On List Size Exponents in Rate-Distortion Coding

Neri Merhav
Computer Systems Laboratory and HP-ISC*

Keywords: rate-distortion theory, $D-$semifaithful codes, memoryless sources, list size exponents.

Abstract

We charactrize and investigate the highest achievable exponential growth rate of the expected number of rate-distortion codewords that are within a certain distance from a given source vector.

*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. Email: merhav@hp.technion.ac.il

1 Introduction

Consider a lossy source coding scheme that operates strictly above that rate-distortion function of a given source. In other words, some degree of redundancy is present, which is normally the case. Although undesirable from the strict rate-distortion tradeoff point of view, this redundancy often provides certain degrees of freedom to design codebooks that possess other desirable properties. One well known example is that of codebooks that achieve the optimum source coding error exponent (see, e.g., [1],[2]), that is, the fastest exponential decay of the probability of the nonexistence of codewords within a given distance from the random source vector.

In this paper we consider a different, though somewhat related, approach for utilizing the degrees of freedom provided by the coding redundancy. We seek codebooks that are optimal in the sense of maximizing the *expected* number of codewords that are within the given distance from the input vector. We shall refer to the set of codewords within a sphere of a given radius, centered at the source vector $\boldsymbol{X}$ as a *list* of codewords for $\boldsymbol{X}$, or simply as a list for $\boldsymbol{X}$. There are a few motivations for maximizing the expected list size. One motivation is that a typically large list size is likely to yield reduction in the number of needed metric computations until finding the first codeword that represents the source within the desired distortion level. Another motivation is that if there is a large variety of candidate codewords for encoding the source within a certain distortion, one has the flexibility to select among them the best one in the sense of having some other desirable features. For instance, one would like to minimize distortion discontinuities near the block edges, or to minimize another fidelity criterion, and so on. The goal of maximizing lists in lossy source coding is somewhat dual to that of minimizing lists in channel decoding (see, e.g., [3],[4] and references therein). The motivations, however, are substantially different.

The main contribution of this paper is in deriving a single-letter characterization of the best achievable exponential growth rate of the expected list size as the block length tends to infinity. In particular, in Section 3 we provide a coding theorem and a converse theorem that characterize the highest attainable exponential rate. In Section 4, we discuss some properties of this exponential growth rate as a function of the coding rate and the distortion level, along with some examples.

2 Problem Definition

Let $\mathcal{X} = \{1, 2, ..., J\}$ be the alphabet of a memoryless source characterized by a vector of letter probabilities $P = \{p(1), p(2), ..., p(J)\}$, and let $\mathcal{Y} = \{1, 2, ..., K\}$ denote the reproduction alphabet. Let $\rho(x, y) \geq 0$, $x \in \mathcal{X}$, $y \in \mathcal{Y}$, denote a single-letter distortion measure. Let $\mathcal{X}^n$ and $\mathcal{Y}^n$ denote the nth Cartesian powers of $\mathcal{X}$ and $\mathcal{Y}$, respectively. The distortion between a source vector $\boldsymbol{x} = (x_1, ..., x_n) \in \mathcal{X}^n$ and a reproduction vector $\boldsymbol{y} = (y_1, ..., y_n) \in \mathcal{Y}^n$ is defined as $\rho_n(\boldsymbol{x}, \boldsymbol{y}) = \sum_{i=1}^{n} \rho(x_i, y_i)$. Let $\mathcal{C}$ denote a codebook of M code vectors of length n, $\boldsymbol{y}_1, \boldsymbol{y}_2, ..., \boldsymbol{y}_M$ in $\mathcal{Y}^n$, where $M = e^{nR}$, R being the coding rate in nats per source symbol. Throughout the sequel, random variables will be denoted by capital letters while specific values they may take on will be denoted by the respective lower case letters. Thus, $\boldsymbol{X} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n) \in \mathcal{X}^n$ will denote a random source vector drawn by the given memoryless source P, and $\boldsymbol{Y} = (\boldsymbol{Y}_1, ..., \boldsymbol{Y}_n) \in \mathcal{Y}^n$ will denote the random reproduction vector corresponding to $\boldsymbol{X}$ for a given codebook $\mathcal{C}$.

Let $R(D)$ denote the rate-distortion function of the source w.r.t $\rho(\cdot, \cdot)$. We shall assume that the coding rate R is strictly larger than $R(D)$ for a given allowable per-letter distortion level D. For a given input vector $\boldsymbol{x} \in \mathcal{X}^n$ and codebook $\mathcal{C}$, let $L(\boldsymbol{x})$ denote the number of code vectors $\{\boldsymbol{y}_i\}$ that satisfy $\rho_n(\boldsymbol{x}, \boldsymbol{y}_i) \leq nD$, i.e.,

$$L(\boldsymbol{x}) = \sum_{i=1}^{M} I\{\rho_n(\boldsymbol{x}, \boldsymbol{y}_i) \leq nD\}, \tag{1}$$

where $I\{\cdot\}$ denotes the indicator function.

Our goal is to characterize the best achievable exponential growth rate of $EL(\boldsymbol{X})$ as $n \to \infty$ among all rate R codebooks that satisfy $E\rho_n(\boldsymbol{X}, \boldsymbol{Y}) \leq nD$.

