RESEARCH PAPER P-160

TRANSMISSION OF COMMAND INFORMATION

Leonard Farkas

October 1965

INSTITUTE FOR DEFENSE ANALYSES
RESEARCH AND ENGINEERING SUPPORT DIVISION

IDA/HQ 65-4327
Copy 23 of 50

RESEARCH PAPER P-160

TRANSMISSION OF COMMAND INFORMATION

Leonard Farkas

October 1965

IDA

INSTITUTE FOR DEFENSE ANALYSES
RESEARCH AND ENGINEERING SUPPORT DIVISION

Contract SD-50
Task T-32

*The **Institute for Defense Analyses** produces three kinds of publication for distribution, entitled Report, Study, and Research Paper.*

*A **Report** embodies the results of a major research project undertaken by IDA and is intended to be an authoritative contribution on its subject.*

*A **Study** is a less formal document and less comprehensive in scope than a Report. It may be the result of a smaller and more narrowly defined research project or it may be a supporting technical paper prepared in connection with a major project.*

*A **Research Paper** represents the work of one or more named authors but is subject to review comparable to that for publication in a professional journal.*

CONTENTS

FIGURES

BLANK PAGE

SUMMARY

The purpose of this paper is to reconsider the concept of an information measure with respect to a command transmission system. Such a system differs from other communication systems in two fundamental ways. First, a command channel may be established to transmit one and only one command message; thus, the model of a stochastic source transmitting a large number of symbols on a probabilistic basis is not a valid one. Second, for each independent command transmission, it is possible to ascribe a cost or value as a measure of the effectiveness of the received transmission.

As an introduction, we briefly consider the basis for Shannon's entropy function. In addition, we illustrate some of the difficulties which arise when we attempt to use either Shannon's single-letter distortion measure or local distortion measure. This leads us to an approach which characterizes the command transmission channel not by its individual symbols but rather by the information content of the total message.

A model of a command transmission system is proposed which differs fundamentally from the usual communication channel model in that the information source does not select symbols or messages on a probabilistic basis. Instead, by means of a conceptual feedback channel, the source estimates the effectiveness of each possible message. Desirable properties for a measure of command information are discussed. First, it is shown that it is reasonable to consider a single command transmission system as the basis for such an information measure provided we later establish the information content of a sequence of independent commands. Second, if all commands, including the null-command, have either zero or the same cost, a transmission is basically not required and the information measure should

yield a zero. Third, since the amount of information transfer is a function of the state of knowledge of the source, the information measure must be a function of a probability distribution which specifies the a priori probability of selecting each message. A condition of unity probability for a particular command only implies perfect knowledge by the source; an information transfer is still necessary. For a condition of maximum uncertainty specified by equal a priori probabilities and a single desired command, we require that the information measure be a monotone function of the total number of commands.

As an approach to the redefinition of the entropy of an information system, a thermodynamic analogy is considered. In the Boltzmann formulation of statistical mechanics, the entropy S is defined by

$$S = k \log W_{max}$$

where

$$W = \frac{n}{\prod_i n_i!} \prod_i p_i^{n_i}$$

and the following conservation conditions are imposed

$$\sum_i n_i = n$$

$$\sum_i n_i E_i = E$$

The analogy presented is based upon the following interpretation of the variables:

Variable	Thermodynamic Theory	Information Theory
E_i	Energy of a particle in the i^{th} cell.	Effectiveness (cost or value) of a message resulting in a state X_i.
n_i	Number of particles in the i^{th} cell.	Number of messages with effectiveness E_i.
p_i	A priori probability that a particle is in the i^{th} cell.	A priori probability that a message in the i^{th} cell will be transmitted.

The entropy is calculated and is given by

$$S = \log \sum_i p_i e^{-\beta E} + \beta \sum_i (n_i/n) E_i$$

where β is a Lagrange multiplier which is proportional to temperature in a thermodynamic system and is later redefined for a command transmission system.

Next, the implications of measuring the information content of a command transmission by means of the Boltzmann entropy function are explored. It is shown that the functional form of the entropy is independent of an additive constant in the cost function and that a constant or zero cost results in zero entropy as required. Large positive values of the entropy arise when the source has knowledge of the zero cost command message and there exists a high risk situation, i.e., in situations where effective transmission is possible and required. Uncertainty by the source of the desired message tends to reduce the effectiveness and need for transmission and, correspondingly, reduces the entropy. Furthermore, the role of an observation is immediately evident since it is shown that the entropy of the source increases and a more effective transmission is possible if the probability of selecting a zero cost command is increased.

Two approaches to the definition of the Lagrange multiplier β are taken. First, it is shown that if β is treated as a transform variable we obtain the important result that the entropy of command information is the logarithm of the characteristic function. As a direct consequence of this realization, it is easily shown that the entropy of independent command transmissions is additive. The second approach taken in the definition of β is an extension of the thermodynamic analogy and constitutes a definition for the "temperature" of a command transmission system. Since the thermodynamic temperature is a measure of an average kinetic energy, it appears reasonable to relate $\theta = 1/\beta$ to the ease or affinity of a system toward change. This results in a consideration of transition probabilities $p_{ij} = p(E_i | E_j)$ of a command transmission. In keeping with the functional form relating temperature to an average kinetic energy, θ is defined as

$$\theta = \sum_j p_j (E_j - \bar{E}_j)^2$$

where

$$\bar{E}_j = \sum_i p_{ij} E_i$$

The term $(E_j - \bar{E}_j)^2$ is indicative of the tendency of the channel toward change and plays the role of $(\text{velocity})^2$; the probability p_j establishes the relative importance of each term in analogy with the mass term in the kinetic energy. However, the importance of the above definition lies in its implications. If a channel causes significant changes in the transmissions, the information temperature is high, resulting in a low entropy indicating that an effective information transfer is not possible. Furthermore, although small values of θ signify reliable transmission, a low value of entropy is still possible if the probability of selecting a desired command is low. It is also shown that if we consider both

$$Z_0 = \sum p_i e^{-\beta \varepsilon_i} \qquad (0 < Z_0 \le 1)$$

and

$$\bar{E} = \sum p_i E_i$$

as design parameters for a command transmission system, then the temperature (and channel) is specified by

$$\theta \le \frac{\bar{E}}{\log \frac{1}{Z_0}}$$

A minimum value for θ (expressing a minimum effect of transitions caused by the channel) is shown to occur when the channel is designed so that the total expected cost upon reception is equal to the total cost of the transmission, i. e., when

$$\sum_j E_j = \sum_j \bar{E}_j$$

Since we strongly emphasize the relative effectiveness of the different messages which are possible, we next consider the application of correlation techniques in the achievement of appropriate transitional probabilities. It is shown that for a command situation in which "removal of uncertainty" is not an adequate measure of information, Viterbi's independent variable $ST/N_0 \log_2 M$ is not appropriate. By merely considering ST/N_0 as the independent variable, the conclusions regarding the behavior of the error probability as a function of the number of messages M are reversed. We then calculate the probability that the i^{th} command will be chosen rather than the j^{th} command in a situation in which the k^{th} command is actually the one which was transmitted:

$$P\left[(Z_i > Z_j) | Z_k\right] = \operatorname{erf}\left[\sqrt{\frac{ST}{N_0}}\left(\frac{\rho_{ki} - \rho_{kj}}{\sqrt{1-\rho_{ij}}}\right)\right]$$

Using the above, we obtain the probability of correct decoding and an upper bound for the error probability. In addition, we show that for codes with non-constant correlation between messages, we can consider an effective energy of the i^{th} command with respect to the k^{th} command as being equal to the true energy times the factor $(1-\rho_{ik})$. Finally, we compare a bound obtained by Fano with Viterbi's calculations. We show that as a result of the bound being poor, the conclusion is reached that a large number of messages (1024) are required for low error probability when, in fact, only a reasonable number (64) are required.

INTRODUCTION

In this paper we will consider the transmission of command information and we will establish that the effectiveness of such information can be incorporated into a definition of an information measure.

The development of information theory from its inception (Shannon, 1948) to the present rests upon the assignment of probabilities to a set of symbols irrespective of the ultimate use of such symbols; it is the function of the information theorist to assure the correct reception of symbols and not ideas. Within the framework of the present theory, a question concerning the ultimate use of an information bit is an improper one. The assumption is that the ultimate use of information cannot be assigned a numerical weight and hence is not amenable to mathematical analyses. Weaver[1] discusses the basic problem of communication and explains the nature of Shannon's contributions. He defines the following three levels of communication problems:

Level A: The technical problem

How accurately can the symbols of communication be transmitted?

---

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

Level B: The semantic problem

How precisely do the transmitted symbols convey the desired meaning?

Level C: The effectiveness problem

How effectively does the received meaning affect conduct in the desired way?

Our basic goal is the extension of the theory of information to include, not only Level A, but also Level B and Level C problems. However, let us next briefly review the entropy concept in the solution of the Level A problem.

Shannon's Entropy Concept

Shannon considers a discrete source and characterizes it as a stochastic process in which successive symbols are chosen according to certain probabilities, i.e., as a Markoff process. This probabilistic representation of the source is prompted by a desire to transmit semantic language information through the channel and it is shown that it is reasonable to consider such information as a complex Markoff process. Thus, it is assumed that the source of information is completely defined by a statistical model. Perhaps this point has been most clearly stated by Khinchin[1], who approaches the subject by defining the

---

[1] A. I. Khinchin, Mathematical Foundations of Information Theory. Translated by R. A. Silverman and M. D. Friedman (New York, Dover Publications, Inc., 1957), p. 2.

entropy of a finite scheme. That is, given the finite scheme

$$(X) = (X_1, X_2 \cdots X_i \cdots X_N) \text{ and a}$$

corresponding probability function

$$(P) = \left[P(X_1),\ P(X_2) \cdots P(X_i) \cdots P(X_2)\right]$$

the entropy of the finite scheme (X) is defined as

$$H(X) = -\sum_{i=1}^{N} \left[P(X_i) \log P(X_i)\right]$$

where this definition for H is justified on the basis of desirable properties which this measure should have.

Once it is agreed that symbol or event probabilities are the only measures of importance in an information transmission system, then by merely requiring the following three properties we arrive at a remarkable conclusion regarding the information measure H(X). Thus, if we require that

(1) H(X) be continuous in $P(X_i)$,

(2) H(X) be a montonic increasing function of N if $P(X_i) = \frac{1}{N}$, and

(3) the total H(X) be the weighted sum of the individual H's in the case of successive choices,

then Shannon proves that, within a positive multiplicative constant, the above H(X) is the only function satisfying all three requirements.

The following is the only consideration or justification of the use of the word "entropy" which was given by Shannon in "The Mathematical Theory of Communication:"

"Quantities of the form $H = -\Sigma P_i \log P_i$ play a central role in information theory as measures of information, choice and uncertainty. The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics[1] where P_i is the probability of a system being in cell i of its phase space. H is, then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\Sigma P_i \log P_i$ the entropy of the set of probabilities $P_1, \cdots P_n$."

As a consequence of this definition, the Shannon theory is concerned with the situation in which there initially exists a noiseless channel between the source and destination which is used to establish a mutual language (encoding) for the subsequent transmission of information. The fact that an information source with zero entropy does not need a channel must be interpreted as meaning that such a source may utilize the channel which is normally used to establish the encoding as a means of conveying its single message to the destination. Once this is done, it is unnecessary to

---

[1] Shannon references R. C. Tolman, Principles of Statistical Mechanics (London: Oxford University Press, 1938).

establish a second channel. The entropy of the source plays an important role in establishing the need for the source and destination to continue their communications subsequent to the dissolution of the noiseless channel available to them in order to agree upon a "language." However, given a data channel, the fact that the entropy of the source is zero does not imply that the given channel is not needed or that information cannot be transferred from the source to the destination. Both Jackson and Blundell point out that

> "the term 'rate of transmission' is somewhat unfortunate. It suggests that information is really transmitted at this rate, which ... is a meaningless statement."[1]

The "rate of transmission" actually corresponds to a potential rate if the source and destination can optimize the encoding for the given statistics.

Measures of Effectiveness

It is interesting to note that the necessity to consider some measure of fidelity or effectiveness in an information transmission system was recognized by Shannon in his basic paper "The Mathematical Theory of Communication"

---

[1] W. Jackson, Communication Theory (London: Butterworths Scientific Publications, 1953) p. 18.

and was further developed in a subsequent paper "Coding Theorems for a Discrete Source with a Fidelity Criterion."[1]

In his basic paper Shannon introduces a "fidelity evaluation function" for a channel which transmits a continuously variable quantity since such a channel is generally required only to transmit information to within a specified accuracy requirement.

However, in spite of the above recognition, the basic Shannon philosophy that a consideration of meaning or effectiveness is outside of the technical realm took hold so strongly that we find essentially no theoretical development along these lines between 1948 and 1960. In 1960, Shannon published his paper "Coding Theorems for a Discrete Source with a Fidelity Criterion" in apparent recognition that a consideration of the theory of information transmission is basically incomplete unless a cost or value measure is associated with the information. Let us examine next Shannon's "fidelity criterion" or "measure of distortion" and some of the problems which arise as a result of its definition.

Consider a source which is capable of transmitting words composed of letters M_i. The source thus produces a sequence of letters $\{M_i\}$ and designates a sequence of length t, $\{M_i\}_t$

---

[1] In Information and Decision Processes, edited by R. E. Machol (New York: McGraw-Hill Book Co., Inc., 1960).

as a word M; i.e., $M = \{M_i\}_t$. The sequence $\{M_i\}_t$ is transmitted over a noisy channel and received as a sequence $\{Z_i\}_t$. We assume for simplicity that there exists a unique correspondence between the M_i and Z_i alphabets. That is, for each distinct M_i there is a corresponding distinct Z_i. In addition, we order the symbols of the two alphabets so that the correct reception of the i^{th} symbol of the transmitted alphabet produces the i^{th} symbol of the receiver's alphabet.

A single-letter distortion measure, d_{ij}, is defined as the cost associated with the reception of the i^{th} symbol of the source alphabet as the j^{th} symbol of the receiver's alphabet. The only restrictions placed on d_{ij} are

$$d_{ii} = 0$$

$$d_{ij} > 0 \qquad i \neq j$$

That is, costs are defined as positive quantities, there is no cost associated with the correct reception of a given symbol, and there is no upper limit defined for d_{ij}. The distortion of a given transmitted word, M, with respect to a given received word, Z, is defined as

$$D\,(M,Z) = \frac{1}{t} \sum_{k=1}^{t} d_{M_k}\, z_k$$

and the overall distortion of the system is defined as

$$D = \sum_{M,Z} P\,(M,Z)\; D\,(M,Z) = \sum_{M,Z} P(M)\; P(Z\,|M)\; D(M,Z)$$

where P (M,Z) is the joint probability of words M and Z. (Note that in the above notation M_k and Z_k are not corresponding symbols of the M and Z alphabet -- they are corresponding symbols of the M and Z received words so that $d_{M_k Z_k}$ is not identically zero.)

Let us review the implicit assumption made previously that the total distortion is adequately represented by the average of the individual letter distortions. It is easily illustrated that this assumption is not valid in many situations. For example, assume that a particular transmission system is being used to transmit decimal digits between 10 and 99 and we have decided that a suitable single letter distortion measure is the magnitude of the numerical difference between a transmitted and received digit. If the number "80" is transmitted and is received as either "70" or "81", the single-letter distortion measure obtains the unrealistic value of ½ in both cases. Furthermore, even if we attempt to weight the cost of successive digits, we run into difficulties. In the above example, if we weight the cost of a first digit error by a factor of ten over that of the second digit, we obtain the apparently satisfactory distributions of 5 and ½. But consider the situation if "80" is transmitted and received as either "89" or "79". In the unweighted case, $D_{89} = 4.5$ and $D_{79} = 5$ -- which is obviously undesirable. However, even if we attempt to weight the successive numerals, we obtain $D_{89} = 4.5$ and

$D_{79} = 9.5$ -- in this case our attempt to improve things by weighting has yielded poorer results. Thus, in this practical example and in many other cases, a single-letter distortion measure is inadequate. Apparently a recognition of this difficulty led Shannon to the definition of a "local distortion measure" which we shall consider next.

