The Institution is not, as a body, responsible for the opinions expressed by individual authors or speakers

THE PROCEEDINGS OF
THE INSTITUTION OF ELECTRICAL ENGINEERS

EDITED UNDER THE SUPERINTENDENCE OF W. K. BRASHER, C.B.E., M.A., M.I.E.E., SECRETARY

VOL. 99. PART III (RADIO AND COMMUNICATION ENGINEERING), No. 58. MARCH 1952

621.391 : 519.21 : 530.162

Paper No. 1239
RADIO SECTION

INFORMATION THEORY AND INVERSE PROBABILITY IN TELECOMMUNICATION

By P. M. WOODWARD, B.A., and I. L. DAVIES, M.A., Graduate.

(*The paper was first received* 11*th October, and in revised form* 10*th December*, 1951.)

SUMMARY

The foundations of information theory are presented as an extension of the theory of inverse probability. By postulating that information is additive and taking suitable averages, all the essential definitions of Shannon's theory for discrete and continuous communication channels, with and without noise, are obtained. The theory is based on the idea that receiving a communication, or making an observation, merely changes the relative probabilities of the various possible messages. The whole process of reception can therefore be regarded as a means of evaluating *a posteriori* probabilities, and this leads to the idea that the optimum receiver in any telecommunication problem can always be specified, in principle, by inverse probability. The simplest instance is the correlation receiver for detecting very weak signals in the presence of noise, and its theory is briefly discussed The paper concludes with an answer to possible criticisms of the use of inverse probability.

(1) INTRODUCTION

One of the fundamental concepts of physics is "uncertainty," not only because of Heisenberg's principle, but because of random heat motion which gives rise to much of the noise in electronic systems. Whenever it is necessary to work with communication signals which are not very large compared with a background of noise, a study of uncertainty becomes important. But in telecommunication, uncertainty plays an even more fundamental part, for it can be argued that the whole object of a communication system is to remove uncertainty at the receiving end. In fact, Shannon's theory of communication[1] is based entirely on the concept that information is the opposite of uncertainty. Thus, in any communication system there is a conflict between the engineer's attempts to remove uncertainty and the opposing tendency of natural phenomena towards randomness.

The language of uncertainty is the theory of probability, of which the theory of information must be regarded as an extension. When Boltzmann first related entropy to probability, he opened up a whole field of statistical mechanics and—unwittingly—of communication theory. This new application of an old idea crystallizes very precisely the intuitive notions of communication engineers.

A history and bibliography of information theory has already been published,[2] and further historical remarks would here be redundant. The object of the present paper is twofold. First, it may serve as a detailed introduction to Shannon's theory,[1] although it will not deal at all with problems of coding. The method of setting up a definition of information here differs from Shannon's, for it conforms more closely to the methods used by Boltzmann. An attempt is also made to present the foundations of the subject in a form which can at every stage be applied equally well to problems involving discrete and continuous probability distributions. For the purpose of exposition, Shannon treats the discrete case first, but his procedure introduces certain difficulties of subsequent generalization which are avoided in the present paper.

In the second place, the authors wish to draw further attention to the theory of inverse probability[3] and its practical applications. Information theory emphasizes the significance of the relative probabilities of possible "messages" before and after receiving a communication. These last probabilities embody all the information in the received communication, and may be computed from it by using the theorem of inverse probability. In principle, this computation can be carried out by the receiving apparatus itself, and so we obtain the specification of an optimum device for extracting all the available signal information remaining in a waveform to which noise has been added. An indication of how this theory may be applied to simple radar and communication problems is given in the later Sections of the paper.

(2) INVERSE PROBABILITY AND COMMUNICATION

When a communication is received, the state of knowledge of the observer is obviously changed; thus, ways must first be considered of defining his state of knowledge. Now any communicable message represents a selection from a collection or ensemble of possible messages which may either be discrete or may merge continuously one into another. In communication theory, the meaning of the various possible messages does not matter and the state of knowledge before reception may therefore be specified simply by saying that each message has a

Written contributions on papers published without being read at meetings are invited for consideration with a view to publication.

Mr. Woodward and Mr. Davies are at the Telecommunications Research Establishment.

certain probability of occurrence. After reception, it is to be hoped that one particular message will have been singled out from the collection; in other words, its *a priori* probability will have been changed into an *a posteriori* probability of unity. But owing to distortion by random interference, the received communication will not always indicate the transmitted message with complete certainty, and the *a posteriori* probability will not then be entirely concentrated on one message but will be distributed amongst several. In general, therefore, unit probability is distributed amongst the messages in one way before receiving the communication and in a different way afterwards. This is the essence of communication theory. Intuitively, it may be said that any increase in the probability of the true message represents a gain in information, but before we can be more precise it is necessary to discuss the *a priori* and *a posteriori* distributions themselves, and this leads immediately into the subject of inverse probability.

The topic of inverse probability is in some respects a controversial one. This is unfortunate, because it tends to cast general suspicion on the theorem of inverse probability, which is by itself entirely uncontroversial. Perhaps the best way in which to describe this theorem is to take first a simple example, which will make it self-evident. Suppose that a simple telegraph system is used to convey one of two messages, "yes" or "no," and that these messages are represented by two different signals indicated by a green and red lamp at the receiver. Suppose further that, over a large number of occasions when a message is communicated, random interference causes a proportion of what ought to be red indications to show green, and vice versa. For the sake of generality, these proportions may be taken to be different from each other, say two-fifths of the greens becoming red and one-third of the reds becoming green. Finally, suppose that the transmitted message is "yes" more often than "no," in the proportion five to three. The whole of the data can be represented schematically as shown in Table 1.