3 Main Result

Before we present the main result, we first define some additional notation.

For a given vector $\boldsymbol{x} \in \mathcal{X}^n$, the empirical probability mass function (EPMF) is the vector $P_{\boldsymbol{x}} = (p_{\boldsymbol{x}}(x), x \in \mathcal{X})$, where $p_{\boldsymbol{x}}(x) = n_{\boldsymbol{x}}(x)/n$, $n_{\boldsymbol{x}}(x)$ being the number of occurrences of the letter x in the vector $\boldsymbol{x}$. The type $T_{\boldsymbol{x}}$ of a vector $\boldsymbol{x}$ is the set of all vectors $\boldsymbol{x}' \in \mathcal{X}^n$ such that $P_{\boldsymbol{x}'} = P_{\boldsymbol{x}}$. The empirical entropy associated with $\boldsymbol{x}$ is the entropy associated with its EPMF $P_{\boldsymbol{x}}$, i.e.,

$$H_{\boldsymbol{x}}(X) = H(P_{\boldsymbol{x}}) = -\sum_{x \in \mathcal{X}} p_{\boldsymbol{x}}(x) \ln p_{\boldsymbol{x}}(x), \tag{2}$$

where both notations will be used interchangeably. Similarly, the EPMF of the sequence $\boldsymbol{y}$ will be denoted $Q_{\boldsymbol{y}} = (q_{\boldsymbol{y}}(y), y \in \mathcal{Y})$, where $q_{\boldsymbol{y}}(y)$ is the relative frequency of y in $\boldsymbol{y}$. The respective type will be denoted by $T_{\boldsymbol{y}}$, and the associated empirical entropy will be denoted by $H_{\boldsymbol{y}}(Y)$ or $H(Q_{\boldsymbol{y}})$.

In the same manner, for sequence pairs $(\boldsymbol{x}, \boldsymbol{y}) \in \mathcal{X}^n \times \mathcal{Y}^n$ the joint EPMF is the matrix $V_{\boldsymbol{xy}} = \{v_{\boldsymbol{xy}}(x, y), x \in \mathcal{X}, y \in \mathcal{Y}\}$ where $v_{\boldsymbol{xy}}(x) = n_{\boldsymbol{xy}}(x, y)/n$, $n_{\boldsymbol{xy}}(x, y)$ being the number of joint occurrences of $x_i = x$ and $y_i = y$. The joint type $T_{\boldsymbol{xy}}$ of $(\boldsymbol{x}, \boldsymbol{y})$ is the set of all pair sequences $(\boldsymbol{x}', \boldsymbol{y}') \in \mathcal{X}^n \times \mathcal{Y}^n$ for which $V_{\boldsymbol{x}'\boldsymbol{y}'} = V_{\boldsymbol{xy}}$. The empirical joint entropy denoted as $H_{\boldsymbol{xy}}(X, Y)$ or $H(V_{\boldsymbol{xy}})$ is the joint entropy associated with the EPMF $V_{\boldsymbol{xy}}$.

A conditional type $T_{\boldsymbol{x}|\boldsymbol{y}}$ for a given $\boldsymbol{y}$ is the set of all sequences $\boldsymbol{x}'$ in $\mathcal{X}^n$ for which $(\boldsymbol{x}', \boldsymbol{y}) \in T_{\boldsymbol{xy}}$. The conditional EPMF is defined as the matrix $W_{\boldsymbol{x}|\boldsymbol{y}} = \{w_{\boldsymbol{x}|\boldsymbol{y}}(x|y), x \in \mathcal{X}, y \in \mathcal{Y}\}$, where $w_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \stackrel{\triangle}{=} v_{\boldsymbol{xy}}(x, y)/q_{\boldsymbol{y}}(y)$ if $q_{\boldsymbol{y}}(y) > 0$ and zero otherwise. The empirical conditional entropy of $\boldsymbol{x}$ given $\boldsymbol{y}$ is given by

$$H_{\boldsymbol{x}|\boldsymbol{y}}(X|Y) = H_{\boldsymbol{xy}}(X, Y) - H_{\boldsymbol{y}}(Y) = -\sum_{x \in \mathcal{X}} \sum_{y \in \mathcal{Y}} v_{\boldsymbol{xy}}(x, y) \ln w_{\boldsymbol{x}|\boldsymbol{y}}(x|y). \tag{3}$$

Now, for a given output EPMF $Q_{\boldsymbol{y}}$ and distortion level D, let

$$G_n(D; Q_{\boldsymbol{y}}) = \left\{ W_{\boldsymbol{x}|\boldsymbol{y}} : \sum_{y \in \mathcal{Y}} q_{\boldsymbol{y}}(y) \sum_{x \in \mathcal{X}} w_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \rho(x, y) \leq D \right\}, \tag{4}$$

and for a given memoryless source P, let

$$F_n(D;Q_{\boldsymbol{y}}) \stackrel{\triangle}{=} \min_{W_{\boldsymbol{x}|\boldsymbol{y}} \in G_n(D;Q_{\boldsymbol{y}})} \sum_{y \in \mathcal{Y}} q_{\boldsymbol{y}}(y) \sum_{x \in \mathcal{X}} w_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \ln \frac{w_{\boldsymbol{x}|\boldsymbol{y}}(x|y)}{p(x)}, \tag{5}$$

where the minimum is understood as infinity if $G_n(D;Q_{\boldsymbol{y}})$ is empty. Finally, let

$$E_n(R,D) \stackrel{\triangle}{=} \max_{Q_{\boldsymbol{y}}} [R \wedge H(Q_{\boldsymbol{y}}) - F_n(D;Q_{\boldsymbol{y}})], \tag{6}$$

where the maximum is taken over all EPMFs of vectors in $\mathcal{Y}^n$.

