{x}$ and $\mathbf{y}$ have a given empirical joint PMF $Q_{\mathbf{xy}}$. There exists a set $\mathcal{C}(\mathbf{y}) \subset \hat{\mathcal{X}}^N$ such that for any $\mathbf{x}' \in T_{\mathbf{x}|\mathbf{y}}$ and $D \ge 0$,*

$$\min_{\hat{\mathbf{x}} \in \mathcal{C}(\mathbf{y})} d(\mathbf{x}', \hat{\mathbf{x}}) \le ND, \tag{82}$$

and at the same time

$$\frac{1}{N} \ln |\mathcal{C}(\mathbf{y})| \le R_{X|Y}(D, Q_{\mathbf{xy}}) + o(N). \tag{83}$$

The proof is a straightforward extension of the proof of the ordinary type covering lemma and hence omitted.

Analogously to Theorem 4, we also have the following parametric form for the rate-distortion guessing exponent with side information:

$$E_{X|Y}(D,\rho) = \sup_{s \ge 0} \max_{f} \left\{ \ln \sum_{y \in \mathcal{Y}} \left[\sum_{x \in \mathcal{X}} p(x,y)^{1/(1+\rho)} f(x|y)^{\rho/(1+\rho)} \right]^{1+\rho} - sD \right\}, \tag{84}$$

where $f = \{f(x|y)\}$ are nonnegative numbers satisfying $\sum_{x \in \mathcal{X}} f(x|y) e^{-s\,d(x,\hat{x})/\rho} \le 1$ for each $\hat{x}, y$. Necessary and sufficient conditions for a given f to achieve the maximum in (84) are that there exists a set of nonnegative numbers $w(\hat{x}|x,y)$ satisfying $\sum_{\hat{x}} w(\hat{x}|x,y) = 1$ such that

$$w(\hat{x}|x,y) q(x|y) = m(\hat{x}|y) f(x|y) e^{-s\,d(x,\hat{x})/\rho} \tag{85}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, where $m(\hat{x}|y) = \sum_x w(\hat{x}|x,y)$ and $q(x|y) = cp(x|y)^{1/(1+\rho)} f(x|y)^{\rho/(1+\rho)}$, with c chosen so that $\sum_x q(x|y) = 1$.

The large deviations exponent is given by $\min D(Q||P)$, where both Q and P are joint PMFs on $\mathcal{X} \times \mathcal{Y}$, and the minimum is over all Q such that $R_{X|Y}(D,Q) \ge R$.

7 Conclusion and Future Work

We have provided a single-letter characterization to the optimum ρth order guessing exponent theoretically attainable for memoryless sources at a given distortion level. We have then studied the basic properties of this exponent as a function of the distortion level D and

the moment order ρ, along with its relation to the source coding error exponent. Finally, we gave a few extensions of our basic results to other cases of interest.

A few problems that remain open and require further work are the following.

General continuous-alphabet memoryless sources. Our first comment in the discussion that follows Theorem 1, naturally suggests to extend part (b) of this theorem to the continuous alphabet case. Obviously, if the source has bounded support, then after a sufficiently fine quantization, we are back in the situation of a finite alphabet source, and so every D-admissible guessing strategy for the quantized source is also $(D+\epsilon)$-admissible for the original source, where ϵ is controlled by the quantization. Thus, the proof of the direct part of Theorem 1 for the case of continuous alphabet with bounded support may rely on the finite alphabet case provided that the sequence of guessing exponents, corresponding to the sequence of quantized sources and their induced distortion measures, tends, in the high resolution limit, to the corresponding function $E(D,\rho)$ of the continuous source. However, the interesting and difficult case is that of unbounded support for which infinite guessing lists are always required. Moreover, in this case, quantization cannot be made uniformly fine unless the alphabet is countably infinite, but then the method of types is not directly applicable.

