INFORMATION THEORY OF MULTIPLE DESCRIPTIONS

BY

ABBAS A. EL GAMAL and THOMAS M. COVER

TECHNICAL REPORT NO. 43
SEPTEMBER 1980

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT ECS 78-23334

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
INFORMATION THEORY OF MULTIPLE DESCRIPTIONS

by

Abbas A. El Gamal
Thomas M. Cover

TECHNICAL REPORT NO. 43
SEPTEMBER 1980

PREPARED UNDER THE AUSPICES
OF
NATIONAL SCIENCE FOUNDATION
GRANT ECS 78-23334

DEPARTMENT OF STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA
INFORMATION THEORY OF MULTIPLE DESCRIPTIONS

by

Abbas A. El Gamal
Thomas M. Cover

Abstract

Consider a sequence of i.i.d. random variables \( X_1, X_2, \ldots, X_n \) and a distortion measure \( d(X_i, \hat{X}_i) \) on the estimates \( \hat{X}_i \) of \( X_i \). Two descriptions \( i(X) \in \{1,2,\ldots,2^{-nR_1}\} \) and \( j(X) \in \{1,2,\ldots,2^{-nR_2}\} \) are given of the sequence \( X = (X_1, X_2, \ldots, X_n) \). From these two descriptions, three estimates \( \hat{X}_1(i(X)), \hat{X}_2(j(X)) \), and \( \hat{X}_0(i(X),j(X)) \) are formed, with resulting expected distortions

\[
E \frac{1}{n} \sum_{k=1}^{n} d(X_k, \hat{X}_{mk}) = D_m, \quad m = 0,1,2.
\]

We find that the distortion constraints \( D_0, D_1, D_2 \) are achievable if there exists a probability mass distribution \( p(x)p(\hat{X}_1, \hat{X}_2, \hat{X}_0|x) \) with \( Ed(X, \hat{X}_m) \leq D_m \) such that

\[
R_1 > I(X; \hat{X}_1) \\
R_2 > I(X; \hat{X}_2) \\
R_1 + R_2 > I(X; \hat{X}_1, \hat{X}_2, \hat{X}_0) + I(\hat{X}_1; \hat{X}_2),
\]

where \( I(\cdot) \) denotes Shannon mutual information.

These rates are shown to be optimal for deterministic distortion measures.

†At Stanford University on leave from University of Southern California. This work was partially supported under the Army Research Office, Grant No. DAAG29-79-C-0054.

‡‡Stanford University. This work was partially supported by National Science Foundation Grant ENG78-23334 and Stanford Research Institute Contract DAHC-15-C-0187.
1. **Introduction**

The following problem of jointly good descriptions was posed by Wolf, Wyner, Ziv, Witsenhausen, and Ozarow at the September 1979 IEEE Information Theory Workshop. Suppose we wish to send a description of a stochastic process to a destination through a communication network. Assume that there is a chance that the description will be lost. Therefore we send two descriptions and hope that one of them will get through. Each description should be individually good. However, if both get through, then we wish the combined descriptive information to be as large as possible.

The difficulty is that good individual descriptions must be close to the process, by virtue of their goodness, and necessarily must be highly dependent. Thus the second description will contribute little extra information beyond one alone. On the other hand, two independent descriptions must be far apart and thus cannot in general be individually good.

The more general problem, stated precisely in the next section, is as follows. Consider a stochastic process $X_1, X_2, \ldots$ where the $X_i$'s are i.i.d. according to some known distribution $p(x)$. Two individuals must describe $X$ at respective rates $R_1$ and $R_2$ bits per transmission. Three single letter distortion measures $d_1, d_2,$ and $d_0$ are given. The question is, "What information should be sent at rates $R_1$ and $R_2$ so that a receiver given only description 1 can recover $X$ with distortion $D_1$, a receiver seeing only description 2 can recover $X$ with distortion $D_2$, and a receiver seeing both descriptions can recover $X$ with distortion $D_0$?" For fixed distortions $D_0, D_1,$ and $D_2$, what is the set of $(R_1, R_2)$ necessary and sufficient to achieve these distortions?

In this paper, we shall exhibit an achievable rate region of $(R_1, R_2)$
pairs as a function of the distortion vector \( D = (D_0, D_1, D_2) \). See Figure 1.

**Figure 1.**

Multiple Descriptions
The following examples consistent with Figure 1 complete our discussion of the motivation for a general theory of joint descriptions.

**Communication network:** A communication network is used to send descriptions of \( X \) to New York with distortion \( D_1 \), at a cost \( c_1 \) dollars/bit, and to Boston at distortion \( D_2 \) and cost \( c_2 \) dollars/bit. Given an acceptable distortion \( D_0 \) for the best estimate of \( X \) from the combined data base, minimize the cost \( c = c_1 R_1 + c_2 R_2 \).

**Data base:** We wish to store the data \( X = (X_1, \ldots, X_n) \) with distortion \( D_1 \) in a New York computer with memory capacity \( nR_1 \) and store \( X \) with distortion \( D_2 \) in a Boston computer with memory capacity \( nR_2 \). How should we do this so that when we combine the data we recover \( X \) with minimal distortion \( D_0 \)?

**Manager:** A manager instructs two survey teams to gather information about \( X \) for their own use and for the subsequent use of the manager. What should he ask them to report?
2. Definitions

We shall first introduce the basic definitions of rate distortion theory and state Shannon's rate distortion theorem. Then the definitions will be extended to multiple descriptions and their incurred distortions.