We are now ready to state the main theorem.

Theorem 1 *Let P be a memoryless source with a rate-distortion function $R(D)$ w.r.t a distortion measure $\rho(\cdot,\cdot)$.*

(a) Converse part: For any rate R block code of length n,

$$EL(\boldsymbol{X}) \leq (n+1)^{JK-1} e^{nE_n(R,D)}. \tag{7}$$

(b) Direct part: If $R > R(D)$ there exists a rate $(R + (\ln 2)/n)$ block code of length n such that $E\rho_n(\boldsymbol{X},\boldsymbol{Y}) \leq nD$ for large n, and at the same time

$$EL(\boldsymbol{X}) \geq (n+1)^{-(JK-1)} e^{nE_n(R,D)}. \tag{8}$$

The theorem is meaningful if $E_n(R,D)$ converges to a limit $E(R,D)$ as will be discussed in the next section.

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

Proof. We begin with the converse part. First, observe that

$$EL(\boldsymbol{X}) = \sum_{i=1}^{M} \Pr\{\rho_n(\boldsymbol{X},\boldsymbol{y}_i) \leq nD\}. \tag{9}$$

Therefore, we shall primarily focus on the behavior of the probability of the event $S(\boldsymbol{y}) \stackrel{\triangle}{=} \{\boldsymbol{x} : \ \rho_n(\boldsymbol{x},\boldsymbol{y}) \leq nD\}$ for a given $\boldsymbol{y} \in \mathcal{Y}^n$.

$$\begin{aligned}
\Pr\{S(\boldsymbol{y})\} &= \sum_{\boldsymbol{x} \in S(\boldsymbol{y})} P(\boldsymbol{x}) \\
&= \sum_{T_{\boldsymbol{x}|\boldsymbol{y}} \subset S(\boldsymbol{y})} |T_{\boldsymbol{x}|\boldsymbol{y}}| \exp[n \sum_{x \in \mathcal{X}} p_{\boldsymbol{x}}(x) \ln p(x)] \\
&\leq \max_{T_{\boldsymbol{x}|\boldsymbol{y}} \subset S(\boldsymbol{y})} (n+1)^{(J-1)K} \exp[nH_{\boldsymbol{x}|\boldsymbol{y}}(X|Y)] \exp[n \sum_{x \in \mathcal{X}} p_{\boldsymbol{x}}(x) \ln p(x)] \\
&= (n+1)^{(J-1)K} \max_{W_{\boldsymbol{x}|\boldsymbol{y}} \in G_n(D;Q_{\boldsymbol{y}})} \exp\left[-n \sum_{y \in \mathcal{Y}} q_{\boldsymbol{y}}(y) \sum_{x \in \mathcal{X}} w_{\boldsymbol{x}|\boldsymbol{y}}(x|y) \ln \frac{w_{\boldsymbol{x}|\boldsymbol{y}}(x|y)}{p(x)}\right] \\
&= (n+1)^{(J-1)K} \exp[-nF_n(D;Q_{\boldsymbol{y}})],
\end{aligned} \tag{10}$$

where we have used the facts [6] that $|T_{\boldsymbol{x}|\boldsymbol{y}}| \leq e^{nH_{\boldsymbol{x}|\boldsymbol{y}}(X|Y)}$ and that the number of distinct conditional types $\{T_{\boldsymbol{x}|\boldsymbol{y}}\}$ associated with pair sequences of length n is upper bounded by $(n+1)^{(J-1)K}$. Therefore, by substituting eq. (10) into eq. (9), we get

$$\begin{aligned}
EL(\boldsymbol{X}) &\leq (n+1)^{(J-1)K} \sum_{i=1}^{e^{nR}} \exp[-nF_n(D; Q_{\boldsymbol{y}_i})] \\
&= (n+1)^{(J-1)K} \sum_{T_{\boldsymbol{y}}} |T_{\boldsymbol{y}} \bigcap \mathcal{C}| \cdot \exp[-nF_n(D; Q_{\boldsymbol{y}})] \\
&\leq (n+1)^{(J-1)K} \sum_{T_{\boldsymbol{y}}} \exp[n(R \wedge H(Q_{\boldsymbol{y}}))] \cdot \exp[-nF_n(D; Q_{\boldsymbol{y}})] \\
&\leq (n+1)^{JK-1} \exp\{n \max_{Q_{\boldsymbol{y}}} [R \wedge H(Q_{\boldsymbol{y}}) - F_n(D; Q_{\boldsymbol{y}})]\} \\
&= (n+1)^{JK-1} \exp\{nE_n(R, D)\}, \qquad (11)
\end{aligned}$$

where $s \wedge t \stackrel{\triangle}{=} \min\{s, t\}$. The second inequality follows from the fact that number of codewords in $T_{\boldsymbol{y}}$ cannot exceed either $M = e^{nR}$ or $|T_{\boldsymbol{y}}| \leq e^{nH(Q_{\boldsymbol{y}})}$ [6], and the third inequality follows from the fact that the number of distinct types of sequences in $\mathcal{Y}^n$ is less than or equal to $(n+1)^{K-1}$. This completes the proof of the converse part.

