Guessing Subject to Distortion

Erdal Arikan
Bilkent University*

and

Neri Merhav
Computer Systems Laboratory and HP-ISC†

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

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 error exponent. Finally, we provide extensions of our main results to the Gaussian case, guessing with side information, and nonmemoryless sources.

*Electical-Electronics Engineering Department, Bilkent University, 06533 Ankara, Turkey. E-mail: arikan@ee.bilkent.edu.tr.

†On sabbatical leave at HP Laboratories. Current address: Hewlett-Packard Laboratories, 1501 Page Mill Road, Palo Alto, CA 94304, U.S.A. Email: merhav@hpl.hp.com. Permanent address: HP Israel Science Center, Technion City, Haifa 32000, Israel. E-mail: merhav@hp.technion.ac.il.

1 Introduction

Consider the following game: Bob selects a single secret random variable X. Then, Alice, who does not see X, presents to Bob a (fixed) sequence of guesses $\hat{x}(1)$, $\hat{x}(2)$, and so on. As long as the game continues, for each guess $\hat{x}(i)$, $i = 1, 2, ...$, Bob checks whether it is close enough to X in the sense that $d(X, \hat{x}(i)) \leq D$, for some distortion measure d and distortion level D. If the answer is affirmative, Bob says "Fine", Alice is scored by the number of guesses thus far $G(X)$, and the game ends. Otherwise, the game continues, Bob examines the next guess, and so on. 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 [1] 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.

The motivation of this problem becomes apparent if we think of the random variable X to be guessed as 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. Thus, the problem of guessing is best thought of as good *ordering* of good code words in a codebook.

In an earlier related work, driven by a similar motivation among others, Merhav [9] 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 [1] 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 [1] 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 rate-distortion 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, $D(Q||P)$ is the relative entropy between Q and P, and the maximum is over all probability distributions Q of memoryless sources. 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 [10] of order $1/(1 + \rho)$ (see [1] for more detail). In view of this, $E(D, \rho)/\rho$, for $D > 0$, can be thought of as Rényi's analogue to the rate-distortion function. 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 distortion below D. The roots of this phenomenon 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 critertion, $\Pr\{G(\mathbf{X}) \geq e^{NR}\}$ for a given $R > R(D,P)$. Obviously, the exponential behavior of this probability is given by the source coding error exponent $F(R,D)$ [7], [3] for memoryless sources. It turns out, indeed, that there is an intimate relation between the rate-distortion 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, and the source coding error exponent $F(R,D)$ as a function of R, are the one-sided Fenchel-Legendre transform of each other. Moreover, since the above mentioned universal guessing scheme minimizes all moments of $G(\mathbf{X})$ at the same time, it also gives the best attainable large deviations performance, universally for every memoryless source P and every $R > R(D,P)$.

In Section 5, we study some basic properties of the function $E(D,\rho)$, such as monotonicity, convexity in both arguments, continuity, differentiability, asymptotics, and others. Since $E(D,\rho)$ does not have a closed-form expression 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 nonmemoryless case, and 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 specific (deterministic) values they may take will be denoted by the respective lower case letters. A similar convention will apply to random and deterministic N-dimensional vectors, which will be denoted by the bold type font. 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 $\mathbf{x} \in \mathcal{X}^N$ 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 information-theoretic 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

$$\int_{\mathcal{X}} q(dx) \int_{\hat{\mathcal{X}}} w(d\hat{x}|x) d(x,\hat{x}) \leq D. \tag{6}$$

We will assume throughout that Q and P are such that there exists a *reference symbol* $\hat{x}_0 \in \hat{\mathcal{X}}$ so that $\mathbf{E}d(X,\hat{x}_0) < \infty$ (see also [2, eq. (4.2.7)]).

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*

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

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 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 D-admissible a 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 $T_{\mathbf{xy}}$ 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 $\mathbf{y}$ 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, and that Q has a reference symbol. 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 $\mathcal{X}$, having a reference symbol in the reconstruction alphabet $\hat{\mathcal{X}}$ w.r.t. a given distortion measure d.*

(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 assymetric 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 [1] 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 [10] of order θ, defined as

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

Another point that was mentioned briefly in the Introduction and should be emphasized is 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 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 since the criterion is $\mathbf{E}G_N(\mathbf{X})^\rho$, it 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, where the contribution of nontypical sequences can be ignored by using the asymptotic equipartition property (AEP), here the major contribution is provided by *nontypical* 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.

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 [11] from the lossless to the lossy case, as universal ordering of sequences in decreaing 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. 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 code words have different statistics than optimum ordinary rate-distortion code words 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 generating a list of code words that interlaces between the code words of a good rate-distortion codebook and a good guessing list. Since the location of each code word is at most doubled by this interlacing procedure, 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 candidiate 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 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,

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

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 source 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

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

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 source coding theorem,

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

which then gives

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

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

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

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

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

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

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

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 [4, p. 181] (see also, [12] 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 *([4], [12]): 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 [4] 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 Relation to the Source Coding Exponent

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 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* [7], [3, 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 D. 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 the sense that

$$\lim_{N\to\infty} \frac{1}{N} \ln \Pr\{G_N^*(\mathbf{X}) \geq e^{NR}\} = -F(R,D). \tag{28}$$

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 is result not very surprising if we recall that $\mathcal{G}_N^*$ asymptotically minimizes all the 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 us that for a fixed distortion level D, the guessing exponent $E(D,\rho)$ as a function of ρ, and the source coding error exponent $F(R,D)$ as a function of R, are related via the one-sided Fenchel-Legendre transform. This relation will help us to study some of the properties of the guessing exponent function in the next section.

