Coding Theorems for a Discrete Source With a Fidelity Criterion*

Claude E. Shannon**

Abstract

Consider a discrete source producing a sequence of message letters from a finite alphabet. A single-letter distortion measure is given by a non-negative matrix (d_{ij}). The entry d_{ij} measures the "cost" or "distortion" if letter i is reproduced at the receiver as letter j. The average distortion of a communications system (source-coder-noisy channel-decoder) is taken to be $d = \sum_{i,j} P_{ij} d_{ij}$ where P_{ij} is the probability of i being reproduced as j. It is shown that there is a function $R(d)$ that measures the "equivalent rate" of the source for a given level of distortion. For coding purposes where a level d of distortion can be tolerated, the source acts like one with information rate $R(d)$. Methods are given for calculating $R(d)$, and various properties discussed. Finally, generalizations to ergodic sources, to continuous sources, and to distortion measures involving blocks of letters are developed.

In this paper a study is made of the problem of coding a discrete source of information, given a *fidelity criterion* or a *measure of the distortion* of the final recovered message at the receiving point relative to the actual transmitted message. In a particular case there might be a certain tolerable level of distortion as determined by this measure. It is desired to so encode the information that the maximum possible signaling rate is obtained without exceeding the tolerable distortion level. This work is an expansion and detailed elaboration of ideas presented earlier [1], with particular reference to the discrete case.

We shall show that for a wide class of distortion measures and discrete sources of information there exists a function $R(d)$ (depending on the particular distortion measure and source) which measures, in a sense, the equivalent rate R of the source (in bits per letter produced) when d is the allowed distortion level. Methods will be given for evaluating $R(d)$ explicitly in certain simple cases and for evaluating $R(d)$ by a limiting process in more complex cases. The basic results are roughly that it is impossible to signal at a rate faster than $C / R(d)$ (source letters per second) over a memoryless channel of capacity C (bits per second) with a distortion measure less than or equal to d. On the other hand, by sufficiently long block codes it is possible to approach as closely as desired the rate $C / R(d)$ with distortion level d.

Finally, some particular examples, using error probability per letter of message and other simple distortion measures, are worked out in detail.

The Single-Letter Distortion Measure. Suppose that we have a discrete information source producing a sequence of letters or "word" $m = m_1, m_2, m_3, \ldots, m_t$, each chosen from a finite alphabet. These are to be transmitted over a channel and reproduced, at least

* Institute of Radio Engineers, *International Convention Record*, vol. 7, 1959.

** This work was supported in part by the U.S. Army (Signal Corps), the U.S. Air Force (Office of Scientific Research, Air Reserve and Development Command), and the U.S. Navy (Office of Naval Research).

approximately, at a receiving point. Let the reproduced word be $Z = z_1, z_2, \ldots, z_t$. The z_i letters may be from the same alphabet as the m_i letters or from an enlarged alphabet including, perhaps, special symbols for unknown or semi-unknown letters. In a noisy telegraph situation m and Z might be as follows:

$m =$ I HAVE HEARD THE MERMAIDS SINGING...

$Z =$ I H?V$\overset{?}{\mathrm{T}}$ HEA?D TSE B?RMAID$\overset{?}{\mathrm{Z}}$??NGING...

In this case, the Z alphabet consists of the ordinary letters and space of the m alphabet, together with additional symbols "?", "$\overset{?}{\mathrm{A}}$", "$\overset{?}{\mathrm{B}}$", etc., indicating less certain identification. Even more generally, the Z alphabet might be entirely different from the m alphabet.

Consider a situation in which there is a measure of the fidelity of transmission or the "distortion" between the original and final words. We shall assume first that this distortion measure is of a very simple and special type, and later we shall generalize considerably on the basis of the special case.

A *single-letter distortion measure* is defined as follows. There is given a matrix d_{ij} with $d_{ij} \geq 0$. Here i ranges over the letters of the m alphabet of, say, a letters (assumed given a numerical ordering), while j ranges over the Z alphabet. The quantity d_{ij} may be thought of as a "cost" if letter i is reproduced as letter j.

If the Z alphabet includes the m alphabet, we will assume the distortion between an m letter and its correct reproduction to be zero and all incorrect reproductions to have positive distortion. It is convenient in this case to assume that the alphabets are arranged in the same indexing order so that $d_{ii} = 0, d_{ij} > 0 \; (i \neq j)$.

The distortion d, if *word* m is reproduced as word Z, is to be measured by

$$d(m, Z) = \frac{1}{t} \sum_{k=1}^{t} d_{m_k\, z_k} .$$

If, in a communication system, word m occurs with probability $P(m)$ and the conditional probability, if m is transmitted, that word Z will be reproduced, is $P(Z|m)$, then we assume that the *over-all distortion of the system* is given by

$$d = \sum_{m,Z} P(m)\, P(Z|m)\, d(m, Z) .$$

Here we are supposing that all messages and reproduced words are of the same length t. In variable-length coding systems the analogous measure is merely the over-all probability that letter i reproduced as j, multiplied by d_{ij} and summed on i and j. Note that $d = 0$ if and only if each word is correctly reproduced with probability 1, otherwise $d > 0$ (in cases where the Z alphabet includes the m alphabet).

Some Simple Examples. A distortion measure may be represented by giving the matrix of its elements, all terms of which are non-negative. An alternative representation is in terms of a line diagram similar to those used for representing a memoryless noisy channel. The lines are now labeled, however, with the values d_{ij} rather than probabilities.

A simple example of a distortion measure, with identical m and Z alphabets, is the error probability per letter. In this case, if the alphabets are ordered similarly, $d_{ij} = 1 - \delta_{ij}$. If there were three letters in the m and and Z alphabets, the line diagram would be that shown in Fig. 1(a). Such a distortion measure might be appropriate in measuring the fidelity of a teletype or a remote typesetting system.

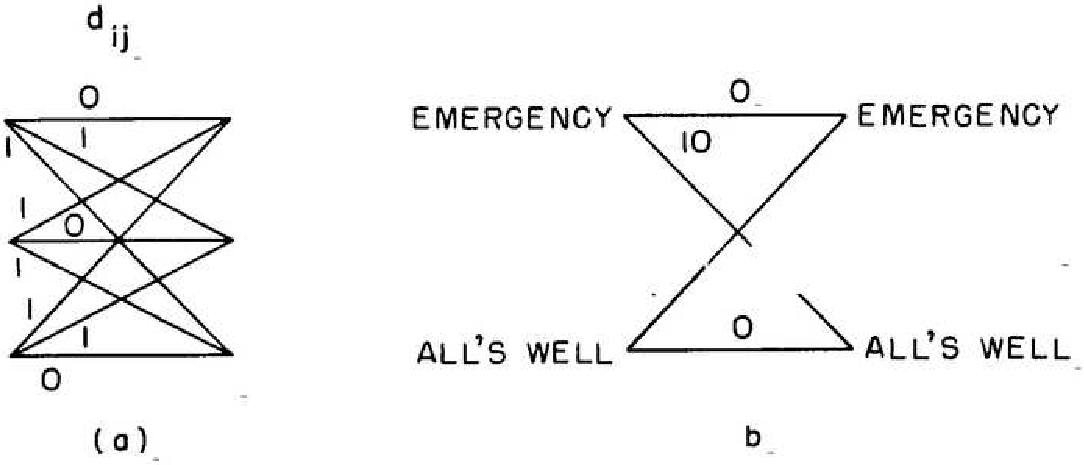

Fig.

Another example is that of transmitting the quantized position of a wheel or shaft. Suppose that the circumference is divided into five equal arcs. It might be only half as costly to have an error of plus or minus one segment as larger errors. Thus the distortion measure might be

$$d_{ij} = \left| \begin{array}{lll} 0 & i = j & \\ \tfrac{1}{2} & |i - j| = 1 & (\text{mod } 5), \\ & |i - j| > 1 & (\text{mod } 5). \end{array} \right.$$

A third example might be a binary system sending information each second, either "all's well" or "emergency," for some situation. Generally, it would be considerably more important that the "emergency" signal be correctly received than that the "all's well" signal be correctly received. Thus if these were weighted 10 to 1, the diagram would be as shown in Fig. 1(b).

A fourth example with entirely distinct m and Z alphabets is a case in which the m alphabet consists of three possible readings, -1, 0 and $+1$. Perhaps, for some reasons of economy, it is desired to work with a reproduced alphabet of two letters, $-\tfrac{1}{2}$ and $+\tfrac{1}{2}$. One might then have the matrix that is shown in Fig. 2.

	$-\frac{1}{2}$	$+\frac{1}{2}$
·I	I	2
0	I	
	2	

Fig. 2

The Rate-Distortion Function R(d). Now suppose that successive letters of the message are statistically independent but chosen with the same probabilities, P_i being the probability of letter i from the alphabet. This type of source we call an *independent letter source.*

Given such a set of probabilities P_i and a distortion measure d_{ij}, we define a *rate-distortio* curve as follows. Assign an arbitrary set of transition probabilities $q_i(j)$ for transitions from to j. (Of course, $q_i(j) \geq 0$ and $\sum_j q_i(j) = 1$.) One could calculate for this assignment tw things: first, the distortion measure $d(q_i(j)) = \sum_{ij} P_i\, q_i(j)\, d_{ij}$ if letter i were reproduced as with conditional probability $q_i(j)$, and, second, the average mutual information between i and if this were the case, namely

$$R(q_i(j)) = E \log \frac{q_i(j)}{\sum_k P_k\, q_k(j)}$$

$$= \sum_{i,j} P_i\, q_i(j) \log \frac{q_i(j)}{\sum_k P_k\, q_k(j)} \; .$$

The rate-distortion function $R(d^*)$ is defined as the greatest lower bound of $R(q_i(j))$ when the $q_i(j)$ are varied subject to their probability limitations and subject to the average distortion d being less than or equal to d^*.

Note that $R(q_i(j))$ is a continuous function of the $q_i(j)$ in the allowed region of variation of $q_i(j)$ which is closed. Consequently, *the greatest lower bound of R is actually attained as a minimum* for each value of R that can occur at all. Further, from its definition it is clear that $R(d)$ is a monotonically decreasing function of d.