Hierarchical structures of guessing strategies. We mentioned in the Introduction that the guessing exponent serves as a measure of the search effort associated with lossy source coding, for a simple class of search schemes that is based on a fixed order of trials. A natural interesting extension would include classes of more sophisticated search schemes that take greater advantage of the distortion information obtained at each step. For example, if we revisit the Bob-and-Alice guessing game described in the Introduction, then what will happen if in order to achieve a target distortion level D, Alice is now allowed to first make guesses w.r.t. a larger distortion D', and then after her first success, to direct her guesses to the desired distortion level D? Thus, the next step is to extend the scope to that of multi-stage guessing strategies. In the limit of many stages corresponding to many distortion level thresholds, we are eventually taking full advantage of the exact distortion level information after each trial.

Joint source-channel guessing. It would be interesting to extend the guessing problem to the more complete setting of a communication system, that is, joint source-channel guessing. Here the problem is to jointly design a source-channel encoder at the transmitter side and a

guessing scheme at the receiver side, so as to minimize $\mathbf{E}G(\mathbf{X})^\rho$ for a prescribed end-to-end distortion level D. Besides the natural question of characterizing the guessing exponent for a given source and channel, it would be interesting to determine whether the separation principle of information theory applies in this context as well.

These issues among some others are currently under investigation.

Appendix

Proof of Lemma 2

First, we prove that $f^{**} \leq f$.

$$\begin{aligned} f^{**}(x) &= \sup_{y\geq 0}[xy - f^*(y)] && \text{(A.1)} \\ &= \sup_{y\geq 0}\inf_{x'\geq 0}[y(x-x') + f(x')] && \text{(A.2)} \\ &\leq \inf_{x'\geq 0}\sup_{y\geq 0}[y(x-x') + f(x')] && \text{(A.3)} \\ &= \inf_{x'\geq 0}\begin{cases} \infty & \text{if } x > x', \\ f(x') & \text{if } x \leq x' \end{cases} && \text{(A.4)} \\ &= f(x). && \text{(A.5)} \end{aligned}$$

By the saddle-point theorem, we have equality in (A.3) if $f(x)$ is convex. Equality (A.5) is due to the nondecreasing property of f.

Since f^{**} is the FLT of f^*, it is convex. So, to prove that f^{**} is equal to the lower convex hull of f, denoted $\hat{f}$, it suffices to prove the inequality $f^{**} \geq \hat{f}$. By rewriting the above equations for the convex function $\hat{f}$, we have $\hat{f}^{**} = \hat{f}$. Next, note that $\hat{f} \leq f$ implies $\hat{f}^* \geq f^*$, which in turn implies $\hat{f}^{**} \leq f^{**}$. Thus, we have

$$\hat{f} = \hat{f}^{**} \leq f^{**} \leq f, \qquad \text{(A.6)}$$

and the proof is complete.

Proof of Proposition 1

(a): Nonnegativity follows by the fact that $G_N(\mathbf{x}) \geq 1$ for every $\mathbf{x}$. The expression of $E(0,\rho)$ is obtained from standard maximization of $[\rho H(Q) - D(Q||P)]$ w.r.t. Q (see also [2]). $E(D,0) = 0$ since $G_N(\mathbf{x})^0 = 1$, P-almost everywhere for every D-admissible strategy. As for the expression of $D_0(\rho)$, we seek the supremum of D such that $E(D,\rho) = \sup_{R\geq 0}[\rho R - F(R,D)] > 0$. This means that there is $R \geq 0$ such $\rho R - F(R,D) > 0$, or equivalently,

$F(R,D)/R < \rho$. But the existence of such R in turn means that $\inf_{R\geq 0} F(R,D)/R$, which is defined as $a(D)$, must be less than ρ.

(b): Both monotonicity and convexity w.r.t. D follow immediately from the same properties of the rate-distortion function. Convexity and monotonicity also imply strict monotonicity in the indicated range.

(c): Nondecreasing monotonicity w.r.t. ρ follows from the monotonicity of $E(D,\rho,Q)$ w.r.t. ρ for every fixed D and Q. Convexity follows from the fact the $E(D,\rho)$ is the maximum over a family of affine functions $\{E(D,\rho,Q)\}$ w.r.t. ρ. Again, strict monotonicity follows from monotonicity and convexity.