As for the direct part, let Q^* denote an EPMF on $\mathcal{Y}$ that achieves $E_n(R, D)$. Also, let Q_0 denote the output PMF corresponding to the test channel [5] associated with $R(D)$. Now, there are two possibilities: $H(Q^*) \leq R$ or $H(Q^*) > R$. In each case we will use a slightly different strategy.

1. If $H(Q^*) \leq R$, construct a codebook that includes *all* sequences in the type T^* corresponding to Q^* plus e^{nR} codewords drawn at random according to Q_0.
2. If $H(Q^*) > R$, construct a codebook that consists of e^{nR} *arbitrary* codewords from T^* and e^{nR} random codewords drawn from Q_0.

In both cases the number of codewords is less than or equal to $2 \cdot e^{nR}$ which means that the coding rate is at most $R + (\ln 2)/n$ nats per source symbol. Since in both cases the codebook consists of e^{nR} random codewords governed by Q_0 and since $R > R(D)$ then by the standard random coding argument (see, e.g., [5]) there exists a set of e^{nR} codewords that yields an average per-letter distortion less than or equal to D, provided that n is sufficiently large. In both cases, the number of codewords from T^* is lower bounded by $|T^*| \wedge e^{nR} \geq (n+1)^{-(K-1)} e^{nR \wedge H(Q^*)}$ [6].

Similarly as above, by using the method of types [6] it is easy to show that

$$\Pr\{S(\boldsymbol{y})\} \geq (n+1)^{-(J-1)K} \exp[-nF_n(D; Q_{\boldsymbol{y}})]. \qquad (12)$$

Thus, for the above described code

$$EL(\boldsymbol{X}) = \sum_{i=1}^{e^{nR}} \Pr\{S(\boldsymbol{y}_i)\}$$

$$
\begin{aligned}
&\geq (n+1)^{-K(J-1)} \sum_{\{i: \boldsymbol{y}_i \in T^*\}} \exp[-nF_n(D;Q^*)] \\
&\geq (n+1)^{-(JK-1)} \exp\{n[R \wedge H(Q^*) - F_n(D;Q^*)]\} \\
&= (n+1)^{-(JK-1)} \exp[nE_n(R,D)], \qquad (13)
\end{aligned}
$$

proving the direct part.

4 Properties of the List Size Exponent Function and Examples

In this section we study some properties of the asymptotic exponent of the expected list size. We first have to establish the convergence of $E_n(R,D)$ to a limit $E(R,D)$ as $n \to \infty$.

The asymptotic list size exponent function $E(R,D)$. Let $Q = \{q(y), y \in \mathcal{Y}\}$ be an arbitrary probability mass function (PMF) on the reproduction alphabet, and let $W = \{w(x|y), x \in \mathcal{X}, y \in \mathcal{Y}\}$ denote an arbitrary conditional PMF on $\mathcal{X}$ given y. Now, let

$$
G(D;Q) = \left\{ W: \sum_{y \in \mathcal{Y}} q(y) \sum_{x \in \mathcal{X}} w(x|y)\rho(x,y) \leq D \right\}, \qquad (14)
$$

and for the given source P, let

$$
F(D;Q) \triangleq \inf_{W \in G(D;Q)} U(Q,W), \qquad (15)
$$

where

$$
U(Q,W) = \sum_{y \in \mathcal{Y}} q(y) \sum_{x \in \mathcal{X}} w(x|y) \ln \frac{w(x|y)}{p(x)}. \qquad (16)
$$

Finally, let

$$
\begin{aligned}
E(R,D) &\triangleq \sup_Q [R \wedge H(Q) - F(D;Q)] \\
&= \sup_Q \sup_{W \in G(D;Q)} [R \wedge H(Q) - U(Q,W)], \qquad (17)
\end{aligned}
$$

where $H(Q) = -\sum_{y \in \mathcal{Y}} q(y) \ln q(y)$ and where the first supremum is taken over all PMFs on $\mathcal{Y}$.