Convexity of the *R(d)* Curve. Suppose that two points on the $R(d)$ curve are (R, d) obtained with assignment $q_i(j)$ and (R', d') attained with assignment $q_i'(j)$. Consider a mixture of these assignments $\lambda q_i(j) + (1 - \lambda)\, q_i'(j)$. This produces a d'' (because of the linearity of d) not greater than $\lambda d + (1 - \lambda)\, d'$. On the other hand, $R(q_i(j))$ is known to be a convex downward function (the rate for a channel as a function of its transition probabilities). Hence $R'' \leq \lambda R + (1 - \lambda)\, R'$. The minimizing $q_i''(j)$ for d'' must give at least this low a value o R''. *Hence the curve R as a function of d (or conversely) is convex downward.*

The minimum possible d value clearly occurs if, for each i, $q_i(j)$ is assigned the value 1 fo the j having the minimum d_{ij}. Thus the lowest possible d is given by

$$d_{\min} = \sum_i P_i \min_j d_{ij} \; .$$

If the m alphabet is imaged in the Z alphabet, then $d_{\min} = 0$, and the corresponding R value is the ordinary entropy or rate for the source. In the more general situation, $R(d_{\min})$ may be readily evaluated if there is a unique $\min_j d_{ij}$ by evaluating R for the assignment mentioned Otherwise the evaluation of $R(d_{\min})$ is a bit more complex.

On the other hand, $R = 0$ is obtained if and only if $q_i(j) = Q_j$, a function of j only. Thi is because an average mutual information is positive unless the events are independent. For a given Q_j giving $R = 0$, the d is then $\sum_{ij} P_i\, Q_j\, d_{ij} = \sum_j Q_j \sum_i P_i\, d_{ij}$. The inner sum is non-negative. If we wish the minimum d for $R = 0$, this would result by finding a j that gives a minimum $\sum_i P_i\, d_{ij}$ (say j^*) and making $Q_{j^*} = 1$. This can be done by assigning $q_i(j^*) = 1$ (all other $q_i(j)$ are made 0).

Summarizing, then, $R(d)$ is a convex downward function as shown in Fig. 3 running from $R(d_{\min})$ at $d_{\min} = \sum_i P_i \min_j d_{ij}$ to zero at $d_{\max} = \min_j \sum_i P_i\, d_{ij}$. It is continuous both ways (R as a function of d or d as a function of R) in the interior of this interval because of its convexity. For $d \geq d_{\max}$, we have $R = 0$. The curve is strictly monotonically decreasing from $d_{\min}$ to $d_{\max}$. Also it is easily seen that in this interval the assignment of $q_i(j)$ to obtain any point $R(d^*)$ must give a d satisfying the equality $d = d^*$ (not the inequality $d < d^*$). For $d^* > d_{\max}$ the inequality will occur for the minimizing $q_i(j)$. Thus the minimizing problem can be limited to a consideration of minima in the subspace where $d = d^*$, except in the range $d^* > d_{\max}$ (where $R(d^*) = 0$).

The convex downward nature of R as a function of the assigned $q_i(j)$ is helpful in evaluating the $R(d)$ in specific cases. It implies that any local minimum (in the subspace for a fixed d) is the absolute minimum in this subspace. For otherwise we could connect the local and absolute minima by a straight line and find a continuous series of points lower than the local minimum along this line. This would contradict its being a local minimum.

Furthermore, the functions $R(q_i(j))$ and $d(q_i(j))$ have continuous derivatives interior to the allowed $q_i(j)$ set. Hence ordinary calculus methods (e.g., Lagrangian multipliers) may be used to locate the minimum. In general, however, this still involves the solution of a set of simultaneous equations.

Solution for *R(d)* in Certain Simple Cases. One special type of situation leads to a simple explicit solution for the $R(d)$ curve. Suppose that all a input letters are equiprobable: $P_i = 1/a$. Suppose further that the d_{ij} matrix is square and is such that each row has the same set of entries and each column also has the same set of entries, although, of course, in different order.

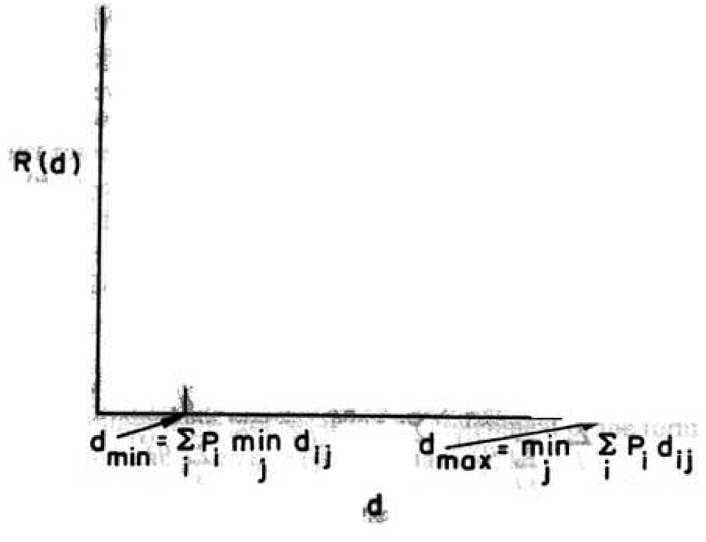

Fig. 3.

An example of this type is the positioning of a wheel mentioned earlier if all positions are equally likely. Another example is the simple error probability distortion measure if all letters are equally likely.

In general, let the entries in any row or column be $d_1, d_2, d_3, \ldots, d_a$. Then we shall show that the minimizing R for a given d occurs when all lines with distortion assignment d_k are given the probability assignment

$$q_k = \frac{e^{-\lambda d_k}}{\sum e^{-\lambda d_i}}$$

Here λ is a parameter ranging from 0 to ∞ which determines the value of d. With this minimizing assignment, d and R are given parametrically in terms of λ:

$$d = \frac{\sum_i d_i\, e^{-\lambda d_i}}{\sum_i e^{-\lambda d_i}}$$

$$R = \log \frac{a}{\sum e^{-\lambda d_i}} - \lambda d$$

When $\lambda = 0$ it can be seen that $d = \frac{1}{a} \sum_i d_i$ and $R = 0$. When $\lambda \to \infty$, $d \to d_{\min}$ and $R \to \log \frac{a}{k}$ where k is the number of d_i with value $d_{\min}$.

This solution is proved as follows. Suppose that we have an assignment $q_i(j)$ giving a certain d^* and a certain R^*. Consider now a new assignment where each line with d_{ij} value d_1 is assigned the average of the assignments for these lines in the original assignment. Similarly, each line labeled d_2 is given the average of all the d_2 original assignments, and so on. Because of the linearity of d, this new assignment has the same d value, namely d^*. The new R is the same as or smaller than R^*. This is shown as follows. R may be written $H(m) - H(m|Z)$. $H(m)$ is not changed, and $H(m|Z)$ can only be increased by this averaging. The latter fact can be seen by observing that because of the convexity of $-\sum x_i \log x_i$ we have

$$\sum_j \alpha_j \sum_t x_j^{(t)} \log x_j^{(t)} \geq - \sum_t \left. \sum_j \alpha_j x_j^{(t)} \right] \log \sum_j \alpha_j x_j^{(t)}$$

where for a given t, $x_j^{(t)}$ is a set of probabilities, and α_j is a set of weighting factors. In particular

$$\sum_j \frac{\sum_s q_j^{(s)}}{\sum_{s,i} q_i^{(s)}} \sum_t \frac{q_j^{(t)}}{\sum_s q_j^{(s)}} \log \frac{q_j^{(t)}}{\sum_s q_j^{(s)}}$$
$$\geq - \sum_t \frac{\sum_j q_j^{(t)}}{\sum_{s,i} q_i^{(s)}} \log \frac{\sum_j q_j^{(t)}}{\sum_{s,i} q_i^{(s)}},$$

where $q_j^{(s)}$ is the original assignment to the line of value d_j from letter s. But this inequality can be interpreted on the left as $H(m|Z)$ after the averaging process, while the right-hand side is $H(m|Z)$ before the averaging. The desired result then follows.

Hence, for the minimizing assignment, all lines with the same d value will have equal probability assignments. We denote these by q_i corresponding to a line labeled d_i. The rate R and distortion d can now be written

$$d = \sum_i q_i d_i ,$$

$$R = \log a + \sum_i q_i \log q_i$$

since all z's are now equiprobable, and $H(m) = \log a, H(m|Z) = -\sum_i q_i \log q_i$. We wish, by proper choice of the q_i, to minimize R for a given d and subject to $\sum_i q_i = 1$. Consider then, using Lagrange multipliers,

$$U = \log a + \sum_i q_i \log q_i + \lambda \sum_i q_i d_i + \mu \sum_i q_i .$$

$$\frac{}{\partial q_i} = 1 + \log q_i + \lambda d_i + \mu = 0 ,$$

$$q_i = A e^{-\lambda d_i}$$

If we choose $A = \frac{}{\sum_i e^{-\lambda d_i}}$ we satisfy $\sum_i q_i =$ This then gives a stationary point and by the convexity properties mentioned above it must be the absolute minimum for the corresponding value of d. By substituting this probability assignment in the formulas for d and R we obtain the results stated above.

Rate for a Product Source with a Sum Distortion Measure. Suppose that we have two independent sources each with its own distortion measure, d_{ij} and $d'_{i'j'}$, and resulting in rate distortion functions $R_1(d_1)$ and $R_2(d_2)$. Suppose that each source produces one letter each second. Considering ordered pairs of letters as single letters the combined system may be called the *product source*. If the total distortion is to be measured by the sum of the individual distortions, $d = d_1 + d_2$, then there is a simple method of determining the function $R(d)$ for the product source. In fact, we shall show that $R(d)$ is obtained by adding both coordinates of the curves $R_1(d_1)$ and $R_2(d_2)$ at points on the two curves having the same slope. The set of points obtained in this manner is the curve $R(d)$. Furthermore, a probability assignment to obtain any point of $R(d)$ is the product of the assignments for the component points.

We shall first show that given any assignment $q_{i,i'}(j, j')$ for the product source, we can do at least as well in the minimizing process using an assignment of the form $q_i(j)\, q'_{i'}(j')$ where q and q' are derived from the given $q_{i,i'}(j, j')$. Namely, let

$$q_i(j) = \sum_{i',j'} P'_{i'} q_{i,i'}(j,j')$$

$$q'_{i'}(j') = \sum_{i,j} P_i\, q_{i,i'}(j, j')$$

We see that these are non-negative and, summed on j and j' respectively, give 1, so they are satisfactory transition probabilities. Also the assignment $q_i(j)\, q'_{i'}(j')$ gives the same total distortion as the assignment $q_{i,i'}(j,j')$. The former is

$$\sum_{i,i'\,j,j'} P_i P'_{i'}\, q_i(j)\, q'_{i'}(j')[d_{ij} + d'_{i'j'}$$
$$= \sum_{i,j} P_i q_i(j) d_{ij} + \sum_{i'j'} P'_{i'} q'_{i'}(j')\, d'_{i'j'}$$
$$= \sum_{\substack{i,j\\ i,i'}} P_i P'_{i'}\, q_{i,i'}(j,j')[d_{ij} + d'_{i'j'}] \,.$$

This last may be recognized as the distortion with $q_{i,i'}(j,j')$.