Table 1

Yes	Yes	Yes	Yes	Yes	No	No	No
Green	Green	Green	Red	Red	Red	Red	Green

This Table is simply an enumeration of all the equally likely possibilities, the top line being the transmitted message and the second line the corresponding received indication. It enables any required probabilities to be read off; for instance, the top line alone displays the *a priori* probability distribution,

$$p(\text{yes}) = \tfrac{5}{8}, \quad p(\text{no}) = \tfrac{3}{8}$$

However, if a green indication has appeared, we must select only those possibilities shown as green in the second line. Of these, three out of four correspond to yes and the *a posteriori* probability of yes when green is received is therefore three-quarters. This is, in fact, an application of inverse probability.

Formally, the general theorem is derived as follows. Let x be the transmitted message and y the received indication. Then, by the product law for probabilities, the joint probability $p(x, y)$ that a value of x and a value of y will occur together is given by

$$p(x, y) = p(x)p_x(y) = p(y)p_y(x) \quad . \quad . \quad . \quad (1)$$

where the conditional probability $p_x(y)$ is the probability of a value of y given the value of x. These important relationships may easily be verified from Table 1; thus when x is yes and y is green, we have

$$\tfrac{3}{8} = \tfrac{5}{8} \times \tfrac{3}{5} = \tfrac{4}{8} \times \tfrac{3}{4}$$

Of the five probabilities in eqns. (1), two are of especial importance in communication theory, namely $p(x)$, which is the *a priori* probability of a message x, and $p_y(x)$, which is the *a posteriori* probability of x upon receiving y. At the receiver, all values of x have in general to be considered, so $p(x)$ and $p_y(x)$ are often described as distributions of probability. Their sums over x are, of course, equal to unity. The theorem of inverse probability is simply an expression for the *a posteriori* distribution obtained from eqn. (1), namely

$$p_y(x) = \frac{p(x)p_x(y)}{p(y)} \quad . \quad . \quad . \quad . \quad . \quad (2)$$

The observer is presumed to know the *a priori* probabilities $p(x)$ and to know the statistical properties of the interfering noise, specified by $p_x(y)$, which represents a whole family of distributions. Upon receiving an indication y he can use eqn. (2) to assess the relative probabilities that each message x was the transmitted one. Since y will then be fixed, $p(y)$ is a constant which may be evaluated by ensuring that $p_y(x)$ is normalized, i.e. that its sum over x is unity. Eqn. (2) may therefore be written

$$p_y(x) = kp(x)p_x(y) \quad . \quad . \quad . \quad . \quad . \quad (3)$$

where k is a constant independent of x. Statistically, $p_y(x)$ represents the relative frequency with which x is actually transmitted in a large number of communications which give the same indication y at the receiver. For instance, when green is indicated,

$$p_{\text{green}}(\text{yes}) = k \times \tfrac{5}{8} \times \tfrac{3}{5} = \tfrac{3}{8}k$$

$$p_{\text{green}}(\text{no}) = k \times \tfrac{3}{8} \times \tfrac{1}{3} = \tfrac{1}{8}k$$

The *a posteriori* distribution is normalized by putting k equal to 2, and becomes

$$p_{\text{green}}(\text{yes}) = \tfrac{3}{4}, \quad p_{\text{green}}(\text{no}) = \tfrac{1}{4}$$

which was obtained by inspection earlier. Similarly, when red is indicated, the *a posteriori* distribution is

$$p_{\text{red}}(\text{yes}) = \tfrac{1}{2}, \quad p_{\text{red}}(\text{no}) = \tfrac{1}{2}$$

It should be clear that in general there will be a different *a posteriori* distribution for every different received indication.

In many problems, the various received indications will not be discrete and finite in number, but members of a continuous ensemble. This could easily happen if y were a voltage at some fixed instant of time. The theorem of inverse probability must then be interpreted in a slightly different way. If y is a member of a continuous ensemble, its probability distribution becomes a continuous distribution of probability density, $p(y)dy$ being the probability of finding y in the interval $(y, y + dy)$. However, upon substituting such probabilities into eqn. (2), the dy's cancel out and eqn. (3) thus remains valid when $p_x(y)$ is a probability density. It is also valid if x is continuous, for then dx cancels out also. The theorem of inverse probability may therefore be used in the form of eqn. (3) whether x or y or both are continuous variables, probability densities being used wherever they are appropriate.

It has been seen that each possible received indication gives rise to its own *a posteriori* message probability distribution. If these distributions are averaged according to their probability of occurring, namely $p(y)$, it is of interest to note that the *a priori* distribution is regained. Thus integration (or summation) of eqn. (1) over y gives

$$\int p(y)p_y(x)dy = \int p(x)p_x(y)dy = p(x) \quad . \quad . \quad (4)$$

The left-hand side is the average *a posteriori* distribution and the right-hand side is the *a priori* distribution. In an abbreviated notation, the average over all values of some variable z is denoted by Av_z and eqn. (4) then becomes

$$\text{Av}_y p_y(x) = p(x) \quad . \quad . \quad . \quad . \quad . \quad (5)$$

In the example, it will be seen from Table 1 that the two received indications, green and red, happen to occur equally often in the total ensemble. Consequently an average, with equal weights, of the two *a posteriori* distributions $(\frac{3}{4}, \frac{1}{4})$ and $(\frac{1}{2}, \frac{1}{2})$ yields the *a priori* distribution $(\frac{5}{8}, \frac{3}{8})$.

