A NEW APPROACH TO RATE-DISTORSION THEORY (*)

by JÁNOS KÖRNER (in Budapest) and ANDREA SGARRO (in Trieste) (**)

SOMMARIO. - *La teoria della distorsione di Shannon non è sufficientemente generale per i problemi di codifica di sorgente a più utenti. Diciamo che due criteri di fedeltà sono complementari per una data sorgente quando da una qualunque coppia di codici che li soddisfino si ricava un terzo codice che riproduce la sorgente essenzialmente senza errori. (Ciò porta a una versione non cooperativa del problema della descrizione multipla). Ora il complemento di un criterio di fedeltà alla Shannon non è sempre di questo tipo: ne viene la necessità di una nuova teoria. In questo lavoro proponiamo una tale teoria e proviamo un teorema di codifica diretto.*

SUMMARY. - *The classical rate-distorsion theory of Shannon is not general enough for multi-terminal source coding problems. We would call two fidelity criteria complementary with respect to a given source if any two codes satisfying these two respective criteria can be combined into a third code that guarantees an essentially error-free reproduction of the source. (This situation leads to a non-cooperative version of the multiple description problem). It can be shown that the complement of a Shannon-type fidelity criterion is not necessarily a Shannon-type criterion: hence the need for a new theory. In this paper we put forward such a theory; a direct theorem is proved.*

(*) Pervenuto in Redazione il 3 marzo 1986. Lavoro eseguito durante una visita del primo Autore nell'allora Istituto di Matematica dell'Università di Trieste.

(**) Indirizzi degli Autori: Mathematical Institute of the Hungarian Academy of Sciences, Budapest (Hungary) - Dipartimento di Scienze Matematiche, Università degli Studi, Trieste (Italy).

I. Non-Cooperative Multiple Description - An Introduction

Let us be given a discrete memoryless stationary information source $\{X_i\}_{i=1}^{\infty}$, i.e., the X_i 's are i.i.d. random variables taking their values in the finite set $\mathcal{X}$. Suppose that an encoder is producing n-length block codes $\{f_n, \varphi_n\}_{n=1}^{\infty}$ of this source that meet the known fidelity criterion (d, Δ). Suppose this is all you know about the codes $\{f_n, \varphi_n\}_{n=1}^{\infty}$. Then, at another location, you have another encoder whom you would like to produce some codes that, in a sense, complement the unknown $\{f_n, \varphi_n\}_{n=1}^{\infty}$. In other words, you are trying to devise a code $\{g_n, \psi_n\}_{n=1}^{\infty}$ such that (f_n, g_n) would allow for an essentially error-free reproduction of X^n, the first n outputs of the source. This is always possible, if you set $g_n(X^n) = X^n$, but you would like to do it economically. Is there an optimal way to do it? Can one achieve

(1) $$\frac{1}{n} \log \|f_n\| + \frac{1}{n} \log \|g_n\| \to H(X_1)$$

if $\lim_{n\to\infty} \frac{1}{n} \log \|f_n\|$ is known to be its minimum, the rate-distorsion value $R(X_1, d, \Delta)$? Is there a fidelity criterion (d^*, Δ^*) optimal satisfaction of which by $\{g_n, \psi_n\}_{n=1}^{\infty}$ would automatically guarantee that $\{f_n, g_n\}$ contain enough information for an essentially error-free reproduction of X^n, and even (1) holds? These are some of the questions we would like to answer. First, some definitions. Here and in the sequel log's and exp's are to the base 2.