The intent of the "local distortion measure" is to remove some of the difficulties associated with the single-letter distortion measure by comparing corresponding blocks of transmitted and received symbols rather than individual symbols. A "local distortion measure of span g" is a cost function $(M_1, M_2 \cdots M_g; Z_1, Z_2 \cdots Z_g)$ of transmitted and received sequences of length g (i.e., g symbols). The distortion D(M,Z) between words $M = \{M_i\}_t$ and $Z = \{Z_i\}_t$ where $t \geq g$ is defined as

$$D(M,Z) = \frac{1}{t-g+1} \sum_{k=1}^{t-g+1} d\left(M_k, M_{k+1} \cdots M_{k+g-1}; Z_k, Z_{k+1} \cdots Z_{k+g-1}\right)$$

and the distortion of a block code is defined as

$$D = \sum_{M,Z} P(M,Z)\, D(M,Z)$$

This measure requires that we associate a cost to successive blocks of g transmitted and received symbols. Instead of averaging the cost of successive (single) symbols as done previously, we now average the cost of successive groups of g symbols. However, a difficulty arises as a result of the definition of the local distortion measure of span g as a function of g symbols of both the uncoded and the coded messages. Assume for the moment that g = t; i.e., the span is equal to the length of the uncoded word. We must then determine the cost of receiving this complete word as a fraction of a coded word which may not even uniquely define a word. The situation is further complicated if $g < t$ and $t < t'$. In this case we must somehow associate a cost (for each possible span of length g) given only a portion of both the source and channel words which may not uniquely define either word.

Problems of the above nature arise when we attempt to characterize the communications channel by its individual symbols, or by groups of symbols, when in fact the information for the most part is contained in the word. As a result of a desire for the theory to embrace a generalized communications system with no restrictions on the structure or content of the words, the price paid is the requirement that for this communications system we must be able to specify a very general cost matrix.

In many situations a matrix formulation is utilized or proposed although the basis for this method lies in a vector formulation. The following example will help clarify this point. Consider a lunar impact vehicle which has a specified desired impact point and a cost function which is equal to the distance from the desired impact point, R, for $0 \leq R \leq 1$ and which is equal to 1 for $R > 1$. Assume that we have available four possible command messages which can cause a change in the vehicle impact point of 0, 0.5, 1, and 5, respectively. For this situation we can write the following cost matrix:

Change Desired	Change Achieved: 0	0.5	1	5
0	0	0.5	1	1
0.5	0.5	0	0.5	1
1	1	0.5	0	1
5	1	1	1	0

For example, if a change of 1 is desired to bring the vehicle to the desired impact point and a change of 5 is actually achieved, then the vehicle is 4 from the impact point and this corresponds to a cost of 1. Thus, it appears that a cost matrix is required. However, consider the following cost vector, which is derived from the given cost function:

State Achieved R_i	Cost $C(R_i)$
0	0
0.5	0.5
1	1
4	1
4.5	1
5	1

Each entry in the cost matrix can be obtained by referring to the cost vector; in the previous example $C_{5,1} = C(4) = 1$. The significant point of the above discussion is that, although a matrix formulation is possible, it is neither a necessary nor a desirable approach for our model of a command transmission system.

In conclusion, the distinction between the matrix and vector formulations may be considered to lie in the "information content" of the source. In the matrix formulation there is no information available at the source; we are really specifying the desirability of receiving the i^{th} message if a situation arose in which the j^{th} message was the desired one. The source does not know the state of affairs it desires to achieve. In the vector formulation, the source can specify the results it wishes to achieve; it knows the relative importance of the different states (perhaps as the result of an observation) and it wishes to utilize the channel in order to convey this information to the destination. Since we wish to consider transmission systems which may be used to transmit only a single command, it is necessary that our measure of information represent the amount of information which is transferred by a message from the source to the destination. But we can only have an information transfer if the information which is to be transferred is available at the source; and it is the vector form of the cost function which provides the source with the information which must be transferred.

Our approach to the definition of an information measure differs from that of Shannon's in three fundamental areas. First, we do not require our theory to include a general and undefined message source; we consider only those channels whose transmissions always result in a distinct action with a measurable end result, i.e., a command transmission system. Second, from the outset we emphasize that the basic commodity of such a system is the complete command word and not the individual symbols of the word. These two assumptions are designed to enable us to avoid some of the above problems which arise in Shannon's paper by limiting ourselves to situations in which the assignment of cost or value is much less ambiguous. Our third and most basic difference from both of Shannon's papers involves the fundamental measure of information content. Whereas Shannon measures the importance of an information transmission by the extent to which it reduces uncertainty, our measure will depend on the extent to which the received message results in the desired outcome.

Although we will recognize the existence of an a priori probability distribution which specifies the probability that a given message will be the one which is transmitted. We will additionally require that the information measure include a quantitative measure of the effectiveness of a transmission. In order to achieve this new emphasis, it is necessary for us to re-interpret the entropy concept introduced by Shannon. But first we must define a model of a command transmission system and consider desirable properties of an information measure for such a system.

A MODEL OF A COMMAND TRANSMISSION SYSTEM

We will next modify Shannon's model of a general communication system in order to obtain a representation more suitable for a command transmission system. We consider such a system to be one in which every message transmission causes an action on the part of the destination which results in a change of state.

The transmission of command information consists of the performance of the following sequence of events which may seem obvious but must be presented here to clarify several important points. Given a command transmission system, it must:

1. Establish the relative desirability of all possible results (states) and select the most desired state.
2. Observe and/or estimate the state of the system with no command transmission (i.e., with the null-command).
3. Estimate the action required to achieve the desired state on the basis of a knowledge of the systems' characteristics.
4. Select and transmit the symbols of the message which will accomplish the required action.
5. Receive the symbols and determine the corresponding message.
6. Perform the action specified by the received message.
7. Arrive at a new state.

We distinguish between messages, M_i; actions, A_i; and states, X_i by considering that a message causes an action which results in a change of state. The state is defined to be any variable in which we are able to express a cost or effectiveness measure. For example, if an action causes a burn-out velocity change in a minute, thereby causing a change in the impact location, we are free to consider the state variable as either the velocity at burn-out or the impact location. Our choice of state variable is completely dependent upon the coordinate in which we wish to express a cost or effectiveness and may even reduce to message or action variables.

It is not desirable to assume that the purpose of the channel is to transmit a large number of such commands; in fact, in many applications a command channel is established to transmit one and only one command message. Consequently, the Shannon model of a stochastic source transmitting a large number of symbols on a probabilistic basis is not a valid one. The information measure we seek should be capable of describing the quantity of information transferred by a system which may only be called upon to transmit a single message. Once we have accomplished this, then we can investigate the properties of the information measure as a function of successive commands.

Is it reasonable to consider a single command transmission system as the basis for an information measure? Consider for the moment a two-command system in which the destination can arrive at a state X as a result of the first command and the state Y as a result of the second command. Two situations may arise; Y may be dependent on X or Y may be independent of X.

If Y is dependent on X, then we may consider an equivalent single message consisting of all possible combinations of the first and second messages as resulting in the final state, Y. Our information measure is then valid for this single equivalent to the two-command transmission. If Y is independent of X, then it is indeed reasonable to first consider the information measure on the basis of X and Y individually, and then to consider the amount of information contained in a sequence of independent (single or combined) commands.

Our model of a command transmission system contains all of the elements of the general communication channel (Fig. 1). In addition, since the function of the command is to cause a desired action, we must provide the "information source" with the information necessary for it to select a particular message from all of the possible messages. A convenient mechanism for doing this is to provide a conceptual feedback channel which enables the source to estimate the state of the destination for each of the possible received messages. In this context we consider conceptual feedback channel in that each possible message results in a possible state which becomes known to the source. Furthermore, to assist the source in the determination of the state resulting from each possible message, we provide a means for observing (i.e., measuring) the system.

The observed coordinates may be the state variable itself or they may require a transformation (prediction) to yield the state variable.

The procedure by means of which the information source selects the particular message to be transmitted is one in

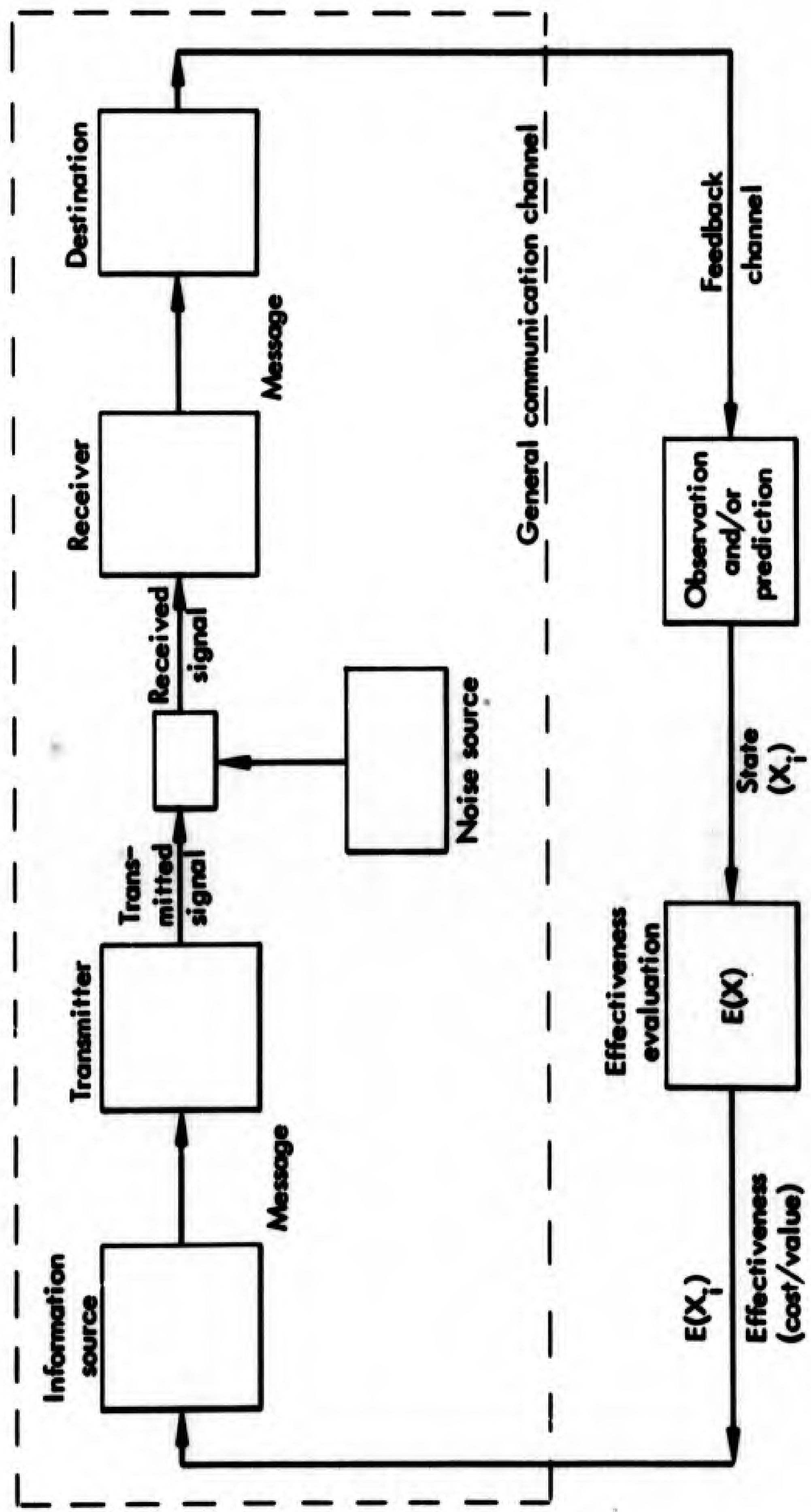

FIGURE 1 Command Transmission Channel

R11-19-66-1

which the desirability of each possible transmission is evaluated. We can consider an Effectiveness Function (a term suggested by Weaver's "Level C - the effectiveness problem") incorporated within the feedback channel which assigns a numerical "cost" or "value" to the successive states which are predicted as a consequence of the conceptual transmissions of all possible messages. (Since it is immaterial for the present whether we use a "cost" or a "value," let us consider the Effectiveness Function a generic term for either of the two.) It is possible that the same numerical value will be assigned to many of the possible states. That is, there will generally exist groupings of states for which the same cost or value may be assigned to all members of each group. For a single command message the Effectiveness Function depends only on the evaluation of the relative desirability of having the destination arrive in a particular state. Since this may involve subjective judgments, the Effectiveness Function is maintained separate from the "Information source" of Shannon's general communication system in order to preserve the basic intent of an information source which is free of any subjective connotation. Thus, a source is represented here as a device which, on the basis of the knowledge supplied by the feedback channel, determines and provides the appropriate message to the channel. The extent of the knowledge which can be supplied by the feedback channel will be related to the source's a priori uncertainty to the desired message.

A most important distinction between our model of a command transmission system and Shannon's model of a general communication system arises directly as a result of the feedback channel.

As a consequence of the feedback channel, we must consider that there is an input to Shannon's "information source" and an output from the "destination." The source is no longer considered a purely stochastic process supplying symbols only on the basis of some probability distribution. In fact, we find ourselves with the somewhat curious terminology that "we must provide the information source with the information." The "destination" is not merely a recipient of symbols; it is an active agent whose actions depend on received messages. Of course we do not gain these additional degrees of freedom without some corresponding loss. Whereas Shannon's general communication system may be used to transmit an endless stream of symbols restricted only by the probability distribution, the command transmission system considered above is used only to transmit a single message in order to accomplish a desired objective.

Desirable Properties for a Measure of Command Information

We must always bear in mind that the function of a command transmission system is to produce a desired response at the receiver. The number of symbols, the type of symbols, or the language used to achieve the desired result should be considered of secondary importance; thus, the information measure should apply to the entire message whether the command be a number, a sentence, or a paragraph. Using this approach, we can encompass the type of problem which has been posed by Brillouin with respect to a telegram which has been transmitted with the following code agreed upon by the source and receiver. If the last symbol of the telegram is a zero, the entire message is to be ignored; if it is a one, then the message is to be accepted. Brillouin asks, how much information is conveyed by the last symbol? It must somehow be related to the information content of the telegram since that one symbol can negate the entire telegram. This situation is analogous to one which arises quite commonly in the transmission of digital commands. In many cases, a command is transmitted for storage in a digital memory unit. At some later time, a second, simpler message is transmitted which sends the stored message to the controlled system for action. Thus, this second message is completely analogous to the zero or one at the end of the telegram. If it is not received, then the previous message is negated; if it is received, then use is made of the message. The two dependent messages "telegram plus zero" may be considered to constitute a single message equivalent to what we have previously called the null-command. Since we are emphasizing the information measure of the total message, the fact that one part of the message is a complete telegram and the second part is but a single symbol is relatively unimportant. Let us explore next certain functions of a command transmission system, the parameters which characterize these functions and how these parameters should effect a measure of information. Since it is our viewpoint that an information measure should describe the ability of the system to select, transmit and utilize information effectively, we will consider desirable properties of our information measure with respect to an effectiveness function which assigns a numerical cost or value to each of the states which can arise as the result of our transmission. The selection of a cost representation seems more convenient on the basis of a natural correspondence between a zero cost and the most desirable state. However, neglecting for a moment the convenience of a zero cost, it seems that the reference level for the cost function is somewhat arbitrary. Aside from the fact that it is more natural to consider costs as positive quantities, the exact numerical cost assignments for the different possible states are based upon subjective judgements. Thus, we require that the functional form of the information measure, which we choose to base on a cost function, should be independent of a constant or reference level of the cost.