In applying inverse probability to a communication problem, the first major step is the evaluation of $p_x(y)$, which expresses the statistical properties of the noise in their most relevant form. This will often be more complicated than might appear from the above account, for y will not always be a simple received indication such as red or green, nor even a voltage which can assume any value selected from a continuous range of amplitudes. In general, y will be a complete waveform, such as a morse signal or a radar echo, plus noise. However, this does not raise any conceptual difficulty, for it is possible to treat a waveform as a sequence of voltage ordinates and to evaluate $p_x(y_1, y_2, y_3 \ldots)$ as a joint distribution in many variables, as described in more detail in Section 4.

(3) INFORMATION THEORY

It is now widely known that information may be so defined that a definite choice between two equally likely events represents one binary unit of information, and a choice between n equally likely events $\log_2 n$ binary units, or "bits." Complete certainty of the result is assumed in this definition, but frequently the effect of a communication is merely to change the relative probabilities of a number of events without singling one out as being certain. It is therefore necessary to develop a more general definition which will reduce to the simpler definition when there is no *a posteriori* uncertainty. This has been done in a classic paper by Shannon.[1] The present treatment is based on that of Shannon but differs in its approach and initial postulates, because actual quantities of information are first considered rather than average quantities or mean rates. However, Shannon's definitions are readily deduced and some of the principal results of his work are briefly mentioned.

A start is made with the following pair of axioms concerning the addition of information:

(*a*) If two communications representing the same message are sent, and the observer regards his *a posteriori* probability after the first communication as the *a priori* before the second, the total gain in information concerning this message is equal to the sum of the gains from each communication.

(*b*) If two communications representing two independent messages are sent, the total gain in information concerning them is the sum of the gains when each communication is considered separately.

From these two axioms, it is possible to develop the whole mathematical theory. Let us denote the message considered in the first axiom by x_i. Three probabilities for this message occur: first its *a priori* probability $p(x_i)$, secondly its *a posteriori* probability $p_y(x_i)$ after the first communication and, thirdly, taking $p_y(x_i)$ as the *a priori* probability for the second communication, we may denote the final probability after the second communication by $p_z(x_i)$. Suppose now that after receiving y and z the observer is completely certain that the ith message was sent, so that $p_z(x_i) = 1$. Information has been gained because the original uncertainty implied by $p(x_i)$ has been entirely removed, and the total gain from y and z therefore depends only on $p(x_i)$. Similarly, the gain from the second communication alone depends only on $p_y(x_i)$. Consequently, the information gained from the first communication depends only on $p(x_i)$ and $p_y(x_i)$. Thus, when uncertainty remains after a given message x_i has been transmitted and received in the presence of noise, the information is a function of the *a priori* and *a posteriori* probabilities of this message alone, and may be written in the form

$$J[p(x_i), p_y(x_i)] \quad . \quad . \quad . \quad . \quad . \quad (6)$$

The first axiom then states

$$J[p(x_i), p_y(x_i)] + J[p_y(x_i), p_z(x_i)] \equiv J[p(x_i), p_z(x_i)] \quad . \quad (7)$$

It is shown in the Appendix that, to satisfy this identity, J must be of the form

$$J[p(x_i), p_y(x_i)] \equiv j[p(x_i)] - j[p_y(x_i)] \quad . \quad . \quad (8)$$

In order now to determine the functional form of j, it is necessary to use the second axiom. Here the two independent messages may be denoted by x_i and x_k, and the corresponding received indications again by y and z. Since the joint probability of two independent events is the product of their separate probabilities, the second axiom gives

$$\begin{aligned} &j[p(x_i)p(x_k)] - j[p_y(x_i)p_z(x_k)] \\ &\equiv j[p(x_i)] + j[p(x_k)] - j[p_y(x_i)] - j[p_z(x_k)] \quad . \quad . \quad (9) \end{aligned}$$

From this identity it is shown in the Appendix that $j(p)$ must be of the form

$$j(p) = A \log p + B \quad . \quad . \quad . \quad . \quad (10)$$

where A and B are constants which may be chosen arbitrarily. Thus from eqns. (6), (8) and (10), the quantity of information J may be written in the form

$$J[p(x_i), p_y(x_i)] = - A \log \frac{p_y(x_i)}{p(x_i)} \quad . \quad . \quad (11)$$

and in order to make an increase in the probability of the true message represent positive information, A is made equal to -1. Finally, it is convenient to simplify the notation by writing $I_{x,y}$ instead of J, to indicate that it is the quantity of communicated information when a given x is transmitted and a given y received. Thus, with the help of eqn. (1),

$$I_{x,y} = \log \frac{p_y(x)}{p(x)} = \log \frac{p(x, y)}{p(x)p(y)} \quad . \quad . \quad . \quad (12)$$

When the logarithmic base is 2, the unit of information is called a "bit"; when it is ϵ, it is called a "natural unit." This is the basic expression for a quantity of information which is implicit in Shannon's theory.

Unfortunately, when a communication, rendered ambiguous by random interference is received, the observer, not realizing which is the true message, will be unable to express his gain of information so simply. Furthermore, from the observer's point of view it would seem intuitive that the same received y should always represent the same quantity of information, regardless of the message actually transmitted. The only way, therefore, to define the observer's gain of information is to average $I_{x,y}$ over all the situations in which y alone is fixed. Now it has been shown in Section 2 that on these occasions the transmitted messages x occur with relative frequencies given by $p_y(x)$. Thus the observer's gain of information I_y may be defined by averaging $I_{x,y}$ with $p_y(x)$ as a weighting factor, giving

$$I_y = \text{Av}_x I_{x,y} = \sum_x p_y(x) \log \frac{p_y(x)}{p(x)} \quad . \quad . \quad . \quad (13)$$

if the messages are discrete.

