On the Rate-Distortion Function of Random Vectors and Stationary Sources with Mixed Distributions *

András György Tamás Linder Kenneth Zeger

To appear in: *IEEE Trans. on Information Theory*
Submitted: June 4, 1998
Final Version: March 26, 1999

Abstract

The asymptotic (small distortion) behavior of the rate-distortion function of an n-dimensional source vector with mixed distribution is derived. The source distribution is a finite mixture of components such that under each component distribution a certain subset of the coordinates have a discrete distribution while the remaining coordinates have a joint density. The expected number of coordinates with a joint density is shown to equal the rate-distortion dimension of the source vector. Also, the exact small distortion asymptotic behavior of the rate-distortion function of a special but interesting class of stationary information sources is determined.

Index Terms: source coding, rate distortion theory, quantization.

*A. György is with the Faculty of Electrical Engineering and Informatics, Technical University of Budapest, H-1521 Budapest, Hungary (email: `gya@szit.bme.hu`). T. Linder is with the Department of Mathematics and Statistics, Queen's University, Kingston, Ontario, Canada K7L 3N6 (email: `linder@mast.queensu.ca`). K. Zeger is with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla CA 92093-0407 (email: `zeger@ucsd.edu`). The research was supported in part by the National Science Foundation.

1 Introduction

Consider a random vector $X^n = (X_1, \dots, X_n)$ taking values in the n-dimensional Euclidean space $\mathbb{R}^n$. The *rate-distortion function* [1] of X^n relative to the normalized *squared error* (expected squared Euclidean distance) criterion is defined for all $D > 0$ by

$$R_{X^n}(D) = \inf_{n^{-1}\mathbf{E}\|X^n - Y^n\|^2 \le D} \frac{1}{n} I(X^n; Y^n)$$

where the infimum of the normalized mutual information $\frac{1}{n}I(X^n;Y^n)$ (computed in bits) is taken over all joint distributions of X^n and $Y^n = (Y_1, \dots, Y_n)$ such that

$$\frac{1}{n}\mathbf{E}\|X^n - Y^n\|^2 = \frac{1}{n}\sum_{i=1}^{n} \mathbf{E}[(X_i - Y_i)^2] \le D.$$

Except for a few special cases, closed form analytic expressions for $R_{X^n}(D)$ are not known, and only upper and lower bounds are available. Arguably, the most important of these bounds is the well known *Shannon lower bound* [1]. For X^n having an absolutely continuous distribution with density f and a finite *differential entropy*

$$h(X^n) = -\int f(x)\log f(x)dx$$

the Shannon lower bound states that

$$R_{X^n}(D) \ge \frac{1}{n}h(X^n) - \frac{1}{2}\log(2\pi e D)$$

where the logarithm is base 2. The right hand side equals $R_{X^n}(D)$ if and only if X^n can be written as a sum of two independent random vectors, one of which has independent and identically distributed (i.i.d.) Gaussian components with zero mean and variance D. In more general cases, the Shannon lower bound is strictly less than $R_{X^n}(D)$ for all $D > 0$, but it becomes tight in the limit of small distortions in the sense that

$$R_{X^n}(D) = \frac{1}{n}h(X^n) - \frac{1}{2}\log(2\pi e D) + o(1) \tag{1}$$

where $o(1) \to 0$ as $D \to 0$ ([2] [3] [4]).

One important feature of the Shannon lower bound is that it easily generalizes to stationary sources. Let $\mathcal{X} = \{X_i\}_{i=1}^{\infty}$ be a real stationary source and for each n, let X^n denote the vector of the first n samples of $\mathcal{X}$. The rate-distortion function of $\mathcal{X}$ is defined by

$$R_{\mathcal{X}}(D) = \lim_{n\to\infty} R_{X^n}(D) \tag{2}$$

(the limit is known to always exist [1]). The quantity $R_{\mathcal{X}}(D)$ represents the minimum achievable rate in lossy coding $\mathcal{X}$ with distortion D (see, e.g., [5]). Let $X^n = (X_1, \dots, X_n)$ have a density and finite differential entropy $h(X^n)$ for all n, and assume that the differential entropy rate $h(\mathcal{X}) = \lim_{n\to\infty} \frac{1}{n} h(X^n)$ is finite. Then the *generalized Shannon lower bound* [1] is

$$R_{\mathcal{X}}(D) \geq h(\mathcal{X}) - \frac{1}{2}\log(2\pi e D) \tag{3}$$

and just as in the finite dimensional case, this lower bound becomes asymptotically tight in the limit of small distortions ([3] [4]).

For source distributions without a density the Shannon lower bound has no immediate extension. However, Rosenthal and Binia [6] have demonstrated that the asymptotic behavior of the rate-distortion function (which for sources with a density is given by (1)) can still be determined for more general distributions. They considered the case when the distribution of X^n is a mixture of a discrete and a continuous component with nonnegative weights $1-\alpha$ and α, respectively, where the continuous component is concentrated on an L-dimensional linear subspace of $\mathbb{R}^n$ and has a density with respect to the Lebesgue measure on that subspace. Equivalently, we are given an n-dimensional random vector $X^{(1)}$ with a discrete distribution, and another n-dimensional random vector $X^{(2)}$ which is obtained by applying an orthogonal transformation to $X' = (X'_1, \dots, X'_L, 0, \dots, 0)$, where the L-dimensional random vector $(X'_1, \dots, X'_L)$ has a density. Let ν be a binary random variable with distribution $\mathbf{P}(\nu = 0) = 1-\alpha$ and $\mathbf{P}(\nu = 1) = \alpha$, and let ν be independent of $(X^{(1)}, X^{(2)})$. It is assumed that X^n can be written in the form