(d): Continuity w.r.t. each one of the variables at strictly positive values follows from convexity. Continuity w.r.t. D at $D=0$ follows from continuity of $R(D,Q)$ both w.r.t. D and Q and continuity of $D(Q||P)$ w.r.t. Q. Continuity w.r.t. ρ at $\rho=0$ is immediate (see also part (e) below).

(e): By definition of $E(D,\rho)$, we have $E(D,\rho) \geq E(D,\rho,P) = \rho R(D,P)$, which proves the first part, and the fact that $\liminf_{\rho\to 0} E(D,\rho)/\rho \geq R(D,P)$. To complete the proof of the second part, it suffices to establish the fact that $\limsup_{\rho\to 0} E(D,\rho)/\rho \leq R(D,P)$. This in turn follows from the following consideration. Let $\{\rho_n\}_{n\geq 1}$ be an arbitrary positive sequence that tends to zero, and let $\{Q_n^*\}_{n\geq 1}$ be a corresponding sequence of maximizers of $E(D,\rho_n,Q)/\rho_n = R(D,Q) - D(Q||P)/\rho_n$. Now, obviously, Q_n^* must tend to P, otherwise $E(D,\rho_n,Q_n^*)/\rho_n$ would have a subsequence that tends to $-\infty$, contradicting the fact that $E(D,\rho)/\rho \geq R(D,Q)$ for all $\rho \geq 0$. Therefore,

$$\begin{aligned}\limsup_{n\to\infty} \frac{E(D,\rho_n)}{\rho_n} &= \limsup_{n\to\infty}\left[R(D,Q_n^*) - \frac{D(Q_n^*||P)}{\rho_n}\right] \\ &\leq \limsup_{n\to\infty} R(D,Q_n^*) \\ &= R(D,P). \end{aligned} \tag{A.7}$$

(f): The upper bound follows immediately by the fact that $E(D,\rho,Q) \leq \rho R(D,Q)$, and by taking the maximum w.r.t. Q. It then also implies that $\limsup_{\rho\to\infty} E(D,\rho)/\rho \leq R_{\max}(D)$. The converse inequality, $\liminf_{\rho\to\infty} E(D,\rho)/\rho \geq R_{\max}(D)$, follows from the following consideration. Without loss of generality, $p_{\min} \stackrel{\triangle}{=} \min_{x\in\mathcal{X}} p(x) > 0$, as if this was not the case, the alphabet $\mathcal{X}$ could have been reduced in the first place. Therefore, $\max_Q D(Q||P) \leq \ln(1/p_{\min})$, and so,

$$E(D,\rho) = \max_Q [\rho R(D,Q) - D(Q||P)]$$

$$\geq \max_{Q} \left[\rho R(D,Q) - \ln(\frac{1}{p_{\min}})\right]$$
$$= \rho R_{\max}(D) - \ln(\frac{1}{p_{\min}}). \tag{A.8}$$

Dividing by ρ and passing to the limit as $\rho \to \infty$, gives the desired result.

Proof of Lemma 4

Monotonicity in each argument is obvious from (35). Concavity in $s \geq 0$ for fixed $\rho \geq 0$: We shall use the geometric interpretation of $E(D,\rho)$ as the vertical axis intercept of the supporting line of slope D to the curve $K(s,\rho)$ vs. s. For a proof by contradiction, suppose $K(s,\rho)$ is not concave in s. Then, there exists D_1, $0 \leq s_1 < s_2$, $0 < \lambda < 1$ such that the supporting line of slope D_1 is tangential to $K(s,\rho)$ at s_1, s_2 and lies strictly above it at $s_\lambda \stackrel{\triangle}{=} \lambda s_1 + (1-\lambda)s_2$, i.e.,