Obviously, for $D \geq D_{\min}(Q) \triangleq \sum_{y \in \mathcal{Y}} q(y) \min_{x \in \mathcal{X}} \rho(x,y)$, the set $G(D;Q)$ is nonempty as it contains at least the conditional PMF $w(x|y)$ that puts all its mass on an $x \in \mathcal{X}$ that minimizes $\rho(x,y)$ for a given y. Since the function $(R \wedge H(Q) - U(Q,W))$ is bounded and continuous in both Q and W and since the set $\{(Q,W) : W \in G(D;Q)\}$ is closed, the above supremum is actually a maximum. Let (Q^*, W^*) attain $E(R,D)$. Clearly, $E_n(R,D) \leq E(R,D)$ for all n as the latter is defined as a supremum over the continuum of pairs (Q,W) rather than rational PMFs with denominator n only. On the other hand, since for $D \geq$

$\inf_Q D_{\min}(Q) \stackrel{\triangle}{=} D_{\min} = \min_{x\in\mathcal{X}} \min_{y\in\mathcal{Y}} \rho(x,y)$, the double supremum in the second line of eq. (17) it taken over a set of pairs (Q,W) that is equal to the closure of its interior [7, pp. 293-294] and since the set of pairs of rational EPMFs $(Q_{\boldsymbol{y}}, W_{\boldsymbol{xy}})$ with denominator n, $n = 1, 2, ...$, is dense in the set of all pairs (Q,W), we can find a sequence of such rational pairs (Q_n, P_n) in $\{(Q,W) : W \in G(D;Q)\}$ such that

$$R \wedge H(Q_n) - U(Q_n, W_n) \to R \wedge H(Q^*) - U(Q^*, W^*) = E(R,D), \tag{18}$$

due to continuity. Thus, we have established the fact that $\lim_{n\to\infty} E_n(R,D) = E(R,D)$ for $D \geq D_{\min}$. For $D < D_{\min}$, $E_n(R,D) = E(R,D) = -\infty$. However, since $D_{\min}$ can be assumed zero without loss of generality, this range in not interesting.

A Parametric Representation to $E(R,D)$. We next show that $E(R,D)$ can be represented as

$$E(R,D) = \max_Q \max_{\theta\geq 0} [H(Q) \wedge R + \sum_{y\in\mathcal{Y}} q(y)\Lambda_\theta(y) + \theta D], \tag{19}$$

where

$$\Lambda_\theta(y) = \ln \sum_{x\in\mathcal{X}} p(x) e^{-\theta\rho(x,y)}. \tag{20}$$

First, observe that $F(D;Q)$ is a monotone decreasing convex function of D for fixed Q. The monotonicity is obvious as $D_1 > D_2$ implies $G(D_1;Q) \supset G(D_2;Q)$. As for convexity, let W_1 achieve $F(D_1;Q)$ and let W_2 achieve $F(D_2;Q)$. For $0 \leq \alpha \leq 1$, obviously $\alpha W_1 + (1-\alpha)W_2 \in G(\alpha D_1 + (1-\alpha)D_2; Q)$, and so

$$\begin{aligned} F(\alpha D_1 + (1-\alpha)D_2; Q) &\leq U(Q, \alpha W_1 + (1-\alpha)W_2) \\ &\leq \alpha U(Q, W_1) + (1-\alpha)U(Q, W_2) \\ &= \alpha F(D_1; Q) + (1-\alpha)F(D_2; Q). \end{aligned} \tag{21}$$

where the second ineqwuality follows from the fact $U(Q,W)$ is convex in W for fixed Q. Thus, $F(D;Q)$ is *strictly* decreasing in the range

$$D_{\min}(Q) < D < D_{\max}(Q) \stackrel{\triangle}{=} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)q(y)\rho(x,y), \tag{22}$$

and hence the minimizing conditional PMF W^* must be such that the inequality constraint that defines the set $G(D;Q)$ is met with equality. Therefore, $F(D;Q)$ can be calculated by minimizing the Lagrange function

$$L(Q) = \sum_{y\in\mathcal{Y}} q(y) \sum_{x\in\mathcal{X}} w(x|y) \left[\ln \frac{w(x|y)}{p(x)} + \theta\rho(x,y)\right], \tag{23}$$

which leads to

$$w^*(x|y) = \frac{p(x)e^{-\theta\rho(x,y)}}{\sum_{x'\in\mathcal{X}} p(x')e^{-\theta\rho(x',y)}}, \tag{24}$$

and θ is chosen such that

$$\sum_{y\in\mathcal{Y}} q(y) \sum_{x\in\mathcal{X}} w^*(x|y)\rho(x,y) = D, \tag{25}$$

or, equivalently,

$$\sum_{y\in\mathcal{Y}} q(y)\dot{\Lambda}_\theta(y) = -D, \tag{26}$$

where $\dot{\Lambda}_\theta(y)$ is the derivative of $\Lambda_\theta(y)$ w.r.t θ. This equation has always a unique solution $\theta = \theta_0 \geq 0$ because of the following consideration. First, observe that the second derivative of $\Lambda_\theta(y)$ w.r.t θ is positive as it can be interpreted as the variance of $\rho(X, y)$ w.r.t $w^*(x|y)$. Therefore, the first derivative $\dot{\Lambda}_\theta(y)$ is a continuous monotone increasing function of θ for fixed y. Next, it is easy to check that

$$\lim_{\theta\to 0} \sum_{y\in\mathcal{Y}} q(y)\dot{\Lambda}_\theta(y) = -D_{\max}(Q) \tag{27}$$

and

$$\lim_{\theta\to\infty} \sum_{y\in\mathcal{Y}} q(y)\dot{\Lambda}_\theta(y) = -D_{\min}(Q). \tag{28}$$