On the other hand, the mutual information R is decreased or left constant if we use $q_i(j)\, q'_{i'}(j)$ instead of $q_{i,i'}(j,j')$. In fact, this average mutual information can be written in terms of entropies as follows (using asterisks for entropies with the assignment $q_i(j)\, q'_{i'}(j')$ and none for the assignment $q_{i,i'}(j,j')$). We have

$$r = H(i,i') - H(i,i'|j,j')$$
$$\geq H(i,i') - H(i|j) - H(i'|j')$$
$$= H(i,i') - H^*(i|j) - H^*(i'|j') \,.$$

Here we use the fact that with our definition of $q_i(j)$ and $q'_{i'}(j')$ we have $Pr^*(i|j) = Pr(i|j)$ and $Pr^*(i'|j') = pr(i'|j')$. (This follows immediately on writing out these probabilities.) Now, using the fact that the sources are independent, $H(i,i') = H(i) + H(i') = H^*(i) + H^*(I')$. Hence our last reduction above is equal to R^*. This is the desired conclusion.

It follows that any point on the $R(d)$ curve for the product source is obtained by an independent or product assignment $q_i(j)\, q'_{i'}(j')$, and consequently is the sum in both coordinates of a pair of points on the two curves. The best choice for a given distortion d is clearly given by

$$R(d) = \min_t\, [R_1(t) + R_2(d-t)] \,,$$

and this minimum will occur when

$$\frac{d}{dt} R_1(t) = \frac{d}{dt} R_2(d-t)$$

Thus the component points to be added are points where the component curves have the same slope. The convexity of these curves insures the uniqueness of this pair for any particular d.

The Lower Bound on Distortion for a Given Channel Capacity. The importance of the $R(d)$ function is that it determines the channel capacity required to send at a certain rate and with a certain minimum distortion. Consider the following situation. We have given an independent letter source with probabilities P_i for the different possible letters. We have given a single-letter distortion measure d_{ij} which leads to the rate distortion function $R(d)$. Finally, there is a memoryless discrete channel K of capacity C bits per second (we assume that this channel may be used once each second). We wish to transmit words of length t from the source over the channel with a block code. The length of the code words in the channel is n. What is the lowest distortion d that might be obtained with a code and a decoding system of this sort?

Theorem 1. Under the assumptions given above it is not possible to have a code with distortion d smaller than the (minimum) d^* satisfying

$$R(d^* = C$$

or, equivalently, in any code, $d \geq \phi \left| \frac{n}{t} C \right.$ where ϕ is the function inverse to $R(d)$.

This theorem, and a converse positive result to be given later, show that $R(d)$ may be thought of as the equivalent *rate of the source for a given distortion d*. Theorem 1 asserts that for the distortion d and t letters of text, one must supply in the channel at least $t\,R(d)$ total bits of capacity spread over the n uses of the channel in the code. The converse theorem will show that by taking n and t sufficiently large and with suitable codes it is possible to approach this limiting curve.

To prove Theorem 1, suppose that we have given a block code which encodes all message words of length t into channel words of length n and a decoding procedure for interpreting channel output words of length n into Z words of length t. Let a message word be represented by $m = m_1, m_2, \ldots, m_t$. A channel input word is $X = x_1, x_2, \ldots, x_n$. A channel output word is $Y = y_1, y_2, \ldots, y_n$ and a reproduced, or Z, word is $Z = z_1, z_2, \ldots, z_t$. By the given code and decoding system, X is a function of m and Z is a function of Y. The m_i are chosen independently according to the letter probabilities, and the channel transition probabilities give a set of conditional probabilities $P(y|x)$ applying to each x_i, y_i pair. Finally, the source and channel are independent in the sense that $P(Y|m, X) = P(Y|X)$.

We wish first to show that $H(m|Z) \geq H(m) - nC$. We have that $H(m|Z) \geq H(m|Y)$ (since Z is a function of Y) and also that $H(m|Y) \geq H(X|Y) - H(X) + H(m)$. This last is because, from the independence condition above, $H(Y|m, X) = H(Y|X)$, so $H(Y, m; X) - H(m, X) = H(X, Y) - H(X)$. But $H(m, X) = H(m)$, since X is a function of m, and for the same reason $H(m, X, Y) = H(m, Y)$. Hence, rearranging, we have

$$\begin{aligned} H(X, Y &= H(m \quad) + H(X) - H(m, X \\ &= H(m, Y + H(X) - H(m) \\ H(X|Y) &\leq H(m|Y) + H(X) - H(m) \end{aligned}$$

Here we used $H(m, x) = H(m)$ and then subtracted $H(Y)$ from each side. Hence $H(m|Z) \geq H(X|Y) - H(X) + H(m)$.

Now we show that $H(X|Y) \geq nC$. This follows from a method we have used in other similar situations, by considering the *change* in $H(X|Y)$ with each received letter. Thus (using Y_k for the first k of the y letters, etc.),

$$\begin{aligned} \Delta H(X|Y) &= H(X|y_1, y_2, \ldots, y_k) - H(X|y_1, y_2, \ldots, y_{k+1}) \\ &= H(X, X_k) - H(Y_k) - H(X, Y_k, y_k \quad) + H(Y_k \\ &= H(y_{k+1}|Y_k) - H(y_{k+1}|X, Y_k) \\ &= H(y_{k+1}|Y_k) - H(y_{k+} \quad x_{k+} \\ &\leq H(y_{k+1}) - H(y_{k+1}|x_{k+} \\ &\leq C \end{aligned}$$

Here we used the fact that the channel is memoryless, so $P(y_{k+1}|X, Y_k) = P(y_{k+1}|x_{k+1})$ and therefore $H(y_{k+1}|X, Y_k) = H(y_{k+1}|x_{k+1})$. Finally, C is the maximum possible $H(y) - H(y|X)$, giving the last inequality.

Since the incremental change in $H(X|Y_k)$ is bounded by C, the total change after n steps is bounded by nC. Consequently, the final $H(X|Y)$ is at least the initial value $H(X)$ less nC. Therefore

$$\begin{aligned} H(m|Z) &\geq H(X|Y) - H(X) + H(m) \\ &\geq H(X) - nC - H(X) + H(m) , \\ H(m|Z) &\geq H(m) - nC . \end{aligned} \tag{1}$$

We now wish to *overbound* $H(m|Z)$ in terms of the distortion d. We have

$$\begin{aligned} H(m|Z) &= H(m_1\, m_2\, \ldots\, m_t | z_1\, z_2\, \ldots\, z_t) \\ &\leq \sum_i H(m_i|z_i) \\ &= \sum_i H(m_i) - \sum_i (H(m_i) - H(m_i|z_i)) \end{aligned}$$

The quantity $H(m_i) - H(m_i|z_i)$ is the average mutual information between original message letter m_i and the reproduced letter z_i. If we let d_i be the distortion between these letters, then $R(d_i)$ (the rate-distortion function evaluated for this d_i) satisfies

$$R(d_i) \leq H(m_i) - H(m_i|z_i) ,$$

since $R(d_i)$ is the minimum mutual information for the distortion d_i Hence our inequality may be written

$$H(m|Z) \leq \sum_{i=1}^{t} H(m_i) - \sum_{i=1}^{t} R(d_i)$$

Using now the fact that $R(d)$ is a convex downward function, we have

$$H(m|Z) \leq \sum_i H(m_i) - t\, R\left[\sum_i \frac{d_i}{t} \right]$$

But $\sum_i \frac{d_i}{t} = d$, the overall distortion of the system, so

$$H(m|Z) \leq \sum H(m_i) - t\, R(d)$$

Combining this with our previous inequality (1) and using the independent letter assumption, we have $H(m) = \sum_i H(m_i)$, so

$$\begin{aligned} H(m) - nC &\leq H(m) - t\, R(d) , \\ nC &\geq t\, R(d) . \end{aligned}$$

This is essentially the result stated in Theorem 1.

It should be noted that the result in the theorem is an assertion about the minimum distortion after any finite number n of uses of the channel. It is not an asymptotic result for large n. Also, as seen by the method of proof, it applies to any code, block or variable length, provided only that after n uses of the channel, t (or more) letters are reproduced at the receiving point, whatever the received sequence may be.

The Coding Theorem for a Single-Letter Distortion Measure. We now prove a positive coding theorem corresponding to the negative statements of Theorem 1; namely, that it is possible to approach the lower bound of distortion for a given ratio of number n of channel letters to t message letters. We consider then a source of message letters and single-letter distortion measure d_{ij}. More generally than Theorem 1, however, this source may be ergodic; it is not necessarily an independent letter source. This more general situation will be helpful in a later generalization of the theorem. For an ergodic source there will still, of course, be letter probabilities P_i, and we could determine the rate distortion function $R(d)$ based on these probabilities as though it were an independent letter source.

We first establish the following result.

Lemma 1. Suppose that we have an ergodic source with letter probabilities P_i, a single-letter distortion measure d_{ij}, and a set of assigned transition probabilities $q_i(j)$ such that

$$\sum_{i,j} P_i \, q_i(j) \, d_{ij} = d^* ,$$

$$\sum_{i,j} P_i \quad (j) \log \frac{q_i(j)}{\sum_k P_k \, q_k(j)} = R$$

Let $Q(Z)$ be the probability measure of a sequence Z in the space of reproduced sequences if successive source letters had independent transition probabilities $q_i(j)$ into the Z alphabet. Then, given $\varepsilon > 0$, for all sufficiently large block lengths t, there exists a set α of messages of length t from the source with total source probability $P(\alpha) \geq 1 - \varepsilon$, and for each m belonging to α a set of Z blocks of length t, say β_m, such that

$$\begin{array}{lll} 1) & d(m, Z) \leq d^* + \varepsilon & \text{for } m \in \alpha \text{ and } Z \in \beta_m , \\ 2) & Q(\beta_m) \geq e^{-t(R+\varepsilon)} & \text{for any } m \in \alpha . \end{array}$$

In other words, and somewhat roughly, long messages will, with high probability, fall in a certain subset α. Each member m of this subset has an associated set of Z sequences β_m. The members of β_m have only (at most) slightly more than d^* distortion with m and the logarithm of the total probability of β_m in the Q measure is underbounded by $e^{-t(R+\varepsilon)}$.

To prove the lemma, consider source blocks of length t and the Z blocks of length t. Consider the two random variables, the distortion d between an m block and a Z block and the (unaveraged) mutual information type of expression below:

$$d = \frac{1}{t} \sum_i d_{m_i z_i} ,$$

$$I(m; Z) = \frac{1}{t} \log \frac{Pr(Z|m)}{Q(Z)} \qquad \frac{1}{t} \sum_i \log \frac{Pr(\ |m_i}{Q(z_i)}$$

Here m_i is the i^{th} letter of a source block m, and z_i is the i^{th} letter of a Z block. Both R and d are random variables, taking on different values corresponding to different choices of m and Z. They are both the sum of t random variables which are identical functions of the joint (m, Z) process except for shifting along over t positions.