$$X^n = (1-\nu)X^{(1)} + \nu X^{(2)}. \tag{4}$$

The main result of [6] shows that as $D \to 0$, the rate-distortion function of X^n with such a mixed distribution is given asymptotically by the expression

$$R_{X^n}(D) = \frac{1}{n}H(\nu) + \frac{1-\alpha}{n}H(X^{(1)}) + \frac{\alpha}{n}h(X') - \frac{\alpha L}{2n}\log\left(\frac{2\pi e n D}{\alpha L}\right) + o(1) \tag{5}$$

where $H(\nu)$ and $H(X^{(1)})$ denote discrete entropies and $h(X')$ is the differential entropy of X'. We note here that Rosenthal and Binia made an error in the derivation (see equation (27) in [6]) and in fact arrived at an incorrect formula instead of the correct expression (5). Their asymptotic expression exceeds (5) by the nonnegative constant $\frac{\alpha L}{2n}\log\left(\frac{1}{\alpha}\right)$.

Although the mixture model Rosenthal and Binia considered can be very useful for modeling memoryless signals encountered in certain practical situations, its use in modeling information sources with memory and mixed marginals is rather limited. In particular, it is easy to see that a source $\{X_i\}_{i=1}^{\infty}$ cannot be ergodic if for all n, the samples $X^n = (X_1, \dots, X_n)$ have a mixture distribution in the form of (4) with $0 < \alpha < 1$. Thus in general (5) cannot be used to obtain the asymptotic behavior of $R_{\mathcal{X}}(D)$ for stationary and ergodic sources with memory and mixed marginals, although such source models are of practical interest, for example, in lossy coding of sparse images [7].

In this correspondence we propose a more general mixture model and provide an extension of (5) to this class of source distributions. Our model has the advantage of allowing stationary and ergodic information sources. We assume that the distribution of X^n is a mixture of finitely many component distributions such that each component has a certain number of coordinates with a discrete distribution while the remaining coordinates have a joint density. More formally, let $\{X^{(j)}, j = 1, \dots, N\}$ be a finite collection of random n-vectors such that for each j exactly d_j coordinates of $X^{(j)}$ have a discrete distribution (the d_j-dimensional vector formed by these "discrete coordinates" is denoted $\widehat{X}^{(j)}$) and the remaining $c_j = n - d_j$ coordinates have a joint density (the c_j-dimensional vector formed by these "continuous coordinates" is denoted $\tilde{X}^{(j)}$). As explained in the next section, we can assume without loss of generality that $X^{(j)}$ and $X^{(j')}$ do not have all their discrete coordinates in the same positions if $j \neq j'$. Let V be a random variable taking values in $\{1, \dots, N\}$ which is independent of the $X^{(j)}$. Our model for X^n assumes that $X^n = X^{(V)}$, that is, if $V = j$, then $X^n = X^{(j)}$. Note that V is a function of X^n with probability 1.

Let $h(\tilde{X}^{(j)}|\widehat{X}^{(j)})$ denote the conditional differential entropy of the continuous coordinates of $X^{(j)}$ given its discrete coordinates, and let $H(\widehat{X}^{(j)})$ denote the entropy of the discrete coordinates of $X^{(j)}$. Our main result, Theorem 1, shows that as $D \to 0$,

$$R_{X^n}(D) = \frac{1}{n}H(V) + \frac{1}{n}\sum_{j=1}^{N}\alpha_j H(\widehat{X}^{(j)}) + \frac{1}{n}\sum_{j=1}^{N}\alpha_j h(\tilde{X}^{(j)}|\widehat{X}^{(j)}) - \frac{\Lambda}{2}\log(2\pi e D/\Lambda) + o(1) \tag{6}$$

where $\alpha_j = \mathbf{P}(V = j)$ and $\Lambda = \frac{1}{n}\sum_{j=1}^{N}\alpha_j c_j$. Note that the quantity $n\Lambda$ is the average number of "continuous coordinates" of X^n. Formula (6) proves that $n\Lambda$ is also the so-called rate-distortion dimension of X^n [8].

To illustrate the application of this result to sources with memory, let $\mathcal{Z} = \{Z_i\}_{i=1}^{\infty}$ be an arbitrary binary stationary source. We construct another stationary source $\mathcal{X} = \{X_i\}_{i=1}^{\infty}$ in the following manner. If $Z_i = 0$, let X_i have a fixed discrete distribution P, while if $Z_i = 1$, let X_i have a density f. We assume that the generating procedure is memoryless so that the X_i are conditionally independent given $\{Z_i\}_{i=1}^{\infty}$. Then the process $\{X_i\}_{i=1}^{\infty}$ is stationary. Note that the distribution of X^n does not have the binary mixture form of (4) if $n \geq 2$. Thus (5) cannot be used to obtain the asymptotic behavior of $R_{X^n}(D)$ for $n \geq 2$ except when $\{Z_i\}$ is memoryless, in which case $R_{X^n}(D) = R_{X_1}(D)$. On the other hand, for all n, the distribution of X^n has a mixture form for which (6) applies. As a consequence of this fact, Corollary 1 shows that as $D \to 0$,

$$R_{\mathcal{X}}(D) = H(\mathcal{Z}) + (1-\alpha)H(P) + \alpha\, h(f) - \frac{\alpha}{2}\log(2\pi e D/\alpha) + o(1) \tag{7}$$

where $H(\mathcal{Z}) = \lim_n \frac{1}{n}H(Z^n)$ is the entropy rate of $\mathcal{Z}$, $H(P)$ and $h(f)$ are the discrete and differential entropies of P and f, respectively, and $\alpha = \mathbf{P}(Z_i = 1)$.