We assume that \( X_i, i = 1, 2, \ldots \) is a sequence of i.i.d. discrete random variables drawn according to a probability mass function \( p(x) \). We are given a reconstruction space \( \hat{X} \) together with an associated distortion measure \( d: X \times \hat{X} \rightarrow \mathbb{R} \). The distortion measure on \( n \)-sequences in \( X^n \times \hat{X}^n \) is defined by the average per symbol distortion

\[
d(x, \hat{x}) = \left( \frac{1}{n} \right) \sum_{i=1}^{n} d(x_i, \hat{x}_i).
\]

A description of \( x \in X^n \) is a map \( i: X^n \rightarrow \{1, 2, \ldots, 2^{nR}\} \), where \( R \) is the rate of the description in bits per symbol of \( x \). A reconstruction of \( X \) is a map \( \hat{x}: \{1, 2, \ldots, 2^{nR}\} \rightarrow \hat{X}^n \), and is said to incur distortion \( D^{(n)} \) defined by

\[
D^{(n)} = E d(X, \hat{x}(i(X)))
= E \left( \frac{1}{n} \right) \sum_{k=1}^{n} d(X_k, \hat{x}_k(i(X))).
\]

The distortion \( D \) is said to be achievable with rate \( R \) for the stochastic process \( \{X_i\}_{i=1} \) if there exists a sequence of rate \( R \) descriptions \( i: X^n \rightarrow \{1, 2, \ldots, 2^{nR}\} \) and reconstructions \( \hat{x}: \{1, 2, \ldots, 2^{nR}\} \rightarrow \hat{X}^n \), such that \( D^{(n)} \leq D \), for all \( n \) sufficiently large.

Let the rate distortion function \( R(D) \) be the infimum of all rates \( R \) achieving distortion \( D \) on a given stochastic process \( \{X_i\} \).
Theorem (Shannon [1]): If $X_i$, $i = 1, 2, \ldots$, are i.i.d. discrete finite alphabet r.v.'s with probability mass function $p(x)$, then

$$R(D) = \inf I(X; \hat{X})$$

where

$$I(X; \hat{X}) = \sum_{x, \hat{x}} p(x, \hat{x}) \log \frac{p(x, \hat{x})}{p(x)p(\hat{x})},$$

and the infimum is taken over all joint probability mass functions $p(x)p(\hat{x}|x)$ such that

$$\sum_{x, \hat{x}} p(x, \hat{x})d(x, \hat{x}) \leq D.$$

Now we define the problem of multiple descriptions shown in Figure 1.

We are given three finite reconstruction spaces $\hat{X}_1$, $\hat{X}_2$, and $\hat{X}_0$, together with associated distortion measures

$$d_m : \hat{X} \times \hat{X}_m \rightarrow R, \ m = 0, 1, 2.$$

The distortion measure on $n$-sequences in $\hat{X}_m^n \times \hat{X}_m^n$ is defined by the average per letter distortion

$$d_m(x, \hat{x}_m) = \left(\frac{1}{n}\right) \sum_{i=1}^{n} d_m(x_i, \hat{x}_{m_i}).$$

Definition: We shall say that $(R_1, R_2)$ is an achievable rate for distortion $D = (D_1, D_2, D_0)$ if there exists a sequence, indexed by $n$, of pairs of descriptions $i(x) \in \{1, 2, \ldots, 2^nR_1\}$, $j(x) \in \{1, 2, \ldots, 2^nR_2\}$, and reconstruction functions $\hat{x}_1(i)$, $\hat{x}_2(j)$, $\hat{x}_0(i, j)$ such that, for sufficiently large $n$,

$$E\left(\frac{1}{n}\right) \sum_{i=1}^{n} d_m(X_i, \hat{X}_{m_i}) \leq D_m, \ m = 0, 1, 2.$$

Definition: The rate distortion region $R(D)$ for distortion $D =$
$(D_1,D_2,D_0)$ is the closure of the set of achievable rate pairs $(R_1,R_2)$ inducing distortion $\leq D$. An achievable rate region is any subset of the rate distortion region.

In the next section, we shall exhibit an achievable rate region for the multiple description problem.
3. **Theorem and Construction**

The following achievable rate region for multiple descriptions has an information theoretic characterization. Specializations will be given in later sections.

**Theorem 1**: Let \( X_1, X_2, \ldots \) be a sequence of i.i.d. finite alphabet r.v.'s drawn according to a probability mass function \( p(x) \). Let \( d_i(\cdot, \cdot) \) be bounded. An achievable rate region for distortion \( D = (D_1, D_2, D_0) \) is given by the convex hull of all \((R_1, R_2)\) such that

\[
R_1 > I(X; \hat{X}_1) \\
R_2 > I(X; \hat{X}_2) \\
R_1 + R_2 > I(X; \hat{X}_0, \hat{X}_1, \hat{X}_2) + I(\hat{X}_1; \hat{X}_2)
\]

for some probability mass function \( p(\hat{X}_0, \hat{X}_1, \hat{X}_2) = p(x)p(\hat{X}_0, \hat{X}_1, \hat{X}_2 | x) \) such that

\[
D_1 \geq E d_1(X; \hat{X}_1) \\
D_2 \geq E d_2(X; \hat{X}_2) \\
D_0 \geq E d_0(X; \hat{X}_0).
\]