A further extension of this requirement may be deduced by considering a command transmission system in which a zero cost has been assigned to each possible command. It is clear that in such a situation we can dispense with the channel since every, or any, state is a desirable one. Furthermore, the assignment of the same numerical cost to each possible command does not change the above conclusions. That is, if the cost is a constant one, irrespective of the message transmitted, then there is essentially no need for the channel. If our information measure is to represent the effectiveness of a channel then it seems appropriate to require that it yield a zero of information for the above situations in which a transmission can have no effectiveness as a result of the cost assignments.

In considering the arbitrary nature of the cost function with respect to the zero or most desirable level, we must also mention the degree of freedom available in the scale factor.

That is, our cost scale can range from zero to 0.1, or to 1, or up to 100, etc. Since the information measure must be a function of the spread between the lowest and highest cost states, it appears convenient to require that in the information measure itself be incorporated a normalizing factor to account for the possible scale changes.

We have been emphasizing very strongly the need to include a cost or effectiveness function in the information measure. However, we do wish to retain the concept of "removal of uncertainty" as one of the functions of a communications system. We are faced with the problem of defining what we mean by uncertainty, or more specifically, of defining a probability function for a command transmission system. Since we wish to deal with complete messages rather than individual symbols, we must define a probability function over a message space rather than over a symbol space. In the Shannon theory the probability of a symbol is quite naturally associated with the relative frequency of that symbol in a typical long sequence of symbols. Our situation is not analogous: we will not be transmitting a long sequence of complete messages and thus cannot interpret the probability of a message in terms of its relative frequency. In order to obtain a suitable interpretation, we must reconsider the sequence of events occurring prior to the transmission of a command.

It is assumed that the source has established the relative desirability or cost for all possible outcomes and has selected the most desirable outcome. A finite number of possible command transmissions are available and the source must select one of these command messages in order to achieve the desired result. To assist in the selection of the appropriate command, the source must have some means of observing the present state of the controlled system in order to estimate the action required for the accomplishment of the desired result; i.e., we assume a measurement capability. Since the final state which is achieved is a function of this measurement capability, it is necessary to include some characteristic of it in our information measure. That is, if the source had imperfect knowledge of the controlled system (or the dynamics of its behavior), then an incorrect command would be selected and transmitted. The net result of an error-free transmission would be somewhat equivalent to a situation in which the source had perfect knowledge but an error occurred in the transmission. In both cases, the most desired state is not achieved. We can consider that the measurement supplies information to the source which then must transmit it to the controlled system. In any event, the measurement or observation function is inextricably tied to the information transmission process; the question we next raise is what parameter best characterizes the observation function?

We approach this with the somewhat pragmatic viewpoint of questioning the influence of the measurement on the selection of the desired command. Prior to a measurement or observation, the source may be considered to have some knowledge of the controlled system. It can be assumed that a tentative identification of the respective costs associated with each possible command can be made before an observation on the basis of the a priori knowledge. However, there is naturally a degree of uncertainty associated with the cost assignments. For example, if, before an observation, we assign a certain cost to a particular command we must recognize that as the result of the observation the cost assignments may change. The effectiveness of the observation should be related to our a priori uncertainty of cost assignments. This uncertainty can be expressed in terms of the probability for each command; that as a result of the observation the source will select that command as the desired transmission. The assignment of a probability equal to one particular command implies a certainty with respect to the result of the observation and, hence, no need for the measurement. We can readily extend the above concept to include, not only the uncertainty over the outcome of the measurement, but also the uncertainty over the dynamic behavior of the controlled system. Even if there is no uncertainty over the measurement of the present state of the system, we may be unable to specify exactly the influence of a command on the system. That is, although we can specify the desired result, we are uncertain of the command which will achieve this result. This may also be considered part of the a priori uncertainty. It is unnecessary for us to consider the mechanism by which the probabilities are arrived at, just as it is unnecessary for Shannon to consider the derivation of the probabilities of the symbols used as inputs to his channel. It suffices to say that we may consider the a priori assignment of both cost and probability on a subjective basis (with the probability assignment restricted, of course, by the usual laws of probability). We can note the behavior we should require of the information measure as a function of this a priori uncertainty. Since it basically represents the state of knowledge of the source, the fact that a particular command may have a probability of one only implies perfect knowledge by the source; an information transfer is still necessary from the source to the controlled system and thus the measure of information should not be zero in this situation.

The specification of a situation of maximum uncertainty requires some care since we require an a priori estimation of both the cost associated with each possible command message and a probability for each command that it will represent the desired command after we have observed the controlled system. However, a condition of maximum uncertainty would preclude our estimating a tentative cost for each command. We can circumvent this problem by agreeing that a situation of maximum uncertainty will be characterized by both equal a priori probabilities and a cost function which merely specifies that there exists a single desired command, all other commands being relatively undesirable. With this agreement, it is appropriate to require that the information measure be a monotonic function of the total number of commands available. The question as to whether the information measure be a monotonically increasing or a monotonically decreasing function is unimportant, being only a matter of convention in the basic definitions. That is, Shannon requires that his H function be a monotonically increasing function of the number of symbols when they are all equally probable since an increasing H is identified with an increasing uncertainty. However, to accomplish this, he merely defines H as the negative of the expected value of the logarithm of symbol probabilities.

Still another property which must be established for an information measure is that of decomposability, which is analogous to Shannon's third property discussed in the beginning of this section. However, since our viewpoint is one of considering the information content of a complete message, not of the component parts of a message, we need only establish the rules which relate the information measures of a sequence of independent command messages.

At first glance it may appear as if the work here falls within the realm of decision theory; although this may indeed occur, it certainly is not our intent. The reason for the identification of this work with decision theory may be traced to our use of a cost function and value judgements. However, the basic problem of decision theory, as the name implies, is to provide a rational set of rules which may be used to govern our actions based upon criteria which are dependent upon functions of our subjective value assignments. The problem which we attack is that of defining a suitable measure for the quantity of information which is transmitted from point to point. Once the quantity of information is defined, then it may be appropriate to investigate what can be done to maximize the rate at which this quantity can be transported.

A REDEFINITION OF THE ENTROPY OF AN INFORMATION SYSTEM

Thermodynamic Analogy

In the section entitled The Entropy Concept in Statistical Mechanics we discussed alternate approaches which have been taken historically to the definition of entropy. We showed that Shannon chose to model his information measure on the Gibbs formulation; we wish to revert to the original Boltzmann formulation in order to include additional, non-probabilistic, parameters in our information measure.

Since we wish to emphasize the information flow or transfer property of a command transmission system, we can provide a more intuitive picture by a slight change of terminology. Thus, let us use the terms "source," "medium," and "sink" to correspond to what is generally called the information "source," "channel," and "receiver." We can consider the source, medium, and sink as representing a closed system in which an "energy balance" is maintained. If we agree that the Effectiveness Function, $E(X_i)$, assigns a cost to each possible state, X_i, of the sink then $E(X_i)$ may be considered as the "energy level" of the sink with respect to the source. The object of a command transmission is not merely to reduce uncertainty but also to lower the energy level of the sink. The transfer of information from the source to the sink results in a decrease in the energy level of the sink which we can imagine as an energy flow back to the source. In a sense, the information flow "balances" the energy flow and thus may be considered an "information energy". It is desired as a result of a command to cause a transfer of information energy from the source to the sink by decreasing the energy level of the sink. Thus, the sink may be considered to possess potential energy capable of doing work upon receipt of a message; the transmission converts the potential energy of the system to useful work. Although the received symbols may differ significantly from those transmitted, the success which has been achieved will be measured by the information energy transfer. Even if the desired minimum cost state has not been achieved, if the received message corresponds to a state which is close to the one with minimum cost then a significant energy transfer has indeed taken place.

Thus, in order to pursue the thermodynamic analogy suggested by the extensive use of the word "entropy" in information theory, it appears useful to consider the Effectiveness or Cost Function, $E(X_i)$, as representing an "information energy". We consider a controlled system capable of arriving in any one of n possible states, X_i, and we associate with each state an energy $E(X_i)$ where $E(X_i) \geq 0$. The assignment of the $E(X_i)$ may be arbitrary and subjective but it may be useful to assume that there exists at least one $i = k$ such that $E(X_k) = 0$; i.e., there is at least one state selected of the possible states which is desired as a result of a control action. In a typical situation it is likely that we will wish to assign the same energy or cost to many different states; for example, we may wish to consider that there is only one desirable state with all others being equally undesirable. Recall that in a thermodynamic system the "condition" of a system is specified by giving the number of particles in each cell of the μ-space; in Sommerfeld's[1] terminology, "Any distribution n_i determines a definite microstate of the gas." In keeping with the thermodynamic terminology, we can consider that we divide the energy space into cells and that the assignment of the different command messages into these cells represents the selection of a particular "microstate" or "condition." Thus, we divide the total energy range into energy cells E_1, E_2, ···, E_k, ···, E_t, and assign n_1 states an energy E_1, n_2 states an energy E_2, ···, and n_k states an energy E_k. One such energy distribution is called a "microstate." The information which is to be transferred from the source to the sink represents more than the designation of one state with $E_k = 0$; at the very least all n_k states with $E_k = 0$ should be designated. Furthermore, we must allow that some information has been transferred if a state with $E_j \approx 0$ has been received; thus some information is associated with a knowledge of the states n_j. We are led to conclude that once the E_i and corresponding n_i have been selected, the information to be transferred is some function of the entire microstate; that is, of the relative costs associated with each of the states possible

[1] A. Sommerfeld, Thermodynamics and Statistical Mechanics (New York: Academic Press, Inc., 1956), p. 214.

as the result of a control action. Furthermore, in a command transmission system we assume that it is possible to specify, prior to an observation, the probability, p_i, that a message in the i^{th} cell of the energy space will be selected for transmission. Since the assignment of the different messages to a particular energy cell is subjective, the "external behavior" of the system with respect to information content is unchanged if different messages are assigned to the energy cells, provided that we maintain the same number of command messages in each cell of the energy space. Recall that this is the analog of the state vs. condition representations of a thermodynamic system. The thermodynamic probability or weight, W^1, of a state is defined as the number of different ways that a given external behavior can be realized; it is the number of combinations possible in a total of n elements when they are taken $n_1, n_2, \cdots, n_t$ at a time. As we discussed elsewhere, the thermodynamic probability (not normalized to unity) is given by

$$W^1 = \frac{n!}{n_1!\ n_2!\ n_3! \cdots, n_t!}$$

In a thermodynamic system we can specify the a priori probability, P_i, that a particle will be in the i^{th} cell of its μ-space. The probability of a distribution in which n_1 particles are in cell 1, n_2 particles in cell 2, etc., is given by

$$P = P_1^{n_1} P_2^{n_2}, \ldots, P_t^{n_t} = \prod_{i=1}^{t} P_i^{n_i}$$

The function W^1 is then modified[1] and becomes

$$W = \frac{n!}{\prod_{i=1}^{t} n_i!} \prod_{i=1}^{t} P_i^{n_i}$$

W serves as the basis for the calculation of the thermodynamic entropy. In addition, two conservation conditions are required; namely, conservation of mass ($\sum_i n_i = n$) and conservation of energy ($\sum_i n_i E_i = E$). We will later show that the entropy is explicitly expressed in terms of the variables n_i, P_i, and E_i.

The following table summarizes the meaning we associate with each of the variables for both thermodynamic and information systems:

Variable	Thermodynamic Theory	Information Theory
E_i	Energy of a particle in the i^{th} cell.	Effectiveness (cost or value) of a message resulting in a state X_i.
n_i	Number of particles in the i^{th} cell.	Number of messages with effectiveness E_i.
P_i	A priori probability that a particle is in the i^{th} cell.	A priori probability that a message in the i^{th} cell will be transmitted.

As Shannon points out, the identification of his information measure is not necessary for the development of the theory. In his words, the identification is "given chiefly to lend a certain plausibility to some of our later definitions. The real justification of the definitions, however, will reside in their implications."

---

1 R. B. Lindsay, Concepts and Methods of Theoretical Physics (New York: D. Van Nostrand Co., 1951).

By considering the n_i and p_i in the above equation as corresponding to parameters of a command transmission system rather than a thermodynamic system, we can consider W as the informational probability or weight of the command transmission system.

In classical thermodynamics, entropy is defined as a property of the system and the entropy difference between two states is calculated by imagining that the system moves between these two states in an infinite sequence of contiguous equilibrium states; that is, the thermodynamic system is always considered as being in an equilibrium condition. In keeping with this general philosophy, the Boltzmann approach to statistical mechanics defines the entropy concept in terms of equilibrium conditions. Equilibrium, in turn, is considered as the state toward which the thermodynamic system tends to move if left by itself; that is, the state of maximum thermodynamic probability or weight. It may be considered to correspond to the "most frequent arrangement" of molecules in a thermodynamic system. In a command transmission system it may be considered as the distribution of messages over an energy or cost space which results in the maximum informational probability. Thus, Boltzmann's Principle defines the entropy, S', by

$$S' = k \log W_{max}$$

where k is Boltzmann's constant.

We next will calculate the values of the n_i's which result in a maximization of W, and we will explicitly compute the entropy. Since we assume that in a given command situation both the number of commands available and the costs associated with the different states are fixed, we are formally faced with the problem of maximizing a given function subject to certain subsidiary conditions. This problem is readily handled by the method of Lagrange multipliers; however, we require that the number of particles or commands in each cell be large. This restriction assures us that the non-integral values for the n_i, which may be obtained as a result of the maximization procedure, can be adequately approximated by what must be an integral number of particles or commands. (The above restriction that the n_i be large can, however, be relaxed to require only that the total number of particles be large - see Sommerfeld, p. 219).

Thus, to determine the entropy we will maximize

$$S/k = \log W = \log n! - \sum_i \log n_i! + \sum_i n_i \log p_i$$

subject to

$$\sum_i n_i = n$$

and

$$\sum_i n_i E_i = E$$

In order to find an extremum of a function

$$f(X_1, X_2 \cdots X_i \cdots X_t)$$

with respect to the t variables X_i when they are connected by the S subsidiary conditions

$$\Phi_j(X_1 \cdots X_t) = 0 \qquad 1 \le j \le S$$

we can introduce s Lagrange multipliers $\lambda_1, \lambda_2 \cdots \lambda_s$ and let

$$F = f + \sum_{j=1}^{s} \lambda_j \Phi_j \, (X_1 \cdots X_t)$$

The equations

$$\frac{\partial F}{\partial X_i} = 0 \qquad 1 \le j \le t$$

$$\Phi_j(X_1, \cdots X_t) = 0 \qquad 1 \le j \le s$$

represent a system of s+t equations for the s+t unknowns

$$X_1, \cdots X_t$$

$$\lambda_1 \cdots \lambda_s$$

which must be satisfied at every extremum of the function $f(X_1, \cdots X_t)$. Thus, in our case we require two Lagrange multipliers and obtain

$$F = \left(\log n! - \sum_i \log n_i! + \sum_i n_i \log p_i\right) + \lambda_1\left(n - \sum_i n_i\right) + \lambda_2\left(E - \sum_i n_i E_i\right)$$

The assumption of large n_i allows us to use the Sterling approximation

$$X! \cong \left(\frac{X}{e}\right)^x$$

and we obtain

$$F = \left(n \log n - \sum_i n_i \log n_i + \sum_i n_i \log p_i\right) + \lambda_1 \left(n - \sum_i n_i\right) + \lambda_2 \left(E - \sum_i n_i E_i\right)$$

$$\frac{\partial F}{\partial n_i} = (-\log n_i - 1 + \log p_i) - \lambda_1 - \lambda_2 E_i = 0$$

Thus, redefining the Lagrange multipliers by

$$\lambda_1 + 1 = \alpha$$

$$\lambda_2 = \beta$$

we obtain for the n_i which extremizes S

$$n_i = e^{-\alpha} p_i \; e^{-\beta E_i}$$

Since $\sum_i n_i = n$ we also obtain

$$n = e^{-\alpha} \sum_i p_i e^{-\beta E_i}$$

The summation $\sum_i p_i e^{-\beta E_i}$ is a form of the "partition function" of statistical mechanics and is significant because all of the thermodynamic properties of a system can be derived from a knowledge of it. In terms of the definition

$$Z_0 = \sum_i p_i e^{-\beta E_i} = \text{partition function}$$

the Lagrange multiplier α is given by

$$\alpha = \log (Z_0/n)$$

Thus, we see that one of the two multipliers need not appear in our final equations.