The above construction can be used to model the formation of sparse images which have a large number of zero-valued pixels [7]. In this case, P is concentrated on the single value zero (i.e., $X_i = 0$ if $Z = 0$) and the fraction of nonzero pixels is controlled by the parameter $\alpha = \mathbf{P}(Z_i = 1)$. The wide range of possible choices for the stationary binary process $\{Z_i\}$ and the density f makes it possible to accurately model the image characteristics. Then formula (7) can be used to compare the performance of a practical coding scheme with the ideal performance given by rate-distortion function.

2 Sources with Mixed Distribution

Let $\{X^{(j)} = (X_1^{(j)}, \dots, X_n^{(j)}),\ j = 1, \dots, N\}$ be a finite collection of $\mathbb{R}^n$-valued random vectors such that each $X^{(j)}$ has d_j coordinates which have discrete distribution, and $c_j = n - d_j$ coordinates which have a joint density. More formally, let $A^j = \{a_1^j, \dots, a_{d_j}^j\}$ be a subset of $\{1, \dots, n\}$ of size d_j such that $a_1^j < a_2^j < \cdots < a_{d_j}^j$, and let $B^j = \{b_1^j, \dots, b_{c_j}^j\} = \{1, \dots, n\} \setminus A^j$, $b_1^j < b_2^j < \cdots < b_{c_j}^j$ be the complement of A^j in $\{1, \dots, n\}$. We assume that the d_j-dimensional random vector

$$\widehat{X}^{(j)} = (X_{a_1^j}^{(j)}, \dots, X_{a_{d_j}^j}^{(j)}) \tag{8}$$

which is chosen from among the coordinates of $X^{(j)}$ by the index set A^j, has a discrete distribution with a finite or countably infinite number of atoms, while the c_j dimensional random vector

$$\tilde{X}^{(j)} = (X_{b_1^j}^{(j)}, \dots, X_{b_{c_j}^j}^{(j)}) \tag{9}$$

has an absolutely continuous distribution with a density. We also allow $d_j = n$ ($X^{(j)}$ has a discrete distribution) and $d_j = 0$ ($X^{(j)}$ has an n-dimensional density).

Let the source vector X^n have a distribution which is a mixture of the distributions of the $X^{(j)}$ with nonnegative weights $\alpha_1, \dots, \alpha_N$ ($\sum_{j=1}^N \alpha_j = 1$). This means that for any measurable $B \subset \mathbb{R}^n$,

$$\mathbf{P}(X^n \in B) = \sum_{j=1}^{N} \alpha_j \mathbf{P}(X^{(j)} \in B). \tag{10}$$

Equivalently, we can define an index random variable V taking values in $\{1, \ldots, N\}$, which is independent of the $X^{(j)}$ and has the distribution $\mathbf{P}(V = j) = \alpha_j$, $j = 1 \ldots, N$. If X^n is defined by

$$X^n = X^{(V)} \tag{11}$$

(i.e., if $V = j$, then $X^n = X^{(j)}$) then X^n has a distribution given by (10).

Without loss of generality we will assume that if $j \neq j'$, then $X^{(j)}$ and $X^{(j')}$ do not have their discrete (and consequently their continuous) coordinates at the same positions, i.e.,

$$A^j \neq A^{j'} \text{ if } j \neq j'. \tag{12}$$

For otherwise, by mixing the distributions of $X^{(j)}$ and $X^{(j')}$ with weights $\alpha_j/(\alpha_j + \alpha_{j'})$ and $\alpha_{j'}/(\alpha_j + \alpha_{j'})$, one would obtain a new distribution which, when assigned the weight $\alpha_j + \alpha_{j'}$, could replace $X^{(j)}$ and $X^{(j')}$ in the definition of X^n. Therefore, we can assume that $N \leq 2^n$ since there are 2^n different possibilities for choosing discrete coordinates.

In what follows we require that X^n satisfy the following mild conditions.

(a) All $X^{(j)}$ have finite second moments $\mathbf{E}\|X^{(j)}\|^2 < \infty$, $j = 1, \ldots, N$.

(b) For each $X^{(j)}$, $j = 1, \ldots, N$, the conditional differential entropy $h(\tilde{X}^{(j)}|\widehat{X}^{(j)})$ is finite, and the entropy of the discrete coordinates $H(\widehat{X}^{(j)})$ is finite.

The next theorem is proved in Section 3.