Often, instead of being discrete as considered so far, the messages form a continuum, e.g. ranges of aircraft or meter readings. The probability distributions in x are then continuous curves of probability density, but the theory can without any difficulty be extended to cover this. If $p(x)$ is a density distribution, $p(x)\delta x$ is the probability that x lies between x and $x + \delta x$. When the range of x is split into cells of width δx, the discrete theory may be applied and then, by letting δx approach zero,

$$I_y = \int p_y(x) \log \frac{p_y(x)}{p(x)} dx \quad . \quad . \quad . \quad . \quad (14)$$

It should be noted that the gain of information given by eqns. (13) or (14) is additive in the sense of axiom (*b*) but not of axiom (*a*). This is because the observer's gain of information is an *a posteriori* average, and averages after the first and second communications are evaluated over different ensembles. As might be expected, it can be shown that I_y (unlike $I_{x,y}$) is never negative and is equal to zero only when the distributions $p(x)$ and $p_y(x)$ are identical, i.e. when the communication leaves the observer's state of knowledge completely unchanged.

The formulae obtained above can be illustrated by applying them to the example given in Section 2. If "yes" is sent and green received, the information communicated is given by eqn. (12) and is

$$\log p_{\text{green}}(\text{yes}) - \log p(\text{yes}) = \log \tfrac{3}{4} - \log \tfrac{5}{8} = \log \tfrac{6}{5}$$

Similarly, if "no" is sent and green again received, the value of $I_{x,y}$ is

$$\log p_{\text{green}}(\text{no}) - \log p(\text{no}) = \log \tfrac{1}{4} - \log \tfrac{3}{8} = \log \tfrac{2}{3}$$

This is negative because the probability of "no," which was actually sent, has diminished at the receiver as the result of the communication. Neither of the above expressions is of much value to the observer, however, for his only knowledge of what was sent is the *a posteriori* distribution

$$p_{\text{green}}(\text{yes}) = \tfrac{3}{4},\ p_{\text{green}}(\text{no}) = \tfrac{1}{4}$$

His gain of information upon receiving green is therefore given by

$$I_{\text{green}} = \tfrac{3}{4} \log \tfrac{6}{5} + \tfrac{1}{4} \log \tfrac{2}{3} = 0{\cdot}0510 \text{ bits}$$

which is an example of eqn. (13). In exactly the same way, his gain upon receiving red is

$$I_{\text{red}} = \tfrac{1}{2} \log \tfrac{4}{5} + \tfrac{1}{2} \log \tfrac{4}{3} = 0{\cdot}0466 \text{ bits}$$

Both I_{green} and I_{red} are inevitably positive.

The quantities of information evaluated above apply to particular received communications, but when the information-handling capacity of a communication channel is required, the important quantity is the mean information I, in which neither x nor y is specified. This is the quantity used by Shannon, and it is obtained by taking the average of the gains given by each received communication, weighted according to its probability of occurring. In the example, red and green happen to occur equally often (see Table 1), and therefore

$$I = \tfrac{1}{2}(0{\cdot}0510 + 0{\cdot}0466) = 0{\cdot}0488 \text{ bits}$$

The general expression is obtained by applying the operator Av_y to eqn. (14), giving

$$I = \text{Av}_y I_y = \int p(y) \int p_y(x) \log \frac{p_y(x)}{p(x)} dxdy \quad . \quad . \quad (15)$$

This is the same as the average of $I_{x,y}$ over all x and all y, which from eqn. (12) gives the symmetrical form

$$I = \text{Av}_{x,y} I_{x,y} = \int\int p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy \quad . \quad (16)$$

Yet another form of I, which is more directly physical in its interpretation, is obtained by first separating the logarithm in eqn. (15) into two parts, thus

$$I = \text{Av}_y \int p_y(x) \log p_y(x) dx - \int\int p(y) p_y(x) \log p(x) dxdy$$

The integration with respect to y may be carried out by means of eqn. (4), whence

$$I = H(x) - H_y(x) \quad (17)$$

where

$$H(x) = -\int p(x) \log p(x) dx \quad . \quad . \quad . \quad (18)$$

$$H_y(x) = -\text{Av}_y \int p_y(x) \log p_y(x) dx \quad . \quad . \quad (19)$$

Shannon describes $H(x)$ as the entropy of the distribution $p(x)$, and eqn. (17) then states that the mean quantity of information per communication is the difference between the *a priori* and *a posteriori* entropies. From the symmetry of eqn. (16), it will be seen that I may also be written in the reciprocal form

$$I = H(y) - H_x(y) \quad . \quad . \quad . \quad . \quad . \quad (20)$$

by interchanging x and y.

It is from this last expression that Shannon obtains his Theorem 17—one of the most important results of communication theory. Briefly, it states that the average quantity of information which can be conveyed in a time T and bandwidth W, in the presence of white Gaussian noise of mean power N, can approach but may never exceed

$$I_{max} = WT \log \left(1 + \frac{P}{N}\right) \quad . \quad . \quad . \quad . \quad (21)$$

where P is the mean received signal power. It is a theorem of great significance, for it shows how the capacity[1] of a communication channel is fundamentally limited. Usually the noise power increases in proportion to the bandwidth and it is possible to write $N = WN_0$ where N_0 is the mean noise power per unit bandwidth. The value of I_{max} then increases with W, but only to the limit

$$\lim_{W \to \infty} I_{max} = PT/N_0 = E/N_0 \text{ natural units} \quad . \quad . \quad (22)$$

where E is the total received signal energy. Eqns. (21) and (22) enable us, in principle, to measure the efficiency of any communication system either in the presence or absence of a bandwidth limitation. It is, of course, extremely difficult to evaluate in this way the efficiency of a complete system conveying, for instance, music from a studio to a radio listener, but such systems as pulse code modulation (P.C.M.), pulse position modulation (P.P.M.) and radar, are capable of comparatively simple analysis. It has been shown[4] that P.C.M. and P.P.M. require about 8 db more power than an ideal system satisfying eqn. (21), whilst at its best the measurement of range by radar[5] comes very close to the ideal of eqn. (22).