A Shannon-type fidelity criterion (d, Δ) is given in terms of a function $d: \mathcal{X} \times \mathcal{Y} \to \mathbf{R}^+$ that assigns a non-negative value to every pair of elements $x \in \mathcal{X}$, $y \in \mathcal{Y}$, where $\mathcal{Y}$ is an arbitrary finite set called the reproduction alphabet. Then $d(x, y)$ is interpreted as a measure of the loss resulting from one's representing x by y. Concentrating on block codes, we now extend d by introducing

$$d_n: \mathcal{X}^n \times \mathcal{Y}^n \to \mathbf{R}^+$$

through

$$d_n(\underline{x}, \underline{y}) = \frac{1}{n} \sum_{i=1}^{n} d(x_i, y_i), \quad \underline{x} = x_1 \dots x_n, \quad \underline{y} = y_1 \dots y_n.$$

An n-length block code is a pair of mappings (f_n, φ_n) such that the composite mapping $\varphi_n(f_n)$ maps $\mathcal{X}^n$ into $\mathcal{Y}^n$. The code $\{f_n, \varphi_n\}_{n=1}^{\infty}$ is said to meet the fidelity criterion (d, Δ) if

(2) $$\Pr\{d_n(X^n, \varphi_n(f_n(X^n))) > \Delta\} \to 0.$$

Information theory is devoted to the study of cheap ways of efficient communication. In the present model we ask for the least complex f_n that achieves (2). Let $\|f_n\|$ denote the number of different values f_n takes. Rate-distortion theory tells us that

$$\liminf_{n\to\infty} \frac{1}{n} \log \|f_n\| = R(d, \Delta) \tag{3}$$

whenever $\{f_n, \varphi_n\}$ meet the fidelity criterion (d, Δ). Here $R(d, \Delta)$ is the rate-distortion function of the source $\{X_i\}_{i=1}^{\infty}$, and we have

$$R(d, \Delta) = R(X_1, d, \Delta) = \min_{\substack{X : P_X = P_{X_1} \\ Ed(X,Y) \leqslant \Delta}} I(X \wedge Y). \tag{4}$$

For more detail on this particular result of Shannon and information theory in general, we refer the reader to [1] or [2]. Time has proved that the above is a useful model of data compression. This does not mean, however, that one should not look for other, equally meaningful models.

In the past decade or so, information-theoretic research has been concentrating on multi-terminal communication problems. A particularly intriguing question several authors have investigated is that of multiple descriptions, [3] - [6]. We shall concentrate on a characteristic special case of this still unsolved problem, in which the basic difficulty is already present. In this case, a discrete memoryless stationary information source $\{X_i\}_{i=1}^{\infty}$ with alphabet $\mathcal{X}$ has to be reconstructed within small probability of error on the basis of two separately encoded versions of the same source. At the same time these two codes have to satisfy certain fidelity criteria, one for each. Formally, let (f_n, φ_n) be an n-length block code of $\{X_i\}$ that ε_n-satisfies the fidelity criterion (d', Δ') where d' is defined on $\mathcal{X} \times \mathcal{Y}$. Similarly, let (g_n, ψ_n) be another n-length block code of $\{X_i\}$ ε_n-satisfying the fidelity criterion (d'', Δ'') where d'' is defined on $\mathcal{X} \times \mathcal{Z}$, i.e.,

$$\begin{aligned} &\Pr\{d'(X^n, \varphi_n(f_n(X^n))) > \Delta'\} \leqslant \varepsilon_n \\ &\Pr\{d''(X^n, \psi_n(g_n(X^n))) > \Delta'\} \leqslant \varepsilon_n . \end{aligned} \tag{5}$$

Further, let the function ω_n be defined on the Cartesian product of the ranges of f_n and g_n so that the inequality

$$\Pr\{X^n \neq \omega_n(f_n(X^n), g_n(X^n))\} \leqslant \varepsilon_n \tag{6}$$

holds. We say that (R', R'') is an achievable rate pair for the (cooperative) multiple description problem if for every $\delta > 0$ and $n \geqslant n_0(\delta)$ there exist n-length block codes (f_n, φ_n), (g_n, ψ_n), a corresponding common decoder ω_n satisfying (5)-(6) for some $\varepsilon_n < \delta$ and

$$\frac{1}{n} \log \|f_n\| < R' + \delta \quad \frac{1}{n} \log \|g_n\| < R'' + \delta . \tag{7}$$

The determination of all achievable rate pairs for given $\{X_i\}$, (d', Δ') and (d'', Δ'') remains an open problem.