The entropy

$$S = k \log W \quad = k\, n \log n - \sum_i n_i \log n_i + \sum_i n_i \log p_i$$

is readily calculated using the above relationship.

$$S = k\left(n \log n - \sum n_i \log \frac{n_i}{p_i}\right)$$

$$S = k\left(n \log n - \sum n_i(-\alpha-\beta E_i)\right)$$

$$S = k\left(n \log n + \alpha n + \beta E\right)$$

$$S = k\left(n \log n + n \log (Z_0/n) + \beta E\right)$$

Therefore, the entropy S is given by

$$S = k\, n \log Z_0 + k\beta E$$

It will be convenient to define

$$k \stackrel{D}{=} 1$$

and to consider a "normalized entropy" or an "entropy per message" as

$$S = S/n = \log Z_0 + \beta/nE$$

Thus,

$$S = \log \sum_i p_i e^{-\beta E_i} + \beta \sum_i (n_i/n) E_i$$

If we let $n_i/n = q_i$ represent the probability that a randomly chosen command will have energy E_i, then we can write

$$S = \log Z_0 + \beta \overline{E}_q$$

Properties of the Entropy Function

We will next turn to a consideration of the implications of representing a command transmission system with the Boltzmann entropy function. We will show the remarkable fact that just as the Gibbs formulation was appropriate to represent a purely probabilistic communications transmission, so also can the Boltzmann formulation be used to represent a command transmission.

We first observe that the functional form of the entropy is independent of an additive constant in the cost function. For, if we let

$$E_i = G_i + C$$

we obtain

$$S = \log\sum_i p_i e^{-\beta(G_i + C)} + \beta\sum_i \left(n_i/n\right)\left(G_i + C\right)$$

$$S = \log\sum_i p_i e^{-\beta G_i} + \log e^{-\beta C} + \beta\sum_i \left(n_i/n\right) G_i + \beta C$$

$$S = \log\sum_i p_i e^{-\beta G_i} + \beta \overline{G}_q$$

In addition, we could let β represent a scale factor for the cost function by merely defining

$$F_i = \beta E_i$$

We would thus obtain

$$S = \log \sum p_i e^{-F_i} + \bar{F}_q$$

However, in a later section we will attempt to establish a more useful and physical interpretation for β.

As we pointed out earlier, the assignment of a constant cost or value to every possible command message must be interpreted to mean that a transmission is not required. For this situation we find that

$$S = \log \sum_i p_i e^{-\beta E_c} + \beta \sum_i \left(n_i/n \right) E_c$$

$$S = -\beta E_c + \beta E_c = 0$$

Thus, we find the important property that the Boltzmann entropy function assumes a value of zero in a command situation for which a transmission is not required.

We next inquire as to the behavior of the entropy as a function of the a-priori probability, p_i . First, we note that since

$$\sum_i p_i e^{-\beta E_i} \geq \sum p_i = 1 \qquad \text{(for } \beta \geq 0\text{)}$$

$$\log \sum_i p_i e^{-\beta E_i} \geq 0$$

Thus, S increases when $\log \sum p_i e^{-\beta E_i} \longrightarrow 0$

or

$$\sum_i p_i e^{-\beta E_i} \longrightarrow 1$$

The maximum value of S with respect to p_i occurs when $p_i = 1$ for $E_i = 0$ and this maximum value is equal to $\beta\bar{E}_q$. When the source has a perfect knowledge of the zero-cost command, our entropy function yields a value proportional to what can be considered the expected cost if the sink were to select a command message at random. Thus, we can consider large positive values of the entropy as corresponding to high risk command situations in which there is a strong need for a transmission.

But what if the source has mistakenly selected a non-zero cost command for transmission. That is $p_j = 1$ has been assigned to a message with cost E_j. In this case, we obtain

$$S = \log\sum_i p_i e^{-\beta E_i} + \beta\bar{E}_q = \beta\left(\bar{E}_q - E_j\right)$$

Thus, if $E_j > \bar{E}_q$ then the information entropy for a command message assumes negative values. We interpret situations giving rise to negative entropies as representing ineffective (and possibly harmful) message transmissions. Thus, if the message which the source elects to transmit has a cost greater than the expected cost, $\bar{E}_q$, of a randomly selected message, the negative value obtained by our entropy function indicates an inability to transfer effective information. It implies that we might have been more effective had the sink selected a message at random rather than the command which we elected to transmit.

We will next discuss the role which an observation plays in our measure of information. Let us assume a command situation in which the zeroth command is the one which will yield zero cost and all other commands correspond to high cost states. For this situation we obtain

$$S = \log \sum_i p_i^{-\beta E_i} + \beta \bar{E}_q \approx \log p_0 + \beta \bar{E}_q$$

where p_0 is the probability that the source, as the result of the observation, will select a command corresponding to a zero cost state. The term $\log p_0$ is a measure of the source's uncertainty regarding the appropriate command transmission. Again, we see that a high average cost indicates that a transmission is required but this must be balanced by the fact that a lack of knowledge as to the appropriate command negates the capability of the source to transmit useful information.

Since we have seen that a constant cost function which corresponds to the lack of a need for the channel results in a zero entropy, we may consider that a natural rule balancing the need and effectiveness of a transmission is given by

$$\log p_0 + \beta E_q = 0$$

In Figure 2 the region underneath the curve - $\log p_0$ corresponds to an ineffective and consequently unworthwhile transmission.

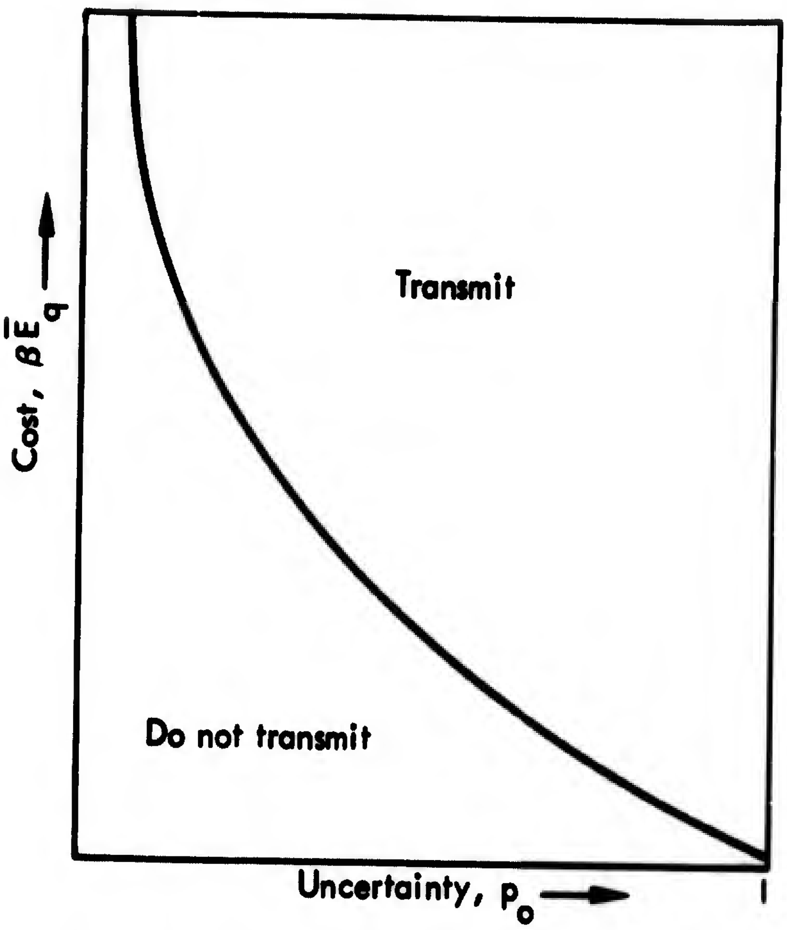

FIGURE 2 Cost-Uncertainty Transmission Rule

In the context of a command transmission system it seems that the principal function of an observation is to increase the probability of selecting, for transmission, a zero cost command. Consequently, a sequence of observations will give rise to a sequence of numbers, p_{o_1}, p_{o_2}, p_{o_3},···, representing the variation in the probability of correctly selecting a zero cost command. Prior to the i^{th} observation, we have entropy

$$S_i = \log p_{o_i} + \beta\bar{E}_q$$

and, as a result of the i^{th} observation, the entropy becomes

$$S_{i+1} = \log p_{0_{i+1}} + \beta \bar{E}_q$$

Thus, the effectiveness of the observation can immediately be related to change in entropy which has occurred as a result of the i^{th} observation

$$\Delta S_i = S_{i+1} - S_i = \log p_{0_{i+1}} / p_{0_i}$$

If, as a result of the observation, the probability of selecting a zero cost command has increased (i.e., - $p_{0_{i+1}} > p_{0_i}$) then the entropy of the source has been increased - that is, a more effective transmission is possible.

Although the above situation corresponds to the case of a desired state and all other states relatively undesirable, the results can easily be extended. A sequence of observations will, in general, give rise to the sequence of probability distributions

$$(p_{01}, p_{11}, p_{21} \cdots p_{t1})$$
$$(p_{02}, p_{12}, p_{22} \cdots p_{t2})$$
$$(p_{03}, p_{13}, p_{23} \cdots p_{t3})$$
etc.

The entropy prior to the j^{th} observation is given by

$$S_j = \log \sum_i p_{ij} e^{-\beta E_i} + \beta \bar{E}_q$$

and the entropy prior to the k^{th} observation is given by

$$S_k = \log \sum_i p_{ik} e^{-\beta E_i} + \beta \bar{E}_q$$

The entropy change caused by all observations from the j^{th} to the k-1 st is given by

$$\Delta S_{kj} = S_k - S_j = \log \left(\sum_i p_{ik} e^{-\beta E_i} \Big/ \sum_i p_{ij} e^{-\beta E_i} \right)$$

The observations result in an increase in the potential effectiveness of the transmission if

$$\Delta S_{kj} > 0$$

That is, if

$$\sum_i p_{ik} e^{-\beta E_i} > \sum_i p_{ij} e^{-\beta E_i}$$

or

$$\sum_i \left(p_{ik} - p_{ij} \right) e^{-\beta E_i} = \sum_i \left(\Delta p_i \right) e^{-\beta E_i} > 0$$

We can easily gain some insight into the effectiveness of a change in a priori probabilities by letting

$$F = \sum_i \left(\Delta p_i \right) e^{-\beta E_i}$$

and considering

$$\frac{\partial F}{\partial (\Delta p_i)} = e^{-\beta E_i}$$

Since an increase in F corresponds to an increase in the entropy, we achieve the greatest increase in entropy if the observations cause large positive changes in those a priori probabilities corresponding to the lowest cost states. Note that since

$$\sum_i \Delta p_i = \sum_i p_{ik} - \sum_i p_{ij} = 0$$

and increase in Δp_i for some $i = 1$ always causes a decrease in Δp_i for some $i = m$, thus partially offsetting the increase achieved in F. The greatest increase in entropy results when

$$E_M > E_l \quad ;$$

that is, when the observations cause a transfer of high a priori probabilities from undesirable to desirable states.

Thus, we see that the Boltzmann entropy function provides us with a suitable model for the representation of the influence which an observation has on the information content of a message. This is particularly important since our entire development is motivated by a desire to measure information content by the effectiveness of a transmission - not only its surprise value.

Entropy as a characteristic function

In the following, we will explore the relationship between entropy and characteristic functions of probability theory. First, let us change the form of our entropy function in the following manner:

Since

$$S = \log \sum_i p_i e^{-\beta E_i} + \beta \bar{E}_q$$

we can write

$$S = \log \sum_i p_i e^{-\beta E_i} + \log e^{\beta \bar{E}_q}$$

which becomes

$$S = \log \sum_i p_i e^{-\beta (E_i - \bar{E}_q)}$$

Letting $E_i - E_q = \Delta E_i$, we obtain

$$S = \log \sum_i p_i e^{-\beta \Delta E_i}$$

The constant β, which was originally introduced as a Lagrange multiplier, may now be considered a transform variable. That is, we can consider that

$$\sum_i p_i e^{-\beta \Delta E_i} = F(\beta)$$

is essentially the characteristic function[1] of the discrete

---

[1] H. Cramer, Mathematical Methods of Statistics (Princeton: Princeton University Press, 1957), p. 89.

probability distribution function of the p_i's. Since β has been completely arbitrary to this point, we could let

$$\beta = -iw$$

without changing any of our previous results. In this way

$$\sum_i p_i^{+iw\Delta\varepsilon_i} = F(-iw)$$

would correspond precisely to the characteristic function; however, for purposes of the following development it appears unnecessary to modify the definition of β .

We are now ready to consider the entropy of a sequence of commands. Assume that we are given a command transmission system which will be used to transmit two distinct commands. The first command is specified by

(1) the possible states $X = (X_1, X_2 \cdots X_t)$

(2) the costs associated with each state
$E(X) = E(X_1), E(X_2) \cdots E(X_t)$

(3) the a priori probabilities $p(X) = p(X_1), p(X_2) \cdots p(X_t)$

and the second command is similarly specified by $Y, E(Y), p(Y)$. The total cost of a dual command transmission is merely the sum of the individual costs, $E(X_i) + E(Y_j)$, and the probability of selecting both X_i and Y_j as the desired states is represented by $p(X_i, Y_j)$. Thus, the entropy of the combined command transmission is given by

$$S(X,Y) = \log \sum_i \sum_i p(X_i, Y_j) e^{-\beta(\Delta E(X_i) + \Delta E(Y_j))}$$

where $\Delta E(X_i) = E(X_i) - \sum_i \frac{n_i}{n} E(X_i)$

$$\Delta E(Y_j) = E(Y_j) - \sum_i \frac{n_i}{n} E(Y_j)$$

If the successive commands are independent, that is, if transmission of a command causing a change in X does not influence the subsequent selection of the desired state in Y, then

$$p(X_i, Y_j) = p(X_i)\, p(Y_j).$$

In this case we obtain

$$S(X,Y) = \log \sum_i \sum_j p(X_i)\, p(Y_j) e^{-\beta(DE(X_i) + DE(Y_j))}$$

$$S(X,Y) = \log \sum_i p(X_i) e^{-\beta \Delta E(X_i)} \sum_j p(Y_j) e^{-\beta \Delta E(Y_j)}$$

$$S(X,Y) = \log \sum_i p(X_i) e^{-\beta \Delta E(X_i)} + \log \sum_j p(Y_j) e^{-\beta \Delta E(Y_j)}$$

$$S(X,Y) = S(X) + S(Y) \quad \text{when X and Y are independent.}$$

Thus, we have established the additive nature of our entropy function for independent commands and have consequently justified our previous contention that a single command transmission is a suitable basis for an information measure.

In concluding this section, it seems remarkable to us that a simple reinterpretation of the parameters in the Boltzmann entropy function yields a host of results applicable to information transmission. Even more remarkable is the fact that the two different entropy functions both yield useful results for information theory.