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

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

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

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

The converse relationship (30) follows from (29) and the facts that $F(R,D)$ is a nondecreasing, convex function of $R \geq 0$ [3, p. 233]. Specifically,

$$\begin{aligned}
\sup_{\rho\geq 0}[\rho R - E(R,D)] &= \sup_{\rho\geq 0}\{\rho R - \sup_{R'\geq 0}[\rho R' - F(R',D)]\} \\
&= \sup_{\rho\geq 0}\inf_{R'\geq 0}[\rho(R-R') + F(R',D)] \\
&= \inf_{R'\geq 0}\sup_{\rho\geq 0}[\rho(R-R') + F(R',D)] \qquad (32) \\
&= \inf_{R'\geq 0}\begin{cases} F(R',D) & \text{if } R'\geq R \\ \infty & \text{if } 0\leq R' < R \end{cases} \\
&= F(R,D), \qquad (33)
\end{aligned}$$

where eq. (32) follows from the saddle-point theorem for convex-concave functions, and eq. (33) follows by the nondecreasing monotonocity of $F(R,D)$ in $R \geq 0$. This completes the proof of Theorem 2. □

Finally, for completeness, we would like to mention another related problem, which has not received attention so far to the best of our knowledge: 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.,

$$\mathcal{J}(R,s) = \lim_{N\to\infty}\frac{1}{N}\min_{\mathcal{C}_N}\ln \mathbf{E}\{e^{sd(\mathbf{X},\mathcal{C}_N)}\}, \qquad (34)$$

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

$$\mathcal{J}(R,s) = J(R,s) = \max_{Q}[sD(R,Q) - D(Q||P)], \qquad (35)$$

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 Fenchel-Legendre transform pair for fixed D, it is easy to see that $J(R,s)$ and $F(R,D)$ are also a one-sided Fenchel-Legendre transform pair, but for fixed R. Thus, $J(R,s)$ and $E(D,\rho)$ can be thought of as a two-dimensional Fenchel-Legendre transform pair.

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*

$$D_0(\rho) = \sup\{D: \ a(D) < \rho\}, \tag{36}$$

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

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

(c) $E(D,\rho)$ *is a strictly increasing, convex function of* ρ *for fixed* $0 \leq D \leq D_0(\rho)$*.*

(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.

The function $E(D,\rho)$ does not have a closed-form expression 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 [2, Sect. 4.3.1]. The Shannon lower bound to $R(D,Q)$ applies for difference distortion measures, i.e., distortion measures where $d(x,\hat{x})$ depends 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,*

$$E(D,\rho) \geq \max\{0, \rho H_{1/(1+\rho)}(P) - \rho\mathcal{H}(D)\}, \tag{37}$$

where $\mathcal{H}(D)$ *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* $R(D,Q) \geq \max\{0, H(Q) - \mathcal{H}(D)\}$ *is met with equality for every* Q*.*

Proof.

$$\begin{aligned}
E(D,\rho) &= \max_Q[\rho R(D,Q) - D(Q||P)] \\
&\geq \max_Q[\rho\max\{0, H(Q) - \mathcal{H}(D)\} - D(Q||P)] \qquad (38)\\
&= \max_Q \max\{-D(Q||P), \rho[H(Q) - \mathcal{H}(D)] - D(Q||P)\} \\
&= \max\{\max_Q[-D(Q||P)], \max_Q[\rho[H(Q) - \mathcal{H}(D)] - D(Q||P)]\} \qquad (39)\\
&= \max\{0, \rho H_{1/(1+\rho)}(P) - \rho\mathcal{H}(D)\}. \qquad (40)
\end{aligned}$$

□

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

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

Fig. 1 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 $R(D,P) = 0$. This means that exponentially many guesses may still be required even if the source does not convey any information at a certain distortion level.

As mentioned earlier, $E(D,\rho)$ does not always have a closed-form expression. The following theorem provides a parametric representation of $E(D,\rho)$ as a double maximization problem, which may be more suitable for iterative procedures for calculating $E(D,\rho)$.

Theorem 4 *For every $D \geq 0$ and $\rho \geq 0$ the guessing exponent $E(D,\rho)$ can be expressed as*

$$E(D,\rho) = \max_{s\geq 0}[K(s,\rho) - sD] \tag{42}$$

where

$$K(s,\rho) = (1+\rho)\max_{f\in\mathcal{F}_s}\ln\sum_{x\in\mathcal{X}} p(x)^{1/(1+\rho)} f(x)^{\rho/(1+\rho)}, \tag{43}$$

and where $\mathcal{F}_s$ is the set of all $f = \{f(x) \geq 0 : x \in \mathcal{X}\}$ with nonnegative elements such that

$$\sum_{x\in\mathcal{X}} f(x)e^{-sd(x,\hat{x})/\rho} \leq 1 \tag{44}$$

for all $\hat{x} \in \hat{\mathcal{X}}$. Necessary and sufficient conditions for a given $\{f(x)\}$ to achieve the maximum are that first, thre exists a set of nonnegative numbers $\{m(\hat{x}),\ \hat{x} \in \hat{\mathcal{X}}\}$ that satisfy

$$\frac{f(x)}{q^*(x)}\sum_{\hat{x}\in\hat{\mathcal{X}}} m(\hat{x})e^{-sd(x,\hat{x})/\rho} = 1 \tag{45}$$

for all $x \in \mathcal{X}$, and that secondly, eq. (44) is satisfied with equality for each $\hat{x}$ with $m(\hat{x}) > 0$. The numbers $q^*(x)$ in (45) are given by

$$q^*(x) = cp(x)^{1/(1+\rho)} f(x)^{\rho/(1+\rho)}, \tag{46}$$

where c is a normalizing constant so that $\sum_{x\in\mathcal{X}} q^*(x) = 1$.

Proof. This theorem is largely a restatement of [6, Theorem 9.4.1, p. 459], which states that for any $D \geq 0$ and Q

$$R(D,Q) = \max_{r\geq 0}\max_{f}\left[H(Q) + \sum_{x\in\mathcal{X}} q(x)\ln f(x) - rD\right], \tag{47}$$

where the maximum is over all $f = \{f(x), \in \mathcal{X}\}$ with nonnegative elements satisfying $\sum_{x\in\mathcal{X}} f(x)e^{-rd(x,\hat{x})} \leq 1$ for all $\hat{x} \in \hat{\mathcal{X}}$. Using (47) in the definition of $E(D,\rho)$, we have

$$E(D,\rho) = \max_{Q} \max_{r\geq 0} \max_{f} \left\{ \rho \left[H(Q) + \sum_{x\in\mathcal{X}} q(x) \ln f(x) - rD \right] - D(Q||P) \right\}. \tag{48}$$

We now substitute $s = \rho r$ and carry out the maximization in (48), first w.r.t. Q. For any fixed s and f in the constraint set, it is easy to see that the right-hand side of (48) is concave in Q and that the maximizing distribution Q^* is given by (46). Substituting Q^* into (48) yields (42). The necessary and sufficient conditions for f to achieve the maximum are a restatement of similar conditions by Gallager, stated for the distribution Q^* here. This completes the proof of Theorem 4. $\square$

This theorem 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 (42) that

$$E(0,\rho) = (1+\rho)\ln\left[\sum_{x} P(x)^{1/(1+\rho)}\right], \tag{49}$$

which agrees with the result in [1].

