A MATHEMATICAL THEORY OF COMMUNICATION

by
C. E. Shannon
Bell Telephone Laboratories, Inc.

I. The Discrete Case

Introduction

The recent development of various methods of modulation such as PCM and PPM which exchange bandwidth for signal to noise ratio has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist[1] and Hartley[2] on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. The significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

---

[1] Nyquist, H., "Certain Factors Affecting Telegraph Speed", Bell System Technical Journal, April 1924, p.324.
"Certain Topics in Telegraph Transmission Theory", A.I.E.E. Trans., v.47, April 1948, p.617.

[2] Hartley, R.V.L., "Transmission of Information", Bell System Technical Journal, July 1928, p.535.

If the number of messages in the set is finite then this number of any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons.

1. It is practically more useful. Parameters of engineering importance such as time, bandwidth, number of relays, etc., tend to vary linearly with the logarithm of the number of possibilities. For example, adding one relay to a group doubles the number of possible states of the relays. It adds 1 to the base two logarithm of this number. Doubling the time roughly squares the number of possible messages, or doubles the logarithm, etc.
2. It is nearer to our intuitive feeling as to the proper measure. This is closely related to 1 since we intuitively measure entities by linear comparison with common standards. One feels, for example, that two punched cards should have twice the capacity of one for information storage, and two identical channels twice the capacity of one for transmitting information.

3. It is mathematically more suitable. Many of the limiting operations are simple in terms of the logarithm but would require clumsy restatement in terms of the number of possibilities.

The choice of a logarithmic base corresponds to the choice of a unit for measuring information. If the base two is used the resulting units may be called binary digits, or more briefly bits a word suggested by J. W. Tukey. A device with two stable positions such as a relay or a flip-flop circuit can store one bit of information. N such devices can store N bits, since the total number of possible states is 2^N and $\log_2 2^N = N$. If the base 10 is used the units may be called decimal digits. Since

$$\log_2 M = \log_{10} M \quad \log_2 10$$
$$= 3.32 \quad \log_{10} M,$$

a decimal digit is about 3-1/3 bits. A digit wheel on a desk computing machine has ten stable positions and therefore has a storage capacity of one decimal digit. In analytical work where integration and differentiation are involved the base e is sometimes useful. The resulting units of information will be called natural units. Change from the base a to base b merely requires multiplication by $\log_b a$.

By a communication system we will mean a system of the type indicated schematically in Fig. 1. It consists of essentially five parts:

1. An information source which produces a message or sequence of messages to be communicated to the receiving terminal. The message may be of various types, e.g. (a) A sequence of letters as in a telegraph or teletype system. (b) A single function of time f(t) as in radio or telephony. (c) A function of time and other variables as in black and white television. Here the message may be thought of as a function f(x,y,t) of two space coordinates and time, the light intensity at point (x,y) and time t on a pickup tube plate. (d) Two or more functions of time, say f(t), g(t), h(t). This is the case in "three dimensional" sound transmission or if the system is intended to service several individual channels in multiplex. (e) Several functions of several variables. In color television the message consists of three functions f(x,y,t), g(x,y,t), h(x,y,t) defined in a three dimensional continuum. We may also think of these three functions as components of a vector field defined in the region. Similarly several black and white television sources would produce "messages" consisting of a number of functions of three variables. (f) Various combinations also occur, for example in television with an associated audio channel.

2. A transmitter which operates on the message in some way to produce a signal suitable for transmission over the channel. In telephony this operation consists merely of changing sound pressure into a proportional electrical current. In telegraphy we

have an encoding operation which produces a sequence of dots, dashes and spaces on the channel corresponding to the message. In a multiplex PCM system the different speech functions must be sampled, compressed, quantized and encoded, and finally interleaved properly to construct the signal. Vocoder systems, television, and frequency modulation are other examples of complex operations applied to the message to obtain the signal.

3. The channel is merely the medium used to transmit the signal from transmitter to receiver. It may be a pair of wires, a coaxial cable, a band of radio frequencies, a beam of light, etc.

4. The receiver ordinarily performs the inverse operation of that done by the transmitter, reconstructing the message from the signal.

5. The destination is the person (or thing) for whom the message is intended.

We wish to consider certain general problems involving communication systems. To do this it is first necessary to represent the various elements involved as mathematical entities, suitably idealized from their physical counterparts. We may roughly classify communication systems into three main categories, discrete, continuous and mixed. By a discrete system we will mean one in which both the message and the signal are a sequence of discrete symbols. A typical case is telegraphy where the message is a sequence of letters and the signal a sequence of dots, dashes and spaces. A continuous system is one in which

the message and signal are both treated as continuous functions e.g. radio or television. A mixed system is one in which both discrete and continuous variables appear, e.g., PCM transmission of speech.

We first consider the discrete case. This case has applications not only in communication theory, but also in cryptography, the theory of computing machines, the design of telephone exchanges and other fields.

PART I: DISCRETE NOISELESS SYSTEMS

1. The Discrete Noiseless Channel

Teletype and telegraphy are two simple examples of a discrete channel for transmitting information. Generally, a discrete channel will mean a system whereby a sequence of choices from a finite set of elementary symbols $S_1,\ldots,S_n$ can be transmitted from one point to another. These symbols S_i are assumed to each have a certain duration in time t_i seconds (not necessarily the same for different S_i, for example the dots and dashes in telegraphy). It is not required that all possible sequences of the S_i be capable of transmission on the system; certain sequences only may be allowed. These will be possible signals for the channel. Thus in telegraphy suppose the symbols are: (1) A dot, consisting of line closure for a unit of time and then line open for a unit of time. (2) A dash, consisting of three time units of closure and one unit open. (3) A letter space consisting of three units of line open. (4) A word space of six units of line open. We might place the restriction on allowable sequences that no spaces follow each other (for if two letter spaces are adjacent, it is identical with a word space). The question we now consider is how one can measure the capacity of such a channel to transmit information.

In the teletype case where all symbols are of the same duration, and any sequence of the 32 symbols is allowed, the answer is easy. Each symbol represents 5 bits of information. If the system transmits n symbols per second it is natural to say that

the channel has a capacity of 5n bits per second. This does not mean that the teletype channel will always be transmitting information at this rate---this is the maximum possible rate and whether or not the actual rate reaches this maximum depends on the source of information which feeds the channel as will appear later.

In the more general case with different lengths of symbols and constraints on the allowed sequences, we make the following definition:

Definition: The capacity C of a discrete channel is given by

$$C = \lim_{T \to \infty} \frac{\log N(T)}{T}$$

where N(T) is the number of allowed signals of duration T.

It is easily seen that in the teletype case this reduces to the previous result. It can be shown that the limit in question will exist as a finite number in most cases of interest. Suppose all sequences of the symbols $S_1 \ldots S_n$ are allowed and these symbols have durations $t_1 \ldots t_n$. What is the channel capacity? If N(t) represents the number of sequences of duration t we have

$$N(t) = N(t-t_1) + N(t-t_2) + \ldots + N(t-t_n)$$

The total number is equal to the sum of the numbers of sequences ending in S_1, S_2, ..., S_n and these are $N(t-t_1)$, $N(t-t_2)$, ..., $N(t-t_n)$, respectively. According to a well known result in finite differences, N(t) is then asymptotic for large t to X_o^t where X_o is the largest real solution of the characteristic equation:

$$X^{-t_1} + X^{-t_2} + \ldots + X^{-t_n} = 1$$

and therefore

$$C = \frac{\log X_o^T}{T} = \log X_o$$

In case there are restrictions on allowed sequences we may still often obtain a difference equation of this type and find C from the characteristic equation. In the telegraphy case mentioned above

$$N(t) = N(t-2) + N(t-4) + N(t-5) + N(t-7) + N(t-8) + N(t-10)$$

as we see by counting sequences of symbols according to the last or next to the last symbol occurring. Hence C is $-\log \mu_o$ where μ_o is the positive root of $1 = \mu^2 + \mu^4 + \mu^5 + \mu^7 + \mu^8 + \mu^{10}$.

A very general type of restriction which may be placed on allowed sequences is the following. We imagine a number of possible states $a_1, a_2 \ldots a_m$. For each state only certain symbols from the set $S_1 \ldots S_n$ can be transmitted (different subsets for the different states). When one of these has been transmitted the state changes to a new state depending both on the old state and the particular symbol transmitted. The telegraph case is a simple example of this. There are two states depending on whether a space was the last symbol transmitted or not. If so then only a dot or a dash can be sent next and the state always changes. If not, any symbol can be transmitted and the state changes if a space is sent, otherwise remaining the same. The conditions can be indicated in a linear graph as shown in Fig. 2. The junction

points correspond to the states and the lines indicate the symbols possible in a state and the resulting state. In appendix 1 it is shown that if the conditions on allowed sequences can be described in this form C will exist and can be calculated in accordance with the following result.

Theorem 1: Let $b_{ij}^{(s)}$ be the duration of the s^{th} symbol which is allowable in state i and leads to state j. Then the channel capacity C is equal to log W where W is the largest real root of the determinant equation:

$$\left|\sum_s W^{-b_{ij}^{(s)}} - \delta_{ij}\right| = 0$$

where $\delta_{ij} = 1$ if $i = j$ and is zero otherwise.

2. The Discrete Source of Information

We have seen that under very general conditions the logarithm of the number of possible signals in a discrete channel increases linearly with time. The capacity to transmit information can be specified by giving this rate of increase, the number of bits per second required to specify the particular signal used.

We now consider the information source. How is an information source to be described mathematically, and how much information in bits per second is produced in a given source? The main point at issue is the effect of statistical knowledge about the source in reducing the required capacity of the channel, by the use of proper encoding of the information. In telegraphy, for example, the messages to be transmitted consist of sequences

of letters. These sequences, however, are not completely random. In general, they form sentences and have the statistical structure of, say, English. The letter E occurs more frequently than Q, the sequence TH more frequently than XP, etc. The existence of this structure allows one to make a saving in time (or channel capacity) by properly encoding the message sequences into signal sequences. This is already done to a limited extent in telegraphy by using the shortest channel symbol, a dot, for the most common English letter E, while the infrequent letters, Q, X, Z are represented by longer sequences of dots and dashes. This idea is carried still further in certain commercial codes where common words and phrases are represented by four or five letter code groups with a considerable saving in average time. The standardized greeting and anniversary telegrams now in use extend this to the point of encoding a sentence or two into a relatively short sequence of numbers.

We can think of a discrete source as generating the message symbol by symbol. It will choose successive symbols according to certain probabilities depending, in general, on preceding choices as well as the particular symbols in question. A physical system, or a mathematical model of a system which produces such a sequence of symbols governed by a set of probabilities is known as a stochastic process. We may consider a discrete source, therefore, to be represented by a stochastic process. Conversely, any stochastic process which produces a discrete

sequence of symbols chosen from a finite set may be considered a discrete source. This will include such cases as:

1. Natural written languages such as English, German, Chinese.
2. Continuous information sources that have been rendered discrete by some quantizing process. For example, the quantized speech from a PCM transmitter, or a quantized television signal.
3. Mathematical cases where we merely define abstractly a stochastic process which generates a sequence of symbols. The following are examples of such sources.

   (A) Suppose we have 5 letters A, B, C, D, E which are chosen each with probability .2, successive choices being independent. This would lead to a sequence of which the following is a typical example.

   B D C B C E C C C A D C B D D A A E C E E A

   A B B D A E E C A C E E B A E E C B C E A D

   This was constructed with the use of a table of random numbers.*

   (B) Using the same 5 letters let the probabilities be .4, .1, .2, .2, .1 respectively, with successive choices independent. A typical message from this source is then:

   A A A C D C B D C E A A D A D A C E D A

   E A D C A B E D A D D C E C A A A A A D

---

*Kendall and Smith, "Tables of Random Sampling Numbers", Cambridge, 1939.

(C) A more complicated structure is obtained if successive symbols are not chosen independently but their probabilities depend on preceding letters. In the simplest case of this type a choice depends only on the preceding letter and not on ones before that. The statistical structure can then be described by a set of transition probabilities $p_i(j)$, the probability that letter i is followed by letter j. The indices i and j range over all the possible symbols. A second equivalent way of specifying the structure is to give the "digram" probabilities p(i,j), i.e., the relative frequency of the digram i j. The letter frequencies p(i), the transition probabilities $p_i(j)$ and the digram probabilities p(i,j) are related by the following formulas.

$$p(i) = \sum_j p(i,j) = \sum_j p(j,i) = \sum_j p(j)p_j(i)$$

$$p(i,j) = p(i)\, p_i(j)$$

$$\sum_j p_i(j) = \sum_i p(i) = \sum_{i,j} p(i,j) = 1$$

As a specific example suppose there are three letters A, B, C with the probability tables:

$p_i(j)$		j A	B	C
	A	0	$\frac{4}{5}$	$\frac{1}{5}$
i	B	$\frac{1}{2}$	$\frac{1}{2}$	0
	C	$\frac{1}{2}$	$\frac{2}{5}$	$\frac{1}{10}$

	p(i)
A	$\frac{9}{27}$
B	$\frac{16}{27}$
C	$\frac{2}{27}$

p(i,j)		j A	B	C
	A	0	$\frac{4}{15}$	$\frac{1}{15}$
i	B	$\frac{8}{27}$	$\frac{8}{27}$	0
	C	$\frac{1}{27}$	$\frac{4}{135}$	$\frac{1}{135}$

A typical message from this source is the following.