Since the joint process is ergodic, we may apply the ergodic theorem and assert that when t is large, d and R will, with probability nearly 1, be close to their expected values. In particular, for any given ε_1 and δ, if t is sufficiently large, we will have with probability $\geq 1 - \delta^2/2$ that

$$d \le \sum_{i,j} P_i \, q_i(j) \, d_{ij} + \varepsilon_1 = d^* + \varepsilon_1$$

Also, with probability at least $1 - \delta^2/2$ we will have

$$I \le \sum_{i,j} P_i \, q_i(j) \log \frac{q_i(j)}{Q_j} + \varepsilon_1 = R(d^*) + \varepsilon_1$$

Let γ be the set of (m, Z) pairs for which *both* inequalities hold. Then $Pr(\gamma) \ge 1 - \delta^2$ because each of the conditions can exclude, at most, a set of probability $\delta^2/2$. Now for any m_1 define β_{m_1} as the set of Z such that (m_1, Z) belongs to γ.

We have

$$Pr(\beta_m | m) \ge 1 - \delta$$

on a set of α of m whose total probability satisfies $Pr(\alpha) \ge 1 - \delta$. This is true, since if it were not we would have a total probability in the set complementary to γ of at least $\delta \cdot \delta = \delta^2$, a contradiction. The first δ would be the probability of m not being in α, and the second δ the conditional probability for such m's of Z not being in β_m. The product gives a lower bound on the probability of the complementary set to γ.

If $Z \in \beta_{m_1}$, then

$$\frac{1}{t} \log \frac{Pr(Z|m_1)}{Q(Z)} \le R(d^*) + \varepsilon_1 ,$$
$$Pr(Z|m_1) \le Q(Z) \, e^{t(R(d^*) + \varepsilon_1)}$$
$$Q(Z) \ge Pr(Z|m_1) e^{-t(R(d^*) + \varepsilon_1)}$$

Sum this inequality over all $Z \in \beta_{m_1}$:

$$Q(\beta_m) = \sum_{Z \in \beta_{m_1}} Q(Z)$$
$$\ge e^{-t(R + \varepsilon_1)} \sum_{Z \in \beta_{m_1}} Pr(Z|m_1) .$$

If $m_1 \in \alpha$ then $\displaystyle\sum_{Z \in \beta_{m_1}} Pr(Z|m_1) \ge 1 - \delta$ as seen above. Hence the inequality can be continued to give

$$Q(\beta_{m_1}) \ge (1 - \delta) e^{-t(R + \varepsilon_1)} , \quad m_1 \in \alpha .$$

We have now established that for any $\varepsilon_1 > 0$ and $\delta > 0$ there exists a set α of m's and sets β_m of Z's defined for each m with the three properties

1) $Pr(\alpha) \ge 1 - \delta,$

2) $d(Z, m) \le d^* + \varepsilon_1,$ if $Z \in \beta_m,$

3) $Q(\beta_m) \ge (1 - \delta) \, e^{-t(R + \varepsilon_1)},$ if $m \in \alpha,$

provided that the block length t is sufficiently large. Clearly, this implies that for any $\varepsilon > 0$ and sufficiently large t we will have

$$\begin{aligned} &1) \quad Pr(\alpha) \geq 1 - \varepsilon, \\ &2) \quad d(Z, m) \leq d^* + \varepsilon, \qquad \text{if } Z \in \beta_m, \\ &\quad\;\; Q(\beta_m) \geq e^{-t(R + \varepsilon)} \end{aligned}$$

since we may take the ε_1 and δ sufficiently small to satisfy these simplified conditions in which we use the same ε. This concludes the proof of the lemma.

Before attacking the general coding problem, we consider the problem indicated schematically in Fig. 4. We have an ergodic source and a single-letter distortion measure that gives the rate distortion function $R(d)$. It is desired to encode this by a coder into sequences u in such a way that the original messages can be reproduced by the reproducer with an average distortion that does not exceed d^* (d^* being some fixed tolerable distortion level). We are considering here block coding devices for both boxes. Thus the coder takes as input successive blocks of length t produced by the source and has, as output, corresponding to each possible m block, a block from a u alphabet.

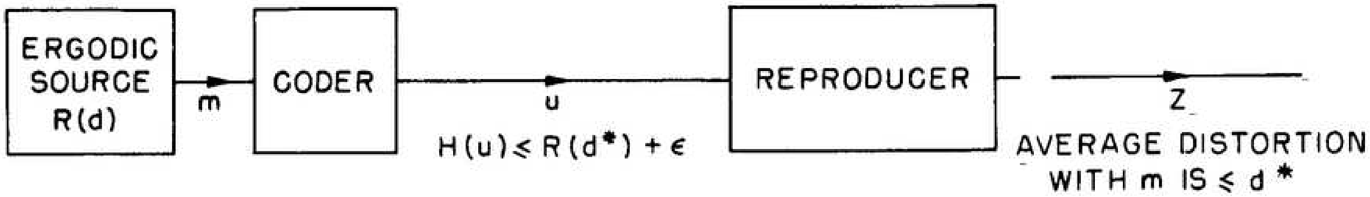

Fig. 4.

The aim is to do the coding in such a way as to keep the entropy of the u sequences as low as possible, subject to this requirement of reproducibility with distortion d^* or less. Here the entropy to which we are referring is the entropy per letter of the original source. Alternatively, we might think of the source as producing one letter per second and we are then interested in the u entropy per second.

We shall show that, for any d^* and any $\varepsilon > 0$, coders and reproducers can be found that are such that $H(u) \leq R(d^*) + \varepsilon$. As $\varepsilon \to 0$ the block length involved in the code in general increases. This result, of course, is closely related to our interpretation of $R(d^*)$ as the equivalent rate of the source for distortion d^*. It will follow readily from the following theorem.

Theorem 2. Given an ergodic source, a distortion measure d_{ij}, and rate distortion function $R(d)$ (based on the single-letter frequencies of the source), given $d^* \geq d_{\min}$ and $\delta > 0$, for any sufficiently large t there exists a set Λ containing M words of length t in the Z alphabet with the following properties:

1) $\dfrac{1}{t} \log M \leq R(d^*) + \delta$,

2) the average distortion between an m word of length t and its nearest (i.e., least distortion) word in the set Λ is less than or equal to $d^* + \delta$.

This theorem implies (except for the δ in property (2) which will later be eliminated) the results mentioned above. Namely, for the coder, one merely uses a device that maps any m word into its nearest member of Λ. The reproducer is then merely an identity transformation.

The entropy per source letter of the coded sequence cannot exceed $R(d^*) + \delta$, since this would be maximized at $\frac{1}{t} \log M$ if all of the M members of Λ were equally probable and $\frac{1}{t} \log M$ is, by the theorem, less than or equal to $R(d^*) + \delta$.

This theorem will be proved by a random coding argument. We shall consider an ensemble of ways of selecting the members of Λ and estimate the average distortion for this ensemble. From the bounds on the average it will follow that at least one code exists in the ensemble with the desired properties.

The ensemble of codes is defined as follows. For the given d^* there will be a set of transition probabilities $q_i(j)$ that result in the minimum R, that is, $R(d^*)$. The set of letter probabilities, together with these transition probabilities, induce a measure $Q(Z)$ in the space of reproduced words. The Q measure for a single Z letter, say letter j, is $\sum_i P_i q_i(j)$. The Q measure for a Z word consisting of letters $j_1, j_2, \ldots, j_t$ is $Q(Z) = \prod_{k=1}^{t} \left[\sum_i P_i q_i(j_k) \right]$.

In the ensemble of codes of length t, the integers from 1 to M are mapped into Z words of length t in all possible ways. An integer is mapped into a particular word Z_1, say, with probability $Q(Z_1)$, and the probabilities for different integers are statistically independent. This is exactly the same process as that of constructing a random code ensemble for a memoryless channel, except that here the integers are mapped into the Z space by using the $Q(Z)$ measure. Thus we arrive at a set of codes (if there are f letters in the Z alphabet there will be $f^{\,tM}$ different codes in the ensemble) and each code will have an associated probability. The code in which integer i is mapped into Z_i has probability $\prod_{i=1}^{M} Q(Z_i)$.

We now use Lemma 1 to bound the average distortion for this ensemble of codes (using the probabilities associated with the codes in calculating the average). Note, first, that in the ensemble of codes if $Q(\beta)$ is the Q measure of a set β of Z words, then the probability that this set contains no code words is $[1 - Q(\beta)]^M$, that is, the product of the probability that code word 1 is not in β, that for code word 2, etc. Hence the probability that β contains at least one code word is $1 - [1 - Q(\beta)]^M$. Now, referring to Lemma 1, the average distortion may be bounded by

$$\bar{d} \le \varepsilon\, d_{\max} + [1 - Q(\beta_m)]^M\, d_{\max} + (d^* + \varepsilon) \; .$$

Here $d_{\max}$ is the largest possible distortion between an M letter and a Z letter. The first term, $\varepsilon d_{\max}$, arises from message words m which are not in the set α. These have total probability less than or equal to ε and, when they occur, average distortion less than or equal to $d_{\max}$. The second term overbounds the contribution that is due to cases in which the set β_m for the message m does not contain at least one code word. The probability in the ensemble of this is certainly bounded by $[1 - Q(\beta_m)]^M$, and the distortion is necessarily bounded by $d_{\max}$. Finally, if the message is in α and there is at least one code word in β_m, the distortion is bounded by $d^* + \varepsilon$, according to Lemma 1. Now, $Q(\beta_m) \ge e^{-t(R(d^*) + \varepsilon)}$. Also, for $0 < x \le 1$,

$$(1-x)^{\frac{1}{x}} = e^{\frac{1}{x}\log(1-x)} \le e^{\frac{1}{x}\left[-x + \frac{x^2}{2}\right]}$$

$$= e^{-1 + \frac{x}{2}} \le e^{-1/2}$$

(using the alternating and monotonically decreasing nature of the terms of the logarithmic expansion). Hence

$$[1 - Q(\beta_m)]^M \leq \quad e^{-t(R(d^*) + \varepsilon)})^M$$

and replacing the exponent by

$$M\, e^{t(R(d^*) + \varepsilon)}\, e^{-t(R(d^*) + \varepsilon)}$$

we see that this is

$$\leq \exp \{ -\tfrac{1}{2}\, e^{-t(R(d^*) + \varepsilon)}\, M.$$

If we choose for M, the number of points, the value $e^{t(R(d^*) + 2\varepsilon)}$ (or, if this is not an integer, the smallest integer exceeding this quantity), then the expression given above is bounded by $\exp \{ -\tfrac{1}{2}\, e^{t\varepsilon} \}$. Thus the average distortion is bounded with this choice of M by

$$\begin{aligned} \bar{d} &\leq \varepsilon d_{\max} + \exp \{ -\tfrac{1}{2}\, e^{t\varepsilon} \}\, d_{\max} + d^* + \varepsilon \\ &\leq d^* + \delta\,, \end{aligned}$$

provided that ε in Lemma 1 is chosen small enough to make $(\varepsilon d_{\max} + 1) \leq \delta/2$ and then t is chosen large enough to make $\exp \{ -\tfrac{1}{2} e^{t\varepsilon} \}\, d_{\max} \leq \delta/2$. We also require that ε be small enough and t large enough to make M , the integer just greater than or equal to $e^{t(R(d^*) + 2\varepsilon)}$, less than or equal to $e^{t(R(d^*) + \delta)}$. Since Lemma 1 holds for all sufficiently large t and any positive ε, these can all be simultaneously satisfied.