In an early paper [7], G. Longo and the first author have introduced a non-cooperative variant of this problem. In the previous setting, one was looking for optimal pairs of (f_n, φ_n) and (g_n, ψ_n) along with a suitable ω_n. Thus, although the actual coding was supposed to be carried out at separate locations not being able to communicate with one another, the design of the overall system of codes was supposed to be done in collaboration. In fact, typically, an optimal system would consist of (f_n, φ_n), (g_n, ψ_n) which are suboptimal with respect to their own fidelity criteria. Some individual rate has to be sacrified in the interest of the common goal of achieving (6). In other words, typically, for an achievable rate pair (R', R'') one has

$$R' + R'' > R(d', \Delta') + R(d'', \Delta'').$$

What happens if no cooperation in designing the codes is possible? What is the minimum rate the encoder for (d'', Δ'') has to choose if all it knows about the other code is that it satisfies the criterion (d', Δ')? It can always encode the whole source, i.e., $R'' = H(X_1)$ is a safe choice, and hence the problem is always meaningful. What, however, is the smallest value of R'' if $R' = R(d', \Delta')$? A particularly interesting case is when

$$R' = R(d', \Delta') \quad R'' = R(d'', \Delta'')$$

is an achievable rate pair, and in fact, to every code (f_n, φ_n) and (g_n, ψ_n) satisfying (5) there exists an ω_n satisfying (6). Clearly, this is a property of the fidelity criteria (d', Δ') and (d'', Δ''). We shall call them *complementary* if they have this property. To every (d', Δ') there are several choices of a complementary (d'', Δ''). The simplest choice is $Z = X$,

$$d''(x, z) = 1 - \delta_{xz}, \quad \Delta'' = 0.$$

What is the condition for equality? We shall not answer these complementary (d'', Δ'') minimizes $R(X_1, d'', \Delta'')$? In the case of simple fidelity criteria, those for which the distortion measure takes only one non-zero value and for distortion level Δ this problem has been studied in [7]. Clearly, if (d', Δ') and (d'', Δ'') are complementary, then

$$R(X_1, d', \Delta') + R(X_1, d'', \Delta'') \geqslant H(X_1).$$

What is the condition for equality? We shall not answer these questions here. We have shown in [7] that some fidelity criteria have a natural (optimum) complement, which, however, cannot be defined in terms of a distortion measure d and distortion threshold Δ in the previous sense. In the present paper, we shall introduce a new kind of fidelity criteria for which the optimal complementary criterion also belongs to the same class. The corresponding multiple description problem will be addressed in a subsequent paper.

II. An intrinsic distortion measure

We shall restrict ourselves to block codes. An n-length block encoder of the source $\{X_i\}_{i=1}^{\infty}$ is an arbitrary mapping f with domain $\mathcal{X}^n$. Its aim is to compress the source within tolerable distortion. Compression creates a situation whereby certain source outputs — those belonging to the same level set of f — become indistinguishable. This is an unavoidable side effect of data compression. In order to evaluate it, we shall introduce the concept of an intrinsic distortion measure (IDM). A function $d : \mathcal{X} \times \mathcal{X} \to \mathbf{R}^+$ where $\mathbf{R}^+$ is the set of the non-negative real numbers is an IDM if

$$d(x', x'') = d(x'', x') \text{ and } d(x, x) = 0$$

for every x', x'' and x in $\mathcal{X}$. We extend d to sequences in the usual way:

$$d(\underline{x}', \underline{x}'') = \frac{1}{n} \sum_{i=1}^{n} d(x'_1, x''_1), \text{ where } \underline{x}' = x'_1 x'_2 \ldots x'_n \text{ and}$$

$$\underline{x}'' = x''_1 x''_2 \ldots x''_n .$$

Given an n-length block code $f : \mathcal{X}^n \to \mathcal{M}$ where $\mathcal{M}$ is an arbitrary finite set, all the elements of the level set $f^{-1}(m), m \in \mathcal{M}$, the full inverse image of m, can be confounded with one another. Our fidelity criterion is given in terms of a threshold Δ. We declare an error if two sequences confounded by the code f have an intrinsic distortion larger than Δ. Let us introduce

$$d(A) = \max_{\substack{\underline{x} \in \mathcal{A} \\ \underline{y} \in \mathcal{A}}} d(\underline{x}, \underline{y}).$$