$$K(s_i,\rho) = E(D_1,\rho) + s_i D_1, \quad \text{for } i = 1,2, \tag{A.9}$$

and

$$K(s_\lambda,\rho) < E(D_1,\rho) + s_\lambda D_1. \tag{A.10}$$

Observe that, from (A.9), $E(D_1,\rho)$ is upper bounded by $K(s_2,\rho)/s_2$. It is easy to see that $K(s,\rho)/s$ is a decreasing function of s and approaches $D_0(\rho)$ as $s \to 0$. So, we have $K(s_2,\rho)/s_2 \leq D_0(\rho)$ since by assumption $s_2 > 0$. Now, let (Q_1,W_1) achieve $E(D_1,\rho)$, i.e., $E(D_1,\rho) = \rho I(Q_1,W_1) - D(Q_1||P)$. Since $0 \leq D_1 \leq D_0(\rho)$, we must also have $\Delta(Q_1,W_1) = D_1$. ¿From (A.9), the pair (Q_1,W_1) is a saddle-point of (35) for $s = s_1, s_2$. Then, it is easy to see that (Q_1,W_1) must be a saddle-point of (35) for $s = s_\lambda$ as well, which implies $K(s_\lambda,\rho) = \lambda K(s_1,\rho) + (1-\lambda)K(s_2,\rho) = E(D_1,\rho) - s_\lambda \rho D_1$, contradicting (A.10). Proof of concavity in $\rho \geq 0$ for fixed $s \geq 0$ is similar, with $J(R,s)$ playing the role of $E(D,\rho)$, and will be omitted.

Proof of Proposition 2.

We first prove uniqueness of the PMF that achieves the maximum in (35). Let $s \geq 0$ be fixed. Note that the function $g(Q,W) \stackrel{\triangle}{=} \rho I(Q,W) + s\Delta(Q,W) - D(Q||P)$ is concave in Q and convex in W. So, any (Q_0,W_0) achieving $K(s,\rho)$ in (35) is a saddle-point of g, i.e., $g(Q,W_0) \leq g(Q_0,W_0) \leq g(Q_0,W)$ for all Q and W. Assume there exist two saddle-points (Q_0,W_0) and (Q_1,W_1) both achieving $K(s,\rho)$ with $Q_0 \neq Q_1$. Then, $g(Q_0,W_0) \leq g(Q_0,W_1) \leq g(Q_1,W_1)$, hence $g(Q_0,W_1) = K(s,\rho)$. By the strict concavity of g in Q, for any $0 < \lambda < 1$, we have $g(\lambda Q_0 + (1-\lambda)Q_1, W_1) > \lambda g(Q_0,W_1) + (1-\lambda)g(Q_1,W_1) =$

$K(s,\rho) = g(Q_1, W_1)$. This contradicts the assumption that (Q_1, W_1) is a saddle-point, and establishes the uniqueness of the PMF achieving (35), denoted in the rest of the proof as Q_s.

Next, fix $0 \leq D < D_0(\rho)$, and let Q^* be a PMF achieving $\max_Q[\rho R(D,Q) - D(Q||P)]$. Since $E(D,\rho) > 0$, $R(D,Q^*) > 0$ and there exists W^* such that $R(D,Q^*) = I(Q^*,W^*)$ and $\Delta(Q^*,W^*) = D$. For any $s_0 \in \mathcal{S}(D)$, we have $K(s_0,\rho) = E(D,\rho) + s_0 D = \rho I(Q^*,W^*) - D(Q^*||P) + s_0\Delta(Q^*,W^*)$. Thus, Q^* solves the maximization problem (35) for $s = s_0$, and hence, is uniquely determined as Q_{s_0}. Since s_0 is an arbitrary point in $\mathcal{S}(D)$, $Q_s = Q^*$ for all $s \in \mathcal{S}(D)$, as claimed.

Next, fix $s \geq 0$ and consider the equality (42) with $r = s/\rho$. Multiply each side by ρ, and subtract the term $D(Q_s||P)$. The resulting expression on the left side equals $K(s,\rho)$ iff $Q = Q_s$. We deduce that Q_s is the unique PMF that achieves the maximum in (44). It follows that Q^* achieves (44) for every $s \in \mathcal{S}(D)$.

Finally, to see that the maximum in (44) is achieved by a unique f_s, substitute the unique Q_s that maximizes the right side (which equals $Q^* = Q^*(D)$ for any D such that $s \in \mathcal{S}(D)$) and note that the resulting function of f is strictly concave in f.