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 any $s \geq 0$, a feasible f satisfying (44) is

$$f(x) = f_s \stackrel{\triangle}{=} \frac{1}{1+(K-1)e^{-s/\rho}}, \quad \text{all } x \in \mathcal{X}. \tag{50}$$

With this choice of f, eq. (42) is maximized over $s \geq 0$ by

$$s^* = \rho \ln \frac{(K-1)(D-1)}{D}, \tag{51}$$

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 (42) (without maximizing over s and f), we have for any $\rho \geq 0$, and $0 \leq D \leq (K-1)/K$,

$$E(D,\rho) \geq \rho[H_{1/(1+\rho)}(P) - h(D) - D\ln(K-1)], \tag{52}$$

where $h(D) = -D\ln(D) - (1-D)\ln(1-D)$. It is easy to see that the condition (45) for equality in (52) will be satisfied if and only if

$$q^*(x) \geq \frac{1}{e^{s^*/\rho} + (K-1)} = \frac{D}{K-1}, \quad \text{all } x \in \mathcal{X}, \tag{53}$$

where $q^*(x)$ is as defined in (46). Thus, equality holds in (52) 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. (52) coincides with the Shannon lower bound as for the Hamming distortion measure, $\mathcal{H}(D) = h(D) + D\ln(K-1)$.

We next provide several technical claims concerning some basic properties of the representation of Theorem 4. These properties will then be useful to establish the differentiability of $E(D,\rho)$ w.r.t. both arguments. The proofs all appear in the Appendix.

Lemma 2 *$K(s,\rho)$ is a concave nondecreasing function of $s \geq 0$.*

Let $\mathcal{S}(D)$ be the set of $s \geq 0$ achieving the maximum in (42), i.e., $\mathcal{S}(D) \stackrel{\triangle}{=} \{s \geq 0 : E(D,\rho) = K(s,\rho) - sD\}$. In general, $\mathcal{S}(D)$ need not be a singleton; however, by the concavity of $K(s,\rho)$ in s, $\mathcal{S}(D)$ must be in the form of an interval $[s_1, s_2]$. Though the maximizing s need not be unique, there is uniqueness for f.

Lemma 3 *For any fixed $D \geq 0$, the maximum over f in (43) is achieved by a unique f simultaneously for all $s \in \mathcal{S}(D)$.*

Corollary 2 *For fixed $D \in [0, D_0(\rho))$, $E(D,\rho,Q)$ is maximized by a unique PMF Q^*.*

The uniqueness of Q^* facilitates the proof of differentiability of $E(D,\rho)$ w.r.t. ρ.

Proposition 2 *$E(D,\rho)$ is differentiable w.r.t. ρ for all $D \geq 0$ with the possible exception of $D = D_0(\rho)$. The derivative is given by $R(D,Q^*)$, where Q^* is the maximizer of $E(D,\rho,Q)$.*

The proof appears in the Appendix.

Note that parts (c), (e), and (f) of Proposition 1 together with Proposition 2, imply that if $R(D,P) < R_{\max}(D)$, then the slope of $E(D,\rho)$ as a function of ρ for fixed D, grows monotonically and continuously in the range $(R(D,P), R_{\max}(D))$ as ρ grows from zero to infinity.

In the remaining part of this section, we examine the differentiability of $E(D,\rho)$ w.r.t. D. We prove that for finite distortion measures $E(D,\rho)$ is differentiable w.r.t. D. For unbounded distortion measures we show by an example that $E(D,\rho)$ need not be differentiable w.r.t. D. We begin by giving a geometrical interpretation to $K(s,\rho)$.

Lemma 4 *$K(s,\rho)$ is the vertical axis ($D = 0$) intercept of the supporting line of slope $-s$ to $E(D,\rho)$ vs. D curve, i.e.,*

$$K(s,\rho) = \inf_{D \geq 0}[E(D,\rho) + sD]. \tag{54}$$

The supporting line of slope $-s$ to the $E(D,\rho)$ vs. D curve is given by $K(s,\rho) - sD$. This line touches the curve at $(D^*, E(D^*,\rho))$ iff $E(D^*,\rho) = K(s,\rho) - sD^*$, i.e., iff $s \in \mathcal{S}(D^*)$. Thus, $\mathcal{S}(D)$ has the geometrical interpretation of being the set of s such that the supporting line of slope $-s$ meets $E(D,\rho)$ at D^*. In general, $\mathcal{S}(D^*)$ is an interval $[s_1, s_2]$, as already noted before. By geometrical considerations it is obvious that $-s_1$ (resp., $-s_2$) equals the right (resp., left) derivative of $E(D,\rho)$ w.r.t. D at $D = D^*$ (which always exist by convexity). Thus, $E(D,\rho)$ is differentiable w.r.t. D at $D = D^*$ iff $\mathcal{S}(D^*)$ is a singleton ($s_1 = s_2$).

Lemma 5 *$\mathcal{S}(D^*)$ is a singleton $\{s^*\}$ iff $K(s,\rho)$ is strictly concave over $s \in \mathcal{S}(D^*)$.*

Lemma 6 *If the distortion measure is finite, i.e., if $\max_{x,\hat{x}} d(x,\hat{x}) < \infty$, and $D_0(\rho) > 0$, then $\mathcal{S}(D)$ is a singleton for all $0 \le D < D_0(\rho)$.*

Corollary 3 *If the distortion measure is finite, then: (i) $E(D,\rho)$ is differentiable w.r.t. D for all $D \ge 0$, with the possible exception of $D = D_0(\rho)$, and the derivative $\partial E(D,\rho)/\partial D$ is given by $-s^*$, where s^* is the unique maximizer of eq. (42). (ii) $\lim_{D\to 0} \partial E(D,\rho)/\partial D = -\infty$.*

Part (i) of Corollary 3 follows directly from the preceding lemma. Part (ii) follows from the following consideration. The slope of $E(D,\rho)$ as $D \to 0$ will be finite only if there exists some $D \ge 0$ with $\mathcal{S}(D)$ having the form $[s,\infty)$. But this contradicts the fact that $\mathcal{S}(D)$ is a singleton.

The following example shows that if the distortion function is unbounded, then $E(D,\rho)$ may not be differentiable w.r.t. D.