The temperature of a command transmission system

In the previous section we have shown that certain useful results can be obtained by considering the Lagrange multiplier, β, as a transform variable. However, a second approach is possible in which our original thermodynamic analogy is extended. In the development of thermodynamics from the viewpoint of statistical mechanics it is shown that the constant β is inversely proportional to the temperature of a thermodynamic system; more precisely,

$$k\beta = \frac{1}{T}$$

where k is the Boltzmann's constant. Let us see if we can attach an analogous significance to a constant $\theta = 1/\beta$; that is, we will attempt to establish a practical significance to the "temperature," θ, of a command transmission system as it appears in the entropy function

$$S = \log \sum_{i} p_i e^{-E_i/\theta} + \frac{1}{\theta} \bar{E}_q$$

Before attempting to determine the significance of the constant θ which appears in the above expression for entropy, it is worthwhile to consider the physical significance of the thermodynamic temperature. Although there are various approaches to the definition of temperature, it appears that the concept of temperature as it arises in the kinetic theory of gases will lead to useful concepts and definitions with respect to the transmission of command information.

If we consider a perfect gas, we find that the temperature of the gas is a measure of the average kinetic energy per molecule[1]; i.e., $\frac{1}{2}\,m\overline{v^2} = \frac{3}{2}kT$. Recalling the previous discussion of the microstate concept, we see that the temperature is indicative of the volatility of a microstate. If at some instant of time the gas molecules have an initial position-velocity distribution (which is represented by a vector in a space with a sufficient number of dimensions to represent all position and velocity coordinates), then a high temperature indicates a high average velocity per molecule, i.e., in a very short time the system will be in a different microstate. Thus, in the context of an information system it appears that it is appropriate to relate the constant θ to the velocity with which the system changes state. θ is thus a measure of the ease with which a change of state occurs, or perhaps the affinity of a system toward change.

What parameters of a command transmission system characterize that systems' tendency towards change? It appears that the effect of the channel on the message transmission is an appropriate point of departure. We begin by noting an interesting possibility regarding the effect of a message change caused by the channel. In our formulation we have allowed that there may exist a probability less than one of transmitting the most desirable command; that is, an inaccurate knowledge of the physical situation may result in the transmission of an incorrect command. It is thus conceptually possible for the channel to

---

1 J.E. Mayer and M.G. Mayer, Statistical Mechanics (New York: John Wiley & Sons, Inc., 1940) p. 9.

cause, by means of a random modification of the message, an increase in the effectiveness of a transmission. A consideration of the channel in a command transmission system should recognize this conceptual possibility.

From the viewpoint which considers a channel as a means of conveying useful information the most important characteristic of a channel is its transitional probability function. That is, the probability, p_{ij}, of the reception of a message with cost E_i given that one with cost E_j was transmitted. The expected cost of the received message when E_j is transmitted is given by

$$\bar{E}_j = \sum_i p_{ij} E_i$$

and the expected change for message E_j is

$$\delta E_j = E_j - \bar{E}_j$$

Note that as we discussed above, δE_j can be positive or negative; that is, the channel can either increase or decrease the effectiveness of the transmission. In keeping with the functional form relating temperature to an average kinetic energy, and in order to maintain a positive value for our Lagrange multiplier, it appears appropriate to define an information temperature, θ, as

$$\theta = \sum_j p_j \left(\delta E_j\right)^2 = \sum_j p_j \left(E_j - \bar{E}_j\right)^2$$

Thus θ is the expected value (averaged over the a priori transmission probabilities) of the square of the expected cost changes for the individual messages. The term $(\delta E_j)^2$ is indicative of a tendency towards change and plays the role of $(\text{velocity})^2$ and the probability, p_j, establishes the relative importance of each term in analogy with the mass term in the kinetic energy.

Let us examine the implications of the above definition on our entropy function. Assuming that there exists a zero cost command with a priori probability p_0, and that all other commands have relatively high cost, we can write the entropy as

$$S = \log \left[p_0 + \sum_{i=1} p_i e^{-E_i/\theta} \right] + 1/\theta \, \bar{E}_q$$

which for small values of θ is approximately

$$S \approx \log p_0 + 1/\theta \, \bar{E}_q \; .$$

We see that as θ decreases, implying a more reliable transmission channel, the entropy increases. That is, the effectiveness of the transmission increases. However, this increase is dependent upon the ability of the source to select a zero cost message. If the probability of selecting a desired command is low, then the effectiveness of the transmission is low in spite of a faithful reproduction of the input message at the channel output. However, if the probability of selecting the desired command is high, then the effectiveness of the transmission depends upon the reliability of the channel; small values of θ

correspond to high reliability or effectiveness, consequently high entropy values.

High values of the information temperature, θ , result in an entropy which is approximately

$$S \approx \log \sum_i p_i = 0$$

If the channel causes significant changes in our transmissions, then the effectiveness decreases until transmission is no longer worthwhile, since the channel has the ultimate say as to which message reaches the sink.

In the following, we will demonstrate the manner in which the effectiveness which we demand of a transmission imposes restrictions on the information temperature or transitional probabilities of the channel. We will make use of the inequality[1]

$$\sum_i x_i \log y_i/x_i \leq 0$$

for x_i and y_i satisfying

$$x_i \geq 0 \qquad y_i \geq 0$$

and $\sum_i x_i = 1 \quad \sum_i y_i = 1$

Let $x_i = p_i$

and $$y_i = \frac{p_i e^{-E_i/\theta}}{\sum p_i e^{-E_i/\theta}} = \frac{p_i e^{-E_i/\theta}}{Z_0}$$

1 N. Abramson, Information Theory and Coding (New York: McGraw-Hill Book Co., Inc., 1963), p. 16.

where Z_0 is the "partition function" of statistical mechanics.

Then,

$$-\sum_i p_i \log p_i \leq -\sum_i p_i \log p_i + \frac{1}{\theta} \sum_i p_i E_i + \log Z_0$$

Thus, we obtain the inequality

$$\theta \leq \frac{\bar{E}}{\log \frac{1}{Z_0}}$$

where $\bar{E} = \sum_i p_i E_i$

Note that Z_0, which ranges from zero to unity, is itself a measure of the potential effectiveness of a transmission. A value of unity corresponds to a certainty that a zero cost command will be selected for transmission and a value of zero corresponds to a certainty that a very high cost command will be selected. We may consider both Z_0 and $\bar{E}$ as a design parameter of a command transmission system in that we can specify both a number from zero to unity representing the effectiveness and also an average cost which we desire. These quantities then fix, by means of the above inequality, the maximum value we can permit for the information temperature. That is, we must design the channel in a manner which limits the changes caused by the channel in accordance with the above inequality.

It is interesting to explore still another implication of the information temperature on the channel design. Since smaller values of θ imply a more effective transmission, it is reasonable to consider the minimization of θ with respect to the transitional probabilities of the channel. Ignoring for the moment the obvious solution that an absolute minimum of $\theta = 0$ is achieved when $E_j = \bar{E}_j$, we seek to determine the channel characteristics corresponding to the condition

$$\frac{\partial \theta}{\partial p_{ij}} = 0$$

From the definition of θ we obtain

$$\theta = \sum_j p_j E_j^2 - \partial \sum_i \sum_j p_{ij} E_i E_j + \sum_j \left(\sum_i p_{ij} E_i \right)^2$$

$$\frac{\partial \theta}{\partial p_{ij}} = - \partial \sum_i \sum_j E_i E_j + \sum_j \partial \left(\sum_i p_{ij} E_i \right) \sum_i E_i$$

Setting this equal to zero and simplifying, we obtain

$$\sum_j E_j = \sum_j \bar{E}_j$$

Thus, it is not necessary to require for each individual message that the expected cost of the received message be equal to the cost of the transmitted message. It is sufficient to design the channel so that the total expected cost upon reception is equal to the total cost of the transmission; obviously, this requirement is also fulfilled if the costs are equal on a message-by-message basis. However, we should recognize that it is usually not possible to insure equality between the transmitted cost and the expected cost of the received message for every message. For example, if the j^{th} message corresponds to a zero cost command, we would require that

$$E_j = \sum_i p_{ij} E_i = 0$$

Since all terms in the summations are positive, this is only possible for an error-free channel; that is, when

$$p_{ij} = 1 \qquad \text{for } i = j$$

This concludes our demonstration of the two approaches which can be taken to establish a physical significance to the Lagrange multiplier appearing in our entropy function.

CORRELATION TECHNIQUES FOR COMMAND TRANSMISSIONS

The design of a command transmission channel which will be used for only a single message or a small number of messages involves considerations not necessarily the same as those of a general transmission channel. First, the "information rate" measured in bits per second is no longer a meaningful concept. What is the information rate of a communication system which can only transmit a single "destruct" command to a faulty rocket? Second, even the entropy $H = -\sum_i P_i \log P_i$ cannot be used as a design criterion. The fact that a particular rocket may have a very low probability of requiring a "destruct" command and a correspondingly low entropy in no way influences the design of a system which must transmit such a command. Third, the specification of an error probability must be carefully considered. It is not necessarily desirable that a command transmission system (which can transmit several commands) have a uniformly low error probability over all possible messages if error probability is defined as a transitional probability between two messages. The messages and error probabilities should be designed to maximize the probability of receiving an effective message. We desire a low probability that noise will cause a transition from the transmitted command to a high cost command.

This leads into our next topic, in which we consider a technique for the design of a system which allows us the freedom of specifying the parameters which yield a required set of transitional probabilities. We shall demonstrate that a correlation decoder is a useful device not only in the classic case of minimizing the probability of error but also in our case where we wish to control the error probabilities for each individual message. Our initial approach is modeled after that of Viterbi's[1]; however, as a consequence of our application we will show that some of the conclusions we obtain are diametrically opposed to those obtained for the general information transmission applications.

---

[1]Golomb, S. W., et al., Digital Communications (Englewood Cliffs, New Jersey: Prentice-Hall, Inc., 1964) pp. 106-134.

In Figure 3 we illustrate an information source selecting a message M_i on the basis of an effectiveness evaluation for the corresponding state χ_i, $E(\chi_i)$, and the subsequent encoding and transmission of an appropriate waveform $S^i(t)$. The received signal, $R(t)$, is assumed to be corrupted by white gaussian noise $N(t)$ with power spectral density N_0; that is,

$$R(t) = S^i(t) + N(t)$$

We will, arbitrarily at this point, consider that a correlation detection process is to be used. That is, for each possible waveform $S^i(t)$ of duration T, the number

$$Z_i = \frac{1}{T}\int_0^T R(t)S^i(t)dt$$

is computed and the decision device selects that $S^i(t)$ as the channel output which corresponds to the maximum Z_i. It is well known that this represents the maximum likelihood solution in the presence of additive gaussian noise.

It will be convenient to approach the determination of the quantity we have previously called $p(E_i|E_j)$ by considering in our new notation the probability of receiving $R(t)$ given that $S^j(t)$ was transmitted; i.e. $p(R|S^j)$.

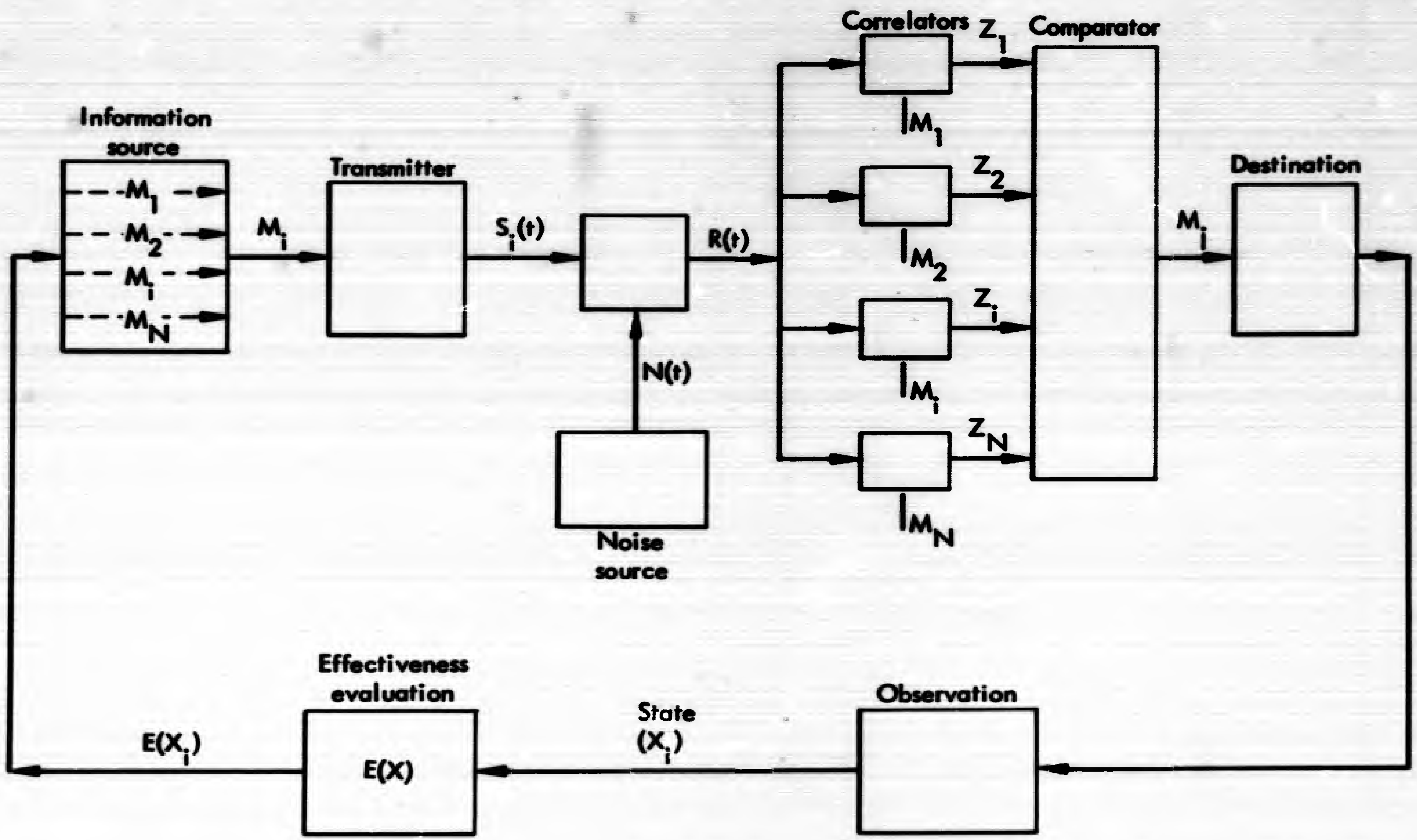

FIGURE 3 A Correlation Command Transmission Channel

RII-10-65-3

Since Z_i is a linear function of a gaussian random variable, it follows that Z_i is also a gaussian random variable whose statistics are completely determined by the mean and variance. The expected value of Z_i, given that S^j was transmitted, is

$$E(Z_i | S^j) = E\left(\frac{1}{T}\int_0^T \left[S^j(t) + n(t)\right] S^i(t)dt\right)$$

$$= \frac{1}{T}\int_0^T S^j(t)S^i(t)dt$$

Similarly, the covariances are straightforwardly calculated:

$$E\left[(Z_i - \bar{Z}_i)(Z_k - \bar{Z}_k) | S^j\right] = \frac{N_0}{2T^2}\int_0^T S^i(t)S^k(t)dt$$

which is independent of the transmitted message. It is convenient to define

$$\rho_{ij} = \frac{\int_0^T S^i(t)S^j(t)dt}{\left[\int_0^T {S^i}^2 dt \int_0^T {S^j}^2 dt\right]^{1/2}} = \frac{\int_0^T S^i(t)S^j(t)dt}{ST}$$

Then

$$E(Z_i | S^j) = S\rho_{ij}$$

$$E\left[(Z_i - \bar{Z}_i)(Z_k - \bar{Z}_k) | S^j\right] = \frac{N_0 S}{2T} \rho_{ik}$$

If i = k, we obtain the variance

$$\sigma^2_{z_i} = \frac{N_0 S}{2T}$$

If the message to be transmitted is a binary sequence, then it is convenient to consider that the signal $S^i(t)$ is generated as a sequence of two signals $\chi_0(t)$ and $\chi_1(t)$ of duration T/n (where n is the number of binary digits in a message) normalized so that

$$\int_{(T/n)(r-1)}^{(T/n)r} \chi_0(t)^2 dt = \int_{(T/n)(r-1)}^{(T/n)(r)} \chi_1(t)^2 dt = S$$

and $$\int_{(T/n)(r-1)}^{(T/n)(r)} \chi_0(t)\chi_1(t) dt = -S$$

For example, if $\chi_0(t) = \sqrt{2S}\cos W_0 t$

$$\chi_1(t) = -\sqrt{2S}\cos W_0 t$$

then

$$\rho_{ij} = \frac{1}{ST}\sum_{r=1}^{n}\int_{(T/n)(r-1)}^{(T/n)(r)} S^i(t)S^j(t)dt$$

$$\rho_{ij} = \frac{\text{No. of binary agreements - No. of binary disagreements}}{n}$$

Thus, the correlation between messages is reduced to a property of the binary coded words which are used to represent the messages.