We have shown, then, that the conditions of the theorem are satisfied by the average distortion of the ensemble of codes. It follows that there exists at least one specific code in the ensemble whose average distortion is bounded by $d^* + \varepsilon$. This concludes the proof.

Corollary: Theorem 2 remains true if δ is replaced by 0 in property (1). It also remains true if the δ in property (1) is retained and the δ in property (2) is replaced by 0, provided in this case that $d^* > d_{\min}$, the smallest d for which $R(d)$ is defined.

This corollary asserts that we can attain (or do better than) one coordinate of the $R(d)$ curve and approximate, as closely as desired, the other, except possibly for the $d_{\min}$ point. To prove the first statement of the corollary, note first that it is true for $d^* \geq d_1$, the value for which $R(d_1) = 0$. Indeed, we may achieve the point $\bar{d} = d_1$ with $M = 1$ and a code of length 1, using only the Z word consisting of the single Z letter which gives this point of the curve. For $d_{\min} \leq d^* < d_1$, apply Theorem 2 to approximate $d^{**} = d^* + \delta/2$. Since the curve is strictly decreasing, this approximation will lead to codes with $\bar{d} \leq d^* + \delta$ and $\frac{1}{t} \log M \leq R(d^*)$, if the δ in Theorem 2 is made sufficiently small.

The second simplification in the corollary is carried out in a similar fashion, by choosing a d^{**} slightly smaller than the desired d^* that is such that $R(d^{**}) = R(d^*) + \delta/2$, and by using Theorem 2 to approximate this point of the curve.

Now suppose we have a memoryless channel of capacity C. By the coding theorem for such channels it is possible to construct codes and decoding systems with rate approximating C (per use of the channel) and error probability $\leq \varepsilon_1$ for any $\varepsilon_1 > 0$. We may combine such a code for a channel with a code of the type mentioned above for a source at a given distortion level d^* and obtain the following result.

Theorem 3. Given a source characterized by $R(d)$ and a memoryless channel with capacity $C > 0$, given $\varepsilon > 0$ and $d^* > d_{\min}$, there exists, for sufficiently large t and n, a block code that maps source words of length t into channel words of length n and a decoding system that maps channel output words of length n into reproduced words of length t which satisfy

$$1)\ \bar{d} \le d^* \ ,$$

$$2)\ \frac{nC}{t} \le R(d^*) + \varepsilon$$

Thus we may attain a desired distortion level d^* (greater than $d_{\min}$) and at the same time approximate using the channel at a rate corresponding to $R(d^*)$. This is done, as in the corollary stated above, by approximating the $R(d)$ curve slightly to the left of d^*, say, at $R(d^*) - \delta$. Such a code will have $M = e^{t(R(d^* - \delta) + \delta_1)}$ words, where δ_1 can be made small by taking t large. A code for the channel is constructed with M words and of length n, the largest integer satisfying $\frac{nC}{t} \le R(d^* - \delta) + \delta_1$. By choosing t sufficiently large, this will approach zero error probability, since it corresponds to a rate less than channel capacity. If these two codes are combined, it produces an over-all code with average distortion at most d^*.

Numerical Results for Some Simple Channels. In this section some numerical results will be given for certain simple channels and sources. Consider, first, the binary independent letter source with equiprobable letters and suppose that the distortion measure is the error probability (per digit). This falls into the class for which a simple explicit solution can be given. The $R(d)$ curve, in fact, is

$$R(d) = 1 + d \log_2 d + (1 - d) \log_2 (1 - d)$$

This, of course, is the capacity of a symmetric binary channel with probabilities d and $(1 - d)$, the reason being that this is the probability assignment $q_i(j)$ which solves the minimizing problem.

This $R(d)$ curve is shown in Fig. 5. Also plotted are a number of points corresponding to specific simple codes, with the assumption of a noiseless binary channel. These will give some idea of how well the lower bound may be approximated by simple means. One point, $d = 0$, is obtained at rate $R = 1$ simply by sending the binary digits through the channel. Other simple codes which encode 2, 3, 4 and 5 message letters into one channel letter are the following. For the ratio 3 or 5, encode message sequences of three or five digits into 0 or 1 accordingly as the sequence contains more than half zeros or more than half ones. For the ratios 2 and 4, the same procedure is followed, while sequences with half zeros and half ones are encoded into 0.

At the receiving point, a 0 is decoded into a sequence of zeros of the appropriate length and a 1 into a sequence of ones. These rather degenerate codes are plotted in Fig. 5 with crosses. Simple though they are, with block length of the channel sequences only one, they still approximate to some extent the lower bound.

Plotted on the same curve are square points corresponding to the well-known single-error correcting codes with block lengths 3, 7, 15 and 31. These codes are used backwards here – any message in the 15-dimensional cube, for example, is transmitted over the channel as the *eleven message* digits of its nearest code point. At the receiving point, the corresponding fifteen-digit message is reconstructed. This can differ at most in one place from the original message. Thus for this case the ratio of channel to message letters is $\frac{11}{15}$, and the error probability is easily found to be $\frac{1}{16}$. This series of points gives a closer approximation to the lower bound.

It is possible to fill in densely between points of these discrete series by a technique of *mixing codes*. For example, one may alternate in using two codes. More generally, one may mix them in proportions λ and 1 − λ, where λ is any rational number. Such a mixture gives a code with a new ratio R of message to channel letters, given by $\frac{1}{R} = \frac{\lambda}{R_1} + \frac{(1-\lambda)}{R_2}$, where R_1 and R_2 are the ratios for the given codes, and with new error probability

$$P_e \quad \frac{\lambda R_1 P_{e1} + (1-\lambda) R_2 P_{e2}}{\lambda R \qquad \lambda) R}$$

This interpolation gives a convex upward curve between any two code points. When applied to the series of simple codes and single-error correcting codes in Fig. 5, it produces the dotted-line interpolations indicated.

Another channel was also considered in this connection, namely, the binary symmetric channel of capacity $C = ½$. This has probabilities 0.89 that a digit is received correctly and 0.11 incorrectly. Here the series of points (Fig. 6) for simple codes actually touches the lower bound at the point $R = ½$. This is because the channel itself, without coding, produces just this error probability. Any symmetric binary channel will have one point that can be attained exactly by means of straight transmission.

Figure 7 shows the $R(d)$ curve for another simple situation, a binary independent letter source but with the reproduced Z alphabet consisting of three letters, 0, 1, and ?. The distortion measure is zero for a correct digit, one for an incorrect digit, and 0.25 for ?. In the same figure is shown, for comparison, the $R(d)$ curve without the ? option.

Figure 8 shows the $R(d)$ curves for independent letter sources with various numbers of equiprobable letters in the alphabet (2, 3, 4, 5, 10, 100). Here again the distortion measure is taken to be error probability (per digit). With b letters in the alphabet the $R(d, b)$ curve is given by

$$R(d, b) = \log_2 b + d \log_2 d + (\quad - d) \log_2 \frac{1-d}{b-}$$

Generalization to Continuous Cases. We will now sketch briefly a generalization of the single-letter distortion measure to cases where the input and output alphabets are not restricted to finite sets but vary over arbitrary spaces.

Assume a message alphabet $A = \{m\}$ and a reproduced letter alphabet $B = \{z\}$. For each pair (m, z) in these alphabets let $d(m, z)$ be a non-negative number, the distortion if m is reproduced as z. Further, we assume a probability measure P defined over a Borel field of subsets of the A space. Finally, we require that, for each z belonging to B, $d(m, z)$ is a measurable function with finite expectation.

Consider a finite selection of points z_i $(i = 1, 2, \ldots, l)$ from the B space, and a measurable assignment of transition probabilities $q(z_i|m)$. (That is, for each i, $q(z_i|m)$ is a measurable function in the A space.) For such a choice of z_i and assignment $q(z_i|m)$, a mutual information and an average distortion are determined:

$$R = \sum_i \int q(z_i|m) \log \frac{q(z_i|m)}{\int q(z_i|m)\, dP(m)}\, dP(m)$$

$$d = \sum \int d(m, z_i)\, q(z_i|m)\, dP(m)$$

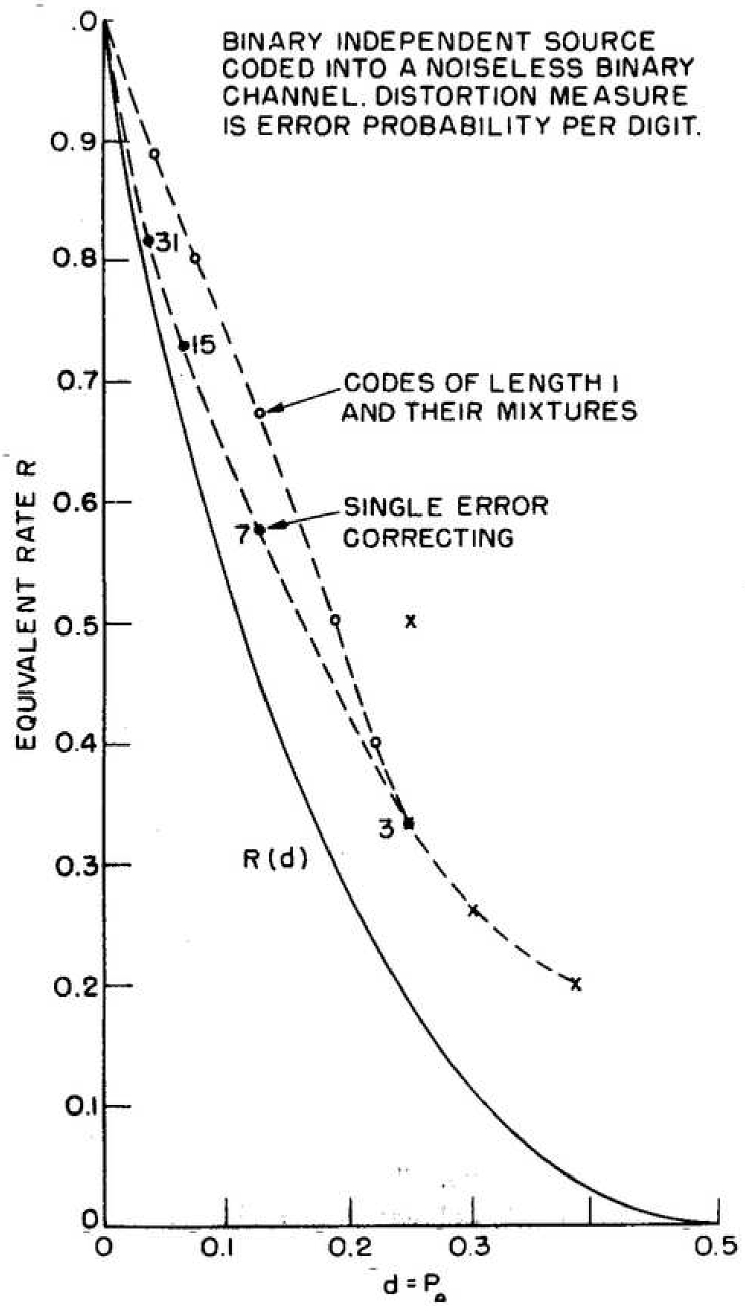
BINARY INDEPENDENT SOURCE CODED INTO A NOISELESS BINARY CHANNEL. DISTORTION MEASURE IS ERROR PROBABILITY PER DIGIT.
CODES OF LENGTH 1 AND THEIR MIXTURES
SINGLE ERROR CORRECTING
31
15
7
3
R(d)
EQUIVALENT RATE R
.0
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.5
d = P_e

Fig. 5.

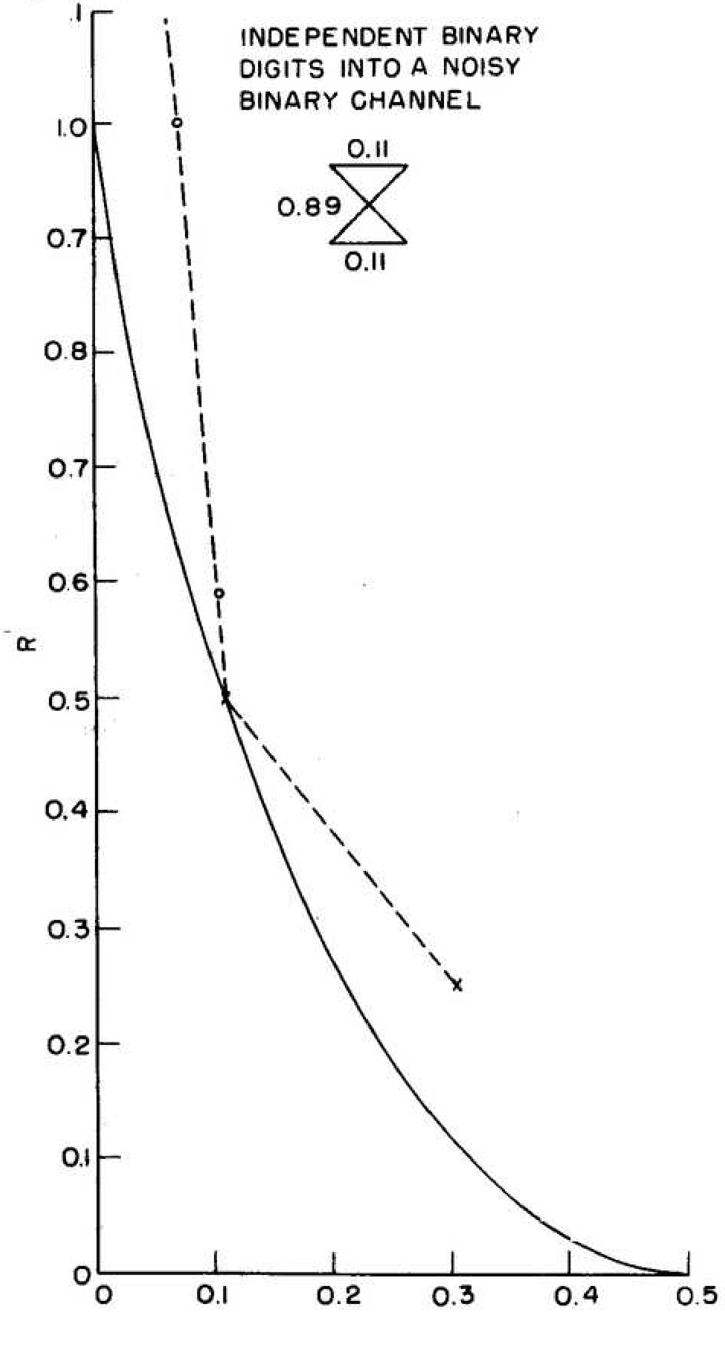
INDEPENDENT BINARY DIGITS INTO A NOISY BINARY CHANNEL
0.11
0.89
0.11
R
.1
1.0
0.7
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
0
0.1
0.2
0.3
0.4
0.5

Fig. 6.

We define the rate distortion function $R(d^*)$ for such a case as the greatest lower bound of R when the set of points z_i is varied (both in choice and number) and the $q(z_i|m)$ is varied over measurable transition probabilities, subject to keeping the distortion at the level d^* or less.

Most of the results we have found for the finite alphabet case carry through easily under this generalization. In particular, the convexity property of the $R(d)$ curve still holds. In fact, if $R(d)$ can be approximated to within ε by a choice z_i and $q(z_i|m,)$ and $R(d')$ by a choice of z_i' and $q'(z_i'|m)$, then one considers the choice z_i'' consisting of the union of the points z_i and z_i', together with $q''(z_i''|m) = \frac{1}{2}[q(z_i''|m) + q'(z_i''|m)]$ (using zero if $q(z''|m)$ or $q'(z''|m)$ is undefined). This leads, by the convexity of R and by the linearity of d, to an assignment for $d'' = \frac{1}{2}d + \frac{1}{2}d'$, giving an R'' within ε of the midpoint of the line joining $d\,R(d)$ and $d'R(d')$. It follows, since ε can be made arbitrarily small, that the greatest lower bound of $R(d'')$ is on or below this midpoint.

In the general case it is, however, not necessarily true that the $R(d)$ curve approaches a finite end-point when d decreases toward its minimum possible value. The behavior may be as indicated in Fig. 9 with $R(d)$ going to infinity as d goes to $d_{\min}$. On the other hand, under the conditions we have stated, there is a finite $d_{\max}$ for which $R(d_{\max}) = 0$. This value of d is given by

$$d_{\max} = \underset{z}{g.l.b.}\, E[d(m,z)]$$

The negative part of the coding theorem goes through in a manner essentially the same as the finite alphabet case, it being assumed that the only allowed coding functions from the source sequences to channel inputs correspond to measurable subsets of the source space. (If this assumption were not made, the average distortion would not, in general, even be defined.) The various inequalities may be followed through, changing the appropriate sums in the A space to integrals and resulting in the corresponding negative theorem.

For the positive coding theorem also, substantially the same argument may be used with an additional ε involved to account for the approximation to the greatest lower bound of $R(d)$ with a finite selection of z_i points. Thus one chooses a set of z_i to approximate the $R(d)$ curve to within ε, and then proceeds with the random coding method. The only point to be noted is that the $d_{\max}$ term must now be handled in a slightly different fashion. To each code in the ensemble one may add a particular point, say z_0, and replace $d_{\max}$ by $E(d(m, z_0))$, a finite quantity. The results of the theorem then follow.

Difference Distortion Measure. A special class of distortion measures for certain continuous cases of some importance and for which more explicit results can be obtained will now be considered. For these the m and z spaces are both the sets of all real numbers. The distortion measure $d(m, z)$ will be called a *difference distortion measure* if it is a function only of the difference $m - z$, thus $d(m, z) = d(m - z)$. A common example is the squared error measure, $d(m, z) = (m - z)^2$ or, again, the absolute error criterion $d(m, z) = |m - z|$

We will develop a lower bound on $R(d)$ for a difference distortion measure. First we define a function $\phi(d)$ for a given difference measure $d(u)$ as follows. Consider an arbitrary distribution function $G(u)$ and let H be its entropy and d the average distortion between a random variable with a given distribution and zero. Thus

$$H = -\int_{-\infty}^{\infty} \log dG(u)\, dG(u) \ ,$$

$$d = \int_{-\infty}^{\infty} d(u)\, dG(u) \ .$$

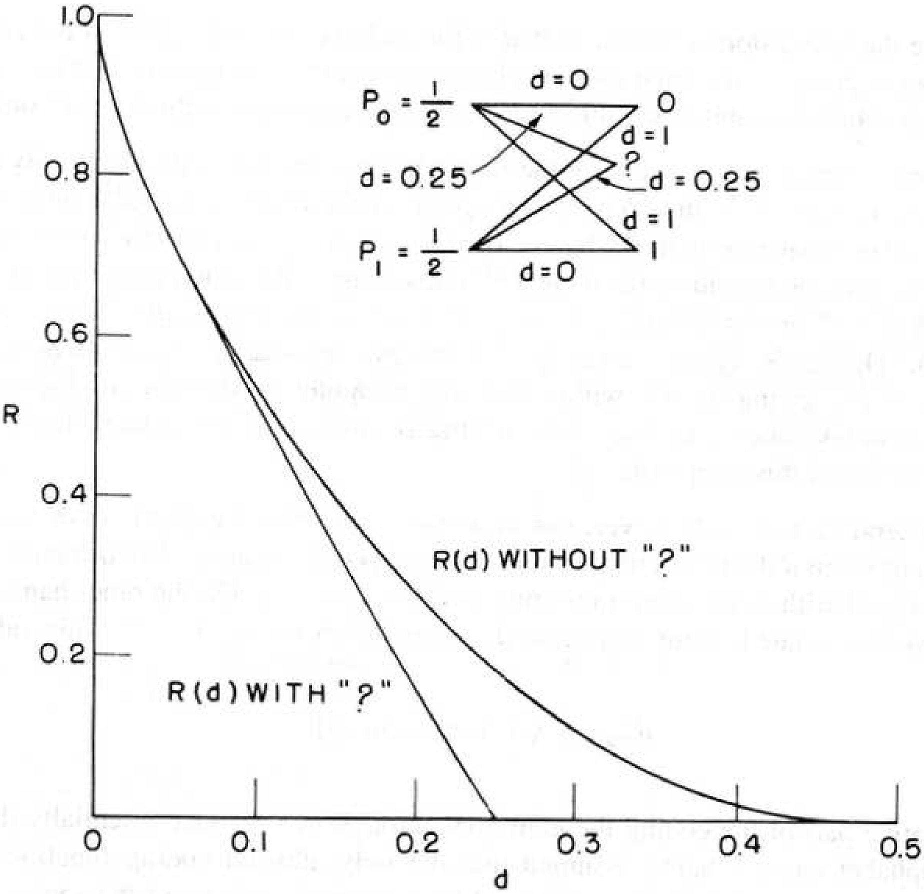

Fig. 7.