Theorem 1 *Assume X^n is of the mixture form (11) such that each component $X^{(j)}$ has d_j coordinates with a discrete distribution and $c_j = n - d_j$ coordinates with a joint density. Suppose the $X^{(j)}$ satisfy* **(a)** *and* **(b)**. *Then the asymptotic behavior of the rate-distortion function of X^n relative to the normalized squared error is given as $D \to 0$ by*

$$R_{X^n}(D) = \frac{1}{n}H(V) + \frac{1}{n}\sum_{j=1}^{N} \alpha_j H(\widehat{X}^{(j)}) + \frac{1}{n}\sum_{j=1}^{N} \alpha_j h(\tilde{X}^{(j)}|\widehat{X}^{(j)}) - \frac{\Lambda}{2}\log(2\pi e D/\Lambda) + o(1) \tag{13}$$

where $\Lambda = \frac{1}{n}\sum_{j=1}^{N} \alpha_j c_j$ and $o(1) \to 0$ as $D \to 0$.

Remark Kawabata and Dembo [8] defined the *rate-distortion dimension* of X^n by

$$\lim_{D\to 0}\frac{nR_{X^n}(D)}{-\frac{1}{2}\log(D)}$$

provided the limit exists. The rate-distortion dimension of X^n with an n-dimensional density is n by (1). It is easy to see that if X^n has a discrete distribution, its rate-distortion dimension is zero. The result of Rosenthal and Binia in (5) demonstrates that if the continuous component of X^n has an L-dimensional density and weight α, then its rate-distortion dimension is αL. Theorem 1 shows that for the mixed distributions we consider, the rate-distortion dimension is

$$\lim_{D\to 0}\frac{nR_{X^n}(D)}{-\frac{1}{2}\log(D)} = n\Lambda$$

where $n\Lambda = \sum_{j=1}^{N}\alpha_j c_j$. Thus the expected number of the continuous coordinates of X^n is also the effective dimension of X^n in the rate-distortion sense.

Example One immediate application of Theorem 1 concerns processes which are obtained by passing a binary stationary source through a memoryless channel. Let $\mathcal{Z} = \{Z_i\}_{i=1}^{\infty}$ be an arbitrary stationary source taking values in $\{0,1\}$, and consider a time-invariant memoryless channel with binary input and real valued output. The output of the channel has a discrete distribution P if the input is 0, and an absolutely continuous distribution with density f if the input is 1. We will assume that f and P have finite variance and that $H(P)$ and $h(f)$ are finite.

Suppose the stationary process $\mathcal{X} = \{X_i\}_{i=1}^{\infty}$ is generated as the output of this channel if the input is $\{Z_i\}_{i=1}^{\infty}$. Fix $n \geq 1$. Since the channel is memoryless, $X_1,\ldots,X_n$ are conditionally independent given Z^n. For $z^n \in \{0,1\}^n$, let $X^{(z^n)}$ be a random n-vector having distribution equal to the conditional distribution of X^n given $Z^n = z^n$, and let $d(z^n)$ and $c(z^n)$ denote the number of 0's and 1's, respectively, in the binary string z^n. Then the coordinates $X_i^{(z^n)}$ for which $z_i = 0$, form a $d(z^n)$-dimensional i.i.d. random vector $\widehat{X}^{(z^n)}$ with a discrete marginal distribution P, and the $X_i^{(z^n)}$ for which $z_i = 1$, form a $c(z^n)$-dimensional i.i.d. random vector $\tilde{X}^{(z^n)}$ with marginal density f. It follows that X^n has the type of mixture distribution considered in Theorem 1 with 2^n components $X^{(z^n)}$ indexed by z^n, where $X^{(z^n)}$ has weight $\mathbf{P}(Z^n = z^n)$. Therefore, we can apply Theorem 1 with $V = Z^n$ and $\alpha(z^n) = \mathbf{P}(Z^n = z^n)$ to obtain that as $D \to 0$,

$$\begin{aligned} R_{X^n}(D) &= \frac{1}{n}H(Z^n) + \frac{1}{n}\sum_{z^n\in\{0,1\}^n}\mathbf{P}(Z^n = z^n)H(\widehat{X}^{(z^n)}) \\ &\quad + \frac{1}{n}\sum_{z^n\in\{0,1\}^n}\mathbf{P}(Z^n = z^n)h(\tilde{X}^{(z^n)}|\widehat{X}^{(z^n)}) - \frac{\alpha}{2}\log(2\pi e D/\alpha) + o(1) \qquad (14) \end{aligned}$$

where

$$\alpha = \frac{1}{n} \sum_{z^n \in \{0,1\}^n} \mathbf{P}(Z^n = z^n) c(z^n) = \frac{1}{n} \mathbf{E}[c(Z^n)] = \mathbf{P}(Z_i = 1)$$

since $\{Z_i\}$ is stationary. Moreover, by independence, we have $H(\widehat{X}^{(z^n)}) = d(z^n)H(P)$ and $h(\tilde{X}^{(z^n)}|\widehat{X}^{(z^n)}) = c(z^n)h(f)$. Since we also have $\frac{1}{n}\sum_{z^n \in \{0,1\}^n} \mathbf{P}(Z^n = z^n)d(z^n) = 1 - \alpha$, (14) can be simplified to

$$R_{X^n}(D) = \frac{1}{n}H(Z^n) + (1-\alpha)H(P) + \alpha\, h(f) - \frac{\alpha}{2}\log(2\pi e D/\alpha) + o(1). \tag{15}$$

From this the following corollary of Theorem 1 is almost immediate.