Example 3: Nondifferentiable $E(D,\rho)$ (cf. [6, 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

$$\begin{bmatrix} 0 & \infty & \infty & \infty & 1 & \infty & 3 \\ \infty & 0 & \infty & \infty & 1 & \infty & 3 \\ \infty & \infty & 0 & \infty & \infty & 1 & 3 \\ \infty & \infty & \infty & 0 & \infty & 1 & 3 \end{bmatrix}. \tag{55}$$

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

$$f_s = \begin{cases} 0.25e^{(3s/\rho)} & \text{if } 0 \le s \le \frac{\rho}{2}\ln(2), \\ 0.5e^{(s/\rho)} & \text{if } \frac{\rho}{2}\ln(2) < s \le \rho\ln(2), \\ 1 & \text{if } s > \rho\ln(2). \end{cases} \tag{56}$$

Substituting the resulting $K(s,\rho)$ in (42) and taking the supremum over $s \ge 0$, we obtain

$$E(D,\rho) = \begin{cases} \rho(2-D)\ln(2) & \text{if } 0 \le D \le 1, \\ \frac{1}{2}\rho(3-D)\ln(2) & \text{if } 1 < D \le 3, \\ 0 & \text{if } D > 3. \end{cases} \tag{57}$$

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 analogue 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*

$$\mathcal{E}(D, \rho) = E(D, \rho), \tag{58}$$

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

$$D(Q||P) = \frac{1}{2}\left(\frac{\sigma_q^2}{\sigma_p^2} - \ln\frac{\sigma_q^2}{\sigma_p^2} - 1\right). \tag{59}$$

Since

$$R(D, Q) = \max\left\{0, \frac{1}{2}\ln\frac{\sigma_q^2}{D}\right\} \tag{60}$$

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

$$E(D, \rho) = \max\left\{0, \frac{1}{2}\left[\rho\ln\frac{\sigma_p^2}{D} + (1+\rho)\ln(1+\rho) - \rho\right]\right\}. \tag{61}$$

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 the case (see Proposition 1(f)).

Proof. 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 straightforward extension of the proof of Theorem 1(b), provided that we have a suitable version of the type covering lemma for Gaussian sources.

To this end, let us first define the notion of a Gaussian type class. For given $\epsilon > 0$ and $\sigma^2 > 0$, a *Gaussian type class* $T^\epsilon(\sigma^2)$ is defined as the set of all N-vectors $\mathbf{x}$ with the property $|\mathbf{x}^t\mathbf{x} - N\sigma^2| \le N\epsilon$, where $\mathbf{x}$ is understood as a column vector and the superscript t

denotes vector transposition. It is easy to show, by applying Stirling's approximation to the formula of the volume of the N-dimensional sphere, that the volume of $T^\epsilon(\sigma^2)$ is given by

$$\mathrm{Vol}\{T^\epsilon(\sigma^2)\} = \exp\left\{\frac{N}{2}[\ln(2\pi e(\sigma^2+\epsilon)) + o(N)]\right\}. \tag{62}$$

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

$$\hat{X} = \begin{cases} (1-\frac{D}{\sigma_q^2})X + V & \text{if } D < \sigma_q^2 \\ 0 & \text{if } D \geq \sigma_q^2 \end{cases}, \tag{63}$$

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

$$T^\epsilon_{\mathbf{x}}(W) = \left\{\hat{\mathbf{x}}:\ \hat{\mathbf{x}} = (1-\frac{D}{\sigma_q^2})\mathbf{x} + \mathbf{v};\ \mathbf{v}^t\mathbf{v} \leq N(D - D^2/\sigma_q^2),\ |\mathbf{v}^t\mathbf{x}| \leq N\epsilon\right\}. \tag{64}$$

It is easy to show that

$$\mathrm{Vol}\{T^\epsilon_{\mathbf{x}}(W)\} = \exp\left\{\frac{N}{2}[\ln[2\pi e(D-\frac{D^2}{\sigma_q^2})] + o(N)\right\}, \tag{65}$$

since the probability of $T^\epsilon_{\mathbf{x}}(W)$ with respect to W given $\mathbf{x}$ is close to unity (see also [8]).

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 $\hat{\mathbf{x}} \equiv 0$ suffices. Assume next, that $\sigma_q^2 > D$. Let us construct a grid S of all vectors in 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 $(\mathbf{x}_0 - \hat{\mathbf{x}}(i))^t(\mathbf{x}_0 - \hat{\mathbf{x}}(i)) > ND$ for all $i = 1, ..., M$, namely, cubes of $T^\epsilon(\sigma_q^2)$ whose centers are not covered by $\mathcal{C}$.

Consider the following random coding argument. Let $\hat{\mathbf{X}}(1), \ldots, \hat{\mathbf{X}}(M)$ denote i.i.d. vectors drawn uniformly in $T^\epsilon(\sigma_q^2 - D)$. 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, and therefore $T_\epsilon(\sigma_q^2)$ is entirely covered by M spheres of radius $\sqrt{N(D+\delta^2)}$. Now,

$$\begin{aligned}\mathbf{E}|U(D)| &= \mathbf{E}\left\{\sum_{\mathbf{x}_0 \in S\cap T^\epsilon(\sigma_q^2)} \prod_{i=1}^M 1\{(\mathbf{x}_0 - \hat{\mathbf{X}}(i))^t(\mathbf{x}_0 - \hat{\mathbf{X}}(i)) > ND\}\right\} \\ &= \sum_{\mathbf{x}_0 \in S\cap T^\epsilon(\sigma_q^2)} [1 - \Pr\{(\mathbf{x}_0 - \hat{\mathbf{X}}(1))^t(\mathbf{x}_0 - \hat{\mathbf{X}}(1)) \leq ND\}]^M. \end{aligned} \tag{66}$$