We wish to vary the distribution $G(u)$, keeping $d \leq d^*$, and seek the maximum H. The least upper bound, if finite, is clearly actually attained as a maximum for some distribution. This maximum H for a given d^* we call $\phi(d^*)$, and a corresponding distribution function is called a maximizing distribution for this d^*.

Now suppose we have a distribution function for the m space (generalized letter probabilities) $P(m)$, with entropy $H(m)$. We wish to show that

$$R(d) \geq H(m) - \phi(d) .$$

Let z_i be a set of z points and $q(z_i|m)$ an assignment of transition probabilities. Then the mutual information between m and z may be written

$$R = H(m) - \sum_i Q_i \, H(m|z_i) ,$$

where Q_i is the resulting probability of z_i. If we let d_i be the average distortion between m and z_i, then

$$H(m|z_i) \leq \phi(d_i) .$$

This is because $\phi(d)$ was the maximum H for a given average distortion and also because the distortion is a function only of the difference between m and z, so that this maximizing value applies for any z_i. Thus

$$R \geq H(m) - \sum_i Q_i \, \phi(d_i) .$$

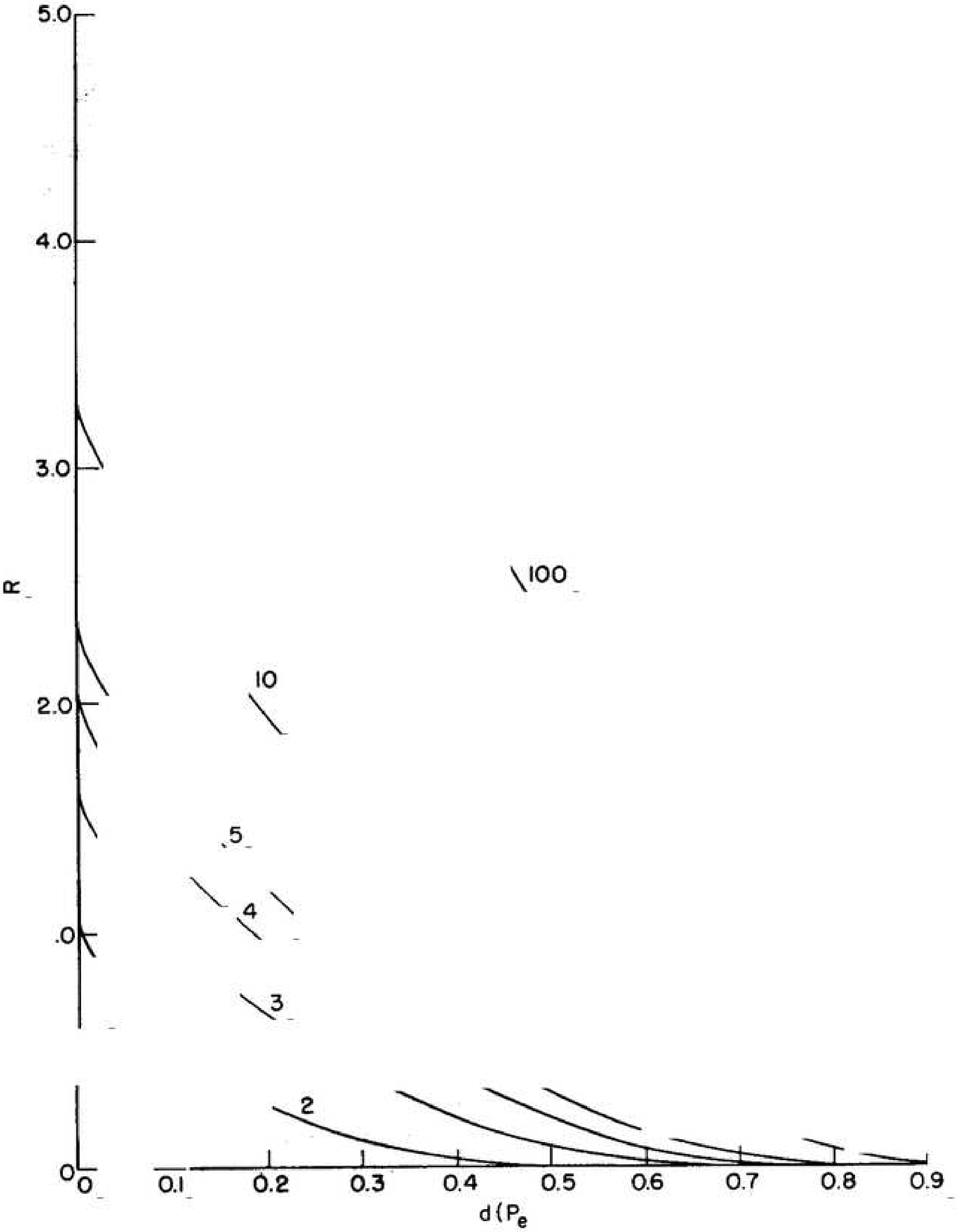

Fig. 8.

Now $\phi(d)$ is a concave function. This is a consequence of the concavity of entropy considered as a function of a distribution function and the linearity of d in the same space of distribution functions, by an argument identical with that used previously. Hence $\sum_i Q_i \, \phi(d_i) \leq \phi(\sum_i Q_i \, d_i) = \phi(d)$, where d is the average distortion with the choice z_i and the assigned transition probabilities. It follows that

$$r \geq H(m) \quad \phi(d)$$

This is true for any assignment z_i and $q(z_i | m)$, and proves the desired result.

If, for a particular $P(m)$ and $d(u)$, assignments can be made which approach this lower bound, then, of course, this is the $R(d)$ function. Such is the case, for example, if $P(m)$ is Gaussian and $d(u) = u^2$ (mean square error measure of distortion). Suppose that the message has variance σ^2, and consider a Gaussian distribution of mean zero and variance $\sigma^2 - d$ in the z space. (If this is zero or negative, clearly $R(d) = 0$ by using only the z point zero.) Let the

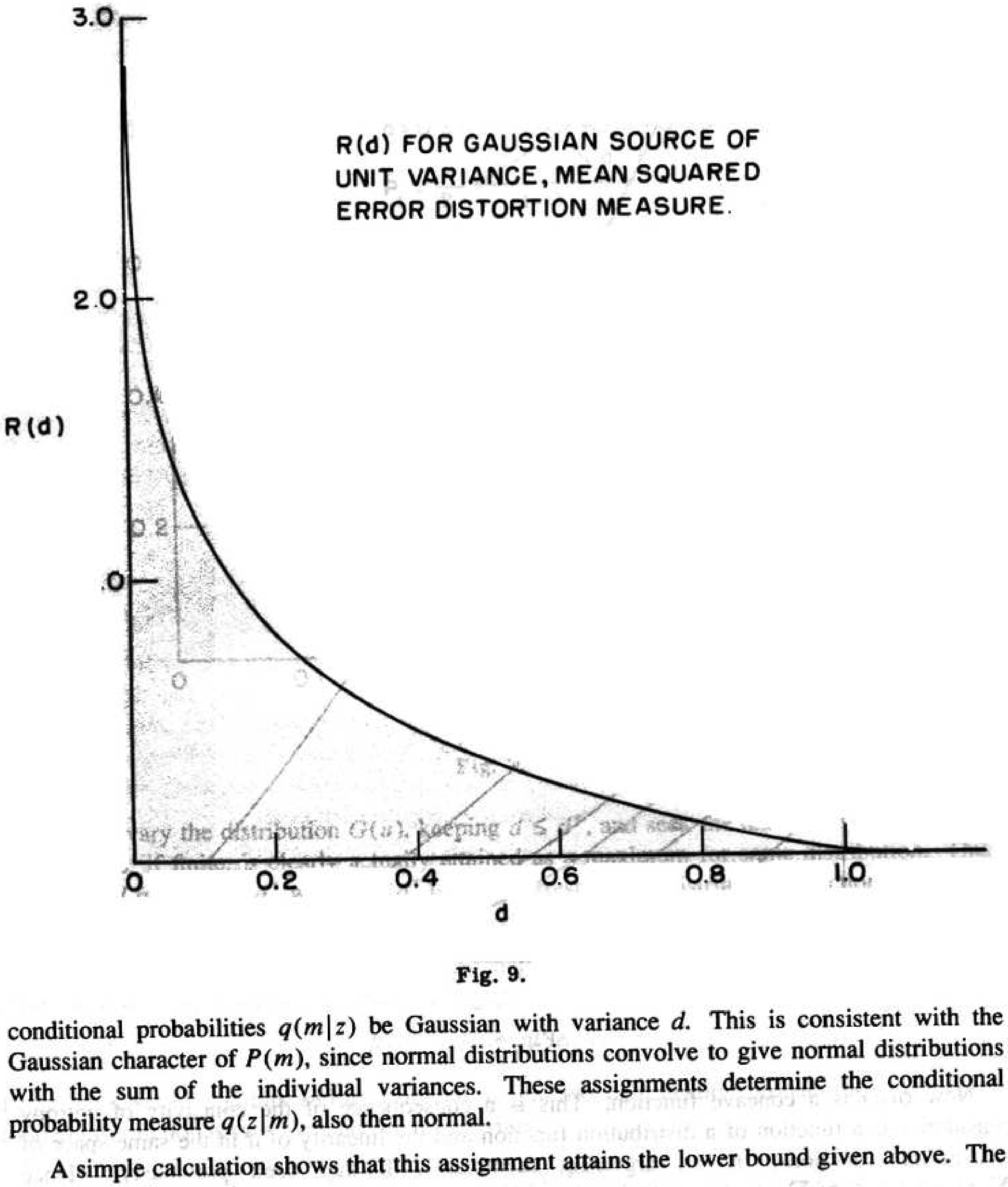

Fig. 9.

conditional probabilities $q(m|z)$ be Gaussian with variance d. This is consistent with the Gaussian character of $P(m)$, since normal distributions convolve to give normal distributions with the sum of the individual variances. These assignments determine the conditional probability measure $q(z|m)$, also then normal.