One of the features of the above approach to the fundamental definitions of information theory is that all the formulae apply equally well to discrete probabilities and to continuous distributions of probability density, provided that sums are suitably replaced by integrals. Fundamentally, this is because the logarithms involve only probability (density) ratios, which always remain dimensionless. Merely by postulating that information is additive, it has been shown that information can be measured as a logarithmic change of probability. The formal definition of $I_{x,y}$ in eqn. (12) is of little direct value, however, for it presupposes knowledge both of the transmitted message x and the received indication y. When an observer has received a communication, he is more interested in the average of $I_{x,y}$ over all the messages which from his point of view might have been transmitted. The resulting quantity has been called the observer's gain of information and denoted by I_y. Finally, I_y may be averaged over all values of y to give the average quantity of information per communication, denoted by I. This can be written in a variety of forms, one of which [eqn. (17)] is an expression for reduction of entropy. The concept of entropy is of course borrowed from thermodynamics, and in two recent papers Brillouin[6, 7] gives an interesting discussion of the identity, apart from a factor of Boltzmann's constant, between the notions of information in communication theory and negative entropy in physics. In this connection, eqn. (22) is perhaps the most significant result of Shannon's theory.

(4) NOISE AND THE *A POSTERIORI* DISTRIBUTION

In order to apply any of the foregoing theory to a practical problem, it is necessary to consider how the *a posteriori* distribution is constructed from the received communication y. It is presumed that the *a priori* distribution $p(x)$ is known and the *a posteriori* distribution $p_y(x)$ may then be obtained directly from eqn. (3) once $p_x(y)$ has been evaluated. The conditional probability $p_x(y)$ describes the effect of noise on the system. It represents the unpredictable nature of the received communication when the transmitted message is fixed. If the received communication is a waveform $y(t)$, consisting of the sum of a signal waveform $u_x(t)$, representing a message x, and a white Gaussian noise waveform, the evaluation of $p_x(y)$ is fairly straightforward and will now be described.

If $u_x(t)$ is the waveform which would be received in the absence of noise, the probability distribution for the resultant waveform $y(t)$, when noise has been added, is of the form

$$p_x(y) = G(y - u_x) \qquad . \quad . \quad . \quad . \quad (23)$$

where $G(n)$ is the probability density for a noise waveform $n(t)$. A simple way of formulating G precisely is by means of waveform-sampling analysis, which for the sake of completeness is here briefly summarized.

Sampling analysis rests on a well-known mathematical theorem[1] that if a function of time $f(t)$ contains no frequencies greater than W, then

$$f(t) \equiv \sum_r f(r/2W)\,\mathrm{sinc}\,(2Wt - r) \quad . \quad . \quad . \quad (24)$$

where $\mathrm{sinc}\,x$ is an abbreviation for the function $(\sin \pi x)/\pi x$. This function occurs so often in Fourier analysis and its applications that it does seem to merit some notation of its own. Its most important properties are that it is zero when x is a whole number but unity when x is zero, and that

$$\int_{-\infty}^{\infty} \mathrm{sinc} x\, dx = 1$$

and
$$\int_{-\infty}^{\infty} \mathrm{sinc}\,(x - r)\,\mathrm{sinc}\,(x - s)\,dx = \begin{cases} 1, & r = s \\ 0, & r \neq s \end{cases}$$

r and s both being integers. The importance of the identity (24) is that it enables a continuous function of time to be specified uniquely in terms of sample values at intervals $1/2W$, where W is an arbitrary frequency greater than any which occurs in the frequency spectrum of $f(t)$.

To apply this analysis, it is necessary to assume that all the waveforms under consideration have been passed through a low-pass filter with a frequency response which is uniform up to W and zero for all higher frequencies. If W is chosen sufficiently large, the signal will be unaffected and no loss of generality results. The filter is simply a mathematical artifice for it will be seen later that the precise value of W is immaterial and eventually disappears from the analysis. It does, however, permit the use of sampling values both for signal and noise waveforms. Consider first a white Gaussian noise waveform $n(t)$; at the output of the filter each sampling value $n(r/2W)$, or more briefly n_r, has by definition the probability distribution

$$p(n_r) \propto \exp\,(-\,n_r^2/2N) \quad . \quad . \quad . \quad . \quad (25)$$

where N is the mean square value of $n(t)$, or the mean noise power. It can be shown that these noise samples are statistically independent, provided that the noise spectrum extends uniformly over all frequencies up to W, and hence the joint probability distribution for a whole set of samples is the product of each separate distribution. Since the samples determine the waveform, this product gives the probability density for the waveform itself. Thus

$$G(n) \propto \exp\,(-\,\textstyle\sum n_r^2/2N) \quad . \quad . \quad . \quad . \quad (26)$$

By squaring the fundamental identity (24), integrating over time, and using the properties of $\mathrm{sinc}\,x$, the sum in the exponent can be expressed as an integral and then

$$G(n) \propto \exp\left[-\,\frac{1}{N_0}\int n^2(t)dt\right] \quad . \quad . \quad . \quad (27)$$

Here N_0 is the mean noise power per unit bandwidth, which has the dimensions of energy and is the fundamental noise parameter.