**Remark**: Since we have been unable to prove in general that the region in Theorem 1 is convex, it may be necessary to convexify this region by time-sharing. For each choice of \( p = p(\hat{X}_0, \hat{X}_1, \hat{X}_2 | x) \) let

\[
\mathcal{V}(p) = (I(X; \hat{X}_1), I(X; \hat{X}_2), I(X; \hat{X}_0, \hat{X}_1, \hat{X}_2) + I(\hat{X}_1; \hat{X}_2), Ed_1(X; \hat{X}_1), Ed_2(X; \hat{X}_2), Ed_0(X; \hat{X}_0)).
\]

Let \( \mathcal{V} \) be the convex hull of \( \mathcal{V}(p) \). (An arbitrary point in the convex hull can be obtained by mixing at most 7 \( \mathcal{V}(p) \)'s.) Let \((R, D) = \)
\((R_1, R_2, R_1 + R_2, D_1, D_2, D_0)\). Then it follows that \((R, D)\) is achievable if there exists a \(V \in \mathcal{V}\) such that

\[
(R, D) > V.
\]  

(3.4)

Remark about method of proof: Let \(p(x)\) be given, and fix a choice of \(p(\hat{x}_1, \hat{x}_2, \hat{x}_0 | x)\). Sequences \(x\) and \(\hat{x}_0\) are said to be jointly typical if the empirical joint composition of \((x, \hat{x}_0)\) is approximately equal to \(p(x, \hat{x}_0)\). A subset \(B_0\) of \(\hat{x}_0^n\) is said to cover \(\bar{x}^n\) if, with high probability, a randomly drawn \(X \in \bar{x}^n\) is jointly typical with some \(\hat{x}_0 \in B_0\). It can be shown that \(2^{nI(\overline{x}, \hat{x}_0)}\)'s are both necessary and sufficient to cover \(\bar{x}^n\). Moreover, if \(\hat{x}_0\) is jointly typical with \(x\), then the joint composition \(p(x, \hat{x}_0)\) guarantees that \(d_0(\overline{x}, \hat{x}_0)\) will be close to \(D_0\).

Two extra ingredients in addition to the notion of covering are found in the proof of Theorem 1. The distribution \(p(x, \hat{x}_1, \hat{x}_2, \hat{x}_0)\) is used to find a covering \((\hat{x}_1, \hat{x}_2)\) of \(\bar{x}^n\) so that the \(\hat{x}_1\)'s individually cover \(\bar{x}^n\), the \(\hat{x}_2\)'s individually cover \(\bar{x}^n\), and the \((\hat{x}_1, \hat{x}_2)\)'s cover \(\bar{x}^n\). Finding the best rate region for the product covering involves a new result on lists. Second, for each \((\hat{x}_1, \hat{x}_2)\), a minimal conditional covering of \(\bar{x}^n\) is found. The conditional \(x_0\) information is to be distributed to the two describers and increases their rates beyond that which is necessary to individually recover \(\hat{x}_1\) and \(\hat{x}_2\) at distortions \(D_1\) and \(D_2\). However, the extra information efficiently adds information to the two descriptions to yield a third description \(\hat{x}_0\) with distortion \(D_0\).

Before proceeding to the proof in Section 5, we give an example in the next section.
4. Example

Let $X_1, X_2, \ldots$ be independent identically distributed normal random variables with mean 0 and variance 1. From rate distortion theory \([1,2]\), we know that $R(D) = (1/2) \log 1/D$, is the rate distortion function for a squared error distortion

$$ d(x, \hat{x}) = (x - \hat{x})^2. \quad (4.1) $$

Thus $(n/2) \log 1/D$ bits are necessary and sufficient to describe a random sequence $(X_1, X_2, \ldots, X_n)$ with expected squared error

$$ E \left\{ \frac{1}{n} \sum_{i=1}^{n} (X_i - \hat{X}_i)^2 \right\} \leq D. \quad (4.2) $$

Consequently, for the simultaneous description problem, an obvious inner bound on the rate region is

$$ R_1 \geq (1/2) \log 1/D_1 $$
$$ R_2 \geq (1/2) \log 1/D_2 $$
$$ R_1 + R_2 \geq (1/2) \log 1/D_0. \quad (4.3) $$

The third inequality follows from the realization that the description rate is $R_1 + R_2$ for the $\hat{X}_0$ reconstruction. Surprisingly, for high distortions $(D_1 + D_2 - D_0 \approx 1)$ these rates are actually achievable. The detailed calculations have been provided to the authors by M. Aref. We examine the region of Theorem 1 for the joint normal distribution on $(\hat{X}, \hat{X}_1, \hat{X}_2, \hat{X}_0)$ obtained from $X \sim N(0,1)$,

$$\hat{X}_1 = \alpha_1 \hat{X}_0 + Z_1,$$
$$\hat{X}_2 = \alpha_2 \hat{X}_0 + Z_2,$$
$$\hat{X}_0 = \alpha_0 X + Z_0, \quad (4.4)$$
where \((Z_0, Z_1, Z_2)\) are jointly normal with mean 0 and covariance matrix

\[
K = \begin{bmatrix}
k_0 & 0 & 0 \\
0 & k_1 & k_{12} \\
0 & k_{12} & k_2
\end{bmatrix},
\]  

(4.5)

and \((Z_0, Z_1, Z_2)\) is independent of \(X\). Optimizing (3.1) in Theorem 1 over \(\alpha_0, \alpha_1, \alpha_2, K\) subject to the distortion constraints

\[
E(X - \hat{X}_1)^2 = D_1 \\
E(X - \hat{X}_2)^2 = D_2 \\
E(X - \hat{X}_0)^2 = D_0
\]

yields the region described below.