Proof of eq. (59).

Consider an auxiliary zero-mean Gaussian memoryless source with variance $\sigma^2(1+\epsilon)$. Then,

$$\begin{aligned} 1 &\geq \int_{|\mathbf{x}^t\mathbf{x} - N\sigma^2| \leq N\epsilon\sigma^2} [2\pi\sigma^2(1+\epsilon)]^{-N/2} \exp\left[-\frac{\mathbf{x}^t\mathbf{x}}{2\sigma^2(1+\epsilon)}\right] d\mathbf{x} \\ &\geq \int_{|\mathbf{x}^t\mathbf{x} - N\sigma^2| \leq N\epsilon\sigma^2} [2\pi\sigma^2(1+\epsilon)]^{-N/2} \exp\left[-\frac{N\sigma^2(1+\epsilon)}{2\sigma^2(1+\epsilon)}\right] d\mathbf{x} \\ &= [2\pi e\sigma^2(1+\epsilon)]^{-N/2} \mathrm{Vol}\{T^\epsilon(\sigma^2)\} \end{aligned} \tag{A.11}$$

which completes the proof of eq. (59).

Proof of eq. (62).

First observe that eq. (61) defines a set of vectors $\hat{\mathbf{x}}$, which for a given $\mathbf{x}$, are just shifted versions of vectors $\mathbf{v}$. Therefore, the volume of $T^\epsilon_{\mathbf{x}}(W)$ is identical to the volume of the set $\tilde{T}^\epsilon_{\mathbf{x}}(W)$ of vectors $\mathbf{v}$ that satisfy the indicated constraints on $\mathbf{v}^t\mathbf{v}$ and $\mathbf{v}^t\mathbf{x}$. To lower bound the volume of this set, consider an auxiliary Gaussian random N-vector $\mathbf{V}$ with zero-mean uncorrelated components of variance $(D - D^2/\sigma_q^2)$. The probability that $\mathbf{V}$ would fall in

$\tilde{T}^{\epsilon}_{\mathbf{x}}(W)$ is upper bounded by

$$\begin{aligned}
\Pr\{\tilde{T}^{\epsilon}_{\mathbf{x}}(W)\} &= \int_{\tilde{T}^{\epsilon}_{\mathbf{x}}(W)} [2\pi(D-D^2/\sigma_q^2)]^{-N/2} \exp\left[-\frac{\mathbf{v}^t\mathbf{v}}{2(D-D^2/\sigma_q^2)}\right] d\mathbf{v} \\
&\leq \int_{\tilde{T}^{\epsilon}_{\mathbf{x}}(W)} [2\pi(D-D^2/\sigma_q^2)]^{-N/2} e^{-N(1-\epsilon)/2} d\mathbf{v} \\
&= [2\pi e^{1-\epsilon}(D-D^2/\sigma_q^2)]^{-N/2} \mathrm{Vol}\{\tilde{T}^{\epsilon}_{\mathbf{x}}(W)\}. \qquad \text{(A.12)}
\end{aligned}$$

On the other hand, this probability is lower bounded by the union bound and Chebychev's inequality as follows.

$$\begin{aligned}
1 - \Pr\left\{\tilde{T}^{\epsilon}_{\mathbf{x}}(W)\right\} &\leq \Pr\left\{|\mathbf{V}^t\mathbf{V} - N(D-D^2/\sigma_q^2)| > N\epsilon(D-D^2/\sigma_q^2)\right\} + \\
&\quad \Pr\left\{|\mathbf{V}^t\mathbf{x}| > \epsilon\sqrt{N(D-D^2/\sigma_q^2)\mathbf{x}^t\mathbf{x}}\right\} \\
&\leq \frac{\mathbf{E}[\mathbf{V}^t\mathbf{V} - N(D-D^2/\sigma_q^2)]^2}{N^2\epsilon^2(D-D^2/\sigma_q^2)^2} + \frac{\mathbf{E}(\mathbf{V}^t\mathbf{x})^2}{N\epsilon^2(D-D^2/\sigma_q^2)\mathbf{x}^t\mathbf{x}} \\
&= \frac{2N(D-D^2/\sigma_q^2)^2}{N^2\epsilon^2(D-D^2/\sigma_q^2)^2} + \frac{(D-D^2/\sigma_q^2)\mathbf{x}^t\mathbf{x}}{N\epsilon^2(D-D^2/\sigma_q^2)\mathbf{x}^t\mathbf{x}} \\
&= \frac{3}{N\epsilon^2}. \qquad \text{(A.13)}
\end{aligned}$$

Combining now eqs. (A.12) and (A.13) gives eq. (62).