A B B A B A B A B A B A B A B B B A B B B B B A B

A B A B A B A B B B A C A C A B B A B B B B A B B

A B A C B B B A B A

The next increase in complexity would involve trigram frequencies but no more. The choice of a letter would depend on the preceding two letters but not on the message before that point. A set of trigram frequencies p(i,j,k) or equivalently a set of transition probabilities $p_{ij}(k)$ would be required. Continuing in this way one obtains successively more complicated stochastic processes. In the general n-gram case a set of n-gram probabilities $p(i_1, i_2, \ldots, i_n)$ or of transition probabilities $p_{i_1,i_2, \ldots, i_{n-1}}(i_n)$ is required to specify the statistical structure.

(D) Stochastic processes can also be defined which produce a text consisting of a sequence of "words". Suppose there are 5 letters A, B, C, D, E and 16 "words" in the language with associated probabilities:

.10 A	.16 BEBE	.11 CABED	.04 DEB
.04 ADEB	.04 BED	.05 CEED	.15 DEED
.05 ADEE	.02 BEED	.08 DAB	.01 EAB
.01 BADD	.05 CA	.04 DAD	.05 EE

Suppose successive "words" are chosen independently and are separated by a space. A typical message might be:

DAB EE A BEBE DEED DEB ADEE ADEE EE DEB BEBE BEBE

BEBE ADEE BED DEED DEED CEED ADEE A DEED DEED BEBE

CABED BEBE BED DAB DEED ADEB

If all the words are of finite length this process is equivalent to one of the preceding type, but the description may be simpler in terms of the word structure and probabilities. We may also generalize here and introduce transition probabilities between words, etc.

These artificial languages are useful in constructing simple problems and examples to illustrate various possibilities. We can also approximate to a natural language by means of a series of simple artificial languages. The zero order approximation is obtained by choosing all letters with the same probability and independently. The first order approximation is obtained by choosing successive letters independently but each letter having the same probability that it does in the natural language. Thus in the first order approximation to English E is chosen with probability .12 (its frequency in normal English) and W with probability .02, but there is no influence between

adjacent letters and no tendency to form the preferred digrams such as TH, ED, etc. In the second order approximation digram structure is introduced. After a letter is chosen, the next one is chosen in accordance with the frequencies with which the various letters follow the first one. This requires a table of digram frequencies $p_i(j)$. In the third order approximation trigram structure is introduced. Each letter is chosen with probabilities which depend on the preceding two letters.

3. The Series of Approximations to English

To give a visual idea of how this series of processes approaches a language, typical sequences in the approximations to English have been constructed and are given below. In all cases we have assumed a 27 symbol "alphabet", the 26 letters and a space.

1. Zero order approximation (symbols independent and equiprobable).

   XFOML RXKHRJFFJUJ ZLPWCFWKCYJ FFJEYVKCQSGXYD
   QPAAMKBZAACIBZLHJQD

2. First order approximation (symbols independent but with frequencies of English text).

   OCRO HLI RGWR NMIELWIS EU LL NBNESEBYA TH EEI ALHENHTTPA
   OOBTTVA NAH BRL

3. Second order approximation (digram structure as in English).

   ON IE ANTSOUTINYS ARE T INCTORE ST BE S DEAMY ACHIN D
   ILONASIVE TUCOOWE AT TEASONARE FUSO TIZIN ANDY TOBE
   SEACE CTISBE

4. Third order approximation (trigram structure as in English).

   IN NO IST LAT WHEY CRATICT FROURE BIRS GROCID PONDENOME OF DEMONSTURES OF THE REPTAGIN IS REGOACTIONA OF CRE

5. 1st Order Word Approximation. Rather than continue with tetragram, ..., n-gram structure it is easier and better to jump at this point to word units. Here words are chosen independently but with their appropriate probabilities.

   REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.

6. 2nd Order Word Approximation. The word transition probabilities are correct but no further structure is included.

   THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED

The resemblance to ordinary English text increases quite noticeably at each of the above steps. Note that these samples have reasonably good structure out to about twice the range that is taken into account in their construction. Thus in (3) the statistical process insures reasonable text for two-letter sequence, but four-letter sequences from the sample can

usually be fitted into good sentences. In (6) sequences of 4 or more words can easily be placed in sentences without unusual or strained constructions. The particular sequence of twn words "attack on an English writer that the character of this" is not at all unreasonable. It appears then that a sufficiently complex stochastic process will give a satisfactory representation of a discrete source.

The first two samples were constructed by the use of a book of random numbers in conjunction (for example 2) with a table of letter frequencies. This method might have been continued for (3), (4), and (5), since digram, trigram, and word frequency tables are available, but a simpler equivalent method was used. To construct (3), for example, one opens a book at random and selects a letter at random on the page. This letter is recorded. The book is then opened to another page and one reads until this letter is encountered. The succeeding letter is then recorded. Turning to another page this second letter is searched for and the succeeding letter recorded, etc. A similar process was used for (4), (5), and (6). It would be interesting if further approximations could be constructed, but the labor involved becomes enormous at the next stage.

4. Graphical Representation of a Markoff Process

Stochastic processes of the type described above are known mathematically as discrete Markoff processes and have

been extensively studied in the literature.* The general case can be described as follows. There exist a finite number of possible "states" of a system; A_1, A_2, ..., A_n. In addition there is a set of transition probabilities; $p_i(j)$ the probability that if the system is in state A_i it will next go to state A_j. To make this Markoff process into an information source we need only assume that a letter is produced for each transition from one state to another. The states will correspond to the "residue of influence" from preceding letters.

The situation can be represented graphically as shown in Figs. 3, 4 and 5. The "states" are the junction points in the graph and the probabilities and letters produced for a transition are given beside the corresponding line. Fig. 3 is for the example B in Section 2, while Fig. 4 corresponds to the example C. In Fig. 3 there is only one state since successive letters are independent. In Fig. 4 there are as many states as letters. If a trigram example were constructed there would be at most n^2 states corresponding to the possible pairs of letters preceding the one being chosen. Fig. 5 is a graph for the case of word structure in example D. Here S corresponds to the "space" symbol.

---

* For a detailed treatment see M. Frechet, "Methods des fontions arbitraires. Theorie des événements en chaine dans le cas d'un nombre fini d'états possibles." Paris, Gauthier-Villars, 1938.

5. Ergodic and Mixed Sources

As we have indicated above a discrete source for our purposes can be considered to be represented by a Markoff process. Among the possible discrete Markoff processes there is a group with special properties of significance in communication theory. This special class consists of the "ergodic" processes and we will call the corresponding sources ergodic sources. Although a rigorous definition of an ergodic process is somewhat involved, the general idea is simple. In an ergodic process every sequence produced by the process is the same in statistical properties. Thus the letter frequencies, digram frequencies, etc., obtained from particular sequences will, as the lengths of the sequences increases, approach definite limits independent of the particular sequence. Actually this is not true of every sequence but the set for which it is false has probability zero. Roughly the ergodic property means statistical homogeneity.

All the examples of artificial languages given above are ergodic. This property is related to the structure of the corresponding graph. If the graph has the following two properties the corresponding process will be ergodic.

1. The graph does not consist two isolated parts A and B such that it is impossible to go from junction points in part A to junction points in part B along lines of the graph in the direction of arrows and also impossible to go from junctions in part B to junctions in part A.

2. A closed series of lines in the graph with all arrows on the lines pointing in the same orientation will be called a "circuit". The "length" of a circuit is the number of lines in it. Thus in Fig. 5 the series BEBES is a circuit of length 4. The second property required is that the greatest common divisor of the lengths of all circuits in the graph be one.

If the first condition is satisfied but the second one violated by having the greatest common divisor equal to $d > 1$, the sequences have a certain type of periodic structure. The various sequences fall into d different classes which are statistically the same apart from a shift of the origin (i.e. which letter in the sequence is called letter 1). By a shift of from 0 up to d - 1 any sequence can be made statistically equivalent to any other. A simple example with $d = 2$ is the following. There are three possible letters a, b, c. Letter a is followed with either b or c with probabilities $\frac{1}{3}$ and $\frac{2}{3}$ respectively. Either b or c is always followed by letter a. Thus a typical sequence is

a b a c a c a c a b a c a b a b a c a c.

This type of situation is not of much importance for our work.

If the first condition is violated the graph may be separated into a set of subgraphs each of which satisfies the first condition. We will assume that the second condition is also satisfied for each subgraph. We have in this case what may

be called a "mixed" source made up of a number of pure components. The components correspond to the various subgraphs. If L_1, L_2, L_3,... are the component sources we may write

$$L = p_1 L_1 + p_2 L_2 + p_3 L_3 + \cdots$$

where p_i is the a priori probability of the component source L_i.

Physically the situation represented is this. There are several different sources L_1, L_2, L_3, ... which are each of homogeneous statistical structure (i.e., they are ergodic). We do not know a priori which is to be used, but once the sequence starts in a given pure component L_i it continues indefinitely recording to the statistical structure of that component.

As an example one may take two of the processes defined above and assume $p_1 = .2$ and $p_2 = .8$. A sequence from the mixed source

$$L = .2\, L_1 + .8\, L_2$$

would be obtained by choosing first L_1 or L_2 with probabilities .2 and .8 and after this choice generating a sequence from whichever was chosen.

Except when the contrary is stated we shall assume a source to be ergodic. This assumption enables one to identify averages along a sequence with averages over the ensemble of possible sequences (the probability of a discrepancy being zero). For example the relative frequency of the letter A in a particular infinite sequence will be, with probability one, equal to its relative frequency in the ensemble of sequences.

If P_i is the probability of state i and $p_i(j)$ the transition probability to state j, then for the process to be stationary it is clear the P_i must satisfy equilibrium conditions:

$$P_j = \sum_i P_i \, p_i(j)$$

In the ergodic case it can be shown that with any starting conditions the probabilities $P_j(N)$ of being in state j after N symbols, approach the equilibrium values as $N \rightarrow \infty$.

6. Choice Uncertainty and Entropy

We have represented a discrete information source as a Markoff process. Can we define a quantity which will measure, in some sense, how much information is "produced" by such a process, or better, at what rate information is produced?

Suppose we have a set of possible events whose probabilities of occurrence are $p_1, p_2, \ldots, p_n$. These probabilities are known but that is all we know concerning which event will occur. Can we find a measure of how much "choice" is involved in the selection of the event or of how uncertain we are of the outcome?

If there is such a measure, say $H(p_1, p_2, \ldots, p_n)$, it is perhaps reasonable to require of it the following properties:

1. H should be continuous in the p_i.
2. If all the p_i are equal, $p_i = \frac{1}{n}$, then H should be a monotonic increasing function of n. With equally likely events there is more choice, or uncertainty, when there are more possible events.
3. If a choice be broken down into two successive choices, the original H should be the weighted sum of the individual values of H. The meaning of this is illustrated in Fig. 6. At the left we have three possibilities $p_1 = \frac{1}{2}$, $p_2 = \frac{1}{3}$, $p_3 = \frac{1}{6}$. On the right we

first choose between two possibilities each with probability $\frac{1}{2}$, and if the second occurs make another choice with probabilities $\frac{2}{3}$, $\frac{1}{3}$. The final results have the same probabilities as before. We require, in this special case, that

$$H(\tfrac{1}{2}, \tfrac{1}{3}, \tfrac{1}{6}) = H(\tfrac{1}{2}, \tfrac{1}{2}) + \tfrac{1}{2} H(\tfrac{1}{3}, \tfrac{2}{3})$$

The coefficient $\frac{1}{2}$ is because this second choice only occurs half the time.

In Appendix 2, the following result is established.

Theorem 2: The only H satisfying the three above assumptions is of the form:

$$H = -K \sum_{i=1}^{n} p_i \log p_i$$

where K is a positive constant.

This theorem, and the assumptions required for its proof, are in no way necessary for the present theory. It is given chiefly to lend a certain plausibility to some of our later definitions. The real justification of these definitions, however, will reside in their implications.

Quantities of the form $H = -\Sigma p_i \log p_i$ (the constant K merely amounts to a choice of a unit of measure) 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 where p_i is the probability of a system being cell i of

its phase space. H is then, for example, the H in Boltzmann's famous H theorem. We shall call $H = -\sum p_i \log p_i$ the entropy of the set of probabilities $p_1, \ldots, p_n$. If x is a chance variable we will write H(x) for its entropy; thus x is not an argument of a function but a label for a number, to differentiate it from H(y) say, the entropy of the chance variable y.

The entropy in the case of two possibilities with probabilities p and q = 1-p, namely

$$H = -(p \log p + q \log q)$$

is plotted in Fig. 7 as a function of p.

The quantity H has a number of interesting properties which further substantiate it as a reasonable measure of choice or information.

1. H = 0 if and only if all the p_i but one are zero, this one having the value unity. Thus only when we are certain of the outcome does H vanish. Otherwise H is positive.

2. For a given n, H is a maximum and equal to log n when all the p_i are equal (i.e., $\frac{1}{n}$). This is also intuitively the most uncertain situation.