Case 1: (High distortion) \(D_1 + D_2 - D_0 \geq 1\).

The following region is achievable:

\[
R_1 > (1/2) \log 1/D_1 \\
R_2 > (1/2) \log 1/D_2 \\
R_1 + R_2 > (1/2) \log 1/D_0 .
\]  

(4.6)

This is trivially optimal by our previous remarks.

Case 2: (Low distortion) \(D_1 + D_2 - D_0 < 1\).

The following region is achievable:

\[
R_1 \geq (1/2) \log 1/D_1 \\
R_2 \geq (1/2) \log 1/D_2 ,
\]  

(4.7)
$$R_1 + R_2 \geq \frac{1}{2} \log \frac{1}{D_0} + \frac{(1-D_0)^2}{(1-D_0)^2 - ((1-D_1)^{1/2}(1-D_2)^{1/2} - (D_1 - D_0)^{1/2}(D_2 - D_0)^{1/2})^2}$$

Apparently, for low distortions, the consequent dependence of the descriptions causes an increase in the total description rate $R_1 + R_2$ beyond the $(1/2) \log 1/D_0$ necessary for independent descriptions. In contrast to Case 1, we do not have a converse establishing Case 2 as the optimal rate region.
5. **Proof of Theorem 1**

Throughout this proof we shall assume that \( \mathbf{X} \) and the reconstruction spaces \( \mathbf{X}_i \) are finite sets.

Before proceeding with the proof, we shall define the idea of joint typicality.

Let \( \{ x^{(1)}, x^{(2)}, \ldots, x^{(k)} \} \) denote a finite collection of discrete random variables with some fixed joint distribution \( p(x^{(1)}, x^{(2)}, \ldots, x^{(k)}) \), \( (x^{(1)}, x^{(2)}, \ldots, x^{(k)}) \in X^{(1)} \times X^{(2)} \times \ldots \times X^{(k)} \). Let \( S \) denote an ordered subset of these r.v.'s, and consider \( n \) independent copies of \( S \). Thus,

\[
\Pr \{ S = s \} = \prod_{i=1}^{n} P \{ S_i = s_i \}, \quad s \in S^n. \tag{5.1}
\]

Let \( N(s; s) \) be the number of indices \( i \in \{1, 2, \ldots, n\} \) such that \( S_i = s \). By the law of large numbers, for any subset \( S \) of random variables and for all \( s \in S \),

\[
\left( \frac{1}{n} \right) N(s; s) \to p(s). \tag{5.2}
\]

Also,

\[
- \frac{1}{n} \log p(S_1, S_2, \ldots, S_n) = - \frac{1}{n} \sum_{i=1}^{n} \log p(S_i) \to H(S). \tag{5.3}
\]

Convergence in (5.2) and (5.3) take place simultaneously with probability one for all \( 2^k \) subsets

\( S \subseteq \{ x^{(1)}, x^{(2)}, \ldots, x^{(k)} \} \).

**Definition:** The set \( T_\varepsilon \) of **strongly \( \varepsilon \)-typical** \( n \)-sequences

\( \{ x^{(1)}, x^{(2)}, \ldots, x^{(k)} \} \) is defined by
\[ T_\varepsilon(x^{(1)}, x^{(2)}, \ldots, x^{(k)}) = \]

\[ \{(x^{(1)}, x^{(2)}, \ldots, x^{(k)}) : |\frac{1}{n} N(s; \overline{s}) - p(s)| < \varepsilon \ \forall \ s \in S \] \]

and \( \forall \ S \subseteq \{x^{(1)}, x^{(2)}, \ldots, x^{(k)}\} \), \hspace{1cm} (5.4)

where \( \| A \| \) is the cardinality of the set \( A \).

Let \( T_\varepsilon(S) \) denote the restriction of \( T_\varepsilon \) to the coordinates corresponding to \( S \). We need the following well known properties of \( T_\varepsilon \). For any \( \varepsilon \to 0 \), there exists a \( \delta(\varepsilon) \to 0 \) as \( \varepsilon \to 0 \), such that for sufficiently large \( n \),

(i) \( \Pr \{ T_\varepsilon(S) \} \geq 1 - \varepsilon, \ \forall \ S \subseteq \{x^{(1)}, \ldots, x^{(k)}\} \), \hspace{1cm} (5.5)

(ii) \( 2^{n(H(S) - \delta)} \leq \| T_\varepsilon(S) \| \leq 2^{n(H(S) + \delta)} \) \hspace{1cm} (5.6)

(iii) \( 2^{-n(H(S) + \delta)} \leq p(\overline{s}) \leq 2^{-n(H(S) - \delta)} \),

for all \( \overline{s} \in T_\varepsilon(S) \), \hspace{1cm} (5.7)

