THE BELL SYSTEM
TECHNICAL JOURNAL

Volume 47 July–August 1968 Number 6

The Transmission Distortion of a Source as a Function of the Encoding Block Length*

By R. J. PILC

(Manuscript received December 15, 1967)

This paper is concerned with the transmission of a discrete, independent letter information source over a discrete channel. A distortion function is defined between source output letters and decoder output letters and is used to measure the performance of the system for each transmission. The coding block length is introduced as a variable and its influence upon the minimum attainable transmission distortion is investigated.

The lower bound to transmission distortion is found to converge to the distortion level d_C *(C is the channel capacity) algebraically as* a/n. *The nonnegative coefficient* a *is a function of both the source and channel statistics, which are interrelated in such a way as to suggest the utility of this coefficient as a measure of "mismatch" between source and channel, the larger the mismatch the slower the approach of the lower bound to the asymptote* d_C. *For noiseless channels* $a = \infty$ *and for this case the lower bound is shown to converge to* d_C *as* $a_1(\ln n)/n$.

For noisy channels the upper bound to transmission distortion is found to converge to the asymptote d_C *algebraically as* $b[(\ln n)/n]^{\frac{1}{2}}$. *For noiseless channels, the upper bound converges to* d_C *as* $a_1(\ln n)/n$.

---

* The material presented in this paper is based upon the author's thesis, "Coding Theorems for Discrete Source-Channel Pairs," presented to the Massachusetts Institute of Technology in November 1966 in partial fulfillment of the requirements for the degree of Doctor of Philosophy.

I. INTRODUCTION

By now the results originally obtained by Shannon[1] relating reliability and channel capacity are well known. Roughly speaking, they state that perfect transmission can be achieved if, and only if, the capacity of the channel in the transmission link is greater than the information content of the source. For amplitude and time discrete sources the information content is the entropy of the source, but for amplitude continuous sources the entropy and the information content are not the same since the information content is infinite. This, of course, implies that perfect transmission of amplitude continuous sources, or discrete sources with an entropy that is "too large," is impossible with a given finite capacity channel. Yet this is just the situation that is often presented to the communication engineer who must then try to reduce the average distortion to the lowest possible, or practicable, level.

For communication systems in which the capacity of the channel is not sufficient to allow perfect transmission, there are two obvious questions to ask:

(*i*) How small can the average distortion be made if any transmission strategy at all is allowed?
(*ii*) How much does the system complexity, or cost, increase when you are required to get "closer" to this minimum?

To answer the first question, Shannon generalized his results in a later paper[2] in which the channel requirements are found that are necessary and sufficient to allow transmission at a given level of distortion, or a given error rate. It is our purpose here to consider the second question. We use the coding block length to measure the complexity of the system, and study the behavior of the minimum attainable transmission distortion as the block length is increased.

In the work we restrict our attention to sources and channels that are discrete in amplitude and time, and that are constant and memoryless. This means that successive events are independent and are governed by the same probability distributions. The encoder is a block encoder that we describe later in this section. To measure the distortion in the system, we introduce a nonnegative function $d(w,z)$ which gives the distortion in the event letter z is presented to the user at the decoder output when letter w was transmitted. Normally, this function would be specified by the user of the system to reflect how undesirable any particular misinterpretation of the source output is to him. We will assume that the distortion between two sequences of letters is the averaged sum of the composing letter distortions.

Shannon's theory associates with each source and distortion function a rate-distortion curve which expresses the minimum attainable transmission distortion in terms of the maximum allowable mutual information in the system. Associated with each point (d_R, R) on the rate-distortion curve is a particular set of transition probabilities, called the "test channel," which has the significance that among all channels that transmit the given source with distortion d_R or less, it operates at the lowest transmission rate, R. Equivalently, the test channel is that channel which yields the lowest distortion d_R among those that transmit information from the source at a rate R or less. It is in this sense the cheapest channel one could use and meet a distortion criterion. The rate R can also be interpreted as the equivalent information content of the source when a distortion d_R is tolerable.

That the rate-distortion curve gives the channel capacity sufficient to allow a prescribed performance is shown by Shannon through the intermediate step of proving that the rate-distortion curve actually expresses the entropy and resultant distortion in the "best" discrete representation of an output sequence from the original source. This discrete representation can then be transmitted with no further distortion, if its entropy is less than the channel capacity, by the use of suitable channel coding techniques.

Shannon has found the rate-distortion curves for many discrete sources and an explicit expression for this curve for time discrete gaussian sources. These results, together with Shannon's work with vector sources, were used to get rate-distortion curves for gaussian random processes.[3, 4] Bounds to the rate-distortion curve for non-gaussian sources have also been obtained.[5, 6]

However, all of the rate-distortion results derived for both continuous and discrete sources are limiting results, that is, they can be approached in general only when arbitrarily complex operations on very long sequences of source output are allowed before transmitting the "message" through a correspondingly large use of the channel. T. Goblick was the first to study the rate of approach to these limiting results as the source output block length increases, but limited his work to source representation or source encoding, with a deterministic map between the source and its representation.[7] Our work includes a noisy channel, or probabilistic function, between the source and user.

A performance curve $d(n)$ will be introduced for each source-channel pair as the minimum possible average distortion obtainable using a modulator that encodes a string of n successive source outputs into an input signal acceptable by a channel composed of n uses of the original channel. For a source with the rate-distortion curve of Fig. 1 and a channel with capacity C, the performance curve might look like the one shown in Fig. 2.

From Shannon's theory it is known that the performance curve starts at d_0, the zero-rate distortion, and decreases to asymptotically approach d_C, the distortion corresonding to the information rate C on the rate-distortion curve. The curve, of course, has meaning only for integral values of n. Not all modulators and decoders provide a distortion curve that approaches d_C for large n, but this curve obviously must lie above the performance curve which alternately could have been defined as the lower envelope to the set of distortion curves corresponding to all encoder-decoder pairs.

II. THE LOWER BOUND

Upper and lower bounds to the performance curve have been derived.[8] We present the lower bound in the first part of this paper, and the upper bound in Sections XI through XVII. Most of our effort concerning the lower bound was directed toward finding information about the rate of approach of the performance curve to its asymptote. In particular, we tried to relate the source and channel statistics, as well as the method of encoding that is used, to the rate of approach of $d(n)$ to d_C.

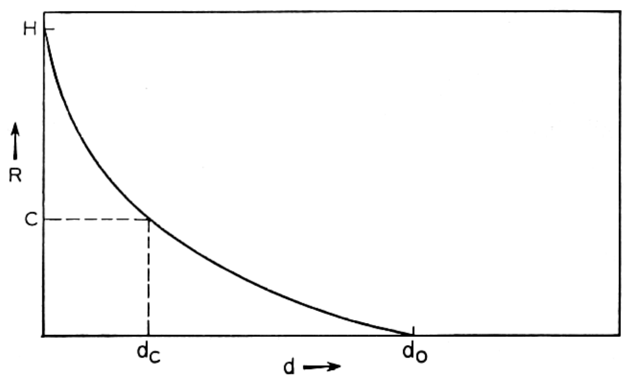

Fig. 1 — The rate distortion curve for S.

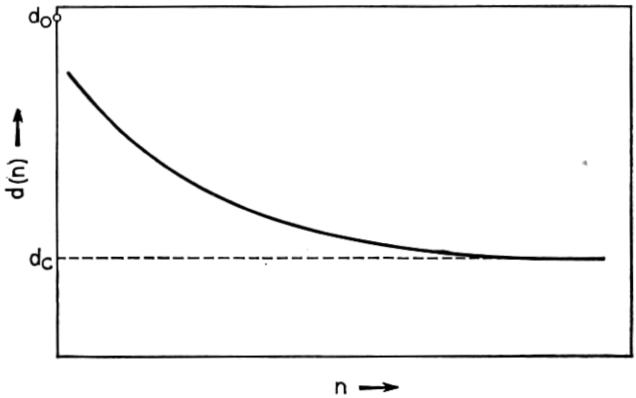

Fig. 2 — The performance curve for $\mathcal{S}$ and $\mathcal{C}$.

Concerning this rate of approach, several interesting situations are known to exist. For one, there are some source-channel pairs for which the minimum attainable transmission distortion is independent of the encoding block length, with the consequence that it is possible to attain the distortion level d_C with a coding block length of one. One example of such a pair is a binary symmetric source (equally likely binary letters with $d(i,j) = 1 - \delta_{ij}$, $i,j = 1,2$) used with a binary symmetric channel, where the optimum encoder is a direct connection. Another example is a gaussian source used with an additive gaussian noise channel, where the optimum encoder is simply an amplifier.[9]

When the source-channel pair is such that the minimum attainable distortion is independent of the coding block length we shall say that the source and channel are "matched." For the more common situation wherein the minimum attainable transmission distortion decreases with increasing encoding block length to asymptotically approach the distortion level d_C, we say that there is a "mismatch" between the source and channel, and suggest as a measure of this mismatch the "slowness" of the approach of the distortion to d_C.

Another interesting situation occurs when there is a choice of using one of several channels of different capacity. Although the channel of highest capacity would be the best choice when one is willing to use infinite block length coding, it might not be the best choice with finite length coding. This could easily happen if the high capacity channel were very much more mismatched to the source than some lower capacity channel.

III. SYSTEM MODEL

Figure 3 is a detailed illustration of the transmission system that we work with. The source $\mathcal{S}$ produces a sequence of letters $\boldsymbol{\omega} = \omega_1, \omega_2, \cdots, \omega_n$, each chosen from the alphabet $W = \{w_1, \cdots, w_H\}$, which is mapped by the encoder into a sequence of channel input letters $\boldsymbol{\xi} = \xi_1, \xi_2, \cdots, \xi_n$, each a member of $X = \{x_1, \cdots, x_K\}$. The channel then transforms the channel input word $\boldsymbol{\xi}$ into a sequence of channel output letters $\boldsymbol{\eta} = \eta_1, \eta_2, \cdots, \eta_n$ which are members of $Y = \{y_1, \cdots, y_L\}$, and $\boldsymbol{\eta}$ in turn is decoded by the receiver into a sequence $\boldsymbol{\zeta} = \zeta_1, \zeta_2, \cdots, \zeta_n$ of letters from the decoding space $Z = \{z_1, \cdots, z_J\}$.

The source and channel are both assumed to be constant and memoryless; therefore, successive events on each are independent and governed by the same probability distributions. In particular we have

$$p_{\omega}(\mathbf{w}) = \prod_{m=1}^{n} p_{\omega_m}(w^m)$$

$$p_{\eta|\xi}(\mathbf{y} \mid \mathbf{x}) = \prod_{m=1}^{n} p_{\eta_m|\xi_m}(y^m \mid x^m),$$

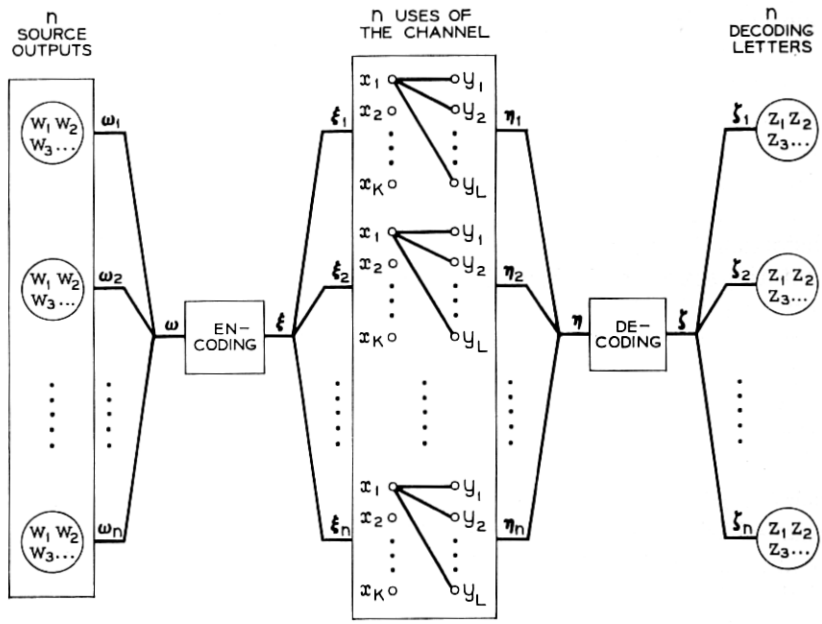

Fig. 3 — Block diagram of the encoding and decoding.

where the superscript on w^m, x^m, y^m is used to denote the m'th letter in the n-letter words **w**, **x**, **y** respectively, and is not to be confused with the particular letters w_m, x_m, and y_m in the alphabets W, X, and Y. The subscripts on the probability distribution are hereafter dropped whenever no confusion will occur.

The distortion in the system when the source word **w** is transmitted but received as **z** is taken to be the normalized sum of the n letter distortions, or

$$d(\mathbf{w}, \mathbf{z}) = \frac{1}{n} \sum_{m=1}^{n} d(w^m, z^m). \tag{1}$$

Finally, although we have set up the problem so that a sequence of n source letters is transmitted as a sequence of n channel letters, different block lengths at the source output and channel input can be allowed by considering a new source and channel that are products of the original ones, with the order of each product adjusted to obtain the desired block length ratio n_s/n_c.

IV. THE SPHERE PACKING ARGUMENT

A generalization of the sphere-packing concept is used to derive the lower bound. We assume the coding block length is n and derive a bound conditioned on the event that a particular source word **w** has occurred at the source output. We further assume that the channel input word **x** is used to transmit **w**, but delay the selection of **x** until the end of the derivation when the result is optimized over all possible choices. The total lower bound to distortion is found by averaging this conditioned lower bound over all source words in W^n. The asymptotic form of this bound is studied in detail and from it a measure of mismatch between the source and channel is defined.

The idea involved can be described with the following simple, but poor, bound which is subsequently improved. Remembering that the source word **w** is assumed transmitted by the channel input word **x**, we list all possible channel output words, **y**, ordered in decreasing conditional probability $p(\mathbf{y} \mid \mathbf{x})$, and pair with each the decoder output word $\mathbf{z}(\mathbf{y})$ to which it is decoded by the optimum decoder. The resulting (conditional) distortion,

$$d(\mathbf{w}) = \sum_{Y^n} p(\mathbf{y} \mid \mathbf{x})\, d[\mathbf{w}, \mathbf{z}(\mathbf{y})], \tag{2}$$

is seen to equal the sum of conditional probability-distortion products on this list. If the set of distortion values that appear on this list is now rearranged (with the list of conditional probabilities fixed) to be ordered according to increasing distortion values, the resulting sum of conditional probability-distortion products must be smaller than, or at most equal to, the sum in equation 2. It therefore provides a lower bound.

The improved lower bound uses the same sort of orderings and rearrangements but includes a probability function, $f(\mathbf{y})$, in the ordering of the channel output words. This function is defined over the set of channel output words, Y^n, and is later chosen to optimize the result. The channel output words are now ordered according to increasing values of the information difference $I(\mathbf{x}, \mathbf{y}) = (1/n) \ln [f(\mathbf{y})/p(\mathbf{y} \mid \mathbf{x})]$ and each is again paired with the decoder output word $\mathbf{z}(\mathbf{y})$ to which it is decoded by the optimum decoder.

The rearrangement of decoder output words is also slightly different. To describe this rearrangement we visualize each channel output word, $\mathbf{y}$, as "occupying" an interval of width $f(\mathbf{y})$ along the line $[0, 1]$. The decoder output word, $\mathbf{z}(\mathbf{y})$, that is paired with a particular channel output word $\mathbf{y}$ is also viewed as occupying the same region along $[0, 1]$ as $\mathbf{y}$, but, because any particular word $\mathbf{z}_o$ might be the decoding result of several channel output words, the region along $[0, 1]$ occupied by $\mathbf{z}_o$ could be a set of separated intervals. The rearrangement of decoder output words is this time a rearrangement of occupancies in $[0, 1]$ toward the desired configuration wherein the decoder words are ordered in increasing distortion along this line, and each occupies the same total width in $[0, 1]$ as it did before the ordering. Thus two monotone nondecreasing functions can be defined along the line $[0, 1]$; one, $I(h)$, giving the information difference $I(\mathbf{x}, \mathbf{y})$ at the point h, $0 \leqq h \leqq 1$, and the other, $d(h)$, giving the distortion $d(\mathbf{w}, \mathbf{z})$ at h. The first theorem presents a lower bound to the single word distortion in terms of these two functions.