Since $T^\epsilon_{\mathbf{x}}(W)$ is a subset of $T^\epsilon(\sigma_q^2 - D)$ for $\mathbf{x} \in T^\epsilon(\sigma_q^2)$, and it includes only $\hat{\mathbf{x}}$-vectors with $(\mathbf{x} - \hat{\mathbf{x}})^t(\mathbf{x} - \hat{\mathbf{x}}) \leq ND$, then

$$\begin{aligned}\Pr\{(\mathbf{x}_0 - \hat{\mathbf{X}}(1))^t(\mathbf{x}_0 - \hat{\mathbf{X}}(1)) \leq ND\} &\geq \frac{\mathrm{Vol}\{T^\epsilon_{\mathbf{x}}(W)\}}{\mathrm{Vol}\{T^\epsilon(\sigma_q^2 - D)\}} \\ &= \exp\{-N[R(D,Q) + o(N)]\}, \qquad (67)\end{aligned}$$

where we have used the above expressions for volumes of types. Thus,

$$\begin{aligned}\mathbf{E}|U(D)| &\leq |S \cap T^\epsilon(\sigma_q^2)|\,[1 - \exp\{-N[R(D,Q) + o(N)]\}]^M \\ &\leq \delta^{-N} \exp[\frac{N}{2}\ln(2\pi e(\sigma_q^2 + \epsilon))] \cdot \\ &\quad \exp[-M\exp\{-N[R(D,Q) + o(N)]\}], \qquad (68)\end{aligned}$$

where we have used the fact that the number of cubes in $T^\epsilon(\sigma_q^2)$ cannot exceed the volume ratio, and the fact that $1 - u \leq e^{-u}$. It is readily seen that $\mathbf{E}|U(D)| \to 0$ if M is of any exponential order larger than $\exp[NR(D,Q)]$.

Having the covering lemma in its continuous version, we now proceed as in the discrete case, where we now divide the range $\sigma_q^2 > 0$ into infinitesimally small intervals of size ϵ, and collect the contributions of all these intervals. It is easy to show that the probability of $T^\epsilon(\sigma_q^2)$ is exponentially $e^{-ND(Q||P))}$. The summation is over infinitely many such intervals but it is exponentially dominated by the maximum term. This completes the proof of Theorem 5. □

6.2 Nonmemoryless Sources

A natural extension of Theorem 1 is to certain classes of stationary nonmemoryless sources. 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 $\mathbf{v} \in \mathcal{X}^n$,

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

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*

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

where

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

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 nonoverlapping 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

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

Since this is true for every positive integer k, then

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

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

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

which then leads to

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

Combining eqs. (74) and (77), we conclude that both $N^{-1} \inf_{\mathcal{G}} \ln \mathbf{E}\{G_N(\mathbf{X})^\rho\}$ 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 the broader class of ψ-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 a 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

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

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 $P = \{p(x, y),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$, $Q = \{q(x, y),\ x \in \mathcal{X},\ y \in \mathcal{Y}\}$ for some joint PMF q, $D(Q||P)$ is defined as the relative entropy between the pair PMF's, and $R_{X|Y}(D, Q)$ is the rate-distortion function with side information of X given Y both being governed by Q. More precisely, $R_{X|Y}(D, Q)$ is 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 $E_{X|Y}(D, \rho) \leq E_X(D, \rho)$ with equality when X and Y are independent under P.

Again, for the proof of the direct part, we need to modify the type covering lemma. The suitable version of the type covering lemma is as follows.

Lemma 7 *Let $T_{\mathbf{x}|\mathbf{y}}$ be a conditional type of $\mathbf{x}$ given $\mathbf{y}$ which have a given empirical joint PMF $Q_{\mathbf{x}\mathbf{y}}$. 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 $D \geq 0$,*

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

and at the same time

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

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:

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

where $f = \{f(x|y)\}$ are nonnegative numbers satisfying, for each $\hat{x}$ and y,

$$\sum_{x\in\mathcal{X}} f(x|y) e^{-sd(x,\hat{x})/\rho} \leq 1. \tag{85}$$

Necessary and sufficient conditions for a given f to achieve the maximum in (84) are that first, there exists a set of nonnegative numbers $\{m(\hat{x}|y)\}$ satisfying

$$\frac{f(x|y)}{q^*(x|y)} \sum_{\hat{x}\in\hat{\mathcal{X}}} m(\hat{x}|y) e^{-sd(x,\hat{x})/\rho} = 1, \tag{86}$$

for all $x \in \mathcal{X}$, $y \in \mathcal{Y}$, and (85) is satisfied with equality for each $\hat{x}$ with $m(\hat{x}|y) > 0$. Here, $q^*(x|y)$ is given by

$$q^*(x|y) = cp(x|y)^{1/(1+\rho)} f(x|y)^{\rho/(1+\rho)}, \tag{87}$$

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 $R_{X|Y}(D,Q) \geq R$. Other results from the unconditional case extend as well to the setting with side information.

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 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.

Transform Relations. We have seen four bivariate functions $E(D,\rho)$, $F(R,D)$, $J(R,s)$, and $K(s,\rho)$, all related via the Fenchel-Legendre transform, or similar operators. The Fenchel-Legendre transform relationships among the first three functions have been mentioned in Section 4. The function $K(s,\rho)$ of Section 5 can be thought of being related to $-E(D,\rho)$ by an operator that is similar to the Fenchel-Legendre transform, except the maximization over the nonnegative reals is replaced by minimization (see Theorem 4 and Lemma 4). Fig. 2 summarizes these relationships. It will be interesting to complete the fourth edge of the depicted square, that is, to find a suitable direct transform relation between $K(s,\rho)$ and $J(R,s)$ for fixed s. This will result in a dual to Theorem 4 for representing $J(R,s)$ in a parametric form.

Hierarchical structures of guessing strategies. We mentioned earlier that the guessing exponent can be interpreted as a measure of computational effort associated with lossy source coding. Many practical source coding schemes are based on hierarchical tree structures which may substantially reduce this computatonal effort, sometimes without even sacrificing rate-distortion optimality [5]. This motivates to extend the scope to that of multi-stage guessing strategies. 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?

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 Proposition 1

(a): Nonnegativity follows by the fact that $G_N(\mathbf{x}) \geq 1$ for every $\mathbf{x}$. The expression for $E(0,\rho)$ is obtained from standard maximization of $[\rho H(Q) - D(Q||P)]$ w.r.t. Q (see also [1]). $E(D,0) = 0$ 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. Since $F(R,D)/R = \sup_{\rho\geq 0}[\rho - E(D,\rho)/R]$ is a monotonically increasing function of $R > 0$, then the infimum of this function is attained at the limit as $R \to 0$.