Corollary 1 *Let $\mathcal{X} = \{X\}_{i=1}^n$ be the stationary process of the previous example and let $H(\mathcal{Z}) = \lim_n \frac{1}{n}H(Z^n)$ be the entropy rate of the generating binary stationary source $\mathcal{Z} = \{Z_i\}_{i=1}^{\infty}$. Then as $D \to 0$,*

$$R_{\mathcal{X}}(D) = H(\mathcal{Z}) + (1-\alpha)H(P) + \alpha\, h(f) - \frac{\alpha}{2}\log(2\pi e D/\alpha) + o(1).$$

Proof. Using more precise notation, (15) can be rewritten as

$$R_{X^n}(D) = \frac{1}{n}H(Z^n) + (1-\alpha)H(P) + \alpha\, h(f) - \frac{\alpha}{2}\log(2\pi e D/\alpha) + \epsilon(n, D) \tag{16}$$

where $\epsilon(n, D) \to 0$ as $D \to 0$ for all n. Since we do not claim that $\epsilon(n, D)$ converges to zero uniformly for all n, we cannot simply take the limit as $n \to \infty$ of both sides of (16) to obtain the asymptotic behavior of $R_{\mathcal{X}}(D) = \lim_n R_{X^n}(D)$. Fortunately, it is known [9] that

$$|R_{X^n}(D) - R_{\mathcal{X}}(D)| \le \frac{1}{n} I(X^n; X_0, X_{-1}, \dots)$$

where $X_0, X_{-1}, \dots$ are samples from the two-sided stationary extension of $\{X_i\}_{i=1}^n$. Therefore if $\lim_n \frac{1}{n}I(X^n; X_0, X_{-1}, \dots) = 0$, then $R_{X^n}(D)$ converges to $R_{\mathcal{X}}(D)$ *uniformly* for all D. Since each Z_i is a function of X_i with probability 1, and since the X_i are conditionally independent given $\{Z_i\}$, we have

$$I(X^n; X_0, X_{-1}, \dots) = I(Z^n; Z_0, Z_{-1}, \dots).$$

Thus $\lim_n \frac{1}{n}I(X^n; X_0, X_{-1}, \dots) = 0$ if $\lim_n \frac{1}{n}I(Z^n; Z_0, Z_{-1}, \dots) = 0$, which always holds because the Z_i have a finite alphabet (see, e.g., [5, Corollary 6.4.1]).

On the other hand, denoting

$$R_n(D) = \frac{1}{n}H(Z^n) + (1-\alpha)H(P) + \alpha\, h(f) - \frac{\alpha}{2}\log(2\pi e D/\alpha)$$

and

$$R(D) = H(\mathcal{Z}) + (1-\alpha)H(P) + \alpha\, h(f) - \frac{\alpha}{2}\log(2\pi e D/\alpha)$$

we obviously have that $R_n(D)$ converges to $R(D)$ uniformly for all D as $n \to \infty$. These two facts readily imply that

$$\lim_{D\to 0}\left(R_{\mathcal{X}}(D) + \frac{\alpha}{2}\log(2\pi e D/\alpha) - H(\mathcal{Z}) - (1-\alpha)H(P) - \alpha\, h(f)\right) = 0$$

which is equivalent to the claim of Corollary 1. □

Corollary 1 suggests a method that is near optimal for encoding $\{X_i\}$ with small distortion. Since Z^n is a function of X^n it can be losslessly encoded using approximately $H(Z^n)$ bits. The binary vector Z^n specifies the positions of the "discrete" and "continuous" samples of X^n. Therefore the $d(Z^n)$ discrete samples can be losslessly encoded using approximately $d(Z^n)H(P)$ bits and the $c(Z^n)$ continuous samples can be encoded with overall squared distortion $c(Z^n)D/\alpha$ using a vector quantizer which is optimal for the $c(Z^n)$-dimensional i.i.d. random vector with marginal density f. By (1), the vector quantizer will need approximately $c(Z^n)h(f) - \frac{c(Z^n)}{2}\log(2\pi e D/\alpha)$ bits. The normalized expected squared error of this scheme is D, while for large n and small D, the per sample expected rate will be close to $H(\mathcal{Z}) + (1-\alpha)H(P) + \alpha\, h(f) - \frac{\alpha}{2}\log(2\pi e D/\alpha)$. Intuition tells us, and Corollary 1 proves it formally, that this strategy is asymptotically optimal.

3 Proofs

The proof of Theorem 1 is given in two parts. First we show in Lemma 1 that the right hand side of (13) is an asymptotic lower bound on $R_{X^n}(D)$. Then a matching upper bound is proved in Lemma 2. Our method of proof is based partially on [6], but with the help of techniques developed in [4] and [10], we have managed to give simpler proofs of more general results.