Theorem 1: The average transmission distortion, $d(\mathbf{w})$, *conditioned on the occurrence of the source word* $\mathbf{w}$ *and its transmission using the channel input word* $\mathbf{x}$, *satisfies*

$$d(\mathbf{w}) \geqq \int_0^1 d(h) e^{-nI(h)} \, dh. \tag{3}$$

Proof: Figure 4 is used to help prove the inequality. The distortion resulting from optimum decoding is given by equation 2; the conditional probability-distortion products on the previous list *before* rearrangement of the decoder output words. For convenience this is

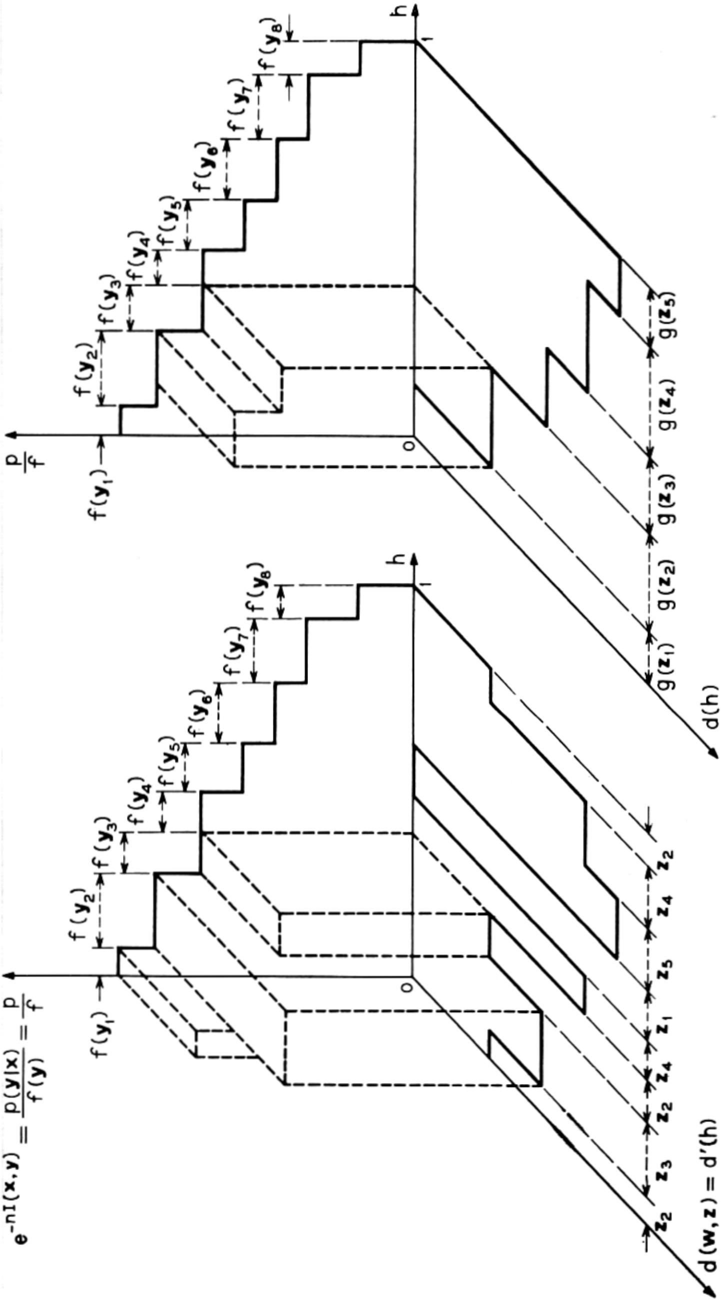

Fig. 4 — The geometry for theorem 1.

rewritten here as

$$d(\mathbf{w}) = \sum_{Y^n} d[\mathbf{w}, \mathbf{z}(\mathbf{y})]\left[\frac{p(\mathbf{y} \mid \mathbf{x})}{f(\mathbf{y})}\right]f(\mathbf{y}) \tag{4}$$

which can be seen equal to the "volume" in Fig. 4a enclosed by the two "amplitude functions" d' and p/f and the "width measure" f.

The rearrangement of the decoder output words to obtain the monotone function $d(h)$ from $d'(h)$ can be accomplished by a sequence of interchanges of the following type. We consider any two points in $0 \leqq h \leqq 1$, say h_1 and h_2, for which $d'(h_2) \leqq d'(h_1)$ and $p/f(h_2) \leqq p/f(h_1)$. If we consider an interval Δh around each point in which both amplitude functions are single valued and interchange amplitude values of d' in the two intervals, we effect a volume transformation that decreases (or leaves unchanged) the total volume since

$$\begin{aligned}
&\text{initial volume—final volume} \\
&= \left[d'(h_1)\frac{p}{f}(h_1) + d'(h_2)\frac{p}{f}(h_2)\right]\Delta h \\
&\quad - \left[d'(h_2)\frac{p}{f}(h_1) + d'(h_1)\frac{p}{f}(h_2)\right]\Delta h \\
&= [d'(h_1) - d'(h_2)]\left[\frac{p}{f}(h_1) - \frac{p}{f}(h_2)\right]\Delta h \\
&\geqq 0.
\end{aligned}$$

Volume interchanges of this type are repeated until the desired monotonic function $d(h)$ is obtained. The resulting volume configuration is then as shown in Fig. 4b. As each interchange of Δh width volumes decreases the total volume, or leaves it unchanged, the total volume in Fig. 4b is certainly no larger than that in Fig. 4a. We need now only notice that $p/f(h) = \exp - nI(h)$ to recognize that the integral in equation 3 is equal to the volume in Fig. 4b, and, therefore, to establish the inequality claimed in the theorem.

To be sure, the construction in Fig. 4b, and the calculation of the lower bound in equation 2 requires some knowledge of the structure of the optimum decoder. Fortunately, this knowledge is minimal; it is only the total width along [0, 1] occupied by each member, $\mathbf{z}$, of the decoding space Z^n. We refer to this occupancy as the "size" of the decoding set for $\mathbf{z}$ and denote it by $g(\mathbf{z})$.

From the construction of the lower bound volume in Fig. 4b, we see that

$$g(\mathbf{z}) = \sum_{Y(\mathbf{z})} f(\mathbf{y})$$

where $Y(z)$ is the set of channel output words that are decoded into $\mathbf{z}$ by the optimum decoder. Indeed, if we assume unique decoding by the optimum decoder we have

$$\sum_{Z^n} g(\mathbf{z}) = \sum_{Z^n} \sum_{Y(\mathbf{z})} f(\mathbf{y}) = \sum_{Y^n} f(\mathbf{y}) = 1,$$

or that $g(\mathbf{z})$ is also a probability function. Even this function, though, is unknown in the general case or at least is impractical to calculate. The idea of the lower bound development, therefore, is to retain this unknown probability function for the present and subsequently replace it with another such function which minimizes the final lower bound expression. Within this step an approximation involving the form of $g(\mathbf{z})$ is required which is detailed in Section 6.2.

V. FURTHER EVALUATION OF THE LOWER BOUND IN THEOREM 1

The integral in equation 3 can be simplified if we suppress the intermediate variable h and relate the variables d and I directly. The pairings of d and I through a common value of h, $d(h) = I(h)$, does not by itself define a function because several different values of d could be paired with a given value of I, and vice versa. However, we will use the properties that exist among these pairs to define a distortion function $d(I)$ which has the property that for any I, the dependent variable d is at least as small as the smallest $d(h)$ among the pairs that have $I(h) = I$.

To do this, we reinterpret the monotone nondecreasing functions $d(h)$ and $I(h)$. First, we view the distortion $d(\mathbf{w}, \mathbf{z})$ as a random variable on Z^n governed by $g(\mathbf{z})$. Its cumulative distribution function

$$G(d) = \sum_{\substack{\mathbf{z} \ni \\ d(\mathbf{w},\mathbf{z}) \leq d}} g(\mathbf{z}) \tag{5}$$

is then seen to be the "inverse" of $d(h)$. (Strictly speaking, the inverse of a staircase function does not exist, so the term inverse is used here only as an aid in relating $d(h)$ and $G(d)$ pictorially.) In a similar way we also view the information difference $I(\mathbf{x}, \mathbf{y})$ as a random variable on Y^n governed by $f(\mathbf{y})$. Its cumulative distribution function is given by

$$F_1(I) = \sum_{\substack{\mathbf{y} \ni \\ I(\mathbf{x},\mathbf{y}) \leq I}} f(\mathbf{y}), \tag{6}$$

or the "inverse" of $I(h)$. The desired function $d(I)$ can now be defined in terms of $G(d)$ and $F_1(I)$ by relating to any information difference value I the distortion value that satisfies

$$F_1(I^-) = G(d). \tag{7}$$

The following geometric interpretation of $d(I)$ might be helpful. If each size, or "volume," $g(\mathbf{z})$ of the decoding sets is successively placed about the volume $g(\mathbf{z}_1)$ of the decoded word with minimum distortion $d(\mathbf{w}, \mathbf{z}_1)$, and each size, or "volume," $f(\mathbf{y})$ of the channel output words successively placed about the volume $f(\mathbf{y}_1)$ of the channel output word with minimum information difference $I(\mathbf{x}, \mathbf{y}_1)$, the total volume included by a point in the first construction at a distortion "radius" d is $G(d)$ and that included by a point in the second construction at an information difference "radius" I is $F_1(I)$. The function $d(I)$ then gives (except for edge effects) the correspondence between the radii that include the same volume in both geometrical constructions. Figure 5a illustrates the construction of $d(I)$ through the chain $I \rightarrow F_1(I^-) = G(d) \rightarrow d$.

It is convenient at this point to introduce a second random variable of information difference; one which is governed by $p(\mathbf{y} \mid \mathbf{x})$ rather than $f(\mathbf{y})$. Its cumulative distribution function is

$$F_2(I) = \sum_{\substack{\mathbf{y}; \\ I(\mathbf{x},\mathbf{y}) \leq I}} p(\mathbf{y} \mid \mathbf{x}). \tag{8}$$

To distinguish the two information difference variables, we will denote by I_1 the variable that has the distribution function in equation 6 and by I_2 the variable that has the distribution function in equation 8.

We are now in a position to rewrite the bound in Theorem 1 in terms of functions that involve only d and I. The distortion function $d(I)$ has been constructed to lower bound all $d(h)$ with $I(h) = I$, thus we can replace $d(h)$ in equation 3 with $d[I(h)]$. As this substitution replaces $d(h)$ with a distortion function that is single valued over subintervals of [0,1] in which I is a constant, we can perform the integration in equation 3 by simply multiplying the integrand in each such constant I interval by the interval width, $dF_1(I)$, and summing. Therefore, we can continue the inequality in equation 3 with

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} d(I) \exp(-nI)\, dF_1(I),$$

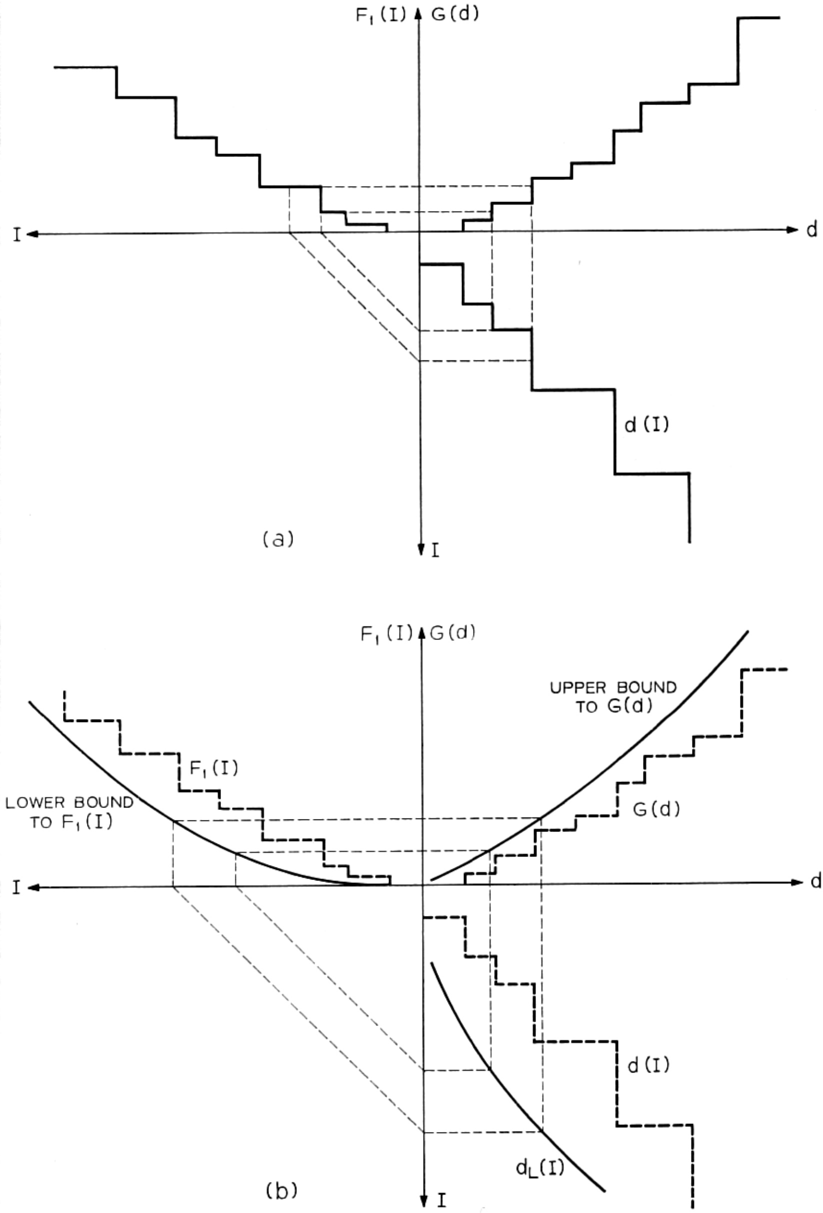

Fig. 5 — The construction of (a) $d(I)$ and (b) $d_L(I)$.

which, upon using $p(\mathbf{y} \mid \mathbf{x}) = \exp(-nI)f(\mathbf{y})$, establishes the lower bound in the next theorem.

Theorem 2: The average transmission distortion, d(**w**), *conditioned on the occurrence of the source word* **w** *and its transmission using the channel input word* **x**, *satisfies*

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} d(I)\, dF_2(I). \tag{9}$$

VI. AN ESTIMATE OF THE FUNCTION $d(I)$

6.1 *The Random Variables* I_1 *and* I_2

To obtain an estimate of $d(I)$ we require an estimate of the two distribution functions, $G(d)$ and $F_1(I)$, from which $d(I)$ was defined. We first focus on $F_1(I)$ and the random variable I_1. Since the lower bounds in Theorems 1 and 2 can be derived for any choice of $f(\mathbf{y})$, we choose a form of $f(\mathbf{y})$ that simplifies the following arguments. We specify that $f(\mathbf{y})$ factors as

$$f(\mathbf{y}) = \prod_{m=1}^{n} f(y^m). \tag{10}$$

One consequence of this assumed form is that the information difference $I(\mathbf{x}, \mathbf{y})$ is given as a sum of n letter information differences:

$$I(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{m=1}^{n} \ln \frac{f(y^m)}{p(y^m \mid x^m)} = \frac{1}{n} \sum_{m=1}^{n} I(x^m, y^m). \tag{11}$$

Among these n letter information differences, however, there are different types, depending on the corresponding transmitted letter x^m in **x**. To separate these, we introduce the vector **c** to denote the letter composition of the channel input word **x**, letting $\mathbf{c} = c_1, c_2, \cdots, c_K$ when there are nc_1 appearances of the letter x_1 in **x**, nc_2 appearances of x_2 in **x**, and so on. Thus we can write the information difference in equation 10 as

$$I(\mathbf{x}, \mathbf{y}) = \frac{1}{n} \sum_{k=1}^{K} \sum_{r=1}^{nc_k} I_{kr} \tag{12}$$

in which I_{kr} is used to denote the information difference between the r'th appearance of the letter x_k in **x** and the corresponding letter in **y**. The interpretation of the I_{kr} as letter information difference random variables on Y governed by the letter probability function $f(y)$ can now be seen to be consistent with the previous interpretation of I_1

as a word information difference random variable on Y^n governed by $f(\mathbf{y})$. Using the abbreviations

$$f(y_l) = f_l$$
$$p(y_l \mid x_k) = p_{kl} ,$$

the probability distribution function of I_{kr} can be written as

$$P_{1,I_{kr}}\left[\ln \frac{f_l}{p_{kl}}\right] = f_l \; ; \qquad 1 \leqq r \leqq nc_k \; ; \qquad 1 \leqq k \leqq K. \tag{13}$$

What this has accomplished is to cast I_1 as the sum of n independent random variables, a step that enables us to use large number laws to estimate $F_1(I)$.[10–13]

In an almost identical way, the random variable I_2 can be cast as a sum of n independent random variables. This can be done if we associate with the variable I_{kr} the probability distribution function

$$P_{2,I_{kr}}\left[\ln \frac{f_l}{p_{kl}}\right] = p_{kl} \; ; \qquad 1 \leqq r \leqq nc_k \; ; \qquad 1 \leqq k \leqq K \tag{14}$$

instead of that in equation 13. With this distribution the word information difference variable $I(\mathbf{x}, \mathbf{y})$ in equation 12 can be seen to be governed by the probability function $p(\mathbf{y} \mid \mathbf{x})$, therefore, it is equal to the random variable I_2 .

6.2 *The Random Variable* d

In the work so far, the function $g(\mathbf{z})$ is that probability function induced on Z^n by $f(\mathbf{y})$ through the optimum decoder function and cannot, therefore, be freely chosen once $f(\mathbf{y})$ is chosen. On the other hand its precise calculation from the optimum decoder is impractical. The only alternative is to retain the unknown function $g(\mathbf{z})$ in the lower bound expressions and to minimize the final lower bound to distortion over all possible probability functions on Z^n. Since $g(\mathbf{z})$ is one such probability function the inequality in the lower bound is continued. Unfortunately, when this is done it cannot, in general, be shown that the function which minimizes the lower bound factors into n letter probabilities, a form which we were permitted to assume for $f(\mathbf{y})$. However, to proceed beyond the bounds in Theorems 1 and 2, it is necessary to approximate this $g(\mathbf{z})$ by such a product, as in

$$g(\mathbf{z}) = \prod_{m=1}^{n} g(z^m). \tag{15}$$

The necessity for an approximation of this type is, of course, because of the requirement that an estimate be made for the distribution function $G(d)$. The assumed form for $g(\mathbf{z})$ in equation 15, will again allow us to use large number laws to obtain this estimate.

More specifically, the assumed product form for $g(\mathbf{z})$ allows us to cast the word distortion random variable $d(\mathbf{w}, \mathbf{z})$ as a sum of n independent letter variables. This is done in the following way. Among the letter distortions $d(w^m, z^m)$ that sum to the total word distortion there are H different types, corresponding to each of the different letters w_i, $1 \leqq i \leqq H$, that appear in the source word $\mathbf{w}$.