(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 $\limsup_{\rho\to 0}E(D,\rho)/\rho\leq R(D,P)$, or equivalently, for every $\epsilon>0$, there exists $\rho>0$ below which $E(D,\rho)/\rho\leq R(D,P)+\epsilon$. This in turn follows from the following consideration. By Theorem 2, we have

$$\frac{E(D,\rho)}{\rho}=\sup_{R\geq 0}\left[R-\frac{F(R,D)}{\rho}\right], \tag{A.1}$$

and let us denote by $B(R)$ the bracketed expression for a given D and ρ. Since $F(R,D)$ is convex in R [3, Theorem 6.6.2], then $B(R)$ is concave. Thus, if $B(R_2)<B(R_1)$ for some $R_2>R_1$, then $\sup_{R\geq 0}B(R)=\sup_{0\leq R\leq R_2}B(R)$. Now, fix $\epsilon>0$, and let $\rho=F(R(D,P)+\epsilon,D)/(2\epsilon)$. Then, $B(R(D,P))=R(D,P)$, but $B(R(D,P)+\epsilon)=R(D,P)-\epsilon<B(R(D,P))$, and so,

$$\begin{aligned}\frac{E(D,\rho)}{\rho} &= \sup_{R\geq 0}B(R)\\ &= \sup_{0\leq R\leq R(D,P)+\epsilon}B(R)\\ &\leq \sup_{0\leq R\leq R(D,P)+\epsilon}R\\ &= R(D,P)+\epsilon.\end{aligned} \tag{A.2}$$

(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,

$$\begin{aligned}E(D,\rho) &= \max_Q[\rho R(D,Q)-D(Q||P)]\\ &\geq \max_Q\left[\rho R(D,Q)-\ln(\frac{1}{p_{\min}})\right]\\ &= \rho R_{\max}(D)-\ln(\frac{1}{p_{\min}}).\end{aligned} \tag{A.3}$$

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

Proof of Lemma 2

$K(s,\rho)$ is nondecreasing since the constraint set $\mathcal{F}_s$ becomes larger as s increases. To prove concavity, let $s_2 > s_1 \geq 0$ be arbitrary and let f_i achieve $K(s_i,\rho)$, $i = 1,2$. Since the function $\alpha e - \beta$ has a positive semidefinite Hessian w.r.t. α and β, it is convex in the pair (α,β), and so, $f \stackrel{\triangle}{=} \lambda f_1 + (1-\lambda) f_2$ belongs to constraint set $\mathcal{F}_s$, where $s \stackrel{\triangle}{=} \lambda s_1 + (1-\lambda)s_2$. Specifically, for any fixed $(x,\hat{x})$, we have

$$f(x)e^{-sd(x,\hat{x})/\rho} \leq \lambda f_1(x)e^{-s_1 d(x,\hat{x})/\rho} + (1-\lambda) f_2(x) e^{-s_2 d(x,\hat{x})}, \tag{A.4}$$

which when summed over x, confirms the membership of f in $\mathcal{F}_s$. Thus,

$$\begin{aligned} K(s,\rho) &\geq (1+\rho)\ln \sum_{x\in\mathcal{X}} p(x)^{1/(1+\rho)} f(x)^{\rho/(1+\rho)} && \text{(A.5)} \\ &\geq \lambda K(s_1,\rho) + (1-\lambda)K(s_2,\rho), && \text{(A.6)} \end{aligned}$$

where (A.6) follows by the concavity of the right side of (A.5). Note that since we have strict concavity here, equality holds in (A.6) if and only if $f_1 = f_2$.

Proof of Lemma 3

Since the right side of (43) is strictly concave in f for fixed s, then the maximizing f for this s must be unique. Thus, the assertion of the lemma is trivially true if $\mathcal{S}(D)$ is a singleton. Assume then, that $\mathcal{S}(D)$ is not a singleton and contains two elements s_1 and s_2 with non-identical maximizers f_1 and f_2, respectively. For any $0 < \lambda < 1$, we have $s \stackrel{\triangle}{=} \lambda s_1 + (1-\lambda)s_2 \in \mathcal{S}(D)$ and $f \stackrel{\triangle}{=} \lambda f_1 + (1-\lambda) f_2 \in \mathcal{F}_s$. So, again by the strict concavity of the right side of (43) in f,

$$K(s,\rho) > \lambda K(s_1,\rho) + (1-\lambda)K(s_2,\rho). \tag{A.7}$$

Substituting

$$K(s_i,\rho) = E(D,\rho) + s_i D, \quad i = 1,2, \tag{A.8}$$

in (A.7), we obtain

$$K(s,\rho) > E(D,\rho) + sD, \tag{A.9}$$

which contradicts eq. (42).

Proof of Proposition 2

Let $\{\delta_n\}$ be an arbitrary real sequence that tends to zero. Let Q^* be the unique maximizer of $E(D,\rho,Q)$, and let Q_n^* be the maximizer of $E(D,\rho+\delta_n,Q)$. Then,

$$\begin{aligned} \liminf_{n\to\infty} \frac{E(D,\rho+\delta_n) - E(D,\rho)}{\delta_n} &= \liminf_{n\to\infty} \frac{E(D,\rho+\delta_n,Q_n) - E(D,\rho,Q^*)}{\delta_n} \\ &\geq \liminf_{n\to\infty} \frac{E(D,\rho+\delta_n,Q^*) - E(D,\rho,Q^*)}{\delta_n} \\ &= R(D,Q^*). && \text{(A.10)} \end{aligned}$$

To prove the converse inequality, consider next a subsequence $\epsilon_k = \delta_{n_k}$, $k = 1,2,..$, such that $[E(D,\rho+\epsilon_k) - E(D,\rho)]/\epsilon_k$ converges to $\limsup_{n\to\infty}[E(D+\delta_n,\rho) - E(D,\rho)]/\delta_n$. Let

$\tilde{Q}_k^* = Q_{n_k}^*$ denote the maximizer of $E(D, \rho + \epsilon_k, Q)$. Now, by the continuity of $E(D, \rho)$ w.r.t. ρ, the continuity of $E(D, \rho, Q)$ w.r.t. Q, and the uniqueness of Q^*, the sequence $\tilde{Q}_k^*$ must converge to Q^*. Thus,

$$\begin{aligned}
\limsup_{n\to\infty} \frac{E(D, \rho + \delta_n) - E(D, \rho)}{\delta_n} &= \lim_{k\to\infty} \frac{E(D, \rho + \epsilon_k) - E(D, \rho)}{\epsilon_k} \\
&\le \limsup_{k\to\infty} \frac{E(D, \rho + \epsilon_k, \tilde{Q}_k) - E(D, \rho, \tilde{Q}_k)}{\epsilon_k} \\
&= \limsup_{k\to\infty} R(D, \tilde{Q}_k) \\
&= R(D, Q^*), \qquad \text{(A.11)}
\end{aligned}$$

where the last equality follows from continuity of $R(D, Q)$ w.r.t. Q for a bounded distortion measure. Combining the upper bound and the lower bound, we conclude that $\partial E(D, \rho)/\partial \rho$ exists and is given by $R(D, Q^*)$.