Viterbi computes the probability of correctly decoding the j^{th} signal as the probability that $Z_j >$ Max $(Z_1, \cdots Z_{i \neq j} \cdots Z_m)$. He shows that

$$1 - P_{E_j} = P_{C_j} = \int_{-\infty}^{\infty} dV_j \int_{-\infty}^{V_j + \sqrt{2ST/N_0}(1-\rho_{j1})} \int_{-\infty}^{V_j + \sqrt{2ST/N_0}(1-\rho_{jm})}$$

$$\frac{1}{(2\pi)^{M/2}|\rho_{ij}|} \exp\left(-\frac{1}{2}V^T[\rho_{ij}]V\right) dV_1 \cdots dV_{i\neq j} \cdots dV_M$$

where $|\rho_{ij}|$ is the determinant of the correlation matrix $[\rho_{ij}]$.

Viterbi presents numerical results for the error probabilities for the following:

(1) Orthogonal codes in which $\rho_{ij} = 0$ $i \neq j$ and the number of binary digits in a message (n) is equal to the number of messages (M).

(2) Trans-orthogonal codes in which $\rho_{ij} = -\frac{1}{M-1}$ and n = M-1.

(3) Bi-orthogonal codes in which $\rho_{ij} = 0$ but $\rho_{i,\pm i} = \pm 1$ and n = M/2.

It is significant to note that a uniformly low cross-correlation is demanded between the transmitted message and all other messages. In addition, Viterbi assumes that each of the M possible messages is equally probable; as a consequence of the assumption he considers the basic parameter of the system to be

$$\frac{ST_B}{N_0} \text{ where } T_B = \frac{T}{\log_2 M}$$

since $\frac{T}{\log_2 M}$ represents the time to transmit a single bit. In Figure 4 we have reproduced Viterbi's results for orthogonal code signals. From the given curves we infer that by increasing the number of equally probable messages which we transmit, we decrease the error probability. However, this conclusion is not valid for a situation in which we have a fixed number of command messages which are not equally probable; in this case the parameter $T_B = \frac{T}{\log_2 M}$ is not meaningful.

In Figure 5 we have redrawn Viterbi's data with ST/N_0 as the independent variable. (ST/N_0 represents the total energy per message divided by the noise power spectral density.) Now we see that if the energy per message and noise are constant, then increasing the number of code

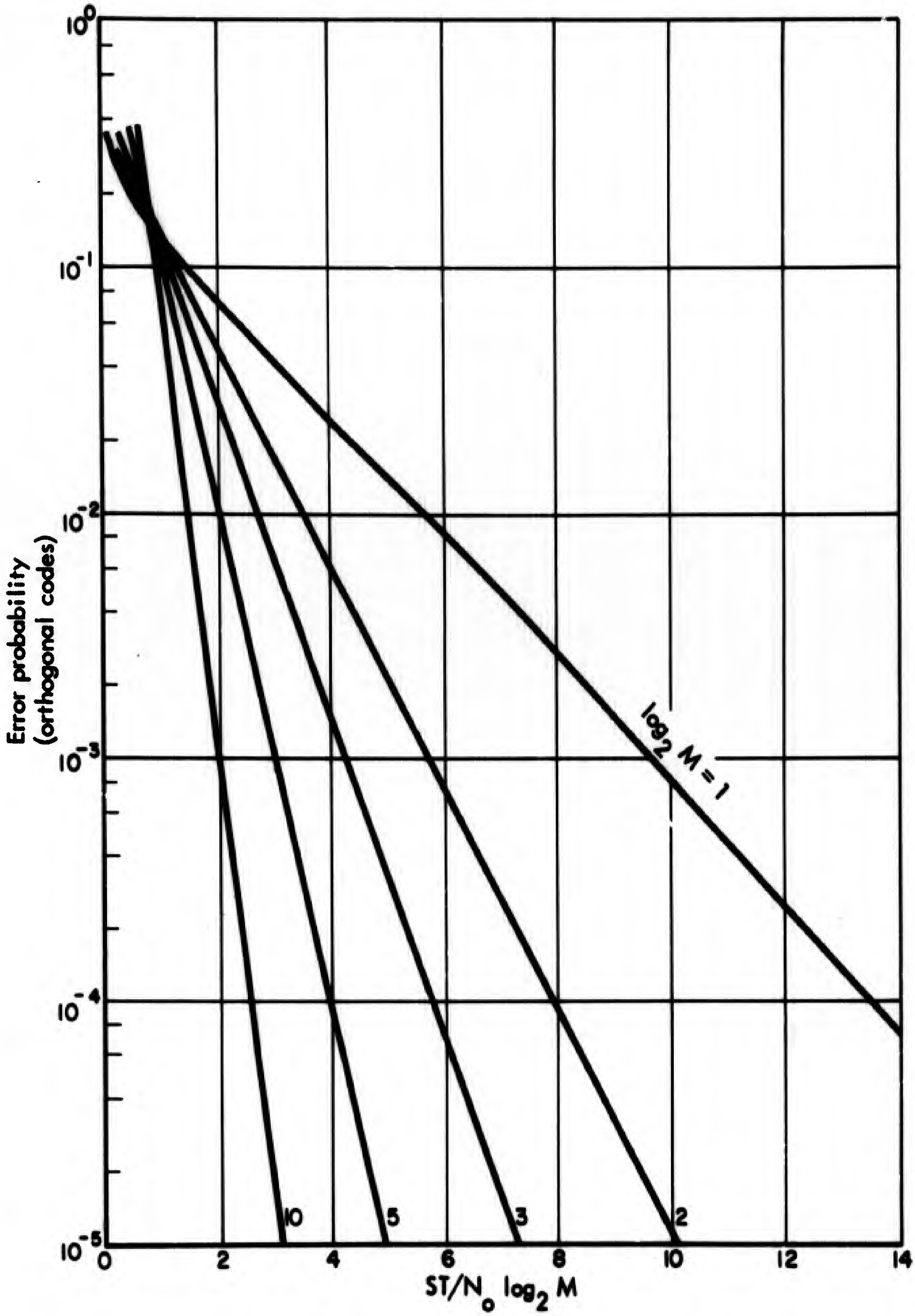

FIGURE 4 Error Probability vs. $ST/N_o \; Log_2 M$

R11-19-65-4

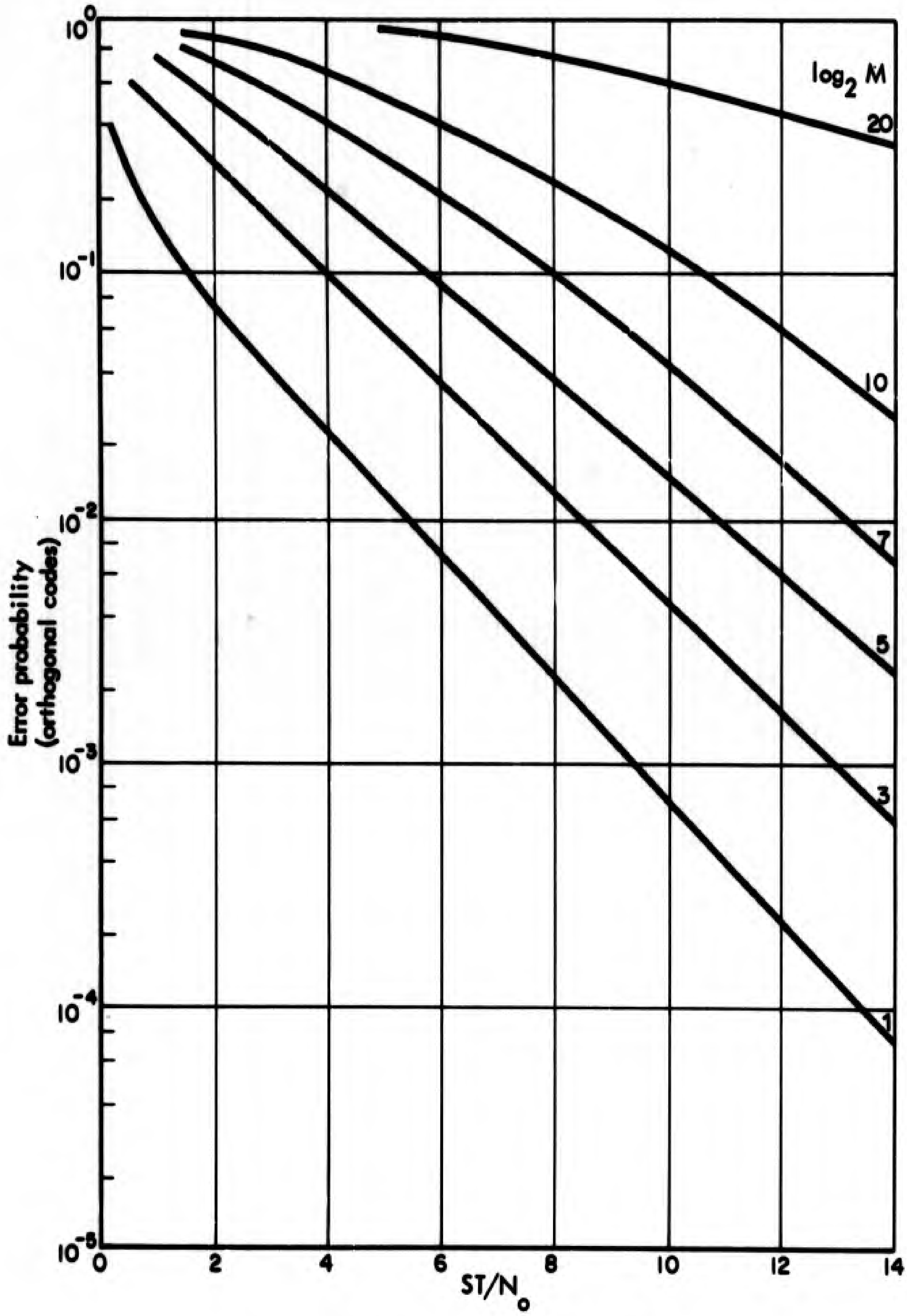

FIGURE 5 Error Probability vs. ST/N_o

R11-19-65-5

words results in a greater error probability. It is also important to note that we can make no inferences regarding the number of binary digits used in the orthogonal code words. (Of course we must be capable of representing the M distinct messages.)

The above conclusion is verified by considering the equation for the error probability given by

$$1 - P_e = P_c = \int_{-\infty}^{\infty} \frac{\exp - V_1^2/2}{\sqrt{2\pi}} \left[\int_{-\infty}^{V_1 + \sqrt{2ST/N_o}} \frac{\exp - V^2/2}{\sqrt{2\pi}} \, dV \right]^{M-1} dV_1$$

which is obtained by setting $[\rho_{ij}] = 1$ in the general equation given previously. Note that properties of the individual code words do not appear; in addition to ST/N_o, the only other parameter is M, the number of messages. Furthermore

$$0 < \left[\int_{-\infty}^{V_1 + \sqrt{2ST/N_o}} \frac{\exp - V^2/2}{\sqrt{2\pi}} \, dV \right]^{M-1} < 1$$

Since

$$\exp - V_1^2/2 \geq 0$$

it is apparent that as M increases the value of the integral decreases; this corresponds to an increase in the error probability, P_e. Thus, as we increase the number of messages in an orthogonal code dictionary, we increase the probability of incorrectly decoding the received message.

As we have discussed elsewhere, in a command situation we are often interested in the relative chance of decoding the i^{th} message as the j^{th} message. This is particularly important when a command message is composed of two parts, an address and a numerical value; for example, a guidance command such as "Pitch - 10 degrees." In fact, even the sign of a command may be considered an address; for example, in the commands "increase velocity - 10 fps" or "decrease velocity - 10 fps," "increase" and "decrease" are essentially addresses which convey a completely different importance than does the actual numerical value. A few illustrative addresses for a missile or space vehicle might be

Roll

Pitch

Yaw

Velocity change

Destruct

Power on

It is unreasonable to consider the statistics of these messages; what we are really interested in is the probability of decoding "Roll" instead of "Pitch" or "Destruct" instead of "Power-on." That is, we must calculate $p(Z_i > Z_j)$ where Z_i and Z_j are the outputs of the correlators representing the address portion of the command.

Since Z_i and Z_j are gaussian random variables, their joint probability density function is given by

$$p(Z_i,Z_j) = \frac{1}{2\pi\sigma_i\sigma_j\sqrt{1-\rho_{ij}^2}} \exp - \frac{1}{2}\frac{1}{(1-\rho_{ij}^2)}\left[\left(\frac{Z_i-M_i}{\sigma_i}\right)^2 - \frac{2\rho_{ij}(Z_i-M_i)(Z_j-M_j)}{\sigma_i\sigma_j} + \left(\frac{Z_j-M_j}{\sigma_j}\right)^2\right]$$

Furthermore, we can perform a linear transformation of Z_i and Z_j to the uncorrelated variables X_i and X_j by means of the transformation

$$X_i = (Z_i-M_i)\cos\Phi + (Z_j-M_j)\sin\Phi$$

$$X_j = -(Z_i-M_i)\sin\Phi + (Z_j-M_j)\cos\Phi$$

where

$$\tan 2\Phi = \frac{2\rho_{ij}\sigma_i\sigma_j}{\sigma_i^2-\sigma_j^2}$$

The inverse transformation is given by

$$Z_i-M_i = X_i\cos\Phi - X_j\sin\Phi$$

$$Z_j-M_j = X_i\sin\Phi + X_j\cos\Phi$$

Thus,

$$P(Z_i > Z_j) = P\left[M_i + X_i \cos\Phi - X_j \sin\Phi > M_j + X_i \sin\Phi + X_j \cos\Phi\right]$$

$$P(Z_i > Z_j) = P\left[X_j < \left(\frac{\cos\Phi - \sin\Phi}{\cos\Phi + \sin\Phi}\right) X_i + \frac{M_j - M_i}{\cos\Phi + \sin\Phi}\right]$$

We assume $\sigma_i^2 = \sigma_j^2$ and obtain

$$P(Z_i > Z_j) = P\left[X_j < \frac{M_j - M_i}{\sqrt{2}}\right]$$

But X_j is gaussian with

$$E(X_j) = 0$$

$$E(X_j^2) = \sigma_i^2 \sin^2\Phi - 2\,\rho_{ij}\sigma_i\sigma_j \sin\Phi \cos\Phi + \sigma_j^2 \cos^2\Phi$$

$$= \sigma^2(1-\rho_{ij}) \quad \text{when} \quad \sigma_i^2 = \sigma_j^2 = \sigma^2$$

Thus

$$P(Z_i > Z_j) = \frac{1}{\sqrt{2\pi}\,\sigma\sqrt{1 - \rho_{ij}}} \int_{-\infty}^{\frac{M_j - M_i}{\sqrt{2}}} \exp\left[\frac{-X_j^2}{2\sigma^2(1-\rho_{ij})}\right] dX_j$$

We have shown previously that

$$\sigma = \sqrt{\frac{N_0 S}{2T}}$$

and if the k^{th} message is transmitted, then

$$M_i = E(Z_i) = S\rho_{ki}$$

$$M_j = E(Z_j) = S\rho_{kj}$$

Thus,

$$P\left[(Z_i > Z_j)|Z_k\right] = \frac{1}{\sqrt{2\pi}\,\sigma\sqrt{1-\rho_{ij}}} \int_{-\infty}^{\frac{S}{\sqrt{2}}(\rho_{ki}-\rho_{kj})} \exp \frac{-X_j^2}{2\sigma^2(1-\rho_{ij})}\, dX_j$$

We can let

$$Z = \frac{X_j}{\sigma\sqrt{1-\rho_{ij}}}$$

then we obtain

$$P\left[(Z_i > Z_j)|Z_k\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{\frac{ST}{N_o}}\left(\frac{\rho_{ki}-\rho_{kj}}{\sqrt{1-\rho_{ij}}}\right)} \exp \frac{-Z^2}{2}\, dZ$$

$$= \operatorname{erf}\left[\sqrt{\frac{ST}{N_o}}\left(\frac{\rho_{ki}-\rho_{kj}}{\sqrt{1-\rho_{ij}}}\right)\right]$$

Note that the function $P\left[(Z_i > Z_j)|Z_k\right]$ is the probability that the i^{th} command will be chosen rather than the j^{th} command in a situation in which the k^{th} command is actually the one which was transmitted. A consideration of the factor

$$\frac{\rho_{ki}-\rho_{kj}}{\sqrt{1-\rho_{ij}}}$$

appearing in the upper limit of the integral shows that the i^{th} (erroneous) command will be selected rather than the j^{th} if its correlation with the true command is greater than the correlation of the j^{th} command with the true command.