3. Suppose there are two events x and y in question with m possibilities for the first and n for the second. Let p(i,j) be the probability of the joint occurrence of i for the first and j for the second. The entropy of the joint event is

$$H(x,y) = \sum_{i,j} p(i,j) \log p(i,j)$$

while

$$H(x) = \sum_{i,j} p(i,j) \log \sum_{i} p(i,j)$$

$$H(y) = \sum_{i,j} p(i,j) \log \sum_{j} p(i,j)$$

It is easily shown that

$$H(x,y) \leq H(x) + H(y)$$

with equality only if the events are independent (i.e., $p(i,j) = p(i)\ p(j)$). The uncertainty of a joint event is less than or equal to the sum of the individual uncertainties.

4. Any change toward equalization of the probabilities $p_1, p_2, \ldots, p_n$ increases H. Thus if $p_1 < p_2$ and we increase p_1, decreasing p_2 an equal amount so that p_1 and p_2 are more nearly equal, then H increases. More generally, if we perform any "averaging" operation on the p_i of the form

$$p_i' = \sum_{j} a_{ij}\ p_j$$

where $\sum_{i} a_{ij} = \sum_{j} a_{ij} = 1$, and all $a_{ij} \geq 0$, then H increases (except in the special case where this transformation amounts to no more than a permutation of the p_j with H of course remaining the same).

5. Suppose there are two chance events x and y as in 3, not necessarily independent. For any particular value i that x can assume there is a conditional probability $p_i(j)$ that y has the value j. This is given by

$$p_i(j) = \frac{p(i,j)}{\sum_{j} p(i,j)}$$

We define the conditional entropy of y, $H_x(y)$ as the average of the entropy of y for each value of x, weighted according to the probability of getting that particular x. That is,

$$H_x(y) = -\sum_{i,j} p(i,j) \log p_i(j)$$

This quantity measures how uncertain we are of y on the average when we know x. Substituting the value of $p_i(j)$ we obtain

$$H_x(y) = -\sum_{i,j} p(i,j) \log p(i,j) + \sum_{i,j} p(ij) \log \sum_{j} p(i,j)$$

$$= H(x,y) - H(x)$$

or

$$H(x,y) = H(x) + H_x(y)$$

The uncertainty of the joint event x,y is the uncertainty of x plus the uncertainty of y knowing x.

6. From 3 and 5 we have

$$H(x) + H(y) \geq H(x,y) = H(x) + H_x(y)$$

Hence

$$H(y) \geq H_x(y)$$

The uncertainty of y is never increased by knowledge of x. It will be decreased unless x and y are independent events, in which case it is not changed.

7. The Entropy of an Information Source

Consider a discrete source of the finite state type considered above. For each possible state i there will be a set of probabilities $p_i(j)$ of producing the various possible symbols j. Thus there is an entropy H_i for each state. The

entropy of the source will be defined as the average of these H_i weighted in accordance with the probability of occurrence of the states in question:

$$H = \sum_i P_i H_i$$

$$= -\sum_{ij} P_i p_i(j) \log p_i(j)$$

This is the entropy of the source per symbol of text. If the Markoff process is proceeding at a definite time rate there is also an entropy per second

$$H' = \sum_i f_i H_i$$

where f_i is the average frequency (occurrences per second) of state i. Clearly

$$H' = mH$$

where m is the average number of symbols produced per second. H or H' measure the amount of information generated per symbol or per second by the source. If the logarithmic base is two they will represent bits per symbol or per second.

If successive symbols are independent then H is simply $-\Sigma p_i \log p_i$ where p_i is the probability of symbol i. Suppose in this case we consider a long message of N symbols. It will contain with high probability about p_1N occurrences of the first symbol, p_2N occurrences of the second etc. Hence the probability of this particular message will be roughly

$$p = p_1^{p_1N} \; p_2^{p_2N} \cdots p_n^{p_nN}$$

or

$$\log p \doteq N \sum_i p_i \log p_i$$

$$\log p \doteq -N H$$

$$H \doteq \frac{\log \frac{1}{p}}{N}$$

H is thus approximately the logarithm of the reciprocal probability of a typical long sequence divided by the number of symbols in the sequence. The same result holds for any source. Stated more precisely we have (see appendix 3):

Theorem 3: Given any $\varepsilon > 0$ and $\delta > 0$, we can find an N_0 such that the sequences of any length $N \geq N_0$ fall into two classes.

1. A set whose total probability is less than ε.
2. The remaining set all of whose members have probabilities satisfying the inequality

$$\left| \frac{\log p^{-1}}{N} - H \right| < \delta$$

In other words we are almost certain to have $\frac{\log p^{-1}}{N}$ very close to H when N is large.

A closely related result deals with the number of sequences of various probabilities. Consider again the sequences of length N and let them be arranged in order of decreasing probability. We define n(q) to be the number we must take from this set starting with the most probable one in order to accumulate a total probability q for those taken.

Theorem 4:

$$\lim_{N\to\infty} \frac{\log n(q)}{N} = H$$

when q does not equal 0 or 1.

We may interpret log n(q) as the number of bits required to specify the sequence when we only consider the most probable sequences with a total probability q. Then log n(q)/N is the number of bits per symbol for the specification. The theorem says that for large N this will be independent of q and equal to H. The rate of growth of the logarithm of the number of reasonably probable sequences is given by H, whether "reasonably probable" means excluding the .1% which are least probable or the 99.9%. Due to these results, which are proved in appendix 3, it is possible for most purposes to treat the long sequences as though there were just 2^{HN} of them, each with a probability 2^{-HN}.

The next two theorems show that H can be determined by limiting operations directly from the statistics of the message sequences, without reference to the states and transition probabilities between states.

Theorem 5: Let $p(B_i)$ be the probability of a sequence B_i of symbols from the source. Let

$$G_N = -\frac{1}{N} \sum_i p(B_i) \log p(B_i)$$

where the sum is over all sequences B_i containing N symbols. Then G_N is a monotonic decreasing function of N and

$$\lim_{N\to\infty} G_N = H.$$

Theorem 6. Let $p(B_i,S_j)$ be the probability of sequence B_i followed by symbol S_j and $p_{B_i}(S_j) = p(B_i,S_j)/p(B_i)$ be the conditional probability of S_j after B_i. Let

$$F_N = -\sum_{i,j} p(B_iS_j) \log p_{B_i}(S_j)$$

where the sum is over all blocks B_i of N-1 symbols and over all symbols S_j. Then F_N is a monotonic decreasing function of N,

$$F_N = NG_N - (N-1)\, G_{N-1}$$

$$F_N \leq G_N$$

and $\lim_{N\to\infty} F_N = H.$

These results are derived in appendix 3. They show that a series of approximations to H can be obtained by considering only the statistical structure of the sequences extending over 1, 2, ... N symbols. F_N is the better approximation. In fact F_N is the entropy of the $N^{\underline{th}}$ order approximation to the source of the type discussed above. If there are no statistical influences extending over more than N symbols, that is if the conditional probability of the next symbol knowing the preceeding (N-1) is not changed by a knowledge of any before that, then $F_N = H$. F_N of course is the conditional entropy of the next symbol knowing the (N-1) preceeding ones, while G_N is the entropy per symbol of blocks of N symbols.

The ratio of the entropy of a source to the maximum value it could have while still restricted to the same symbols will be called its relative entropy. This is the maximum compression possible when we encode into the same alphabet. One minus the relative entropy is the redundancy. The redundancy of ordinary English, not considering statistical structure over greater distances than about eight letters is roughly 50%. This means that when we write English half of what we write is determined by the structure of the language and half is chosen freely. The figure 50% was found by several independent methods which all gave results in this neighborhood. One is by calculation of the entropy of the approximations to English. A second method is to delete a certain fraction of the letters from a sample of English text and let someone attempt to restore them. If they can be restored when 50% are deleted the redundancy must be greater than 50%. A third method depends on certain known results in cryptography.

Two extremes of redundancy in English prose are represented by Basic English and by James Joyces' book "Finigans Wake." The Basic English vocabulary is limited to 850 words and the redundancy is very high. This is reflected in the expansion that occurs when a passage is translated into Basic English. Joyce on the other hand enlarges the vocabulary and is alleged to achieve a compression of semantic content.

8. Representation of the Encoding and Decoding Operations

We have yet to represent mathematically the operations performed by the transmitter and receiver in encoding and decoding the information. Either of these will be called a discrete transducer. The input to the transducer is a sequence of input symbols and its output a sequence of output symbols. The transducer may have an internal memory so that its output depends not only on the present input symbol but also on the past history. We assume that the internal memory is finite, i.e., there exists a finite number m of possible states of the transducer and that its output is then a function of the present state and the present input symbol. The next state will be a second function of these two quantities. Thus a transducer can be described by two functions

$$y_n = f(x_n\ \alpha_n)$$

$$\alpha_{n+1} = g(x_n\ \alpha_n)$$

where x_n is the $n^{\underline{th}}$ input symbol

α_n is the state of the transducer when the $n^{\underline{th}}$ input symbol is introduced

y_n is the output symbol (or sequence of output symbols) produced when x_n is introduced if the state is α_n.

If the output symbols of one transducer can be identified with the input symbols of a second, they can be connected in tandem and the result is also a transducer. If there exists a

second transducer which operates on the output of the first and recovers the original input, the first transducer will be called non-singular and the second will be called its inverse.

Theorem 7. The output of a finite state transducer driven by a finite state statistical source is a finite state statistical source, with entropy (per unit time) less than or equal to that of the input. If the transducer is non-singular, equality obtains.

Let α represent the state of the source, which produces a sequence of symbols x_i and let the state of the transducer be β, which produces in its output blocks of symbols y_j. The combined system can be represented by the "product state space" of pairs (α,β). Two points in the space, (α_1,β_1) and (α_2,β_2) are connected by a line if α_1 can produce an x which changes β_1 to β_2, and this line is given the probability of that x in this case. The line is labelled with the block of y_j symbols produced by the transducer. The entropy of the output can be calculated as the weighted sum over the states. If we sum first on β each resulting term is less than or equal to the corresponding term for α, hence the entropy is not increased. If the transducer is non-singular let its output be connected to the inverse transducer. If H_1', H_2' and H_3' are the output entropies of the source, the first and second transducers respectively then $H_1' \geq H_2' \geq H_3' = H_1'$ and therefore $H_1' = H_2'$.

Suppose we have a system of constraints on possible sequences of the type which can be represented by a linear graph as in Fig. 2. If probabilities $p_{ij}^{(s)}$ were assigned to the various lines connecting state i to state j this would become a source. There is one particular assignment which maximizes the resulting entropy given by the following result (Appendix 4).

Theorem 8: Let the system of constraints considered as a channel have a capacity C. If we assign

$$p_{ij}^{(s)} = \frac{B_j}{B_i} C^{-\ell_{ij}^{(s)}}$$

where $\ell_{ij}^{(s)}$ is the duration of the s^{th} symbol leading from state i to state j and the B_i satisfy

$$B_i = \sum_{s,j} B_j C^{-\ell_{ij}^{(s)}}$$

then H is maximized and equal to C.

By proper assignment of the transition probabilities the entropy of symbols on a channel can be maximized at the channel capacity.

9. The Fundamental Theorem for a Noiseless Channel

We will now justify our interpretation of H as the rate of generating information by proving that H determines the channel capacity required with most efficient coding.

Theorem 9: Let a source have entropy H (bits per symbol) and a channel have a capacity C (bits per second). Then it is possible to encode the output of the source in such a way as to transmit at the average rate $\frac{C}{H} - \epsilon$ symbols per second over the channel where ϵ is arbitrarily small. It is not possible to transmit at an average rate greater than $\frac{C}{H}$.

The converse part of the theorem, that $\frac{C}{H}$ cannot be exceeded may be proved by noting that the entropy of the channel input per second is equal to that of the source, since the transmitter must be non-singular, and also this entropy cannot exceed the channel capacity. Hence $H' \leq C$ and the number of symbols per second $= H'/H \leq C/H$.

The first part of the theorem will be proved in two different ways. The first method is to consider the set of all sequences of N symbols produced by the source. For N large we can divide these into two groups, the first containing less than $2^{(H+\eta)N}$ members and the second containing less than 2^{RN} members and having a total probability less than μ. As N increases η and μ approach zero. The number of signals of duration T in the channel is greater than $2^{(C-\theta)T}$ with θ small when T is large. If we choose

$$T = (\frac{H}{C} + \lambda)N$$

then there will be a sufficient number of sequences of channel symbols for the high probability group when N and T are sufficiently large (however small λ) and also some additional ones. The high probability group is coded in an arbitrary one to one way into this set. The remaining sequences are represented by larger sequences, starting and ending with one of sequences not used for the high probability group. This special sequence acts as a start and stop signal for a different code. In between a sufficient time is allowed to give enough different sequences for all the low probability messages. This will require

$$T_1 = (\frac{R}{C} + \varphi)N$$

where φ is small. The mean rate of transmission in message symbols per second will then be greater than

$$\left[(1-\delta)\ \frac{T}{N} + \delta\ \frac{T_1}{N}\right]^{-1}$$

$$= \left[(1-\delta)\ (\frac{H}{C} + \lambda) + \delta\ (\frac{R}{C} + \varphi)\right]^{-1}$$

As N increases δ, λ and φ approach zero and the rate approaches $\frac{C}{H}$.