Hence by continuity there must exist a unique solution θ_0 to eq. (26). By substituting eq. (24) into eq. (16), we get

$$\begin{aligned} F(D;Q) &= \sum_{y\in\mathcal{Y}} q(y)[\theta_0\dot{\Lambda}_{\theta_0}(y) - \Lambda_{\theta_0}(y)] \\ &= -\theta_0 D - \sum_{y\in\mathcal{Y}} q(y)\Lambda_{\theta_0}(y). \end{aligned} \tag{29}$$

Note that $(\theta D + \sum_y q(y)\Lambda_\theta(y))$ is a convex function of θ as it has a positive second order derivative, and the first derivative vanishes at $\theta = \theta_0$. Therefore, θ_0 minimizes this function, and so $F(D;Q)$ can also be expressed as

$$F(D;Q) = -\min_{\theta\geq 0}[\theta D + \sum_{y\in\mathcal{Y}} q(y)\Lambda_\theta(y)], \tag{30}$$

which in turn leads to eq. (19).

Concavity. The function $E(R, D)$ is concave in the pair (R, D). In other words, for every two points in the $R - D$ plane, (R_1, D_1) and (R_2, D_2) and for every $0 \leq \alpha \leq 1$,

$$E(\alpha R_1 + (1-\alpha)R_2, \alpha D_1 + (1-\alpha)D_2) \geq \alpha E(R_1, D_1) + (1-\alpha)E(R_2, D_2). \tag{31}$$

This can be shown by a standard time-sharing argument. Let $\mathcal{C}_n^*$ be a rate R_1 codebook of length n and let $\mathcal{C}_m^*$ be a rate R_2 codebook of length m, both optimal in the sense of maximizing the expected list size for distortion levels D_1 and D_2, respectively. Consider now the block-code of length $(n + m)$ formed by applying $\mathcal{C}_n^*$ to the first n components and $\mathcal{C}_m^*$ to the last m components of $(x_1, ..., x_{n+m})$. The rate of this code is $\alpha R_1 + (1-\alpha)R_2$, where $\alpha = n/(n + m)$. The size of the list for $(x_1, ..., x_{n+m})$ at $D = \alpha D_1 + (1-\alpha)D_2$ is at least as large as the product of the size of the list for $(x_1, ..., x_n)$ at D_1 and the size of the list for $(x_{n+1}, ..., x_{n+m})$ at D_2. This is true because encoding the first m components at distortion nD_1 and the last m components at distortion mD_2, guarantees an overall per-letter distortion $\alpha D_1 + (1-\alpha)D_2$. Thus, for a list-optimal codebook of rate $\alpha R_1 + (1-\alpha)R_2$, distortion $\alpha D_1 + (1-\alpha)D_2$, and length $n+m$,

$$\begin{aligned} EL(X_1, ...X_{n+m}) &\geq E[L(X_1, ..., X_n)L(X_{n+1}, ..., X_{n+m})] \\ &= EL(X_1, ..., X_n) \cdot EL(X_{n+1}, ..., X_{n+m}) \end{aligned} \tag{32}$$

where the second line follows from the memorylessness of source. Since the left-most side is of exponential order

$$\exp[(n+m)E(\alpha R_1 + (1-\alpha)R_2, \alpha D_1 + (1-\alpha)D_2)]$$

and the right-most side is of exponential order

$$\exp[nE(R_1, D_1)] \cdot \exp[mE(R_2, D_2)],$$

concavity is established.

Simple bounds to $E(R, D)$. The function $E(R, D)$ satisfies following inequalities.

$$H(Q_0) \wedge R - R(D) \leq E(R, D) \leq R - F(D), \tag{33}$$

where Q_0 is the output PMF induced by the test channel corresponding to $R(D)$ and $F(D) \stackrel{\triangle}{=} \min_Q F(D; Q)$. Both bounds are tight in the sense that there are situations in which they hold with equality. These inequalities imply that the rate $R_0(D)$ at which $E(R, D)$ crosses zero for a given D, is bounded by

$$F(D) \leq R_0(D) \leq R(D). \tag{34}$$

The lower bound in (33) follows from the following consideration. First, observe that $F(D; Q_0) \leq R(D)$ because the backward test channel corresponding to $R(D)$ is a member of $G(D; Q_0)$. Therefore,

$$\begin{aligned} E(R, D) &= \max_Q [H(Q) \wedge R - F(D; Q)] \\ &\geq H(Q_0) \wedge R - F(D; Q_0) \\ &\geq H(Q_0) \wedge R - R(D). \end{aligned} \tag{35}$$

Note that $R(D) < H(Q_0)$ thus, in the range $R(D) < R < H(Q_0)$ the lower bound is given by $R - R(D)$, and for $R \geq H(Q_0)$ the lower bound is $H(Q_0) - R(D)$ independently of R.