From the integral above, we easily obtain the probability of correctly decoding the received message. Letting i = k and using the fact that $\rho_{kk} = 1$, we obtain

$$P\left[(Z_k > Z_j)|Z_k\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{\frac{ST}{N_0}(1-\rho_{kj})}} \exp - \frac{Z^2}{2}\, dZ$$

If Z_j represents the message which has maximum correlation with the transmitted message (i.e., $\rho_{kj} = \rho_k(\max)$), then the probability of correctly decoding will be given by the probability that the output of the k^{th} correlator exceeds the output of the correlator for this maximally correlated command. Thus,

$$P_c(Z_k) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\sqrt{\frac{ST}{N_0}\left(1-\rho_k(\max)\right)}} \exp \frac{-Z^2}{2}\, dZ$$

We next will obtain an upper bound for the probability of an error for the usual case of low error probability. We make use of the inequality

$$\frac{1}{\sqrt{2\pi}} \int_{X}^{\infty} \exp - \frac{t^2}{2}\, dt \le \frac{1}{\sqrt{2\pi}X} \exp - \frac{X^2}{2} \text{ for } X>>0$$

(In Figure 6 we have plotted the percentage error incurred by using the right hand side as an approximation to the integral. We see that for $X>4$ the error is less than three percent.)

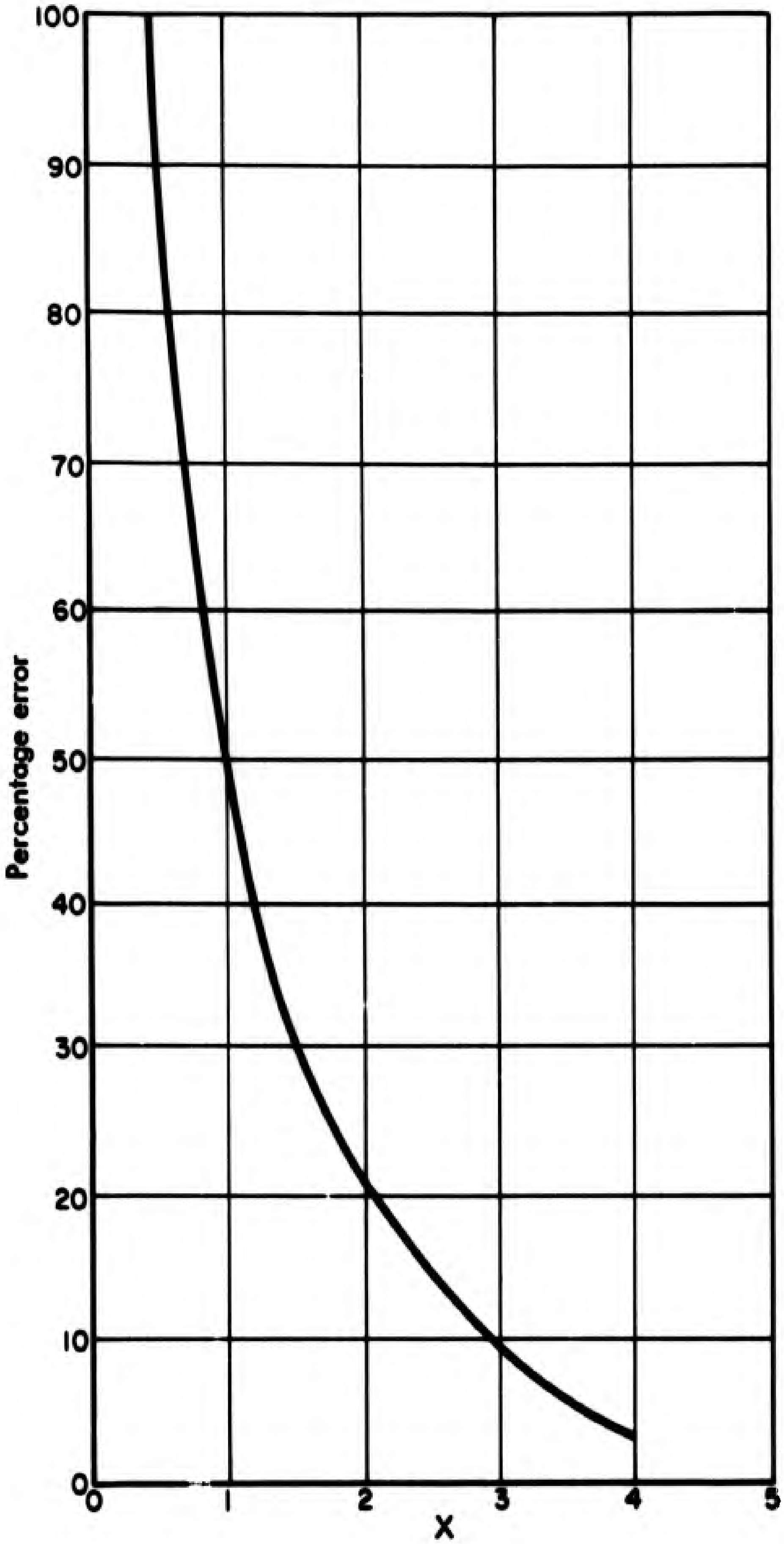

FIGURE 6 Approximation Error

R11-19-68-6

The probability of selecting the i^{th} command when the k^{th} command was transmitted is given by

$$P\left[(Z_i > Z_k)|Z_k\right] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-\sqrt{\frac{ST}{N_o}}(1-\rho_{ik})} \exp -\frac{Z^2}{2}\, dZ$$

$$= \frac{1}{\sqrt{2\pi}} \int_{\sqrt{\frac{ST}{N_o}}(1-\rho_{ik})}^{\infty} \exp -\frac{Z^2}{2}\, dZ$$

$$< \frac{1}{\sqrt{2\pi}\sqrt{\frac{ST}{N_o}(1-\rho_{ik})}} \exp -\frac{ST}{2N_o}(1-\rho_{ik})$$

Letting

$$A = \frac{ST}{2N_o}(1-\rho_{ik})$$

we obtain

$$P\left[(Z_i > Z_k)|Z_k\right] < \frac{1}{\sqrt{4\pi A}} e^{-A}$$

It is interesting to compare one of our results with a similar one obtained by Viterbi. He shows that the error probability for the class of codes with constant correlation coefficients ($\rho_{ij} = \rho$ for all $i \neq j$) is related to the error probability for the orthogonal code ($\rho = 0$) by

$$P_E\left[\frac{ST}{N_o}, \rho\right] = P_E\left[\frac{ST}{N_o}(1-\rho), 0\right]$$

That is, the effect of correlation is to reduce the effective energy of the signal. The above is a special case of the result which we obtained. We have shown that

$$P\left[(Z_i > Z_j) | Z_k\right] = \operatorname{erf}\left[\sqrt{\frac{ST}{N_0}}\left(\frac{\rho_{ki} - \rho_{kj}}{1 - \rho_{ij}}\right)\right]$$

from which we obtained

$$P\left[(Z_i > Z_j) | Z_k\right] = \operatorname{erf}\left[-\sqrt{\frac{ST}{N_0}(1 - \rho_{ik})}\right]$$

Thus, for codes with non-constant correlation between messages we can consider the effective energy of the i^{th} command with respect to the k^{th} command as being equal to the true energy times the factor $(1-\rho_{ik})$. Viterbi's result is obtained by letting $\rho_{ik} = \rho$.

Fano[1] considers the problem of correlation decoding of orthogonal signals in white, gaussian noise but obtains upper bounds on the probability of error instead of calculating it as Viterbi does. We next compare Fano's bound with Viterbi's calculations and show that the bound obtained is rather poor and leads to an erroneous conclusion. An upper bound to the probability of error is given by equations 6.88 to 6.92 in Fano.

---

[1]R. M. Fano, Transmission of Information (Cambridge: The M.I.T. Press, 1961), Chapter 6.

For the case when

$$\frac{ST}{N_0 \log_2 M} \geq 3$$

we obtain

$$P(e) < KMe^{-\frac{ST}{2N_0}}$$

where

$$K = \frac{1}{\sqrt{4\pi \frac{ST}{N_0}}}\left[\frac{\sqrt{2}}{\sqrt{1-\left(\frac{2}{\log_2 e}\right)\frac{N_0 \log_2 M}{ST}}} + \frac{1}{\sqrt{2} - \sqrt{1-\left(\frac{2}{\log_2 e}\right)\frac{N_0 \log_2 M}{ST}}}\right]$$

Fano finds that for

$$\frac{ST}{N_0 \log_2 M} \approx 3$$

and

$$M = 1024$$

that $Pe < 2 \times 10^{-4}$

Viterbi's calculations show that for these parameters

$$Pe \approx 10^{-5}$$

and that $Pe = 2 \times 10^{-4}$ is achieved with

$$\frac{ST}{N_0 \log_2 M} = 3 \quad \text{and} \quad M = 64$$

As a result of this example, Fano concludes that

> "the number of messages M must be rather large (1024) to permit efficient transmission with a reasonably low probability of error."

We have illustrated that using the exact figures rather than the bound reduces the number of messages by a factor of 16, to M = 64.

BIBLIOGRAPHY

1. Abramson, N.: Information Theory and Coding, New York: McGraw-Hill Book Company, Inc., 1963.

2. Bell, D.A.: Information Theory and Its Engineering Applications, London: Sir Isaac Pitman and Sons, Inc., 1953.

3. Bongard, M.M.: The Concept of Useful Information, Problems of Cybernetics, Vol. 9, 1963. Translation Available, Office of Technical Services: 63-31941, F.P.R.S.: 21,448.

4. Brillouin, L.: Science and Information Theory, New York: Academic Press, 1962.

5. Cherry, C.: On Human Communication: A Review, A Survey, and A Criticism, The Technology Press of Massachusetts Institute of Technology and John Wiley and Sons, Inc., New York, 1957.

6. Cramer, H.: Mathematical Methods of Statistics, Princeton: Princeton University Press, 1946.

7. Davenport, W.B. and Rust, W.L.: An Introduction to the Theory of Random Signals and Noise, New York: McGraw-Hill Book Company, Inc., 1958.

8. Fano, R.M.: Transmission of Information, Cambridge, Mass.: The M.I.T. Press, 1961.

9. Fisher, R.A.: The Design of Experiments, Sixth Edition, London: Oliver and Boyd, 1951.

10. Golomb, S.W., et, al.: Digital Communications with Space Applications, New Jersey: Prentice-Hall, Inc., 1964.

11. Harman, W.W.: Principles of the Statistical Theory of Communication, New York: McGraw-Hill Book Company, Inc., 1963.

12. Jackson, W., ed.: Communication Theory, London: Butterworths Scientific Publications, 1953.

13. Jaynes, E.T." "Information Theory and Statistical Mechanics," Physical Review, 106:4, May 15, 1957 and 108:2, Oct 15, 1957.

14. Jaynes, E.T.: Note on Unique Decipherability, I.R.E. Transactions on Information Theory, IT-5:3, Sept 1959.

15. Khinchin, A.I.: Mathematical Foundations of Information Theory, New York: Dover Publications, Inc., 1957.

16. Kullback, S.: Information Theory and Statistics, New York: John Wiley and Sons, Inc., 1959.

17. Lindsay, R.B.: Concepts and Methods of Theoretical Physics, New York: D. Van Nostrand Company, Inc., 1951.

18. Machol, R.E., ed.: Information and Decision Processes, New York: McGraw-Hill Book Company, Inc., 1960.

19. Machol, R.E., and Gray, P., eds.: Recent Developments in Information and Decision Processes, New York: Macmillan Company, 1962.

20. Mathison, R.P.: Mariner Mars 1964 Telemetry and Command System, T.R. 32-684, Jet Propulsion Laboratory, June 1, 1965.

21. Mayer, J.E. and Mayer M.G.: Statistical Mechanics, Second Edition, New York: John Wiley and Sons, Inc., 1940.

22. Shannon, C.E. and Weaver, W.: The Mathematical Theory of Communication, Urbana: The University of Illinois Press, 1949.

23. Sommerfeld, A.: Thermodynamics and Statistical Mechanics, New York: Academic Press Inc., 1956.

24. Springett, J.C.: Telemetry and Command Techniques for Planetary Spacecraft, T.R. 32-495, Jet Propulsion Laboratory, Jan 15, 1965.

25. Tolman, R.C.: The Principles of Statistical Mechanics, First Edition, Oxford University Press, 1938.

26. Vinuze, I.: An Interpretation of the I-Divergence of Information Theory, Transactions of the Second Prague Conference, Pub. House of the Czechoslovak Acad. of Sciences , Prague, 1960.

27. Wiener, N.: Cybernetics, Second Edition, New York: The M.I.T. Press and John Wiley and Sons, Inc., 1961.

BLANK PAGE

APPENDIX

THE ENTROPY CONCEPT IN STATISTICAL MECHANICS

We have discussed previously Shannon's justification of the importance of functions of the form $-\sum_i P_i \log P_i$ and of their subsequent use as the measure of information content. We will next consider the basis and background for the selection of the word "entropy" to identify such functions in information theory. In so doing we hope to provide some insight as to the nature and extent of the relationship between "information entropy" and "thermodynamic entropy."

First, we will quote the only consideration or justification of the use of the word "entropy" which was given by Shannon in "The Mathematical Theory of Communication:"

> "Quantities of the form $H = -\Sigma P_i \log P_i$ play a central role in information theory as measures of information, choice and uncertainty. The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics[1] where P_i is the probability of a system being in cell i of its phase space. H is, then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\Sigma P_i \log P_i$ the entropy of the set of probabilities $P_1, \cdots P_n$."

We next refer to Tolman [1] and Sommerfeld [2] for a better understanding of the entropy concept.