Another method of performing this coding and proving the theorem can be described as follows: Arrange the messages of length N in order of decreasing probability and suppose their probabilities are $p_1 \geq p_2 \geq p_3 \cdots \geq p_n$. Let $P_s = \sum_1^{s-1} p_i$; that is P_s is the cumulative probability up to, but not including, p_s.

We first encode into a binary system. The binary code for message s is obtained by expanding P_s as a binary number. The expansion is carried out to m_s places, where m_s is the integer satisfying:

$$\log_2 \frac{1}{P_s} \leq m_s < 1 + \log_2 \frac{1}{p_s}$$

Thus the messages of high probability are represented by short codes and those of low probability by long codes. From these inequalities we have

$$\frac{1}{2^{m_s}} \leq P_s < \frac{1}{2^{m_s}-1}$$

The code for P_s will differ from all succeeding ones in one or more of its m_s places, since all the remaining P_i are at least $\frac{1}{2^{m_s}}$ larger and their binary expansions therefore differ in the first m_s places. Consequently all the codes are different and it is possible to recover the message from its code. If the channel sequences are not already sequences of binary digits, they can be ascribed binary numbers in an arbitrary fashion and the binary code thus translated into signals suitable for the channel.

The average number H' of binary digits used per symbol of original message is easily estimated. We have

$$H' = \frac{1}{N} \Sigma\, m_s p_s$$

But,

$$\frac{1}{N} \Sigma(\log \frac{1}{p_s})\, p_s \leq \frac{1}{N} \Sigma\, m_s p_s < \frac{1}{N} \Sigma\, (1 + \log_2 \frac{1}{p_s})\, p_s$$

and therefore,

$$-\Sigma p_s \log p_s \leq H' < \frac{1}{N} - \Sigma p_s \log p_s$$

As N increases $-\Sigma p_s \log p_s$ approaches H, the entropy of the source and H' approaches H.

We see from this that the inefficiency in coding when only a finite delay of N symbols is used, need not exceed $\frac{1}{N}$ plus the difference between the true entropy H, the entropy H_N calculated for sequences of length N. The per cent excess time needed over the ideal is therefore less than

$$\frac{G_N}{H} + \frac{1}{HN} - 1.$$

This method of encoding is substantially the same as one found independently by R. M. Fano. His method is to arrange the messages of length N in order of decreasing probability. Divide this series into two groups of as nearly equal probability as possible. If the message is in the first group its first binary digit will be 0, otherwise 1. The groups are similarly divided into subsets of nearly equal probability and the particular subset determines the second binary digit. This process is continued until each subset contains only one message. It is easily seen that apart from minor differences (generally in the last digit) this amounts to the same thing as the arithmetic process described above.

10. Discussion

In order to obtain the maximum power transfer from a generator to a load a transformer must in general be introduced so that the generator as seen from the load has the load resistance. The situation here is roughly analogous. The transducer which does the encoding should match the source to the channel in a statistical sense. The source as seen from the channel through the transducer should have the same statistical structure as the source which maximizes entropy in the channel. The content of Theorem 9 is that although an exact match is not in general possible, we can approximate it as closely as desired. The ratio of the actual rate of transmission to the capacity C may be called the efficiency of the coding system. This is of course equal to the ratio of the actual entropy of the channel symbols to their maximum possible entropy.

In general ideal or nearly ideal encoding requires a long delay in the transmitter and receiver. In the noiseless case which we have been considering, the main function of this delay is to allow reasonably good matching of probabilities to corresponding lengths of sequences. With a good code the logarithm of the reciprocal probability of a long message must be proportional to the durection of the corresponding signal, in fact

$$\left| \frac{\log p^{-1}}{T} - C \right|$$

must be small for all but a small fraction of the long messages.

If a source can produce only one particular message its entropy is zero, and no channel is required. For example, a computing machine set up to calculate the successive digits of π produces a definite sequence with no chance element. No channel is required to "transmit" this to another point. One could construct a second machine at this point to compute the same sequence. However, this may be impractical. In such a case we can choose to ignore some or all of the statistical knowledge we have of the source. We might consider the digits of π to be a random sequence in that we construct a system capable of sending any sequence of digits. In a similar way we may choose to use some of our statistical knowledge of English in constructing a code, but not all of it. In such a case we consider the source with the maximum entropy subject to the statistical conditions we wish to retain. The entropy of this source determines the channel capacity which is necessary and sufficient. In the π example the only information retained is that all the digits are chosen from the set 0, 1, ..., 9. In the case of English one might wish to use the statistical savings possible due to letter frequencies, but nothing else. The maximum entropy source is then the 1st approximation to English and its entropy determines the required channel capacity.

11. Examples

As a simple example of some of these results consider a source which produces a sequence of letters chosen from among A, B, C, D, with probabilities $\frac{1}{2}$, $\frac{1}{4}$, $\frac{1}{8}$, $\frac{1}{8}$, successive symbols being chosen independently. We have

$$H = -(\frac{1}{2} \log \frac{1}{2} + \frac{1}{4} \log \frac{1}{4} + \frac{2}{8} \log \frac{1}{8})$$

$$= \frac{7}{4} \text{ bits per symbol.}$$

Thus we can approximate a coding system to encode messages from this source into binary digits with on the average 7/4 binary digit per symbol. In this case we can actually achieve the limiting value by the following code (obtained by the method of the second proof of Theorem 8):

A	0
B	10
C	110
D	111

The average number of binary digits used in encoding a sequence of N symbols will be

$$N (\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{2}{8} \times 3) = \frac{7}{4} N$$

It is easily seen that the binary digits 0,1 have probabilities $\frac{1}{2}$, $\frac{1}{2}$ so the H for the coded sequences is 1 bit per symbol.

Since on the average we have $\frac{7}{4}$ binary symbols per original letter the entropies on a time basis are the same. The maximum possible entropy for the original set is log 4 = 2, occurring when A, B, C, D have probabilities $\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4}$. Hence the relative entropy is $\frac{7}{8}$. We can translate the binary sequences into the original set of symbols on a two to one basis by the following table:

00	A'
01	B'
10	C'
11	D'

This double process then encodes the original message into the same symbols but with an average compression ratio $\frac{7}{8}$.

As a second example consider a source which produces a sequence of A's and B's with probability p for A and q for B. If $p << q$ we have

$$H = - \log p^p (1-p)^{1-p}$$
$$= - p \log p (1-p)^{\frac{1-p}{p}}$$
$$\doteq p \log \frac{e}{p}$$

In such a case one can construct a fairly good coding of the message on a 0,1 channel by sending a special sequence, say 0000, for the infrequent symbol A and then a sequence indicating the <u>number</u> of B's following it. This could be indicated by the binary representation with all numbers containing the special sequence deleted. All numbers up to 16 are represented as usual,

16 is represented by the next binary number after 16 which does not contain four zeros, namely 17 = 10001, etc.

It can be shown that as $p \rightarrow 0$ the coding approaches ideal providing the length of the special sequence is properly adjusted.

PART II The Discrete Channel With Noise

11. Representation of a Noisy Discrete Channel

We now consider the case where the signal is perturbed by noise during transmission or at one or the other of the terminals. This means that the received signal is not necessarily the same as that sent out by the transmitter. Two cases may be distinguished. If a particular transmitted signal always produces the same received signal; i.e., the received signal is a definite function of the transmitted signal, then the effect may be called distortion. If this function has an inverse, no two transmitted signals producing the same received signal, distortion may be corrected, at least in principle, by merely performing the inverse functional operation on the received signal.

The case of interest here is that in which the signal does not always undergo the same change in transmission. In this case we may assume the received signal E to be a function of the transmitted signal S and a second variable, the noise N.

$$E = f(S,N)$$

The noise is considered to be a chance variable just as the message was above. In general it may be represented by a suitable stochastic process. The most general type of noisy discrete channel we shall consider is a generalization of the finite state noise free channel described previously. We

assume a finite number of states and a set of probabilities

$$p_{\alpha,i}(\beta,j) .$$

This is the probability if the channel is in state α and symbol i is transmitted that symbol j will be received and the channel left in state β. Thus α and β range over the possible states, i over the possible transmitted symbols and j over the possible received symbols In the case where successive symbols are independently perturbed by the noise there is only one state, and the channel is described by the set of transition probabilities $p_i(j)$, the probability of transmitted symbol i being received as j.

If a noisy channel is fed by a source there are two statistical processes as work; the source and the noise. Thus there are a number of entropies that can be calculated. First there is the entropy H(x) of the source or of the input to the channel (these will be equal if the transmitter is nonsingular). The entropy of the output of the channel, i.e., the received signal will be denoted by H(y). In the noiseless case H(y) = H(x). The joint entropy of input and output will be H(x,y). Finally there are two conditional entropies $H_x(y)$ and $H_y(x)$, the entropy of the output when the input is known and conversely. Among these quantities we have the relations

$$H(x, y) = H(x) + H_x(y) = H(y) + H_y(x).$$

All of these entropies can be measured on a per second or a per symbol basis.

12. Equivocation and Channel Capacity

If the channel is noisy it is not in general possible to reconstruct the original message or the transmitted signal with certainty by any operation on the received signal E. There are, however, ways of transmitting the information which are optimal in combating noise. This is the problem which we now consider.

Suppose there are two possible symbols 0 and 1, and we are transmitting at a rate of 1000 symbols per second with probabilities $p_0 = p_1 = \frac{1}{2}$. Thus our source is producing information at the rate of 1000 bits per second. During transmission the noise introduces errors, so that on the average 1 in 100 is received incorrectly (a 0 as 1, or 1 as 0). What is the rate of transmission of information? Certainly less than 1000 bits per second since about 1% of the received symbols are incorrect. Our first impulse might be to say the rate is 990 bits per second, merely subtracting the expected number of errors. This is not satisfactory since it fails to take into account the recipient's lack of knowledge of where the errors occur. We may carry it to an extreme case and suppose the noise so great that the received symbols are entirely independent of the transmitted symbols. The probability of receiving 1 is 1/2 whatever was transmitted and similarly for 0. Then about half of the received symbols are correct due to chance alone and we would be giving the system credit for transmitting 500 bits per second while actually no information is being transmitted

at all. Equally "good" transmission would be obtained by dispensing with the channel entirely and flipping a coin at the receiving point.

Evidently the proper correction to apply to the amount of information transmitted is the amount of this information which is missing in the received signal, or alternatively the uncertainty when we have received a signal of what was actually sent. From our previous discussion of entropy as a measure of uncertainty it seems reasonable to use the conditional entropy of the message knowing the received signal as a measure of this missing information. This is indeed the proper definition as we will see later. Following this idea the rate of actual transmission, R, would be obtained by subtracting from the rate of production (i.e., the entropy of the source) the average rate of conditional entropy.

$$R = H(x) - H_y(x) \ .$$

The conditional entropy $H_y(x)$ will for convenience be called the <u>equivocation</u>. It measures the average ambiguity of the received signal.

In the example considered above if a 0 is received the a posteriori probability that a 0 was transmitted is .99, and that a 1 was transmitted is .01. These figures are reversed if a 1 is received. Hence

$$H_y(x) = -[.99 \log .99 + 0.01 \log 0.01]$$
$$= .081 \text{ bits/symbol}$$

or 81 bits per second. The system is therefore transmitting at a rate 1000 - 81 = 919 bits per second. In the extreme case where a 0 is equally likely to be received as a 0 or 1 and similarly for 1, the a posteriori probabilities are $\frac{1}{2}$, $\frac{1}{2}$ and

$$H_y(x) = -\left[\frac{1}{2} \log \frac{1}{2} + \frac{1}{2} \log \frac{1}{2}\right]$$
$$= 1 \text{ bit per symbol}$$

or 1000 bits per second. The rate of transmission is then 0 as it should be.

The following theorem gives a direct intuitive interpretation of the equivocation and also serves to justify it as the unique appropriate measure. We consider a communication system and an observer (or auxiliary device) who can see both what is sent and what is recovered (with errors due to noise). This observer notes the errors in the recovered message and transmits data to the receiving point over a "correction channel" to enable the receiver to correct the errors. The situation is indicated schematically in Fig. 8

Theorem 10: If the correction channel has a capacity equal to $H_y(x)$ it is possible to so encode the correction data as to send it over this channel and correct all but an arbitrarily small fraction ε of the errors. This is not possible if the channel capacity is less than $H_y(x)$.

Roughly then $H_y(x)$ is the amount of additional information that must be supplied per second at the receiving point to correct the received message.

To prove the first part, consider long sequences of received message M' and corresponding original message M. There will be logarithmically $TH_y(x)$ of the M's which could reasonably have produced each M'. Thus we have $TH_y(x)$ binary digits to send each T seconds. This can be done with ϵ frequency of errors on a channel of capacity $H_y(x)$.

The second part can be proved by noting first that for any discrete chance variables x, y, z

$$H_y(x,z) \geq H_y(x)$$

The left-hand side can be expanded to give

$$H(z) + H_{yz}(x) > H_y(x)$$

$$H_{yz}(x) > H_y(x) - H(y).$$

If we identify x as the output of the source, y as the received signal and z as the signal sent over the correction channel, then the right-hand side is the equivocation less the rate of transmission over the correction channel. If the capacity of this channel is less than the equivocation the right-hand side will be greater than zero and $H_{yz}(x) > 0$. But this is the uncertainty of what was sent knowing both the received signal and the correction signal. If this is greater than zero the frequency of errors cannot be arbitrarily small.