Proof of eq. (68).

First observe that since the the function $U(x) = \ln(Ax+B) - Cx$ is monotonically decreasing beyond a certain value of x, the maximum over real x, and hence also over the integers $x = i$, must exist. Let then $U_{\max}$ be the maximum of $U(i)$, and let I be the smallest integer such that for all $i \geq I$, we have $\ln(Ai+B)/i \leq C/2$. Also, let J be the smallest integer i for which $-iC/2 < U_{\max}$, and let $K = \max\{I, J\}$. Clearly, $U_{\max}$ must be achieved for $i < K$, and so,

$$\begin{aligned}
\sum_{i=1}^{\infty} e^{N[\ln(Ai+B)-Ci]} &= \sum_{i=1}^{K} e^{N[\ln(Ai+B)-Ci]} + \sum_{i>K} e^{Ni[\ln(Ai+B)/i-C]} \\
&\leq K e^{NU_{\max}} + \sum_{i \geq K} e^{-NiC/2} \\
&= K e^{NU_{\max}} + \frac{e^{-NKC/2}}{1-e^{-NC/2}} \\
&\leq K e^{NU_{\max}} + \frac{e^{NU_{\max}}}{1-e^{-NC/2}} \qquad \text{(A.14)}
\end{aligned}$$

which is clearly of exponential order of $e^{NU_{\max}}$. On the other hand, the series in question is trivially lower bounded by its maximum term $e^{NU_{\max}}$. This completes the proof of eq. (68).

8 Acknowledgement

We are grateful to the anonymous reviewers for their very useful comments.

References

[1] R. Ahlswede, "Extremal properties of rate-distortion functions," *IEEE Trans. Inform. Theory*, vol. IT-36, no. 1, pp. 166-171, January 1990.

[2] E. Arikan, "An inequality on guessing and its application to sequential decoding," *IEEE Trans. Inform. Theory*, vol. IT-42, no. 1, pp. 99-105, January 1996.

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

[4] R. E. Blahut, *Principles and Practice of Information Theory.* Reading, MA: Addison-Wesley, 1987.

[5] I. Csiszár. "Generalized cutoff rates and Rényi's information measures," *IEEE Trans. Inform. Theory*, vol. 41, no. 1, pp. 26-34, January 1995.

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

[7] A. Dembo and O. Zeitouni, *Large Deviations Techinques and Applications,* Jones and Bartlett Publishers, 1993.

[8] W. H. R. Equitz and T. M. Cover, "Successive refinement of information," *IEEE Trans. Inform. Theory,* vol. IT-37, pp. 269-274, Mar. 1991.

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

[10] P. A. Humblet, "Generalization of Huffman coding to minimize the probability of buffer overflow," *IEEE Trans. Inform. Theory*, vol. 27, no. 2, pp. 230-232, March 1981.

[11] F. Jelinek, "Buffer overflow in variable length coding of fixed rate sources," *IEEE Trans. Inform. Theory*, vol. 14, no. 3, pp. 490-501, May 1968.

[12] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 197-199, 1974.

[13] N. Merhav, "Universal decoding for memoryless Gaussian channels with a deterministic interference," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 1261-1269, July 1993.

[14] N. Merhav, "On list size exponents in rate-distortion coding," submitted for publication, 1995.

[15] N. Merhav, "Universal coding with minimum probability of code word length overflow," *IEEE Trans. Inform. Theory*, Vol. IT-37, No. 3, pp. 556–563, May 1991.