If the composition of this word is $\mathbf{q} = q_1, q_2, \cdots q_H$, that is, if there are nq_1 appearances of w_1 in $\mathbf{w}$, nq_2 appearances of w_2, and so on, the normalized word distortion can be written as

$$d(\mathbf{w}, \mathbf{z}) = \frac{1}{n} \sum_{i=1}^{H} \sum_{r=1}^{nq_i} D_{ir} . \tag{16}$$

In this expression D_{ir} is used to denote the distortion between the r'th appearance of the letter w_i in $\mathbf{w}$ and the corresponding letter in $\mathbf{z}$. Equation 15 now allows the interpretation of the D_{ir} as independent random variables, having the probability distributions

$$\begin{aligned} P_{D_{ir}}(d_{ij}) &= g_j ; \qquad 1 \leqq r \leqq nq_i , \quad 1 \leqq i \leqq H \\ d(w_i , z_j) &= d_{ij} \\ g(z_j) &= g_j , \end{aligned} \tag{17}$$

with the result that $G(d)$ is an n-fold convolution of elementary distribution functions for which there exist many estimating forms.[10–13]

We realize that the approximation in equation 15 is not entirely satisfactory because it eliminates nonproduct probability functions from the minimization of the lower bound and, as far as we know, one of these functions could provide the minimization. However, there is good reason to believe that this approximation does not significantly affect the bound when n is reasonably large. For example, in the next several sections we derive a lower bound to distortion that uses the product from in equation 15. For this bound the required minimization over all probability functions $g(\mathbf{z})$ is reduced to one over all J dimensional vectors $\mathbf{g}$. It can be shown that if in the limit as n becomes large, the product form requirement for $g(\mathbf{z})$ is relaxed, and the minimization of this lower bound is again made over all probability functions $g(\mathbf{z})$, then the optimizing function $g_o(\mathbf{z})$ still has the product form.

Even more significant is the asymptotic form of the lower bound that is derived using equation 15. We later show that it is *only* the final value of the minimizing decoder set size vector $\mathbf{g}_o(n = \infty)$ that affects *both* the asymptote of the lower bound, d_C, and the next lowest order term, which is one proportional to $1/n$. Values of the minimizing vector for finite n, $\mathbf{g}_o(n < \infty)$, affect only terms of $o(1/n)$.

Further, it can be shown that a similar conclusion is reached even if the independence property assumed over letters in equation 15 is generalized to be over blocks of length r, that is if

$$g(\mathbf{z}) = \prod_{m=1}^{n/r} g(\mathbf{z}'^m)$$

$$\mathbf{z}'^m = z_j, z_{j+1}, \cdots, z_{j+r-1}; \qquad j = mr - r + 1.$$

When $g(\mathbf{z})$ is assumed to have this form, the minimization of the lower bound over all decoder set sizes is a minimization over all probability functions $g(\mathbf{z}')$ on Z^r. The conclusion that can be made from the bound derived using this assumption is that it is again *only* the value of the minimizing decoder set size function at $n = \infty$, $g_o(\mathbf{z}', \infty)$, that influences both the asymptote and the term proportional to $1/n$. And, at $n = \infty$, the minimizing decoder set size function on Z^r, $g_o(\mathbf{z}', \infty)$, factors into a product of single letter probability functions on Z. When this solution is substituted in the bound (that uses $r \geqq 1$) the *asymptotic* form is the same for *every* choice of the constant r. Only lower order terms differ for different values of r.

There is one situation in which the assumed product form in equation 15 does not represent an approximation. That is the case of a doubly uniform source, which is a source that has a uniform probability distribution over its letters and has a distortion matrix in which each row and column is the respective permutation of another row and column. For such a source it has been shown[8] that the probability distribution $g(\mathbf{z})$ which minimizes the lower bound in Theorem 1 is uniform for all n, thus has the factorability property in equation 15.

6.3 *A Lower Bound to* d(I)

We now seek an approximation to $d(I)$ that we can substitute in equation 9 and preserve the inequality. A safe approximation to $d(I)$ can be had if, instead of equating $F_1(I^-)$ to $G(d)$ as in equation 7, we equate a lower bound estimate of $G(d)$ to an upper bound estimate of $F_1(I^-)$. Figure 5b illustrates this construction. The result is another distortion function, $d_L(I)$, that satisfies

$$d_L(I) \leqq d(I) \tag{18}$$

which can be used in equation 9 to obtain

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} d_L(I)\, dF_2(I). \tag{19}$$

Since the random variable I_2 is a normalized sum of n independent random variables, its variance is proportional to $1/n$. Consequently, when n becomes large the distribution function $F_2(I)$ has almost all of its "rise" around the mean of I_2, which we denote by $\bar{\text{I}}$. In this region, $I \approx \bar{\text{I}}$, $d \approx d(\bar{\text{I}})$, the values of both distribution functions $G(d)$ and $F_1(I)$ are exponentially small. Therefore, the bounds to the tails of distribution functions[10–13] are applicable to the estimation of $G(d)$ and $F_1(I)$ in this region. Indeed, it was with the intended use of these powerful bounds that we formed both the distortion and information difference random variables as sums of n independent letter random variables. All of the bounds, though, are parametric in form and allow only a parametric representation of $d_L(I)$.

We have elsewhere[8] applied strict upper and lower bounds to $G(d)$ and $F_1(I)$, respectively, to obtain the function $d_L(I)$. However, when these bounds are used, the resulting total lower bound to transmission distortion, though applicable for all block lengths n, does not reveal the correct asymptotic behavior inherent to the sphere-packing procedure which has been used. (This happens because the strict bounds to $G(d)$ and $F_1(I)$ themselves do not have the correct asymptotic form to large n.)

In addition, the resulting lower bound to the total distortion is very complex and so does not provide much insight into the factors which affect the rate of approach of the performance curve to its asymptote. For these reasons, we instead use Shannon's[11] and Gallager's[13] asymptotic forms for the tails of distribution functions to bound $G(d)$ and $F_1(I)$. These are:

$$G(d) \leqq \left[\frac{1}{\sqrt{2\pi n s^2 \mu''(s)}} + A_U(n, s) \right] \exp n[\mu(s) - s\mu'(s)] \tag{20a}$$

$$\mu'(s) = d \tag{20b}$$

with

$$0 < d \leqq E(d \mid \mathbf{q}) = \sum_{Z^n} d(\mathbf{w}, \mathbf{z} \mid \text{comp}\,\mathbf{w} = \mathbf{q}) g(\mathbf{z}),$$

and

$$F_1(I) \geqq \left[\frac{1}{\sqrt{2\pi n t^2 \gamma''(t)}} + A_L(n, t)\right] \exp n[\gamma(t) - t\gamma'(t)] \tag{21a}$$

$$\gamma'(t) = I \tag{21b}$$

with

$$I_{\min} < I \leqq E(I_1 \mid \mathbf{c}) = \sum_{Y^n} I(\mathbf{x}, \mathbf{y} \mid \text{comp } \mathbf{x} = \mathbf{c}) f(\mathbf{y}).$$

In these bounds, $A_U(n, s)$ and $A_L(n, t)$ are sums of rather difficult integrals but each has been shown by Shannon and Gallager to be

$$o\left(\frac{1}{\sqrt{n}}\right).$$

Also within the previous bounds, we have used $\mu(s)$ to denote the semi-invariant moment generating function of the variable d,

$$\begin{aligned} \mu(s) &= \sum_{i=1}^{H} q_i \mu_i(s) \\ &= \sum_{i=1}^{H} q_i \ln \sum_{j=1}^{J} g_j \exp s\, d_{ij} \,, \end{aligned} \tag{22}$$

and $\gamma(t)$ to denote the semi-invariant moment generating function of the variable I,

$$\begin{aligned} \gamma(t) &= \sum_{k=1}^{K} c_k \gamma_k(t) \\ &= \sum_{k=1}^{K} c_k \ln \sum_{l=1}^{L} f_l^{1+t} p_{kl}^{-t} \,. \end{aligned} \tag{23}$$

To guarantee the boundedness of $\gamma(t)$, we restrict the vector $\mathbf{f}$ to have nonzero components. This does not affect the resulting bound. (Actually, these bounds strictly apply only when the variables d and I are nonlattice. For lattice variables the corresponding bounds[11,13] have in their coefficient a quantity Δ which does not change continuously with the argument of the distribution function, and cannot be used within our derivation. One alternative would be to decrease one assigned letter distortion $d(w, z)$ by an arbitrarily small irrational number, and similarly, to change two transition probabilities on the channel in a way consistent with a lower bound to distortion. The new variables d' and I' would then be nonlattice.)

The desired distortion function, $d_L(I)$, can now be defined by equating the two bounds in equations 20 and 21. It can be constructed through the chain: $I^- \to t \to s \to d$ in which the superscript could now be dropped since the bound to $F_1(I)$ is continuous in I. It is important to notice that the region of validity of the previous two bounds allows definition of the function $d_L(I)$ only in a subinterval $[I_a, I_b]$ of $[I_{\min}, I_{\max}]$ with

$$I_{\min} < I_a < \bar{I} < I_b \leqq E(I_1 \mid \mathbf{c}), I[E(d \mid \mathbf{q})].$$

Outside the interval $[I_a, I_b]$ we can define $d_L(I)$ equal to zero and write the lower bound in equation 19 as

$$d(\mathbf{w}) \geqq \int_{I_a}^{I_b} d_L(I)\, dF_2(I). \tag{24}$$

We are now faced with the difficult integration of a doubly parametric expression. Rather than integrate directly, we use the following Taylor series expansion for $d_L(I)$ within $[I_a, I_b]$:

$$\begin{aligned} d_L(I) &= d_L(\bar{I}) + d_L'(\bar{I})(I - \bar{I}) + \tfrac{1}{2}\, d_L''(\bar{I})(I - \bar{I})^2 + \tfrac{1}{6}\, d_L'''(I')(I - \bar{I})^3 \\ &\equiv TS(d_L) \end{aligned}$$

with $I_a \leq I' \leq I_b$. (The indicated derivatives can be shown to exist within the restricted interval $[I_a, I_b]$.) Using this form for $d_L(I)$ within equation 24 we see that if the region of integration were $[I_{\min}, I_{\max}]$ instead of $[I_a, I_b]$, the resulting form would be a sum of central moments of I_2 with the Taylor series derivatives as coefficients. To restore this form we rewrite equation 24 as

$$d(\mathbf{w}) \geqq \int_{I_{\min}}^{I_{\max}} \cdots - \int_{I_{\min}}^{I_a} \cdots - \int_{I_b}^{I_{\max}} TS(d_L)\, dF_2(I). \tag{25}$$

In these integrals, the lower limit $I_{\min}$ is finite since f_l is assumed nonzero for all l, and $I_{\max}$ can be taken as the largest finite value of $\ln f_l/p_{kl}$ since this is the largest value of I for which the random variable I_2 has nonzero probability. Therefore the function $TS(d_L)$ is bounded in $[I_{\min}, I_a]$ and $[I_b, I_{\max}]$ with the result that the last two integrals in equation 25 are exponentially small in n. The first integral in this equation has the desired form, involving the central moments of I_2:

$$\begin{aligned} \int_{I_{\min}}^{I_{\max}} TS(d_L)\, dF_2(I) = d_L(\bar{I}) + d_L'(\bar{I})E(I - \bar{I}) + \tfrac{1}{2}\, d_L''(\bar{I})E[(I - \bar{I})^2] \\ + \tfrac{1}{6}\, d_L'''(I')E[(I - \bar{I})^3]. \end{aligned}$$

In the above equation the second term is zero since we have specified that $\bar{I}$ is the expected value of I_2, and the last term can be shown to be proportional to $(1/n)^2$. This establishes the result in the next theorem.

Theorem 3: The conditional average transmission distortion, $d(\mathbf{w})$, satisfies

$$d(\mathbf{w}) \geqq d_L(\bar{I}) + \tfrac{1}{2}\, d_L''(\bar{I}) \operatorname{var}(I_2) + o\left(\frac{1}{n}\right). \tag{26}$$

Compared with the last low order term, the variance of I_2 is proportional to $1/n$.

The simplicity in the form of the last result is due to the use of the Taylor series expansion which not only has allowed us to evaluate a difficult integral, but has provided a natural way of separating the important terms in the lower bound to distortion.

6.4 *The Evaluation of* $d_L(\bar{I})$ *and* $d_L''(\bar{I})$

We shall denote by s_o and t_o the parameter values consistent with $I = \bar{I}$ in equations 20 and 21. Since

$$\gamma'(-1) = \sum_{k=1}^{K} \sum_{l=1}^{L} p_{kl} \ln f_l/p_{kl}\,,$$

which is seen equal to $E(I_2) = \bar{I}$, we can conclude that $t_o = -1$. We also note here for future use that

$$\gamma(-1) = 0.$$

The first of the two significant terms in equation 26 is immediate:

$$d_L(\bar{I}) = \mu'(s_o).$$

Next, elementary differentiation of the parametric expressions in equations 20 and 21 provides

$$d_L'(\bar{I}) = \frac{t}{s}\bigg|_{t_o, s_o} = -\frac{1}{s_o}$$

and

$$d_L''(\bar{I}) = \frac{1}{s}\left[\frac{1}{\gamma''(t)} - \frac{t^2}{s^2\mu''(s)}\right]\bigg|_{t_o, s_o} = \frac{1}{s_o}\left[\frac{1}{\gamma''(-1)} - \frac{1}{s_o^2\mu''(s_o)}\right].$$

Finally, the variance of I_2 is seen from equation 12 to equal

$$\begin{aligned}\mathrm{Var}\,(I_2) &= \frac{1}{n}\sum_{k=1}^{K} c_k\,\mathrm{Var}\,(I_{kr})\\ &= \frac{1}{n}\sum_{k=1}^{K} c_k\left[\sum_{l=1}^{L} p_{kl}(\ln f_l/p_{kl})^2 - \left(\sum_{l=1}^{L} p_{kl}\ln f_l/p_{kl}\right)^2\right]\\ &= \gamma''(-1).\end{aligned}$$

With the substitution of these terms in equation 26 we obtain the result in the next theorem.

Theorem 4: The conditional average transmission distortion, $d(\mathbf{w})$, *satisfies*

$$d(\mathbf{w}) \geqq \mu'(s_o) - \frac{1}{2ns_o}\left[\frac{\gamma''(-1)}{s_o^2\mu''(s_o)} - 1\right] + o\left(\frac{1}{n}\right) \tag{27}$$

in which s_o *is given by*

$$\mu(s_o) - s_o\mu'(s_o) = \bar{\mathrm{I}} - \frac{1}{2n}\ln\frac{\gamma''(-1)}{s_o^2\mu''(s_o)} + o\left(\frac{1}{n}\right). \tag{28}$$

It remains to average this lower bound over the entire source space W^n.

VII. THE AVERAGE OVER THE SOURCE SPACE

To average the lower bound in Theorem 4 over the source space W^n we assume that channel input words of equal composition are used for all transmissions. It has been shown[8] that this assumption does not affect the asymptotic form of the lower bound to distortion. We first notice that the lower bound in Theorem 4 depends upon the source word $\mathbf{w}$ only through its composition $\mathbf{q}$ which enters in the form of $\mu(s)$. Therefore, we can average $d(\mathbf{w})$ over the set of all compositions for $\mathbf{w}$ rather than over all of W^n. As all composition vectors for $\mathbf{w}$ are probability vectors, they are all located on an $H - 1$ dimensional hyperplane, termed the composition space Q^H, which is in the "first quadrant" of R^H and intersects each axis q_i at one. Not all points in Q^H are possible word compositions for any particular n. For example, with $H = 2$ and $n = 2$ there are only three possible compositions. But as n increases, the points in Q^H that are source word compositions become quite dense.

The probability that any particular composition $\mathbf{q}$ occurs at the source output is

$$P(\mathbf{q}) = N(\mathbf{q}) \prod_{i=1}^{H} p_i^{nq_i} \tag{29}$$

in which $N(\mathbf{q})$ is the number of distinct source sequences with the composition $\mathbf{q}$ and the product is the probability of each. The number $N(\mathbf{q})$ is given by

$$N(\mathbf{q}) = \frac{n\,!}{\prod_{i=1}^{H} (nq_i)\,!}.$$

We now write the total average source distortion, $d(\mathcal{S})$, as

$$d(\mathcal{S}) = \sum_{\substack{\text{all source}\\ \text{compositions}}} d(\mathbf{q})P(\mathbf{q})$$

which we can lower bound by substituting for $d(\mathbf{q})$ the lower bound found in Theorem 4. Rather than write out the entire expression each time we want to use it, we let $d_L(\mathbf{q})$ denote the right side of equation 27, thus have

$$d(\mathcal{S}) \geqq \sum_{\substack{\text{all source}\\ \text{compositions}}} d_L(\mathbf{q})P(\mathbf{q}). \tag{30}$$

Viewed as a function over Q^H, $P(\mathbf{q})$ is a set of impulses. This allows us to consider the distortion function $d_L(\mathbf{q})$ a continuous function over all Q^H, rather than a function defined only at composition points, and to write

$$d(\mathcal{S}) \geqq \int \cdots \int_{Q^H} d_L(\mathbf{q})P(\mathbf{q})\, d\mathbf{q}. \tag{31}$$

Again because the expression for $d_L(\mathbf{q})$ in equations 27 and 28 is parametric, we use a Taylor series expansion of this distortion function to evaluate the integral. The point chosen for the expansion is $\mathbf{p}$, the probability vector characterizing the source. The reason for this choice is that the components of this vector are the means of the coordinates of $\mathbf{q}$ when the latter are considered (dependent) random variables governed by $P(\mathbf{q})$. The Taylor series then contains terms of the type $(q_i - p_i)$, $(q_i - p_i)(q_j - p_j)$, and so on, which, when averaged by $P(\mathbf{q})$, are the central moments of the components of $\mathbf{q}$.