Example:

Suppose the errors occur at random in a sequence of binary digits; probability p that a digit is wrong and q = 1 - p that it is right. These errors can be corrected if their position is known. Thus the correction channel need only send information as to these positions. This amounts to transmitting from a source which produces binary digits with probability p for 1 (correct) and q for 0 (incorrect). This requires a channel of capacity

$$- [p \log p + q \log q]$$

which is the equivocation of the original system.

The rate of transmission R can be written in two other forms due to the identities noted above. We have

$$\begin{aligned} R &= H(x) - H_y(x) \\ &= H(y) - H_x(y) \\ &= H(x) + H(y) - H(x,y). \end{aligned}$$

The first defining expression has already been interpreted as the amount of information sent less the uncertainty of what was sent. The second measures the amount received less the part of this which is due to noise. The third is the sum of the two amounts less the joint entropy and therefore in a sense is the number of bits per second common to the two. Thus all three expressions have a certain intuitive significance.

The capacity C of a noisy channel should be the maximum possible rate of transmission, i.e., the rate when the source is properly matched to the channel. We therefore define the channel capacity by

$$C = \text{Max}\,(H(x) - H_y(x))$$

where the maximum is with respect to all possible information sources used as input to the channel. If the channel is noiseless $H_y(x) = 0$. The definition is then equivalent to that already given for a noiseless channel since the maximum entropy for the channel is its capacity.

13. The Fundamental Theorem for a Discrete Channel With Noise

It may seem surprising that we should define a definite capacity C for a noisy channel since we can never send certain information in such a case. It is clear, however, that by sending the information in a redundant form the probability of errors can be reduced. For example by repeating the message many times and by a statistical study of the different received versions of the message the probability of errors could be made very small. One would expect, however, that to make this probability of errors approach zero, the redundancy of the encoding must increase indefinitely, and the rate of transmission therefore approach zero. This is by no means true. If it were, there would not be a very well defined capacity, but only a capacity for a given frequency of errors, or a given equivocation; the capacity going down as the error

requirements are made more stringent. Actually the capacity C defined above has a very definite significance. It is possible to send information at the rate C through the channel with as small a frequency of errors or equivocation as desired by proper encoding. This statement is not true for any rate greater than C. If an attempt is made to transmit at a higher rate than C, say $C + R_1$, then there will necessarily be an equivocation equal to or greater than the excess R_1. Nature takes payment by requiring just that much uncertainty, so that we are not actually getting any more than C through correctly.

The situation is indicated in Fig. 9. The rate of information into the channel is plotted horizontally and the equivocation vertically. Any point above the heavy line in the shaded region can be attained and those below cannot. The points on the line cannot in general be attained, but there will usually be two points on the line that can.

These results are the main justification for the definition of C and will now be proved.

Theorem 11. Let a discrete channel have the capacity C and a discrete source the entropy per second H. If $H \leq C$ there exists a coding system such that the output of the source can be transmitted over the channel with an arbitrarily small frequency of errors (or an arbitrarily small equivocation). If $H > C$ it is possible to encode the source so that the equivocation is less than $H - C + \varepsilon$ where ε is arbitrarily small.

There is no method of encoding which gives an equivocation less than H - C.

The method of proving the first part of this theorem is not by exhibiting a coding method having the desired properties, but by showing that such a code must exist in a certain group of codes. In fact we will average the frequency of errors over this group and show that this average can be made less than ε. If the average of a set of numbers is less than ε there must exist at least one in the set which is less than ε. This will establish the desired result.

The capacity C of a noisy channel has been defined as

$$C = \text{Max } (H(x) - H_y(x))$$

where x is the input and y the output. The maximization is over all sources which might be used as input to the channel.

Let S_0 be a source which achieves the maximum capacity C. If this maximum is not actually achieved by any source let S_0 be a source which approximates to giving the maximum rate. Suppose S_0 is used as input to the channel. We consider the possible transmitted and received sequences of a long duration T. We will have:

1. The transmitted sequences fall into two classes, a high probability group with about $2^{T\,H(x)}$ members and the remaining sequences of small total probability.

2. Similarly the received sequences have a high probability set of about $2^{T\,H(y)}$ members and a low probability set of remaining sequences.

3. Each high probability output could be produced by about $2^{T\,H_y(x)}$ inputs. The probability of all other causes has a small total.

All the ε's and δ's implied by the words "small" and "about" in these statements approach zero as we allow T to increase and S_o to approach the maximizing source.

The situation is summarized in Fig. 10 where the input blocks are points on the left and output blocks points on the right. The fan of cross lines represents the range of possible causes for a typical output.

Now suppose we have another source producing information at rate R with $R < C$. In the period T this source will have $2^{T\,R}$ high probability outputs. We wish to associate these with a selection of the possible channel inputs in such a way as to get a small frequency of errors. We will set up this association in all possible ways (using, however, only the high probability group of inputs as determined by the source S_o) and average the frequency of errors for this large class of possible coding systems. This is the same as calculating the frequency of errors for a random association of the messages and channel inputs of duration T. Suppose a particular output y_1

is observed. What is the probability of more than one message in the set of possible causes of y_1? There are 2^{TR} messages distributed at random in $2^{TH(x)}$ points. The probability of a particular point being a message is thus

$$2^{T(R-H(x))}$$

The probability that none of the points in the fan is a message (apart from the actual originating message) is

$$P = \left[1 - 2^{T(R-H(x))}\right]^{2^{TH_y(x)}}$$

Now $R < R(x) - H_y(x)$ so $R-H(x) = -H_y(x) - \eta$ with η positive. Consequently

$$P = \left[1 - 2^{-TH_y(x)2-T\eta}\right]^{2^{TH_y(x)}}$$

approaches (as $T \to \infty$)

$$1 - 2^{-T\eta}$$

Hence the probability of an error approaches zero and the first part of the theorem is proved.

The second part of the theorem is easily shown by noting that we could merely send C bits per second from the source completely neglecting the remainder of the information generated. At the receiver the neglected part gives an

equivocation H(x)-C and the part transmitted need only add ϵ. This limit can also be attained in many other ways as will be shown when we consider the continuous case.

The last statement of the theorem is a simple consequence of our definition of C. Suppose we can encode a source with R = C + a in such a way as to obtain an equivocation less than a. Then $H_y(x) < a$ and $R = H(x) = C + a$

$$H(x) - H_y(x) > C$$

This contradicts the definition of C as the maximum of $H(x) - H_y(x)$.

Actually more has been proved than was stated in the theorem. If the average of a set of positive numbers is within ϵ of zero, a fraction of at most $\sqrt{\epsilon}$ can be greater than $\sqrt{\epsilon}$. Since ϵ is arbitrarily small we can say that almost all the systems are arbitrarily close to the ideal.

14. Discussion

The demonstration of theorem 11, while not a pure existence proof, has some of the deficiencies of such proofs. An attempt to obtain a good approximation to ideal coding by following the method of the proof is generally impractical. In fact apart from some rather trivial cases and certain limiting situations no explicit description of a series of approximation to the ideal has been found. Probably this is no accident but is related to the difficulty of giving an explicit construction for a good approximation to a random sequence.

An approximation to the ideal would have the property that if the signal is altered in a reasonable way by the noise, the original can still be recovered. In other words the alteration will not in general bring it closer to another reasonable signal than the original. This is accomplished at the cost of a certain amount of redundancy in the coding. The redundancy must be introduced in the proper way to combat the particular noise structure involved. However, any redundancy in the source will usually help if it is utilized at the receiving point. In particular, if the source already has a certain redundancy and no attempt is made to eliminate it in matching to the channel, this redundancy will help combat noise. For example, in a noiseless teletype channel one could save about 50% in time by proper encoding of the messages. This is not done and the redundancy of English remains in the channel symbols. This has the advantage, however, of allowing considerable noise in the channel. A sizable fraction of the letters can be received incorrectly and still reconstructed by the context. In fact this is probably not a bad approximation to the ideal in many cases, since the statistical structure of English is rather involved and the reasonable English sequences not too far (in the sense required for the theorem) from a random selection.

As in the noiseless case a delay is generally required to approach the ideal encoding. It now has the additional function of allowing a large sample of noise to affect

the signal before any judgment is made at the receiving point as to the original message. Increasing the sample size always sharpens the possible statistical assertions.

15. Example of a Discrete Channel and Its Capacity

A simple example of a discrete channel is indicated in Fig. 11. There are three possible symbols. The first is never affected by noise. The second and third each have probability p of coming through undisturbed, and q of being changed into the other of the pair. We have (letting $\alpha = -[p \log p + q \log q]$, P be the probability of the first symbol and Q that of the second and third),

$$H(x) = -P \log P - 2Q \log Q$$

$$H_y(x) = 2Q\alpha$$

We wish to choose P and Q in such a way as to maximize $H(x) - H_y(x)$, subject to the constraint $P + 2Q = 1$. Hence we consider

$$U = -P \log P - 2Q \log Q - 2Q\alpha + \lambda(P + 2Q)$$

$$\frac{\partial U}{\partial P} = -1 - \log P + \lambda = 0$$

$$\frac{\partial U}{\partial Q} = -2 - 2 \log Q - 2\alpha + 2\lambda = 0$$

Eliminating λ

$$\log P = \log Q + \alpha$$

$$P = Qe^{\alpha} = Q\beta$$

$$P = \frac{\beta}{\beta+2} \quad Q = \frac{1}{\beta+2}$$

The channel capacity is then

$$C = \log \frac{\beta+2}{\beta}$$

Note how this checks the obvious values in the cases $p = 1$ and $p = \frac{1}{2}$. In the first $\beta = 1$ and $C = \log 3$ which is correct since the channel is then noiseless with 3 possible symbols. If $p = \frac{1}{2}$ $\beta = 2$ and $C = \log 2$. Here the second and third symbols cannot be distinguished at all and act together like one symbol. The first symbol is used with probability $p = \frac{1}{2}$ and the second and third together with probability $\frac{1}{2}$. This may be distributed in any desired way and still achieve the maximum capacity.

For intermediate values of p the channel capacity will lie between log 2 and log 3. The distinction between the second and third symbols conveys some information but not as much as in the noiseless case. The first symbol is used somewhat more frequently than the other two because of its freedom from noise.

16. The Channel Capacity in Certain Special Cases

If the noise affects successive channel symbols independently it can be described by a set of transition probabilities p_{ij}. This is the probability if symbol i is sent that j will be received. The channel capacity C is then given by the maximum of

$$\sum_{i,j} P_i p_{ij} \log \sum_i P_i p_{ij} - \sum_{i,j} P_i p_{ij} \log p_{ij}$$

where we vary the P_i subject to $\Sigma P_i = 1$. This leads by the method of Lagrange to the equations,

$$\sum_j P_{sj} \log \frac{p_{sj}}{\sum_i p_i p_{ij}} = \mu \qquad s = 1, 2, \ldots$$

Multiplying by P_s and summing on s shows that $\mu = -C$. Let the inverse of p_{sj} (if it exists) be h_{sj} so that $\sum_s h_{st} p_{sj} = \delta_{tj}$. Then:

$$\sum_{sj} h_{st} p_{sj} \log p_{sj} - \log \sum_i p_i p_{it} = -C \sum_s h_{st}$$

Hence:

$$\sum_i P_i p_{it} = \exp \left[C \sum_s h_{st} + \sum_{s,j} h_{st} p_{sj} \log p_{sj} \right]$$

or,

$$P_i = \sum_t h_{it} \exp \left[C \sum_s h_{st} + \sum_{s,j} h_{st} p_{sj} \log p_{sj} \right]$$

This is the system of equations for determining the maximizing values of P_i, with C to be determined so that $\Sigma P_i = 1$. When this is done C will be the channel capacity, and the P_i the proper probabilities for the channel symbols to achieve this capacity.

If each input symbol has the same set of probabilities on the lines emerging from it, and the same is true of each output symbol, the capacity can be easily calculated.

Examples are shown in Fig. 12. In such a case $H_x(y)$ is independent of the distribution of probabilities on the input symbols, and is given by $-\Sigma\, p_i \log p_i$ where the p_i are the values of the transition probabilities from any input symbol. The channel capacity is

$$\text{Max}\,[H(y) - H_x(y)]$$

$$= \text{Max}\, H(y) + \Sigma\, p_i \log p_i .$$

The maximum of H(y) is clearly log m where m is the number of output symbols since it is possible to make them all equally probable by making the input symbols equally probable. The channel capacity is therefore

$$C = \log m + \Sigma\, p_i \log p_i$$

In Fig. 12a it would be

$$C = \log 4 - \log 2 = \log 2$$

This could be achieved by using only the 1st and 3rd symbols. In Fig. 12b

$$C = \log 4 - \frac{2}{3} \log 3 - \frac{1}{3} \log 6$$

$$= \log 4 - \log 3 - \frac{1}{3} \log 2$$

$$= \log \frac{1}{3}\, 2^{\frac{5}{3}}$$

In Fig. 12c we have

$$C = \log 3 - \frac{1}{2}\log 2 - \frac{1}{3}\log 3 - \frac{1}{6}\log 6$$

$$= \log \frac{3}{2^{\frac{1}{2}}\, 3^{\frac{1}{3}}\, 6^{\frac{1}{6}}}$$

Suppose the symbols fall into several groups such that the noise never causes a symbol in one group to be mistaken for a symbol in another group. Let the capacity for the nth group be C_n when we only use the symbols in this group. Then it is easily shown that for best use of the entire set, the total probability P_n of all symbols in the nth group should be

$$P_n = \frac{2^{C_n}}{\Sigma 2^{C_n}}$$

Within a group the probability is distributed just as it would be if these were the only symbols being used. The channel capacity is

$$C = \log \Sigma 2^{C_n} .$$

17. An Example of Efficient Coding

The following example, although somewhat unrealistic, is a case in which exact matching to a noisy channel is possible. There are two channel symbols, 0 and 1, and the noise affects them in blocks of seven symbols. A block of seven is either transmitted without error, or exactly one symbol of the seven is incorrect. These eight possibilities are equally likely. We have

$$C = \text{Max}\,[H(y) - H_x(y)]$$

$$= \frac{1}{7}\left[7 + \frac{8}{8}\log\frac{1}{8}\right]$$

$$= \frac{4}{7} \text{ bits/symbol}$$

An efficient code, allowing complete correction of errors, and transmitting at the rate C is the following (found by a method due to R. Hamming).