[16] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Symp. on Math. Statist. Probability*, Berkeley, CA, 1961, vol. 1, pp. 547-561.

[17] R. T. Rockafellar, *Convex Analysis,* Princeton University Press, 1970.

[18] M. J. Weinberger, J. Ziv and A. Lempel, "On the optimal asymptotic performance of universal ordering and of discrimination of individual sequences," *IEEE Trans. Inform. Theory*, vol. IT-38, pp. 380-385, Mar. 1992.

[19] A. D. Wyner, "On the probability of buffer overflow under an arbitrary bounded input-output distribution," *SIAM J. Appl. Math.*, vol. 27, no. 4, pp. 544-570, December 1974.

[20] B. Yu and T. P. Speed, "A rate of convergence result for a D-semifaithful code," *IEEE Trans. Inform. Theory*, vol. IT-39, pp. 813-820, May 1993.

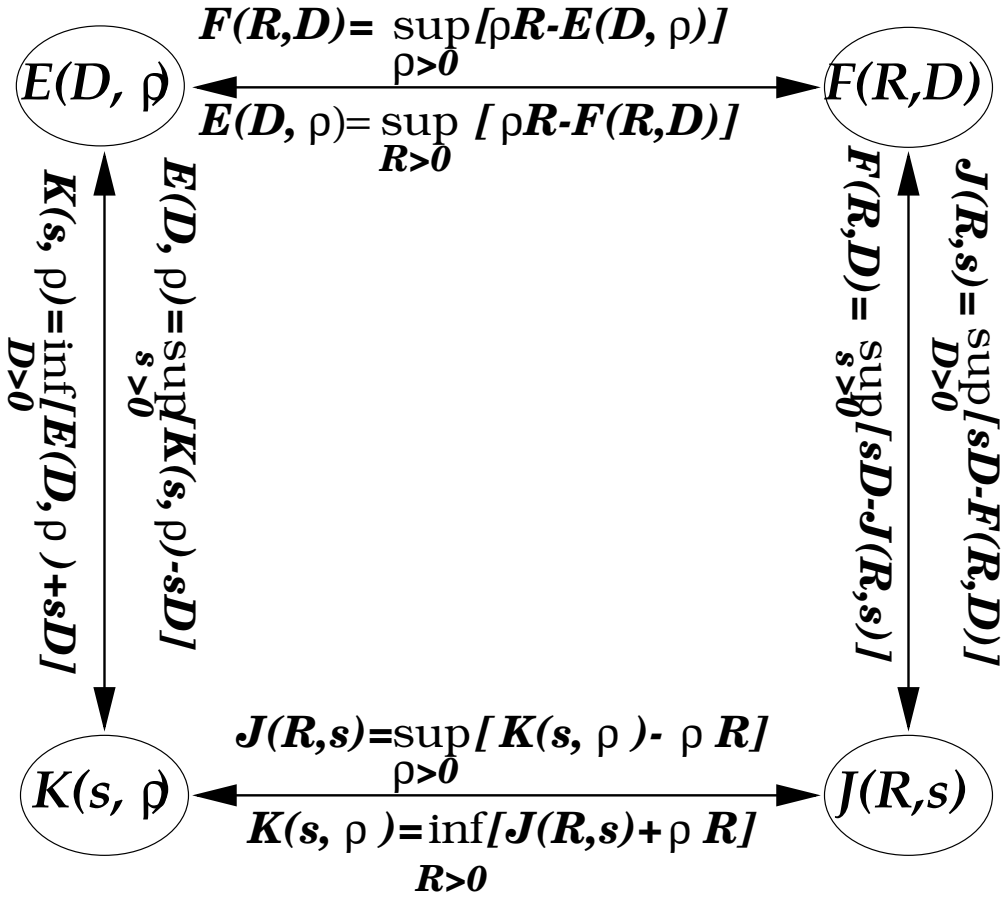

Figure 1: Transform relations among $E(D,\rho)$, $F(R,D)$, $J(R,s)$, and $K(s,\rho)$.