A simple calculation shows that this assignment attains the lower bound given above. The resulting $R(d)$ curve is

$$R(d) = \begin{cases} \mathrm{lo} \quad \sigma & d \leq \sigma^2, \\ 0 & d > \sigma^2 \end{cases}$$

This is shown for $\sigma^2 = 1$ in Fig. 9.

Definition of a Local Distortion Measure. Thus far we have considered only a distortion measure d_{ij} (or $d(m, z)$) which depends upon comparison of a message letter with the corresponding reproduced letter, this letter-to-letter distortion to be averaged over the length of message and over the set of possible messages and possible reproduced messages. In many practical cases, however, this type of measure is not sufficiently general. The seriousness of a particular type of error often depends on the context.

Thus in transmitting a stock market quotation, say: "A.T.&T. 5900 shares, closing 194," an error in the 9 of 5900 shares would normally be much less serious than an error in the 9 of the closing price.

We shall now consider a distortion measure that depends upon local context and, in fact, compares blocks of g message letters with the corresponding blocks of g letters of the reproduced message.

A *local distortion measure of span* g is a function $d(m_1, m_2, \ldots, m_g; z_1, z_2, \ldots, z_g)$ of message sequences of length g and reproduced message sequences of length g (from a possibly different or larger alphabet) with the property that $d \geq 0$. The distortion between $m = m_1, m_2, \ldots, m_t$ and $z = z_1, z_2, \ldots, z_t$ $(t \geq g)$ is defined by

$$d(m, Z) = \frac{1}{t-g} \sum_{k=1}^{t-g+1} d(m_k, m_{k+1}, \ldots, m_{k+g-1} \; ; z_k, z_{k+1}, \ldots, z_{k+g-1}) \; .$$

The distortion of a *block code* in which message m and reproduced version Z occur with probability $P(m, Z)$ is defined by

$$d = \sum_{m,Z} P(m, Z) \; d(m, Z) \; .$$

In other words, we assume, with a local distortion measure, that the evaluation of an entire system is obtained by averaging the distortions for all block comparisons of length g each with its probability of occurrence a weighting factor.

The Functions $R_n(d)$ and $R(d)$ for a Local Distortion Measure and Ergodic Source. Assume that we have given an ergodic message source and a local distortion measure. Consider blocks of m message letters with their associated probabilities (as determined by the source) together with possible blocks Z of reproduced message of length n. Let an arbitrary assignment of transition probabilities from the m blocks to the Z blocks, $q(Z|m)$, be made. For this assignment we can calculate two quantities: 1) the average mutual information per letter $R = \frac{1}{n} E\left[\log \frac{q(Z|m)}{Q(z)}\right]$ and 2) the average distortion if the m's were reproduced as Z's with the probabilities $q(Z|m)$. This is $d = \sum_{m,Z} P(m, Z) \; d(m, Z)$. By variation of $q(Z|m)$, while holding $d \leq d^*$, we can, in principle, find the minimum R for each d^*. This we call $R_n(d^*)$.

The minimizing problem here is identical with that discussed previously if we think of m and Z as individual letters in a (large) alphabet, and various results relating to this minimum can be applied. In particular, $R_n(d)$ is a convex downward function.

We now define the *rate distortion function* for the given source relative to the distortion measure as

$$R(d) = \liminf_{n \to \infty} R_n(d) \; .$$

It can be shown, by a direct but tedious argument that we shall omit, that the "inf" may be deleted from this definition. In other words, $R_n(d)$ approaches a limit as $n \to \infty$.

We are now in a position to prove coding theorems for a general ergodic source with a local distortion measure.

The Positive Coding Theorem for a Local Distortion Measure.

Theorem 4. Suppose that we are given an ergodic source and a local distortion measure with rate distortion function $R(d)$. Let K be a memoryless discrete channel with capacity C, let d^* be a value of distortion, and let ε be a positive number. Then there exists a block code with distortion less than or equal to d^* ε, and a signaling rate at least $\left[\frac{C}{R} - \varepsilon\right]$ message letters per channel letter.

Proof. Choose an n_1 so that $R_{n_1}(d^*) - R(d^*) < \frac{\varepsilon}{3}$ and, also, so large that $\frac{g}{n_1} d_{\max} < \frac{\varepsilon}{3}$. Now consider blocks of length n_1 and "letters" of an enlarged alphabet. Using Theorem 3 we can construct a block code using sufficiently long sequences of these "letters" signaling at a rate close to (say within ε/3 of) $R_{n_1}(d^*)/C$ (in terms of original message letters) and with distortion less than $d^* + \frac{\varepsilon}{3}$. It must be remembered that this distortion is based on a single "letter" comparison. However, the distortion by the given local distortion measure will differ from this only because of overlap comparisons (g for each n_1 letters of message) and hence the discrepancy is, at most, $\frac{g}{n_1} d_{\max} < \frac{\varepsilon}{3}$. It follows that this code signals at a rate within ε of $R(d^*)$ and at a distortion within ε of d^*.

The Converse Coding Theorem.

Theorem 5. Suppose that we are given an ergodic source and a local distortion measure with rate distortion function $R(d)$. Let K be a memoryless discrete channel with capacity C, let d^* be a value of distortion, and let ε be a positive number. Then there exists t_0 which is such that any code transmitting $t \geq t_0$ message letters with n uses of the channel at distortion d^*, or less, satisfies

$$\frac{n}{t} C \quad R(d^* \quad - \varepsilon$$

That is, the channel capacity bits used per message letter must be nearly $R(d^*)$ for long transmissions.

Proof. Choose t_0 so that for $t \geq t_0$ we have $R_t(d) \geq R(d) - \varepsilon$. Since $R(d)$ was defined as $\liminf\limits_{t \to \infty} R_t(d)$, this is possible. Suppose that we have a code for such a $t \geq t_0$ which maps sequences m consisting of t message letters into sequences X of n channel letters and decodes sequences Y of n channel output letters into sequences Z of reproduced messages. The channel will have, from its transition probabilities, some $P(Y|X)$. Furthermore, from the encoding and decoding functions, we shall have $X = f(m)$ and $Z = g(Y)$. Finally there will be, from the source, probabilities for the message sequences $P(m)$. By the encoding function $f(m)$ this will induce a set of probabilities $P(X)$ for input sequences. If the channel capacity is C, the average mutual information $R(X, Y)$ between input and output sequences must satisfy

$$R(X, Y) = E \log \frac{P(X|Y)}{P(X)} \leq nC \ ,$$

since nC is the maximum possible value of this quantity when $P(X)$ is varied. Also, since X is a function of m and Z is a function of Y, we have

$$R(m, Z) = E \log \frac{P(m|Z)}{P(m)} \leq R(X, Y) \leq nC \, .$$

The coding system in question amounts, overall, to a set of conditional probabilities from m sequences to Z sequences as determined by the two coding functions and the transition probabilities. If the distortion of the overall system is less than or equal to d^*, then $t\,R_t(d^*) = \min\limits_{P(Z|m)} R(m, Z)$ is certainly less than or equal to the particular $R(m, Z)$ obtained with the probabilities given by the channel and coding system. (The t factor is present because $R_t(d)$ is measured on a per message letter basis, while the $R(m, Z)$ quantities are for sequences of length t.) Thus

$$tR_t(d^*) \leq R(m, Z) \leq nC$$

$$t(R(d^*) - \varepsilon) \leq nC \, ,$$

$$\frac{n}{t} C \geq R(d^*) - \varepsilon \, .$$

This is the conclusion of the theorem.

Notice from the method of proof that again the code used need not be a block code, provided only that, after n uses of the channel, t recovered letters are written down. If one has some kind of variable-length code and, starting at time zero, uses this code continually, the inequality of the theorem will hold for any finite time after t_0 message letters have been recovered; and of course as longer and longer blocks are compared, $\varepsilon \to 0$. It is even possible to generalize this to variable-length codes in which, after n uses of the channel, the number of recovered message letters is a random variable depending, perhaps, on the particular message and the particular chance operation of the channel. If, as is usually the case in such codes, there exists an average signaling rate with the properties that after n uses of the channel then, with probability nearly one, t letters will be written down, with t lying between $t_1(1 - \delta)$ and $t_1(1 + \delta)$ (the $\delta \to 0$ as $n \to \infty$), then essentially the same theorem applies, using the mean t_1 for t.

Channels with Memory. Finally we mention that while we have, in the above discussion, assumed the channel to be memoryless, very similar results, both of positive and negative type, can be obtained for channels with memory.

For a channel with memory one may define a capacity C_n for the first n use of the channel starting at state s_0. This C_n is $\frac{1}{n}$ times the maximum average mutual information between input sequences of length n and resulting output sequences when the probabilities assigned the input sequences of length n are varied. The lower bound on distortion after n uses of the channel is that given by Theorem 1 using C_n for C.

We can also define the capacity C for such a channel as $C = \limsup\limits_{n \to \infty} C_n$. The positive parts of the theorem then state that one can find arbitrarily long block codes satisfying Theorem 3. In most channels of interest, of course, historical influences die out in such a way as to make $C_n \to C$ as $n \to \infty$. For memoryless channels, $C_n = C$ for all n.

Duality of a Source and a Channel. There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel. This duality is enhanced if we consider channels in which there is a ''cost'' associated with the different input letters, and it is desired to find the capacity subject to the constraint that the expected cost not exceed a certain quantity. Thus input letter i might have cost a_i and we wish to find the capacity with the side condition $\sum_i P_i a_i \le a$, say, where P_i is the probability of using input letter i. This problem amounts, mathematically, to *maximizing* a mutual information under variation of the P_i with a linear inequality as constraint. The solution of this problem leads to a capacity cost function $C(a)$ for the channel. It can be shown readily that this function is *concave* downward. Solving this problem corresponds, in a sense, to finding a source that is just right for the channel and the desired cost.

In a somewhat dual way, evaluating the rate distortion function $R(d)$ for a source amounts, mathematically, to *minimizing* a mutual information under variation of the $q_i(j)$, again with a linear inequality as constraint. The solution leads to a function $R(d)$ which is *convex* downward. Solving this problem corresponds to finding a channel that is just right for the source and allowed distortion level. This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus we may have knowledge of the past but cannot control it; we may control the future but have no knowledge of it.

BIBLIOGRAPHY

[1] C. E. Shannon and W. Weaver, *The Mathematical Theory of Communication*, University of Illinois Press, 1949.

[2] R. W. Hamming, "Error-Detecting and Error-Correcting Codes," *Bell System Technical Journal*, Vol. 29, 1950, p. 147.