Let a block of seven symbols be $X_1, X_2, \ldots X_7$. Of these X_3, X_5, X_6 and X_7 are message symbols and chosen arbitrarily by the source. The other three are redundant and calculated as follows:

X_4 is chosen to make $\alpha = X_4 + X_5 + X_6 + X_7$ even

X_2 " " " " $\beta = X_4 + X_5 + X_6 + X_7$ "

X_1 " " " " $\gamma = X_1 + X_3 + X_5 + X_7$ "

When a block of seven is received, α, β and γ are calculated and if even called zero, if odd called one. The binary number $\alpha\ \beta\ \gamma$ then gives the subscript of the X_i that is incorrect (if 0 there was no error).

APPENDIX 1

The Growth of the Number of Blocks of Symbols With A Finite State Condition

Let $N_i(L)$ be the number of blocks of symbols of length L ending in state i. Then we have

$$N_j(L) = \sum_{is} N_i(L - b_{ij}^{(s)})$$

where $b_{ij}^{1}, b_{ij}^{2}, \ldots b_{ij}^{m}$ are the lengths of the symbols which may be chosen in state i and lead to state j. These are linear difference equations and the behavior as $L \to \infty$ must be of the type

$$N_j = A_j W^L$$

Substituting in the difference equation

$$A_j W^L = \sum_{i,s} A_i \, W^{L - b_{ij}^{(s)}}$$

or

$$A_j = \sum_{is} A_i \, W^{-b_{ij}^{(s)}}$$

$$\sum_i \left(\sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right) A_i = 0$$

For this to be possible the determinant

$$D(W) = |a_{ij}| = \left| \sum_s W^{-b_{ij}^{(s)}} - \delta_{ij} \right|$$

must vanish and this determines W, which is, of course, the largest real root of D = 0.

The quantity C is then given by

$$C = \lim_{L \to \infty} \frac{\log \Sigma A_j W^L}{L} = \log W$$

and we also note that the same growth properties result if we require that all blocks start in the same (arbitrarily chosen) state.

APPENDIX 2

Derivation of $H = -\Sigma\, p_i \log p_i$

Let $H(\frac{1}{n}, \frac{1}{n}, \ldots, \frac{1}{n}) = A(n)$. From condition (3) we can decompose a choice from s^m equally likely possibilities into a series of m choices each from s equally likely possibilities and obtain

$$A\,(s^m) = m\,A(s)$$

Similarly

$$A\,(t^n) = n\,A(t)$$

We can choose n arbitrarily large and find an m to satisfy

$$s^m \leq t^n < s^{(m+1)}$$

Thus, taking logarithms and dividing by a log s,

$$\frac{m}{n} \leq \frac{\log t}{\log s} \leq \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{\log t}{\log s} \right| < \varepsilon$$

where ε is arbitrarily small.

Now from the monotonic property of $A(n)$

$$A(s^m) \leq A(t^n) \leq A(s^{m+1})$$

$$m\,A(s) \leq nA(t) \leq (m+1)\,A(s)$$

Hence, dividing by $nA(s)$,

$$\frac{m}{n} \leq \frac{A(t)}{A(s)} \leq \frac{m}{n} + \frac{1}{n} \quad \text{or} \quad \left| \frac{m}{n} - \frac{A(t)}{A(s)} \right| < \varepsilon$$

$$\left| \frac{A(t)}{A(s)} - \frac{\log t}{\log s} \right| \leq 2\varepsilon \qquad A(t) = -K \log t$$

where K must be positive to satisfy (2).

Now suppose we have a choice from n possibilities with commeasurable probabilities $p_i = \frac{n_i}{\Sigma n_i}$ where the n_i are integers. We can break down a choice from Σn_i possibilities into a choice from n possibilities with probabilities $p_i \ldots p_n$ and then, if the ith was chosen, a choice from n_i with equal probabilities. Using condition 3 again, we equate the total choice from Σn_i as computed by two methods

$$K \log \Sigma n_i = H(p_1, \ldots, p_n) + K\Sigma\, p_i \log n_i$$

Hence

$$H = K \left[\Sigma\, p_i \log \Sigma\, n_i - \Sigma\, p_i \log n_i\right]$$

$$= -K\, \Sigma\, p_i \log \frac{n_i}{\Sigma n_i} = -K\, \Sigma\, p_i \log p_i$$

If the p_i are incommeasurable, they may be approximated by rationals and the same expression must hold by our continuity assumption. Thus the expression holds in general. The choice of coefficient K is a matter of convenience and amounts to the choice of a unit of measure.

APPENDIX 3

Theorems on Ergodic Sources

If it is possible to go from any state with $P > 0$ to any other along a path of probability $p > 0$, the system is ergodic and the strong law of large numbers can be applied. Thus the number of times a given path p_{ij} in the network is traversed in a long sequence of length N is about proportional to the probability of being at i and then choosing this path, $P_i p_{ij} N$. If N is large enough the probability of percentage error $\pm \delta$ in this is less than ε so that for all but a set of small probability the actual numbers lie within the limits

$$(P_i p_{ij} \pm \delta)N$$

Hence nearly all sequences have a probability p given by

$$p = \Pi\, p_{ij}^{(P_i p_{ij} \pm \delta)N}$$

and $\frac{\log p}{N}$ is limited by

$$\frac{\log p}{N} = \Sigma(P_i p_{ij} \pm \delta)\log p_{ij}$$

or

$$\left| \frac{\log p}{N} - \Sigma\, P_i p_{ij} \log p_{ij} \right| < \eta$$

This proves theorem 3.

Theorem 4 follows immediately from this on calculating upper and lower bounds for $\eta(q)$ based on the possible range of values of p in Theorem 3.

In the mixed (not ergodic) case if

$$L = \Sigma\, p_i\, L_i$$

and the entropies of the components are $H_1 \geq H_2 \geq \cdots \geq H_n$ we have the

Theorem: $\lim_{N \to \infty} \frac{\text{Log } \eta(q)}{N} = \varphi(q)$ is a decreasing step function,

$$\varphi(q) = H_s \text{ in the interval } \sum_1^{s-1} p_i < q < \sum_1^{s} p_i$$

To prove theorems 5 and 6 first note that F_N is monotonic decreasing because increasing N adds a subscript to a conditional entropy. A simple substitution for $p_{Bi}(s_j)$ in the definition of F_N shows that

$$F_N = N\, G_N - (N-1)\, G_{N-1}$$

and summing this for all N gives $G_N = \frac{1}{N} \Sigma\, F_N$. Hence $G_N \geq F_N$ and G_N monatonic decreasing. Also they must approach the same limit. By using theorem 3 we see that $\lim_{N \to \infty} G_N = H$.

APPENDIX 4

Maximizing the Rate for a System of Constraints

Suppose we have a set of constraints on sequences of symbols that is of the finite state type and can be represented therefore by a linear graph. Let ℓ_{ij}^{s} be the lengths of the various symbols that can occur in passing from state i to state j. What distribution of probabilities P_i for the different states and p_{ij}^{s} for choosing symbol s in state i and going to state j maximize the rate of generating information under these constraints? The constraints define a discrete channel and the maximum rate must be less than or equal to the capacity C of this channel, since if all blocks of large length were equally likely this rate would result, and if possible this would be best. We will show that this rate can be achieved by proper choice of the P_i and $p_i^{s}(j)$.

The rate in question is

$$\frac{-\Sigma P_i p_{ij}^{(s)} \log p_{ij}^{s}}{\Sigma P_{(i)} p_{ij}^{(s)} \ell_{ij}^{s}} = \frac{N}{M}$$

Let $\ell_{ij} = \sum_s \ell_{ij}^{s}$. Evidently for a maximum $p_{ij}^{s} = k \exp \ell_{ij}^{s}$.

The constraints on maximization are $\Sigma P_i = 1$. $\sum_j p_{ij} = 1$

$\Sigma P_i (p_{ij} - \delta_{ij}) = 0$.

Hence we maximize

$$U = \frac{-\Sigma P_i p_{ij} \log p_{ij}}{\Sigma P_i p_{ij} \ell_{ij}} + \lambda \sum_i P_i + \Sigma \mu_i p_{ij} + \Sigma \eta_i P_i (p_{ij} - \delta_{ij})$$

$$\frac{\partial U}{\delta p_{ij}} = - \frac{M P_i (1+\log p_{ij}) + N P_i \ell_{ij}}{M^2} + \lambda + \mu_i + \eta_i P_i = 0$$

Solving for p_{ij}

$$p_{ij} = A_i B_j D^{-\ell_{ij}}$$

Since

$$\sum_j p_{ij} = 1, \quad A_i^{-1} = \sum_j B_j D^{-\ell_{ij}}$$

$$p_{ij} = \frac{B_j D^{-\ell_{ij}}}{\sum_s B_s D^{-\ell_{is}}}$$

The correct value of D is the capacity C and the B_j are solutions of

$$B_i = \Sigma B_j C^{-\ell_{ij}}$$

for then

$$P_{ij} = \frac{B_j}{B_i} C^{-\ell_{ij}}$$

$$\Sigma P_i \frac{B_j}{B_i} C^{-\ell_{ij}} = P_j$$

Hence we maximize

$$U = \frac{-\Sigma P_i p_{ij} \log p_{ij}}{\Sigma P_i p_{ij} \ell_{ij}} + \lambda \sum_i P_i + \Sigma \mu_i p_{ij} + \Sigma \eta_i P_i (p_{ij} - \delta_{ij})$$

$$\frac{\partial U}{\delta p_{ij}} = - \frac{M P_i (1+\log p_{ij}) + N P_i \ell_{ij}}{M^2} + \lambda + \mu_i + \eta_i P_i = 0$$

Solving for p_{ij}

$$p_{ij} = A_i B_j D^{-\ell_{ij}}$$

Since

$$\sum_j p_{ij} = 1, \quad A_i^{-1} = \sum_j B_j D^{-\ell_{ij}}$$

$$p_{ij} = \frac{B_j D^{-\ell_{ij}}}{\sum_s B_s D^{-\ell_{is}}}$$

The correct value of D is the capacity C and the B_j are solutions of

$$B_i = \Sigma B_j C^{-\ell_{ij}}$$

for then

$$P_{ij} = \frac{B_j}{B_i} C^{-\ell_{ij}}$$

$$\Sigma P_i \frac{B_j}{B_i} C^{-\ell_{ij}} = P_j$$

or

$$\Sigma \frac{P_i}{B_i} C^{-\ell_{ij}} = \frac{P_j}{B_i}$$

So that if γ_i satisfy

$$\Sigma \gamma_i C^{-\ell_{ij}} = \gamma_j$$

$$P_i = B_i \gamma_i$$

Both of the sets of equations for B_i and γ_i can be satisfied since C is such that

$$\left| C^{-\ell_{ij}} - \delta_{ij} \right| = 0$$

In this case the rate is

$$- \frac{\Sigma P_i p_{ij} \log \frac{B_j}{B_i} C^{-\ell_{ij}}}{\Sigma P_i p_{ij} \ell_{ij}}$$

$$= C - \frac{\Sigma P_i p_{ij} \log \frac{B_j}{B_i}}{\Sigma P_i p_{ij} \ell_{ij}}$$

but

$$\Sigma P_i p_{ij} (\log B_j - \log B_i)$$

$$= \Sigma_j P_j \log B - \Sigma P_i \log B_i = 0$$

Hence the rate is C and as this could never be exceeded this is the maximum, justifying the assumed solution.