Lemma 1 *Assume X^n is of the mixture form (11) and conditions* **(a)** *and* **(b)** *hold. Let $\Lambda = \frac{1}{n}\sum_{j=1}^{N}\alpha_j c_j$. Then we have*

$$\begin{aligned}\liminf_{D\to 0}&\left(R_{X^n}(D) + \frac{\Lambda}{2}\log(2\pi e D/\Lambda)\right)\\ &\geq \frac{1}{n}H(V) + \frac{1}{n}\sum_{j=1}^{N}\alpha_j H(\widehat{X}^{(j)}) + \frac{1}{n}\sum_{j=1}^{N}\alpha_j h(\tilde{X}^{(j)}|\widehat{X}^{(j)}).\end{aligned}$$

Proof. For each $D > 0$, let Y^n be a random n-vector achieving $R_{X^n}(D)$ in the sense that

$$I(X^n, Y^n) = nR_{X^n}(D) \quad \text{and} \quad \mathbf{E}\|X^n - Y^n\|^2 \le D.$$

Since $\mathbf{E}\|X^n\|^2 < \infty$, such Y^n always exists (see e.g. [11]). Note that we have suppressed the dependence of Y^n on D in the notation. It is readily seen that V is a function of X^n with probability 1 since by (12) the distributions of the $X^{(j)}$, $j = 1, \ldots, N$ are mutually singular. This and the chain rule for mutual information imply

$$\begin{aligned} I(X^n; Y^n) = I(X^n, V; Y^n) &= I(V; Y^n) + I(X^n; Y^n|V) \\ &= I(V; Y^n) + \sum_{j=1}^{N} \alpha_j I(X^{(j)}; Y^{(j)}) \end{aligned} \tag{17}$$

where $Y^{(j)}$ denotes a random n-vector whose distribution is equal to the conditional distribution of Y^n given $V = j$. Lemma 3 given in the Appendix implies that

$$\lim_{D\to 0} I(V; Y^n) = H(V). \tag{18}$$

Next we will consider the terms in the sum in (17) individually. Recall equations (8) and (9) defining $\widehat{X}^{(j)}$ and $\tilde{X}^{(j)}$, the discrete and the continuous coordinates of $X^{(j)}$, respectively. By the chain rule we have

$$\begin{aligned} I(X^{(j)}; Y^{(j)}) &= I(\widehat{X}^{(j)}, \tilde{X}^{(j)}; Y^{(j)}) \\ &= I(\widehat{X}^{(j)}; Y^{(j)}) + I(\tilde{X}^{(j)}; Y^{(j)}|\widehat{X}^{(j)}). \end{aligned} \tag{19}$$

Introducing $\widehat{Y}^{(j)} = (Y^{(j)}_{a^j_1}, \ldots, Y^{(j)}_{a^j_{d_j}})$ and $\tilde{Y}^{(j)} = (Y^{(j)}_{b^j_1}, \ldots, Y^{(j)}_{b^j_{c_j}})$, the first term of (19) is sandwiched as

$$H(\widehat{X}^{(j)}) \ge I(\widehat{X}^{(j)}; Y^{(j)}) \ge I(\widehat{X}^{(j)}; \widehat{Y}^{(j)}) \ge d_j R_{\widehat{X}^{(j)}}(\rho)$$

where $\rho = \frac{1}{d_j}\mathbf{E}\|\widehat{X}^{(j)} - \widehat{Y}^{(j)}\|^2$, and where $R_{\widehat{X}^{(j)}}(\rho)$ is the rate-distortion function of $\widehat{X}^{(j)}$. Since $d_j R_{\widehat{X}^{(j)}}(0) = H(\widehat{X}^{(j)})$ and the rate-distortion function (relative to the squared error) of a discrete random variable is continuous at zero (see, e.g., [11, Thm. 2.4]), the fact that $\rho \to 0$ as $D \to 0$ implies

$$\lim_{D\to 0} I(\widehat{X}^{(j)}; Y^{(j)}) = H(\widehat{X}^{(j)}). \tag{20}$$

For the second term in (19) we have

$$\begin{aligned} I(\tilde{X}^{(j)}; Y^{(j)}|\widehat{X}^{(j)}) &= h(\tilde{X}^{(j)}|\widehat{X}^{(j)}) - h(\tilde{X}^{(j)}|Y^{(j)}, \widehat{X}^{(j)}) \\ &\ge h(\tilde{X}^{(j)}|\widehat{X}^{(j)}) - h(\tilde{X}^{(j)}|\tilde{Y}^{(j)}) && (21) \\ &\ge h(\tilde{X}^{(j)}|\widehat{X}^{(j)}) - \frac{c_j}{2}\log(2\pi e D_j/c_j) && (22) \end{aligned}$$

where $D_j = \mathbf{E}\|\tilde{X}^{(j)} - \tilde{Y}^{(j)}\|^2$. In (21) we used the fact that conditioning reduces differential entropy, and (22) holds because

$$h(\tilde{X}^{(j)}|\tilde{Y}^{(j)}) = h(\tilde{X}^{(j)} - \tilde{Y}^{(j)}|\tilde{Y}^{(j)}) \leq h(\tilde{X}^{(j)} - \tilde{Y}^{(j)})$$

and by a well know result [12], the differential entropy of the c_j-dimensional random vector $Z = \tilde{X}^{(j)} - \tilde{Y}^{(j)}$ is upper bounded as $h(Z) \leq \frac{c_j}{2}\log(2\pi e\mathbf{E}\|Z\|^2/c_j)$. Note that $h(\tilde{X}^{(j)} - \tilde{Y}^{(j)})$ is well defined and finite since $h(\tilde{X}^{(j)})$ and $I(\tilde{X}^{(j)}; \tilde{Y}^{(j)})$ are finite.