Using the notation $d'_{Li}(\mathbf{p})$ to indicate the partial derivative of $d_L(\mathbf{q})$ with the respect to q_i evaluated at $\mathbf{q} = \mathbf{p}$ (and similarly for higher order derivatives), we have

$$d(\mathbb{S}) \geqq \int_{Q^H} \cdots \int \Big[d_L(\mathbf{p}) + \sum_{i=1}^{H} d'_{Li}(\mathbf{p})(q_i - p_i) + \tfrac{1}{2} \sum_{ij} d''_{Lij}(\mathbf{p})(q_i - p_i)(q_j - p_j) + \tfrac{1}{6} \sum_{ijk} d'''_{Lijk}(\varphi)(q_i - p_i)(q_j - p_j)(q_k - p_k) \Big] P(\mathbf{q})\, d\mathbf{q} \qquad (32)$$

with $\varphi \varepsilon Q^H$. The central moments of the components of $\mathbf{q}$ can be found to be

$$E(q_i - p_i) = 0,$$

$$E[(q_i - p_i)(q_j - p_j)] = \frac{1}{n}(p_i\ \delta_{ij} - p_i p_j) \qquad (33)$$

$$E[(q_i - p_i)(q_j - p_j)(q_k - p_k)] = \left(\frac{1}{n}\right)^2 [p_i\ \delta_{ijk} - p_i p_j\ \delta_{ki} - p_j p_k\ \delta_{ij} - p_k p_i\ \delta_{jk} + 2 p_i p_j p_k],$$

which, when substituted in equation 32, yields

$$d(\mathbb{S}) \geqq d_L(\mathbf{p}) + \frac{1}{2n}\Big[\sum_i d''_{Lii}(\mathbf{p}) p_i - \sum_{ij} d''_{Lij}(\mathbf{p}) p_i p_j\Big] + o\left(\frac{1}{n}\right). \qquad (34)$$

Referring to equation 27 we see that the required second derivative need only be taken of $\mu'(s_o)$ as the two $1/n$ coefficients allow other terms to be absorbed in those of $o(1/n)$. The differentiation is lengthy, but straightforward, and yields

$$\frac{\partial}{\partial q_i} \mu'(s_o, \mathbf{q}) = \frac{\mu_i(s_o)}{s_o}$$

and

$$\frac{\partial^2}{\partial q_i\, \partial q_j} \mu'(s_o, \mathbf{q}) = -\frac{\theta_i \theta_j}{s_o^3 \mu''(s_o, \mathbf{p})}$$

where

$$\theta_i \equiv \mu_i(s_o) - s_o \mu_i'(s_o).$$

Upon substitution of these derivatives in equation 34 we obtain

$$d(\mathbb{S}) \geqq d_L(\mathbf{p}) - \frac{1}{2n s_o^3 \mu''(s_o)}\Big[\sum_i p_i \theta_i^2 - \sum_{ij} p_i p_j \theta_i \theta_j\Big] + o\left(\frac{1}{n}\right)$$

$$= d_L(\mathbf{p}) - \frac{1}{2n s_o^3 \mu''(s_o)} \operatorname{Var}(\theta) + o\left(\frac{1}{n}\right).$$

With the final substitution of the expression for $d_L(\mathbf{p})$ in equation 27 we have the result in the next theorem.

Theorem 5: The average transmission distortion of the source $\mathcal{S}$, *when used with the channel* $\mathcal{C}$, *is lower bounded by*

$$d(\mathcal{S}) \geqq \mu'(s_o, \mathbf{p}) - \frac{1}{2ns_o}\left[\frac{\gamma''(-1) + \sigma^2(\theta)}{s_o^2\mu''(s_o, \mathbf{p})} - 1\right] + o\left(\frac{1}{n}\right) \qquad (35)$$

in which s_o *is given by*

$$\mu(s_o, \mathbf{p}) - s_o\mu'(s_o, \mathbf{p}) = \bar{\mathrm{I}} - \frac{1}{2\mathrm{n}} \ln \frac{\gamma''(-1)}{s_o^2\mu''(s_o, \mathbf{p})} + o\left(\frac{1}{n}\right). \qquad (36)$$

In this bound the vector **g** is, for the reasons previously stated, that which minimizes the bound, the vector **f** is chosen to maximize the bound in order to obtain the tightest bound, and the vector **c** is chosen to minimize the bound, that is to use the best composition for the channel input code words. As formidable as the derivations of these extremum appear, we show in the next section that the work involved in establishing the asymptotic behavior of the bound is actually quite simple.

It should be mentioned that these results do *not* apply when $\gamma''(-1) = 0$, which is a situation that occurs when channel $\mathcal{C}$ is noiseless, for the reason that we have divided by and canceled factors equal to $\gamma''(-1)$. The result for this case is derived separately in Section IX.

VIII. THE ASYMPTOTE AND RATE OF APPROACH

8.1 *The Asymptote*

When n becomes large, the limiting form of the bound in Theorem 5 is:

$$d_\infty(\mathcal{S}) \geqq \mu'(s_o, \mathbf{p})$$

in which s_o satisfies

$$\mu(s_o, \mathbf{p}) - s_o\mu'(s_o, \mathbf{p}) = \bar{\mathrm{I}}$$

with

$$\bar{\mathrm{I}} = \sum_{k=1}^{K} c_k \sum_{l=1}^{L} p_{kl} \ln f_l/p_{kl} .$$

The vectors **g**, **f**, and **c** must now be chosen to provide the extremum indicated just after Theorem 5. Since only **f** and **c** enter in the expression for $\bar{I}$, we can minimize $d_\infty(\mathcal{S})$ with respect to $\mathbf{g}$ for a constant $\bar{I}$. This minimization provides precisely the expression[7] for the rate-distortion curve for $\mathcal{S}$ at the information rate $\bar{I}$. It is further shown in the same reference that the value of $\mathbf{g}$ which provides the minimization is the vector that describes the output statistics on the test channel for $\mathcal{S}$ at the point $(d_{\bar{I}}, \bar{I})$ on the rate-distortion curve.

The maximization and minimization of $d_\infty(\mathcal{S})$ with $\mathbf{f}$ and $\mathbf{c}$, respectively, can be accomplished by finding the same extremum of $\bar{I}$. The resulting values for $\mathbf{f}$ and $\mathbf{c}$ are the output and input probabilities, respectively, on channel $\mathcal{C}$ when it is being used to capacity and the value of $\bar{I}$ at the extremum point is $-C$. Therefore, the resulting expression for the asymptote of the lower bound is

$$d(\mathcal{S}) \geqq \min_{\mathbf{g}} \mu'(s_o , \mathbf{p}) = d_C \tag{37}$$

with s_o satisfying

$$\mu(s_o , \mathbf{p}) - s_o\mu'(s_o , \mathbf{p}) = -C. \tag{38}$$

This agrees with what we know to te the correct asymptote of the performance curve.[2,7]

8.2 *The Rate of Approach to the Asymptote*

Since the lower bound in equations 35 and 36 is parametric in s and includes the vectors $\mathbf{f}$, $\mathbf{c}$, and $\mathbf{g}$, which when optimally chosen are functions of n, the complete asymptotic dependence of this lower bound upon the block length n is not obvious. To establish this dependence, we first find the *full* derivative of the lower bound in Theorem 5 with respect to n and then integrate the result between n and infinity.

We first simplify the procedure slightly by using our freedom to choose $\mathbf{f}$ by setting this vector equal to its value at $n = \infty$; $\mathbf{f}(\infty)$. This does not change the end result. We also drop the terms of $o(1/n)$ in equations 35 and 36, because they clearly do not affect the asymptotic result. Denoting the right side of equation 35 by d_L and using the chain rule several times, we can write the desired derivative as

$$\frac{dd_L}{dn} = \left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} + \left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n} \frac{ds}{dn} + \sum_i \left(\frac{\partial d_L}{\partial g_i}\right)_{\substack{g_k \neq i \\ \mathbf{c},n,s}} \frac{dg_i}{dn} + \sum_k \left(\frac{\partial d_L}{\partial c_k}\right)_{\substack{c_l \neq k \\ \mathbf{g},n,s}} \frac{dc_k}{dn}$$

with

$$\frac{ds}{dn} = \left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} + \sum_j \left(\frac{\partial s}{\partial g_j}\right)_{\substack{g_k \neq j \\ \mathbf{c},n}} \frac{dg_j}{dn} + \sum_k \left(\frac{\partial s}{\partial c_k}\right)_{\substack{c_l \neq k \\ \mathbf{g},n}} \frac{dc_k}{dn}.$$

The notations outside each parentheses indicate the variables which are momentarily held constant. Substitution yields:

$$\begin{aligned}\frac{dd_L}{dn} = \left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} &+ \left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n}\left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} \\ &+ \sum_j \left[\left(\frac{\partial d_L}{\partial s}\right)_{\substack{\mathbf{c},\mathbf{g} \\ n}}\left(\frac{\partial s}{\partial g_j}\right)_{\substack{g_k \neq j \\ \mathbf{c},n}} + \left(\frac{\partial d_L}{\partial g_j}\right)_{\substack{g_k \neq j \\ \mathbf{c},s,n}}\right]\frac{dg_j}{dn} \\ &+ \sum_k \left[\left(\frac{\partial d_L}{\partial s}\right)_{\substack{\mathbf{c},\mathbf{g}, \\ n}}\left(\frac{\partial s}{\partial c_k}\right)_{\substack{c_l \neq k \\ \mathbf{g},n}} + \left(\frac{\partial d_L}{\partial c_k}\right)_{\substack{c_l \neq k \\ \mathbf{g},n,s}}\right]\frac{dc_k}{dn}.\end{aligned}$$

The bracketed terms represent the respective partial derivatives of d_L with respect to g_j and c_k with s removed from those quantities held constant. Since $\mathbf{g}(n)$ and $\mathbf{c}(n)$ are chosen for each value of n to minimize the lower bound d_L, these partial derivatives must satisfy

$$\left(\frac{\partial d_L}{\partial g_j}\right)_{\substack{g_k \neq j \\ \mathbf{c},n}} + \lambda = 0 \qquad 1 \leqq j \leqq J \tag{39}$$

$$\left(\frac{\partial d_L}{\partial c_k}\right)_{\substack{c_l \neq k \\ \mathbf{g},n}} + \nu = 0 \qquad 1 \leqq k \leqq K. \tag{40}$$

This presumes that, at least for sufficiently high n, both $\mathbf{g}$ and $\mathbf{c}$ have only nonzero components. This is known to be true for $\mathbf{c}$,[14] which at $n = \infty$ equals the channel input probabilities that use the channel to capacity.

The vector $\mathbf{g}$, though, can at $n = \infty$ have a zero component. For this case, if the approach of $\mathbf{g}(n)$ to $\mathbf{g}(\infty)$ is from within the composition space, that is, if the components of $\mathbf{g}(n < \infty)$ are nonzero, equation 39 is correct as written for all finite n. If, however, the approach of $\mathbf{g}(n)$ to $\mathbf{g}(\infty)$ is along the boundary of the composition space, that is, having one or more components equal to zero for all $n > N$, then equation 39 can be written, not for all $1 \leqq j \leqq J$, but only for the J' nonzero components. Over the region (N, ∞) the other $J - J'$ zero components obviously can be treated as constants and not included in the differentiation process, thus excluded from the previous summations on j. We shall not attempt to deal with the only remaining possibility, which has $\mathbf{g}(n)$ approaching $\mathbf{g}(\infty)$ such that it oscillates between vector values with all nonzero components and values with some zero components, since no example has been found exhibiting this behavior.

We continue the derivation by substituting equations 39 and 40 into the derivative of d_L to obtain

$$\frac{dd_L}{dn} = \left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} + \left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n}\left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} - \lambda \sum_i \frac{dg_i}{dn} - \nu \sum_k \frac{dc_k}{dn}. \quad (41)$$

Finally, since both $\mathbf{g}$ and $\mathbf{c}$ are probability vectors, the last two sums are equal to zero (this is true even when the first sum is only over the J' nonzero components of $\mathbf{g}$). It remains only to find the required partial derivatives from equations 35 and 36. These are given by:

$$\left(\frac{\partial d_L}{\partial n}\right)_{\mathbf{c},\mathbf{g},s} = \frac{1}{2n^2 s}\left(\frac{\gamma'' + \sigma^2}{s^2\mu''} - 1\right),$$

$$\left(\frac{\partial d_L}{\partial s}\right)_{\mathbf{c},\mathbf{g},n} = \mu'' + o(1)$$

$$\left(\frac{\partial s}{\partial n}\right)_{\mathbf{g},\mathbf{c}} = \frac{1}{2n^2 s\mu''} \ln \frac{\gamma''}{s^2\mu''}$$

whence substitution in equation 41 provides

$$\frac{dd_L}{dn} = -\frac{1}{n^2}\frac{1}{2\,|\,s\,|}\left[\left(\frac{\gamma''}{s^2\mu''} - 1\right) - \ln \frac{\gamma''}{s^2\mu''} + \frac{\sigma^2}{s^2\mu''}\right] + o\left(\frac{1}{n^2}\right). \quad (42)$$

At this point, the vectors $\mathbf{g}$, $\mathbf{c}$ and the parameter s are still functions of n chosen to satisfy the prescribed minimizations of Equation 55 and the parametric Equation 35. If, for large n, these functions are written as

$$\mathbf{g}(n) = \mathbf{g}(\infty) + \Delta\mathbf{g}(n)$$

$$\mathbf{c}(n) = \mathbf{c}(\infty) + \Delta\mathbf{c}(n)$$

$$s(n) = s(\infty) + \Delta s(n),$$

the delta terms can be extracted from the first term in Equation 42. Since each has limit zero for large n, they can, together with the $(1/n)^2$ coefficient, be absorbed into the terms of $o(1/n^2)$. Thus, in equation 42, we can use for $\mathbf{g}$, $\mathbf{c}$, and s their *final* values: $\mathbf{g}(\infty)$, $\mathbf{c}(\infty)$, and $s(\infty)$.

Simple integration of equation 42 between n and infinity, and the use of the known final value of $d_L(n)$, $d_L(\infty) = d_C$, provides the final lower bound to distortion. We again point out that the derivation has included the approximation that $g(\mathbf{z})$ factors as in equation 15.

Theorem 6: A lower bound to the minimum attainable transmission distortion in a system that includes the source $\mathcal{S}$ *and the channel* $\mathcal{C}$ *is given by*

$$d(\mathcal{S}) \geqq d_C + \frac{1}{2n\,|\,s\,|}\left[\left(\frac{\gamma''}{s^2\mu''} - 1\right) - \ln\frac{\gamma''}{s^2\mu''} + \frac{\sigma^2}{s^2\mu''}\right] + o\left(\frac{1}{n}\right) \quad (43)$$

in which

C = capacity of $\mathcal{C}$

d_C = the distortion at $R = C$ on the rate-distortion curve for $\mathcal{S}$

$$\mu(s) = \sum_i q_i \ln \sum_j g_j \exp sd_{ij}$$

$$\gamma(t) = \sum_k c_k \ln \sum_l f_l^{1+t} p_{kl}^{-t}$$

$\mathbf{q} = \mathbf{p}$, the source output probabilities
$\mathbf{g}$ = the output probabilities on the test channel for $\mathcal{S}$ at (d_C, C)
$\mathbf{c}, \mathbf{f}$ = the input and output probabilities on $\mathcal{C}$ when it is used to capacity
$t = -1$
s satisfies $\mu - s\mu' = -C$.

The lower bound in equation 43 is seen to approach its limit algebraically as a/n. Since $(w-1)$ is at least as large as $\ln w$ for any w and σ^2 and μ'' are variances, hence nonnegative, the coefficient a cannot be negative. But it can in special cases equal zero. The conditions for this are

$$\gamma'' = s^2\mu''$$

$$\sigma^2 = 0,$$

conditions that are necessarily met when the source and channel are perfectly matched; that is, when $d(\mathcal{S}) = d_C$ for all n.

They do not, however, constitute a sufficient condition for matching since the low order correction terms in equation 43 could still be nonzero. For the more common situations wherein a is nonzero, the form of the lower bound suggests that the larger the value of a, the longer the coding block length must be to obtain a tolerable level of distortion, $d_C + \Delta$. In turn, the more complex the modulator and demodulator must become. These relations all suggest the utility of the coefficient a as a measure of mismatch between the source $\mathcal{S}$ and the channel $\mathcal{C}$; the larger the value of a, the slower the approach of the lower bound to its asymptote and the greater the mismatch between source and channel. Section X gives several numerical examples illustrating different types of mismatch.