The *a posteriori* distribution for the message x can now be written in an explicit form. From eqns. (3), (23) and (27)

$$p_y(x) = kp(x)\exp\left[-\,\frac{1}{N_0}\int (y - u_x)^2 dt\right] \quad . \quad (28)$$

is obtained as the fundamental probability equation for all reception problems in which interference is caused solely by the addition of white Gaussian noise to the signal waveform. The constant k is chosen so as to normalize $p_y(x)$; the integral in the exponent is a definite one, carried over the whole time for which the communication lasts. It will be seen immediately that, apart from the *a priori* weighting factor, the most probable message is the one whose waveform $u_x(t)$ has the least r.m.s. departure from the received waveform $y(t)$, a result which is certainly intuitive.

(5) THE CORRELATION RECEIVER

It might appear that the mathematical processes so far described are useful only for the purpose of calculating the information content of any given communication, but the probability equation (28) has a much wider significance. In principle, it shows how to design an optimum receiver for extracting all the available information from the mixture of signal and noise at the receiver input. The concept upon which this statement rests should now be obvious, for $p_y(x)$ is itself the available information. All that can be reasonably demanded of a receiver is that it shall enable an observer to gauge the relative probabilities that each possible message is the true one. If the receiver actually computes these probabilities, no further problem of interpreting the received waveform remains. Viewed in this way, eqn. (28) specifies mathematically the ideal receiver.

The consequences of this statement will now be examined. Suppose that the integral which occurs in the exponent is expanded thus:

$$\int y^2(t)dt - 2\int y(t)u_x(t)dt + \int u_x^2(t)dt$$

Upon reception, y^2 is fixed and therefore independent of x, which can be regarded as the message under test. The first integral is thus a constant multiplying factor in $p_y(x)$ and can be absorbed into k. In many problems, though by no means all, the integral of u_x^2 will also be independent of x, because the various message waveforms all have the same energy. If so, this term also can be struck out. Eqn. (28) then becomes

$$p_y(x) = kp(x) \exp\left[2q(x)/N_0\right] \quad . \quad . \quad (29)$$

where

$$q(x) = \int y(t)u_x(t)dt \quad . \quad . \quad . \quad (30)$$

To start with, q must be formed by multiplying the received waveform $y(t)$ by each of the possible message waveforms in turn, and integrating with respect to time. This is the heart of eqn. (29) and is an evaluation of the correlation between y and u_x. The function $q(x)$ will tend to be greatest when $u_x(t)$ is the waveform corresponding to the message actually transmitted. The operation of forming q from y is usually an irreversible one, and is in a sense the key operation, because irreversible processes in general destroy information, whilst this particular one does not. It merely destroys some unwanted information about the noise component of $y(t)$.

If all that is finally required is to determine the most probable message, and if the *a priori* probabilities of the messages are all equal, it is not necessary to evaluate anything more than $q(x)$. The rest of eqn. (29) is merely an amplitude distortion of $q(x)$ and, since the distortion is monotonic, the value of x which makes $p_y(x)$ a maximum is the value which makes $q(x)$ a maximum. Nevertheless, the factor $2/N_0$ and the exponential are interesting. The factor $2/N_0$ scales $q(x)$ in such a way that, when the noise is small, the exponential function vastly exaggerates the variations of $q(x)$ with x. This is precisely what would be expected, for when the noise is small there ought to be little doubt as to which is the true message, and $p_y(x)$ ought therefore to show a pronounced maximum representing a high degree of certainty. In the opposite extreme, if the noise is so large that it swamps the signal completely, it will be found that the factor $2/N_0$ makes the exponent very small compared with unity. Thus $p_y(x)$ simply reproduces $p(x)$, and eqn. (14) shows that there is no gain of information. It will be seen that the division between these two extremes of behaviour does not occur when the mean signal power P is roughly equal to the noise power N, but when the total signal energy E is comparable with the noise power per unit bandwidth N_0. In fact P might be very much smaller than N, and herein lies the great virtue of a correlation receiver. The effect is also seen directly from eqn. (22), where it is the value of E/N_0 which determines the maximum quantity of information in an ideal system. However, it should not be thought that correlation is anything more than what is commonly called an integration technique.

It would appear that in practice, correlation methods can be applied only to the very simplest systems. It has been tacitly assumed that the various message waveforms $u_x(t)$ are precisely known before reception, but radiocommunication is usually complicated by the fact that the time-origin of the signal is not known in advance. The theory then becomes more complicated, because the waveform corresponding to a message x has to be written in the form $u_x(t-\tau)$, where τ is an unknown and irrelevant time-delay. The *a posteriori* distribution then becomes an integral over all possible time-delays, thus

$$p_y(x) = k\int p(x, \tau) \exp\left[\frac{2}{N_0}\int y(t)u_x(t-\tau)dt\right]d\tau \quad . \quad . \quad (31)$$

This expression means that the received waveform must first be cross-correlated with all the possible message waveforms at all possible times, and in all but the simplest systems this would lead to an impossible degree of practical complication.

An interesting variation occurs in radiolocation, however, where the message (target range) is represented by the time-delay itself and the waveform is otherwise fixed. It is assumed for simplicity here that there is only one target and that the echo from it is of known strength, independent of range. The *a posteriori* distribution for the time-delay is then given by

$$p_y(\tau) = kp(\tau) \exp\left[\frac{2}{N_0}\int y(t)u(t-\tau)dt\right]. \quad . \quad . \quad (32)$$

In this formula, $y(t)$ and $u(t)$ are radio-frequency waveforms and, apart from the question of limits discussed more fully in another paper,[8] the integral has the form of the output from a linear filter. The input is the received waveform $y(t)$ and the impulsive response is $u(-t)$, the time-reverse of the transmitted waveform. Such a filter, in effect, cross-correlates y and u at the radio frequency, which indeed is what a conventional receiver does, a pulse at a time. The need for detection is not apparent from eqn. (32), because in theory it is unnecessary when determining the range of a stationary target; it only destroys the fine-structure range information obtainable from the carrier. The further implications of eqn. (32) have been discussed more fully elsewhere[5, 8] and need not be pursued here.