In summary, (17)-(22) show that as $D \to 0$,

$$\begin{aligned} nR_{X^n}(D) = I(X^n; Y^n) \quad &\geq \quad H(V) + \sum_{j=1}^{N} \alpha_j H(\widehat{X}^{(j)}) + \sum_{j=1}^{N} \alpha_j h(\tilde{X}^{(j)}|\widehat{X}^{(j)}) \\ &\qquad - \frac{1}{2}\sum_{j=1}^{N} \alpha_j c_j \log(2\pi e D_j/c_j) + o(1) \end{aligned} \tag{23}$$

where (18) and (20) have been incorporated into a single term $o(1)$ which converges to zero as $D \to 0$. Recall that we have defined $\Lambda = \frac{1}{n}\sum_{j=1}^{N}\alpha_j c_j$. Then Jensen's inequality and the convexity of the logarithm imply

$$\frac{1}{2}\sum_{j=1}^{N} \alpha_j c_j \log(2\pi e D_j/c_j) \leq \frac{n\Lambda}{2}\log\left(2\pi e \sum_{j=1}^{N} \frac{\alpha_j c_j}{n\Lambda}\frac{D_j}{c_j}\right) \leq \frac{n\Lambda}{2}\log(2\pi e D/\Lambda), \tag{24}$$

since

$$\frac{1}{n}\sum_{j=1}^{N} \alpha_j D_j = \frac{1}{n}\sum_{j=1}^{N} \alpha_j \mathbf{E}\|\tilde{X}^{(j)} - \tilde{Y}^{(j)}\|^2 \leq D.$$

Substitution of (24) into (23) completes the proof of the lemma. $\square$

Lemma 2 *Assume X^n is of the mixture form (11) and conditions* **(a)** *and* **(b)** *hold. Let $\Lambda = \frac{1}{n}\sum_{j=1}^{N}\alpha_j c_j$. Then we have*

$$\begin{aligned} &\limsup_{D\to 0}\left(R_{X^n}(D) + \frac{\Lambda}{2}\log(2\pi e D/\Lambda)\right) \\ &\qquad \leq \quad \frac{1}{n}H(V) + \frac{1}{n}\sum_{j=1}^{N}\alpha_j H(\widehat{X}^{(j)}) + \frac{1}{n}\sum_{j=1}^{N}\alpha_j h(\tilde{X}^{(j)}|\widehat{X}^{(j)}). \end{aligned} \tag{25}$$

Proof. For each $j \in \{1, \ldots, N\}$ define the n-dimensional random vector $Y^{(j)}$ by setting $\widehat{Y}^{(j)} = \widehat{X}^{(j)}$ and $\tilde{Y}^{(j)} = \tilde{X}^{(j)} + Z^{(j)}$, where $Z^{(j)}$ is a c_j-dimensional i.i.d. Gaussian random vector with zero mean and variance D/Λ. It is assumed that $Z^{(j)}$ is independent of $X^{(j)}$ and the index random variable V. In other words, $Y^{(j)}$ is obtained by adding independent Gaussian noise of variance D/Λ to the continuous coordinates of $X^{(j)}$. Let Y^n be the mixture of these distributions, i.e., define $Y^n = Y^{(V)}$. The expected squared error of Y^n is

$$\begin{aligned} \frac{1}{n}\mathbf{E}\|X^n - Y^n\|^2 &= \frac{1}{n}\sum_{j=1}^{N} \alpha_j \mathbf{E}\|X^{(j)} - Y^{(j)}\|^2 \\ &= \frac{1}{n}\sum_{j=1}^{N} \alpha_j \mathbf{E}\|\tilde{X}^{(j)} - \tilde{Y}^{(j)}\|^2 \\ &= \frac{1}{n}\sum_{j=1}^{N} \alpha_j c_j \frac{D}{\Lambda} = D \end{aligned} \tag{26}$$

and therefore by definition, $R_{X^n}(D) \le \frac{1}{n} I(X^n; Y^n)$. In a similar manner as in (17), we obtain

$$I(X^n; Y^n) = I(V; Y^n) + \sum_{j=1}^{N} \alpha_j I(X^{(j)}; Y^{(j)}) \tag{27}$$

where by (26) and Lemma 3 we have

$$\lim_{D\to 0} I(V; Y^n) = H(V). \tag{28}$$

Using the chain rule we can write

$$\begin{aligned} I(X^{(j)}; Y^{(j)}) &= I(\widehat{X}^{(j)}, \tilde{X}^{(j)}; \widehat{Y}^{(j)}, \tilde{Y}^{(j)}) \\ &= I(\widehat{X}^{(j)}; \widehat{Y}^{(j)}, \tilde{Y}^{(j)}) + I(\tilde{X}^{(j)}; \widehat{Y}^{(j)}, \tilde{Y}^{(j)} | \widehat{X}^{(j)}) \\ &= H(\widehat{X}^{(j)}) + I(\tilde{X}^{(j)}; \tilde{Y}^{(j)} | \widehat{X}^{(j)}) \qquad (29) \\ &= H(\widehat{X}^{(j)}) + h(\tilde{Y}^{(j)} | \widehat{X}^{(j)}) - h(\tilde{Y}^{(j)} | \widehat{X}^{(j)}, \tilde{X}^{(j)}) \\ &= H(\widehat{X}^{(j)}) + h(\tilde{X}^{(j)} + Z^{(j)} | \widehat{X}^{(j)}) - h(\tilde{X}^{(j)} + Z^{(j)} | \widehat{X}^{(j)}, \tilde{X}^{(j)}) \qquad (30) \end{aligned}$$