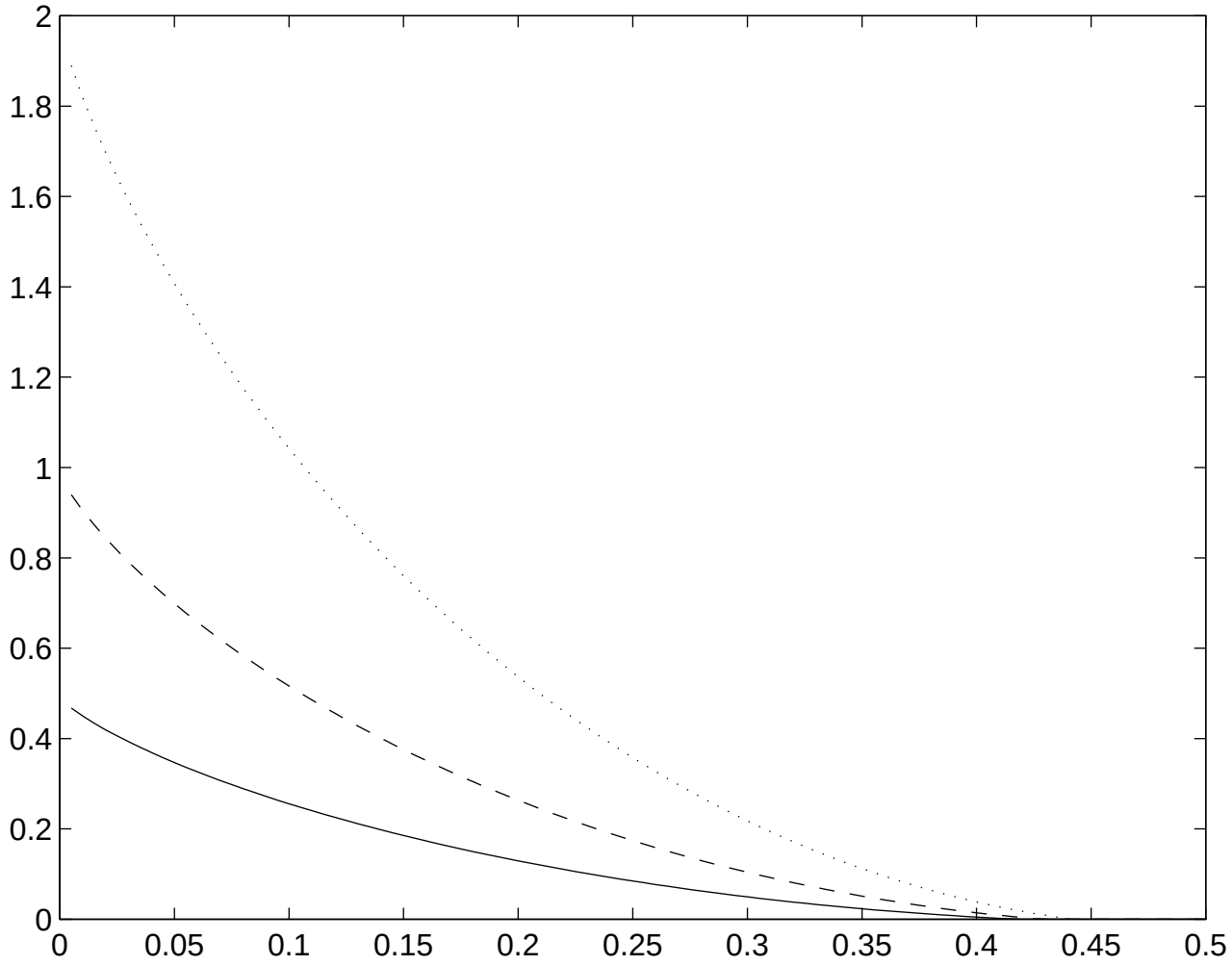

Figure 2: Curves of $E(D, \rho)$ vs. D for a binary source with letter probabilities $p(0) = 1 - p(1) = 0.4$, and the Hamming distortion measure. The solid line corresponds to $\rho = 0.5$, the dashed line to $\rho = 1$, and the dotted line to $\rho = 2$.