In order to illustrate eqn. (32) when the waveforms are all at

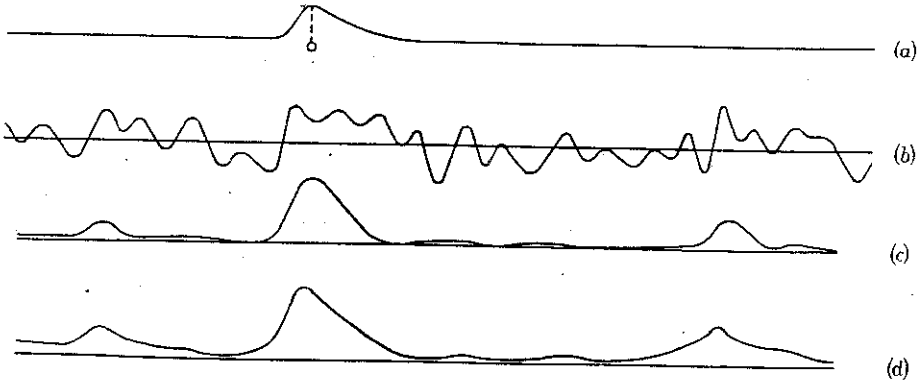

Fig. 1.—An experiment in pulse location.

(*a*) Signal.
(*b*) Signal plus noise.
(*c*) *A posteriori* probability distribution, guessed.
(*d*) *A posteriori* distribution, calculated by inverse probability.

low frequency, a simple experiment has been carried out. Fig. 1(*a*) represents a signal waveform, and a numerical message x is supposed to be encoded by making the time-delay τ of this waveform equal to x. Fig. 1(*b*) illustrates a typical waveform y which might result after the addition of noise, filtered through an arbitrary band W as described earlier. After having been told that the pulse should be considered *a priori* to lie anywhere within the trace with equal probability, an observer was asked to assess, purely by inspection, the *a posteriori* probability distribution for its position. He was shown the shape and amplitude of the pulse, but kept in ignorance of its true position. His subjective curve is shown in Fig. 1(*c*), while the theoretical curve calculated from eqn. (32) is shown in Fig. 1(*d*). If the signal/noise ratio in Fig. 1(*b*) were increased, Fig. 1(*d*) would approach a delta function. An ideal receiver would convert Fig. 1(*b*) into Fig. 1(*d*) electronically, and so indicate the relative probabilities of all possible messages without the need for guesswork.

(6) CONCLUSION AND DISCUSSION

It has been shown that of all probability distributions which may be used to describe the statistical behaviour of telecommunication systems, two are of special significance. They are the distributions of probability amongst the various possible "messages" before and after receiving a communication. Reception may be treated as an event which changes the relative probabilities of the messages, and information theory provides a numerical measure of this change. These concepts lead naturally to the somewhat novel idea that the whole physical process of reception is simply a means of evaluating, or at least of making evident, the *a posteriori* message probabilities. Hence the formula for the *a posteriori* distribution is in itself the specification of an optimum receiver. Rather strangely, the theory of inverse probability seems to have been little used in the analysis of signal and noise problems before the advent of information theory, yet it represents very closely not only the electronic mechanism of reception but also the human mechanism, as is well illustrated by the example shown in Fig. 1.

It will have been observed that in all the expressions for *a posteriori* probability, the *a priori* distribution appears as a weighting factor. This is a fundamental feature of inverse probability and it may give rise to difficulty. In statistics, it frequently happens when an attempt is made to decide between a number of hypotheses (the messages) in the light of some new data (the communication) that there is no obvious way to weight the hypotheses *a priori*. Although this difficulty would scarcely arise in communication systems, it will often occur in observational systems such as radar, where, in circumstances quite different from any which have previously been encountered, it might be necessary to ascribe an *a priori* probability distribution to the range of a target. If there are no statistics on which to base an *a priori* distribution, how can it be defined at all? Here it may sometimes be tempting to postulate some non-committal function[3] simply as "a formal way of expressing ignorance." However, this is at best a somewhat arbitrary procedure, and the objection to it is well expressed by Bartlett[9] who writes, "This substitution of a simple function for a prior probability, which, if it could be evaluated at all, would certainly need all the data d to be enunciated, gives the posterior probability in an exact form which is highly misleading. Moreover, the formula, in trying to make our final inference about a parameter [the message] for us and give us the exact probability of every possible value, is compelled to mix the information which can be got out of a sample [the communication] with what other knowledge or notions we may have." Bartlett here makes two different points and it is the second one which calls particularly for an answer. It is necessary to distinguish the two ways in which inverse probability is used in the present paper. First, it is used as a step in the process of measuring a gain of information in the Shannon sense. Now, although the gain is brought about by the communication itself, its magnitude is found to depend on the *a priori* distribution. Unless an *a priori* distribution can be formulated it is impossible to determine the extent to which the communication merely duplicates existing knowledge. This does not, of course, justify the use of purely subjective *a priori* distributions in Shannon's theory, which is based on the frequency definition of probability, but it does justify the use of *a priori* probabilities whenever prior statistics are available. Secondly, it is suggested in this paper that inverse probability provides a method of specifying the ideal receiver. However it is possible to qualify this statement. An ideal receiver need not complete the computation of the *a posteriori* distribution so long as it is readily obtained from the output. Thus the troublesome *a priori* factor might well be omitted from the receiver specification when it is doubtful and in practice supplied by the observer subjectively.