C. E. SHANNON

April 21, 1948

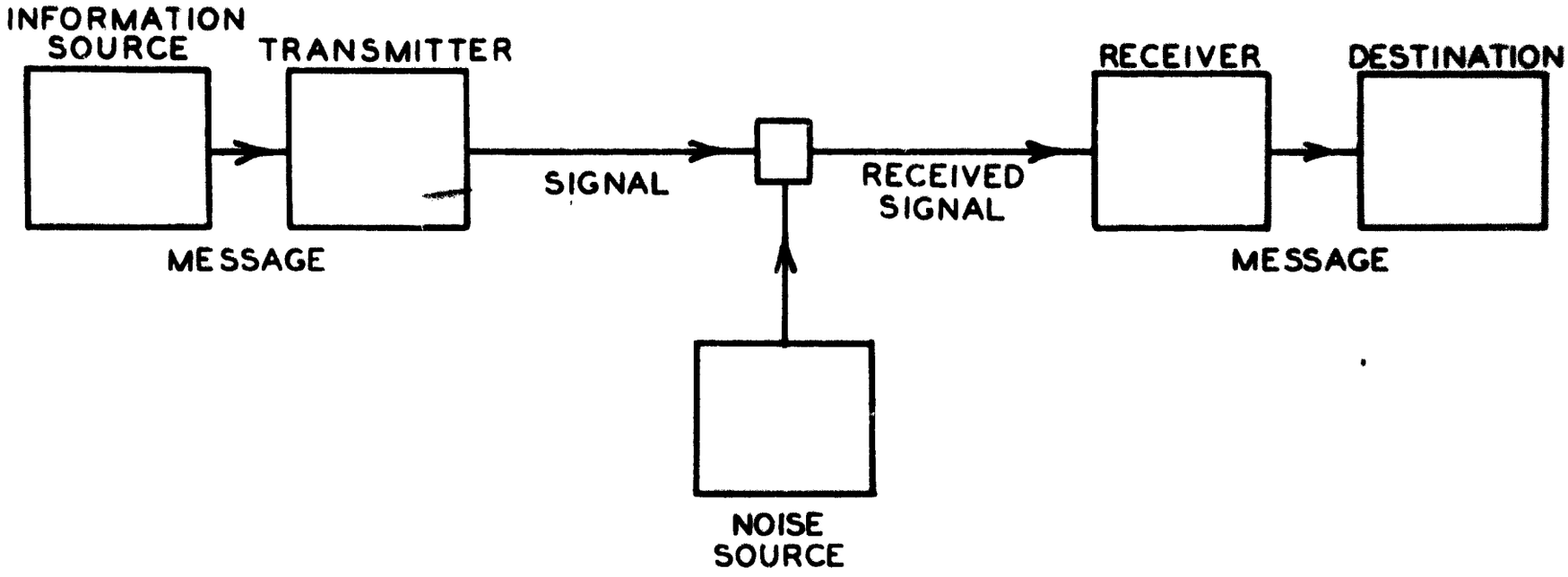

INFORMATION SOURCE
TRANSMITTER
SIGNAL
RECEIVED SIGNAL
RECEIVER
DESTINATION
MESSAGE
MESSAGE
NOISE SOURCE

Fig. 1

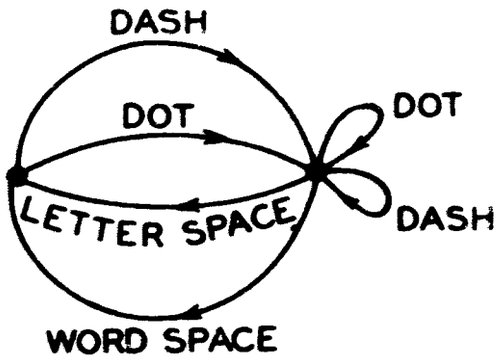


Fig. 2

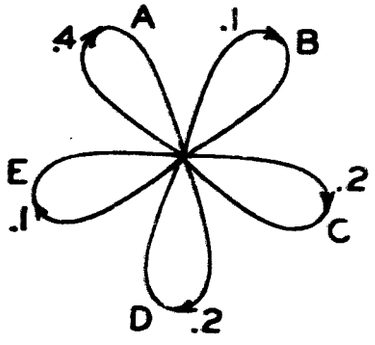


Fig. 3

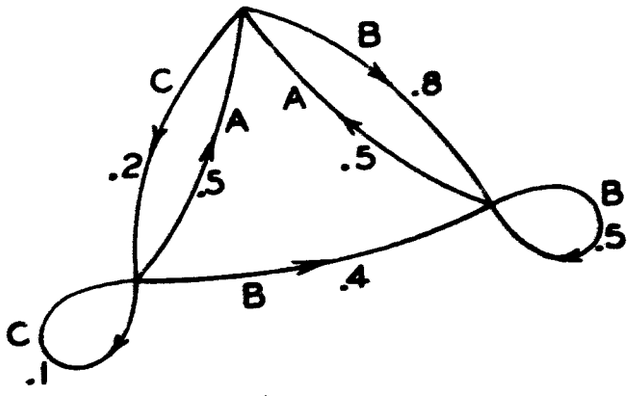


Fig. 4

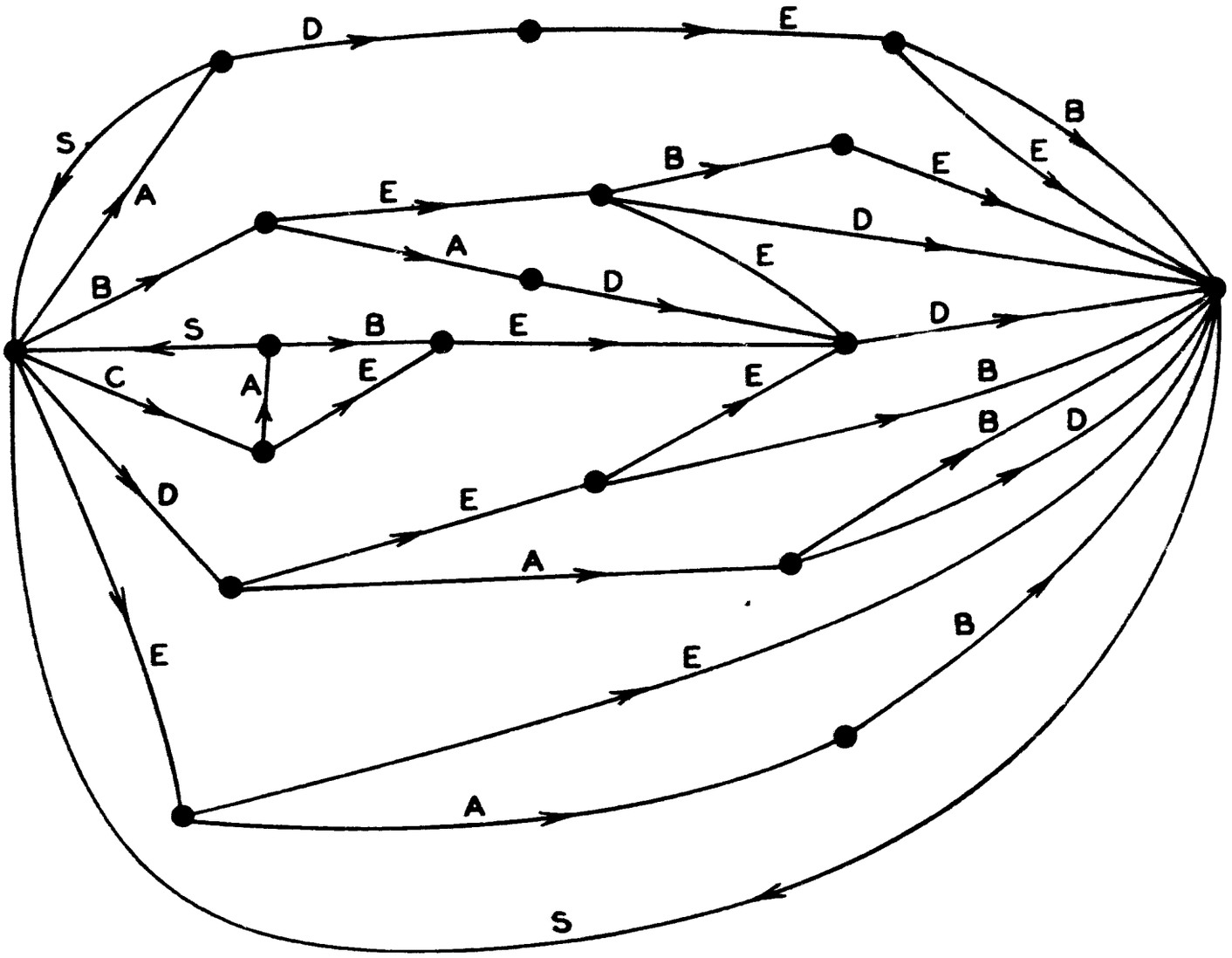


Fig. 5

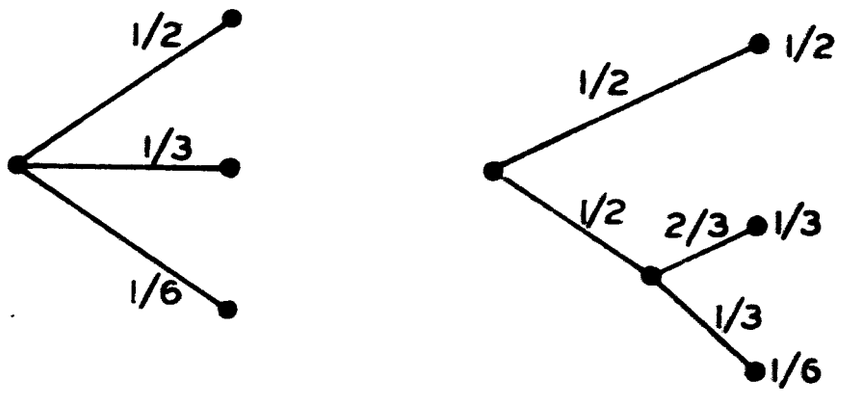


Fig. 6

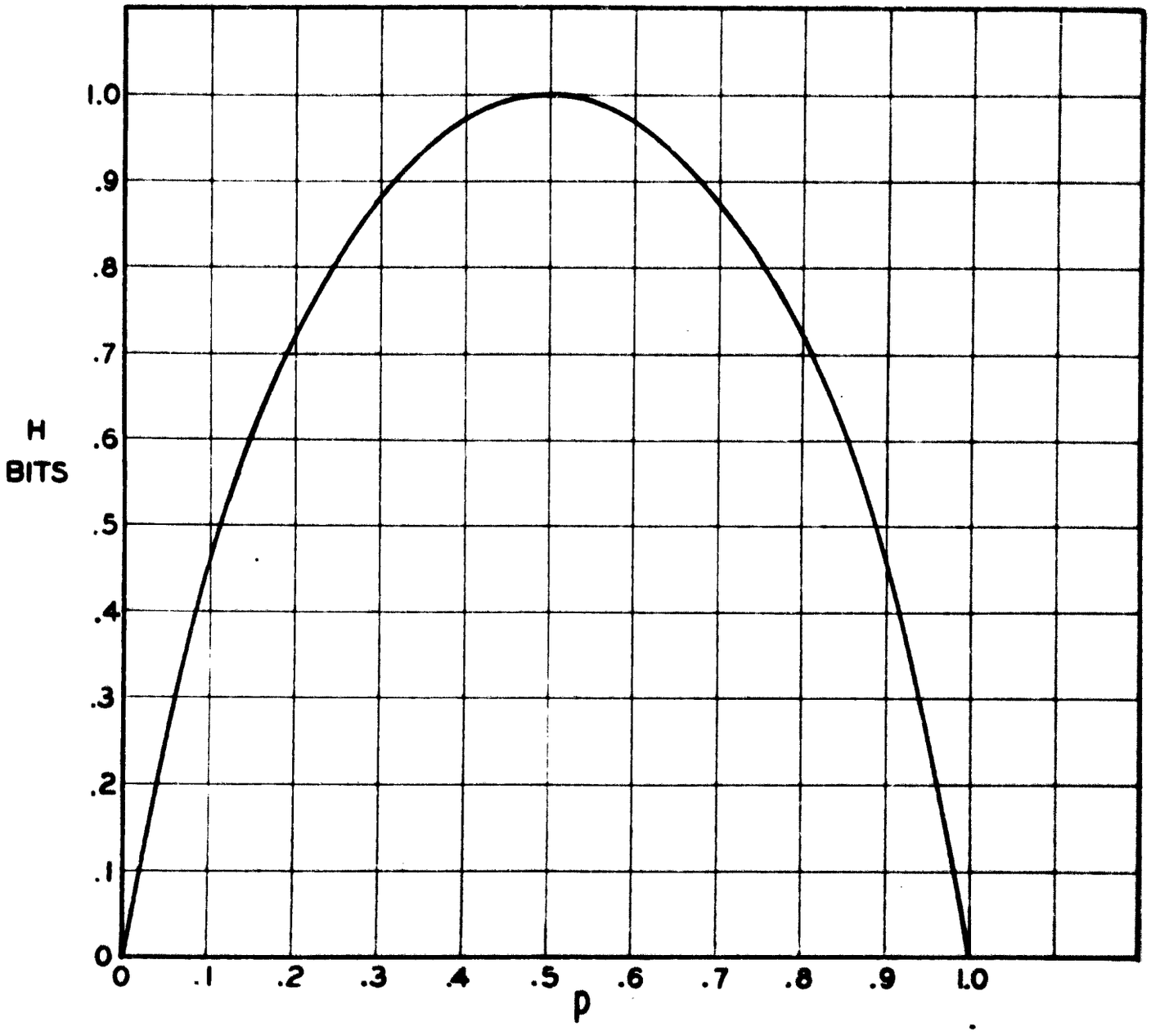


Fig. 7