We shall say that the code f *ε-meets the intrinsic fidelity criterion* (d, Δ) if

$$\Pr\{d(f^{-1}(f(X^n))) > \Delta\} < \varepsilon . \tag{8}$$

This means that except for a set of probability less than ε, no confoundable source outputs have a pairwise distortion larger than Δ. A non-negative number R will be called an *ε-achievable rate* for the source $\{X_i\}$ and the intrinsic fidelity criterion (IFC) (d, Δ) if for any $\delta > 0$ and $n \geqslant n_0(\delta)$ there exist n-length block codes ε-meeting the IFC (d, Δ). Similarly, R will be called an achievable rate for $\{X_i\}_{i=1}^{\infty}$ and (d, Δ) if it is ε-achievable for every $\varepsilon \in (0, 1)$.

For the rest of this paper, we fix the source $\{X_i\}_{i=1}^{\infty}$, and denote by P the common distribution of the X_i's. Similarly, we fix the IDM d and will omit references to both whenever this does not create misunderstandings. Let $R^*_\varepsilon(\Delta)$ denote the infimum of the ε-achievable rates for the fixed source and IDM, and let $R^*(\Delta)$ denote the infimum of the achievable rates in the same case. By definition, we have

$$R^*(\Delta) = \sup_{\varepsilon \in (0,1)} R^*_\varepsilon(\Delta).$$

To state our main result, we need some notation. Let $\mathcal{U}$ be an arbitrary finite set. We set

(9) $$\mathcal{S}(\mathcal{U}, \Delta) = \{P_{UX} : U \in \mathcal{U}, X \in \mathcal{X}, P_X = P, \max_{UX\tilde{X} : P_{UX} = P_{U\tilde{X}}} Ed(X, \tilde{X}) \leq \Delta\},$$

where P_Y denotes the probability distribution of the random variable (RV) Y. Note that $\mathcal{S}(\mathcal{U}, \Delta)$ is non-empty for $|\mathcal{U}| \geqslant |\mathcal{X}|$, since assuming $\mathcal{X} \subseteq \mathcal{U}$ we can always have $U \equiv X$.

For arbitrary $\Delta \geqslant 0$ we write

(10) $$\tilde{R}(\Delta) = \tilde{R}(P, d, \Delta) = \inf_{\substack{|\mathcal{U}| < \infty, \\ P_{UX} \in \mathcal{S}(U, \Delta)}} I(U \wedge X)$$

and

(11) $$R(\Delta) = R(P, d, \Delta) = \min_{\substack{P_{UX} : |\mathcal{U}| \leqslant |\mathcal{X}|^2 + 2 \\ P_{UX} \in \mathcal{S}(U, \Delta)}} I(U \wedge X).$$

Obviously, $\tilde{R}(\Delta) \leqslant R(\Delta)$. A routine argument will show (cf. the Appendix) that

LEMMA 1 - $\tilde{R}(\Delta) = R(\Delta)$.

LEMMA 2 - $R(\Delta)$ is a continuous, convex, non-increasing function of Δ. $R(\Delta) \geqslant 0$ with equality only if $\Delta \geqslant \Delta^*$ where

$$\Delta^* = \max_{X\tilde{X} : P_X = P_{\tilde{X}} = P} Ed(X, \tilde{X}).$$

The proof of this lemma will be postponed to the Appendix.

III. The Coding Theorem

Our main result is

THEOREM

$$R^*_\varepsilon(\Delta) \leq R^*(\Delta) \leq R(\Delta) \text{ for } \varepsilon \in (0,1).$$

Proof. The result is a straightforward consequence of the proof of Berger's Type Covering Lemma [5, Lemma 2.4.1]. In fact, let the RV's UX achieve the minimum in the definition of $R(\Delta)$, cf. (11).