IX. THE SPECIAL CASE OF A NOISELESS CHANNEL

As we have stated, Theorem 5 cannot be applied when $\mathcal{C}$ is noiseless because factors equal to $\gamma''(-1)$ have been canceled within its derivation and, for a noiseless channel, $\gamma''(-1)$ equals zero. We return to the lower bound in equation 3 which is still valid. If the vector $\mathbf{f}$ is chosen uniform over Y^n, we see from the definition of a noiseless channel (L^n outputs) and the definition of information difference in Section IV that $I(\mathbf{x}, \mathbf{y})$ is equal to $\ln (1/L)$ for the output $\mathbf{y}_1$ that has $p(\mathbf{y}_1/\mathbf{x}) = 1$, and is infinite for all other outputs. Since $f(\mathbf{y}_1) = L^{-n}$, $e^{-nI(h)}$ is nonzero only in $0 \leqq h \leqq L^{-n}$, where it is equal to L^n. Therefore, equation 3 can be written as

$$d(\mathbf{w}) \geqq L^n \int_0^{L^{-n}} d(h)\, dh. \tag{44}$$

We remember that the distribution function $G(d)$ is the "inverse" function to $d(h)$ and write

$$d(\mathbf{w}) \geqq L^n \int_0^{d(L^{-n})} [L^{-n} - G(d)]\, dd$$

which can be continued, with any $d_2 \leq d(L^{-n})$, by

$$d(\mathbf{w}) \geqq L^n \int_0^{d_2} [L^{-n} - G(d)]\, dd.$$

Upon dividing the region of integration into two parts, $0 \leq d_1 \leq d_2$, and using the monotonicity of $G(d)$, we have

$$d(\mathbf{w}) \geqq d_2 - L^n\, d_1 G(d_1) - L^n \int_{d_1}^{d_2} G(d)\, dd. \tag{45}$$

A further lower bound results if we use an upper bound to $G(d)$ in each of the last two terms. In particular, we use the asymptotic bound in equation 20 which we denote here by

$$G(d) \leqq H(n, s) \exp n[\mu(s) - s\mu'(s)] \tag{46}$$

$$\mu'(s) = d.$$

We now set d_2 equal to $\mu'(s_o)$ with s_o given by

$$H(n, s_o) \exp n[\mu(s_o) - s_o\mu'(s_o)] = L^{-n} = e^{-nC}. \tag{47}$$

The fact that $G(d_2) \leq L^{-n}$ guarantees the inequality $d_2 \leq d(L^{-n})$ which we have already used. The second term in equation 45 can be shown to be exponentially small in n whenever $d_1 < d_2$; therefore, we also impose this inequality. To bound the last term in the same equation we use the well known Chernov bound inequality:

$$\exp n[\mu(s)-s\mu'(s)] \leqq \exp n[\mu(s_o)-s_o d]$$

$$\mu'(s) = d$$

together with equations 46 and 47 to obtain

$$L^n \int_{d_1}^{d_2} G(d)\, dd \leqq De^{ns_o\mu'(s_o)} \int_{d_1}^{d_2} e^{-ns_o d}\, dd$$

with

$$D = \max_{d_1 \leqq d \leqq d_2} \frac{H(n, s)}{H(n, s_o)}.$$

The resulting bound for $d(\mathbf{w})$, therefore, is

$$d(\mathbf{w}) \geqq \mu'(s_o) + \frac{D}{ns_o}[1 - \exp ns_o(\mu'(s_o) - d_1)] + o\left(\frac{1}{n}\right).$$

If d_1 is chosen in a way to approach $\mu'(s_o)$ with increasing n, this bound becomes:

$$d(\mathbf{w}) \geqq \mu'(s_o) + \frac{1}{ns_o}[1 + o(1)] \tag{48}$$

in which s_o satisfies equation 47, rewritten here as

$$\begin{aligned} \mu(s_o) - s_o\mu'(s_o) &= -C - \frac{1}{n} \ln H(n, s_o) \\ &= -C + \frac{1}{2n} \ln n[1 + o(1)]. \end{aligned} \tag{49}$$

The remaining steps, averaging over the source space and minimizing the resulting bound over all choices of $\mathbf{g}$ (we continue to use the approximation in Equation 15), are identical in procedure to those previously used. We state only the result.

Theorem 7: The minimum attainable transmission distortion of the source $\mathcal{S}$, *when used with a noiseless channel of capacity* C, *satisfies*

$$d(\mathcal{S}) \geqq d_C + \frac{1}{2}\frac{\ln n}{|s_o|\, n}[1 + o(1)] \tag{50}$$

in which s_o satisfies

$$\mu(s_o\,,\,\mathbf{p})\;-\;s_o\mu'(s_o\,,\,\mathbf{p})\;=\;-C. \tag{51}$$

We see by comparing equations 43 and 50 that while the lower bound to distortion with a noisy channel approaches its asymptote, d_C, as $1/n$, the lower bound to distortion with a noiseless channel approaches d_C only as $(\ln n)/n$. These bounds are not inconsistent since for a noiseless channel the variance γ'' is zero with the result that the coefficient of $1/n$ in equation 43 is infinite. A similar limiting statement is also true. If a noisy channel is made to approach a noiseless one by reducing the noisy transition probabilities toward zero, at the same time keeping the channel capacity constant by appropriately reducing either the channel input alphabet size or the channel dimensionality, the coefficient of the $1/n$ term increases and is unbounded. These results therefore suggest than when there is a choice between using a noiseless channel or a noisy one *of equal capacity*, the noisy channel is always the better choice. And, inasmuch as we are using the coefficient of the $1/n$ term to measure the source-channel mismatch, the noiseless channel represents the worst possible match to any source.

X. EXAMPLES

In the first three examples, we illustrate different types of source-channel mismatch and calculate the effect of each upon the coefficient a in the lower bound of equation 43. Each of these examples tends to strengthen the suggestion in the lower bound result that this coefficient is a measure of source-channel mismatch since it increases monotonically as the channel is perturbed away from the matching channel.

Because the channel statistics influence only the first two terms of a, we use in these examples a doubly uniform source for which the σ^2 term equals zero. To further isolate the relative matching properties of the source-channel pairs, we keep constant the channel capacity per source output, C, as the channel is varied. Thus the distortion per source component has the same asymptote, d_C, for all source-channel pairs and the only difference in the lower bound curves, at least asymptotically, is in the coefficient a.

Example 1

This example illustrates a dimensionality, or coding block length, mismatch between a source and channel. We take for the source $\mathcal{S}$

the m_s'th product of a binary symmetric source, defined by $\mathbf{p} = (\frac{1}{2}, \frac{1}{2})$ and $d_{11} = d_{22} = 0$, $d_{12} = d_{21} = 1$. For the channel $\mathcal{C}$ we take the m_c'th product of a binary symmetric channel, each component $\mathcal{C}_i$ having a crossover probability p. The channel capacity per source component is m_c/m_s times the capacity of $\mathcal{C}_i$ and is kept constant as m_c/m_s is varied by appropriately changing the crossover probabilities p.

Figure 6 shows the dependence of a upon m_c/m_s. When comparing the two curves in this figure, notice that the ordinate has been normalized by d_C. We know that for $m_c/m_s = 1$ the source and channel are precisely matched and this is indicated in the figure by the value $a = 0$ at that point. Above this point a increases monotonically in m_c/m_s and can be shown to have the asymptotic form $a \sim k(m_c/m_s)^{\frac{1}{2}}$. Below $m_c/m_s = 1$, a also becomes unbounded as m_c/m_s approaches the ratio that requires each component channel $\mathcal{C}_i$ be noiseless. This is not inconsistent with the noiseless channel result (equation 50) which indicated that the rate of approach of the distortion to d_C was not as a/n but as $(\ln n)/n$.

Example 2

Here we do not change the relative dimensionality, only the form of the channel. The source is a binary symmetric source and the channel a binary nonsymmetric channel of varying asymmetry. The crossover probabilities are again changed in a way that does not vary the capacity. We see in Fig. 7 that a is rather insensitive to small perturbations from a binary symmetric channel and in most cases is affected less by this type of mismatch than a dimensionality mis-

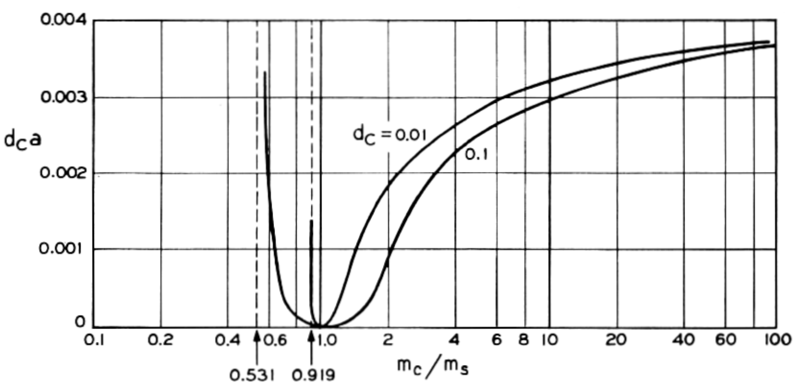

Fig. 6 — The mismatch between a binary symmetric source and a binary symmetric channel of different dimensionality.

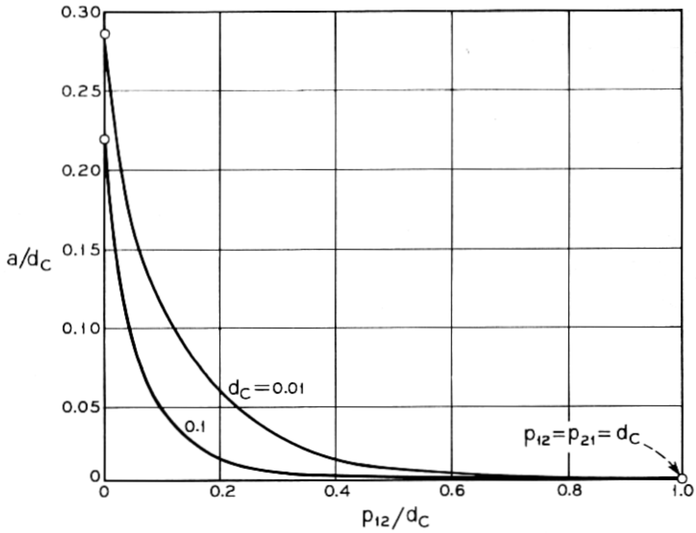

Fig. 7 — The mismatch between a binary symmetric source and a binary nonsymmetric channel.

match. A similar result obtains if the source is also allowed to be nonsymmetric.

Example 3

For this example we use a binary symmetric source and a discrete channel which models the m orthogonal signal modulator used in the next example. The channel has m inputs and m outputs and has from each input one transition of probability $1 - (m-1)p$ and $m - 1$ transitions of probability p. The numbers m and p are varied together in such a way that the capacity of the channel remains constant. We see in Fig. 8 that the mismatch coefficient a is much higher when the binary symmetric source is used with this channel than when it is used with that product binary symmetric channel of Example 1 which has available an input alphabet of equal size. The comparison can be made on Figures 6 and 8 at points for which $m_c/m_s = \log_2 m$.

Example 4

In this, the last example, we include in the system a continuous channel which is to be used by a discrete source with a discrete modulator. Now, as the modulator changes the discrete channel extracted from the actual channel changes and *both* its capacity and its matching characteristics change. It turns out that both properties are not necessarily optimized for the same modulator structure and, therefore, one must strike a compromise (influenced by the block length of interest) between a modulator design that minimizes the asymptote d_C and maximizes the rate of approach to d_C.

To illustrate this we assume the channel to be a band-limited channel with additive white gaussian noise in the allowed bandwidth. During the interval (O,T), the discrete modulator is constrained to transmit one of m orthogonal signals in each of B bauds and altogether an energy no greater than E. To model the bandwidth constraint the mB product is assumed constant, but m and B can otherwise be varied to optimize the system. Thus the equivalent discrete channel is the B'th product of the m input doubly uniform channel of Example 3. The source to be transmitted is a binary symmetric source with an output rate of M_s digits every T seconds.

In Fig. 9 we show the minimum attainable distortion d_C (determined through the channel capacity) and the mismatch coefficient a as a function of m. For the values shown in figure, we see that while d_C is minimized at $m = 15$, the coefficient a is then quite large. And, around $m = 22$, where $a = 0$, the minimum distortion d_C is higher than that which can be realized with a smaller m. The conclusion from this is that the modulator should be designed with $m =$ 15 (to maximize capacity and minimize d_C) only when one is willing to use very long coding block lengths. For shorter block lengths, a larger value of m, and a corresponding smaller value of a, could result in a smaller average distortion even with the larger value of d_C. For

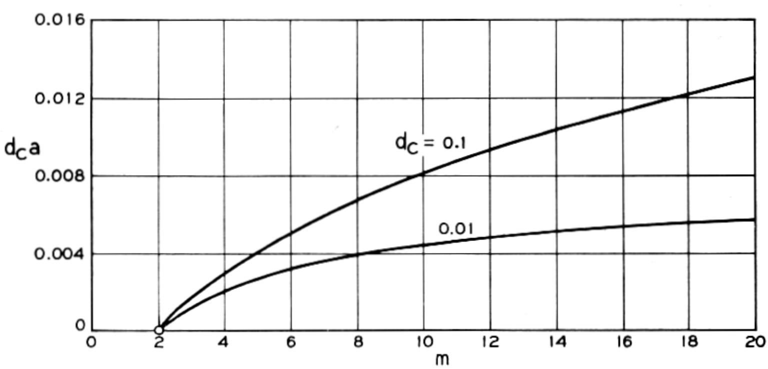

Fig. 8 — The mismatch between a binary symmetric source and the m-orthogonal signal channel.

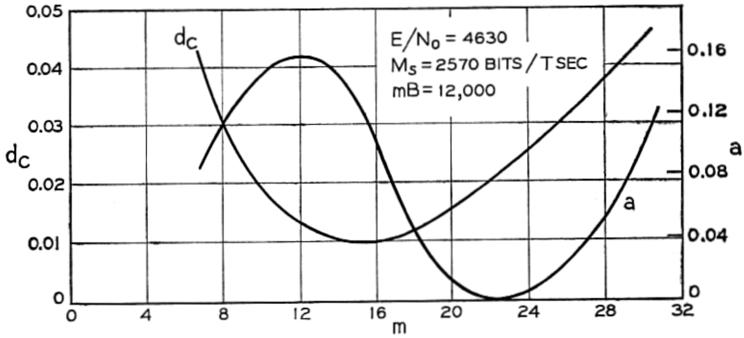

Fig. 9 — The influence of the modulator design in Example 4 on the minimum attainable distortion and the mismatch coefficient.

this example a compromise design with m about 19 would probably be best over a range of intermediate block lengths.

It is interesting to notice in this example that the coefficient a can be zero even when the source and channel are not matched. This is consistent with our previous interpretation of $a = 0$ as a necessary but not sufficient condition for matching. We remember that the coefficient a being zero does not imply that the lower bound in equation 43 is precisely d_C for all n. There are several other terms of $o(1/n)$ in this equation that have not been specified which are not necessarily zero when $a = 0$.

XI. THE UPPER BOUND

Now let us present an upper bound to the minimum attainable transmission distortion as a function of the coding block length. As with the lower bound, the upper bound approaches the asymptote d_C, but only as $[(\ln n)/n]^{1/2}$. The reason for the difference, we believe, is that within the upper bound derivation the transmitting signal set was restricted to contain at most $M = e^{nC}$ members, a restriction that was not necessary to impose in the lower bound. We also present an upper bound to the transmission distortion with a noiseless channel. This bound does agree, asymptotically, with the corresponding lower bound.

XII. THE RANDOM CODING ARGUMENT

All of the upper bound derivations in this paper use random coding arguments. That is, we do not explicitly find the encoder and decoder which, when used with $\mathcal{S}$ and $\mathcal{C}$, provide the distortion in the upper bound, but show that one pair does exist. More specifically, we construct a set of encoder-decoder pairs with a probabilistic rule according to which each system is selected to be used. This defines an ensemble of transmission systems, each with its own distortion, corresponding to all possible coding selections. What we calculate is a bound to the average distortion of this ensemble. Clearly, this provides an upper bound to the minimum distortion in the ensemble, hence to the minimum attainable distortion in any system that includes $\mathcal{S}$ and $\mathcal{C}$.

12.1 *The Construction of the Ensemble*

We denote the set of points on the rate distortion curve for $\mathcal{S}$ by (d_R, R) and assume the capacity of $\mathcal{C}$ to be C. We first choose any point (d^*, R^*) on the rate-distortion curve below (d_C, C) and design the code in such a way that the ensemble average distortion approaches d^* with increasing block length. We know this to be possible from Shannon's results.[2] Moreover, we expect, since the situation is somewhat analogous to a channel coding problem with $R^* < C$, that the distortion can be made to approach d^* exponentially fast. The point (d^*, R^*) is subsequently varied to obtain the best result at any particular block length of interest.

For any selection of (d^*, R^*), we then choose the number of signal points, $M = e^{nR}$, used to transmit $\mathcal{S}$. To attain a transmission distortion level d^*, we certainly must have the number of signal points large enough to represent the source to at least within d^*, and this requires that R be greater than R^*. We also require that R be less than C so that in the limit as n becomes large, we are guaranteed correct decoding among the signal points at the receiver. Therefore we have

$$R^* < R < C \tag{52}$$

and, for the corresponding values of distortion on the rate-distortion curve,

$$d_{\max} \geqq d^* > d_R > d_C \,. \tag{53}$$

The value of R can also later be chosen to optimize the result.

An ensemble of codes of length n is constructed for each selection of R and R^*. We use the probability distribution $p(\mathbf{x}, \mathbf{z})$ to generate the ensemble by picking, according to $p(\mathbf{x}, \mathbf{z})$, M independent pairs $(\mathbf{x}, \mathbf{z})$ from $X^n Z^n$. Thus we have a set of codes containing all possible mappings of the integers 1 through M into pairs of n-letter words $(\mathbf{x}, \mathbf{z})$, or $(JK)^{nM}$ codes in total. (We continue to use here the notation defined in the earlier part of the paper dealing with the lower bound.) Each of these codes has the associated probability

$$\Pr(\text{code}) = \prod_{i=1}^{M} p(\mathbf{x}_i, \mathbf{z}_i).$$

Any probability function $p(\mathbf{x}, \mathbf{z})$ could be used to obtain an upper bound, but we use a distribution that factors into $p(\mathbf{x})g(\mathbf{z})$; therefore, in the ensemble, each set of M decoded words, θ_1, is independent of each set of M channel input words, θ_2. Thus we can write

$$\Pr(\text{code}) = p(\theta_1, \theta_2) = p(\theta_1)p(\theta_2) = \prod_{i=1}^{M} p(\mathbf{x}_i) \prod_{i=1}^{M} g(\mathbf{z}_i).$$

Further, we use for $p(\mathbf{x})$ and $g(\mathbf{z})$ the product forms

$$\prod_{m=1}^{n} p(x^m) \quad \text{and} \quad \prod_{m=1}^{n} g(z^m)$$

in which the letter probability distribution $p(x)$ is that which yields a mutual information C on $\mathcal{C}$ and the letter probability distribution $g(z)$ is that which gives the output statistics on the test channel for $\mathcal{S}$ at the point (d^*, R^*) on the rate-distortion curve.