(iv) If \( s_2 \in T_\varepsilon(S_2) \), then the number of jointly \( \varepsilon \)-typical \( s_1 \)'s, denoted by \( \| T_\varepsilon(S_1|s_2) \| \), is bounded by

\[ 2^{n(H(S_1|S_2) - \delta)} \leq \| T_\varepsilon(S_1|s_2) \| \leq 2^{n(H(S_1|S_2) + \delta)} \] \hspace{1cm} (5.8)

Proof of Theorem 1: We shall choose a net of \( \hat{x}_1 \) sequences covering \( \hat{A}^n \) with distortion \( D_1 \), a net of \( \hat{x}_2 \) sequences covering \( \hat{A}^n \) with distortion \( D_2 \), and a conditional net of \( \hat{x}_0 \) sequences (given \((\hat{x}_1, \hat{x}_2))\) covering \( \hat{A}^n \) with distortion \( D_0 \). A random choice of the nets will suffice. Fix a joint probability mass function of the form \( p(x)p(\hat{x}_1, \hat{x}_2, \hat{x}_0|x) \) on \( \hat{A} \times \hat{x}_1 \times \hat{x}_2 \times \hat{x}_0 \). 

-13-
\( \hat{X}_2 \in \hat{X}_0 \) such that
\[
\begin{align*}
E d_1(X, \hat{X}_1) &< D_1 \\
E d_2(X, \hat{X}_2) &< D_2 \\
E d_0(X, \hat{X}_0) &< D_0.
\end{align*}
\]
(5.9)

Choose real numbers \( R'_1, R'_2, \Delta \geq 0 \).

**Random Coding:**

1. Let \( n \)-vectors \( \hat{x}_1(1), \hat{x}_1(2), \ldots, \hat{x}_1(2^{nR'_1}) \) be drawn independently according to a uniform distribution over the set \( T_\epsilon(\hat{x}_1) \) of \( \epsilon \)-typical \( \hat{x}_1 \) \( n \)-vectors. That is, \( P\{\hat{x}_1(i) = \hat{x}_1\} = 1/\|T_\epsilon(\hat{x}_1)\| \), if \( \hat{x}_1 \in T_\epsilon \), and = 0 otherwise.

2. Similarly, let \( 2^{nR'_2} \) \( n \)-vectors \( \hat{x}_2(1), \hat{x}_2(2), \ldots, \hat{x}_2(2^{nR'_2}) \) be drawn i.i.d. according to a uniform distribution over \( T_\epsilon(\hat{x}_2) \).

3. Finally, for every jointly typical \( (\hat{x}_1(i), \hat{x}_2(j)) \) in the above list, let \( \hat{x}_0(i,j,1), \hat{x}_0(i,j,2), \ldots, \hat{x}_0(i,j,2^{n\Delta}) \) be drawn i.i.d. according to a uniform distribution over the set \( T_\epsilon(\hat{x}_0 | \hat{x}_1(i), \hat{x}_2(j)) \) of conditionally \( \epsilon \)-typical \( \hat{x}_0 \) 's, conditioned on \( (\hat{x}_1(i), \hat{x}_2(j)) \).

**Encoding:** Given an \( x \in \hat{X}^n \), find, if possible, a triple \( (i,j,k) \) such that \( (x, \hat{x}_1(i), \hat{x}_2(j), \hat{x}_0(i,j,k)) \) is in the set \( T_\epsilon \) of all jointly typical sequences. If no such \( (i,j,k) \) exists, simply set \( (i,j,k) = (0,0,0) \).

We now divide the description of \( k \in 2^{n\Delta} \) into two parts, \( k = (k_1, k_2), k_1 \in 2^{n\Delta_1}, k_2 \in 2^{n\Delta_2}, \Delta_1 \geq 0, \Delta_2 \geq 0, \Delta_1 + \Delta_2 = \Delta \). We use the obvious notation \( 2^{n\Delta} \) for the set \( \{1,2,\ldots,[2^{n\Delta}]\} \), where \( [t] \) denotes the greatest integer less than or equal to \( t \). We send \( (i,k_1) \) to decoder 1, \( (j,k_2) \) to decoder 2, and consequently send \( (i,j,k) \) to decoder 0. The resulting rates of transmission become

-14-
\[ R_1 = R_1' + \Delta_1 \]  
\[ R_2 = R_2' + \Delta_2 . \]  

(5.10)

We shall prove that a successful encoding of \( x \) achieving distortions \( \leq (D_1, D_2, D_0) \) is accomplished (with high probability by this random net) provided \( n \) is sufficiently large and

\[ R_1' > I(X; \hat{X}_1) \]
\[ R_2' > I(X; \hat{X}_2) \]
\[ R_1' + R_2' - I(\hat{X}_1; \hat{X}_2) > I(X; \hat{X}_1, \hat{X}_2) \]
\[ \Delta_1' \geq 0 \]
\[ \Delta_2' \geq 0 \]
\[ \Delta_1' + \Delta_2' > I(X; \hat{X}_0) \hat{X}_1, \hat{X}_2) . \]  

(5.11)

Theorem 1 then follows from adding these inequalities with the result

\[ R_1 = R_1' + \Delta_1' \geq I(X; \hat{X}_1) \]
\[ R_2 = R_2' + \Delta_2' \geq I(X; \hat{X}_2) \]
\[ R_1 + R_2 = R_1' + R_2' + \Delta_1' + \Delta_2' \]
\[ > I(X; \hat{X}_1, \hat{X}_2) + I(X; \hat{X}_0 | \hat{X}_1, \hat{X}_2) + I(\hat{X}_1; \hat{X}_2) \]
\[ = I(X; \hat{X}_1, \hat{X}_2, \hat{X}_0) + I(\hat{X}_1; \hat{X}_2) , \]  

(5.12)

where the last equality is an information theoretic identity. These are the conditions of Theorem 1.