Proof of Lemma 4

By eq. (42), the right-hand side of (54) can be written as

$$\inf_{D\ge 0} \sup_{s'\ge 0} [K(s', \rho) - s'D + sD]. \qquad \text{(A.12)}$$

Since the bracketed expression is concave in s and convex (actually affine) in D, there exists a saddle-point and

$$\begin{aligned}
\inf_{D\ge 0} \sup_{s'\ge 0} [K(s', \rho) - s'D + sD] &= \sup_{s'\ge 0} \inf_{D\ge 0} [K(s', \rho) - s'D + sD] \qquad \text{(A.13)} \\
&= K(s, \rho), \qquad \text{(A.14)}
\end{aligned}$$

where equality (A.14) follows by a simple reasoning: For $s' > s$, the right-hand side of (A.13) equals $-\infty$; for $0 \le s' \le s$, it equals $K(s', \rho)$. So, the supremum is achieved at $s' = s$.

Proof of Lemma 5

Necessity: Suppose $\mathcal{S}(D^*) = [s_1, s_2]$ with $s_2 > s_1$. Then,

$$K(s, \rho) = E(D^*, \rho) + sD^* \qquad \text{(A.15)}$$

for all $s \in [s_1, s_2]$, and $K(s, \rho)$ is affine. Thus, if $K(s, \rho)$ is strictly concave over $\mathcal{S}(D^*)$, $\mathcal{S}(D^*)$ must be a singleton.

Sufficiency: Suppose $\mathcal{S}(D^*)$ is a singleton $\{s^*\}$ and that $K(s, \rho)$ is affine over an interval $[s^* - \epsilon, s^* + \epsilon]$, for some $\epsilon > 0$. In other words, assume that, for all $s \in [s^* - \epsilon, s^* + \epsilon]$,

$$K(s, \rho) = K(s^*, \rho) + \beta(s - s^*) \qquad \text{(A.16)}$$

where β is a constant independent of ϵ. We first show that β must be equal to D^*.

Let D_1 be an arbitrary point in $\mathcal{D}(s^*+\epsilon)$. Since $D^* \notin \mathcal{D}(s^*+\epsilon)$ (otherwise, $s^*+\epsilon$ would belong to $\mathcal{S}(D^*)$, violating the singleton hypothesis), we have $D_1 < D^*$. Now,

$$\begin{aligned} E(D_1,\rho) &= K(s^*+\epsilon,\rho) - (s^*+\epsilon)D_1 && \text{(A.17)} \\ &= K(s^*,\rho) + \beta\epsilon - (s^*+\epsilon)D_1 && \text{(A.18)} \\ &= K(s^*,\rho) - s^*D^* + \epsilon(\beta - D_1) + s^*(D^* - D_1) && \text{(A.19)} \\ &= E(D^*,\rho) + \epsilon(\beta - D_1) + s^*(D^* - D_1). && \text{(A.20)} \end{aligned}$$

Since $D_1 < D^*$, we have $E(D^*,\rho) \le E(D_1,\rho)$, and it follows from (A.20) that

$$\epsilon(\beta - D_1) + s^*(D^* - D_1) \ge 0. \tag{A.21}$$

This is possible only if $\beta > D_1$. Since ϵ can be chosen arbitrarily small, taking the infimum we obtain $\beta \ge D^*$. Applying the same type of argument to $D_2 \in \mathcal{D}(s^*-\epsilon)$, we obtain $\beta \le D^*$, proving that $\beta = D^*$.

Thus, $K(s,\rho)$ has the form

$$K(s,\rho) = K(s^*,\rho) + D^*(s - s^*) \tag{A.22}$$

for all $s \in [s^*-\epsilon, s^*+\epsilon]$, for some $\epsilon > 0$. However, this implies

$$E(D^*,\rho) = K(s,\rho) - sD^* \tag{A.23}$$

for all $s \in [s^*-\epsilon, s^*+\epsilon]$, contradicting the hypothesis that $\mathcal{S}(D^*)$ is a singleton. □

Proof of Lemma 6

Suppose to the contrary that for some $D \in (0, D_0(\rho))$, the set $\mathcal{S}(D)$ contains two non-identical elements $s_1 > s_2 > 0$. We know that the same f achieves the maximum in (43) for $s = s_1$ and $s = s_2$. Then, by the condition (44) of Theorem (4),

$$\sum_x f(x) e^{-sd(x,\hat{x})/\rho} \le 1, \tag{A.24}$$

for all $\hat{x}$ and for $s = s_1, s_2$. We must have equality in (A.24) for $s = s_1$ and for some $\hat{x}$, otherwise f would not be a maximizer. Unless $d(x,\hat{x}) = 0$ for all x with $f(x) > 0$, we will then have a violation of the same inequality when $s = s_2$. On the other hand, $f(x)$ must be strictly positive for all x for otherwise the condition (45) for the optimality of f cannot be satisfied by any $m(\hat{x})$. So, we must have $d(x,\hat{x}) = 0$ for all x. This means that the representation letter $\hat{x}$ is at zero distance from all source letters. Thus, $E(D,\rho) = 0$ for all $D \ge 0$, contradicting the assumption that $D_0(\rho) > 0$.

8 References

[1] 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.

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

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

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

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

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

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

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

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

[10] 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.

[11] 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.

[12] 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.