Then, by definition, we have

(12) $$\max_{P_{UX} = P_{U\tilde{X}}} Ed(X, \tilde{X}) \leq \Delta.$$

Let us fix some $\delta > 0$. By the proof of Lemma 2.4.1 in [1], there exists a set $\mathcal{B} \subset T^n_{[U]}$ of U-typical sequences of length n with cardinality

$$|\mathcal{B}| \leqslant \exp\left[n\left(I(U \wedge X) + \frac{\delta}{2}\right)\right] \tag{13}$$

such that, with the notation of [1],

$$T^n_{[X]} \subset \bigcup_{\underline{u} \in \mathcal{B}} T^n_{[X|U]}(\underline{u}).$$

Hence there exists an obvious construction for a function $\tilde{f} : \mathcal{X}^n \to \mathcal{B} \cup \{b_0\}$ with the property

$$\tilde{f}^{-1}(\underline{u}) \subset T^n_{[X|U]}(\underline{u}), \quad \underline{u} \in \mathcal{B}.$$

Let us partition every level set $f^{-1}(\underline{u}), \underline{u} \in \mathcal{B}$ of this function according to the joint type of $\underline{u}$ and $\underline{x} \in f^{-1}(\underline{u})$. More precisely, let $\mathcal{S}_n$ denote the set of all possible joint types of pairs of n-length sequences $\underline{u} \in \mathcal{B}$, $\underline{x} \in T^n_{[X]}$, and let $f : \mathcal{X}^n \to \mathcal{B} \times \mathcal{S}_n$ map every $\underline{x} \in T^n_{[X]}$ into

$$f(\underline{x}) = (\tilde{f}(\underline{x}), P_{\tilde{f}(\underline{x})\underline{x}}),$$

where $P_{\tilde{f}(\underline{x})\underline{x}}$ is the joint type of $\tilde{f}(\underline{x})$ and $\underline{x}$. By the Type Counting Lemma [1], p. 29 and (13),

$$\frac{1}{n}\log\|f\| \leq \frac{1}{n}\log|\mathcal{B}| + |\mathcal{U}|\,|\mathcal{X}|\,\frac{\log(n+1)}{n}. \tag{14}$$

Further, if $b \in \mathcal{B}$, $Q \in \mathcal{S}_n$ and $f^{-1}(b, Q)$ is non-empty, then, obviously, for n large enough, $n \geq n_0(\delta)$, (12) implies

$$d(f^{-1}(b, Q)) \leq \Delta + \delta.$$

Next observe that (cf. Lemma 2.12 in [1])

$$\Pr\{X^n \in T^n_{[X]}\} \to 1$$

and thus, for $n \geq n_\circ(\delta, P)$

$$\Pr\{d(f^{-1}(f(X^n))) > \Delta + \delta\} \to 0.$$

On the other hand, (13) and (14) imply

$$\frac{1}{n}\log\|f\| \leq R(\Delta) + \delta$$

for $n \geq n_\circ(\delta)$. Hence by the continuity of $R(\Delta)$ proved in Lemma 2 the result follows.

IV. Discussion

The above model, and in particular, the evaluation of the rate-distortion trade-off depend entirely on the encoder, inasmuch it is the encoder alone that determines the kind of data compression the encoder-decoder pair achieves.

Formally, we may introduce a decoder in the following way: given an encoder $f : \mathcal{X}^n \to \mathcal{M}$ where $\mathcal{M}$ is an arbitrary finite set as before, we let $g : \mathcal{M} \to 2^{\mathcal{X}^n}$ map the elements of $\mathcal{M}$ into subsets of $\mathcal{X}^n$. Here $2^{\mathcal{X}^n}$ denotes the family of all such subsets. We may choose

$$g(m) = f^{-1}(m).$$

We can define

$$d(\underline{x}, \mathcal{A}) = \begin{cases} \max\limits_{\substack{\underline{x}' \in \mathcal{A} \\ \underline{x}'' \in \mathcal{A}}} d(\underline{x}', \underline{x}'') & \text{if } \underline{x} \in \mathcal{A}. \\ +\infty & \text{else} \end{cases}$$