**Reconstruction:** Decoder 1, given \((i, k_1)\), announces \( \hat{x}_1(i) \) as his reconstruction of \( x \). Similarly decoder 2, given \((j, k_2)\), announces \( \hat{x}_2(j) \).
And decoder 0, given (i,j,k), where k = (k_1,k_2), announces x_0(i,j,k). These reconstructions will be shown to meet the distortion constraints.

**Distortion:** We first note that if (i,j,k) can be found satisfying the encoding step, then the distortions satisfy \( d_m(x, \hat{x}_m) \leq D_m \), \( m = 0,1,2 \), by construction of \( T_\varepsilon \). Let us denote by \( E_i \) the error event that there does not exist (i,j,k) such that \((\hat{x}, \hat{x}_1(i), \hat{x}_2(j), \hat{x}_0(i,j,k)) \in T_\varepsilon \). We wish to show \( P(E) \to 0 \), as \( n \to \infty \).

**Probability of Error:** An error \( E \) will occur if one or more of the following events occurs:

1. \( E_0 : x \notin T_\varepsilon \)
2. \( E_1 : (x, \hat{x}_1(i)) \notin T_\varepsilon \), for all \( i \in 2^{nR_1} \)
3. \( E_2 : (x, \hat{x}_2(j)) \notin T_\varepsilon \), for all \( j \in 2^{nR_2} \)
4. \( E_3 : (x, \hat{x}_1(i), \hat{x}_2(j)) \notin T_\varepsilon \), for all \( (i,j) \in 2^{nR_1} \times 2^{nR_2} \)
5. \( E_4 : \exists (i,j) \in 2^{nR_1} \times 2^{nR_2} \) such that \((x, \hat{x}_1(i), \hat{x}_2(j)) \in T_\varepsilon \), but \((x, \hat{x}_1(i), \hat{x}_2(j), \hat{x}_0(i,j,k)) \notin T_\varepsilon \), for \( k \in 2^{n\Delta} \).

Thus \( E = \cup_{i=0}^{4} E_i \) and the probability of (encoding) error is bounded by

\[
P(E) \leq P(E_0) + \sum_{i=1}^{4} P(E_i \cap E_0^c). \tag{5.13}
\]

Clearly \( P(E_0) \to 0 \), as \( n \to \infty \). Also, it follows from known results in rate distortion theory that

\[
R_1' > I(X; \hat{X}_1) \implies P(E_1 \cap E_0^c) \to 0;
\]
\[
R_2' > I(X; \hat{X}_2) \implies P(E_2 \cap E_0^c) \to 0;
\]
\[
\Delta > I(X; \hat{X}_0 | \hat{X}_1, \hat{X}_2) \implies P(E_4 \cap E_0^c) \to 0. \tag{5.14}
\]
It remains to show that \( P(E_3 \cap E^c_0) \rightarrow 0 \), i.e., we wish to show that there are enough individually typical \( (\hat{X}_1(i), \hat{X}_2(j)) \) so that we can find at least one jointly typical pair. Thus define, for each \( x \in T_E \), the random set
\[
C_x = \{(i,j) \in 2^{nR_1^i} \times 2^{nR_2^j} : (\hat{X}_1(i), \hat{X}_2(j)) \in T_E(\hat{X}_1, \hat{X}_2|x) \}, \tag{5.15}
\]
consisting of those \((i,j)\) pairs for which the random code assignment yields \((X, \hat{X}_1(i), \hat{X}_2(j))\) jointly \(\epsilon\)-typical. It is easily seen that
\[
P(E_3 \cap E^c_0) = P \{ \|C_x\| = 0, x \in T_E \} \leq \max_{x \in T_E} P \{ \|C_x\| = 0 \}. \tag{5.16}
\]
But, for all \( x \in T_E \), \(\epsilon > 1\), we have
\[
P \{ \|C_x\| = 0 \} \leq P \{ \|C_x\| - E[\|C_x\|] \geq \epsilon E[\|C_x\|] \} \leq \text{Var}[\|C_x\|]/\epsilon^2(E[\|C_x\|])^2. \tag{5.17}
\]
Writing \(\|C_x\|\) as the sum of indicator functions, it follows from arguments similar to the proof in [3] that
\[
E[\|C_x\|] \geq 2^{n(R_1^i + R_2^i - H(\hat{X}_1) - H(\hat{X}_2) + H(\hat{X}_1, \hat{X}_2|x) - o(1))} \tag{5.18}
\]
and
\[
\text{Var}[\|C_x\|] \leq 2^{n(R_1^i + R_2^i - H(\hat{X}_1) - H(\hat{X}_2) + H(\hat{X}_1, \hat{X}_2|x) + o(1))}. \tag{5.19}
\]
There, \( P(E_3 \cap E^c_0) \rightarrow 0 \) if
\[
R_1^i + R_2^i > H(\hat{X}_1) + H(\hat{X}_2) - H(\hat{X}_1, \hat{X}_2|x) = I(X; \hat{X}_1, \hat{X}_2) + I(\hat{X}_1; \hat{X}_2). \tag{5.20}
\]
It remains to argue that the expected distortion is asymptotically unaffected when an encoding description \((i,j,k)\) cannot be found, i.e., when the event \(E\) occurs. But \(E^c\) implies \(\lim_{n \to \infty} E \{ d | E^c \} \leq D\), by construction of \(T_e\). Finally,

$$E \, d \leq P(E^c)D + P(E) \, l \, d_{\max},$$

(5.21)

where \(l = (1,1,1)\) and \(d_{\max} = \max_{x \in \hat{X}} d_m(x, \hat{x}_m)\).