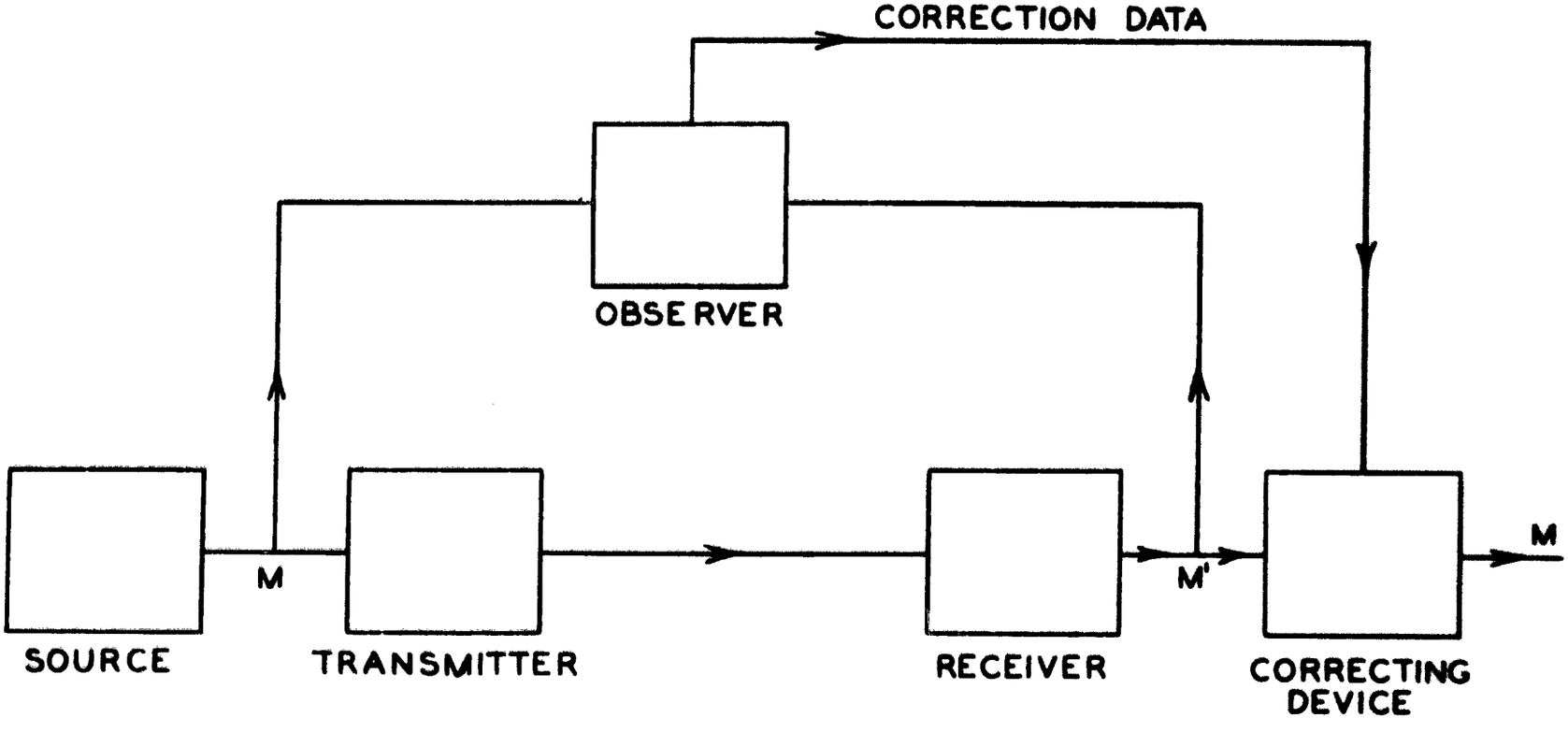

CORRECTION DATA
OBSERVER
SOURCE
M
TRANSMITTER
RECEIVER
M'
CORRECTING DEVICE
M

Fig. 8

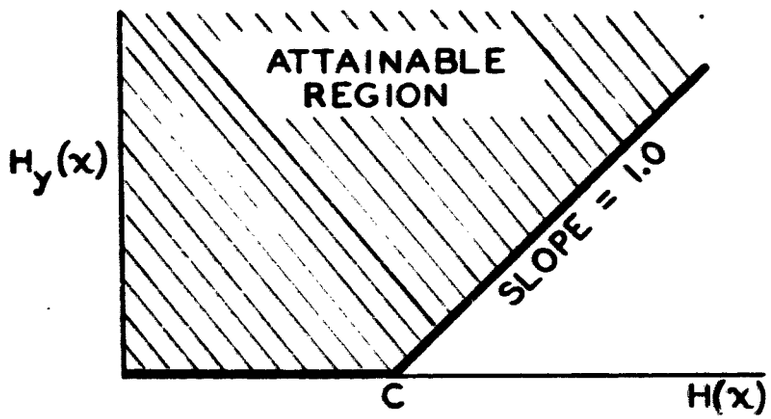


Fig. 9

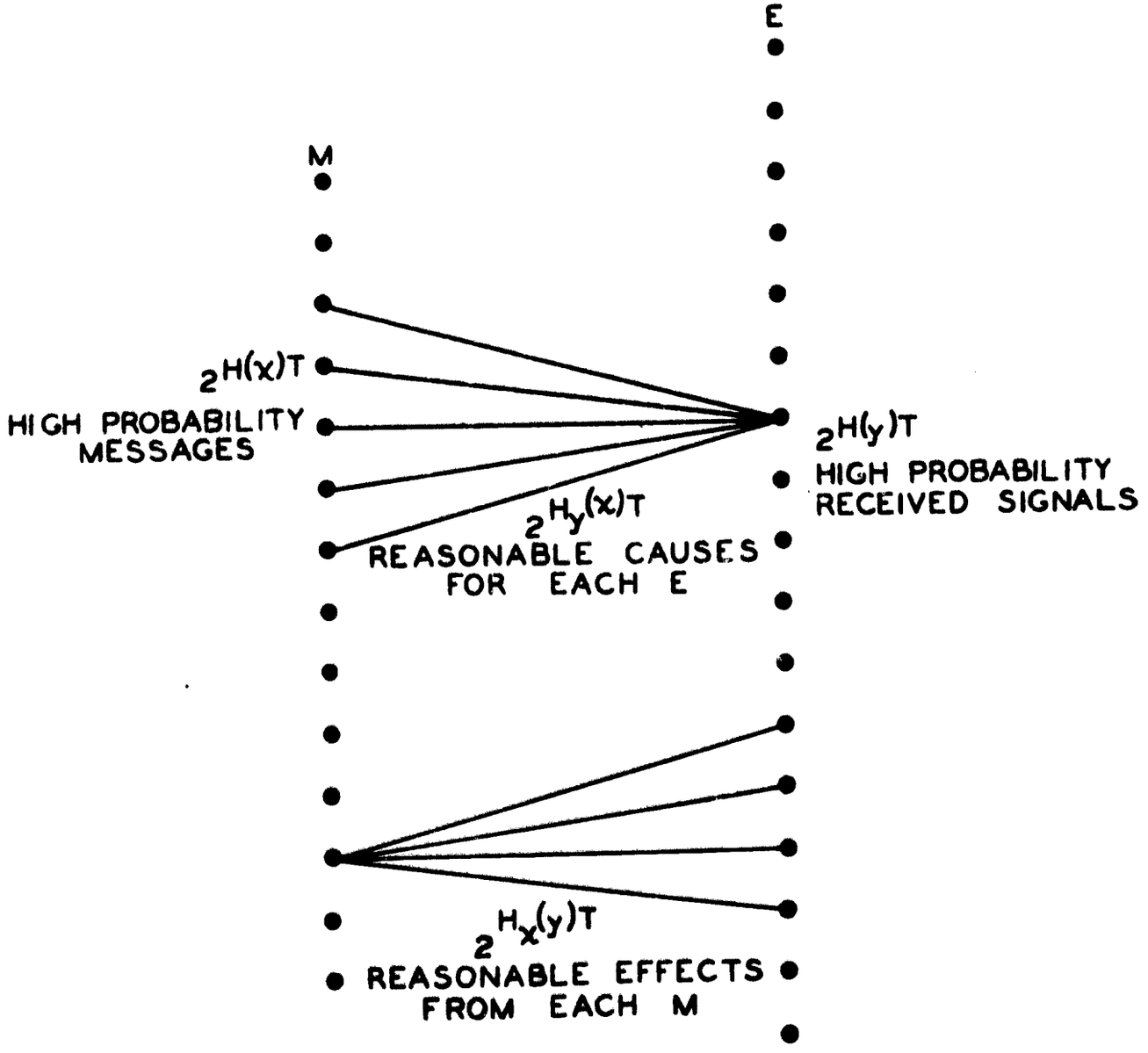

E
M
$2^{H(x)T}$
HIGH PROBABILITY
MESSAGES
$2^{H(y)T}$
HIGH PROBABILITY
RECEIVED SIGNALS
$2^{H_y(x)T}$
REASONABLE CAUSES
FOR EACH E
$2^{H_x(y)T}$
REASONABLE EFFECTS
FROM EACH M

Fig. 10

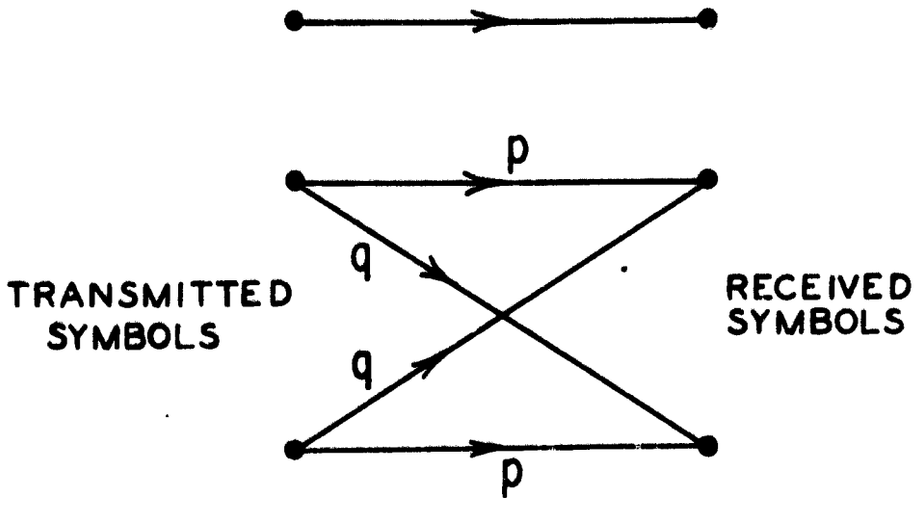
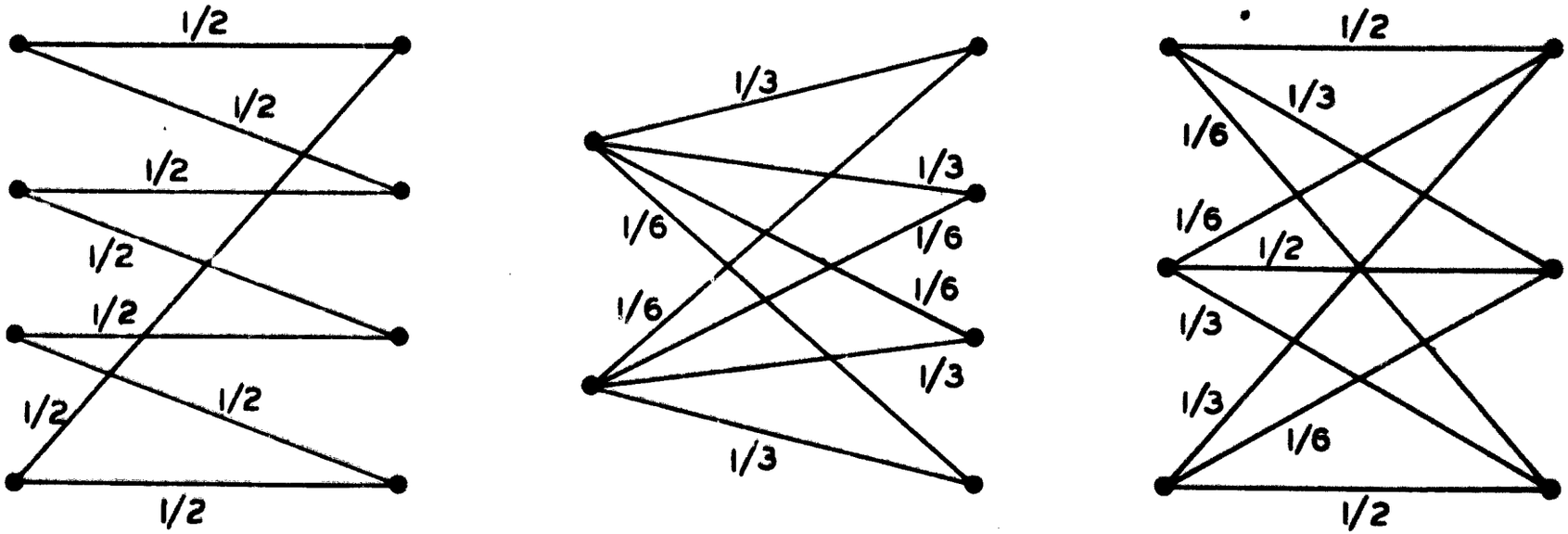


Fig. 11



a b c

Fig. 12