Clearly, this «distortion measure» cannot be derived from a single-letter measure of the same kind and thus our new model is not a special case of the classical one of Shannon. Nevertheless, just as in the case of Shannon's model, our IFC (8) has an alternative. We shall say that the code (f, g) meets the IFC (d, Δ) in the average if with the above extension of d to a distortion measure on $\mathcal{X}^n \times 2^{\mathcal{X}^n}$ we have

$$Ed(X^n, g(f(X^n))) \leq \Delta. \tag{15}$$

Similarly to the foregoing, we shall denote by $S^*(\Delta)$ the infimum of those numbers $R \geq 0$ for which there exists a sequence (f_n, g_n) of n-length block codes meeting (15), and satisfying

$$\liminf_{n\to\infty} \frac{1}{n} \log \| f_n \| = R.$$

One easily sees using the technique applied in this paper that

$$S^*(\Delta) \leq R(\Delta) \quad \text{for } \Delta > 0.$$

Although $S^*(0) \neq R(0)$, the determination of $S^*(0)$ is equivalent to a combinatorial problem solved by McEliece and Posner [8], cf. [1], Problem 2.4.11. In order to keep this paper selfcontained, we do not go into the details for which graph theoretic terminology would be needed.

The special case $\Delta = 0$ of our model can be interpreted within the framework of Shannon's classical theory. This special case has an interesting graph-theoretic interpretation introduced in [9], where the quantity $R^*(0)$ is called graph entropy. Recently, the latter has been used very substantially in combinatorics to derive efficient non-existence bounds for the number of functions necessary in perfect hashing, [10], [11]. Using [10], one can obtain a more explicit formula for $R(0)$. In order to see this, let us take a closer look at the set $\mathcal{S}(\mathcal{U},0)$. Clearly,

$$\mathcal{S}(\mathcal{U},0)=\{P_{UX}: U\in\mathcal{U},\ X\in\mathcal{X},\ P_X=P,\ P_{X\hat{X}}(x',x'')>0 \text{ and } P_{UX}=P_{U\hat{X}} \text{ imply } d(x',x'')=0 \text{ for every } x',x''\in\mathcal{X}\}.$$

This has the consequence that for every $P_{UX}\in\mathcal{S}(\mathcal{U},0)$ the support of every distribution $P_{UX}(\cdot\mid u)$, $u\in\mathcal{U}$ must be a set on which $d(.,.)$ is identically zero. If $\mathcal{A}_u$ is such a set, then $x'\in\mathcal{A}_u$, $x''\in\mathcal{A}_u$ imply $d(x',x'')=0$. Let us now introduce a graph G on $\mathcal{X}$ in the following way: $(x',x'')\in E(G)$, i.e., there is an edge between x' and x'' in G if and only if $d(x',x'')>0$. Let $\mathcal{A}\subset 2^{X^n}$ be the family of all the independent sets of G. For any fixed $P_{UX}\in\mathcal{S}(\mathcal{U},0)$ to every $u\in\mathcal{U}$ there corresponds a unique $A\subset 2^{X^n}$, the supporting set of the distribution $P_{UX}(\cdot\mid u)$. Identifying for every *PD* P_{UX} the set $\mathcal{U}$ with the corresponding subset of it, we can think of P_{UX} as a distribution on $\mathcal{X}\times\mathcal{A}$. Let us denote the family of these distributions as P_{UX} runs over $\mathcal{S}(\mathcal{U},0)$ by $\mathcal{H}$. Then,

$$R(0)=\min_{\substack{P_{UX}\in\mathcal{H}\\ |U|\leqslant|\mathcal{X}|^2+2}} I(U\wedge X).$$

This is the same formula as in [9], with the slight advantage of having a bound on the number of independent sets needed (recall Lemma 1).