As for the upper bound,

$$\begin{aligned} E(R, D) &= \max_Q [H(Q) \wedge R - F(D; Q)] \\ &\leq \max_Q [R - F(D; Q)] \\ &= R - \min_Q F(D; Q) \\ &= R - F(D). \end{aligned} \tag{36}$$

To demonstrate tightness in the above sense, consider the binary symmetric source (BSS) with the Hamming distortion measure. Let $h(\alpha)$, $0 \le \alpha \le 1$, denote the binary entropy function. In the BSS case, it is readily seen that for $D \le 1/2$,

$$\begin{aligned} F(D;Q) &= 1 - \sum_{y=0}^{1} q(y) h(w^*(1-y|y)) \\ &\ge 1 - h(\sum_{y=0}^{1} q(y) w^*(1-y|y)) \\ &\ge 1 - h(D) = R(D). \end{aligned} \tag{37}$$

By setting $w^*(1-y|y) = D$, $y = 0, 1$, both inequalities become equalities, and therefore

$$F(D;Q) = R(D). \tag{38}$$

Note that here $F(D;Q)$ is independent of Q. Thus,

$$E(R,D) = \max_{Q}[H(Q) \wedge R - R(D)] = R - R(D) \tag{39}$$

which coincides with both the upper bound, because $F(D) = R(D)$, and the lower bound because $Q_0 = (1/2, 1/2)$ and then $H(Q_0) = \ln 2 \ge R$.

$E(R,D)$ is constant for large R and fixed D. Let $\hat{Q}$ be the PMF that maximizes $H(Q) - F(D;Q)$. Then, for every $R \ge H(\hat{Q})$ and every output PMF Q, we have

$$\begin{aligned} H(\hat{Q}) \wedge R - F(D;\hat{Q}) &= H(\hat{Q}) - F(D;\hat{Q}) \\ &\ge H(Q) - F(D;Q) \\ &\ge H(Q) \wedge R - F(D;Q), \end{aligned} \tag{40}$$

which means that $\hat{Q}$ also maximizes $H(Q) \wedge R - F(D;Q)$ and hence $E(R,D) = H(\hat{Q}) - F(D;\hat{Q})$ independently of R. In other words, there is no point to increase R beyond $H(\hat{Q})$ because the list size exponent $E(R,D)$ will not improve any further.

We next provide a numerical example. Fig. 1 depicts $E(R, D = 0.2)$ as a function of R for a Bernoulli source with letter probabilities $p(0) = 1 - p(1) = 1/3$ and the Hamming distortion measure. As can be seen, $E(R,D)$ is a monotone increasing, concave function of R, starting from negative values at low rates, where the probability that $L(\boldsymbol{X}) = 0$ is very high, and ending with a plateau as anticipated in the previous paragraph. To demonstrate the numerical procedure for calculating $E(R,D)$, we use the parametric representation provided in eq. (19). In this case,

$$\Lambda_\theta(0) = \ln \frac{1 + 2e^{-\theta}}{3} \tag{41}$$

and

$$\Lambda_\theta(1) = \ln \frac{2 + e^{-\theta}}{3}. \tag{42}$$

Following eq. (26), for a given output PMF parameterized by $q = q(0) = 1 - q(1)$, the value of θ_0 is determined by solving the equation

$$q \cdot \frac{2s}{2s+1} + (1-q) \cdot \frac{s}{s+2} = D, \tag{43}$$

where $s = e^{-\theta}$. The solution is

$$s = \frac{5D - 3q - 1 + \sqrt{(5D - 3q - 1)^2 + 16D(1-D)}}{4(1-D)}. \tag{44}$$

For $D = 0.2$, this amounts to

$$s = \frac{\sqrt{9q^2 + 2.56} - 3q}{3.2}. \tag{45}$$

Consequently, according to eq. (29),

$$F(0.2, q) = 0.2 \ln s - q \ln \frac{2s+1}{3} - (1-q) \ln \frac{s+2}{3} \tag{46}$$

For large R, it is found numerically that

$$E(R, 0.2) = \max_{0 \le q \le 1} [h(q) - F(0.2, q)] = 0.5004. \tag{47}$$

The maximizing PMF is given by $q^* = 0.4$ and hence the last equation is valid for every $R \ge h(0.4) = 0.673$ nats/symbol, which is close to the maximum possible rate in a binary alphabet, i.e., $R = \ln 2 = 0.693$ nats/symbol. According to the construction in the proof of the direct part of Theorem 1, the codebook contains all sequences from T^*, i.e., sequences with 40% zeroes and 60% ones, as well as codewords of the best composition Q_0 corresponding the test channel of $R(D)$, i.e., sequences with 22.2% zeroes and 77.8% ones.

5 Conclusion and Future Research

We have characterized and studied properties of the largest attainable expected list size exponent function in lossy source coding. Possible directions for extending these results are the following.

- *General moments.* Analogously to reported work on list channel decoding (e.g., [3], [4], and others), it would be interesting to obtain single-letter characterizations of $E\{L(\boldsymbol{X})^\rho\}$ for any $\rho > 0$.
- *Large deviations performance.* Another performance measure that would be interesting to investigate is the asymptotic behavior of the probability that $L(\boldsymbol{X}) < e^{n(E(R,D)-\Delta)}$ for a given $\Delta > 0$.
- *Continuous alphabet sources.* Since rate-distortion theory is primarily aimed at efficient representations of continuous alphabet sources, this extension appears natural.