The encoding and decoding is done as follows: In every ensemble member there is a list θ_1 of allowed decoded words and a list θ_2 of usable channel input words. When a source output $\mathbf{w}$ occurs, the encoder scans θ_1 and chooses any member $\mathbf{z}_o$ in this list for which

$$d(\mathbf{w}, \mathbf{z}_o) \leqq d^*. \tag{54}$$

If there are none, the encoder chooses any member at all on the list θ_1, say $\mathbf{z}_1$. Since the lists are chosen together, there corresponds to $\mathbf{z}_o$ or $\mathbf{z}_1$ a particular $\mathbf{x}$ in θ_2, and this word is used to transmit $\mathbf{w}$. The decoder uses a maximum likelihood decision rule to decode $\mathbf{y}$ into a member of θ_2, which is then associated, through the pairings among the two lists, with a member $\mathbf{z}$ in θ_1. The resulting distortion, by definition, is $d(\mathbf{w}, \mathbf{z})$.

12.2 *The Ensemble Average Distortion*

Each member, θ, of the ensemble is a complete transmission system in itself, and has an average transmission distortion dependent upon the codes, θ_1 and θ_2, that are used. This average distortion, which is an average over all possible source and channel events, is equal to

$$d(\theta) = d(\theta_1, \theta_2) = \sum_{W^n} p(\mathbf{w}) \sum_{Y^n} p(\mathbf{y} \mid \mathbf{x})\, d(\mathbf{w}, \mathbf{z}).$$

The ensemble average distortion is obtained by averaging $d(\theta_1, \theta_2)$ over all choices of θ_1 and θ_2, hence

$$\langle d(\theta)\rangle_{av} = \sum_{W^n} p(\mathbf{w}) \sum_{Y^n} [\sum_{\theta_1} \sum_{\theta_2} p(\mathbf{y} \mid \mathbf{x})\, d(\mathbf{w}, \mathbf{z}) p(\theta_1) p(\theta_2)]. \qquad (55)$$

We next separate the events $\mathbf{w}$, θ_1, θ_2, and $\mathbf{y}$ into two sets: (*i*) those quadruples for which *either* there does not exist a $\mathbf{z}$ in θ_1 satisfying equation 54 *or* the received word $\mathbf{y}$ is decoded into a member of θ_2 different from the transmitted word $\mathbf{x}(\mathbf{w})$, and (*ii*) its complement. For quadruples in set one, the distortion $d(\mathbf{w}, \mathbf{z})$ is surely upper bounded by d_{max}, the maximum entry in $\| d(w, z) \|$. For those in the second set, we use equation 54 and the fact that the decoder returns us through $\mathbf{x}(\mathbf{w})$ to $\mathbf{z}_o$ to upper bound the distortion by d^*. Therefore, if the characteristic function Φ is used to indicate the quadruples in set one, we can upper bound the ensemble average with

$$\langle d(\theta)\rangle_{av} \leqq \sum_{W^n} p(\mathbf{w}) \sum_{Y^n} \sum_{\theta_1} \sum_{\theta_2} p(\mathbf{y} \mid \mathbf{x}) p(\theta_1) p(\theta_2) [d^*(1 - \Phi) + d_{max}\Phi]$$

$$= d^* + (d_{max} - d^*) \Pr(\Phi). \qquad (56)$$

Finally, we use the union bound to upper bound $Pr(\Phi)$ and the ensemble average distortion, $\langle d(\theta)\rangle_{av}$, to upper bound the minimum attainable transmission distortion, $d(\mathcal{S})$, and obtain the result in the next theorem.

Theorem 8: The minimum attainable transmission distortion of the source $\mathcal{S}$, *when used with the channel* $\mathcal{C}$, *satisfies*

$$d(\mathcal{S}) \leqq d^* + (d_{max} - d^*)[\Pr(\nexists\, \mathbf{z}_o \text{ in } \theta_1) + \Pr(\text{channel error})] \qquad (57)$$

in which $\exists'$ means "there does not exist," d^* is any distortion greater than d_C, and R (a variable in the bracketed terms) is any rate in the interval $R^* < R < C$. The bound is a function of n through the quantity in the brackets.

The last term in the brackets, the probability of error on the channel, has been approximated by many people, but we will use Gallager's bound[15]

$$\Pr(e) \leqq e^{-nE(R)} \qquad (58)$$

in which $E(R)$ is a positive monotonically increasing function of the difference $C - R$. The next section is devoted to the evaluation of the first term in the brackets, which is the probability that the source word $\mathbf{w}$ and the list θ_1 are such that equation 54 is not satisfied for any $\mathbf{z}$ in θ_1.

XIII. THE PROBABILITY OF FAILURE AT THE ENCODER

We say that failure occurs at the encoder, for the source output **w**, when each of the M allowed decoded words on list θ_1 are at a distortion $d(\mathbf{w}, \mathbf{z})$ from **w** greater than d^*. Because each of the M words in θ_1 is selected independently, we can write the total probability of this failure as

$$\begin{aligned}\Pr(\exists' \mathbf{z}_o \text{ in } \theta_1) &= \sum_{W^n} p(\mathbf{w}) \Pr(\exists' \mathbf{z}_o \text{ in } \theta_1 \mid \mathbf{w}) \\ &= \sum_{W^n} p(\mathbf{w})[1 - \Pr(\mathbf{z} \ni d(\mathbf{w}, \mathbf{z}) \leqq d^* \mid \mathbf{w})]^M.\end{aligned} \tag{59}$$

The last probability is seen equal to the distribution function of the distortion random variable described in Section 6.2 and defined by equations 16 and 17. In these equations $\mathbf{q} = q_1, q_2, \cdots, q_H$ is the composition vector of the source word **w**, and D_{ir} is the letter distortion random variable between the r'th appearance of the letter w_i in **w** and the corresponding letter in **z**.

We again notice that the distribution function of $d(\mathbf{w}, \mathbf{z})$ depends only upon the composition **q** of **w**. Thus we are able to perform the average over W^n in equation 59 as one over all possible compositions of **w**. All possible compositions can be represented as points in the $H - 1$ dimensional hyperplane within the first quadrant of R^H which intersects each axis q_i at one. This hyperplane is called the composition space Q^H. The probability of any composition point is equal to the product of the number of different source words having this composition and the probability of each, therefore, we have

$$\begin{aligned}P(\mathbf{q}) &= N(\mathbf{q}) \prod_{i=1}^{H} p_i^{nq_i} \\ &= \frac{n\,!}{\prod_{i=1}^{H} (nq_i)\,!} \prod_{i=1}^{H} p_i^{nq_i}.\end{aligned}$$

Interpreting $P(\mathbf{q})$ as an impulse function over Q^H we can now write equation 59 as

$$\Pr(\exists' \mathbf{z}_o \text{ in } \theta_1) = \int \underset{Q^H}{\cdots} \int P(\mathbf{q})[1 - G(d^* \mid \mathbf{q})]^M \, d\mathbf{q}. \tag{60}$$

To continue the inequality in equation 57, we require a lower bound to $G(d^*)$. For our present purpose, Fano's lower bound[12] is sufficient:

$$G(d^* \mid \mathbf{q}) \geqq K(n, \mathbf{q}) \exp n[\mu(s, \mathbf{q}) - s\mu'(s, \mathbf{q})] \\ \equiv K(n, \mathbf{q}) \exp - nR(d^*, \mathbf{q}) \tag{61}$$

in which

$$\mu'(s, \mathbf{q}) = d^* \tag{62}$$

$$0 < d^* \leqq E(d \mid \mathbf{q}) \tag{63}$$

$$\mu(s) = \sum_{i=1}^{H} q_i \ln \sum_{j=1}^{J} g_j \exp sd_{ij}$$

and $K(n, \mathbf{q})$ is a rather complex function of $\mathbf{q}$ and n that goes to zero algebraically in n with increasing n. Its precise form is otherwise unimportant in the following derivation. (The bound in equation 61 can still be used for points $\mathbf{q}$ that violate equation 63 if one uses the value of $s = 0$ rather than that which satisfies equation 62.) We can therefore write

$$\Pr(\exists' \mathbf{z}_o \text{ in } \theta_1) \leqq \int_{Q^H} \cdots \int P(\mathbf{q})[1 - K(n, \mathbf{q}) \exp - nR(d^*, \mathbf{q})]^{\exp nR} \, d\mathbf{q}. \tag{64}$$

The next step is to divide the composition space Q^H into two disjoint subspaces, Q and Q', that are defined by

$$Q = \{\mathbf{q}: R(d^*, \mathbf{q}) < R - \delta\} \tag{65}$$

$$Q' = \{\mathbf{q}: R(d^*, \mathbf{q}) \geqq R - \delta\} \tag{66}$$

with δ any positive number satisfying $R^* < R - \delta$. The idea behind this separation is illustrated in Fig. 10. The bracketed term in the integrand of equation 64 has the form $[1 - \exp(-nA)]^{\exp nB}$ which approaches zero with increasing n when $A < B$, and one when $A > B$. In the first region, which, except for the δ, corresponds to the set Q, we shall use the upper bound

$$[1 - \exp(-nA)]^{\exp nB} \leqq \exp[-\exp n(B - A)] \tag{67}$$

and in the second region, corresponding to Q', the (poorer) bound

$$[1 - \exp(-nA)]^{\exp nB} \leqq 1. \tag{68}$$

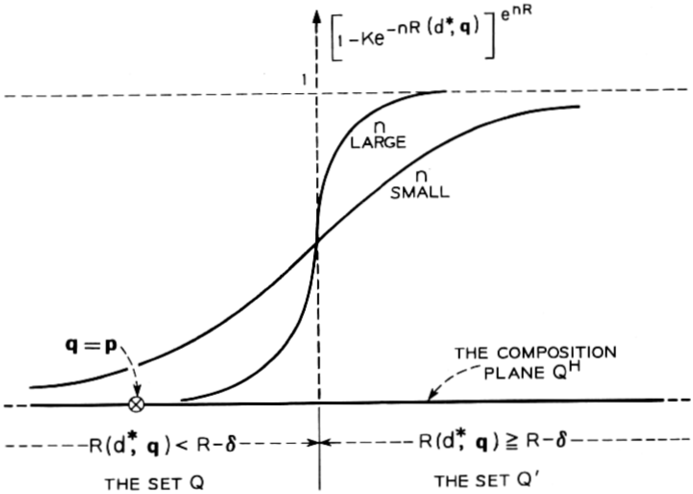

Fig. 10 — The division of the composition plane Q^H into the sets Q and Q'.

The use of these bounds in equation 64 results in

$$
\begin{aligned}
&\Pr\,(\exists\,'\mathbf{z}_o \text{ in } \theta_1) \\
&\quad \leqq \int_Q \cdots \int P(\mathbf{q}) \exp\{-K(n,\mathbf{q}) \exp n[R - R(d^*,\mathbf{q})]\}\, d\mathbf{q} \\
&\quad\quad + \int_{Q'} \cdots \int P(\mathbf{q})(1)\, d\mathbf{q} \\
&\quad \leqq \int_Q \cdots \int P(\mathbf{q}) \exp[-K(n,\mathbf{q})e^{n\delta}]\, d\mathbf{q} + \Pr\,(Q') \\
&\quad \leqq \exp[-K(n)e^{n\delta}] + \Pr\,(Q')
\end{aligned}
\tag{69}
$$

in which $K(n)$ denotes the minimum of $K(n, \mathbf{q})$ over Q. The first term in this upper bound is a double exponential in n which will turn out to be unimportant. Thus it remains to evaluate $Pr\ (Q')$.

We shall use what we call the hypercube method to upperbound $Pr(Q')$. Although the resulting bound is not as tight as others that could be derived (see, for example, the maximum probability point method in Ref. 8), it has the advantage of being simpler both to derive and to use and, in addition, does not seriously degrade the final bound to transmission distortion. What is done is to enclose the set Q' by another set Q_1' that has a relatively simple configuration, and to upper bound $Pr(Q')$ by $Pr(Q_1')$.

We construct in R^H a hypercube of dimension $2u$ centered at $\mathbf{q} = \mathbf{p}$,

$$K^H = \{\mathbf{q}: p_i - u \leqq q_i \leqq p_i + u\},$$

and intersect with it the composition space Q^H. The intersection forms a "solid" Q_1

$$Q_1 = Q^H \cap K^H$$

which contains vertices of the form $\mathbf{q}_v = q_{1v}, q_{2v}, \cdots q_{Hv}$, with the components, of course, summing to one. When H is even, q_{iv} equals either $p_i + u$ or $p_i - u$, and when H is odd, q_{iv} has the same values with the addition of one component equal to p_i. The vertices of Q_1 are joined by straight lines.

At this point we use the fact that Q is a convex set,[8] that is, for $0 \leqq \lambda \leqq 1$, $\lambda\mathbf{q}_a + (1 - \lambda)\mathbf{q}_b$ is a member of Q whenever both $\mathbf{q}_a$ and $\mathbf{q}_b$ are. This property ensures us that whenever the vertices of Q_1 are in the set Q, the entire set Q_1 is in Q,

$$Q_1 \subseteq Q,$$

with the consequence that

$$\Pr(Q') \leqq \Pr(Q_1'). \tag{70}$$

The remaining step is to bound the total probability of the set Q_1'. Because this probability equals the probability that *any* of the dependent events $q_i \varepsilon' [p_i - u, p_i + u]$ occurs, we can use the union bound to upper bound $Pr(Q_1')$ by the sum of the individual probabilities. Thus

$$\Pr(Q_1') \leqq \sum_{i=1}^{H} \Pr[q_i < p_i - u] + \Pr[q_i > p_i + u].$$

These quantities can be further upper bounded by a simple application of Chernov bounds. This has been done for us in Ref. 16, page 102, where the result found is, in our notation,

$$\Pr(Q_1') \leqq \sum_{i=1}^{H} e^{-nX_i} + e^{-nY_i} \tag{71}$$

in which

$$\left.\begin{matrix} X_i \\ Y_i \end{matrix}\right\} = -\ln\left[\left(\frac{p_i}{d_i}\right)^{d_i}\left(\frac{1-p_i}{1-d_i}\right)^{1-d_i}\right]$$

and

$$d_i = p_i - u \quad \text{for} \quad X_i$$
$$= p_i + u \quad \text{for} \quad Y_i .$$

In these bounds, the hypercube dimension $2u$ should be maximized, to obtain the tightest bound, subject only to the constraint that all vertices $\mathbf{q}_v$ be in region Q, that is, that they satisfy equation 65.

The bound in equation 71 can be simplified still further by writing

$$\Pr(Q_1') \leqq 2H \exp\,[-n \min\,(X_i\,,\,Y_i)]$$
$$\equiv K_1 \exp\, -\, nE_s(R). \tag{72}$$

Indeed, it can be shown,[8] that there are two, and not $2H$, candidates for the minimizing quantity in the exponent.

XIV. THE SET OF UPPER BOUNDS

Combining equations 57, 58, 69, and 72, we have the following result:

Theorem 9: The minimum attainable transmission distortion of the source $\mathcal{S}$*, when used with the channel* C*, satisfies*

$$d(\mathcal{S}) \leqq d^* + (d_{\max} - d^*)\{\exp\,[-K(n)e^{n\delta}]$$
$$+ K_1 \exp\,[-nE_s(R)] + \exp\,[-nE(R)]\} \tag{73}$$

for any d^* *and* R *that satisfy*

$$d_{\max} \geqq d^* > d_R > d_C \tag{74}$$
$$R^* < R < C. \tag{75}$$

The freedom provided by equations 74 and 75 can be used to generate a set of upper bounds, corresponding to all possible choices of d^* and R, the properties of which depend upon those of the two exponential functions in equation 73. It has been shown elsewhere[8] that $E_s(R)$ is a positive monotone increasing function of the difference $R - R^*$, that $E_s(R^*) = E_s'(R^*) = 0$, and that $E_s''(R^*) \neq 0$. Comparing these with the corresponding properties of the channel reliability function:[15] $E(R)$ a positive monotone increasing function of the difference $C - R$, $E(C) = E'(C) = 0$, $E''(C) \neq 0$; we see that the two functions are quite similar. Typically, their curves would look like those in Fig. 11.

With these curves, we can examine the behavior of the set of bounds in Theorem 9. As shown in Fig. 12, when d^* is chosen much larger than d_C, the nonzero slope of the rate-distortion curve allows

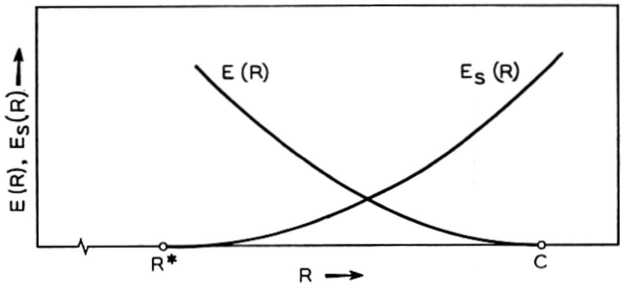

Fig. 11 — Typical behavior of $E_s(R)$ and $E(R)$ near their zero value.

a choice of R that can make both the differences $C - R$ and $R - R^*$ large. In turn, the exponents $E_s(R)$ and $E(R)$ in equation 73 are large and the exponential terms decay very rapidly with n. But for this choice, the asymptote d^* is much greater than the level d_C, which we know can be approached.