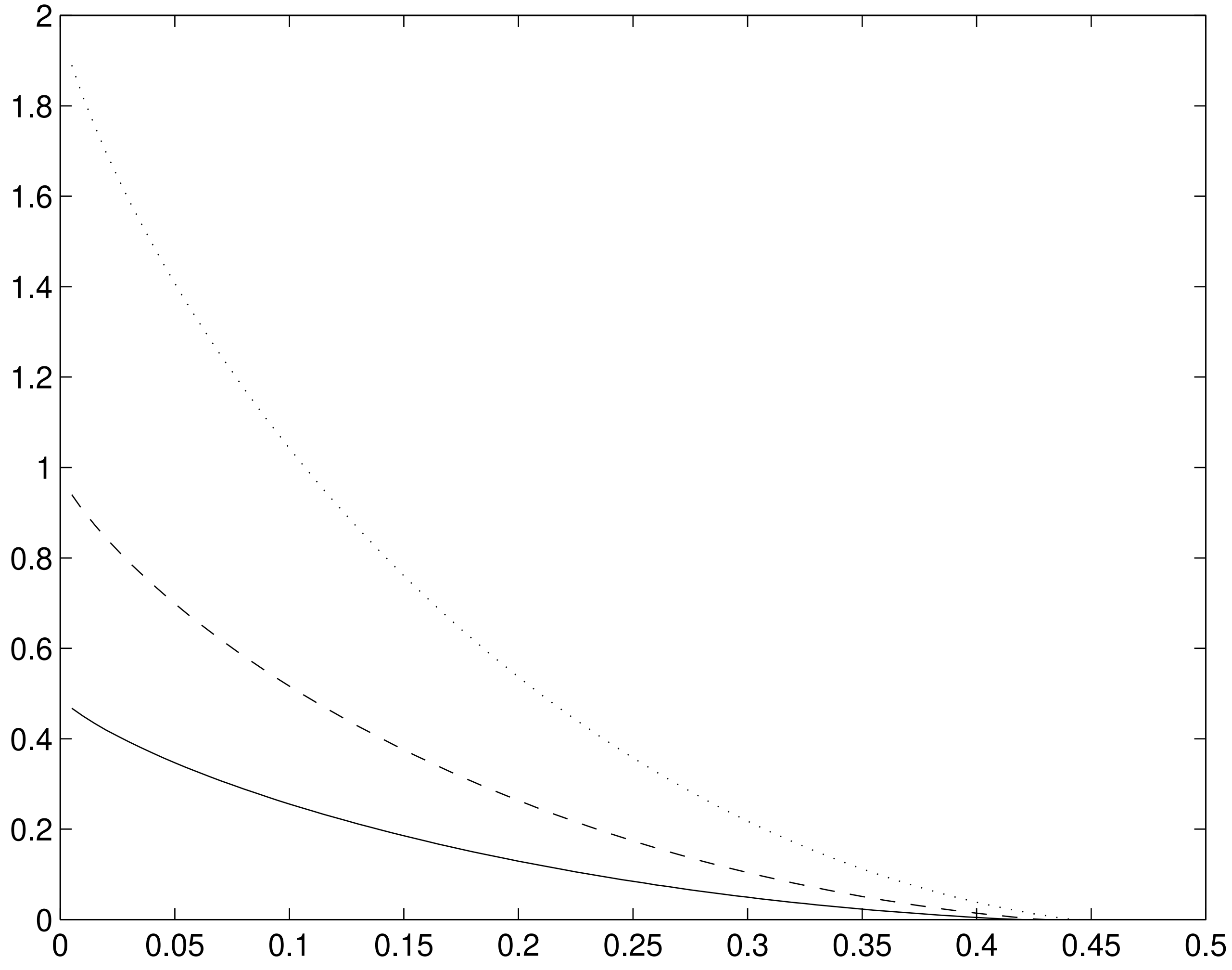

Figure 1: Curves of $E(D,\rho)$ vs. D for a binary source with letter probabilities $p(0) = 1-p(1) = 0.4$, and the Hamming distortion measure. The solid line corresponds to $\rho = 0.5$, the dashed line to $\rho = 1$, and the dotted line to $\rho = 2$.

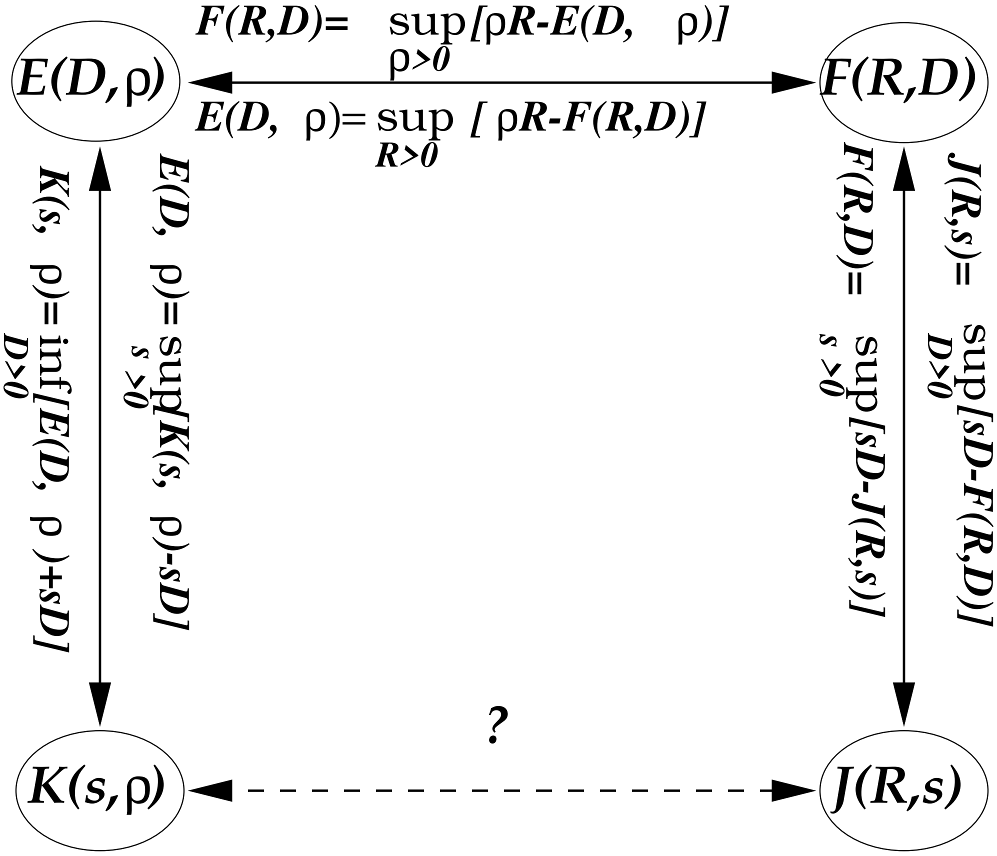

Figure 2: Transform relations among $E(D,\rho)$, $F(R,D)$, $J(R,s)$, and $K(s,\rho)$.