Appendix

Proof of Lemma 1 - This a straightforward application of the Support Lemma [1], Lemma 3.3.4. Let us fix any distribution $P_{UX\hat{X}}\in\mathcal{S}(\mathcal{U},\Delta)$, where $U\in\mathcal{U}$ and $|\mathcal{U}|<\infty$. We would like to prove that there is another *PD* $P_{U'X'\hat{X}'}\in\mathcal{S}(\mathcal{U},\Delta)$ for which

$$Ed(X,\hat{X})=Ed(X',\hat{X})$$

$$I(U\wedge X)=I(U'\wedge X')$$

and

$$U'\in\mathcal{U}' \text{ with } |\mathcal{U}'|\leq|\mathcal{X}|^2+2.$$

Consider the following $|\mathcal{X}|^2+2$ continuous real functions over the set of all the *PD*'s on $\mathcal{X}\times\mathcal{X}$:

$$f_{x',x''}(P_{X\hat{X}})\overset{\Delta}{=}P_{X\hat{X}}(x',x'') \text{ for every } (x',x'')\in\mathcal{X}\times\mathcal{X} \text{ except, } (x_\circ,x_\circ), \text{ say}.$$

$$f_{|\mathfrak{X}|^2}(P_{X\hat{X}}) \stackrel{\Delta}{=} Ed(X, \hat{X})$$

$$f_{|\mathfrak{X}|^2+1}(P_{X\hat{X}}) \stackrel{\Delta}{=} H(X)$$

$$f_{|\mathfrak{X}|^2+2}(P_{X\hat{X}}) \stackrel{\Delta}{=} \sum_{x \in \mathfrak{X}} |P_X(x) - P_{\hat{X}}(x)|.$$

An application of the Support Lemma to these $|\mathfrak{X}|^2 + 2$ functions shows that there is a random triple as asserted.

Proof of Lemma 2 - The monotonicity and non-negativity are obvious.

$R(\Delta) = 0$ is possible if and only if there is a PD in $\mathcal{S}(\mathcal{U}, \Delta)$ for which U and X are independent and thus $Ed(X, \hat{X}) \leq \Delta$ must hold whenever $P_X = P_{\hat{X}} = P$. Clearly, this happens only if $\Delta \geq \Delta^*$.

Let us look at convexity. Let Δ_1, Δ_2 and $\alpha \in (0, 1)$ be arbitrary non-negative numbers and write $\bar{\alpha} = 1 - \alpha$, $\Delta = \alpha\Delta_1 + \bar{\alpha}\Delta_2$. Assume that P_{UX} and P_{VY} achieve the minimum in the definition of $R(\Delta_1)$ and $R(\Delta_2)$, respectively, $P_X = P_Y = P$, $U \in \mathcal{U}$, $V \in \mathcal{V}$. Let us introduce the random triple IWZ by the following conditions. Without loss of generality we can suppose that $\mathcal{U} \cap \mathcal{V} = \phi$. Let W have range $\mathcal{U} \cup \mathcal{V}$ and let Z be conditionally independent of I given W. Further, assume

$$\Pr\{I = 1\} = 1 - \Pr\{I = 2\} = \alpha,$$
$$P_{W|I}(\cdot \mid I = 1) = P_U(\cdot)$$
$$P_{W|I}(\cdot \mid I = 2) = P_V(\cdot).$$

We claim that $P_{WZ} \in \mathcal{S}(\mathcal{U} \cup \mathcal{V}, \Delta)$. Obviously $P_Z = P$. Assume that $\hat{Z}$ is an arbitrary RV with $P_{WZ} = P_{W\hat{Z}}$. Then

$$Ed(Z, \hat{Z}) = \alpha Ed(X, \hat{X}) + \bar{\alpha} Ed(Y, \hat{Y}) \leq \alpha\Delta_1 + \bar{\alpha}\Delta_2 = \Delta,$$

where $\hat{X}$ resp. $\hat{Y}$ are RV's the joint distribution of which with X resp. Y equals the conditional joint distribution of Z and $\hat{Z}$ under the respective conditions $I = 1$ and $I = 2$. By the last inequality and the definition of $R(\Delta)$, we see that

$$R(\Delta) \leq I(W \wedge Z) \leq I(IW \wedge Z) = I(I \wedge Z) + I(Z \wedge W \mid I).$$