- *Universal list-optimal codes.* Can one attain $E(R, D)$ without knowing the source P in advance? If yes, what would be the minimum attainable cost of universality?

- *Side information.* How can side information at the decoder best help in increasing the expected list size exponent?

These topics are currently under investigation.

7 References

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

[2] K. Marton, "Error exponent for source coding with a fidelity criterion," *IEEE Trans. Inform. Theory*, vol. IT-20, no. 2, pp. 197-199, March 1974.

[3] R. Ahlswede, N. Cai, and Z. Zhang,"Erasure, list and detection zero-error capacities for low noise and relation to identification," *Proc. IEEE Int. Symp. on Inform. Theory*, 1994.

[4] I. E. Telatar, "Zero-error list capacities of discrete memoryless channels," *Proc. IEEE Int. Symp. on Inform. Theory*, 1995.

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

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

[7] T. M. Cover and J. A. Thomas, *Elements of Information Theory.* J. Wiley & Sons, 1991.

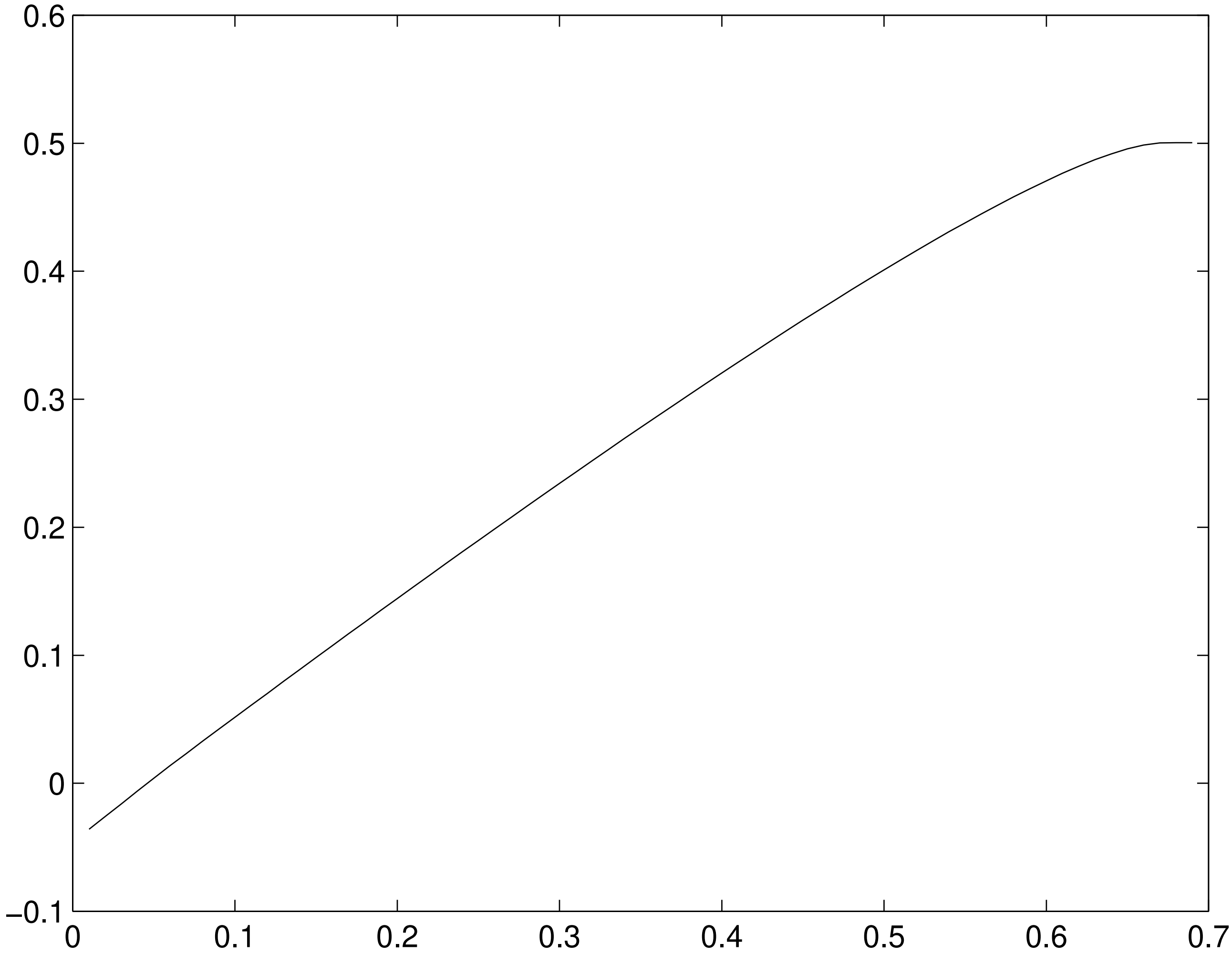

Figure 1: $E(R, 0.2)$ vs. R for $p(0) = 1 - p(1) = 1/3$ and the Hamming distortion measure.