On the other hand, if we choose d^* only slightly greater than d_C, we have an upper bound with an asymptote that is nearly d_C, but now the differences $C - R$ and $R - R^*$, and therefore the exponents $E_s(R)$ and $E(R)$, are much smaller and the rate of approach to the asymptote d^* is correspondingly slower. Thus, in the selection of d^* and R there is a trade-off between a small asymptotic value and a fast rate of approach. This is illustrated in Fig. 13 in which we show a set of curves obtained from the upper bound expressions in equation 73. The best compromise for any value of n is given by the

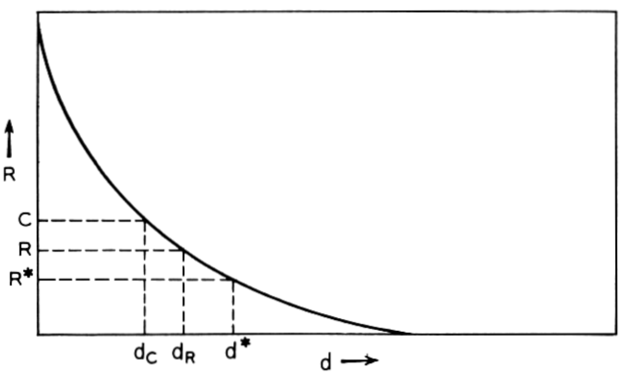

Fig. 12 — The rate-distortion curve for $\mathcal{S}$ illustrating the relations among the parameters in Theorem 9.

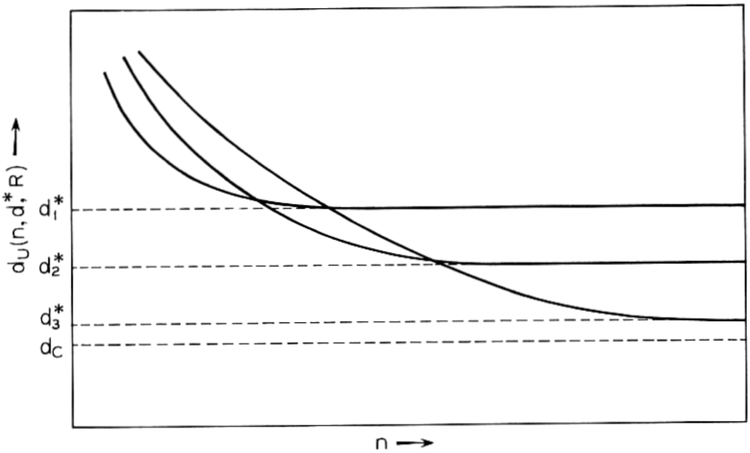

Fig. 13 — The upper bound in Theorem 9 with three different values for d^* and R.

lower envelope to the entire set of bounds in equation 73, therefore we have

Theorem 10: *The minimum attainable transmission distortion of the source* $\mathcal{S}$, *when used with the channel* $\mathcal{C}$, *satisfies*

$$d(\mathcal{S}) \leqq \min_{d^*, R} d_U(n, d^*, R) \equiv d_U(\mathcal{S}) \tag{76}$$

in which the function $d_U(n, d^*, R)$ is used to denote the right side of equation 73.

In the next section we study the asymptotic behavior of the lower envelope. At this point, though, we wish to include an important conclusion that can be established from the set of upper bounds in equation 73. Each individual bound indicates that, in a system where the distortion level d_C is attainable in the limit, if one would tolerate a distortion $d^* = d_C + \Delta$, this level could be approached exponentially fast as the coding block length is increased.

Actually, a much stronger statement is possible. Since the distortion curve for $d^* = d_C + \frac{1}{2}\Delta$ approaches this level in the limit, it must cross, at some finite n, the level $d_C + \Delta$. Because both curves are for the same source and channel, this proves that the distortion level $d_C + \Delta$ is not only approachable exponentially fast, it is in fact attainable with a finite coding block length. This is true for any $\Delta > 0$, no matter how small.

XV. THE ASYMPTOTIC BEHAVIOR OF THE UPPER BOUND

From the previous discussion it is clear that as n increases, the optimum value of d^* must approach d_C and therefore that the exponents $E_s(R)$ and $E(R)$ must approach zero. For this reason we use the Taylor series representations for these functions at R^* and C in equations 73 and 76, respectively, and obtain

$$d_U(\mathcal{S}) \approx \min_{d^*, R} \{d^* + (d_{\max} - d^*)$$

$$\cdot [K_1 \exp - nb_1(R - R^*)^2 + \exp - nb_2(C - R)^2]\} \tag{77}$$

with $b_1 = \frac{1}{2}E_s''(R^*)$ and $b_2 = \frac{1}{2}E''(C)$. In using the Taylor series for $E(R)$ and $E_s(R)$ we have dropped the cubic terms since both $E'''(C)$ and $E_s'''(R^*)$ are finite and $C - R$ and $R - R^*$ are $o(1)$. The double exponential term involving δ is also dropped since it can be shown to contribute nothing important in the asymptotic bound.

We next avoid the minimization on R by choosing that value of R which equates the two exponents:

$$b_1(R - R^*)^2 = b_2(C - R)^2. \tag{78}$$

While this selection of R is nonoptimum for finite n, it can be shown that it asymptotically approaches R_{opt}, and that it does not affect the asymptotic behavior of the upper bound. This particular choice of R allows us to combine the two exponential terms in equation 77. If we start with equation 78 and the obvious equality

$$(C - R) + (R - R^*) = C - R^*,$$

we can establish

$$(C - R) = \frac{\sqrt{b_1}}{\sqrt{b_1} + \sqrt{b_2}}(C - R^*) \tag{79}$$

$$(R - R^*) = \frac{\sqrt{b_2}}{\sqrt{b_1} + \sqrt{b_2}}(C - R^*), \tag{80}$$

which further allows us to write the two exponents in terms of the common difference $C - R^*$.

Next, we wish to express the difference $C - R^*$ in terms of the difference $d_C - d^*$. Taylor's formula with remainder is again used:

$$R(d^*) = R(d_C) + R'(d_C)(d^* - d_C) + o(d^* - d_C)$$

or

$$C - R^* = -R'(d_C)(d^* - d_C) - o(d^* - d_C) \tag{81}$$
$$= -s_o(d^* - d_C) - o(d^* - d_C).$$

In the last equation we have used the fact that the slope of the rate distortion curve at the point (d_C, C) is equal to the value of s which satisfies $\mu(s) - s\mu'(s) = -C$.[7, 8]

Finally, we substitute equations 79, 80, and 81 into equation 77, subtract d_C from both sides of this last equation, and change the minimizing variable to $d^* - d_C$ to obtain

$$d(\mathcal{S}) - d_C \leqq \min_x \,[x + (A - x)K_2 \exp - Bnx^2] \tag{82}$$

in which $x = d^* - d_C$, $A = d_{\max} - d_C$, $K_2 = K_1 + 1 = 2H + 1$, and

$$B = b_1 b_2 s_o^2/(\sqrt{b_1} + \sqrt{b_2})^2.$$

We next find the asymptotic behavior of the lower envelope in equation 82.

If x is considered the parameter, each function of n in the set $f(x, n)$ starts at $f(x, 0) = x + (A - x)K_2$ and decreases exponentially to $f(x, \infty) = x$. For any two parameter values, x_1 and x_2, with $x_1 > x_2$ we have

$$f(x_1, 0) - f(x_2, 0) = (1 - K_2)(x_1 - x_2)$$
$$= -2H[f(x_1, \infty) - f(x_2, \infty)].$$

Consequently, any two curves must cross as in Fig. 14.

It follows that the parameter $x_o(n)$, which identifies the minimum of $f(x, n_o)$ at the value $n = n_o$, must change with n. Since this parameter is the solution of

$$f'_x(x, n) = 0,$$

we have

$$\exp\,(nBx_0^2) - K_2 = 2nK_2Bx_0(A - x_0). \tag{83}$$

Figure 15 shows the required graphical solution which clearly always exists. The substitution of $x_o(n)$ in $f(x, n)$ specifies the single function of n, $f[x_o(n), n]$, which is the desired lower envelope. Unfortunately, an explicit solution is not possible for $x_o(n)$, nor for $f[x_o(n), n]$, but we can obtain bounds to both that are adequate for our purposes.

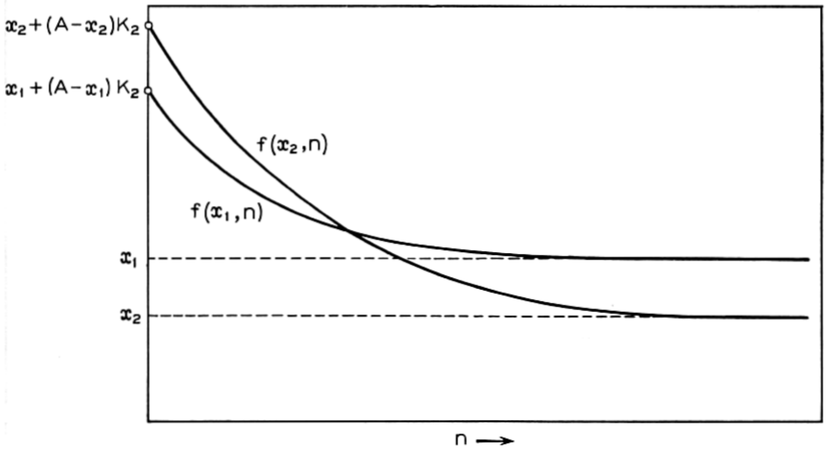

Fig. 14 — Two members of the family of curves: $f(x,n) = x + (A - x)K_2 \exp(-Bnx^2)$.

From the graphical solution in Fig. 15, we see that any conjectured solution, x_o?, must be too large if, in equation 83, the left side exceeds the right and too small if the reverse is true. This criterion could also be used on a trial functional solution $x_o(n)$?. Now, if the left side of equation 83 is functionally stronger in n than the right, we know that our trial solution $x_o(n)$? is too strong in n. Again the reverse is also true.

After several guesses we are led to the trial functional solution $x_o(n) = [a(\ln n)/Bn]^{1/2}$ with which the right side of equation 83 is greater than the left for a $\leq 1/2$, and the reverse is true for $a > 1/2$.

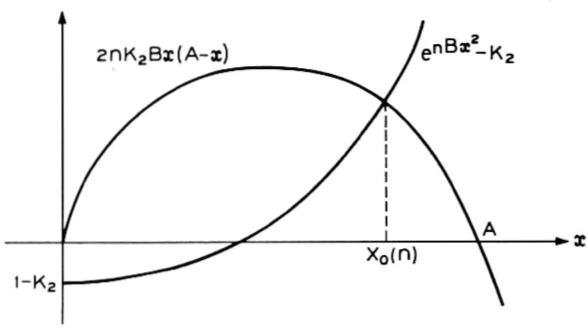

Fig. 15 — The graphical solution of equation 83.

This determines the highest order term of $x_o(n)$ and we can write

$$(\tfrac{1}{2})^{\frac{1}{2}}\left(\frac{\ln n}{Bn}\right)^{\frac{1}{2}}[1 + o(1)] \leqq x_o(n) \leqq (\tfrac{1}{2} + \epsilon)^{\frac{1}{2}}\left(\frac{\ln n}{Bn}\right)^{\frac{1}{2}}[1 + o(1)].$$

It follows that

$$f[x_o(n), n] \geqq \left(\frac{1}{2B}\right)^{\frac{1}{2}}\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}[1 + o(1)]$$

and, since the lower envelope is smaller than any individual $f(x, n)$, that

$$f[x_o(n), n] \leqq f\left[\left(\frac{1}{2B}\right)^{\frac{1}{2}}\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}, n\right] = \left(\frac{1}{2B}\right)^{\frac{1}{2}}\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}[1 + o(1)]. \tag{84}$$

Although only an upper bound to $f(x, n)$ is required, both upper and lower bounds were found to show that the method used to obtain the desired lower envelope provides asymptotically tight results. Continuing the inequality in equation 82 by that in equation 84 provides our final upper bound to transmission distortion.

Theorem 11: The minimum attainable transmission distortion of the source $\mathcal{S}$*, when used with the channel* $\mathcal{C}$*, is upper bounded by*

$$d(\mathcal{S}) \leqq d_C + b\left(\frac{\ln n}{n}\right)^{\frac{1}{2}}[1 + o(1)] \tag{85}$$

in which

$$b = \left(\frac{1}{2B}\right)^{\frac{1}{2}} = \frac{1}{(2)^{\frac{1}{2}}}\frac{1}{\mid s_o \mid}\left[\frac{1}{(b_1)^{\frac{1}{2}}} + \frac{1}{(b_2)^{\frac{1}{2}}}\right]$$

$$b_1 = \tfrac{1}{2}E_s''(R^* = C)$$

$$b_2 = \tfrac{1}{2}E''(C).$$

For a fixed source $\mathcal{S}$, we see from this theorem that the coefficient b is smallest when $\mathcal{S}$ is used with that channel (among those of equal capacity) for which the constant b_2 is largest. In the same way, the coefficient b is seen to be a decreasing function of b_1 when the channel is fixed. Since the constant b_2 is independent of the source and b_1 independent of the channel, our upper bound does not provide an indicator of matching between the source and channel as we obtained in the lower bound. This was actually expected since here we were forced to separate the source and channel with an interface containing at most e^{nC} points.

The coefficient b_1, though, has an interesting significance. It is equal to one-half the derivative $E_s''(R^* = C)$ which can be thought to indicate how fast the boundary of Q' initially moves away from $\mathbf{p}$ with increasing R. In turn, this indicates, in a reciprocal manner, the necessary rate of change of the rate required to handle source words with compositions just around $\mathbf{p}$, which are just less than typical. Thus, we can think of the coefficient b_1 as a type of "stretch factor"[16] for the source.

When the result in equation 85 is compared with the lower bound to distortion, we see that the $[(\ln n)/n]^{\frac{1}{2}}$ rate of approach to d_C is slower than the $1/n$ rate of approach of the lower bound. Mathematically, at least, the reason for the upper bound decreasing more slowly than $(1/n)^{\frac{1}{2}}$ is that, for small arguments, the lowest order term in the two exponents $E(R)$ and $E_s(R)$ is quadratic. Their form for large n, exp $-n(\Delta R)^2$, shows that values of ΔR larger than $(1/n)^{\frac{1}{2}}$ are required to have these terms go to zero with increasing n. Because the slope of the rate-distortion curve is nonzero, the corresponding values of distortion difference (Δd) must also be larger than $(1/n)^{\frac{1}{2}}$.

There is reason to think that this type of exponential term, and the consequential $[(\ln n)/n]^{\frac{1}{2}}$ rate of approach to d_C , is present in the upper bound because we have used threshold devices in the transmission system. One at the encoder leads to the first exponential term in equation 73 (we again disregard the double exponential term). It uses the rule in equation 54 to choose, for each source word $\mathbf{w}$, any decoder word $\mathbf{z}$ in list θ_1 at a distortion less than d^*. When list θ_1 is lacking such an entry, any $\mathbf{z}$ at all on the list is chosen which, since the members of θ_1 are chosen independently, is then independent of $\mathbf{w}$. The resulting distortion in this circumstance is usually much greater than d^*. In the next section we compare the performance of this encoder with another that does not use such a threshold and show that the source encoding alone need only contribute to a rate of approach to d_C equal to $(\ln n)/n$.

A second threshold operation in our system is at the channel decoder, but it is really dependent upon the coding of the entire system. It leads to the second exponential term in equation 73. To isolate its effect on the system performance, we assume that failure has not occurred at the encoder, that is, there does exist a $\mathbf{z}$ on θ_1 with $d(\mathbf{w}, \mathbf{z}) \leqq d^*$. Now if the channel decoder makes no error, we are assured that the resulting distortion is less than d^*. However, if an error is made, the believed channel input word $\mathbf{x}_1$ is different from the actual word $\mathbf{x}$; therefore the decoded word $\mathbf{z}_1$ is different from $\mathbf{z}_o$. Moreover, since the lists θ_1 and θ_2 are chosen independently, $\mathbf{z}_o$ and $\mathbf{z}_1$ are statistically independent. It follows that $\mathbf{z}_1$ and $\mathbf{w}$ are also statistically independent, and in consequence that the distortion $d(\mathbf{w}, \mathbf{z}_1)$ is usually much greater than d^*.

It is this threshold which, it is believed, cannot be eliminated when the signal space is constrained to contain at most $M = e^{nC}$ points, even if the lists θ_1 and θ_2 are chosen dependently. A heuristic argument in Ref. 8 suggests that with such a constrained signal set, the transmission distortion can approach d_C no more rapidly than as $n^{-\frac{1}{2}}$. This, of course, is a slower rate of approach to d_C than the a/n rate of approach of the corresponding lower bound to distortion that was derived using a signal set not constrained in size.

XVI. AN IMPROVED UPPER BOUND FOR NOISELESS CHANNELS

For the special case of a noiseless channel, the previously derived upper bound can be improved. Since such a channel contains e^C noiseless transitions, or "direct" paths, transmission of the encoder output is trivial and the communication problem is only one of source representation. For this representation we are allowed to choose, from an e^C letter representation alphabet, one representation letter for every source output letter. Just as one is allowed n uses of the channel to transmit an n-letter source output, one is allowed an n-letter representation word to approximate an n-letter source word.

We first state that if the threshold source encoder defined by equation 54 is used in the ensemble of representation codes θ_1 of Section XII, the ensemble average representation error is very similar to the ensemble average transmission error derived in the previous sections. The only difference in the derivation is that the Pr(channel error) term is no longer present in equation 57, nor in any succeeding equation, with the only result being that $b_2 = \infty$ in equation 85.

We note here that this particular result is valid only for sources that are not doubly-uniform, that is, having a uniform probability distribution and a distortion matrix in which all rows are permutations of one row vector and all columns are permutations of one column vector. The reason for this exclusion is that for doubly-uniform sources the exponential term in equation 73 involving $E_s(R)$ also vanishes, and the double exponential term involving δ, previously dropped as insignificant, now remains as the only term. It is instructive to delay further evaluation of the bound in this case until after the following upper bound to representation distortion is derived.