It may well be asked why it should be necessary for an ideal receiver to compute the *a posteriori* probabilities of every possible message rather than select the one which is most probable, since this is usually all that will be required in practice. Indeed, many communication systems would be vastly complicated if a (possibly multi-dimensional) *a posteriori* distribution were sought at the receiving end, but it is not always necessary to interpret the theory so literally. The *a posteriori* distribution can, if desired, be regarded simply as a means of determining the most probable message. A receiver can thus be designed to give as its output the value of x which makes $p_y(x)$ a maximum. But sometimes there are objections to this procedure. First, the force of Bartlett's criticism is much enhanced, because the *a priori* distribution is used in an irreversible manner. Secondly, the reliability of the communication ceases to be evident to the observer. And, finally, it may happen in some problems such as radar that further signals concerning the same message will become available after an interim observation. The interim *a posteriori* distribution is then required as the *a priori* distribution for the further communication. It has been shown elsewhere[8] that this is precisely what happens when pulse-to-pulse summation is performed in a radar system. Premature selection of the optimum message makes it impossible to combine several communications in this way, and destroys useful information irretrievably.

(7) ACKNOWLEDGMENT

Acknowledgment is made to the Chief Scientist, Ministry of Supply, and to the Controller, H.M. Stationery Office, for permission to publish the paper.

(8) REFERENCES

(1) SHANNON, C. E.: "A Mathematical Theory of Communication," *Bell System Technical Journal*, 1949, **27**, p. 379 and p. 623.

(2) CHERRY, E. C.: "A History of the Theory of Information," *Proceedings I.E.E.*, 1951, **98**, Part III, p. 383.

(3) JEFFREYS, H.: "Theory of Probability" (Oxford University Press, 1939).

(4) SHANNON, C. E.: "Communication in the Presence of Noise," *Proceedings of the Institute of Radio Engineers*, 1949, **37**, p. 10.

(5) WOODWARD, P. M., and DAVIES, I. L.: "A Theory of Radar Information," *Philosophical Magazine*, 1950, **41**, p. 1001.

(6) Brillouin, L.: "Maxwell's Demon Cannot Operate: Information and Entropy. I," *Journal of Applied Physics*, 1951, **22,** p. 334.

(7) Brillouin, L.: "Physical Entropy and Information. II," *ibid.*, p. 338.

(8) Woodward, P. M.: "Information Theory and the Design of Radar Receivers." *Proceedings of the Institute of Radio Engineers*, 1951, **39,** p. 1521.

(9) Bartlett, M. S.: "Probability and Chance in the Theory of Statistics," *Proceedings of the Royal Society*, A, 1933, **141**, p. 518.

(9) APPENDIX. The form of $J[p(x_i), p_y(x_i)]$

The function $J[p(x_i), p_y(x_i)]$ may be written more simply $J(\xi, \eta)$, a function of the two variables ξ and η. Then the identity (7) states that

$$J(\alpha, \beta) + J(\beta, \gamma) \equiv J(\alpha, \gamma) \quad . \quad . \quad . \quad (33)$$

where α, β and γ are particular values of the variables ξ and η. It is assumed that $J(\xi, \eta)$ is differentiable with respect to ξ. Then, by considering an increase of α to $\alpha + \delta\alpha$ in eqn. (33),

$$\left[\frac{\partial J(\xi, \eta)}{\partial \xi}\right]_{(\alpha, \beta)} \equiv \left[\frac{\partial J(\xi, \eta)}{\partial \xi}\right]_{(\alpha, \gamma)}$$

Since this identity is true for all values of β and γ, $\partial J(\xi, \eta)/\partial \xi$ is independent of η. Integration with respect to ξ therefore yields

$$J(\xi, \eta) \equiv j(\xi) + k(\eta) \quad . \quad . \quad . \quad . \quad (34)$$

where $j(\xi)$ is independent of η and $k(\eta)$ is independent of ξ. But on substituting from eqn. (34) into eqn. (33),

$$j(\beta) \equiv -k(\beta)$$

and hence
$$J(\xi, \eta) \equiv j(\xi) - j(\eta) \quad . \quad . \quad . \quad . \quad (35)$$

Eqn. (9) may now be written in the form

$$j(\alpha\beta) - j(\gamma\delta) \equiv j(\alpha) + j(\beta) - j(\gamma) - j(\delta)$$

Keeping γ and δ constant,

$$j(\alpha\beta) \equiv j(\alpha) + j(\beta) + \text{constant} \quad . \quad . \quad . \quad (36)$$

If $j(\xi)$ is again assumed to be differentiable, considering a small change of α in eqn. (36) gives

$$\beta\left[\frac{dj(\xi)}{d\xi}\right]_{(\alpha\beta)} \equiv \left[\frac{dj(\xi)}{d\xi}\right]_{(\alpha)}$$

Putting α equal to unity,

$$\left[\frac{dj(\xi)}{d\xi}\right]_{(\beta)} \equiv \frac{A}{\beta} \quad . \quad . \quad . \quad . \quad . \quad . \quad (37)$$

where A is a constant equal to $j'(1)$. Since eqn. (37) is an identity,

$$j'(\xi) = A/\xi$$

and hence
$$j(\xi) = A \log \xi + B \quad . \quad . \quad . \quad . \quad (38)$$

where A and B are arbitrary constants, which is the required result quoted in eqn. (10).