\[
\hat{x}_m \in \hat{X}_m, m \in \{0,1,2\}
\]

Here we assume \(d_{\max} < \infty\). Hence \(P(E^c) \to 0\) implies \(\lim_{n \to \infty} E \, d \leq D\), and the theorem is proved.
6. Function Distortion Measures

Suppose that receivers 0, 1, and 2 wish only to know \( f_0(x) \), \( f_1(x) \) and \( f_2(x) \) respectively. Here we have a simpler rate region and a converse.

Let \( \hat{x}_0 = f_0(X) \), \( \hat{x}_1 = f_1(X) \), and \( \hat{x}_2 = f_2(X) \). Then \( (R_1,R_2) \) allows recovery of \( f_1 \) by 1, \( f_2 \) by 2, and \( f_0 \) by 0 if and only if

\[
R_1 > I(X;\hat{x}_1) = H(f_1(X)) \\
R_2 > I(X;\hat{x}_2) = H(f_2(X)) \\
R_1 + R_2 > I(f_1(X),f_2(X),f_0(X);X) + I(f_1(X);f_2(X)) = H(f_1(X)) + H(f_2(X)) + H(f_0(X)|f_1(X),f_2(X)).
\]

**Proof:** Achievability is immediate from Theorem 1.

For the converse, let there be maps

\[
i: \mathcal{X}^n \rightarrow 2^{nR_1},
\]

\[
j: \mathcal{X}^n \rightarrow 2^{nR_2},
\]

and set

\[
I = i(X), \quad J = j(X).
\]

Also given are reconstruction functions \( \hat{x}_0(\cdot), \hat{x}_1(\cdot), \hat{x}_2(\cdot) \), with the property

\[
P(\hat{x}_0(I(X),J(X)) \neq f_0(X)) \leq \varepsilon,
\]

\[
P(\hat{x}_1(I(X)) \neq f_1(X)) \leq \varepsilon,
\]

\[
P(\hat{x}_2(J(X)) \neq f_2(X)) \leq \varepsilon,
\]

where

\[
f_i(x) = (f_i(x_1), \ldots, f_i(x_n)), \quad i = 0,1,2.
\]
We shall show that the rates $R_1, R_2$ must satisfy (7.1), (7.2), and (7.3).

Inequalities (7.1) and (7.2) follow from Shannon. To prove (7.3) we first note that

$$nR_1 + nR_2 \geq H(I) + H(J)$$

$$= H(I, f_1(X)) + H(J, f_2(X)) - H(f_1(X)|I)$$

$$- H(f_2(X)|J).$$

If follows from (7.4) and Fano's inequality that

$$H(f_1(X)|I) \leq (H(I) \epsilon + h(\epsilon)) \leq n \delta_n$$

$$H(f_2(X)|J) \leq (H(J) \epsilon + h(\epsilon)) \leq n \delta_n,$$

where $\delta_n \to 0$ as $\epsilon \to 0$. Therefore

$$nR_1 + nR_2 \geq nH(f_1(X)) + nH(f_2(X)) + H(I|f_1(X))$$

$$+ H(J|f_2(X)) - 2n \delta_n$$

$$\geq nH(f_1(X)) + nH(f_2(X)) + H(I, J, f_0(X)|f_1(X), f_2(X))$$

$$- H(f_0(X)|I, J, f_1(X), f_2(X)) - 2n \delta_n.$$

Again, by (7.4) and Fano's inequality, we have

$$H(f_0(X)|I, J) \leq n \delta_n.$$

Thus

$$nR_1 + nR_2 \geq nH(f_1(X)) + nH(f_2(X)) + nH(f_0(X)|f_1(X), f_2(X))$$

$$+ H(I, J|f_1(X), f_2(X), f_0(X)) - 3n \delta_n. \quad (7.5)$$

Finally, by dropping the fourth term in (7.5), inequality (7.3) follows.
7. **Robust Descriptions**

We shall now describe a problem concerning robust descriptions that seems at first to be unrelated to Theorem 1.