But I and Z are independent, and hence

$$R(\Delta) = I(Z \wedge W \mid I) = \alpha I(U \wedge X) + \bar{\alpha} I(V \wedge Y) = \alpha R(\Delta_1) + \bar{\alpha} R(\Delta_2).$$

Convexity implies that $R(\Delta)$ is continuous for every $\Delta > 0$. For $\Delta = 0$ monotonicity implies

$$R(0) \geq \sup_{\Delta > 0} R(\Delta).$$

We will conclude by proving inequality in the opposite direction. By the well-known Ascoli theorem, cf. [12], $\max_{P_{UX\tilde{X}} : P_{UX} = P_{U\tilde{X}}} Ed(X, \tilde{X})$ is a continuous function of P_{UX}. Hence $R(\Delta)$ is of the form

$$R(\Delta) = \min_{t : g(t) \leq \Delta} f(t)$$

where both $f(\cdot)$ and $g(\cdot)$ are continuous. Let $\Delta_n \to 0$ monotonically. Let t_n achieve minimum in the definition of $R(\Delta_n)$. Then, by the compactness of the range of the t's (the distributions P_{UX}) we have a subsequence of the t_n's converging to some t_∞. Without loss of generality, we can suppose that $t_n \to t_\infty$. Since Δ_n is a decreasing sequence, for every Δ_n we have $g(t_m) \leq \Delta_n$ for every $m \geq n$ and hence $g(t_\infty) \leq \Delta_n$ by the continuity of $g(\cdot)$; and therefore $g(t_\infty) = 0$. This means that $R(0) \leq f(t_\infty)$. However, by the continuity of $f(\cdot)$, $f(t_\infty) = \lim_{n \to \infty} f(t_n) = \lim_{n \to \infty} R(\Delta_n)$. Hence $\lim_{n \to \infty} R(\Delta_n) = R(0)$.

ACKNOWLEDGEMENT

The first author would like to express his thanks to Victor K. Wei for fruitful discussions at Bell Laboratories.

REFERENCES

[1] I. CSISZÁR, J. KÖRNER, *Information Theory. Coding Theorems for Discrete Memoryless Systems.* Akadémiai Kiadó, Budapest, 1981 and Academic Press, New York, 1982.

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

[3] T. COVER, A. EL GAMAL, *Achievable rates for multiple descriptions.* IEEE Trans. Inform. Theory, Vol. IT-28, pp. 851-857, 1982.

[4] L. H. OZAROW, *On a source-coding problem with two channels and three receivers.* BSTJ, Vol. 59, n. 10, 1909-1921, 1980.

[5] R. F. AHLSWEDE, *The rate-distortion region for multiple descriptions without excess rate.* IEEE Trans. Inform. Theory, Vol. 31, pp. 721-726, 1985.

[6] T. BERGER, Z. ZHANG, *New Results in Binary Multiple Descriptions.* IEEE Trans. Inform. Theory, to appear.

[7] J. KÖRNER, G. LONGO, *Two-Step Encoding for Finite Sources.* IEEE Trans. Inform. Theory, Vol. IT-19, pp. 778-782, 1973.

[8] E. C. POSNER, R. J. McELIECE, *Hide and seek, data storage and entropy.* Annals of Math. Stat., Vol. 42, pp. 1706-1716, 1971.

[9] J. KÖRNER, *Coding of an information source having ambiguous alphabet and the entropy of graphs.* In: Proc. 6th Prague Conference on Inform. Theory, Academia, Prague, pp. 411-425, 1973.

[10] M. FREDMAN, J. KOMLÓS, *On the Size of Separating Systems and Perfect Hash Functions.* SIAM J. on Algebraic and Discrete Methods, Vol. 5, n. 1, pp. 61-68, 1984.

[11] J. KÖRNER, *Fredman - Komlós bounds and information theory.* SIAM J. on Algebraic and Discrete Methods, Vol. 7, n. 4, pp. 560-570, 1986.

[12] H. ROYDEN, *Real Analysis.* MacMillan, London, 1968.