where (29) holds because $\widehat{Y}^{(j)} = \widehat{X}^{(j)}$. Recall that the differential entropy of a Gaussian random variable with variance σ^2 is $\frac{1}{2}\log(2\pi e \sigma^2)$ [12]. Therefore the independence of $X^{(j)}$ and $Z^{(j)}$ implies

$$h(\tilde{X}^{(j)} + Z^{(j)} | \widehat{X}^{(j)}, \tilde{X}^{(j)}) = h(Z^{(j)}) = \frac{c_j}{2}\log(2\pi e D/\Lambda) \tag{31}$$

where the last equality follows because $Z^{(j)}$ has c_j coordinates with common variance D/Λ. On the other hand, Lemma 1 of [13] implies[1]

$$\lim_{D\to 0} h(\tilde{X}^{(j)} + Z^{(j)} | \widehat{X}^{(j)}) = h(\tilde{X}^{(j)} | \widehat{X}^{(j)}). \tag{32}$$

[1]The inequality $h(\tilde{X}^{(j)} + Z^{(j)} | \widehat{X}^{(j)}) \ge h(\tilde{X}^{(j)} | \widehat{X}^{(j)})$ is obvious. The reverse direction is proved by extending the result $\limsup_{\sigma\to 0} h(X + \sigma \cdot Z) \le h(X)$, where X and Z are independent random variables [14], to conditional differential entropies.

From (27)-(32) we can conclude that

$$I(X^n; Y^n) = H(V) + \sum_{j=1}^{N} \alpha_j H(\widehat{X}^{(j)}) + \sum_{j=1}^{N} \alpha_j h(\tilde{X}^{(j)}|\widehat{X}^{(j)}) - \frac{\Lambda}{2}\log(2\pi e D/\Lambda) + o(1)$$

where $\Lambda = \frac{1}{n}\sum_{j=1}^{N} \alpha_j c_j$ and $o(1) \to 0$ as $D \to 0$. Since $R_{X^n}(D) \leq \frac{1}{n} I(X^n; Y^n)$ the proof is complete. □

4 Appendix

Lemma 3 *Suppose X^n is of the mixture form (11) and let $\{Y_k\}_{k=1}^{\infty}$ be a sequence of n-dimensional random vectors such that*

$$\lim_{k\to\infty} \mathbf{E}\|X^n - Y_k\|^2 = 0. \tag{33}$$

Then

$$\lim_{k\to\infty} I(V; Y_k) = H(V).$$

Proof. From (12) we have that V is function of X^n with probability 1. Therefore

$$I(V, X^n) = H(V).$$

On the other hand, since (V, Y_k) converges in distribution to (V, X^n) by (33), the lower semicontinuity of the mutual information (see [11, Lemma 2.2]) implies that

$$\liminf_{k\to 0} I(V, Y_k) \geq I(V, X^n) = H(V).$$

Since $I(V, Y_k) \leq H(V)$, the lemma is proved. □

Acknowledgement

The authors wish to thank Ram Zamir for helpful comments.

References

[1] T. Berger, *Rate Distortion Theory.* Englewood Cliffs, New Jersey: Prentice–Hall, 1971.

[2] Y. N. Linkov, "Evaluation of epsilon entropy of random variables for small epsilon," *Problems of Information Transmission*, vol. 1, pp. 12–18, 1965. Translated from Problemy Peredachi Informatsii, Vol. 1, 18-28.

[3] J. Binia, M. Zakai, and J. Ziv, "On the ϵ-entropy and the rate-distortion function of certain non-Gaussian processes," *IEEE Trans. Inform. Theory*, vol. IT-20, pp. 514–524, July 1974.

[4] T. Linder and R. Zamir, "On the asymptotic tightness of the Shannon lower bound," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 2026–2031, Nov. 1994.

[5] R. M. Gray, *Entropy and Information Theory.* New York: Springer-Verlag, 1990.

[6] H. Rosenthal and J. Binia, "On the epsilon entropy of mixed random variables," *IEEE Trans. Inform. Theory*, vol. IT-34, pp. 1110–1114, Sep. 1988.

[7] Y. Bresler, M. Gastpar, and R. Venkataramani, "Image compression on-the-fly by universal sampling in Fourier imaging systems." *Proc. 1999 IEEE Information Theory Workshop on Detection, Estimation, Classification, and Imaging*, (Santa Fe, NM, 1999), p. 48.

[8] T. Kawabata and A. Dembo, "The rate-distortion dimension of sets and measures," *IEEE Trans. Inform. Theory*, vol. IT-40, pp. 1564–1572, Sep. 1994.

[9] A. D. Wyner and J. Ziv, "Bounds on the rate-distortion function for stationary sources with memory," *IEEE Trans. Inform. Theory*, vol. IT-17, pp. 508–513, Sep. 1971.

[10] T. Linder and R. Zamir, "High-resolution source coding for non-difference distortion measures: the rate distortion function." *IEEE Trans. Inform. Theory* under revision, 1998.

[11] I. Csiszár, "On an extremum problem of information theory," *Studia Scientiarum Mathematicarum Hungarica*, pp. 57–70, 1974.

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

[13] T. Linder, R. Zamir, and K. Zeger, "On source coding with side information dependent distortion measures." submitted to *IEEE Trans. Inform. Theory*, 1998.

[14] A. R. Barron, "Entropy and the central limit theorem," *Ann. Probab.*, vol. 14, no. 1, pp. 336–342, 1986.