---

1 Shannon references R. C. Tolman, Principles of Statistical Mechanics (London: Oxford University Press, 1938).

2 A. Sommerfeld, Thermodynamics and Statistical Mechanics (New York: Academic Press, 1956).

The techniques of statistical mechanics were originally developed in order to provide a description of a mechanical (i.e., molecular) system consisting of such a large number of individual particles as to make it impractical to attempt a complete physical specification for each and every particle. Instead, we attempt to determine certain macroscopic thermodynamic variables in terms of statistical properties of the totality of particles of a particular system or in terms of the statistical properties of an ensemble of (similar) systems. In this context an ensemble of systems is considered to be a collection of similarly structured systems which at any particular instant of time may be distributed over a range of permissible states.

It is convenient to consider the following geometric concept as a means of conceptually representing the mechanical variables of a particular system. Each particle of a system can be assumed to possess r degrees of freedom. A point mass will generally have 3 position degrees of freedom to yield an $r = 3$; a general particle can, in addition, possess many other (internal) degrees of freedom. If our system consists of n identical particles, we will have a total of $f = nr$ degrees of freedom. The space of 2 f dimensions consisting of f position coordinate axes and f momentum axes is designated a "phase space." The instantaneous state of a system is thus (conceptually) represented by a single representative point or "phase point" in phase space. It is sometimes convenient to call a phase space for an entire system, as described above, a γ-space in order to differentiate it from a phase space for an individual molecule--called a μ-space. A point in μ-space completely specifies the mechanical state of the corresponding particle; in a system consisting of n identical particles it is convenient to construct n identical μ-spaces.

We can subdivide either the γ or the μ-space into finite volume elements, written as δv_γ or δv_μ, and consider each volume element as an indexed cell in the γ or μ-space. An approximate description of the system being considered (either the entire system or a given particle depending on whether we are in a γ or μ-space) is given by specifying the number of the cell which the system is in as a function of time.

With the above as background, we can proceed to a consideration of the manner in which entropy is defined. We will approach this definition by two distinct routes. The first being the Gibbs' formulation of statistical mechanics which Shannon apparently chose as the basis for his formulation; the second being Boltzmann's conceptually simpler method and its direct consideration of entropy.

Gibbs considers an ensemble of systems each represented by a phase point in its γ-space. The state of the ensemble itself is specified by giving the density of the distribution of phase points. Let F_k be the (normalized) density function which gives the number of system phase points, N_k, in the k^{th} region of the γ-space. If N denotes the total number of systems in the ensenbles we can write

$$N_k = NF_k \ \delta v_\gamma$$

(Tolman carefully points outs that F_κ is a "course grained density" corresponding to the small but finite distinguishable regions, δv_γ, but this distinction is inconsequential to us here.)

In terms of the above formulation by Gibbs the quantity H is defined as

$$H = \Sigma(F_\kappa \log F_\kappa)\ \delta v_\gamma$$

from which we easily obtain

$$H = \Sigma \frac{N_\kappa}{N} \log \frac{N_\kappa}{N} - \log \delta v_\gamma.$$

If we consider that

$$P_\kappa = \frac{N_\kappa}{N}$$

represents the probability that a member of the ensemble is found in the k^{th} volume element, then we have

$$H = \Sigma P_\kappa \log P_\kappa + \text{constant}.$$

Apparently, it was the recognition of this functional form which led Shannon to the definition of

$$H = -\Sigma P_i \log P_i$$

as "the entropy of the set of probabilities $P_i, \cdots P_n$." However, although Shannon's entropy function is of the same form as the H in Boltzmann's H theorem the link between H and the entropy of thermodynamics must be established. This link is established by relating entropy to the value of H obtained for equilibrium conditions of the ensemble under consideration. Thus, if we consider that the phase point for each member of our ensemble is a function of time we can also consider that the density function, F_κ, may be a function of time. If the ensemble consists of isolated systems it is reasonable to expect that a steady state or equilibrium condition will arise in that the distribution of systems of our ensemble in phase space will reach an equilibrium. For this particular situation we represent the thermodynamic entropy, S, by

$$S = -kH$$

where

k is Boltzmann's Constant

and H specifically refers to an ensemble which is in equilibrium. This approach is in keeping with the thermodynamic definition of entropy which considers changes in thermodynamic systems as occurring "through a sequence of states of equilibrium."[1]

In some formulations of statistical mechanics the entropy is defined directly in terms of H without reference to the requirement of equilibrium. The compatibility of this approach with that requiring equilibrium hinges on the fact that for typical thermodynamic systems the numercial value of H is relatively insensitive to whether or not we use the equilibrium ensemble. It is apparent that Shannon, by relating his H to entropy, chose to use the formulation which does not include the equilibrium requirement in the definition of entropy. In the following we will review the Boltzmann approach to the definition of entropy and we will consider in more detail the equilibrium requirement as a possible alternative formulation for the entropy function of information theory.

1 A. Sommerfeld, Thermodynamics and Statistical Mechanics (New York: Academic Press, Inc., 1956), p. 26.

We will begin by distinguishing between the "state" of a system and the "condition" of a system. A state is specified by giving a point or cell in a phase space which has a distinct coordinate axis corresponding to each degree of freedom for each particle. However, insofar as the behavior of the system is concerned, it is unnecessary to distinguish which specific particle of a group of identical particles is in a particular cell of the μ-space. That is, we cannot distinguish between permutations of the same type of particle. The "condition" of a system is specified by merely giving the number of particles, n_i, which are in cell i of the μ-space. (We have implicitly assumed the sub-division of the identical μ-space into identical volume elements.) Although an interchange of two particles which are in different cells of their respective μ-spaces causes a change in the "state" of the system, it does not cause a change in the "condition." To clarify the above, consider that we have divided each μ-space into volume elements or cells and in addition we have chosen to divide the γ-space into cells such that a single γ-cell corresponds to a collection of μ-cells. To denote the volume element which designates the j^{th} cell of the μ-space, representing the i^{th} particle, we can write $\left(\delta v_{\mu_j}\right)_i$. The K^{th} cell of the γ-space can be written as

$$\delta v_{\gamma_K} = \left(\delta v_{\mu_j}\right)_1 \left(\delta v_{\mu_k}\right)_2 \left(\delta v_{\mu_1}\right)_3 \cdots \left(\delta v_{\mu_s}\right)_n$$

The following different (L^{th}) volume element in the γ-space

$$\delta v_{\gamma_L} = \left(\delta v_{\mu_j}\right)_i \left(\delta v_{\mu_1}\right)_2 \left(\delta v_{\mu_k}\right)_3 \cdots \left(\delta v_{\mu_s}\right)_n$$

is obtained by merely interchanging identical particles 2 and 3. Although the external behavior cannot differ in both of these cases, we require two points in the γ-space to represent these two cases.

For illustrative purposes consider two identical particles each with a one dimensional phase space. (This is clearly impossible; the dimensionality of a phase space is always an even number since it is equal to twice the number of degress of freedom of the particle). The μ-space for each particle is a line and the instantaneous state of each particle is represented as a point on the line. (See Fig. A-1.) The γ-space for this system of two particles can be represented by the two dimensional space formed by using each line of the two μ-spaces as a coordinate axis for the γ-space. If $\mu_1 = a$ (i.e., particle 1 is in cell a) and $\mu_2 = b$ then the cell (a,b) in the γ-space represents the state of the two parti le system. On the other hand, if $\mu_1 = b$ and $\mu_2 = a$ then a different cell, (b,a) represents the state of the two particles, even though in both cases the external behavior of the systems are identical. However, the "condition" of each system is specified by;

one particle in cell a

one particle in cell b

Thus, the statement of "condition," although not as complete as the "state" representation, is an adequate and more compact representation of external behavior.

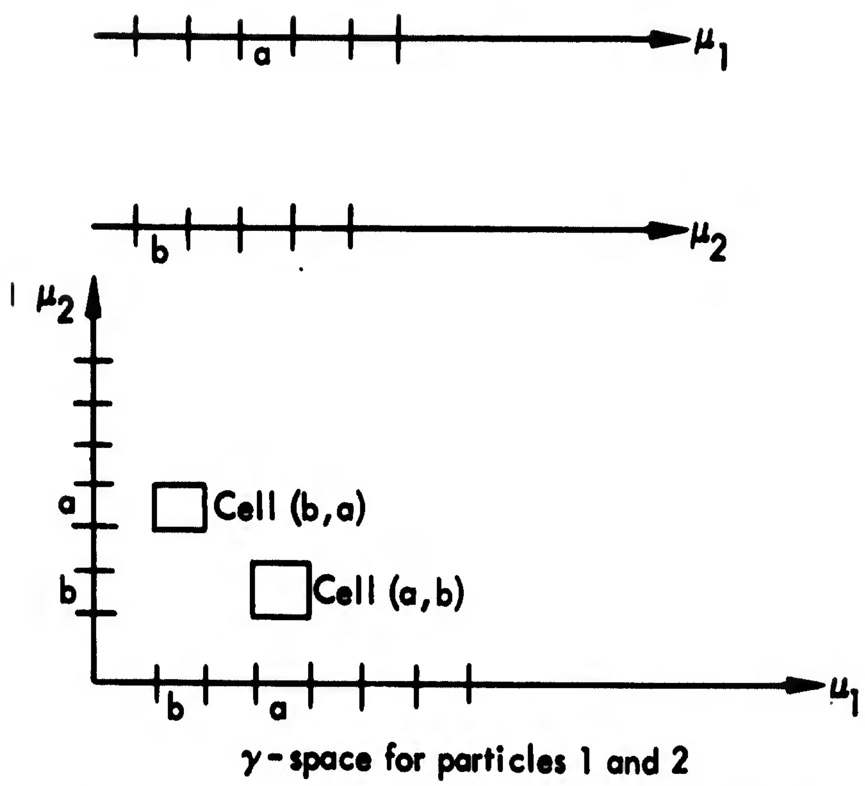

FIGURE A-1 Instantaneous State and γ-Space for Particles

If at a particular instant no two particles are in corresponding cells of their μ-space (i.e., $j \neq k \neq l \cdots \neq s$) then we have defined a specific δv_{γ_M}. However, from an external point of view, we could just as well have any one of the n particles in cell δv_{μ_j}, any one of the remaining n-1 particles in cell δv_{μ_K}, etc. Thus, in this case it requires n! different γ-space cells to represent one particular external behavior of the system. If at any instant of time n_1 of the n particles are in the corresponding cells δv_{μ_s} of their respective μ-spaces, and all other particles are in different cells, then we can have the same external behavior if we have any one of n particles in δv_{μ_j} , any one of n-1 particles in δv_{μ_k} $\cdots$, any one of $n-n_1+1$ in δv_{μ_r} with the n_1 remaining particles in δv_{μ_s}. Thus, there are $(n)(n-1)(n-2)\cdots(n-n_1+1)(1)$ different γ-space cells corresponding to this situation. But

$$(n)(n-1)(n-2)(n-n_1+1) = \frac{n!}{n_1!}$$

If in addition we have n_2 particles in the same δv_{μ_r}, then since we now have initially only a number $\frac{n!}{n_1!}$ γ-space cells we obtain similarly

$$\frac{n!/n_1!}{n_2!} = \frac{n!}{n_1!\; n_2!}$$

different γ-space cells for the same behavior. The "conditon" of the system is specified by giving the number of particles n_1, n_2, $\cdots$ n_i, $\cdots$ n_t in each cell, i, of the μ-space. Thus, it requires

$$W` = \frac{n!}{n_1!\;\; n_2!\;\; n_3!\;\;\;\; n_1!\;\;\;\; n!_t}$$

different volume elements of the γ-space to describe a single "condition" with n_i particles in δv_{μ_i} $(1 \leq i \geq t)$. Equivalently, this represents the number of different ways that the same external behavior can be realized; Sommerfeld calls W` the "thermodynamic probability or weight" of the state.

If there exists an a-priori probability distribution which specifies the probability p_i, that a particle will be in the i^{th} cell, then the a-priori probability of having n particles in cell 1,

n_2 particles in cell 2_a, etc., is given by

$$P_1^{n_1} P_2^{n_2} \cdots P_t^{n_t} = \prod_{i=i}^{t} P_i^{n_i} .$$

The function W is then modified[1] and becomes

$$W = \frac{n!}{\prod_{i=i}^{t} n!} \prod_{i=i}^{t} P_i^{n_i}$$

In accordance with Sommerfeld's interpretation of the Boltzmann entropy concept, "Boltzmann's Principle" defines the thermodynamic entropy, S, directly in terms of what we have called the "condition" or "thermodynamic probability or weight" of a state. Thus we have

$$S = k \log W$$

where W is understood to represent the equilibrium or "most probable" condition (and k is Boltzmann's constant)

i.e. $S = k \log W_{max}$

In addition, we consider that the total energy of the n_i particles in the i^{th} cell is E_i and assume that both the total number of particles and the total energy, U, of the system are prescribed.

$$U = \sum_i n_i E_i \qquad\qquad n = \sum_i n_i$$

1 R. B. Lindsay, Concepts and Methods of Theoretical Physics (New York: D. Van Nostrand Co., 1951).

The maximization procedure (which we will consider elsewhere) is straightforward, and results in

$$S = K \log W_{max} = k\, n \log \sum_i P_i e^{-E_i/kT} + \frac{U}{kT}$$

T is introduced as a Lagrange multiplier and is shown to be the thermodynamic temperature.

Thus, if we relate entropy to the logarithm of the maximum of the thermodynamic probability and impose two conservation requirements (i.e., on U and n), we arrive at an expression for entropy which is far from the form which Shannon chose to model the entropy of an information source.

It is interesting to note Brillouin's derivation of Shannon's entropy concept. If we assume that $S = k \log W^*$ without considering the requirement that W^* is understood to represent an equilibrium or "most probable" condition, then we obtain

$$S = k \log \frac{n!}{\prod_{i=i} n_i!}$$

which by Sterling's approximation reduces to

$$S = k\,(n \log n - \sum_i n_i \log n_i)$$

Thus,

$$S = k \sum n_i (\log n - \log n_i)$$

$$S = -k\, n \sum_i \frac{n_i}{n} \log \frac{n_i}{n}$$

Letting

$$k\,n = K$$

$$n_i/n = P_i$$

we obtain

$$S = -K \sum_i P_i \log P_i$$

which is precisely the form used by Shannon in the definition of entropy. However, we make the following observations with respect to this function. First, as explained by Sommerfeld, the entropy is defined in terms of the logarithm of the most probable condition, W max, not the condition W. Second, $p_i = n_i/n$ is not "the probability of a system being in cell i of its phase space." It can be considered an estimate of the probability that a randomly chosen particle of the system will be in cell i of its μ-space. Thirdly, we cannot ignore the fact that the constant K is not independent of the system under consideration, since K = kn. Thus, for example, it is well known in information theory that the probability function which maximizes S is a uniform one; i.e., $P_i = \frac{1}{n}$. Therefore, S_{max} is usually obtained as

$$S_{max} = -K \sum_i \frac{1}{n} \log \frac{1}{n} = \log n$$

Thus,

$$S_{max} = k\,n \log n$$

If we desire, we can define k = 1 and obtain

$$S_{max} = n \log n$$

This form will yield different results than those usually obtained in the application of the Shannon theory.

In conclusion, we have shown several alternative approaches to the derivation of the thermodynamic entropy function. For the development of a theory of information which is based exclusively on a probabilistic basis, it is clear that the Gibbs formulation with its single variable, P_i, is the logical starting point for the development of an analogy. However, if we wish to include other fundamental parameters in the formulation of a theory of information,then the Gibbs formulation is inadequate if only for the reason that it does not provide us with enough variables for our needs. In order to retain, and even expand, the relationship between information theory and statistical mechanics which arises as a result of the entropy analogy, we will develop a model for a command transmission which can utilize the entropy concept as formulated in Boltzmann's Principle.