16.1 *Optimum Source Encoder*

We now derive an upper bound to the source representation error when an optimum source encoder is used in place of the threshold encoder of the previous section. The resulting upper bound will be seen to approach the asymptote, d_C, as $(\ln n)/n$. This represents an improvement upon the best previously known upper bound to source representation distortion[7] which approached d_C essentially as $n^{-1/2}$.

The coding ensemble used here is very similar to the set of codes, θ_1, used in Section XII. But now the size of the set, M, is set equal to e^{nC} for all n, rather than have it approach this size with increasing n. And, the probability with which each ensemble member is used,

$$\Pr(\text{code}) = p(\theta_1) = \prod_{i=1}^{M} g(\mathbf{z}_i),$$

is now governed by that probability distribution $g(\mathbf{z})$ equal to the output probability distribution of the test channel at the point (d_C, C) on the rate distortion curve for $\mathcal{S}$. Within each ensemble member the encoder chooses, for any occurring source word $\mathbf{w}$, that member $\mathbf{z}$ on θ_1 for which $d(\mathbf{w}, \mathbf{z})$ is minimum. Therefore, for each ensemble member the average distortion over all possible source events is

$$d(\theta_1) = \sum_{W^n} p(\mathbf{w})[\min_{\substack{1 \leq i \leq M \\ \mathbf{z}_i \in \theta_1}} d(\mathbf{w}, \mathbf{z}_i)]. \tag{86}$$

The ensemble average distortion is given by

$$\langle d(\theta_1)\rangle_{\text{av}} = \sum_{W^n} p(\mathbf{w}) \sum_{\theta_1} p(\theta_1)[\min_{\substack{1 \leq i \leq M \\ \mathbf{z}_i \in \theta_1}} d(\mathbf{w}, \mathbf{z}_i)]. \tag{87}$$

The set of quantities $d(\mathbf{w}, \mathbf{z}_i)$ in equation 87 could be thought of as a set of M independent and identically distributed random variables, each conditioned on $\mathbf{w}$ and governed by the word probability distribution $g(\mathbf{z})$. The minimum of this set, $d_{\min}(\mathbf{w})$, is then also a random variable, governed by the code probability distribution $p(\theta_1)$. The inner sum in equation 87 is, therefore, the expected value of $d_{\min}(\mathbf{w})$ and we can write

$$\langle d(\theta_1)\rangle_{\text{av}} = \sum_{W^n} p(\mathbf{w}) \int_0^{d_{\max}} d\, dF_{d_{\min}|\mathbf{w}}(d \mid \mathbf{w})$$

which, upon integration by parts, becomes

$$\langle d(\theta_1)\rangle_{\text{av}} = \sum_{W^n} p(\mathbf{w}) \int_0^{d_{\max}} [1 - F_{d_{\min}|\mathbf{w}}(d \mid \mathbf{w})]\, dd. \tag{88}$$

The conditional distortion random variables $d(\mathbf{w}, \mathbf{z}_i)$ are the same distortion variables used in Section XIII. Since they depend only upon the composition of $\mathbf{w}$, we can again perform the summation in equation

88 by integration over the composition space, thus

$$\langle d(\theta_1)\rangle_{av} = \int_{Q^H} \cdots \int P(\mathbf{q})\, d\mathbf{q} \int_0^{d_{max}} [1 - F_{d_{min}|\mathbf{q}}(d \mid \mathbf{q})]\, dd \tag{89}$$

$$\equiv \int_{Q^H} \cdots \int P(\mathbf{q})\, d\mathbf{q} \langle d_{min}(\mathbf{q})\rangle_{av} . \tag{90}$$

The inner integrand in equation 89 is the probability that all M points on θ_1 have a distortion $d(\mathbf{w}, \mathbf{z})$ from $\mathbf{w}$ greater than d. Using the independence property of the members of θ_1, we can write this probability as

$$1 - F_{d_{min}|\mathbf{q}}(d \mid \mathbf{q}) = [1 - G(d \mid \mathbf{q})]^M. \tag{91}$$

It can be seen from equation 16 that the variance of the variable d is proportional to $1/n$ for every $\mathbf{q}$. Therefore the function $[1 - G(d \mid \mathbf{q})]$, which for every n decreases monotonically from one to zero, approaches, with increasing n, a negative step at the value of distortion $d = E(d \mid \mathbf{q})$.

The same is also true of $[1 - G(d \mid \mathbf{q})]^M$ which approaches a negative step at some lower value of distortion, $d_C(\mathbf{q})$. This can be established using the following asymptotic upper and lower bounds to the distribution function $G(d \mid \mathbf{q})$ which are from Shannon[11] and Gallager[13]:

$$h(n, \mathbf{q}) \exp -nR(d, \mathbf{q}) \leqq G(d \mid \mathbf{q}) \leqq H(n, \mathbf{q}) \exp -nR(d, \mathbf{q}) \tag{92}$$

with

$$R(d, \mathbf{q}) \equiv \mu(s, \mathbf{q}) - s\mu'(s, \mathbf{q}) \tag{93}$$

$$0 < \mu'(s, \mathbf{q}) = d \leqq E(d \mid \mathbf{q})$$

and in which $h(n, \mathbf{q})$ and $H(n, \mathbf{q})$ are algebraically small functions of n. Therefore, within the range $0 < d \leqq E(d \mid \mathbf{q})$, the function in equation 91 can be bounded by

$$[1 - He^{-nR}]^{\exp nC} \leqq [1 - G(d \mid \mathbf{q})]^M \leqq [1 - he^{-nR}]^{\exp nC}; \tag{94}$$

which proves that $[1 - G(d \mid \mathbf{q})]^M$ must approach one when $R(d, \mathbf{q}) > C$ and zero when $R(d, \mathbf{q}) < C$. That the function $R(d, \mathbf{q})$ is monotone decreasing in d within $0 < d \leqq E(d \mid \mathbf{q})$ now establishes the stated limiting step function form of $[1 - G(d \mid \mathbf{q})]^M$ with $d_C(\mathbf{q})$ equal to the distortion value for which

$$R[d_C(\mathbf{q}), \mathbf{q}] = C. \tag{95}$$

The region of integration in equation 89 is thus conveniently divided into two parts: one over $[0, d_C(\mathbf{q}) + \Delta]$ in which the integrand is upper-bounded by unity, and the other $[d_C(\mathbf{q}) + \Delta, d_{\max}]$ in which the integrand is upper-bounded by its value at the lower limit. The result is

$$\langle d_{\min}(\mathbf{q})\rangle_{\text{av}} \leqq d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta][1 - G(d_C(\mathbf{q}) + \Delta \mid \mathbf{q})]^M \tag{96}$$

which, with the use of the lower bound in equation 92, can be continued by

$$\langle d_{\min}(\mathbf{q})\rangle_{\text{av}} \leqq d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta] \cdot \{1 - h \exp[-nR(d_C(\mathbf{q}) + \Delta, \mathbf{q})]\}^{\exp nC}.$$

Equation 67 allows the further continuation of this bound by:

$$\langle d_{\min}(\mathbf{q})\rangle_{\text{av}} \leqq d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta] \cdot \exp(-h \exp\{n[C - R(d_C(\mathbf{q}) + \Delta, \mathbf{q})]\}). \tag{97}$$

Again the monotone decreasing property of $R(d, \mathbf{q})$ in d provides that the quantity $C - R(d_C(\mathbf{q}) + \Delta, \mathbf{q})$ is positive when Δ is positive and, therefore, that the last term in equation (97) is a decreasing double exponential in n.

Equation 97 actually provides, for each $\mathbf{q}$, a set of upper bounds to $\langle d_{\min}(\mathbf{q})\rangle_{\text{av}}$ very similar to the family of curves studied in Section XV. In the choice of the parameter Δ there is once again a trade-off between a small asymptote, $d_C(\mathbf{q}) + \Delta$, and a fast rate of approach. It should, in general, be chosen to optimize the bound at each n. Since we want an upper bound to $\langle d_{\min}(\mathbf{q})\rangle_{\text{av}}$ that approaches $d_C(\mathbf{q})$ with increasing n, the optimizing parameter $\Delta_o(n)$ clearly must approach zero as n increases. But $\Delta_o(n)$ must approach zero in a way that also allows the last term of equation 97 to vanish.

Since an asymptotic bound is our goal, we extract the essential behavior of this term for small Δ by forming a Taylor series of $R(d, \mathbf{q})$ at $d = d_C(\mathbf{q})$:

$$\begin{aligned} C - R(d_C(\mathbf{q}) + \Delta, \mathbf{q}) &= -\Delta R'(d_C(\mathbf{q}), \mathbf{q}) + o(\Delta) \\ &= -s\Delta + o(\Delta). \end{aligned}$$

In this expression s is the parameter value in equation 93 when d equals $d_C(\mathbf{q})$. Thus the lower envelope to the set of bounds in equation 97 can be written, for the purpose of an asymptotic bound, as

$$\langle d_{\min}(\mathbf{q})\rangle_{\text{av}} \leqq \min_{\Delta} \{d_C(\mathbf{q}) + \Delta + [d_{\max} - d_C(\mathbf{q}) - \Delta] \exp(-he^{-sn\Delta})\}.$$

The minimization is found using the same method used in Section XV. In this process, it is important to notice that Shannon's coefficient $h(n, \mathbf{q})$ in equation 92 is proportional to $n^{-\frac{1}{2}}$. The result is that the optimizing parameter satisfies

$$\frac{1}{2}\frac{\ln n}{-sn}[1 + o(1)] \leqq \Delta_o(n) \leqq \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}[1 + o(1)]$$

and that $\langle d_{\min}(\mathbf{q})\rangle_{\text{av}}$ satisfies

$$\langle d_{\min}(\mathbf{q})\rangle_{\text{av}} \leqq d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}[1 + o(1)]. \tag{98}$$

Returning to equation 90, the ensemble average representation error therefore can be upper bounded by

$$\langle d(\theta_1)\rangle_{\text{av}} \leqq \int \cdots \int_{Q^H} P(\mathbf{q})\left[d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}\right] d\mathbf{q}. \tag{99}$$

The above integral is evaluated in the same way similar averages were found for the lower bound. The bracketed quantity is expanded in a Taylor series about $\mathbf{q} = \mathbf{p}$ and is truncated after three terms with a Lagrange remainder term. Upon integration of this expansion we find

$$\begin{aligned}\langle d(\theta_1)\rangle_{\text{av}} \leqq\ & d_C(\mathbf{p}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-s_o n} \\ &+ \sum_i \frac{\partial}{\partial q_i}\left[d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}\right]_{\mathbf{p}} E(q_i - p_i) \\ &+ \sum_{ij} \frac{\partial^2}{\partial q_i\,\partial q_j}\left[d_C(\mathbf{q}) + \left(\frac{1}{2} + \epsilon\right)\frac{\ln n}{-sn}\right]_{\varphi} E[(q_i - p_i)(q_j - p_j)]\end{aligned} \tag{100}$$

with $s_o \equiv s(\mathbf{p})$ and $\varphi \in Q^H$.

Using the following expected values in equation (100),

$$E(q_i - p_i) = 0$$

$$E[(q_i - p_i)(q_j - p_j)] = \frac{1}{n}(p_i\,\delta_{ij} - p_i p_j),$$

we have the following upper bound to the ensemble average distortion and, therefore, to the minimum attainable representation error.

Theorem 12: The minimum attainable transmission distortion (representation distortion) of the source $\mathcal{S}$, *when used with a noiseless channel* *of capacity C, is upper bounded by*

$$d(\mathcal{S}) \leqq d_C + \left(\frac{1}{2} + \epsilon\right) \frac{\ln n}{-s_o n} [1 + o(1)] \tag{101}$$

in which s_o satisfies

$$\mu(s_o, \mathbf{p}) - s_o \mu'(s_o, \mathbf{p}) = -C.$$

Except for the arbitrarily small positive ϵ, the bound in equation 101 agrees precisely with the asymptotic lower bound that we found earlier in this paper.

We see by comparing equation 85 (with $b_2 = \infty$ for the noiseless channel) and equation 101 that the replacement of the threshold source encoder with an optimum encoder increases the rate of approach to the asymptote from $[(\ln n)/n]^{\frac{1}{2}}$ to $(\ln n)/n$. To obtain some feeling for the reason for this improvement, we might think of the optimum encoder as a threshold encoder, *but* with a threshold that varies depending on the particular source output. Indeed, we used this step within the mathematics when we separated all events (equation 96) into two sets with the separation dependent upon the source word. In particular, for any source output word with composition $\mathbf{q}$, we used a threshold, $d_C(\mathbf{q}) + \Delta$, just large enough so that for large n there is almost surely a representation word in θ_1 that is acceptable. It does not require, as does the fixed threshold encoder, that the set of source words not meeting a fixed distortion level of d^* have a total probability that goes to zero with n. This restriction is really more severe than one would think we need, since some of the source words $\mathbf{w}$ discarded by the fixed threshold encoder are just outside $\mathbf{p}$, having characteristics just less than typical, for which some of the distortions $d(\mathbf{w}, \mathbf{z}_i)$ might be only marginally greater than any fixed d^*.

16.2 *The Special Case of a Double Uniform Source*

There is one situation for which both source encoders provide a representation distortion that approaches the limit d_C as $(\ln n)/n$. This is when the source $\mathcal{S}$ is doubly-uniform. Since $\mu(s, \mathbf{q})$ is independent of $\mathbf{q}$ for such a source, $R(d^*, \mathbf{q})$ in equation 61 is also independent $\mathbf{q}$, with the result that the set Q' in equation 66 is always empty. Therefore, $\Pr(Q') = 0$ in equation 69 and we have for the set of upper bounds to representation distortion, using threshold encoders:

$$d(\mathcal{S}) \leqq d^* + (d_{\max} - d^*) \exp(-he^{n\delta}).$$

In this bound we have used the lower bound in equation **92** rather than that in equation 61. It can now be shown, using precisely the same procedure as before, that this set of bounds approaches the limit d_C as $(\ln n)/n$.

XVII. SUMMARY

We have presented upper and lower bounds to the minimum attainable transmission distortion of a source measured by a specified distortion measure. The bounds, which were derived for both noisy and noiseless channels, have all been shown to converge to the same level of distortion, d_C, algebraically in the block length n. The quantity d_C is that level of distortion shown by Shannon to be the minimum attainable transmission distortion when the channel capacity is C and arbitrarily complex transmission methods are allowed.

For noisy channels, the rate of approach of the lower bound to d_C is as a/n and that of the upper bound as $b[(\ln n/n)]^{1/2}$. The nonnegative coefficients a and b are both functions of the statistics of the source and channel, but have different forms. The lower bound coefficient, a, interrelates these statistics in such a way as to suggest its utility as a measure of "mismatch" between the source and channel, the larger a, the slower the rate of approach of the bound to d_C, and the larger the source-channel mismatch. This coefficient is, of course, necessarily equal to zero whenever the source and channel are perfectly matched, that is, whenever the minimum attainable transmission distortion is equal to d_C for all block lengths, n.

The coefficient b in the upper bound, though, does not present an indicator of source-channel mismatch. It is the sum of two terms which separately contain the source statistics and the channel statistics. The cause of this separation is the interface between the source and channel that results from the use of a transmitting signal set constrained to contain at most e^{nC} members, a constraint which we found necessary to introduce in the development of the bound.

For noiseless channels, both the upper and lower bounds to the transmission distortion (or the source representation distortion) have the same form. They both have been shown to approach the asymptote d_C as $a_1\ (\ln n)/n$.

REFERENCES

1. Shannon, C. E., "A Mathematical Theory of Communications," B.S.T.J., 27, Nos. 3 and 4 (July and October 1948), pp. 379–423, 623–656.

2. Shannon, C. E., "Coding Theorems for a Discrete Source with a Fidelity Criterion," IRE National Convention Record, Part 4 (1959), pp. 142–163.
3. Holsinger, J. L., unpublished work.
4. Goblick, T. J., "Theoretical Limitations on the Transmission of Data from Analog Sources," IEEE Trans. Inform. Theory, *IT-11* (October 1965), pp. 558–567.
5. Gerrish, A. M. and Shultheiss, P. M., "Information Rates on Non-Gaussian Processes," IEEE Trans. Inform. Theory *IT-10* (October 1964), pp. 265–271.
6. Pinkston, J. T., "Information Rates of Independent Sample Sources," S. M. Thesis, Department of Electrical Engineering, M.I.T., Cambridge, Massachusetts (1966).
7. Goblick, T. J., "Coding for a Discrete Information Source with a Distortion Measure," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., Cambridge, Massachusetts (1962).
8. Pilc, R. J., "Coding Theorems for Discrete Source-Channel Pairs," Ph.D. Thesis, Department of Electrical Engineering, M.I.T., Cambridge, Massachusetts (1967).
9. Pilc, R. J., unpublished work.
10. Chernov, H., "A Measure of Asymptotic Efficiency for Tests of an Hypothesis Based on a Sum of Observations," Ann. Math. Stat. *23* (1952), pp. 493–507.
11. Shannon, C. E., unpublished work.
12. Fano, R. M., *The Transmission of Information,* New York: Wiley, 1961.
13. Gallager, R. G., "Lower Bounds on the Tails of Probability Distributions," M.I.T. Research Lab. of Electronics, Quart. Progress Report, *77* (April 1965), pp. 277–291.
14. Gallager, R. G., unpublished work.
15. Gallager, R. G., "A Simple Derivation of the Coding Theorem and Some Applications," IEEE Trans. Inform. Theory, *IT-11* (January 1965), pp. 3–18.
16. Wozencraft, J. M. and Jacobs, I. M., *Principles of Communication Engineering,* New York: John Wiley (1965).