Suppose that we wish to describe a source \( \mathbf{x} \in \mathcal{X}^n \) by an index \( i \in 2^{nR} \) in such a manner that the description \( \hat{x}(i) \) is good simultaneously for several given distortion measures. (See Figure 7.1) For example, we may wish to describe a Gaussian source so that it can be recovered with low mean squared error distortion and low absolute error distortion. Or, for a Bernoulli source \( \mathbf{X} \), we might wish \( \hat{X}_1 \) to agree with \( \mathbf{X} \) on the zeros of \( \mathbf{X} \), and for \( \hat{X}_2 \) to agree with \( \mathbf{X} \) on the ones of \( \mathbf{X} \). Incidentally, this does not guarantee that \( \mathbf{X} \) can be recovered perfectly from \( \hat{X}_1 \) and \( \hat{X}_2 \).

\[
\begin{array}{c}
\mathbf{X} \\
(\hat{x}(i) \in 2^{nR})
\end{array} \quad \begin{array}{ccc}
\hat{X}_1(i) : D_1 \\
\hat{X}_2(i) : D_2 \\
\vdots \\
\hat{X}_m(i) : D_m
\end{array}
\]

**Figure 7.1**

Robust Description

A more precise specification of the problem is as follows. Given is a source \( \{X_i\}_{i=1}^\infty \) where \( X_1, X_2, \ldots \) are i.i.d. random variables drawn according to a probability mass function \( p(x), x \in \mathcal{X} \). Also given are \( m \) distortion functions \( d_i(x, \hat{x}_i) \) defined on \( \mathcal{X} \times \hat{\mathcal{X}} \), \( i = 1, 2, \ldots, m \). Distortions \( D = (D_1, D_2, \ldots, D_m) \) are to be achieved. What rate \( R(D) \) is necessary? The solution to this
problem is given in the following theorem. A version of this theorem can
be found in Berger [2].

**Theorem 2:**

\[
R(D) = \min I(X; \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_m), \tag{7.1}
\]

where the minimum is over all \( p(\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m|X) \) satisfying

\[
E d_i(X; \hat{X}_i) \leq D_i, \quad i = 1, 2, \ldots, m. \tag{7.2}
\]

**Remark:** As mentioned before, it appears that a description that can be
used to make no mistakes on the 0's of \( X \) and can also be used to make
no mistakes on the 1's of \( X \) must necessarily make no mistakes whatsoever.
This would require rate \( R = 1 \), but since \( R(D_1, D_2) < 1 \), for all \( D_1, D_2 < 1/2 \),
we see that the above idea is mistaken.

**Remark:** For \( m = 2 \), Theorem 2 is a special case of Theorem 1 obtained
by constraining \( R_2 = 0 \).

We give some examples before proceeding with the proof.

**Example:** Let \( X \) be Bernoulli with parameter 1/2. Let \( d_1 \) and \( d_2 \)
be given by

\[
\begin{array}{c|cc}
 & 0 & 1 \\
\hline
0 & 0 & \infty \\
1 & 1 & 0 \\
\end{array}
\]

\[
\begin{array}{c|cc}
 & 0 & 1 \\
\hline
0 & 0 & 1 \\
1 & \infty & 0 \\
\end{array}
\] \tag{7.3}

Thus \( d_1 \) allows no errors on the 0's of \( X \), and \( d_2 \) allows no errors
on the 1's of \( X \). The optimizing distribution \( p(x_1, x_2|X) \) is given by
inspection of Figure 7.2.

-22-
The resulting rate is

\[ R(D_1, D_2) = I(X; \hat{X}_1, \hat{X}_2) \]

\[ = h(D_1/2, D_2/2, (D_1 + D_2)/2) - (1/2) h(D_1) - 1/2 h(D_2). \]  

**Remark:** It may happen that all distributions satisfying the distortion constraint (7.2) correspond to a Markov chain \( X \rightarrow \hat{X}_1 \rightarrow \hat{X}_2 \) in that order. In such a case, we have

\[ I(X; \hat{X}_1, \hat{X}_2) = I(X; \hat{X}_1). \]  

Thus only the first distortion constraint is active, and a good description for 1 and 2 is given by the best description for 1 alone.

**Proof of Theorem 2:** Achievability follows immediately once the encoding is specified. Fix a choice of \( p(x_1, x_2, \ldots, x_m | x) \) satisfying distortion constraints (7.2). This induces a distribution \( p(x_1, x_2, \ldots, x_m) \). Choose \( 2^{nR} \) vectors \( \hat{X} = (\hat{x}_1, \hat{x}_2, \ldots, \hat{x}_m) \) independently drawn according to

\[ \prod_{i=1}^{n} p(\hat{x}_{1i}, \hat{x}_{2i}, \ldots, \hat{x}_{mi}). \]  

Designate these vectors \( X(k), k \in 2^{nR} \). By the usual arguments, these vectors will "cover" \( \hat{X} \) if
\[ R > I(X; \hat{X}_1, \hat{X}_2, \ldots, \hat{X}_m). \]  

(7.6)

Thus for large \( n \), with probability arbitrarily close to 1, there will be an index \( k \in 2^{nR} \) such that

\[ d_i(X, \hat{X}_i(k)) \leq D_i, \quad i = 1, 2, \ldots, m. \]

The converse is precisely along the lines of the converse for Shannon rate distortion theory.
8. Conclusions

It is a pleasant surprise that two descriptions are "twice as good" as one when the two descriptions need not be individually very good. This conclusion holds (for high distortions) for the Gaussian example in Section 4 and also holds for the Hamming distortion on Bernoulli sequences, an example not given here. Of course when we demand low distortions for the individual descriptions the peculiar tension of this problem reveals itself. When the distortion constraints are sufficiently severe, two descriptions are no better than one because they are in fact the same description.

Acknowledgement

Beyond our obvious debt to the formulators of this problem, we wish to thank M. Aref for working out the Gaussian example and C. Heegard for working out the Bernoulli case with Hamming distortion.
REFERENCES

